การวิเคราะห์เชิงตัวเลข

ผ่านสิ่งที่เราเห็นบนหน้าจอ

จากตัวเลขในเครื่องสู่โลกที่เคลื่อนไหว

ฉบับร่าง · preprint · 2026-10-23



เอกชัย ทวินันท์

การวิเคราะห์เชิงตัวเลขผ่านสิ่งที่เราเห็นบนหน้าจอ

จากตัวเลขในเครื่องสู่โลกที่เคลื่อนไหว

ฉบับร่าง · Preprint · 2026-10-23

© 2026 เอกชัย ทวินันท์ สงวนลิขสิทธิ์ · ฉบับร่างเผยแพร่ให้อ่านฟรี ·
ห้ามคัดลอก ดัดแปลง หรือเผยแพร่ต่อโดยไม่ได้รับอนุญาต

https://ekktha-home.pages.dev

สารบัญ

I คอมพิวเตอร์กับตัวเลข **1**

1 ตัวเลขในเครื่อง **2**

1.1 เครื่องเก็บตัวเลขอย่างไร 3

1.2 ปัดเศษ ขอบเขต และตัวเลขที่เต็ม 14

2 เมื่อความคลาดเคลื่อนเล็ก ๆ โตขึ้น **27**

2.1 สะสมทีละนิด 28

2.2 ลบกันแล้วความแม่นหายไป 38

2.3 ปัญหาไม่ดีหรือวิธีไม่ดี 47

II จากความต่อเนื่องสู่ข้อมูลไม่ต่อเนื่อง **59**

3 เชื่อมจุด **60**

3.1 ค่าตรงกลางอยู่ที่ไหน 62

3.2 ผิดไปเท่าไร และทำไมคีย์เฟรมมากไม่ได้แปลว่าดีกว่า 72

3.3 ต่อเส้นให้เรียบ 84

4 เส้นที่เหมาะกับข้อมูลที่สุด **97**

4.1 เส้นตรงที่ดีที่สุด 98

4.2 เลือกฐานเชิงตั้งฉาก 108

4.3 คลื่นที่ซ่อนอยู่ในเสียงและภาพ 118

5 อนุพันธ์และพื้นที่จากข้อมูลไม่กี่จุด **132**

5.1 ความเร็วจากตำแหน่ง 133

5.2 ระยะทางจากความเร็ว 143

5.3 เลือกจุดให้ฉลาด 155

5.4 รู้ได้อย่างไรว่าแม่นพอ 167

III การแก้สมการ **178**

6 หนึ่งสมการ หนึ่งตัวไม่รู้ **179**

6.1 ล้อมคำตอบไว้ในกรง 180

ฉบับร่าง · Preprint · 2026-10-23

6.2 แทนด้วยเส้นสัมผัส . . . 191
6.3 เมื่อไม่รู้จะเริ่มตรงไหน . . . 203

7 สมการเชิงเส้นหลายตัว 215
7.1 ตัวไม่ทราบหลายตัวพร้อมกัน . . . 216
7.2 ทำไมต้องสลับแถว และทำไมภาพเบลอแก้คืนยาก . . . 226
7.3 เมื่อเมทริกซ์ส่วนใหญ่เป็นศูนย์ . . . 239
7.4 กลับไปสมการไม่เชิงเส้น: หลายตัวแปร . . . 250

8 ข้อมูลมากกว่าตัวไม่รู้ 261
8.1 คำตอบที่พอดีที่สุดของระบบที่มีสมการเกิน . . . 262
8.2 ข้อมูลที่ซ้ำซ้อนกัน . . . 274

9 ทิศทางพิเศษของเมทริกซ์ 284
9.1 ทิศทางที่ไม่เปลี่ยน . . . 285
9.2 ยกกำลังซ้ำ ๆ . . . 296
9.3 ขั้นตอนวิธี QR: หาทุกโหมดพร้อมกัน . . . 305
9.4 ใช้เพียงไม่กี่ทิศ . . . 316

IV ระบบขนาดใหญ่ 328

10 เมื่อเมทริกซ์ใหญ่เกินกว่าจะแยกตัวประกอบ 329
10.1 จากควันสู่เมทริกซ์ . . . 330
10.2 เฉลี่ยกับเพื่อนบ้านซ้ำ ๆ . . . 339
10.3 มองจากไกลก่อน . . . 351

11 ลงหุบเขาให้ฉลาดขึ้น 361
11.1 ซิกแซกลงเขา . . . 362
11.2 เร็วแค่ไหน . . . 373
11.3 ปรับโจทย์ให้ง่ายขึ้น . . . 381
11.4 เมื่อภูมิประเทศไม่สมมาตร . . . 389

V สิ่งที่เปลี่ยนไปตามเวลา 400

12 ก้าวไปตามเวลา 401
12.1 ตามลูกศรไปทีละก้าว . . . 403
12.2 ก้าวให้นิ่งขึ้น . . . 410
12.3 ใช้ตำแหน่งก่อนหน้า . . . 421

13 ก้าวให้ฉลาดขึ้น 432
13.1 ลองหลายจุดต่อหนึ่งก้าว . . . 434
13.2 ให้เครื่องเลือกขนาดก้าวเอง . . . 443

ฉบับร่าง · Preprint · 2026-10-23

14 เมื่อสปริงแข็งเกินไป **453**
14.1 ทำไมก้าวต้องเล็กจนทำอะไรไม่ได้ 454
14.2 ก้าวแบบโดยปริยายและสูตร BDF 464
14.3 วิธีรุงเงอ-คุททาแบบโดยปริยาย 472

VI ข้างหน้า **481**

15 คลื่นบนน้ำและความร้อนที่แผ่ออก **482**
15.1 ริ้วคลื่นและก้าวเวลาที่ปลอดภัย 483
15.2 ความร้อนและภาพเบลอ 492
15.3 ชิ้นเล็ก ๆ ประกอบกัน 501

ดรรชนี (ไทย – English) **516**

ดรรชนี (English – ไทย) **522**

ฉบับร่าง · Preprint · 2026-10-23

ภาค I

คอมพิวเตอร์กับตัวเลข

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 1

ตัวเลขในเครื่อง

สถานการณ์สมมติ: เดินไปไกลแล้วตัวละครไม่ขยับ

คุณคงเคยเล่นเกมโลกเปิดที่เดินไปได้เรื่อย ๆ โดยไม่มีขอบ ลองสมมติว่าเกมนับหน่วยเป็นเมตร และตัวละครของเรามาอยู่ไกลจากจุดกำเนิดของโลกหนึ่งพันกิโลเมตร คือที่พิกัด $x = 1{,}000{,}000$ เมตร ภาพบนจอยังดูปกติทุกอย่าง

ทีนี้ลองนึกว่าโค้ดการเดินเป็นชุดเดียวกับที่ใช้ตรงจุดกำเนิดทุกบรรทัด คือทุกเฟรมเอาความเร็วคูณกับเวลาหนึ่งเฟรม แล้วบวกเข้ากับตำแหน่งเดิม ถ้าเดินเฟรมละ 1 เซนติเมตร (0.01 เมตร) ที่ 60 เฟรมต่อวินาที ตัวละครที่จุดกำเนิดจะไปได้ประมาณ 0.6 เมตรใน 60 เฟรม ตามที่ควรเป็น แต่ตัวละครที่ $x = 10^6$ จะ **ยืนนิ่ง** ไม่ขยับเลยสักเฟรม ทั้งที่เราสั่งให้มันเดิน

ถ้าเร่งเป็นเฟรมละ 4 เซนติเมตร ตัวละครที่ไกลออกไปจะเดินได้ แต่เดินเป็นช่วง ๆ ทุกก้าวกระโดด 6.25 เซนติเมตร และได้ระยะ 3.75 เมตร ทั้งที่ควรได้ 2.4 เมตร คือเร็วเกินจริงราว 1.5625 เท่า ผู้เล่นจะเห็นตัวละครกระตุกเป็นจังหวะ ทั้งที่โค้ดเหมือนกันทุกอย่างยกเว้นตำแหน่งเริ่มต้น

สมมติว่าเกมเก็บพิกัดเป็น**เลขความแม่นเดี่ยว** (single precision) ซึ่งใช้ 32 บิตต่อหนึ่งตัว (ในโค้ดเรียกว่า `float32`) ที่พิกัด 10^6 เลขสองตัวที่เก็บได้ติดกันห่างกันเพียง 0.0625 เมตร ไม่มีเลขตัวใดอยู่ตรงกลาง เอกสารของ Epic Games เกี่ยวกับ Unreal Engine บอกว่า ตัวเลขทศนิยมชนิด 32 บิตจำกัดขนาดของโลกที่ทำได้ และ Unreal Engine 5 เพิ่มตัวเลข 64 บิตสำหรับโลกที่ใหญ่กว่า Games 2026

ตัวเลขทั้งหมดในสถานการณ์นี้มาจากโค้ดของเราเอง ไม่ได้มาจากเกมใดเกมหนึ่ง และเอกสารของ Epic ไม่ได้บอกว่าตัวละครในเกมจะหยุดที่ระยะเท่าใด

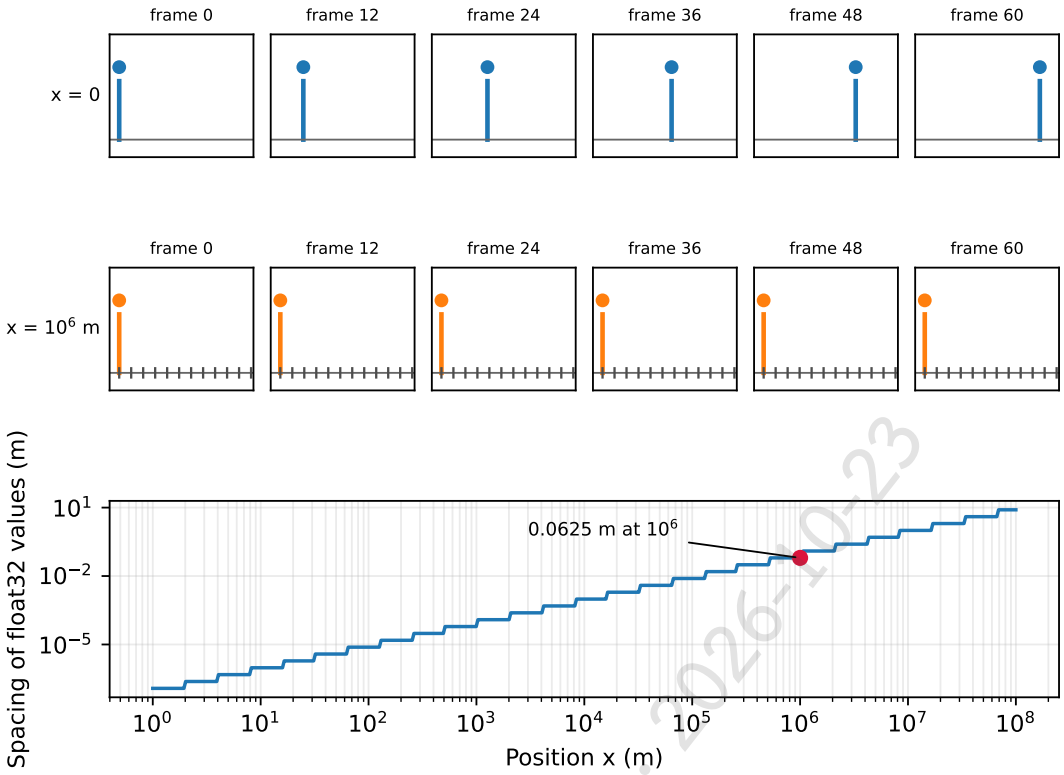
คำถามนำของบทนี้มีข้อเดียวคือ ทำไมระยะห่างระหว่างตัวเลขที่เครื่องเก็บได้จึงกว้างขึ้นเมื่ออยู่ไกลจากจุดกำเนิด

ฉบับร่าง · Preprint · 2026-10-28

ภาพรวมบท

บทนี้มีสี่ความคิด

- ตัวเลขในเครื่องเป็น**เลขจุดลอยตัว** (floating-point number) คือเลขฐานสองที่มีจำนวนบิตจำกัด และจุดฐานลอยไปตามขนาดของเลข
- เลขที่เก็บได้เรียงกันเป็นช่วง ๆ ตามกำลังของ 2 ในแต่ละช่วงช่องว่างเท่ากัน แต่ช่องว่างจะกว้างขึ้นเมื่อพิกัดใหญ่ขึ้น
- จำนวนส่วนใหญ่ไม่ได้อยู่บนช่องเหล่านี้พอดี เราจึงต้องปัดไปยังเลขที่เก็บได้ตัวที่ใกล้ที่สุด ทำให้เกิดความคลาดเคลื่อนเล็ก ๆ ทุกครั้ง



รูปที่ 1.1: (ก) แถบเฟรมของตัวละครสองตัวที่เดินเฟรมละ 0.01 เมตร แถวบนเริ่มที่ $x = 0$ แถวล่างเริ่มที่ $x = 10^6$ ขีดเล็กบนพื้นแถวล่างคือตำแหน่งที่เก็บได้ ห่างกัน 0.0625 เมตร (ข) ระยะห่างระหว่างเลขความแม่นเดี่ยวที่ติดกันเป็นฟังก์ชันของพิกัด x กราฟเป็นขั้นบันไดทุกครั้งที่ x ข้ามเลขยกกำลังของ 2 จุดสีแดงคือ $x = 10^6$ (สร้างจาก `numanal/ch01.py`)

ฉบับร่าง Preprint · 2026-10-23

- ความคลาดเคลื่อนเล็ก ๆ เหล่านี้สะสมได้ และในบางการคำนวณอาจขยายขึ้นมาก

เราเริ่มจากตัวละครที่ยืนนิ่งไกลจากจุดกำเนิด แล้วจะเจออีกด้านหนึ่งของปัญหาเดียวกัน คือกำแพงบางที่ปลายดาบทะลุไปได้ ในหัวข้อ 1.1 เราดูว่าเลขที่เก็บได้หน้าตาเป็นอย่างไร ส่วนหัวข้อ 1.2 พูดถึงการปัดเศษ และขอบเขตของตัวเลขที่เล็กหรือใหญ่เกินไป

1.1 เครื่องเก็บตัวเลขอย่างไร

ภาพรวมก่อนอ่าน

ลองนึกถึงไม้บรรทัดที่มีขีดให้อ่านเพียงบางจุด ถ้าเราวัดระยะที่อยู่ระหว่างขีดสองขีด เราต้องปัดไปยังขีดที่ใกล้ที่สุด เครื่องคอมพิวเตอร์ก็ทำแบบเดียวกัน แต่ไม้บรรทัดของมันไม่ได้มีขีดห่างเท่ากันทุกตำแหน่ง ยิ่งพิกัดใหญ่ ขีดยิ่งห่างกัน ที่ $x = 10^6$ เมตร ขีดห่างกัน 6.25 เซนติเมตร รูปด้านล่างแสดงตัวละครสองตัวเป็นแถบเฟรม และระยะห่างของขีดทั้งหมด

ตารางที่ 1.1: ระยะห่างของตัวเลขที่เก็บได้ในสองแบบ (ตัวเลขจากโค้ดของเรา)

ตำแหน่ง x	เลขจุดคงที่ (8 บิตหลังจุด)	เลขความแม่นเดี่ยว
$x = 1$	0.00390625	$2^{-23} \approx 1.19 \times 10^{-7}$
$x = 10^6$	0.00390625	0.0625
$x = 2^{23}$	0.00390625	1

ตัวเลขฐานสอง

เครื่องคอมพิวเตอร์ใช้**เลขฐานสอง** (binary number) ซึ่งมีเพียงสองหลักคือ 0 กับ 1 ทุกตำแหน่งมีค่าเป็นกำลังของ 2 เหมือนเลขฐานสิบที่ทุกตำแหน่งเป็นกำลังของ 10 ตัวอย่างเช่น $1010.11_2 = 8 + 2 + \frac{1}{2} + \frac{1}{4} = 10.75$ แต่ละหลักของเลขฐานสองเรียกว่า**บิต** (bit) คือหน่วยที่เล็กที่สุดของการเก็บข้อมูล เลขในเครื่องจึงมีจำนวนบิตจำกัด

ปัญหาแรกคือเลขบางตัวที่สั้นในฐานสิบกลับยาวไม่รู้จบในฐานสอง เลข 0.1 เขียนในฐานสองได้เป็น

$$0.1 = 0.0001\,1001\,1001\,1001\ldots_2$$

หลักหลังจุดวนซ้ำ 1001 ไม่รู้จบ เราจึงไม่มีทางเขียน 0.1 ด้วยบิตจำนวนจำกัดให้แม่นเป๊ะ (พิสูจน์ไว้ในส่วนพิสูจน์ด้านล่าง)

เลขจุดคงที่กับเลขจุดลอยตัว

ก่อนดูรูปแบบของเลขความแม่นเดี่ยว ลองพิจารณาวิธีที่ตรงที่สุดก่อน คือให้จุดฐานอยู่ตำแหน่งเดียวกันเสมอ เช่นให้ทุกตัวมี 8 บิตหลังจุดฐานสอง วิธีนี้เรียกว่า**เลขจุดคงที่** (fixed-point number) ซึ่งหมายถึงเลขที่จุดฐานไม่ขยับตามขนาดของตัวเลข

เลขจุดคงที่ 32 บิตที่มี 8 บิตหลังจุดฐานมีระยะห่างระหว่างเลขที่เก็บได้เท่ากันทุกที่ คือ $2^{-8} = 1/256 = 0.00390625$ และเก็บค่าได้ตั้งแต่ประมาณ -8.4 ล้านถึง 8.4 ล้านเมตร ที่ $x = 10^6$ ระยะห่างนี้ละเอียดกว่าเลขความแม่นเดี่ยวถึง 16 เท่า แต่วิธีนี้ก็มีข้อเสีย ลองให้ตัวละครที่ไกลเดินเฟรมละ 0.01 เมตร ก้าวนี้จะถูกปัดเป็น $3/256 = 0.01171875$ เมตร ตัวละครจึงไม่ติดนิ่ง แต่เร็วกว่าที่ควรราว 17 เปอร์เซ็นต์ทุกเฟรม

เลขจุดคงที่ให้ช่องว่างเท่ากันทุกที่ แต่ช่วงที่เก็บได้แคบ ส่วนเลขจุดลอยตัวให้ช่วงกว้างมากโดยที่ช่องว่างเป็นสัดส่วนกับขนาดของ x สำหรับเลขปกติในรูปแบบความแม่นเดี่ยว ช่วงนั้นอยู่ตั้งแต่ราว 1.2×10^{-38} ถึง 3.4×10^{38} Atkinson 1989, §1.2, pp. 11--16 และช่องว่างอยู่ระหว่าง 2^{-24} ถึง 2^{-23} เท่าของ x เสมอ (แบบฝึกหัด 1.1.6) จึงเป็นเหตุผลที่เราเลือกเลขจุดลอยตัว แม้ต้องยอมรับว่าช่องว่างจะไม่คงที่

ลองเอง: เลขจุดคงที่ต้องใช้กี่บิต

ลองให้เลขจุดคงที่เก็บตั้งแต่ 0 ถึง 10^6 เมตร โดยมีช่องห่างละเอียด 10^{-6} เมตร เพื่อให้ตัวละครเดินช้า ๆ ได้ จำนวนช่องคือ 10^{12} ซึ่งเท่ากับ $2^{39.86\ldots}$ จึงต้องใช้อย่างน้อย 40 บิตสำหรับค่าที่ไม่ติดลบ ไม่นับบิตเครื่องหมาย ส่วนเลขจุดลอยตัวความแม่นเดี่ยวใช้ 32 บิตทั้งหมด และให้ความแม่นสัมพัทธ์เกือบเท่ากันทั้งช่วง เพราะช่องห่างปรับตามขนาดของตัวเลข (ที่ 10^6 ช่องห่างคือ 0.0625 เมตร ไม่ละเอียดถึง 10^{-6} แต่ใช้ได้กับทุกขนาดด้วยความแม่นสัมพัทธ์)

ส่วนประกอบของตัวเลขที่เก็บ

แนวคิด. แยกเลขที่ไม่เป็นศูนย์ออกเป็นสามส่วน แล้วเก็บแต่ละส่วนในบิตของมันเอง

เครื่องเก็บเลขที่ไม่เป็นศูนย์เป็นสามส่วน คือ เครื่องหมาย (บวกหรือลบ) **ส่วนยกกำลัง** (exponent) ซึ่งบอกว่าเลขอยู่ในช่วงกำลังของ 2 ใด และ **แมนทิสซา** (significand) ซึ่งบอกตำแหน่งที่ละเอียดขึ้นในช่วงนั้น แมนทิสซารวมหลักนำหน้าที่เป็น 1 ด้วย

ร่างพิมพ์ก่อน · Preprint · 2026-10-23

ในเลขความแม่นเดี่ยว บิตที่เก็บจริงทั้ง 23 บิตเรียกว่า**ส่วนเศษ** (fraction field) ซึ่งไม่รวมหลักนำหน้า เราไม่ต้องเก็บหลักนำหน้า เพราะเลขปกติทุกตัวมีหลักนี้เป็น 1 เสมอ ดังนั้น float32 มีหนึ่งบิตเครื่องหมาย แปดบิตส่วนยกกำลัง และยี่สิบสามบิตส่วนเศษ รวมเป็นสามสิบสองบิต แต่แมนทิสซาให้ความแม่นได้ยี่สิบสี่บิต คือยี่สิบสามบิตที่เก็บจริงบวกหลักนำหน้า

แนวคิดของการแทนเลขแบบนี้เป็นแนวคิดมาตรฐาน อธิบายไว้ใน Atkinson 1989, §1.2, pp. 11--16 ส่วนมาตรฐานที่กำหนดรูปแบบบิตของเลขความแม่นเดี่ยวและความแม่นคู่คือ **มาตรฐาน IEEE 754** (IEEE 754) ซึ่งเป็นมาตรฐานที่กำหนดรูปแบบเลขจุดลอยตัว; บทนี้อธิบายโครงสร้างโดยอ้าง Atkinson และตรวจตัวเลขด้วยโค้ดของเรา ตัวอย่างและการพิสูจน์ในบทนี้เราเขียนขึ้นเอง

วิธีเก็บตัวเลข x ทำได้สามขั้น

1. หาเลข e ที่ทำให้ $2^e \le |x| < 2^{e+1}$
2. เขียน $|x|$ ในรูป $2^e(1+f)$ เมื่อ $0 \le f < 1$
3. เก็บเครื่องหมาย e และ f โดยเก็บ f ได้ 23 บิต ถ้า f ยาวกว่านั้น เราต้องปัด ซึ่งจะพูดถึงในหัวข้อ 1.2

ขั้นที่ 2 ทำให้เห็นว่า ตัวเลขที่ใหญ่ขึ้นจะถูกเก็บโดยใช้ e ที่โตขึ้น ส่วนความละเอียดของ f ยังเท่าเดิม

เก็บ 0.1 ด้วยมือ ทีละขั้น

ลองทำสามขั้นข้างบนกับ $x = 0.1$ ทีละบรรทัด เราจะได้เห็นว่ารูปบิตที่เก็บจริงมาจากไหน

1. **หา e.** เพราะ $2^{-4} = 0.0625 \le 0.1 < 0.125 = 2^{-3}$ จึงได้ $e = -4$
2. **หา f.** เขียน $0.1 = 2^{-4}(1+f)$ ได้ $f = 0.1/0.0625 - 1 = 0.6$
3. **เขียน 0.6 ในฐานสอง.** คูณด้วย 2 แล้วเก็บหลักหน้าไว้ทีละหลัก ค่าที่ได้คือ $0.6 \to 1.2 \to 0.4 \to 0.8 \to 1.6 \to \cdots$ หลักที่เก็บคือ $1, 0, 0, 1$ แล้ววนกลับ จึงได้ $0.6 = 0.1001\,1001\,1001\ldots_2$
4. **ตัดให้เหลือ 23 บิต.** บิตที่ 1 ถึง 23 คือ $1001\,1001\,1001\,1001\,1001\,100$ บิตที่ 24 เป็น 1 และส่วนที่เหลือไม่เป็นศูนย์ จึงเกินครึ่งช่องและปัดขึ้น ได้บิตที่ 23 เป็น 1 กลายเป็น $1001\,1001\,1001\,1001\,1001\,101$
5. **ประกอบบิตทั้งหมด.** เครื่องหมายเป็น 0 ส่วนยกกำลังเก็บเป็น $-4 + 127 = 123$ คือ 01111011 ค่าที่เก็บได้คือ $2^{-4}\big(1 + \frac{5\,033\,165}{2^{23}}\big) = 0.100000001490116\ldots$ ซึ่งมากกว่า 0.1 เล็กน้อย ตรงกับข้อควรระวังด้านบน

ลองเอง: ถอดรหัสบิตกลับเป็นตัวเลข

บิตของเลขความแม่นเดี่ยวคือ `0 10000001 01000000000000000000000` ให้แยกเป็นเครื่องหมาย ส่วนยกกำลัง และส่วนเศษก่อน แล้วเขียนค่าด้วย $(-1)^s 2^e(1+f)$ ตัวเลขนี้คืออะไร ผลคือ $e = 129 - 127 = 2$ และ $f = \frac{1}{4}$ จึงได้ $4 \times 1.25 = 5$

ลองเอง: $0.1 + 0.2$ เท่ากับ 0.3 หรือไม่

เดาก่อนว่าผลของ $0.1 + 0.2$ เท่ากับ 0.3 หรือไม่ ในความแม่นคู่ (`float64`) ผลคือ 0.30000000000000004 จึงไม่เท่ากัน แต่ในความแม่นเดี่ยว ผลของ `float32` เท่ากับค่าที่เก็บได้ของ 0.3 พอดี การเทียบค่าด้วยเครื่องหมายเท่ากับจึงให้ผลต่างกันตามรูปแบบที่เลือก

ฉบับร่าง Preprint · 2026-10-23

ตารางที่ 1.2: รูปแบบบิตของเลขความแม่นเดี่ยวและเลขความแม่นคู่

รูปแบบ	บิตเครื่องหมาย	ส่วนยกกำลัง	ส่วนเศษ	ความแม่น (บิต)	รวม
เลขความแม่นเดี่ยว (float32)	1	8	23	24	32
เลขความแม่นคู่ (float64)	1	11	52	53	64

ตารางที่ 1.3: ผลของการเดิน 60 เฟรมด้วยเลขความแม่นเดี่ยว (หน่วยเมตร)

กรณี	พิกัดเริ่ม	ก้าวต่อเฟรม	ขยับหลัง 60 เฟรม	ควรได้	ขยับต่อเฟรม
จุดกำเนิด	0	0.01	0.5999997	0.6	≈ 0.01
ไกลออกไป (ก้าวเล็ก)	10^6	0.01	0	0.6	0
ไกลออกไป (ก้าวใหญ่)	10^6	0.04	3.75	2.4	0.0625

เลขความแม่นคู่

ถ้าต้องการความละเอียดมากขึ้น เราใช้**เลขความแม่นคู่** (double precision) ซึ่งใช้ 64 บิตต่อหนึ่งตัว (ในโค้ดเรียกว่า `float64`) มีโครงสร้างเหมือนกันทุกประการ ต่างกันแค่จำนวนบิตของแต่ละส่วน ตารางด้านล่างเทียบสองรูปแบบตามมาตรฐาน IEEE 754

ไบอัส (bias) คือค่าที่ถูกบวกเพิ่มเข้าไปในส่วนยกกำลังที่เก็บ เพื่อให้เก็บยกกำลังติดลบได้โดยไม่ต้องใช้บิตเครื่องหมายแยก ค่าไบอัสคือ 127 สำหรับความแม่นเดี่ยว และ 1023 สำหรับความแม่นคู่ ดังนั้นยกกำลัง e ถูกเก็บเป็น $e + 127$ หรือ $e + 1023$ ตามรูปแบบ

ลองเอง: เลข 10^6 เก็บอย่างไร

ตัวเลข $10^6 = 1\,000\,000$ อยู่ระหว่าง $2^{19} = 524\,288$ กับ $2^{20} = 1\,048\,576$ ดังนั้น $e = 19$ และ

$$10^6 = 2^{19} \times 1.9073486328125 = 2^{19}\left(1 + \tfrac{7433}{8192}\right).$$

ส่วนเศษ $\frac{7433}{8192}$ คูณด้วย 2^{23} ได้ $7\,611\,392$ ซึ่งเป็นจำนวนเต็มพอดี จึงเก็บในยี่สิบสามบิตได้ทั้งหมด ดังนั้น 10^6 เก็บได้แม่นเป๊ะ ไม่ต้องปัด

ในเลขความแม่นคู่ ยกกำลังยังเป็น 19 แต่เก็บเป็น $19 + 1023 = 1042$ และส่วนเศษคือ $7433 \cdot 2^{39} = 4\,086\,334\,964\,629\,504$ ซึ่งเป็นจำนวนเต็มเช่นกัน ระยะห่างของเลขความแม่นคู่ที่ 10^6 จึงเป็น $2^{-33} \approx 1.16 \times 10^{-10}$ เมตร เล็กกว่าความแม่นเดี่ยวราว 537 ล้านเท่า

ลองเอง: หนึ่งก้าวด้วยมือ

ที่ 10^6 บวก 0.01 ได้ $1\,000\,000.01$ ตอนนี้เลขที่เก็บได้ที่ใกล้ที่สุดคือ 10^6 เองเพราะช่องว่างคือ 0.0625 และครึ่งหนึ่งของช่องว่างคือ 0.03125 ซึ่งมากกว่า 0.01 ผลจึงปัดกลับมาที่ 10^6 พอดี ก้าวนี้หายไปทั้งก้าว

รันหกสิบเฟรม

เรารันตัวละครสามแบบ แต่ละแบบเดิน 60 เฟรม ทุกการบวกทำด้วย `float32` ตัวเลขจาก `numanal/ch01.py`

แถวสุดท้ายอธิบายได้ด้วยเลขจุดลอยตัวที่เราเพิ่งเห็น ก้าว 0.04 เมตรอยู่ใกล้ 0.0625 มากกว่า 0 ตัวละครจึงขยับหนึ่งช่องทุกเฟรม ได้ $60 \times 0.0625 = 3.75$ เมตร และ $3.75/2.4 = 1.5625$ ซึ่งคือความเร็วที่เกินจริงที่เราเห็นในสถานการณ์

ร่างหนังสือ · Preprint · 2026-10-23

สถานการณ์สมมติ: อีกด้านหนึ่งของปัญหาเดียวกัน (ปลายดาบทะลุกำแพง)

เมื่อกี้ตัวละครที่ไกลจากจุดกำเนิดยืนนิ่ง คราวนี้ลองดูอีกด้านหนึ่ง ตัวละครยังอยู่ที่ $x = 1{,}000{,}000$ เมตร และเงื้อดาบเข้าหากำแพงบางหนา 2 เซนติเมตร ผิวหน้าของกำแพงอยู่ที่ $10^6 + 0.93$ เมตร ผิวหลังอยู่ที่ $10^6 + 0.95$ เมตร ปลายดาบเคลื่อนเข้าหากำแพงเฟรมละ 5 มิลลิเมตร และตำแหน่งของปลายดาบในแต่ละเฟรมคำนวณใหม่จากตำแหน่งของตัวละครบวกระยะของดาบ ไม่ได้บวกสะสมไปเรื่อย ๆ

ตัวตรวจการชนแบบตรงไปตรงมาเป็นแบบที่โปรแกรมหลายตัวเริ่มต้นด้วย คือปลายดาบ "อยู่ในกำแพง" ก็ต่อเมื่ออยู่ระหว่างสองผิวจริง ๆ ที่จุดกำเนิด ตัวตรวจเห็นปลายดาบอยู่ในกำแพงสามเฟรม คือเฟรมที่ 27 ถึง 29 แต่ที่ 10^6 เมตร ตัวตรวจไม่เห็นเลยสักเฟรม ปลายดาบทะลุกำแพงไปโดยที่เครื่องไม่รู้ (เราไม่ได้บอกว่าเกมใดทำแบบนี้)

สาเหตุคือช่องว่างระหว่างตัวเลขที่เก็บได้ ที่ 10^6 ช่องว่างคือ 0.0625 เมตร ผิวทั้งสองจึงถูกปัดไปเป็นตัวเลขตัวเดียวกัน คือ $10^6 + 0.9375$ กำแพงหนา 2 เซนติเมตรจึงถูกเก็บเป็นกำแพงที่หนาศูนย์ และตำแหน่งของปลายดาบก็เป็นได้เฉพาะจำนวนทวีคูณของ 0.0625 ไม่มีตำแหน่งที่เก็บได้ตำแหน่งใดอยู่ระหว่างผิวทั้งสองเลย นี่ไม่ใช่เรื่องความเร็วหรือการชนที่พลาดไประหว่างเฟรม (บทที่ 12) และไม่ใช่ผิวที่สั่นเพราะความลึกของภาพ (บทที่ 2) นี่คือเรื่องความแม่นของตัวเลข

ปัญหาเดียวกันยังทำให้การวัดระยะผิดไปด้วย ลองวัดช่องว่างจากปลายดาบถึงผิวหน้า ซึ่งอยู่ที่ $10^6 + 0.93$ เมตร กำหนดให้ช่องว่างบวกเมื่อปลายดาบยังอยู่ก่อนผิวหน้า และเป็นลบเมื่อผ่านเข้าไปแล้ว เมื่อช่องว่างจริงเป็น $+0.10, +0.06, +0.04, +0.02, 0, -0.02, -0.04, -0.06, -0.10$ เมตร เครื่องรายงานตามลำดับเป็น $+0.125, +0.0625, +0.0625, 0, 0, 0, -0.0625, -0.0625, -0.0625$ เมตร ตัวเลขทั้งหมดคำนวณจากผิวหน้า ไม่ใช่ผิวหลัง

กำแพงที่หายไปกับตัวละครที่ยืนนิ่งมาจากเหตุเดียวกัน คือช่องว่างระหว่างเลขที่เก็บได้ซึ่งกว้างขึ้นเมื่ออยู่ไกลจากศูนย์ ดังนั้นคำถามของเรื่องนี้จึงไม่ใช่ว่า ทำไมตัวละครจึงติด แต่คือ ทำไมสิ่งหนึ่งจึงอยู่ตรงนั้นได้ ทั้งที่เครื่องมองไม่เห็นมัน

Preprint · 2026-10-23

นิยามและทฤษฎีบท

จากตัวละครที่ยืนนิ่งและกำแพงที่หายไป เราต้องตอบให้ได้ว่าตัวเลขแต่ละตัวที่เครื่องเก็บได้หน้าตาเป็นอย่างไร ตอนนี้เรานิยามเซตของตัวเลขเหล่านั้นให้แม่นยำ สิ่งที่ต้องจำคือรูปร่างของแต่ละเลข ไม่ต้องจำค่าตัวเลขทุกตัว

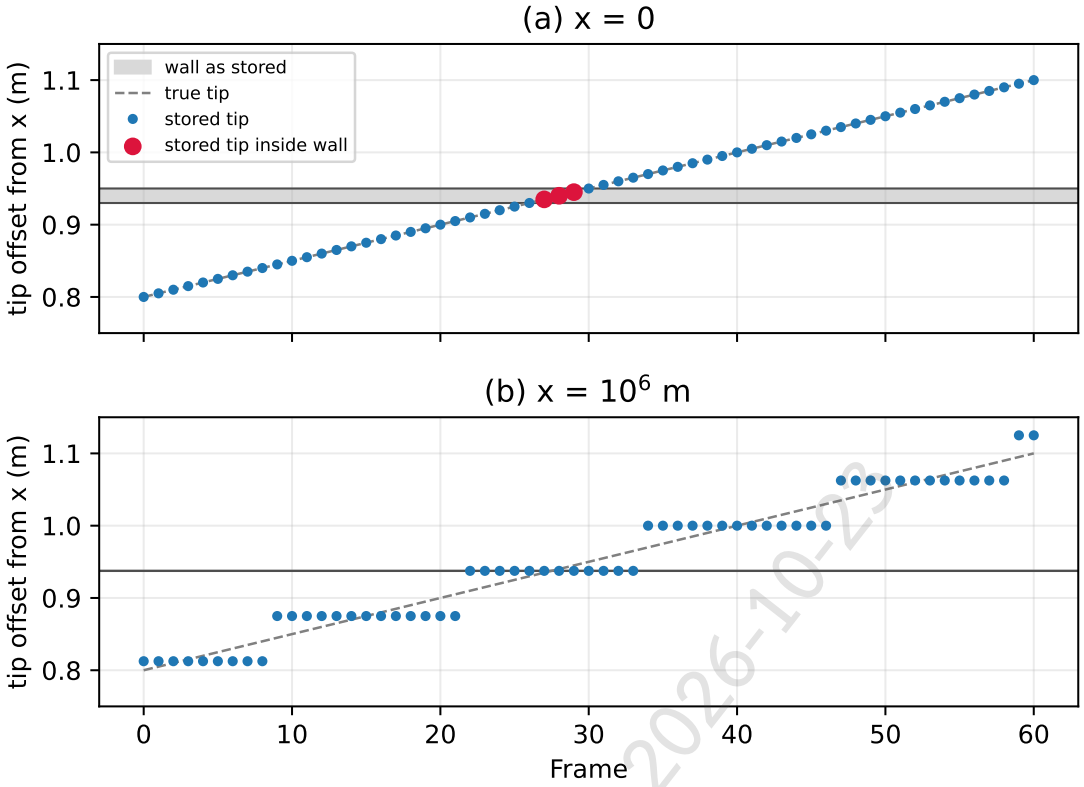
เลขปกติ (normal number) คือเลขที่เก็บแบบเต็มรูปแบบ คือมีหลักนำหน้าเป็น 1 และยกกำลังอยู่ในช่วงที่กำหนด **เลขความแม่นเดี่ยวปกติ** (normal float32) คือเลขปกติที่เก็บในรูปแบบความแม่นเดี่ยว

นิยาม 1.1.1: เลขความแม่นเดี่ยวปกติ

เซต F ของเลขความแม่นเดี่ยวปกติ คือ เซตของจำนวนจริงที่เขียนได้ในรูปนี้

$$x = (-1)^{s}\, 2^{e} \Big(1 + \sum_{i=1}^{23} b_i\, 2^{-i}\Big),$$

เมื่อ $s \in \{0, 1\}$ เป็นเครื่องหมาย e เป็นจำนวนเต็มในช่วง $-126 \le e \le 127$ และ $b_1, \ldots, b_{23} \in \{0, 1\}$ เป็นบิตของส่วนเศษ เซต F นี้ยังไม่รวมเลขศูนย์และเลขพิเศษ ซึ่งเราจะพูดถึงในหัวข้อ 1.2



รูปที่ 1.2: ปลายดาบ 61 เฟรมเทียบกับกำแพง เส้นประคือตำแหน่งจริง จุดสีน้ำเงินคือตำแหน่งที่เก็บได้ จุดสีแดงคือเฟรมที่ตัวตรวจเห็นปลายดาบอยู่ในกำแพง (ก) ที่จุดกำเนิดมี 3 เฟรม (ข) ที่ $x = 10^6$ เมตรไม่มีเฟรมใดเลย (สร้างจาก `numanal/ch01.py`)

ร่างปรับปรุง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

- $(-1)^s$ คือเครื่องหมาย $s = 0$ คือบวก $s = 1$ คือลบ
- 2^e บอกว่าเรากำลังอยู่ในช่วงไหนของขั้นบันได แต่ละช่วงคือ $[2^e, 2^{e+1})$
- $1 + \sum_i b_i 2^{-i}$ คือแมนทิสซา คือหลักนำหน้า 1 บวกกับบิตทั้ง 23 ตัว ส่วน $\sum_i b_i 2^{-i}$ คือส่วนเศษ
- ภายในหนึ่งช่วง ส่วนเศษมี 23 บิต จึงได้เลขเก็บไว้ 2^{23} ตัวเท่ากันพอดี

ตัวอย่าง ในช่วง $[1, 2)$ มีเลขที่เก็บได้ $2^{23} = 8\,388\,608$ ตัว ไม่ใช่ทุกจำนวนจริงในช่วงนั้น ตัวอย่างหนึ่งที่ไม่อยู่ใน F คือ 0.1 เพราะการกระจายฐานสองของมันไม่จบ แม้ว่า 0.1 จะอยู่ใกล้เลขใน F มาก (คือ $0.100000001490116\ldots$)

ทฤษฎีบท 1.1.1: ระยะห่างของเลขที่เก็บได้

ให้ e เป็นจำนวนเต็มในช่วง $-126 \le e \le 126$ เลขในเซต F ที่อยู่ในช่วง $[2^e, 2^{e+1}]$ ห่างกันเท่ากันทุกคู่ที่อยู่ติดกัน และระยะห่างนั้นเท่ากับ 2^{e-23} โดยเฉพาะที่ $x = 10^6$ เรามี $e = 19$ จึงได้ระยะห่างเท่ากับ $2^{-4} = 0.0625$ เมตร

อ่านแบบเข้าใจง่าย

ทฤษฎีบทนี้บอกว่าขั้นบันไดแต่ละขั้นมีความกว้างเท่ากันหมด และขั้นจะกว้างเป็นสองเท่าเมื่อเราข้ามไปยังกำลังของ 2 ถัดไป ช่องว่างจึงขึ้นกับขนาดของตัวเลข ไม่ได้ขึ้นกับตัวเลขตัวนั้นเอง ตัวอย่าง ใกล้ 1 ช่องว่างคือ $2^{-23} \approx 1.19 \times 10^{-7}$ ใกล้ 2^{20} ช่องว่างคือ 0.125

พิสูจน์

แนวคิด. เลขทุกตัวในช่วง $[2^e, 2^{e+1})$ มีรูปร่างเหมือนกัน ต่างกันเพียงตัวคูณ 2^e ดังนั้นเรานับช่องว่างได้จากการนับบิต

1. **เขียนเลขในช่วงให้เป็นรูปเดียว.** ให้ $b_1, \ldots, b_{23}$ เป็นบิตของส่วนเศษ แล้วให้ $k = \sum_{i=1}^{23} b_i 2^{23-i}$ เป็นจำนวนเต็มที่เขียนจากบิตเหล่านั้น บิตยี่สิบสามตัวแทนจำนวนเต็ม $0, 1, \ldots, 2^{23}-1$ ทุกค่า โดยแต่ละค่ามาจากชุดบิตหนึ่งชุดพอดี ดังนั้นเลขใน F ที่อยู่ในช่วงนี้คือ $2^e(1 + k\,2^{-23})$ เมื่อ $k = 0, 1, \ldots, 2^{23}-1$

2. **ระยะห่างระหว่างค่าที่ติดกัน.** สำหรับ $0 \le k < 2^{23}-1$ เรามี

$$2^e\big(1 + (k+1)2^{-23}\big) - 2^e\big(1 + k\,2^{-23}\big) = 2^e \cdot 2^{-23} = 2^{e-23}.$$

   ผลนี้ไม่ขึ้นกับ k จึงระยะห่างเท่ากันทุกคู่ในช่วง

3. **ปลายของช่วง.** ค่าที่มากที่สุดในช่วงคือ $k = 2^{23}-1$ ซึ่งได้ $2^e(2 - 2^{-23}) = 2^{e+1} - 2^{e-23}$ ค่าถัดไปคือ 2^{e+1} ซึ่งเป็นเลขใน F เช่นกันเพราะ $e+1 \le 127$ ระยะห่างคือ $2^{e+1} - (2^{e+1} - 2^{e-23}) = 2^{e-23}$ เท่าเดิม ดังนั้นขั้นบันไดไม่เปลี่ยนที่ขอบช่วง

4. **ใช้กับ 10^6.** เพราะ $2^{19} = 524\,288 \le 10^6 < 1\,048\,576 = 2^{20}$ จึงได้ $e = 19$ และระยะห่างคือ $2^{19-23} = 2^{-4} = 0.0625$

สรุป: ในแต่ละช่วงกำลังของ 2 ช่องว่างเท่ากันทุกช่อง และเมื่อข้ามไปช่วงถัดไป ช่องว่างจะเป็นสองเท่า

ข้อสังเกต (1/10 ไม่มีการกระจายฐานสองจำกัด). ถ้า $\frac{1}{10} = \frac{m}{2^k}$ สำหรับจำนวนเต็ม $m, k \ge 0$ จะได้ $2^k = 10m = 2 \cdot 5 \cdot m$ ดังนั้น 5 หาร 2^k ลงตัว ซึ่งเป็นไปไม่ได้ เพราะตัวประกอบเฉพาะเพียงตัวเดียวของ 2^k คือ 2 จึงไม่มีคู่ (m, k) ดังกล่าว และเลขที่เขียนเป็น $m/2^k$ ได้เท่านั้นที่มีการกระจายฐานสองจำกัด

ในโลกจริง: นาฬิกาของแพทริออต (ดาห์ราน, 1991)

วันที่ 25 กุมภาพันธ์ ค.ศ. 1991 ที่เมืองดาห์ราน ประเทศซาอุดีอาระเบีย ระบบขีปนาวุธแพทริออตไม่สามารถติดตามและสกัดกั้นขีปนาวุธ**สกัดด์** (Scud) ได้ ขีปนาวุธตกที่ค่ายทหารจนทหารอเมริกัน 28 นายเสียชีวิต รายงานของสำนักงานตรวจสอบรัฐบาลสหรัฐฯ (GAO) สรุปว่าสาเหตุคือปัญหาซอฟต์แวร์ นาฬิกาภายในของเครื่องนับเวลาเป็นหนึ่งในสิบวินาทีด้วยจำนวนเต็ม และการแปลงเป็นเวลาจริงถูกจำกัดด้วย**รีจิสเตอร์** (register) ที่มี 24 บิต (ช่องเก็บข้อมูลขนาดเล็กในหน่วยประมวลผล) เมื่อเครื่องเปิดต่อเนื่องเกินหนึ่งร้อยชั่วโมง ความคลาดเคลื่อนเวลาสะสมถึง 0.3433 วินาที และช่วงระยะที่ระบบค้นหาเป้าเลื่อนไป 687 เมตร Office 1992, ภาคผนวก II, น. 15

ระบบนี้ไม่ใช่เลขความแม่นเดี่ยว แต่ใช้หลักการเดียวกัน คือ 0.1 ไม่มีการกระจายฐานสองจำกัด เราลองคำนวณเองว่า ถ้าตัด 0.1 เหลือ 23 หลักหลังจุดฐานสอง ความคลาดเคลื่อนต่อหนึ่งขีดคือราว 9.5×10^{-8} วินาที เมื่อคูณกับ 3,600,000 ขีดที่เกิดใน 100 ชั่วโมง ได้ 0.3433 วินาที ตรงกับตารางของ GAO (การคำนวณนี้เป็นของเราเอง ไม่ใช่ของ GAO)

ร่างพิมพ์ Preprint · 2026-10-23

เกร็ดเพิ่มเติม

เอกสารของ Epic Games ระบุว่าค่า `WORLD_MAX` ของ UE4 คือ 21 กิโลเมตร ซึ่ง UE5 ให้เลือกใช้ได้ผ่านตัวเลือกหนึ่ง ส่วนค่าเริ่มต้นของ UE5 ใหญ่กว่ามาก คือ 88 ล้านกิโลเมตรตามเอกสาร เราเปิดดูเอกสารนี้เมื่อวันที่ 22 ตุลาคม 2569 Games 2026 ชื่อ `float32` และ `float64` ที่ใช้ในโค้ดเป็นชื่อในภาษาโปรแกรม ส่วนรูปแบบบิตที่ใช้จริงกำหนดไว้ใน มาตรฐาน IEEE 754 และเราจะกลับมาดูรายละเอียดเพิ่มในหัวข้อ 1.2

ข้อควรระวัง: ตัวเลขที่พิมพ์ไม่ใช่ตัวเลขที่เก็บ

เมื่อ เรา พิมพ์ 0.1 บน จอ ดู เหมือน เก็บ 0.1 แต่ เลข ความ แม่น เดี่ยว ที่ ใกล้ ที่สุด คือ $0.100000001490116119384765625$ ซึ่ง ไม่ เท่ากับ 0.1 ดัง นั้น การ เทียบ สอง ค่า ด้วย เครื่องหมาย เท่ากับ อาจให้ผลเป็นเท็จ ทั้งที่เราคิดว่าถูกต้อง

ข้อควรระวัง: การบวกที่ไม่เห็นผลไม่ได้แปลว่าไม่มีผล

ที่ 10^6 การบวก 0.01 ทีละเฟรมไม่ทำให้ตัวละครขยับเลย แต่ที่จุดกำเนิดการบวกเดียวกันทำให้ขยับเห็นได้ชัด ตัวเลขเดียวกันมีผลต่างกันตามขนาดของพิกัด จึงต้องดูขนาดของตัวเลขก่อนตัดสินว่าก้าวเล็กหรือใหญ่

ข้อควรระวัง: เอกสารบอกอะไร และเราเห็นอะไร

เอกสารของ Epic บอกเรื่องขนาดโลกของ Unreal Engine เท่านั้น เราไม่รู้ว่าเกมอื่นเก็บพิกัดเป็นตัวเลขชนิดใดภายใน ตัวเลขในสถานการณ์นี้มาจากโค้ดของเราเอง และไม่บอกว่าเกมใดเกมหนึ่งเป็นแบบนี้จริง ตัวละครและกำแพงในสถานการณ์ทั้งสองเป็นของสมมติ

สรุปใจความ

- เครื่องเก็บตัวเลขเป็นเลขฐานสองที่มีจำนวนบิตจำกัด ตัวเลขบางตัว เช่น 0.1 เก็บแม่นเป๊ะไม่ได้
- เลขจุดลอยตัวเรียงเป็นขั้นตามกำลังของ 2 แต่ละช่วงมี 2^{23} เลขสำหรับความแม่นเดี่ยว ช่องว่างกว้างขึ้นเมื่ออยู่ไกลจากศูนย์ ที่ 10^6 ช่องว่างคือ 0.0625 เมตร
- เลขความแม่นคู่ใช้ 64 บิต ส่วนเศษ 52 บิต ช่องว่างจึงเล็กกว่ามาก แต่ยังเป็นขั้นเหมือนกัน
- เลขจุดคงที่มีช่องว่างเท่ากันทุกที่แต่ช่วงแคบ เลขจุดลอยตัวมีช่วงกว้างมากเพราะช่องว่างเป็นสัดส่วนกับ x
- สองสถานการณ์มาจากเหตุเดียวกัน ตัวละครที่ก้าวเล็กกว่าครึ่งช่องยืนนิ่ง และกำแพงหนา 2 เซนติเมตรที่สองผิวอยู่ในช่องเดียวกันถูกเก็บเป็นกำแพงหนาศูนย์

เรากลับไปที่คำถามของบท ระยะห่างกว้างขึ้นเมื่ออยู่ไกลจากศูนย์ เพราะช่องว่างของแต่ละช่วงเป็นสองเท่าทุกครั้งที่ข้ามกำลังของ 2 ส่วนคำถามว่าจะรู้ล่วงหน้าได้อย่างไร จะตอบในหัวข้อ 1.2

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 1.1.1 (⋆) หาระยะห่างระหว่างเลขความแม่นเดี่ยวปกติสองตัวที่อยู่ติดกัน ที่ $x = 1$, $x = 1024$, $x = 2^{20}$ และ $x = 2^{24}$ ตอบเป็นกำลังของ 2

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 1.1.2 (⋆) จงเขียน 0.75 และ 0.2 ในรูปเลขฐานสอง แล้วบอกว่าเลขความแม่นเดี่ยวเก็บแต่ละตัวได้แม่นเป๊ะหรือไม่ อธิบายด้วยรูป $m/2^k$

แบบฝึกหัด 1.1.3 (⋆⋆) ที่ $x = 10^6$ ให้ตัวละครเดินด้วยก้าว d เมตรต่อเฟรม โดย d เป็นเลขความแม่นเดี่ยว และในแต่ละเฟรมเราบวก d เข้ากับตำแหน่ง แล้วปัดผลให้เป็นเลขที่เก็บได้

1. ตัวละครขยับ 0.0625 เมตรในเฟรมเดียว เมื่อ d อยู่ในช่วงใด
2. ที่ $d = 0.03125$ และ $d = 0.09375$ ตัวละครขยับเท่าไร ทำไมจึงต่างจากค่าที่อยู่ใกล้ ๆ

แบบฝึกหัด 1.1.4 (⋆⋆) เลขความแม่นคู่ (`float64`) ใช้ 52 บิตของส่วนเศษ ที่ $x = 10^6$ ระยะห่างของเลขความแม่นคู่เท่าไรในช่วง $[10^6, 10^6 + 1)$ มีเลขที่เก็บได้กี่ตัวในแต่ละแบบ และถ้าตัวละครเดิน 0.01 เมตรต่อเฟรม 60 เฟรมด้วยความแม่นคู่ ระยะที่ได้ต่างจาก 0.6 เมตรเท่าไร ประมาณขอบเขตของความต่างนี้ด้วย

แบบฝึกหัด 1.1.5 (⋆⋆) ลองพิสูจน์ว่า $\frac{3}{10}$ ไม่มีการกระจายฐานสองจำกัด แล้วอธิบายว่าเหตุใด 0.3 จึงเก็บในเครื่องแม่นเป๊ะไม่ได้

แบบฝึกหัด 1.1.6 (⋆ ⋆ ⋆) ให้ x เป็นเลขความแม่นเดี่ยวปกติที่มีค่าบวก และให้ $\mathrm{gap}(x)$ เป็นระยะห่างจาก x ถึงเลขที่เก็บได้ตัวถัดไปที่มากกว่า x ลองพิสูจน์ว่า

$$2^{-24} < \frac{\mathrm{gap}(x)}{x} \leq 2^{-23}$$

และบอกว่าขอบบนเกิดเมื่อใด

ตัวช่วยคิด

ให้ $2^e \leq x < 2^{e+1}$ แล้วใช้ทฤษฎีบทระยะห่างกับ $\mathrm{gap}(x) = 2^{e-23}$ จากนั้นดูว่า x อยู่ระหว่างค่าใดสองค่า

แบบฝึกหัด 1.1.7 (⋆ ⋆ ⋆) ต่อจากสถานการณ์ของบท ตัวละครเดินด้วยเลขความแม่นเดี่ยว

1. ถ้าเดินเฟรมละ 0.01 เมตร และพิกัดอยู่ในช่วง $[2^{17}, 2^{18})$ เมตร ตัวละครขยับเฟรมละเท่าไร
2. ถ้าเดินเฟรมละ 0.01 เมตร ตัวละครเริ่มติดนิ่งทุกเฟรมตั้งแต่พิกัดใด
3. ถ้าเปลี่ยนก้าวเป็น 0.04 เมตร ตัวละครเริ่มติดนิ่งตั้งแต่พิกัดใด

ตัวช่วยคิด

ตัวละครติดนิ่งเมื่อก้าวน้อยกว่าครึ่งหนึ่งของช่องว่างในช่วงนั้น ลองหาว่าช่องว่างในแต่ละช่วงเป็นเท่าไร และดูว่าก้าวของเราเทียบกับครึ่งหนึ่งนั้นอย่างไร

แบบฝึกหัด 1.1.8 (⋆⋆) กลับไปที่สถานการณ์ปลายดาบ เปลี่ยนกำแพงให้หนา 6 เซนติเมตร โดยผิวหน้าอยู่ที่ $10^6 + 0.93$ เมตร และผิวหลังอยู่ที่ $10^6 + 0.99$ เมตร

1. ผิวทั้งสองถูกเก็บเป็นเลขเท่าไร และความหนาที่เก็บได้เป็นเท่าไร
2. ปลายดาบอยู่ที่ $10^6 + 0.80 + 0.005k$ เมตรสำหรับ $k = 0, 1, \ldots, 60$ ในคณิตศาสตร์ที่แม่นเป๊ะ ปลายดาบอยู่ในกำแพงกี่เฟรม และเครื่องเห็นว่าอยู่ในกำแพงกี่เฟรม

แบบฝึกหัด 1.1.9 (⋆ ⋆ ⋆) กำแพงหนา 1 เซนติเมตรมีผิวหน้าที่ a และผิวหลังที่ $a + 0.01$ โดยทั้งสองผิวอยู่ในช่วง $[2^e, 2^{e+1})$ ช่วงเดียวกันของเลขความแม่นเดี่ยว

1. ถ้า $e \leq 16$ จงแสดงว่าผิวทั้งสองถูกเก็บเป็นเลขตัวเดียวกันไม่ได้ ไม่ว่า a จะอยู่ตรงไหนในช่วง
2. จงแสดงว่าที่ $e = 17$ มีค่า a ที่ทำให้ผิวทั้งสองถูกเก็บเป็นเลขตัวเดียวกัน

ฉบับร่าง · Preprint · 2026-10-23

ตัวช่วยคิด

ในช่วง $[2^e, 2^{e+1})$ ระยะห่างของเลขที่เก็บได้คือ 2^{e-23} ถ้าระยะห่างนี้น้อยกว่า 0.01 จะมีจุดกึ่งกลางระหว่างเลขที่เก็บได้สองตัวอยู่ในช่วง $(a, a + 0.01)$ เสมอ

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 1.1.1.

1. $x = 1 = 2^0$ จึงได้ $e = 0$ ระยะห่างคือ $2^{0-23} = 2^{-23}$
2. $x = 1024 = 2^{10}$ จึงได้ $e = 10$ ระยะห่างคือ $2^{10-23} = 2^{-13}$
3. $x = 2^{20}$ จึงได้ $e = 20$ ระยะห่างคือ $2^{20-23} = 2^{-3} = 0.125$
4. $x = 2^{24}$ จึงได้ $e = 24$ ระยะห่างคือ $2^{24-23} = 2$ ดังนั้นจำนวนเต็มที่ต่างกันหนึ่ง เช่น 2^{24} กับ $2^{24} + 1$ ไม่สามารถแยกเก็บได้อีกต่อไป

ตรวจด้วยการคำนวณ: `test_ch01.py`, `test_ex_1_1_1_spacings`.

เฉลย 1.1.2.

1. $\frac{3}{4} = \frac{1}{2} + \frac{1}{4}$ จึงเขียนเป็น 0.11_2 ได้ เป็นรูป $m/2^k$ โดย $m = 3$ และ $k = 2$ ดังนั้น 0.75 เก็บแม่นเป๊ะ
2. $\frac{1}{5} = 0.0011\,0011\,0011\ldots_2$ เพราะ $2^k = 5m$ เป็นไปไม่ได้ (ดูข้อสังเกตในเนื้อหา) ดังนั้น 0.2 ไม่มีการกระจายฐานสองจำกัด
3. ค่าที่เก็บจริงของ 0.2 ในความแม่นเดี่ยวคือ $0.200000002980232238769531 25$ ซึ่งไม่เท่ากับ 0.2 จึงเก็บไม่แม่นเป๊ะ

ตรวจด้วยการคำนวณ: `test_ex_1_1_2_terminating_and_not` ค่าที่เก็บจริงของ 0.2 คำนวณด้วย `Decimal` จากค่า float32 ที่ได้

เฉลย 1.1.3. ให้ $g = 0.0625$ เป็นช่องว่างที่ 10^6 เลขที่เก็บได้ใกล้ 10^6 คือ $10^6 + jg$ สำหรับจำนวนเต็ม j เมื่อบวก d เข้าไป ผลคือ $10^6 + d$ และเราต้องปัดไปยังจุดที่ใกล้ที่สุด

1. ผลไม่เปลี่ยนเมื่อ $d < g/2 = 0.03125$ และขยับหนึ่งช่องเมื่อ $g/2 < d < 3g/2 = 0.09375$ ดังนั้นตัวละครขยับ 0.0625 เมตรในเฟรมเดียวเมื่อ $0.03125 < d < 0.09375$
2. ที่ $d = 0.03125$ เป็นจุดกึ่งกลางพอดีระหว่าง 10^6 กับ $10^6 + g$ เครื่องปัดจุดกึ่งกลางไปยังเลขที่ส่วนเศษเป็นเลขคู่ 10^6 คือ $16\,000\,000\,g$ ซึ่งเป็นเลขคู่ จึงอยู่ที่เดิม ตัวละครไม่ขยับเลย ที่ $d = 0.09375$ จุดนั้นเป็นจุดกึ่งกลางระหว่าง $16\,000\,001\,g$ กับ $16\,000\,002\,g$ เครื่องเลือกเลขคู่คือ $16\,000\,002\,g$ จึงขยับ $2g = 0.125$ เมตร ไม่ใช่ g เหมือนค่าที่อยู่ใกล้ ๆ

ตรวจด้วยการคำนวณ: `test_ch01.py`, `test_ex_1_1_3_one_step_at_one_million` ทดสอบค่าสองปลายด้วยตัวเอง และทดสอบ d อีก 2000 ค่าที่กระจายในช่วง $(0, 0.2)$

เฉลย 1.1.4.

1. เลขความแม่นคู่มีส่วนเศษ 52 บิต ที่ 10^6 ซึ่งอยู่ใน $[2^{19}, 2^{20})$ ระยะห่างคือ $2^{19-52} = 2^{-33} \approx 1.16 \times 10^{-10}$

ร่างต้น · Preprint · 2026-10-23

2. ช่วง $[10^6, 10^6 + 1)$ มีความยาว 1 ดังนั้นจำนวนเลขที่เก็บได้คือ $1/$ระยะห่าง ความแม่นเดี่ยวได้ $1/0.0625 = 16$ ความแม่นคู่ได้ $2^{33} = 8\,589\,934\,592$

3. การบวกแต่ละครั้งปัดผลให้ใกล้ที่สุด ผิดได้ไม่เกินครึ่งหนึ่งของระยะห่าง 2^{-34} เมื่อทำ 60 ครั้ง ค่าคลาดเคลื่อนสัมบูรณ์รวมไม่เกิน $60 \cdot 2^{-34} \approx 3.5 \times 10^{-9}$ เมื่อคำนวณจริงได้ 0.6000000005588 ซึ่งต่างจาก 0.6 ประมาณ 5.6×10^{-10} ซึ่งอยู่ในขอบเขตนี้

ตรวจด้วยการคำนวณ:

`test_ex_1_1_4_float64_gap_and_count`

`test_float64_walk_is_close_to_exact` ค่าความต่างรวมคำนวณด้วยโค้ด

เฉลย 1.1.5. สมมติว่า $\frac{3}{10} = \frac{m}{2^k}$ สำหรับจำนวนเต็ม $m, k \geq 0$ คูณสองข้างด้วย $10 \cdot 2^k$ จะได้ $3 \cdot 2^k = 10m = 2 \cdot 5 \cdot m$ ถ้า $k = 0$ จะได้ $3 = 10m$ ซึ่งไม่มีคำตอบเป็นจำนวนเต็ม ถ้า $k \geq 1$ หารด้วย 2 ได้ $3 \cdot 2^{k-1} = 5m$ ดังนั้น 5 หาร $3 \cdot 2^{k-1}$ ลงตัว แต่ 5 เป็นจำนวนเฉพาะที่ไม่หาร 2 และไม่หาร 3 จึงขัดแย้ง ดังนั้นไม่มี m, k ดังกล่าว

เลข 0.3 จึงไม่อยู่ในเซต F ของเลขความแม่นเดี่ยวปกติ เพราะเลขทุกตัวใน F มีรูป $m/2^k$ และการกระจายฐานสองของ 0.3 คือ $0.0100\,1100\,1100\ldots_2$ ซึ่งไม่จบ ตรวจด้วยการคำนวณ: `test_ex_1_1_5_three_tenths_never_terminates` ทดสอบว่า $\frac{3}{10} \cdot 2^k$ ไม่เป็นจำนวนเต็มสำหรับ k ตั้งแต่ 1 ถึง 60 และทดสอบ 12 หลักแรกของการกระจาย

เฉลย 1.1.6. ให้ x เป็นเลขความแม่นเดี่ยวปกติที่เป็นบวก มี e ที่ทำให้ $2^e \leq x < 2^{e+1}$ จากทฤษฎีบทระยะห่าง $\mathrm{gap}(x) = 2^{e-23}$ ดังนั้น

1. ขอบล่าง: เพราะ $x < 2^{e+1}$ จึงได้ $\dfrac{\mathrm{gap}(x)}{x} > \dfrac{2^{e-23}}{2^{e+1}} = 2^{-24}$

2. ขอบบน: เพราะ $x \geq 2^e$ จึงได้ $\dfrac{\mathrm{gap}(x)}{x} \leq \dfrac{2^{e-23}}{2^e} = 2^{-23}$

3. ขอบบนเกิดเมื่อ $x = 2^e$ พอดี เพราะ $x \geq 2^e$ จะเท่ากันก็ต่อเมื่อ $x = 2^e$ ส่วนขอบล่างไม่มีวันเกิด เพราะ $x < 2^{e+1}$ เสมอ

ตรวจด้วยการคำนวณ: `test_ex_1_1_6_relative_spacing_bounds` สุ่มเลขด้วย seed สามค่า และตรวจว่า $x = 2^{10}$ ให้ค่าบนพอดี

เฉลย 1.1.7.

1. ในช่วง $[2^{17}, 2^{18})$ ช่องว่างคือ $2^{17-23} = 2^{-6} = 0.015625$ ครึ่งช่องคือ 0.0078125 ก้าว 0.01 มากกว่าครึ่งช่องและน้อยกว่า 1.5 ช่อง (0.0234375) จึงขยับหนึ่งช่องพอดี คือ 0.015625 เมตรต่อเฟรม (เร็วกว่า 0.01 เท่ากับ 1.5625 เท่า เช่นเดียวกับในสถานการณ์)

2. ตัวละครติดนิ่งเมื่อ 0.01 น้อยกว่าครึ่งช่อง นั่นคือช่องว่างต้องมากกว่า 0.02 ช่องว่างเป็นกำลังของ 2 ดังนั้นต้องเป็น $0.03125 = 2^{-5}$ ซึ่งเกิดในช่วง $[2^{18}, 2^{19})$ ครึ่งช่องคือ $0.015625 > 0.01$ จึงติดนิ่ง คำตอบคือ $x = 2^{18} = 262\,144$ เมตร และยังติดนิ่งต่อไปเพราะช่องว่างโตขึ้นเรื่อย ๆ

3. ช่องว่างต้องมากกว่า 0.08 เพื่อให้ครึ่งช่องมากกว่า 0.04 ช่องว่างเป็นกำลังของ 2 ที่เกินเกณฑ์นี้เป็นครั้งแรกคือ $0.125 = 2^{-3}$ ซึ่งเกิดเมื่อ $e = 20$ คำตอบคือ $x = 2^{20} = 1\,048\,576$ เมตร เพราะในช่วง $[2^{19}, 2^{20})$ ที่ช่องว่าง 0.0625 ครึ่งช่องคือ $0.03125 < 0.04$ ตัวละครยังขยับได้ 0.0625 เมตรต่อเฟรม

ตรวจด้วยการคำนวณ:

`test_ex_1_1_7_slow_walker_stops_at_2_pow_18`

ฉบับร่าง Preprint 2026-10-23

`test_ex_1_1_7_part_b_fast_walker_stops_at_2_pow_20` ทดสอบสุ่มตำแหน่งในแต่ละช่วงด้วย seed สองค่า

เฉลย 1.1.8.

1. ผิวหน้า $10^6+0.93$ ใกล้ $10^6+0.9375$ (ห่าง 0.0075) มากกว่า $10^6+0.875$ (ห่าง 0.055) จึงเก็บเป็น $10^6+0.9375$ ผิวหลัง $10^6+0.99$ ใกล้ 10^6+1 (ห่าง 0.01) มากกว่า $10^6+0.9375$ (ห่าง 0.0525) จึงเก็บเป็น 10^6+1 ความหนาที่เก็บได้คือ $1-0.9375=0.0625$ เมตร ไม่ใช่ 0.06 เมตร เพราะกำแพงยังหนาอยู่หนึ่งช่องว่าง

2. ในคณิตศาสตร์ที่แม่นเป๊ะ $0.93 < 0.80+0.005k < 0.99$ เทียบเท่ากับ $26 < k < 38$ ดังนั้น $k = 27, 28, \ldots, 37$ รวม 11 เฟรม ส่วนเครื่อง ตำแหน่งที่เก็บได้ของปลายดาบที่ 10^6 เป็นพหุคูณของ 0.0625 จึงไม่มีค่าใดอยู่ในช่วง $(0.9375, 1)$ เลย เครื่องเห็นศูนย์เฟรม ทั้งที่กำแพงหนาขึ้นเป็น 3 เท่า

ตรวจด้วยการคำนวณ: `test_ch01.py`, `test_ex_1_1_8_six_cm_wall_faces_and_frames`

เฉลย 1.1.9.

1. ถ้า $e \le 16$ ระยะห่างของเลขที่เก็บได้ในช่วงนี้คือ $2^{e-23} \le 2^{-7} = 0.0078125 < 0.01$ ช่วง $(a, a+0.01)$ ยาวกว่าระยะห่างนี้ จึงมีจุดกึ่งกลาง m ระหว่างเลขที่เก็บได้ติดกันสองตัว $p < q$ โดย $q-p = 2^{e-23}$ และ $a < m < a+0.01$ ค่า a อยู่ต่ำกว่า m จึงปัดไปยังเลขที่ไม่เกิน p ส่วน $a+0.01$ อยู่สูงกว่า m จึงปัดไปยังเลขที่ไม่ต่ำกว่า q ผิวทั้งสองจึงถูกเก็บเป็นเลขต่างกันเสมอ

2. สำหรับ $e = 17$ ระยะห่างคือ $2^{-6} = 0.015625 > 0.01$ เลือก $a = 2^{17} + 0.0079$ ดังนั้น $a + 0.01 = 2^{17} + 0.0179$ ทั้งสองอยู่ใน $[2^{17}, 2^{18})$ เลขที่เก็บได้ใกล้เคียงคือ 2^{17} และ $2^{17}+0.015625$ จุดกึ่งกลางคือ $2^{17}+0.0078125$ ค่า a อยู่สูงกว่าจุดนี้จึงปัดเป็น $2^{17}+0.015625$ และค่า $a+0.01$ ห่างจาก $2^{17}+0.015625$ เพียง 0.0023 จึงปัดเป็นตัวเดียวกัน ผิวทั้งสองจึงถูกเก็บเป็นเลขตัวเดียวกัน

สรุปได้ว่า ในช่วง $[2^{16}, 2^{17})$ กำแพงหนา 1 เซนติเมตรไม่มีวันถูกเก็บเป็นศูนย์ แต่ในช่วง $[2^{17}, 2^{18})$ มีตำแหน่งที่ถูกเก็บเป็นศูนย์ได้ ตรวจด้วยการคำนวณ:

`test_ex_1_1_9_one_cm_wall_below_2_pow_17` ทดสอบจุดสุ่มด้วย seed สองค่าเป็นการตรวจเสริม และตรวจตัวอย่างของข้อ (ข) ตรงตัว

อ่านต่อ. หัวข้อ 1.2 จะตอบคำถามที่ค้างอยู่ คือ เมื่อปัดเศษแล้วความคลาดเคลื่อนแต่ละครั้งใหญ่แค่ไหน ตัวเลขที่เล็กหรือใหญ่เกินไปจะเป็นอย่างไร และเรารู้ล่วงหน้าได้อย่างไรว่าปัญหาเช่นนี้จะเกิด

1.2 ปัดเศษ ขอบเขต และตัวเลขที่เต็ม

สถานการณ์สมมติ: ตัวนับหยุดที่ 16,777,216

ลองนึกถึงตัวนับคะแนนในเกมที่เก็บเป็นเลขความแม่นเดี่ยวเหมือนในหัวข้อ 1.1 ทุกเฟรมมันบวก 1 เข้าไป ที่ 60 เฟรมต่อวินาที ตัวนับจะเพิ่มขึ้นเรื่อย ๆ แล้วหยุดที่ 16,777,216 ซึ่งคือ 2^{24} หลังจากนั้นมันไม่ขยับอีกเลย ทั้งที่ยังบวก 1 ทุกเฟรม การนับไปถึงจุดนี้ใช้เวลาราว 77.7 ชั่วโมง

นาฬิกาของเกมก็มีปัญหาคล้ายกัน มันบวก $\frac{1}{60}$ วินาทีทุกเฟรม ดูเหมือนถูกต้องทุกเฟรม แต่หลังผ่านไปหนึ่งชั่วโมงมันอ่านได้ 3597.2068 วินาที ไม่ใช่ 3600 วินาที (ตัวเลขทั้งหมดมาจากโค้ดของเรา ไม่ได้มาจากเกมใดเกมหนึ่ง)

คำถามของหัวข้อนี้คือ ทำไมการบวกที่ง่ายที่สุดจึงหยุดหรือคลาดได้ และเครื่องรู้ได้อย่างไรว่าค่าคลาดเคลื่อนสัมบูรณ์แต่ละครั้งมีขนาดเท่าใด

ฉบับร่าง · Preprint · 2026-10-23

ภาพรวมก่อนอ่าน

เครื่องมีช่องเก็บที่ห่างเท่ากันในแต่ละช่วงกำลังของ 2 ถ้าผลที่ต้องการไม่ตรงกับช่องใด เครื่องจะไปช่องที่อยู่ใกล้ที่สุด ค่าคลาดเคลื่อนสัมบูรณ์แต่ละครั้งจึงไม่เกินครึ่งช่อง ถ้าช่องกว้างกว่าค่าที่เราบวก ค่านั้นจะหายไปทั้งก้อน บทนี้ดูสามเรื่องนี้ คือการปัดเศษ จำนวนเต็มที่เก็บได้ตรง ๆ และสิ่งที่เกิดเมื่อเลขใหญ่หรือเล็กเกินช่วง เราจะเห็นว่าความผิดทุกชนิดมีขอบเขตที่คำนวณได้ แต่ขอบเขตนั้นอาจหยาบมาก

ปัดเข้าช่องที่ใกล้ที่สุด

เมื่อผลลัพธ์ของการคำนวณไม่ตรงกับเลขที่เก็บได้ เครื่องต้องเลือกเลขหนึ่งตัวจากเลขเหล่านั้น วิธีมาตรฐานคือเลือกตัวที่อยู่ใกล้ที่สุด เรียกว่า **การปัดเศษแบบใกล้ที่สุด** (rounding to nearest) ถ้าผลอยู่กึ่งกลางระหว่างสองช่องพอดี เครื่องเลือกช่องที่ส่วนเศษเป็นเลขคู่ เรียกว่า **ปัดเข้าหาเลขคู่** (ties to even) เพื่อไม่ให้การปัดกึ่งกลางเอนไปทางขึ้นเสมอ

นิยาม 1.2.1: การปัดเศษแบบใกล้ที่สุด

ให้ x เป็นจำนวนจริงที่ $2^{-126} \le |x| \le \max F$ เมื่อ $\max F = (2 - 2^{-23})\,2^{127}$ ให้ $\mathrm{fl}(x)$ เป็นเลขใน F จากหัวข้อ 1.1 ที่อยู่ใกล้ x มากที่สุด ถ้ามีสองตัวที่อยู่ห่าง x เท่ากัน ให้เลือกตัวที่ส่วนเศษเป็นเลขคู่

อ่านแบบเข้าใจง่าย

- $\mathrm{fl}(x)$ คือค่าที่เครื่องเก็บได้จริงของ x
- ช่วง $2^{-126} \le |x| \le \max F$ คือช่วงของเลขปกติ นอกช่วงนี้มีกติกาอีกแบบ ซึ่งเราจะดูในหัวข้อเรื่องขอบเขต
- ตัวอย่าง ที่ $x = 10^6 + 0.04$ เลขที่เก็บได้ที่อยู่ใกล้สุดคือ $10^6 + 0.0625$ เพราะห่างเพียง 0.0225 ส่วน 10^6 ห่างถึง 0.04 ดังนั้น $\mathrm{fl}(x) = 10^6 + 0.0625$

ลองเอง: ตัวนับที่ค้าง

ตัวนับเริ่มที่ $2^{24} - 8$ แล้วบวก 1 ทุกเฟรม เดาก่อนรันว่าตัวนับจะค้างที่ค่าใด และค้างตั้งแต่เฟรมที่เท่าใด จากนั้นรันโค้ดเพื่อพิมพ์ค่าทั้ง 12 เฟรม ผลที่ได้คือ $2^{24} - 8, \ldots, 2^{24}$ แล้วค้างที่ 2^{24} ตั้งแต่เฟรมที่ 8 เป็นต้นไป

หน่วยปัดและเอปซิลอนของเครื่อง

การปัดทำให้ผิดเท่าไร ต้องวัดเทียบกับขนาดของ x เสมอ เพราะผิด 0.01 เมตรที่ 10^6 เมตร กับผิด 0.01 เมตรที่จุดกำเนิด มีความหมายต่างกันมาก ค่าคลาดเคลื่อนสัมพัทธ์ที่มากที่สุดจากการปัดเรียกว่า **หน่วยปัด** (unit roundoff) (Atkinson เรียกว่า unit round) ใช้สัญลักษณ์ u Atkinson 1989, §1.2, (1.2.11), p. 15 เครื่อง VAX รุ่นเก่าในตารางของ Atkinson ใช้ความแม่น 24 บิตแบบเดียวกัน และมีค่า $u = 5.96 \times 10^{-8}$ Atkinson 1989, §1.2, Table 1.1, p. 13

นิยาม 1.2.2: หน่วยปัดและเอปซิลอนของเครื่อง

สำหรับเลขความแม่นเดี่ยว ให้ $u = 2^{-24}$ เป็นหน่วยปัด และให้ $\varepsilon = 2^{-23}$ เป็น **เอปซิลอนของเครื่อง** (machine epsilon) ซึ่งคือระยะห่างระหว่าง 1 กับเลขที่เก็บได้ตัวถัดไป

ฉบับร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

$\varepsilon = 2u$ เพราะช่องหลัง 1 กว้างเป็นสองเท่าของครึ่งช่องที่ u วัด ลองดูผลเมื่อบวกเลขเล็ก ๆ เข้ากับ 1 ถ้าเลขนั้นไม่เกิน u ผลจะปัดกลับมาที่ 1 ทุกครั้ง แต่ $1 + \varepsilon$ เก็บเป็นเลขถัดไปจาก 1 ได้จริง

ทฤษฎีบท 1.2.1: ค่าคลาดเคลื่อนสัมพัทธ์ของการปัดเศษ

ถ้า $2^{-126} \le |x| \le \max F$ แล้ว

$$|\mathrm{fl}(x) - x| \le u\,|x|$$

ดังนั้นเขียนได้เป็น $\mathrm{fl}(x) = x(1+\delta)$ เมื่อ $|\delta| \le u = 2^{-24}$

อ่านแบบเข้าใจง่าย

ค่าคลาดเคลื่อนสัมพัทธ์ของการปัดไม่เกิน u (คือไม่เกิน u เท่าของขนาด x) ซึ่งประมาณ 6×10^{-8} หรือ 0.000006 เปอร์เซ็นต์ เรื่องนี้เป็นจริงในทุกขนาดของช่วงปกติ แต่ค่าคลาดเคลื่อนสัมบูรณ์เปลี่ยนไปตามขนาดของ x ที่ $x = 10^6$ ค่าคลาดเคลื่อนสัมบูรณ์ที่มากที่สุดที่เกิดได้คือครึ่งช่อง 0.03125 เมตร ซึ่งเล็กกว่าขอบเขต $u|x| \approx 0.0596$ เมตร

ลองเอง: ตรวจทฤษฎีบทด้วย 0.1

ความแม่นเดี่ยวเก็บ 0.1 เป็น $0.100000001490116\ldots$ ดังนั้นค่าคลาดเคลื่อนสัมบูรณ์คือ $|\mathrm{fl}(0.1) - 0.1| \approx 1.49 \times 10^{-9}$ ส่วนขอบเขต $u|0.1| = 2^{-24} \times 0.1 \approx 5.96 \times 10^{-9}$ ค่าคลาดเคลื่อนสัมบูรณ์จริงจึงเป็นราวหนึ่งในสี่ของขอบเขต ทฤษฎีบทใช้ได้ และขอบเขตนี้ไม่หยาบจนไม่มีความหมายในกรณีนี้

พิสูจน์.

แนวคิด. เลขที่ใกล้ $|x|$ ที่สุดอยู่ในช่องที่ $|x|$ ตกอยู่ ช่องนั้นกว้าง 2^{e-23} ค่าคลาดเคลื่อนสัมบูรณ์จึงไม่เกินครึ่งช่อง และครึ่งช่องนั้นไม่เกิน $u|x|$

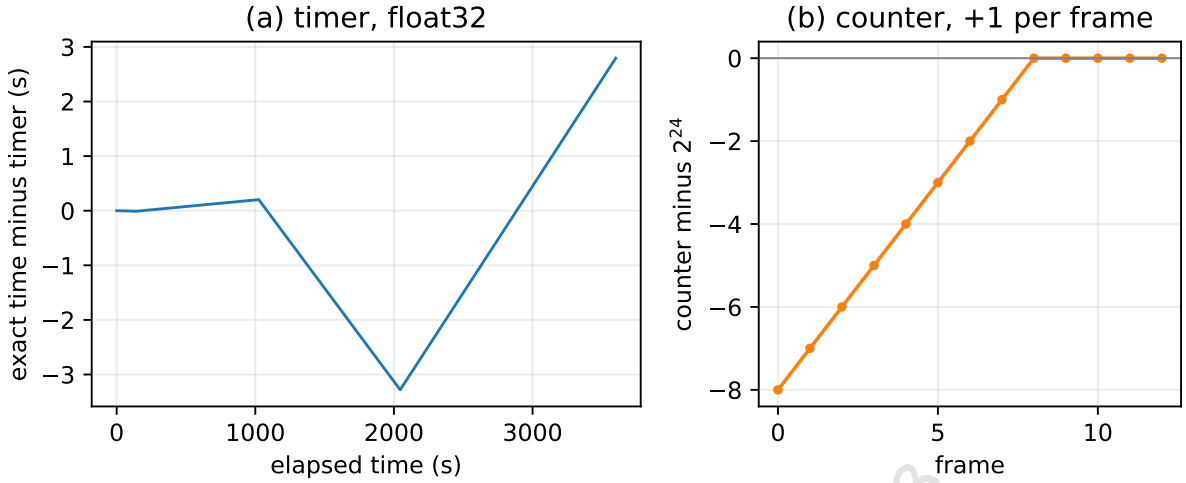
1. **หาช่องที่ $|x|$ อยู่.** ถ้า $|x| \in F$ แล้ว $\mathrm{fl}(x) = x$ และเสร็จ มิฉะนั้นให้ e เป็นจำนวนเต็มที่ $2^e < |x| < 2^{e+1}$ เพราะ $2^{-126} \le |x| \le \max F$ จึงมี $-126 \le e \le 127$ จากทฤษฎีบทระยะห่างในหัวข้อ 1.1 มีเลขสองตัวใน F ที่ติดกัน $p < |x| < q$ โดย $q - p = 2^{e-23}$ (ถ้า $e = 127$ จุด q ไม่เกิน $\max F$ เพราะ $|x| < \max F$ ในกรณีนี้)
2. **ระยะห่าง.** ไม่มีเลขใน F อยู่ระหว่าง p กับ q ดังนั้น $\mathrm{fl}(x)$ ต้องเป็น p หรือ q และ $\min\{|x| - p,\ q - |x|\} \le \frac{1}{2}(q-p) = 2^{e-24}$
3. **เทียบกับขนาด.** เพราะ $|x| \ge 2^e$ จึงได้ $2^{e-24} = 2^{-24} \cdot 2^e \le 2^{-24}|x| = u|x|$ ดังนั้น $|\mathrm{fl}(x) - x| \le u|x|$
4. **รูปคูณ.** ตั้ง $\delta = (\mathrm{fl}(x) - x)/x$ จะได้ $\mathrm{fl}(x) = x(1+\delta)$ และ $|\delta| \le u$

สรุป: การปัดเป็นเลขที่ใกล้สุดทำให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกินครึ่งช่อง และครึ่งช่องนั้นไม่เกิน u คูณขนาดของ x

ข้อควรระวัง: เอปซิลอนไม่ใช่ค่าน้อยที่สุดที่ทำให้ 1 เปลี่ยน

เราอาจคิดว่า ε คือเลขที่เล็กที่สุดซึ่งบวกกับ 1 แล้วได้ผลไม่เท่ากับ 1 แต่ไม่ใช่ เพราะ 2^{-24} เองบวกกับ 1 แล้วได้ 1 เนื่องจากอยู่กึ่งกลางพอดีและปัดเข้าหาเลขคู่ ส่วนเลขใดก็ตามที่มากกว่า 2^{-24} จะทำให้ 1 เปลี่ยน ดังนั้นไม่มีค่าที่เล็กที่สุด ค่า $\varepsilon = 2^{-23}$ คือระยะห่างหลัง 1 ที่ใช้กันทั่วไป

ร่างต้น · Preprint · 2026-10-23



รูปที่ 1.3: (ก) ความคลาดเคลื่อนของนาฬิกาความแม่นเดี่ยว คือเวลาจริงลบค่าที่อ่านได้ ตลอดหนึ่งชั่วโมง (ข) ตัวนับที่บวก 1 ทุกเฟรม เริ่มที่ $2^{24} - 8$ ค้างที่ 2^{24} ตั้งแต่เฟรมที่ 8 (สร้างจาก `numanal/ch01.py`)

จำนวนเต็มที่เก็บได้ตรง ๆ

ตัวนับต้องการจำนวนเต็ม เราจึงถามว่า จำนวนเต็มช่วงไหนเก็บได้ตรงโดยไม่ต้องปัด เรียกว่า **จำนวนเต็มที่เก็บได้ตรง ๆ** (exact integers) คำตอบคือถึง 2^{24} เหนือจุดนั้นจำนวนคี่หายไปจากเซต F Atkinson เรียกค่า 2^{24} นี้ว่า $M = \beta^t$ และทิ้งให้ผู้อ่านพิสูจน์เอง Atkinson 1989, §1.2, (1.2.16), p. 16

ทฤษฎีบท 1.2.2: จำนวนเต็มที่เก็บได้ตรง ๆ

ทุกจำนวนเต็ม m ที่ $1 \le m \le 2^{24}$ เป็นเลขใน F ส่วน $2^{24}+1$ ไม่เป็นเลขใน F และ $\mathrm{fl}(2^{24}+1) = 2^{24}$

พิสูจน์.
แนวคิด. จำนวนเต็มทุกตัวอยู่บนช่องที่ละเอียดพอ ส่วน $2^{24}+1$ อยู่กึ่งกลางพอดีระหว่างสองช่องที่เก็บได้

1. **จำนวนเต็มช่วงแรก.** ถ้า $1 \le m < 2^{24}$ ให้ e เป็นจำนวนเต็มที่ $2^e \le m < 2^{e+1}$ จะได้ $e \le 23$ ช่องว่างในช่วงนี้คือ 2^{e-23} และ $m = (m\,2^{23-e}) \cdot 2^{e-23}$ โดย $m\,2^{23-e}$ เป็นจำนวนเต็ม จึงเป็นพหุคูณของช่องว่างนี้ ดังนั้น $m \in F$
2. **ค่า 2^{24}.** 2^{24} เป็นเลขใน F โดยตรง
3. **ค่า $2^{24}+1$.** อยู่ใน $(2^{24}, 2^{25})$ ซึ่งช่องว่างคือ 2 เลขใน F ที่อยู่ติดกันคือ 2^{24} และ $2^{24}+2$ ห่างจาก $2^{24}+1$ เท่ากัน จึงเป็นกรณีกึ่งกลาง ส่วนเศษของ 2^{24} เป็นศูนย์ (คู่) ส่วนเศษของ $2^{24}+2$ มีบิตสุดท้ายเป็น 1 (คี่) จึงปัดเป็น 2^{24} และ $2^{24}+1$ ไม่อยู่ใน F

สรุป: จำนวนเต็มจนถึง 2^{24} เก็บได้ตรง เหนือจุดนี้ช่องว่างเป็นสอง จำนวนคี่จึงหายไป

สัมมนา · Preprint · 2026-10-23

ตารางที่ 1.4: ค่าที่นาฬิกาอ่านได้และความคลาดเคลื่อน (ตัวเลขจากโค้ดของเรา หน่วยวินาที)

เวลาจริง	อ่านได้ (float32)	คลาดเคลื่อน (float32)	คลาดเคลื่อน (float64)
1 วินาที	0.9999997	3.0×10^{-7}	-1.3×10^{-15}
10 วินาที	10.0000191	-1.9×10^{-5}	-7.6×10^{-14}
1 นาที	59.9992561	7.4×10^{-4}	2.1×10^{-12}
10 นาที	599.8976440	0.1024	2.2×10^{-10}
1 ชั่วโมง	3597.2067871	2.7932	-1.8×10^{-8}

ตารางที่ 1.5: ช่วงที่เครื่องหมายของความคลาดเคลื่อนไม่เปลี่ยน (นับทุกวินาที ตัวเลขจากโค้ดของเรา)

ช่วง (วินาที)	เครื่องหมาย	ความคลาดเคลื่อนที่ปลายช่วง (วินาที)
1–4	บวก	3.1×10^{-6}
5–11	ลบ	-3.8×10^{-6}
12–67	บวก	1.7×10^{-4}
68–180	ลบ	-1.7×10^{-4}
181–1,084	บวก	1.6×10^{-3}
1,085–2,884	ลบ	-3.7×10^{-3}
2,885–3,600	บวก	2.7932

นาฬิกาที่บวกทีละนิด

นาฬิกาของเกมบวก $\frac{1}{60}$ วินาทีทุกเฟรม แต่ก่อนบวกได้ เครื่องต้องเก็บ $\frac{1}{60}$ ก่อน และ $\frac{1}{60}$ มีการกระจายฐานสองไม่จบ เหมือน $\frac{1}{10}$ ในหัวข้อ 1.1 เลขที่เก็บได้คือ $0.0166666675\ldots$ ผิดไปราว 8.7×10^{-10} วินาทีต่อเฟรม ดูน้อยมาก แต่ 216,000 เฟรมรวมกันได้ราว 1.9×10^{-4} วินาที ส่วนที่ใหญ่กว่ามากมาจากการบวกแต่ละครั้งที่ผลบวกต้องถูกปัดให้เข้าช่องของมันเอง ตารางด้านล่างแสดงค่าที่อ่านได้ และรูป 1.3 (ก) แสดงความคลาดเคลื่อนตลอดชั่วโมง

ความคลาดเคลื่อนไม่ได้โตขึ้นเรื่อย ๆ และเครื่องหมายของมันก็เปลี่ยนไปมา เราใช้ความคลาดเคลื่อน = เวลาจริง − ค่าที่อ่านได้ ค่าบวกแปลว่านาฬิกาช้ากว่าเวลาจริง ค่าลบแปลว่าเร็วกว่า ตารางด้านล่างแสดงทุกช่วงที่เครื่องหมายไม่เปลี่ยนในชั่วโมงแรก ความคลาดเคลื่อนต่ำสุดคือ -3.28 วินาทีที่ราว 2,045 วินาที และกลับเป็นบวกอย่างถาวรตั้งแต่ราว 2,885 วินาที จนได้ $+2.7932$ วินาทีที่หนึ่งชั่วโมง

อ่านแบบเข้าใจง่าย: ทำไมการปัดไปทางเดียวกันเป็นช่วง ๆ

ผลบวกที่เก็บได้ในช่วง $[2^e, 2^{e+1})$ เป็นทวีคูณของช่องว่าง 2^{e-23} อยู่แล้ว ดังนั้นการบวก $\frac{1}{60}$ ที่ผลอยู่ในช่วงเดียวกันจะเพิ่มจำนวนช่องว่างเท่ากันทุกครั้ง ใน $[1024, 2048)$ ช่องว่างคือ 2^{-13} และ $\frac{1}{60}$ เท่ากับ 136.533 ช่องว่าง ส่วนเศษ 0.533 มากกว่าครึ่ง ทุกการบวกจึงเพิ่ม 137 ช่องว่าง ซึ่งเกิน $\frac{1}{60}$ อยู่ 5.7×10^{-5} วินาที นาฬิกาจึงเร็วขึ้นทีละนิด ใน $[2048, 3600]$ ช่องว่างคือ 2^{-12} และ $\frac{1}{60}$ เท่ากับ 68.267 ช่องว่าง ส่วนเศษ 0.267 น้อยกว่าครึ่ง ทุกการบวกจึงเพิ่มเพียง 68 ช่องว่าง นาฬิกาจึงช้าลงทีละ 6.5×10^{-5} วินาที ในช่วงแรกของ $[1024, 2048)$ มีการบวกหนึ่งครั้งที่ปัดลง (การบวกที่ข้ามเข้าช่วงนี้) นอกนั้นทั้งหมดปัดขึ้น

ต้นฉบับ · Preprint · 2026-05-23

ตารางที่ 1.6: การปัดของการบวก $\frac{1}{60}$ ในแต่ละช่วงกำลังของ 2 (ตัวเลขจากโค้ดของเรา)

ช่วง $[2^k, 2^{k+1})$	ช่องว่าง	$\frac{1}{60}$ หารด้วยช่องว่าง	ปัดไปทาง	เวลาจริงลบค่าที่อ่านได้ เปลี่ยนต่อการบวก (วินาที)
$[2^6, 2^7)$	2^{-17}	2184.53	ขึ้น	-3.6×10^{-6}
$[2^7, 2^8)$	2^{-16}	1092.27	ลง	4.1×10^{-6}
$[2^8, 2^9)$	2^{-15}	546.13	ลง	4.1×10^{-6}
$[2^9, 2^{10})$	2^{-14}	273.07	ลง	4.1×10^{-6}
$[2^{10}, 2^{11})$	2^{-13}	136.53	ขึ้น	-5.7×10^{-5}
$[2^{11}, 2^{12})$	2^{-12}	68.27	ลง	6.5×10^{-5}

ลองเอง: ทำนายความคลาดเคลื่อนหนึ่งชั่วโมงจากตาราง

ที่ 1,024 วินาที ความคลาดเคลื่อนคือ $+0.2059$ วินาที ในช่วง $[1024, 2048)$ มีการบวก $61{,}231$ ครั้ง แต่ละครั้งลดความคลาดเคลื่อนลง 5.6966×10^{-5} วินาที ในช่วง $[2048, 3600]$ มี $93{,}317$ ครั้ง แต่ละครั้งเพิ่มความคลาดเคลื่อนขึ้น 6.5104×10^{-5} วินาที ดังนั้น

$$0.2059 - 61{,}231 \times 5.6966 \times 10^{-5} + 93{,}317 \times 6.5104 \times 10^{-5} \approx 0.2059 - 3.4881 + 6.0753 \approx 2.7931$$

วินาที ใกล้กับค่าที่นาฬิกาคลาดจริง 2.7932 วินาทีมาก เราทำนายได้โดยไม่ต้องรันทั้งชั่วโมง ถ้าเข้าใจกติกาการปัดของแต่ละช่วง

ทฤษฎีบทค่าคลาดเคลื่อนสัมพัทธ์ใช้กับการบวกทุกครั้งได้ เพราะการบวกแต่ละครั้งคือการปัดผลบวกที่แม่นตรง ผลบวกหลังเฟรมที่ k คือประมาณ $k/60$ การบวกแต่ละครั้งผิดไม่เกิน u คูณผลบวก รวมทั้งหนึ่งชั่วโมงได้ขอบเขตหยาบ

$$\sum_{k=1}^{216\,000} u \, \frac{k}{60} \approx 23.2 \text{ วินาที}$$

ค่าคลาดเคลื่อนสัมบูรณ์จริง 2.79 วินาทีอยู่ในขอบเขตนี้ แต่ขอบเขตนี้หยาบมาก เพราะไม่ได้ใช้ความจริงว่าการปัดในช่วงเดียวกันมีทิศทางเดียวกัน (ค่าคลาดเคลื่อนของ $\frac{1}{60}$ เองรวมกันราว 2×10^{-4} วินาทีจึงไม่อยู่ในขอบเขตนี้) บทที่ 2 จะดูการสะสมค่าคลาดเคลื่อนแบบนี้อย่างละเอียด

เมื่อเลขใหญ่หรือเล็กเกินช่วง

เลขปกติทุกตัวอยู่ในช่วง $2^{-126} \le |x| \le \max F$ ผลลัพธ์ที่ใหญ่เกินช่วงเรียกว่า **โอเวอร์โฟลว์** (overflow) ส่วนผลที่เล็กเกินช่วงเรียกว่า **อันเดอร์โฟลว์** (underflow) ค่า $\max F$ ของความแม่นเดี่ยวประมาณ 3.40×10^{38} และค่าปกติเล็กที่สุดคือ $2^{-126} \approx 1.18 \times 10^{-38}$

ในมาตรฐาน IEEE 754 ผลที่ใหญ่เกินจะกลายเป็น **อนันต์** (infinity) คือ $\pm\infty$ แทนที่จะหยุดโปรแกรม Goldberg 1991 เครื่องรุ่นเก่าที่ Atkinson อธิบายไว้ถือว่าการล้นเป็นความผิดร้ายแรงและหยุดโปรแกรม Atkinson 1989, §1.3, p. 17

ทางด้านเล็ก เลขที่น้อยกว่า 2^{-126} ยังเก็บได้ด้วย **เลขซับนอร์มอล** (subnormal number) ซึ่งไม่มีหลักนำหน้า 1 และช่องว่างคงที่คือ $2^{-149} \approx 1.4 \times 10^{-45}$ เมื่อเลขเล็กกว่านั้นอีกก็ปัดเป็นศูนย์ ดังนั้นค่าจะเข้าใกล้ศูนย์อย่างค่อยเป็นค่อยไป แทนที่จะกระโดดจาก 2^{-126} ไปเป็นศูนย์ทันที Goldberg 1991, section “Denormalized Numbers” ส่วนความหมายของค่าอนันต์และ NaN ตรวจเทียบกับ Goldberg Goldberg 1991, sections “Infinity” and “NaNs”

ฉบับร่าง · Preprint · 2026-10-23

ลองเอง: ค่าที่อยู่กึ่งกลางพอดี

2^{-150} อยู่กึ่งกลางระหว่าง 0 กับ 2^{-149} พอดี การปัดเข้าหาเลขคู่จึงได้ 0 ส่วน 1.5×2^{-150} อยู่ใกล้ 2^{-149} มากกว่า ผลจึงเป็น 2^{-149} ลองคิดก่อนแล้วตรวจด้วยโค้ดในการทดสอบ

ในโลกจริง: ตัวเลขที่ไม่พอช่อง (Ariane 5, เที่ยวบิน 501)

วันที่ 4 มิถุนายน ค.ศ. 1996 จรวด Ariane 5 เที่ยวบิน 501 เบี่ยงออกจากเส้นทางและแตกสลายประมาณ 36.7 วินาที หลังเริ่มลำดับการปล่อย คณะกรรมการสอบสวนอิสระ (Inquiry Board) ระบุว่าค่า BH ซึ่งเป็นค่าความเอนเอียงแนวนอนจากการจัดแนวระบบวัดความเฉื่อย มีค่ามากกว่าที่จำนวนเต็มมีเครื่องหมาย 16 บิตเก็บได้ แต่ซอฟต์แวร์ไม่ได้ป้องกันข้อผิดพลาดตอนแปลงจากเลขทศนิยม 64 บิต เมื่อ SRI 2 หยุดทำงาน คอมพิวเตอร์นำร่องจึงอ่านรูปแบบวินิจฉัยเสมือนเป็นข้อมูลท่าทาง ส่วน SRI สำรองหยุดทำงานไปแล้วในรอบข้อมูลก่อนหน้า Board 1996
เรื่องนี้ไม่ได้เกิดจากความแม่นของตัวเลข แต่เกิดจากขอบเขตของช่องเก็บ ค่าที่อาจใหญ่กว่าช่องต้องถูกตรวจก่อนแปลง ซึ่งเป็นแนวคิดเดียวกับแบบฝึกหัด 1.2.6

จำนวนเต็มก็ล้นได้

จำนวนเต็มที่เก็บด้วยจำนวนบิตคงที่ก็มีช่วงจำกัด เช่นจำนวนเต็มมีเครื่องหมาย 16 บิตเก็บได้ตั้งแต่ $-32\,768$ ถึง $32\,767$ ถ้าบวก $20\,000 + 20\,000$ ผลที่ได้คือ $-25\,536$ เพราะค่าวนกลับไปทางลบ (wrap-around) ต่างจากเลขทศนิยมที่ได้อนันต์ จำนวนเต็มที่ล้นจะให้ตัวเลขธรรมดาที่ดูเหมือนถูกต้อง โดยไม่มีสัญลักษณ์เตือน

ลองเอง: ล้นแล้วเครื่องหมายบอกไหม

เดาก่อนว่าผลของ $20\,000 + 20\,000$ และ $(-30\,000) + (-10\,000)$ มีเครื่องหมายอะไร และคู่ใดล้น ผลจริงคือ $-25\,536$ ซึ่งเป็นลบและล้นจริง ส่วนคู่หลังได้ $+25\,536$ ซึ่งเป็นบวก ทั้งที่ค่าที่ควรได้คือ $-40\,000$ ดังนั้นเครื่องหมายของผลบอกการล้นได้เฉพาะเมื่อตัวถูกดำเนินการไม่เป็นลบ (แบบฝึกหัด 1.2.6)

ข้อควรระวัง: ขั้นกลางล้นได้ แม้ผลลัพธ์ไม่ล้น

ความยาว $\sqrt{x^2 + y^2}$ ของ $x = y = 10^{20}$ คือ $\sqrt{2} \times 10^{20} \approx 1.41 \times 10^{20}$ ซึ่งอยู่ในช่วงเลขความแม่นเดี่ยวสบาย ๆ แต่ $x^2 = 10^{40}$ ล้นไปแล้วก่อนถึงขั้นรากที่สอง การคำนวณตรง ๆ จึงได้อนันต์ แนวทางแก้คือเขียนเป็น $a\sqrt{1 + (b/a)^2}$ โดยให้ a เป็นค่าที่มากกว่า ซึ่งไม่ต้องยกกำลังเลขใหญ่ แนวคิดนี้มาจาก Atkinson 1989, §1.3, p. 23 และโค้ดของเราได้ 1.4142136×10^{20}

ค่าพิเศษ: อนันต์ และค่าไม่ใช่ตัวเลข

ผลที่เครื่องหาความหมายไม่ได้ เช่น $0/0$ หรือ $\infty - \infty$ จะได้ **ค่าไม่ใช่ตัวเลข** (NaN) Goldberg 1991 ค่า NaN ไม่เท่ากับตัวมันเอง จึงตรวจด้วยเครื่องหมายเท่ากับไม่ได้ กติกาของการบวก ลบ คูณ หาร และถอดรากที่สองคือ ถ้าตัวถูกดำเนินการตัวใดเป็น NaN ผลลัพธ์ก็เป็น NaN ดังนั้นค่านี้จะส่งต่อไปยังทุกการคำนวณที่ใช้มันเป็นอินพุตโดยตรง อย่างไรก็ตาม โปรแกรมตรวจว่าค่าเป็น NaN แล้วแยกไปทำอย่างอื่นได้ และบางฟังก์ชัน เช่น `fmax` คืนค่าที่ไม่ใช่ NaN เมื่อมี NaN เพียงตัวเดียว

ร่าง · Preprint · 2026-10-23

ข้อควรระวัง: NaN เปรียบเทียบไม่ได้

ถ้าตรวจว่า x เป็น NaN ด้วย $x = x$ เงื่อนไขนี้จะเป็นเท็จเมื่อ x เป็น NaN ทั้งที่เราอยากให้เป็นจริง ต้องใช้การตรวจแบบอื่น เช่นฟังก์ชันตรวจ NaN โดยเฉพาะ

เมื่อไรความหนาของกำแพงจะหายไป

กลับไปที่กำแพงในหัวข้อ 1.1 กำแพงหนา d จะถูกเก็บเป็นศูนย์เมื่อไร คำตอบขึ้นกับช่องว่างรอบตำแหน่งนั้น ไม่ได้ขึ้นกับความหนาอย่างเดียว

ทฤษฎีบท 1.2.3: เกณฑ์ความหนาที่หายไป

ให้ $d > 0$ และ a เป็นจำนวนจริง ถ้า $\mathrm{fl}(a) = \mathrm{fl}(a + d) = g$ แล้ว

$$d \le \tfrac{1}{2}\,(s_- + s_+)$$

เมื่อ s_- และ s_+ คือช่องว่างของเลขที่เก็บได้ทางซ้ายและทางขวาของ g โดยเฉพาะ ถ้า a และ $a + d$ อยู่ในช่วง $[2^e, 2^{e+1})$ เดียวกัน และ $d > 2^{e-23}$ แล้ว $\mathrm{fl}(a) \neq \mathrm{fl}(a + d)$

พิสูจน์.
แนวคิด. ถ้าสองผิวถูกปัดไปจุดเดียวกัน ระยะระหว่างผิวทั้งสองจะต้องไม่เกินระยะที่จุดนั้นครอบคลุมได้

1. **กรณีคร่อม.** ถ้า $a \le g \le a + d$ แล้ว g ใกล้ a มากที่สุด จึงได้ $|g - a| \le s_-/2$ เพราะเลขที่เก็บได้ฝั่งซ้ายของ g อยู่ห่าง s_- และ g ต้องอยู่ใกล้กว่าครึ่งทาง ในทำนองเดียวกัน $|a+d-g| \le s_+/2$ จึงได้ $d = |g-a| + |a+d-g| \le \frac{1}{2}(s_- + s_+)$
2. **กรณีอยู่ด้านเดียว.** ถ้า $a < a+d < g$ แล้ว $d = |g-a| - |g-(a+d)| \le |g-a| \le s_-/2$ ถ้า $g < a < a+d$ แล้ว $d \le |a + d - g| \le s_+/2$ ทั้งสองกรณีไม่เกิน $\frac{1}{2}(s_- + s_+)$
3. **กรณีช่วงเดียวกัน.** ถ้า $a, a + d \in [2^e, 2^{e+1})$ จุด g อยู่ใน $[2^e, 2^{e+1}]$ ถ้า g อยู่ภายในช่วง $s_- = s_+ = 2^{e-23}$ ถ้า $g = 2^e$ ทั้งสองอยู่ด้านขวา จึงใช้ $s_+ = 2^{e-23}$ ถ้า $g = 2^{e+1}$ ทั้งสองอยู่ด้านซ้าย จึงใช้ $s_- = 2^{e-23}$ ในทุกกรณี $d \le 2^{e-23}$ เมื่อ $d > 2^{e-23}$ จึงเป็นไปไม่ได้ที่จะปัดเป็นจุดเดียวกัน

สรุป: ถ้ากำแพงหนา d เก็บเป็นศูนย์ที่ตำแหน่งหนึ่ง ความหนาต้องไม่เกินครึ่งผลรวมของช่องว่างสองข้างที่ตำแหน่งนั้น

อ่านแบบเข้าใจง่าย

กำแพงหนา d จะเก็บเป็นศูนย์ได้เฉพาะเมื่อช่องว่างรอบตำแหน่งนั้นกว้างพอ นี่เป็นเงื่อนไขจำเป็นสำหรับตำแหน่งนั้น แต่ไม่ได้รับประกันทุกตำแหน่ง ส่วนคำถามว่า ในช่วงกำลังของ 2 เดียวกันมีตำแหน่งที่หายไปหรือไม่ ตอบได้ดังนี้ ถ้า $d < 2^{e-23}$ วางกึ่งกลางกำแพงไว้ที่เลขที่เก็บได้ก็หายไปเสมอ ถ้า $d = 2^{e-23}$ ต้องวางผิวทั้งสองบนจุดกึ่งกลางพอดี และอาศัยการปัดเข้าหาเลขคู่ ถ้า $d > 2^{e-23}$ ไม่มีตำแหน่งใดหายไปเลย ที่ x ในช่วง $[2^{16}, 2^{18})$ ช่องว่างไม่เกิน 0.015625 เมตร ซึ่งแคบกว่า 2 เซนติเมตร กำแพงที่ไม่คร่อม 2^{18} จึงไม่หายไป ที่ $[2^{18}, 2^{19})$ ช่องว่างคือ 0.03125 เมตร กว้างกว่า จึงมีตำแหน่งที่หายไปได้ และที่ 10^6 ช่องว่างคือ 0.0625 เมตร กว้างมากจนกำแพงหายไปได้ ตัวอย่างนี้คือแบบฝึกหัด 1.1.9 ที่ทั่วไปขึ้น

ฉบับร่าง · Preprint · 2026-10-23

ลองเอง: ผิวที่พอดีกึ่งกลาง

ที่ $[2^{18}, 2^{19})$ ช่องว่างคือ 0.03125 เมตร ลองวางกำแพงหนา 0.03125 เมตร โดยให้ผิวหน้าอยู่ที่ $a = 2^{18} + 0.046875$ และผิวหลังอยู่ที่ $a + 0.03125 = 2^{18} + 0.078125$ เดาก่อนว่าทั้งสองผิวจะถูกปัดไปจุดเดียวกันหรือไม่ ผลคือทั้งสองปัดไปที่ $2^{18} + 0.0625$ พอดี เพราะผิวทั้งสองอยู่บนจุดกึ่งกลางของช่องว่าง และการปัดเข้าหาเลขคู่เลือกจุดที่ส่วนเศษเป็นเลขคู่ ถ้าเลื่อนทั้งกำแพงไปเพียง 10^{-6} เมตร ผิวหลังจะพ้นจุดกึ่งกลาง ทั้งสองผิวจึงไม่ปัดไปจุดเดียวกันอีก

กลับไปที่ปลายดาบในหัวข้อ 1.1 ที่ 10^6 เมตร ช่องว่างคือ 0.0625 เมตร กว้างกว่าความหนา 2 เซนติเมตรของกำแพง ผิวทั้งสองจึงถูกปัดเป็นตัวเดียวกัน ส่วนกำแพงที่คร่อม 2^{18} เองก็เกิดได้เช่นกัน ดังที่ข้อควรระวังด้านล่างแสดง

ข้อควรระวัง: กำแพงที่คร่อมกำลังของ 2

ทฤษฎีบทข้างบนใช้ได้เมื่อผิวทั้งสองอยู่ในช่วงเดียวกัน ถ้ากำแพงคร่อม 2^{18} ช่องว่างด้านล่างคือ 0.015625 และด้านบนคือ 0.03125 ครึ่งหนึ่งของผลรวมช่องว่างทั้งสองคือ 0.0234 เมตร ซึ่งมากกว่า 2 เซนติเมตร ตัวอย่างเช่น $2^{18} - 0.005$ และ $2^{18} + 0.015$ ปัดเป็น 2^{18} ทั้งคู่ ดังนั้นกำแพงที่ต่ำกว่า 2^{18} เพียงเล็กน้อยก็หายไปได้

สรุปใจความ

- การปัดเศษแบบใกล้ที่สุดให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกินครึ่งช่อง และค่าคลาดเคลื่อนสัมพัทธ์ไม่เกิน $u = 2^{-24} \approx 6 \times 10^{-8}$ สำหรับเลขความแม่นเดี่ยว
- จำนวนเต็มตั้งแต่ 1 ถึง 2^{24} เก็บได้ตรง ดังนั้นตัวนับที่บวก 1 ทุกเฟรมจะค้างที่ 2^{24} (ราว 77.7 ชั่วโมงที่ 60 เฟรมต่อวินาที)
- การบวก $\frac{1}{60}$ ซ้ำ ๆ ทำให้ค่าคลาดเคลื่อนสะสม ในช่วงกำลังของ 2 ที่ใหญ่ การปัดเกือบทุกครั้งไปทางเดียวกัน หลังหนึ่งชั่วโมงนาฬิกาคลาดไป 2.79 วินาที ขอบเขตหยาบจากทฤษฎีบทคือราว 23 วินาที
- ค่าที่ใหญ่หรือเล็กเกินช่วงจะกลายเป็นอนันต์ ศูนย์ หรือเลขซับนอร์มอล การคำนวณที่ไม่มีความหมายได้ NaN และจำนวนเต็มที่ล้นจะวนกลับโดยไม่มีสัญลักษณ์เตือน
- กำแพงหนา d เก็บเป็นศูนย์ได้เฉพาะเมื่อช่องว่างรอบตำแหน่งนั้นกว้างพอ (เป็นเงื่อนไขจำเป็น) กำแพง 2 เซนติเมตรที่ 10^6 เมตรหายไปได้ เพราะช่องว่างคือ 0.0625 เมตร

เรากลับไปที่สองคำถามของหัวข้อนี้ ตัวนับหยุดเพราะ 1 เป็นครึ่งช่องของ 2 และปัดกลับมาที่ 2^{24} ส่วนนาฬิกาคลาดเพราะการปัดในแต่ละช่วงกำลังของ 2 มักไปทางเดียวกัน (เช่น ปัดขึ้นเกือบทั้งหมดใน $[1024, 2048)$ และปัดลงทั้งหมดใน $[2048, 3600]$) จึงสะสมเป็นก้อน โดยทิศทางเปลี่ยนไปตามช่วงเวลา

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 1.2.1 (⋆) ตัวนับเป็นเลขความแม่นเดี่ยวและมีค่าเป็น 2^{24}

1. หา $\mathrm{fl}(2^{24} + 1)$ และ $\mathrm{fl}(2^{24} + 3)$
2. ถ้าตัวนับบวก 2 ทุกเฟรมจาก 2^{24} ตัวนับจะขยับทุกเฟรมหรือไม่ อธิบาย

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 1.2.2 (★★) ลองพิสูจน์ว่า ถ้า y เป็นเลขใน F แล้ว $\mathrm{fl}(y) = y$ จากนั้นใช้ผลนี้อธิบายว่าทำไมการปัดเศษซ้ำ $\mathrm{fl}(\mathrm{fl}(x))$ จึงไม่เปลี่ยนค่าจาก $\mathrm{fl}(x)$

แบบฝึกหัด 1.2.3 (★★) ตัวนับเริ่มที่ 0 และบวก 2 ทุกเฟรม ที่ 60 เฟรมต่อวินาที

1. ตัวนับไปถึงค่าใดก่อนจะค้าง และใช้เวลาเท่าไร
2. เมื่อค้างแล้วตัวนับค้างที่ค่าใด
3. เปรียบเทียบกับตัวนับที่บวก 1 ทุกเฟรม ทำไมผลต่างกัน

แบบฝึกหัด 1.2.4 (★★) ทำซ้ำนาฬิกาในหัวข้อ 1.2 แต่ใช้เลขความแม่นคู่ (`float64`) ซึ่งมี $u = 2^{-53}$

1. ความคลาดเคลื่อนหลังสิบนาทีและหลังหนึ่งชั่วโมงเป็นเท่าไร
2. ขอบเขตหยาบ $\sum_{k=1}^{216000} u\,\frac{k}{60}$ เป็นเท่าไร และความคลาดเคลื่อนจริงหลังหนึ่งชั่วโมงอยู่ในขอบเขตนี้หรือไม่
3. ทำไมความคลาดเคลื่อนของความแม่นคู่จึงยังไม่เป็นศูนย์ ทั้งที่แม่นกว่าความแม่นเดี่ยวมาก

แบบฝึกหัด 1.2.5 (★★) ใช้เลขความแม่นเดี่ยว ให้ $\sqrt{x^2 + y^2}$ เป็นความยาวของเวกเตอร์ (x, y)

1. ที่ $x = 10^{19}$ และ $y = 0$ การคำนวณ $x^2 + y^2$ ตรง ๆ ล้นหรือไม่
2. หาค่า x ที่เล็กที่สุดโดยประมาณซึ่งทำให้ x^2 ล้น
3. เขียนวิธีคำนวณความยาวของ $(10^{20}, 10^{20})$ ที่ไม่ล้น และบอกผลที่ได้

แบบฝึกหัด 1.2.6 (★) ตัวนับคะแนนเก็บเป็นจำนวนเต็มมีเครื่องหมาย 16 บิต ซึ่งเก็บค่าได้ตั้งแต่ $-32\,768$ ถึง $32\,767$

1. ถ้าบวก $20\,000 + 20\,000$ ในการคำนวณแบบวนกลับ ได้ผลเท่าไร
2. ถ้า $a, b \geq 0$ ทำไมการตรวจว่า "ผลเป็นลบแสดงว่าล้น" จึงใช้ได้กับการบวก $a + b$ และถ้า $a, b < 0$ การตรวจแบบเดียวกันพลาดอย่างไร ให้ใช้ $(-30\,000) + (-10\,000)$ เป็นตัวอย่าง
3. เขียนเงื่อนไขที่ตรวจก่อนบวก $a + b$ ซึ่งบอกได้ว่าจะล้นหรือไม่ ทั้งกรณี $b \geq 0$ และ $b < 0$

แบบฝึกหัด 1.2.7 (★★★) โครงการรวมของบทที่ 1. ต่อจากสถานการณ์ตัวนับ นาฬิกา และปลายดาบ

1. ตัวนับเลขความแม่นเดี่ยวบวก 1 ทุกเฟรม ที่ 60 เฟรมต่อวินาที ตัวนับค้างที่ค่าใด และใช้เวลาเท่าไรก่อนจะค้าง
2. นาฬิกาบวก $\frac{1}{60}$ วินาทีทุกเฟรม หลังหนึ่งชั่วโมงความคลาดเคลื่อนเป็นเท่าไร และความคลาดเคลื่อนมีเครื่องหมายเปลี่ยนที่ช่วงใด
3. แสดงว่าถ้า $a + 0.02 < 2^{18}$ แล้วผิวทั้งสองเก็บเป็นเลขต่างกันเสมอ จากนั้นหาตำแหน่ง a ในช่วง $[2^{18}, 2^{19})$ ที่ทำให้ผิวทั้งสองถูกเก็บเป็นตัวเดียวกัน
4. ที่ 10^6 เมตร ทำไมปลายดาบในสถานการณ์หัวข้อ 1.1 จึงทะลุกำแพงไปได้โดยไม่ถูกตรวจพบ ให้ใช้ผลจากข้อ (ค) ตอบ

ตัวช่วยคิด

สำหรับข้อ (ค) ให้ใช้เกณฑ์ความหนาที่หายไป ดูว่าช่องว่างสองข้างของจุด g ที่ผิวทั้งสองถูกปัดไปเป็นมีค่าเท่าใด และเปรียบเทียบครึ่งหนึ่งของผลรวมนั้นกับ 0.02

ฉบับร่าง · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 1.2.1.

1. ค่า $2^{24}+1$ อยู่ใน $[2^{24}, 2^{25})$ ซึ่งช่องว่างคือ 2 เลขที่เก็บได้ที่อยู่ใกล้สุดคือ 2^{24} และ $2^{24}+2$ ห่างเท่ากันคือ 1 จึงเป็นกรณีกึ่งกลาง ส่วนเศษของ 2^{24} เป็นศูนย์ (คู่) จึงได้ $\mathrm{fl}(2^{24}+1) = 2^{24}$ สำหรับ $2^{24}+3$ เลขสองตัวที่ใกล้สุดคือ $2^{24}+2$ (ส่วนเศษเป็นเลขคี่) และ $2^{24}+4$ (เลขคู่) จึงได้ $\mathrm{fl}(2^{24}+3) = 2^{24}+4$

2. ใช่ เพราะ $2^{24}+2$ เป็นเลขที่เก็บได้ตรง และการบวก 2 เท่ากับช่องว่างพอดี ไม่มีการปัด ตัวนับจึงขยับทุกเฟรม (จนกว่าจะถึง 2^{25} ซึ่งเป็นเรื่องของแบบฝึกหัด 1.2.3)

ตรวจด้วยการคำนวณ: `test_ch01.py`, `test_ex_1_2_1_counter_above_two_to_the_24`

เฉลย 1.2.2. ให้ $y \in F$ เลขที่ใกล้ y มากที่สุดใน F คือ y เอง เพราะระยะห่างเป็นศูนย์ ซึ่งน้อยกว่าระยะถึงเลขตัวอื่นทุกตัว จึงไม่มีกรณีกึ่งกลาง ดังนั้น $\mathrm{fl}(y) = y$ สำหรับจำนวนจริง x ใด ๆ ที่อยู่ในช่วงปกติ $\mathrm{fl}(x)$ เป็นเลขใน F ตามนิยาม จึงได้

$$\mathrm{fl}(\mathrm{fl}(x)) = \mathrm{fl}(x).$$

นั่นคือการปัดซ้ำไม่เปลี่ยนค่า เพราะค่าที่ปัดแล้วอยู่บนช่องที่เก็บได้ตรงอยู่แล้ว ตรวจด้วยการคำนวณ: `test_ch01.py`, `test_ex_1_2_2_rounding_twice_changes_nothing` (สุ่ม 3000 ค่าด้วย seed คงที่)

เฉลย 1.2.3.

1. ค่า $0, 2, 4, \ldots$ เป็นเลขคู่ทั้งหมด และทุกเลขคู่จนถึง 2^{25} เก็บได้ตรง เพราะในช่วง $[2^{24}, 2^{25})$ ช่องว่างคือ 2 จึงตัวนับถึง 2^{25} หลังจาก $2^{25}/2 = 2^{24}$ เฟรม ซึ่งใช้เวลา $2^{24}/60 = 279\,620$ วินาที หรือราว 77.7 ชั่วโมง เท่ากับเวลาที่ตัวนับแบบบวก 1 ไปถึง 2^{24}

2. จาก 2^{25} การบวก 2 ได้ $2^{25}+2$ ซึ่งอยู่กึ่งกลางพอดีระหว่าง 2^{25} (ส่วนเศษเป็นศูนย์ คู่) กับ $2^{25}+4$ (ส่วนเศษเป็นเลขคี่) ปัดเข้าหาเลขคู่จึงได้ 2^{25} ตัวนับค้างที่ 2^{25}

3. ตัวนับค้างเมื่อค่าที่บวกน้อยกว่าครึ่งช่อง หรือเท่ากับครึ่งช่องและการปัดเข้าหาเลขคู่ให้ค่าเดิม ที่ 2^{24} ค่าที่บวก 1 เท่ากับครึ่งของช่อง 2 และปัดกลับมาที่ 2^{24} ที่ 2^{25} ช่องกว้าง 4 ค่าที่บวก 2 ก็เท่ากับครึ่งช่องและปัดกลับมาที่ 2^{25} เป็นกติกาเดียวกัน

ตรวจด้วยการคำนวณ: `test_ch01.py`, `test_ex_1_2_3_counter_with_step_two_stops_at_two_to_the_25` (ข้อ 1 ใช้การคำนวณ $2^{24}/60/3600 \approx 77.7$ ชั่วโมง)

เฉลย 1.2.4.

1. จากโค้ดของเรา นาฬิกา `float64` หลังสิบนาที ($36\,000$ เฟรม) คลาดไปราว 2.2×10^{-10} วินาที และหลังหนึ่งชั่วโมง คลาดไปราว -1.8×10^{-8} วินาที (เครื่องหมายลบหมายความว่านาฬิกาอ่านมากกว่าเวลาจริงเล็กน้อย)

2. ขอบเขตหยาบ
$$\sum_{k=1}^{216\,000} 2^{-53}\,\frac{k}{60} = 2^{-53} \cdot \frac{216\,000 \cdot 216\,001}{120} \approx 4.3 \times 10^{-8}$$
วินาที ความคลาดเคลื่อนจริง 1.8×10^{-8} วินาทีน้อยกว่าขอบเขตนี้ จึงอยู่ในขอบเขต

3. ความแม่นคู่มีช่องเล็กกว่าความแม่นเดี่ยวราว $2^{29} \approx 5 \times 10^8$ เท่า แต่ยังไม่เป็นศูนย์ เพราะ $1/60$ เก็บในความแม่นคู่ก็ยังปัดอยู่ และการบวกทุกครั้งยังถูกปัดเข้าช่องของผลบวก ค่าคลาดเคลื่อนเล็กลงจึงไม่ได้หายไป

สัมมนา · Preprint · 2026-10-23

ตรวจด้วยการคำนวณ: `test_ch01.py`, `test_ex_1_2_4_float64_timer_bound_holds_and_is_loose` และ `test_timer_float64_error_after_one_hour_is_tiny`

เฉลย 1.2.5.

1. ไม่ล้น เพราะ $x^2 = 10^{38}$ ซึ่งน้อยกว่า $\max F \approx 3.40 \times 10^{38}$ ผลคือ $\sqrt{10^{38}} = 10^{19}$ (ปัดเล็กน้อยตามปกติ)

2. x^2 ล้นเมื่อ $x^2 > \max F$ นั่นคือ $x > \sqrt{\max F} \approx 1.8447 \times 10^{19}$ ดังนั้นค่า x ที่เล็กที่สุดโดยประมาณที่ทำให้ล้นคือ 1.8447×10^{19}

3. ใช้วิธี $a\sqrt{1 + (b/a)^2}$ โดย $a = 10^{20}$ เป็นค่าที่มากกว่า และ $b = 10^{20}$ จึงได้ $(b/a)^2 = 1$ และผลคือ $10^{20}\sqrt{2} \approx 1.4142136 \times 10^{20}$ ค่านี้อยู่ในช่วงเลขความแม่นเดี่ยว และการคำนวณไม่มีค่าที่ล้นเลย

ตรวจด้วยการคำนวณ: `test_ex_1_2_5_squares_overflow_for_large_x` และ `test_overflow_of_a_naive_length_at_1e20_and_scaled_fix`

เฉลย 1.2.6.

1. $20\,000 + 20\,000 = 40\,000$ เกินช่วง $32\,767$ เมื่อวนกลับ $40\,000 - 65\,536 = -25\,536$

2. เมื่อ $a, b \geq 0$ ผลรวมอยู่ในช่วง $[0,\ 65\,534]$ ถ้าเกิน $32\,767$ มันจะวนกลับไปอยู่ในช่วง $[-32\,768,\ -2]$ ซึ่งเป็นลบเสมอ จึงได้ว่าผลเป็นลบก็ต่อเมื่อล้น การตรวจเครื่องหมายใช้ได้ในกรณีนี้ แต่เมื่อ $a, b < 0$ ผลรวมที่ต่ำกว่า $-32\,768$ จะวนกลับเป็นบวก เช่น $(-30\,000) + (-10\,000) = -40\,000$ วนกลับเป็น $-40\,000 + 65\,536 = 25\,536$ ซึ่งเป็นบวก การตรวจเครื่องหมายจึงพลาด ส่วนเมื่อเครื่องหมายต่างกันไม่มีการล้นเลย แต่ผลอาจเป็นลบได้ เช่น $3 + (-5) = -2$ ซึ่งไม่ใช่การล้น

3. สำหรับ $b \geq 0$ การบวกล้นเมื่อ $a > 32\,767 - b$ และสำหรับ $b < 0$ ล้นเมื่อ $a < -32\,768 - b$ ตรวจเงื่อนไขนี้ก่อนบวกก็จะรู้ได้ทันทีว่าผลจะวนกลับหรือไม่

ตรวจ ด้วย การ คำนวณ: `test_ex_1_2_6_wrap_and_sign_check` และ `test_ex_1_2_6_precheck_detects_every_overflow` (สุ่มคู่ (a, b) 5000 คู่ด้วย seed คงที่)
ตรวจ ส่วน ที่ เพิ่ม ด้วย `test_sign_check_is_exact_for_nonnegative_operands` และ `test_sign_check_misses_negative_overflow`

เฉลย 1.2.7.

1. ตัวนับค้างที่ $2^{24} = 16\,777\,216$ และไปถึงค่านั้นหลังจาก 2^{24} เฟรม ซึ่งที่ 60 เฟรมต่อวินาทีคือราว 77.7 ชั่วโมง เมื่อถึงค่านั้นแล้วการบวก 1 ไม่มีผลอีก

2. หลังหนึ่งชั่วโมง นาฬิกาอ่านได้ 3597.2068 วินาที คลาดไปราว 2.7932 วินาที เครื่องหมายของความคลาดเคลื่อน (เวลาจริงลบค่าที่อ่านได้) เปลี่ยนทั้งหมด 6 ครั้ง ตามช่วงต่อไปนี้ บวกที่ 1–4 วินาที ลบที่ 5–11 บวกที่ 12–67 ลบที่ 68–180 บวกที่ 181–1,084 ลบที่ 1,085–2,884 และบวกตั้งแต่ 2,885 ถึง 3,600 วินาที ความคลาดเคลื่อนต่ำสุดคือ -3.28 วินาที ที่ราว $2{,}045$ วินาที

3. **ข้อความแรก.** ถ้าผิวทั้งสองถูกปัดเป็นจุด g เดียวกัน ให้ s_- และ s_+ เป็นช่องว่างของเลขที่เก็บได้ทางซ้ายและทางขวาของ g ถ้า $a < g < a + d$ แล้ว $|g - a| \leq s_-/2$ และ $|a + d - g| \leq s_+/2$ จึงได้ $d \leq \frac{1}{2}(s_- + s_+)$ ถ้าทั้งสองอยู่ทางซ้ายของ g ได้ $d \leq s_-/2$ และถ้าทั้งสองอยู่ทางขวาได้ $d \leq s_+/2$ ทั้งสามกรณีได้ $d \leq \frac{1}{2}(s_- + s_+)$

   ข้อความที่สอง. ถ้า $a + 0.02 < 2^{18}$ แล้วจุด g ที่เป็นไปได้ต้องไม่เกิน 2^{18} ทุกจุด $g \leq 2^{18}$ และ $g = 2^{18}$ ถ้าทั้งสองอยู่ทางซ้าย $d \leq s_-/2 = 2^{-7} \approx 0.0078 < 0.02$ ถ้า $g < 2^{18}$ ช่องว่างสองข้างของ g ไม่เกิน $2^{-6} = 0.015625$ ดัง

ฉบับร่าง Preprint, 2026-10-23

นั้น $\frac{1}{2}(s_- + s_+) \le 0.015625 < 0.02$ ในทุกกรณีจึงไม่มีจุดใดที่ปัดเป็นตัวเดียวกันได้ **ตำแหน่ง.** ที่ $a = 2^{18} + 0.02$ ทั้ง a และ $a + 0.02 = 2^{18} + 0.04$ ปัดเป็น $2^{18} + 0.03125$ เพราะช่องว่างหลัง 2^{18} คือ 0.03125 ระยะจาก 0.02 และ 0.04 ถึง 0.03125 คือ 0.01125 และ 0.00875 ซึ่งไม่เกินครึ่งช่อง 0.015625

ข้อควรระวัง. ข้อความที่สองใช้ได้เฉพาะเมื่อ $a + 0.02 < 2^{18}$ ถ้ากำแพงคร่อม 2^{18} ช่องว่างสองข้างต่างกัน คือด้านล่าง 0.015625 และด้านบน 0.03125 จึงได้ $\frac{1}{2}(0.015625 + 0.03125) \approx 0.0234 > 0.02$ ทำให้เกิดได้ เช่น $a = 2^{18} - 0.005$ กับ $a + 0.02 = 2^{18} + 0.015$ ปัดเป็น 2^{18} ทั้งคู่

4. ที่ 10^6 เมตร ช่องว่างคือ 0.0625 มากกว่าความหนา 0.02 เมตร ดังนั้นกำแพงสองผิวอาจถูกปัดเป็นจุดเดียวกันได้ตามข้อ (ค) ที่ $x = 10^6$ ผิวหน้า $10^6 + 0.93$ และผิวหลัง $10^6 + 0.95$ ถูกเก็บเป็น $10^6 + 0.9375$ ทั้งคู่ ปลายดาบที่อยู่ระหว่างสองผิวจริงจึงไม่มีตำแหน่งเก็บได้ที่อยู่ "ในกำแพง" ตัวตรวจจึงไม่เห็นปลายดาบ สาเหตุคือช่องว่างที่กว้างกว่าความหนาของกำแพง ไม่ใช่การเคลื่อนที่เร็วเกินไป

ตรวจ ด้วย การ คำนวณ: `test_ex_1_2_7_mini_project_three_facts` ตรวจ ข้อ (ข) ด้วย `test_timer_error_turns_negative_then_positive_at_one_hour`, `test_timer_error_sign_runs_match_the_table` และ `test_timer_rounding_direction_is_set_by_the_binade`, ตรวจข้อ (ค) ด้วย `test_collapsing_example_at_two_to_the_18` และ `test_wall_straddling_two_to_the_18_can_collapse_below_it`

สรุปบทที่ 1 ในหน้าเดียว

สูตรที่ควรจำจากบทนี้คือ ช่องว่างของเลขปกติในช่วง $[2^e, 2^{e+1})$ เท่ากับ 2^{e-23} ค่าคลาดเคลื่อนสัมบูรณ์จากการปัดไม่เกิน $u|x|$ และกำแพงหนา d เก็บเป็นศูนย์ได้เฉพาะเมื่อ d ไม่เกินช่องว่างรอบตำแหน่งนั้น ตารางต่อไปนี้เทียบค่าสำคัญของสองรูปแบบ

ค่าสำคัญของเลขความแม่นเดี่ยวและเลขความแม่นคู่ (ตัวเลขจากโค้ดของเราและทฤษฎีบทในบทนี้)

ค่า	ความแม่นเดี่ยว (32 บิต)	ความแม่นคู่ (64 บิต)
ความแม่น (บิต รวมหลักนำหน้า)	24	53
เอปซิลอน ε	$2^{-23} \approx 1.19 \times 10^{-7}$	$2^{-52} \approx 2.22 \times 10^{-16}$
หน่วยปัด u	$2^{-24} \approx 5.96 \times 10^{-8}$	$2^{-53} \approx 1.11 \times 10^{-16}$
จำนวนเต็มที่เก็บได้ตรง M	$2^{24} = 16\,777\,216$	$2^{53} \approx 9.0 \times 10^{15}$
ช่องว่างที่ 10^6	0.0625	$2^{-33} \approx 1.16 \times 10^{-10}$
ค่ามากที่สุด $\max F$	$\approx 3.40 \times 10^{38}$	$\approx 1.80 \times 10^{308}$
เลขปกติเล็กที่สุด	$\approx 1.18 \times 10^{-38}$	$\approx 2.23 \times 10^{-308}$

วิธีเก็บตัวเลขเทียบกันด้วยระยะห่างและช่วง

วิธี	ระยะห่างระหว่างเลข	ช่วงที่เก็บได้
เลขจุดคงที่ 32 บิต (8 บิตหลังจุด)	$1/256$ เท่ากันทุกที่	ประมาณ ± 8.4 ล้านเมตร
เลขความแม่นเดี่ยว	2^{e-23} ตามขนาดของ x	ประมาณ $\pm 3.4 \times 10^{38}$
เลขความแม่นคู่	2^{e-52} ตามขนาดของ x	ประมาณ $\pm 1.8 \times 10^{308}$

อ่านต่อ. บทที่ 2 เริ่มจากคำถามที่บทนี้ทิ้งไว้ ถ้าค่าคลาดเคลื่อนสัมบูรณ์แต่ละครั้งเล็กแล้ว ทำไมนาฬิกาที่บวกนับล้านครั้งจึงคลาดได้เป็นวินาที และเมื่อสองค่าที่เกือบเท่ากันถูกลบกัน ความแม่นจะหายไปอย่างไร

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 2

เมื่อความคลาดเคลื่อนเล็ก ๆ โตขึ้น

สถานการณ์สมมติ: นาฬิกาที่เพี้ยนทีละนิด และกล้องที่หายไป

ที่บทที่แล้วเราเห็นนาฬิกาของเกมอ่านได้ 3597.2068 วินาที หลังผ่านไปหนึ่งชั่วโมง ทั้งที่เวลาจริงคือ 3600 วินาที ทุกเฟรมนาฬิกาบวก $\frac{1}{60}$ วินาทีเข้าไป ความผิดของการบวกแต่ละครั้งเล็กมากจนแทบมองไม่เห็น แต่ในหนึ่งชั่วโมงมี 216,000 เฟรม ความผิดเล็ก ๆ เหล่านั้นรวมกันได้เกือบสามวินาที

ลองนึกถึงกล้องในเกมอีกตัวหนึ่ง กล้องหันไปหาเป้าหมายด้วยมุมที่คำนวณจากเวกเตอร์สองเส้น ในบางเฟรมโปรแกรมได้ค่าที่ไม่ใช่ตัวเลข (NaN) ตอนที่คำนวณมุม แล้วกล้องก็หายไปจากจอ เรื่องนี้เป็นสถานการณ์สมมติ เราจะดูตัวเลขของมันในหัวข้อ 2.2

ทั้งสองเรื่องถามเหมือนกันว่า ความผิดเล็ก ๆ ในแต่ละขั้นของการคำนวณโตขึ้นได้แค่ไหน และมันมาจากอะไร

ภาพรวมบท

ที่บทที่แล้วเราเห็นว่าตัวเลขในเครื่องมีช่องว่างจำกัด บทนี้ถามต่อว่า เมื่อเรารวมหรือลบตัวเลขเหล่านั้นหลายครั้ง ความผิดจะเป็นอย่างไร

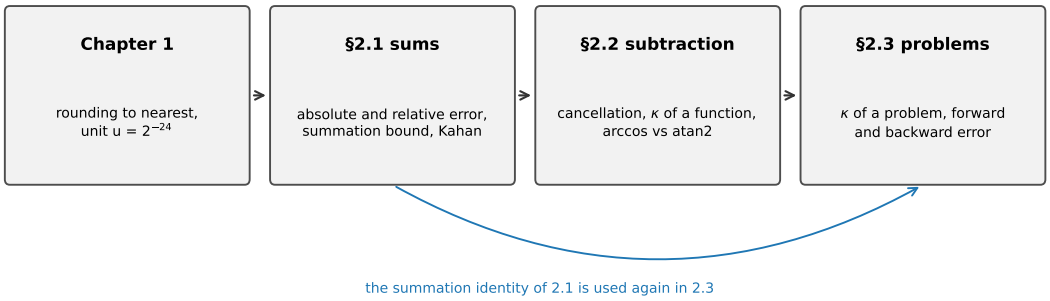
- ความผิดแต่ละครั้งวัดได้สองแบบ คือแบบสัมบูรณ์ซึ่งบอกว่าผิดไปกี่หน่วย และแบบสัมพัทธ์ซึ่งบอกว่าผิดไปเทียบกับขนาดของค่าเท่าไร (หัวข้อ 2.1)
- ความผิดที่ไปทางเดียวกันจะโตเร็วกว่าความผิดที่หักล้างกัน ลำดับการบวกและวิธีรวมก็มีผลด้วย (หัวข้อ 2.1)
- เมื่อลบตัวเลขสองตัวที่ใกล้กันมาก ความแม่นของผลลัพธ์หายไปอย่างรวดเร็ว แม้ตัวเลขแต่ละตัวจะแม่นดี (หัวข้อ 2.2)
- บางครั้งปัญหาเองไวต่อความผิดเล็กน้อยมาก จนไม่มีวิธีใดแก้ได้ดี (หัวข้อ 2.3)

เราเริ่มจากนาฬิกาที่เพี้ยนทีละนิดและกล้องที่หายไป แล้วจะเจออีกด้านหนึ่งของปัญหาเดียวกัน คือผิวสองผิวที่อยู่ใกล้กันมากจนเครื่องแยกไม่ออก ซึ่งจะเจอในหัวข้อ 2.3

ทำไมความผิดขนาดเล็กระดับ 10^{-7} จึงควรสนใจ เกมคำนวณซ้ำทุกเฟรม และในหนึ่งชั่วโมงมี 216,000 เฟรม ความผิดเล็ก ๆ ในแต่ละเฟรมอาจสะสมจนเห็นได้ หรือทำให้ได้ค่าที่ใช้ไม่ได้ เช่น NaN ในกล้องที่หายไป เราจึงต้องรู้สองเรื่อง คือความผิดจะโตเท่าไรเมื่อจำนวนรอบเพิ่มขึ้น และความผิดนั้นมาจากการเขียนสูตร หรือมาจากตัวปัญหาเอง เพราะสองกรณีนี้แก้ไม่เหมือนกัน

ภาพต่อไปนี้แสดงลำดับของหัวข้อในบทนี้ และสิ่งที่แต่ละหัวข้อใช้ต่อจากหัวข้อก่อนหน้า ถ้าอ่านตามลำดับจะเข้าใจง่ายที่สุด

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 2.1: ลำดับของหัวข้อในบทที่ 2 กล่องแต่ละกล่องคือหัวข้อหรือบทก่อนหน้า ลูกศรสีดำคือการพึ่งพาตามลำดับ และลูกศรสีน้ำเงินคือการใช้ผลบวกทีละขั้นของหัวข้อ 2.1 อีกครั้งในหัวข้อ 2.3 (สร้างจาก `numanal/ch02.py`)

2.1 สะสมทีละนิด

สถานการณ์สมมติ: นาฬิกาที่คลาดไปทีละนิด

นาฬิกาของเกมบวก $\frac{1}{60}$ วินาทีเข้าไปทุกเฟรม และเก็บค่าเป็นเลขความแม่นเดี่ยว หลังผ่านไปหนึ่งชั่วโมง ซึ่งมี 216,000 เฟรม นาฬิกาอ่านได้ 3597.2068 วินาที ทั้งที่เวลาจริงคือ 3600 วินาที ความผิดเกือบสามวินาทีเกิดจากการบวกทีละเฟรมแบบตรง ๆ เราจะถามว่าความผิดนี้โตได้แค่ไหน มาจากอะไร และมีวิธีบวกที่ดีกว่าหรือไม่

ภาพรวมก่อนอ่าน

ลองนึกถึงคนที่เดินข้ามสนามทีละก้าว ถ้าทุกก้าวเอียงไปทางซ้ายเล็กน้อย เขาจะเบนไปทางซ้ายมากขึ้นเรื่อย ๆ แต่ถ้าก้าวเอียงสลับซ้ายขวา ส่วนที่เบนจะหักล้างกันไปบ้าง การคำนวณก็เป็นแบบเดียวกัน ความผิดของแต่ละขั้นมีทั้งขนาดและทิศทาง ในหัวข้อนี้เราจะดูว่าเมื่อไรความผิดจะโตขึ้น และทำอย่างไรให้โตช้าลง

ค่าคลาดเคลื่อนสัมบูรณ์และสัมพัทธ์

เมื่อเราเปรียบค่าที่คำนวณหรืออ่านได้กับค่าจริง เราต้องบอกให้ชัดว่าผิดไปเท่าไร และมีสองวิธีที่ใช้กันบ่อย วิธีแรกบอกความผิดเป็นหน่วยเดียวกับตัวเลข เช่นวินาที วิธีที่สองบอกความผิดเทียบกับขนาดของค่าจริง

นิยาม 2.1.1: ค่าคลาดเคลื่อนสัมบูรณ์และสัมพัทธ์

ให้ x เป็นค่าจริง และให้ $\tilde{x}$ เป็นค่าที่คำนวณหรืออ่านได้

- **ค่าคลาดเคลื่อนสัมบูรณ์** (absolute error) คือ $|x - \tilde{x}|$
- **ค่าคลาดเคลื่อนสัมพัทธ์** (relative error) คือ $\dfrac{|x - \tilde{x}|}{|x|}$ เมื่อ $x \neq 0$

อ่านแบบเข้าใจง่าย

- ค่าคลาดเคลื่อนสัมบูรณ์บอกว่าผิดไปกี่หน่วย ถ้า x เป็นวินาที ค่านี้ก็เป็นวินาที
- ค่าคลาดเคลื่อนสัมพัทธ์บอกว่าความผิดเป็นสัดส่วนเท่าไรของค่าจริง มักเขียนเป็นเปอร์เซ็นต์

ตัวอย่าง · Preprint · 2026-10-23

- ตัวอย่าง ค่าจริง 100 อ่านได้ 99 ค่าคลาดเคลื่อนสัมบูรณ์คือ 1 และค่าคลาดเคลื่อนสัมพัทธ์คือ 1% แต่ค่าจริง 0.01 อ่านได้ 0.02 ค่าคลาดเคลื่อนสัมบูรณ์ก็เท่ากับ 0.01 แต่ค่าคลาดเคลื่อนสัมพัทธ์คือ 100% ความผิดเท่ากัน สัดส่วนต่างกันมาก

ลองเอง: นาฬิกาอ่านได้ 1200.4

นาฬิกาของเกมอ่านได้ 1200.4 วินาที ทั้งที่เวลาจริงผ่านไป 1203.1 วินาที ค่าคลาดเคลื่อนสัมบูรณ์คือ 2.7 วินาที ค่าคลาดเคลื่อนสัมพัทธ์คือ $2.7/1203.1 \approx 0.0022442$ หรือประมาณ 0.22 เปอร์เซ็นต์ ตัวเลขนี้ดูเล็กมาก แต่ถ้าค่าคลาดเคลื่อนสัมพัทธ์เท่าเดิมตลอด 10 ชั่วโมง นาฬิกาจะคลาดไปประมาณ 80.8 วินาที เดาก่อนคิดว่าจะคลาดไปเท่าไร แล้วค่อยคิดดู

ค่าคลาดเคลื่อนสัมบูรณ์ที่เล็กในหน่วยหนึ่งอาจใหญ่มากในอีกหน่วยหนึ่ง เรื่องนี้เกี่ยวกับ**เลขนัยสำคัญ** (significant digits) ซึ่งคือจำนวนหลักของค่าที่เราเชื่อได้ นับจากหลักที่มีค่ามากที่สุดที่ไม่เป็นศูนย์ ลองปัดค่า 3597.2068 ให้เหลือสองหลัก จะได้ 3.6×10^{3} หรือ 3600 ซึ่งตรงกับเวลาจริงทุกหลัก แต่ถ้าปัดให้เหลือสี่หลัก จะได้ 3597.0 ซึ่งมีความผิดเพิ่มขึ้นมา

ข้อควรระวัง: ตัวเลขที่พิมพ์ออกมาหลายหลักไม่ได้แปลว่าแม่น

นาฬิกาอ่านได้ 3597.2068 มีแปดหลัก แต่ความผิด 2.79 วินาทีบอกว่าหลักหน่วยและทศนิยมทั้งหมดไม่มีความหมาย และหลักสิบก็ไม่แน่นอนด้วย ก่อนจะดูตัวเลขที่พิมพ์ออกมา ให้ถามก่อนว่าความผิดเป็นเท่าไร

ความผิดของการบวกหนึ่งครั้ง

ทุกครั้งที่เครื่องบวกเลขสองตัว ผลที่ได้อาจไม่ตรงกับเลขที่เก็บได้ เครื่องจึงปัดให้เป็นเลขที่ใกล้ที่สุด ตามที่เราเห็นในบทที่แล้ว ความผิดที่เกิดจากการปัดแต่ละครั้งเรียกว่า**ค่าคลาดเคลื่อนการปัดเศษ** (round-off error) ความผิดนี้ไม่เกินหน่วยปัด $u = 2^{-24}$ คูณกับขนาดของผลที่ปัดนั้น

ให้ $x_1, x_2, \ldots, x_n$ เป็นเลขที่จะรวมกัน และให้ $S_k = x_1 + \cdots + x_k$ เป็นผลรวมที่แม่นตรง เรียก S_k ว่า**ผลบวกย่อย** (partial sum) ในเครื่องเราคำนวณผลบวกย่อยใหม่ทีละขั้น คือ $s_1 = x_1$ และ

$$t_k = s_{k-1} + x_k, \qquad s_k = t_k + r_k, \qquad k = 2, \ldots, n.$$

ที่นี่ t_k คือผลบวกก่อนปัด และ r_k คือค่าคลาดเคลื่อนการปัดเศษของการบวกครั้งที่ k ความผิดรวมคือ $E_n = s_n - S_n$

ลองเอง: บวก 0.1 สิบครั้ง

เลข 0.1 เก็บในเลขความแม่นเดี่ยวได้เป็น $0.100000001490116\ldots$ ซึ่งสูงกว่า 0.1 นิดหนึ่ง ถ้าบวกเลขนี้สิบครั้งในเครื่อง ผลที่ได้คือ $1.0000001192\ldots$ ไม่ใช่ 1 ความผิดทั้งหมด 1.19×10^{-7} แยกได้สองส่วน คือ จากการเก็บ 0.1 ซึ่งทำให้ผลรวมสูงขึ้น $10 \times 1.49 \times 10^{-9} = 1.49 \times 10^{-8}$ และจากการปัดในการบวก ซึ่งเพิ่มอีก 1.04×10^{-7} ลองทายก่อนว่าส่วนไหนใหญ่กว่า แล้วรันโค้ดใน `numanal/ch02.py` ดู

ฉบับร่าง Preprint · 2026-10-23

รวมตรง ๆ ในหนึ่งชั่วโมง

วิธีที่ตรงที่สุดคือรวมจากซ้ายไปขวา เรียกว่า**การรวมแบบตรง** (sequential summation) ตามขั้นตอนนี้

1. ตั้ง $s_1 = x_1$
2. สำหรับ $k = 2, 3, \dots, n$ คำนวณ $t_k = s_{k-1} + x_k$ แล้วปัดให้เป็นเลขความแม่นเดี่ยว ได้ s_k
3. ผลลัพธ์คือ s_n

นาฬิกาของเราทำแบบนี้ทุกเฟรม ทุกเฟรมเอา $\frac{1}{60}$ บวกเข้าไปในตัวนับ ถ้าเราละเรื่องการเก็บ $\frac{1}{60}$ ไว้ก่อน ความผิดของนาฬิกาก็คือ $-E_n$ ของการรวมแบบตรงนี้ ผลลัพธ์ต่อไปนี้ต้องดูในฐานะตัวอย่างของแนวคิด ไม่ใช่ข้อพิสูจน์ของนาฬิกาในเกมใดเกมหนึ่ง

ทฤษฎีบท 2.1.1: ขอบเขตของการรวมแบบตรง

ให้ $x_1, \dots, x_n$ เป็นเลขความแม่นเดี่ยวที่ไม่ติดลบ ให้ $S_k = x_1 + \cdots + x_k$ และให้ s_k เป็นผลบวกย่อยที่คำนวณด้วยการรวมแบบตรง เมื่อ $u = 2^{-24}$ และสมมติว่าการบวกทุกครั้งอยู่ในช่วงเลขปกติ คือไม่เกิดโอเวอร์โฟลว์ และผลไม่ต่ำจนเข้าสู่เลขซับนอร์มอล จะได้

$$|s_n - S_n| \le u\,(1+u)^{n-2} \sum_{k=2}^{n} S_k .$$

อ่านแบบเข้าใจง่าย

ทฤษฎีบทนี้บอกสองอย่าง อย่างแรก ความผิดรวมขึ้นกับจำนวนครั้ง n และขึ้นกับขนาดของผลบวกย่อยทุกตัวด้วย ไม่ใช่แค่จำนวนครั้ง อย่างที่สอง ถ้าผลบวกย่อยโตขึ้น ความผิดของการบวกแต่ละครั้งก็โตขึ้นตาม ตัวอย่างของนาฬิกาคือ $S_k = k/60$ ดังนั้น $\sum_k S_k \approx n^2/120$ เมื่อ $n = 216{,}000$ จะได้ประมาณ 3.9×10^8 คูณด้วย u ได้ประมาณ 23 วินาที

ข้อควรระวัง: ทฤษฎีบทนี้ไม่บอกอะไรเมื่อล้น

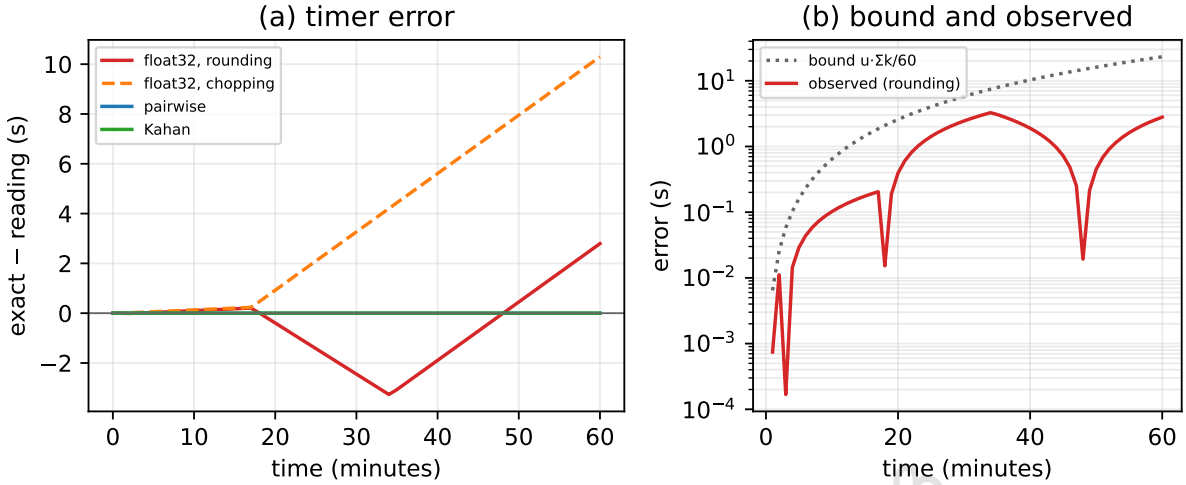
เงื่อนไขเรื่องช่วงไม่ใช่แค่พิธีการ ถ้าผลบวกย่อยเกินค่ามากที่สุดของเลขความแม่นเดี่ยวประมาณ 3.4×10^{38} เครื่องจะได้อนันต์ ตัวอย่างเช่น $3 \times 10^{38} + 3 \times 10^{38} = 6 \times 10^{38}$ ได้อนันต์ในความแม่นเดี่ยว ทั้งที่ค่าจริงเป็นจำนวนจำกัด ด้านเล็ก ถ้าผลเข้าสู่เลขซับนอร์มอล ซึ่งเล็กกว่า $2^{-126} \approx 1.18 \times 10^{-38}$ ความแม่นสัมพัทธ์อาจไม่เป็นไปตาม u อีกต่อไป ดังนั้นทฤษฎีบทนี้ใช้กับกรณีที่ไม่ล้นและไม่ต่ำเกินเท่านั้น

พิสูจน์

แนวคิด. ความผิดรวมคือผลรวมของความผิดจากการปัดแต่ละครั้ง และความผิดแต่ละครั้งไม่เกินหน่วยปัดคูณขนาดของผลบวกย่อยตรงนั้น

1. **เขียนความผิดรวมเป็นผลรวม.** จาก $S_k = S_{k-1} + x_k$ และ $t_k = s_{k-1} + x_k$ จะได้ $t_k - S_k = s_{k-1} - S_{k-1}$ ดังนั้น $s_k - S_k = (t_k - S_k) + r_k = (s_{k-1} - S_{k-1}) + r_k$ เพราะ $s_1 - S_1 = 0$ จึงได้ $E_n = \sum_{k=2}^{n} r_k$
2. **ขนาดของความผิดแต่ละครั้ง.** จากทฤษฎีบทการปัดเศษในหัวข้อ 1.2 เรามี $|r_k| \le u\,|t_k|$ เพราะการบวกแต่ละครั้งอยู่ในช่วงเลขปกติตามสมมติฐาน และเพราะ $x_k \ge 0$ กับ $s_{k-1} \ge 0$ จึงได้ $t_k \ge 0$ ดังนั้น $|E_n| \le u \sum_{k=2}^{n} t_k$

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 2.2: (ก) ความผิดของนาฬิกาทุกนาที (เวลาจริงลบด้วยค่าที่อ่านได้) สีแดงคือการปัดเข้าหาเลขที่ใกล้ที่สุด สีส้มคือการตัดทิ้ง เส้นสีน้ำเงินและเขียวอยู่ใกล้ศูนย์ตลอด (ข) ขอบเขตของการปัดเทียบกับความผิดที่เกิดจริง (แกนตั้งเป็นลอการิทึม และแสดงค่าสัมบูรณ์) เส้นจุดคือขอบเขต $u\sum_k k/60$ ที่ไม่รวมตัวคูณ ความผิดจริงต่ำกว่าขอบเขตตลอดช่วง จุดที่กราฟ (ข) ตกลงต่ำมากคือจุดที่ความผิดผ่านศูนย์ (สร้างจาก `numanal/ch02.py`)

3. **ผลบวกย่อยไม่โตเกินที่ควร.** เราแสดงโดยการเหนี่ยวนำว่า $s_k \le (1+u)^{k-1} S_k$ เมื่อ $k=1$ จริงเพราะ $s_1 = S_1$ สมมติว่าจริงถึง $k-1$ แล้ว จะได้ $t_k = s_{k-1} + x_k \le (1+u)^{k-2} S_{k-1} + x_k \le (1+u)^{k-2} S_k$ เพราะ $(1+u)^{k-2} \ge 1$ จึงได้ $(1+u)^{k-2} S_{k-1} + x_k \le (1+u)^{k-2}(S_{k-1} + x_k)$ จากนั้นปัดได้ $s_k \le (1+u)\,t_k \le (1+u)^{k-1} S_k$ ซึ่งเป็นการเหนี่ยวนำ

4. **รวมเข้าด้วยกัน.** แทนขอบเขตของ t_k จากขั้นที่ 3 ลงในขั้นที่ 2 ได้ $|E_n| \le u\sum_{k=2}^{n}(1+u)^{k-2} S_k \le u(1+u)^{n-2}\sum_{k=2}^{n} S_k$

สรุป: ความผิดรวมเท่ากับผลรวมของความผิดจากการปัดทุกครั้ง และความผิดแต่ละครั้งเป็นสัดส่วนกับผลบวกย่อยตรงนั้น

เกร็ดเพิ่มเติม

ตัวคูณ $(1+u)^{n-2}$ ในทฤษฎีบทเป็นเพียงเศษเล็กน้อยเมื่อ un น้อย ที่นาฬิกาหนึ่งชั่วโมงมันเท่ากับประมาณ 1.013 ขอบเขตจึงประมาณ 23.5 วินาที ส่วนผลรวม $u\sum_{k=2}^{n} k/60$ โดยไม่มีตัวคูณนี้มีค่า 23.17 วินาที ในหนังสือเราใช้ค่าหลังเป็นตัวประมาณ

ผลที่วัดได้ในนาฬิกาของเรา. ตัวเลขทั้งหมดในย่อหน้านี้มาจากโค้ดของเราเอง ให้ x_k เป็นเลข $\frac{1}{60}$ ที่เก็บเป็นความแม่นเดี่ยวทุกเฟรม และอ่านค่าทุกนาที หลังผ่านไปหนึ่งนาทีนาฬิกาอ่าน 59.99926 วินาที หลังสองนาทีอ่าน 120.0112 วินาที ความผิดจึงเปลี่ยนเครื่องหมายได้ ความผิดต่ำสุดคือ -3.28 วินาที ที่ประมาณ $2{,}045$ วินาที หมายถึงนาฬิกาเดินเร็วกว่าเวลาจริง และหลังหนึ่งชั่วโมงนาฬิกาอ่าน 3597.2068 วินาที ทั้งหมดนี้อยู่ในรูป 2.2

ขอบเขตจึงเป็นการรับประกันว่าความผิดจะไม่เกินเท่าไร ไม่ได้บอกว่าจะเกิดเท่าไร ในนาฬิกาของเรา ขอบเขต 23 วินาที กว้างกว่าความผิดจริงราว 8 เท่า หลังหนึ่งชั่วโมงความผิดจริงเพียง 2.79 วินาที

ภัทรกิจ · Preprint · 2026-10-23

ลำดับการบวกมีผล

ลองบวก 1 กับเลข 10^{-8} จำนวนหนึ่งล้านตัว ค่าจริงคือ 1.01 แต่ถ้าบวก 1 ก่อน แล้วบวก 10^{-8} ทีละตัว ผลที่ได้คือ 1.0 เป๊ะ เพราะช่องว่างของเลขความแม่นเดี่ยวที่ 1 คือ $2^{-23} \approx 1.19 \times 10^{-7}$ ครึ่งหนึ่งของช่องนี้มากกว่า 10^{-8} ทุกครั้งที่บวก 10^{-8} เข้าไป ผลจึงปัดกลับมาที่ 1 ทั้งหมด แต่ถ้าบวกเลขเล็กเหล่านั้นก่อน ผลรวมย่อยจะโตจนถึง 0.01 ซึ่งช่องว่างเล็กพอ ผลลัพธ์ที่ได้คือ 1.0099673 ซึ่งใกล้ค่าจริงกว่ามาก

ตัวอย่างนี้ตรงกับคำแนะนำในหนังสือ Atkinson ที่ให้บวกจากเลขเล็กไปใหญ่ Atkinson 1989, §1.5, (1.5.4), p. 30 เหตุผลคือความผิดแต่ละขั้นของการบวกขึ้นกับผลบวกย่อย ถ้าผลบวกย่อยเล็กตั้งแต่ต้น ความผิดก็เล็กตั้งแต่ต้นด้วย

ข้อควรระวัง: เล็กไปใหญ่ไม่ได้รับประกันเสมอไป

การเรียงจากเล็กไปใหญ่ช่วยได้มาก แต่ไม่ได้แก้ทุกอย่าง ลองดู $1 + 2^{-24} + 2^{-24}$ ถ้าบวก 1 ก่อน ผลคือ 1 เพราะ $1 + 2^{-24}$ อยู่กึ่งกลางพอดี และการปัดเข้าหาเลขคู่ทำให้ปัดกลับมาที่ 1 แต่ถ้าบวก 2^{-24} สองตัวก่อนจะได้ 2^{-23} ซึ่งเก็บได้พอดี แล้วบวก 1 ได้ $1 + 2^{-23}$ ซึ่งถูกต้อง ตัวอย่างนี้แสดงว่าการปัดกึ่งกลางอาจทำให้ความผิดทั้งก้อนหายไปได้

ตัดทิ้งหรือปัด: ความเอนเอียง

เครื่องบางแบบไม่ปัด แต่**การตัดทิ้ง** (chopping) คือเก็บเฉพาะหลักที่ทำได้แล้วทิ้งส่วนที่เกินไปเลย ซึ่งทำให้ผลไปทางศูนย์เสมอสำหรับเลขบวก ผลจะเล็กกว่าค่าที่ควรจะเป็นทุกครั้ง ความผิดทุกครั้งจึงมีเครื่องหมายเดียวกัน เราเรียกสิ่งนี้ว่า**ความเอนเอียง** (bias) คือความผิดที่ไปทางเดียวกันซ้ำ ๆ จนผลสะสมไปทางหนึ่ง

Atkinson ให้ข้อสังเกตเชิงสถิติว่า เมื่อปัดเข้าหาเลขที่ใกล้ที่สุด ความผิดรวมมักมีขนาดประมาณ $\sqrt{n}$ แต่เมื่อตัดทิ้ง ความผิดเฉลี่ยจะเป็น n คูณครึ่งหนึ่งของขอบเขตความผิดแต่ละครั้ง Atkinson 1989, §1.5, (1.5.9)–(1.5.10), pp. 31--32 ในนาฬิกาของเรา หลังผ่านไปหนึ่งชั่วโมง การตัดทิ้งให้ความผิด 10.30 วินาที ซึ่งมากกว่าการปัดราว 3.7 เท่า

เรายังตรวจข้อสังเกตของ Atkinson ด้วยการจำลอง โดยสมมติว่าความผิดแต่ละครั้งกระจายสม่ำเสมอบนช่วงหนึ่งหน่วยปัด ซึ่งเป็นสมมติฐาน ไม่ใช่การวัดในเครื่องจริง เมื่อรวม $n = 1000$ ความผิดจากการปัดแล้วทำซ้ำ 4000 ครั้ง ค่าเบี่ยงเบนมาตรฐานของผลรวมคือ 9.21 ใกล้กับ $\sqrt{1000/12} \approx 9.13$ และเมื่อตัดทิ้ง ค่าเฉลี่ยของผลรวมคือ -499.96 ใกล้กับ $-n/2 = -500$ ทั้งสองค่าสอดคล้องกับข้อสังเกตข้างต้น

ในโลกจริง: ดัชนีตลาดแวนคูเวอร์ที่ตัดทิ้งทีละหลัก (1983)

ตลาดหลักทรัพย์แวนคูเวอร์มีดัชนีที่เริ่มในเดือนมกราคม ค.ศ. 1982 ที่ระดับ $1{,}000$ และคำนวณใหม่ทุกครั้งที่ราคาหุ้นเปลี่ยน ในปี 1983 ดัชนีลดลงจนอยู่แถว 520 ผู้บริหารตลาดพบว่าสาเหตุคือการตัดตำแหน่งสุดท้ายทิ้งแทนที่จะปัด หนังสือพิมพ์ The Wall Street Journal (8 พฤศจิกายน 1983) ยกตัวอย่างว่า 540.32567 ควรเก็บเป็น 540.326 แต่ถูกตัดเหลือ 540.325 และรายงานว่าดัชนีลดลงได้ถึงหนึ่งหรือสองจุดทุกวัน Quinn 1983 หนังสือพิมพ์ Toronto Star (29 พฤศจิกายน 1983) รายงานค่าที่คำนวณใหม่ว่า 1098.892 ตัวเลขอื่นในสองฉบับไม่ตรงกันในบางจุด เราจึงใช้เฉพาะเรื่องการตัดทิ้งกับค่าที่คำนวณใหม่ เรื่องนี้ตรงกับความเอนเอียงในหัวข้อนี้ เมื่อค่าเป็นบวก การตัดทิ้งไม่เคยทำให้ค่าสูงขึ้น ความผิดทุกครั้งจึงไปทางเดียวกันและสะสม ฉบับนี้ไม่ได้อ้างว่าตลาดยังคำนวณดัชนีแบบนี้อยู่

การรวมแบบคู่

การรวมแบบคู่ (pairwise summation) คือการจับเลขเป็นคู่ ๆ รวมแต่ละคู่ แล้วรวมผลของคู่เหล่านั้นอีกรอบ ทำซ้ำจนได้ผลเดียว เหมือนการจัดแข่งแบบแพ้คัดออกที่มีรอบเป็นชั้น ๆ ในนาฬิกาหนึ่งชั่วโมงมี 216,000 ตัว ต้นไม้นี้มีความลึกประมาณ 18 ชั้น ดังนั้น

ร่างปลีกย่อย · Preprint · 2026-10-23

ตารางที่ 2.1: ผลของการบวก 216,000 เฟรมด้วยวิธีต่าง ๆ (หน่วยวินาที ตัวเลขจากโค้ดของเรา)

วิธี	ค่าที่อ่านได้หลังหนึ่งชั่วโมง	ความผิด (เวลาจริงลบค่าที่อ่านได้)
ปัดเข้าใกล้ที่สุด (การรวมแบบตรง)	3597.2068	2.7932
ตัดทิ้ง	3589.7031	10.2969
การรวมแบบคู่	3600.000244	-0.000244
การรวมแบบคาฮาน	3600.000244	-0.000244

ผลบวกแต่ละตัวผ่านการบวกไม่เกิน 18 ครั้ง ไม่ใช่ 216,000 ครั้ง

อ่านแบบเข้าใจง่าย: นับการบวกของ x_1

ลองรวม $1, 2, \ldots, 8$ ด้วยการรวมแบบคู่ ชั้นแรกได้ 3, 7, 11, 15 ชั้นที่สองได้ 10 และ 26 ชั้นสุดท้ายได้ 36 ผลบวกย่อยที่มี $x_1 = 1$ อยู่ในนั้นถูกปัดเพียงสามครั้ง คือ $1+2$, $3+7$ และ $10+26$ ส่วนการรวมแบบตรงของตัวเลขเดียวกันมีการปัดต่อเนื่องถึงเจ็ดครั้ง เพราะทุกผลบวกย่อยสืบมาจาก x_1 ต้นไม้จึงทำให้ความลึกของการปัดสั้นลง

ในการทดลองของเรา การรวมแบบคู่ให้ความผิดที่วัดทุกนาทีสูงสุดเพียง 2.44×10^{-4} วินาที ซึ่งเท่ากับช่องว่างของเลขความแม่นเดี่ยวที่ 3600 ซึ่งคือ 2^{-12} เราไม่ได้ทำการวิเคราะห์ความผิดอย่างเข้มงวดของการรวมแบบคู่ในหนังสือเล่มนี้ เราบอกเพียงสิ่งที่วัดได้

การรวมแบบคาฮาน

การรวมแบบคาฮาน (Kahan summation) เป็นการรวมแบบชดเชย (compensated summation) คือเก็บส่วนที่หายไปจากการปัดไว้ แล้วเอาไปบวกคืนในรอบถัดไป แนวคิดนี้มีในบทความของ Kahan ปี 1965 ซึ่งใช้ตัวสะสมตัวที่สอง S_2 เก็บส่วนที่ตกค้างไว้ Kahan 1965 ขั้นตอนด้านล่างเป็น**การนำเสนอของเราเอง**ของการรวมแบบชดเชย (our presentation of compensated summation) ซึ่งใช้ตัวแปรเพียง c ตัวเดียวเก็บส่วนที่หายไป รูปแบบนี้ไม่ได้พิมพ์ไว้ในบทความปี 1965 ทีละบรรทัด ขั้นตอนคือ

1. ตั้ง $s = 0$ และ $c = 0$
2. สำหรับแต่ละ x_k คำนวณ $y = x_k - c$
3. คำนวณ $t = s + y$ แล้วปัด
4. คำนวณ $c = (t - s) - y$ ซึ่งคือส่วนของ y ที่ถูกปัดทิ้งไป
5. ตั้ง $s = t$

แนวคิดของการรวมแบบคาฮาน

ถ้าผลบวก t เก็บส่วนเล็กของ y ไม่ได้ c จะจับความผิดที่เกิดขึ้นไว้ ในรอบถัดไป เราหัก c ออกจาก x_k ก่อนบวก เพื่อชดเชยความผิดของรอบก่อน บทพิสูจน์เต็มว่าวิธีนี้เก็บส่วนที่หายไปได้ถูกต้องเสมออยู่นอกขอบเขตของเล่มนี้ เราแสดงเพียงแนวคิดและตัวเลข

ตัวอย่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย: ลองทำมือสามขั้น

ลองรวม $x_1 = 1$ แล้วตามด้วย $x_2 = x_3 = \cdots = x_{11} = 2^{-24}$ ในความแม่นเดี่ยว ตั้ง $s = 0$ และ $c = 0$ เริ่มจากขั้นแรก

- ใส่ $x_1 = 1$: $y = 1$, $t = 1$ และ $c = 0$
- ใส่ $x_2 = 2^{-24}$: $y = 2^{-24}$ และ $t = 1 + 2^{-24}$ ซึ่งอยู่กึ่งกลางพอดีระหว่าง 1 กับ $1 + 2^{-23}$ เครื่องปัดเข้าเลขคู่จึงได้ $t = 1$ และ $c = (1 - 1) - 2^{-24} = -2^{-24}$ ค่า c เป็นลบหมายความว่าผลที่เก็บไว้น้อยกว่าค่าจริงไป 2^{-24} ขั้นต่อไปจึงหัก c ออกจาก x_k ซึ่งเท่ากับบวกส่วนนั้นคืน
- ใส่ $x_3 = 2^{-24}$: $y = 2^{-24} - (-2^{-24}) = 2^{-23}$ และ $t = 1 + 2^{-23}$ เก็บได้พอดี จึงได้ $c = (t - 1) - 2^{-23} = 0$

ส่วนการบวกตรงจะได้ 1 เป๊ะ เพราะทุกครั้งที่บวก 2^{-24} เข้าไปกับ 1 ผลอยู่กึ่งกลางพอดีแล้วปัดกลับมาที่ 1 สำหรับสิบตัวเล็กหลังเลข 1 การรวมแบบชดเชยให้ $1 + 5 \cdot 2^{-23}$ ซึ่งตรงกับค่าจริงทุกบิต (ทดสอบในโค้ดของเรา)

ในนาฬิกาหนึ่งชั่วโมง การรวมแบบคาฮานให้ 3600.000244 วินาที ซึ่งเป็นค่าที่เก็บได้ใกล้ที่สุดของ 3600 ที่ช่องว่างคือ 2^{-12} ความผิดที่เหลือจริง ๆ คือ 1.8775×10^{-4} วินาที และมาจากการเก็บ $\frac{1}{60}$ เองในเลขความแม่นเดี่ยว ไม่ได้มาจากการบวก ค่าที่ถูกต้องคือ $\frac{1}{60}$ เก็บเป็น $8{,}947{,}849/2^{29}$ ซึ่งเกินกว่า $\frac{1}{60}$ อยู่ $28/(60 \cdot 2^{29})$ ต่อเฟรม เมื่อคูณ $216{,}000$ เฟรมจะได้ $100{,}800/2^{29}$ วินาที

เกร็ดเพิ่มเติม

ถ้าเก็บ $\frac{1}{60}$ และผลรวมเป็นเลขความแม่นคู่แทน ขอบเขตของการรวมแบบตรงในหนึ่งชั่วโมงจะลดเหลือประมาณ 4.3×10^{-8} วินาที คือเล็กกว่าเดิมราว 2^{29} เท่า แบบฝึกหัด 2.1.7 ให้คำนวณตัวเลขนี้เอง

สรุปใจความ

- ค่าคลาดเคลื่อนสัมบูรณ์คือ $|x - \tilde{x}|$ ส่วนสัมพัทธ์คือหารด้วย $|x|$ ค่าคลาดเคลื่อนสัมบูรณ์ที่เล็กในหน่วยหนึ่งอาจใหญ่ในอีกหน่วยหนึ่ง
- ความผิดของการบวกแต่ละครั้งไม่เกินหน่วยปัดคูณผลบวกย่อยตรงนั้น ผลบวกย่อยที่โตขึ้นทำให้ความผิดโตขึ้นด้วย
- นาฬิกาของเราอ่านได้ 3597.2068 วินาทีแบบปัดเข้าใกล้ที่สุด 3589.7 วินาทีแบบตัดทิ้ง ขอบเขตกว้างกว่าความผิดจริงราว 8 เท่า
- การรวมแบบคู่และแบบคาฮานทำให้ความผิดเหลือเพียงระดับช่องว่างของเลขความแม่นเดี่ยว ที่ 3600 คือ 2.44×10^{-4} วินาที
- ลำดับการบวกมีผล ควรบวกจากเล็กไปใหญ่ แต่ไม่ได้รับประกันเสมอไป และความผิดที่ไปทางเดียวกันอันตรายที่สุด

เรากลับไปที่คำถามของบท นาฬิกาเพี้ยนเพราะความผิดจากการปัดแต่ละครั้งรวมกันเป็นวินาที ส่วนว่าเราจะรู้ได้ล่วงหน้าอย่างไร ทฤษฎีบทให้ขอบเขตที่ทำได้ทันทีจากขนาดของผลบวกย่อย ส่วนกล้องที่หายไปจะตอบในหัวข้อ 2.2

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 2.1.1 (⋆) นาฬิกาของเกมอ่านได้ 1200.4 วินาที ทั้งที่เวลาจริงผ่านไป 1203.1 วินาที จงหาค่าคลาดเคลื่อนสัมบูรณ์และค่าคลาดเคลื่อนสัมพัทธ์เป็นเปอร์เซ็นต์ (สี่ตำแหน่ง) จากนั้นถ้าค่าคลาดเคลื่อนสัมพัทธ์นี้คงเดิมตลอด 10 ชั่วโมง นาฬิกาจะคลาดไปกี่วินาที

แบบฝึกหัด 2.1.2 (⋆) ปัดค่า 3597.2068 ให้เหลือ 2, 3 และ 4 หลักนัยสำคัญ ได้ค่าเท่าใดบ้าง ค่าคลาดเคลื่อนสัมบูรณ์ของการปัดเป็น 3600 คือเท่าไร และถ้าเราต้องการรายงานเวลาที่เชื่อได้จริง ควรรายงานกี่หลัก

แบบฝึกหัด 2.1.3 (⋆⋆) บวกเลข 0.3 ที่เก็บเป็นความแม่นเดี่ยว สิบครั้ง ในเครื่องได้ผลเท่าไร ค่าคลาดเคลื่อนสัมบูรณ์เมื่อเทียบกับ 3 เท่าไร จงแยกความผิดนี้เป็นสองส่วน คือส่วนจากการเก็บ 0.3 และส่วนจากการปัดในการบวก แล้วอธิบายว่าทำไมสองส่วนนี้มีเครื่องหมายต่างกัน

แบบฝึกหัด 2.1.4 (⋆⋆) ให้ x_1, x_2, x_3, x_4 เป็นเลขความแม่นเดี่ยวที่ไม่ติดลบ และให้ s_k เป็นผลบวกย่อยที่คำนวณทีละขั้น จงแสดงด้วยมือว่า

$$s_4 - (x_1 + x_2 + x_3 + x_4) = r_2 + r_3 + r_4,$$

เมื่อ $r_k = s_k - (s_{k-1} + x_k)$ คือค่าคลาดเคลื่อนการปัดเศษของการบวกครั้งที่ k จากนั้นรันโค้ดกับเลขสุ่มสี่ตัวจากสามค่าของ seed เพื่อดูว่าเอกลักษณ์นี้เป็นจริง

แบบฝึกหัด 2.1.5 (⋆⋆) บวก 1 กับ 2^{-24} สี่ครั้ง ในความแม่นเดี่ยว ตามลำดับสองแบบ คือ (ก) บวก 1 ก่อน แล้วตามด้วย 2^{-24} สี่ตัว และ (ข) บวก 2^{-24} สี่ตัวก่อน แล้วตามด้วย 1 จงหาผลของแต่ละแบบ เทียบกับค่าจริง $1 + 4 \cdot 2^{-24} = 1 + 2^{-22}$ และอธิบายด้วยการปัดเข้าหาเลขคู่ว่าทำไมแบบ (ก) ให้ผลต่างออกไป

แบบฝึกหัด 2.1.6 (⋆⋆) ใช้ขอบเขต $u \sum_{k=2}^{n} k/60$ โดย $u = 2^{-24}$ (ไม่รวมตัวคูณ $(1+u)^{n-2}$) จงหาจำนวนเฟรมน้อยที่สุด n ที่ทำให้ขอบเขตนี้เกิน 10^{-3} วินาที แล้วรันนาฬิกาที่บวก $\frac{1}{60}$ ทีละเฟรมจนถึง n เฟรม เปรียบเทียบความผิดที่เกิดจริงกับขอบเขต และบอกว่าควรใช้ขอบเขตนี้ตัดสินใจอย่างไร

แบบฝึกหัด 2.1.7 (⋆⋆⋆) ย้อนกลับไปที่นาฬิกาหนึ่งชั่วโมง การรวมแบบคาฮานเหลือความผิด 1.8775×10^{-4} วินาที

1. จงแสดงว่าความผิดนี้มาจากการเก็บ $\frac{1}{60}$ เป็นความแม่นเดี่ยว โดยคำนวณ $216{,}000 \cdot \left(\mathrm{fl}_{32}(\frac{1}{60}) - \frac{1}{60}\right)$ ในรูปเศษส่วนของ 2
2. ถ้าเก็บ $\frac{1}{60}$ เป็นเลขความแม่นคู่ ความคลาดเคลื่อนจากการเก็บในหนึ่งชั่วโมงมีขนาดเท่าใด
3. ถ้าการบวกทั้งหมดเป็นความแม่นคู่ ขอบเขตของความผิดจากการบวกในหนึ่งชั่วโมงคือเท่าใด เมื่อใช้ $u = 2^{-53}$ และขอบเขตนี้เล็กกว่าขอบเขตความแม่นเดี่ยวกี่เท่า

ตัวช่วยคิด

เขียน $\mathrm{fl}_{32}(\frac{1}{60}) = 8{,}947{,}849/2^{29}$ แล้วใช้ตัวส่วนร่วม $60 \cdot 2^{29}$ ผลต่างจะเป็นเศษส่วนที่มีเศษเล็กมาก ส่วนขอบเขตของความแม่นคู่ดึงมาจากตัวอย่างในหัวข้อนี้โดยเปลี่ยน u เท่านั้น

สมการ · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 2.1.1.

1. ค่าคลาดเคลื่อนสัมบูรณ์คือ $|1203.1 - 1200.4| = 2.7$ วินาที
2. ค่าคลาดเคลื่อนสัมพัทธ์คือ $2.7/1203.1 \approx 0.0022442$ คือ 0.2244 เปอร์เซ็นต์ (สี่ตำแหน่ง)
3. 10 ชั่วโมงเท่ากับ $36{,}000$ วินาที ถ้าค่าคลาดเคลื่อนสัมพัทธ์เท่าเดิม ความผิดคือ $36{,}000 \times 0.0022442 \approx 80.8$ วินาที (เป็นการสมมติว่าค่าคลาดเคลื่อนสัมพัทธ์คงที่ ซึ่งเป็นการประมาณ)

ตรวจด้วยการคำนวณ: `test_ch02.py`, `test_ex_2_1_1_clock`.

เฉลย 2.1.2.

1. สองหลักนัยสำคัญ: 3.6×10^3 คือ 3600
2. สามหลักนัยสำคัญ: 3.60×10^3 คือ 3600 เช่นกัน
3. สี่หลักนัยสำคัญ: 3597.0 เพราะหลักถัดไปคือ 2 จึงปัดลง
4. ค่าคลาดเคลื่อนสัมบูรณ์ของ 3600 คือ $|3597.2068 - 3600| = 2.7932$ วินาที
5. ถ้าความผิดไม่เกิน 2.79 วินาที ค่าจริงอยู่ในช่วง $[3594.4,\ 3600.0]$ ค่าทั้งหมดในช่วงนี้ปัดเป็นสองหลักได้ 3.6×10^3 แต่ปัดเป็นสามหลักไม่ได้ เพราะ 3594 ปัดเป็น 3.59×10^3 ดังนั้นเราควรรายงานเพียงสองหลัก

ตรวจด้วยการคำนวณ: `test_ex_2_1_2_digits`.

เฉลย 2.1.3.

1. $\mathrm{fl}_{32}(0.3) = 0.300000011920929\ldots$ สูงกว่า 0.3 อยู่ 1.192×10^{-8} ถ้าบวกค่าที่เก็บไว้สิบครั้งโดยไม่มีการปัด ผลคือ $10 \cdot \mathrm{fl}_{32}(0.3) = 3.000000119\ldots$
2. ผลที่เครื่องได้คือ $2.999999761581421\ldots \approx 2.9999997616$ ค่าคลาดเคลื่อนสัมบูรณ์เมื่อเทียบกับ 3 คือ $3 - 2.9999997616 = 2.384 \times 10^{-7}$
3. แยกความผิดเป็นสองส่วน ส่วนจากการเก็บ 0.3 คือ $10 \cdot \mathrm{fl}_{32}(0.3) - 3 = +1.192 \times 10^{-7}$ และส่วนจากการปัดในการบวกคือ $2.9999997616 - 10 \cdot \mathrm{fl}_{32}(0.3) = -3.576 \times 10^{-7}$ รวมกันได้ -2.384×10^{-7} ซึ่งเป็นผลที่เครื่องได้ลบด้วย 3 ตรงกับข้อ 2 ทั้งสองส่วนมีเครื่องหมายต่างกัน
4. เหตุที่เครื่องหมายต่างกันคือ $\mathrm{fl}_{32}(0.3)$ สูงกว่าค่าจริง ทำให้ค่าที่เก็บไว้รวมกันสูงไปเล็กน้อย แต่การปัดในการบวกครั้งนี้ปัดลงจนผลรวมต่ำกว่าผลรวมของค่าที่เก็บไว้ ส่วนที่ปัดลงมีขนาดมากกว่า ความผิดสุทธิจึงเป็นลบ

ตรวจด้วยการคำนวณ: `test_ex_2_1_3_split`.

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 2.1.4.

1. ตั้ง $s_1 = x_1$ ซึ่งแม่นตรง จึงได้ $s_1 - S_1 = 0$
2. จาก $t_k = s_{k-1} + x_k$ และ $S_k = S_{k-1} + x_k$ จะได้ $t_k - S_k = s_{k-1} - S_{k-1}$
3. จาก $s_k = t_k + r_k$ จะได้ $s_k - S_k = (t_k - S_k) + r_k = (s_{k-1} - S_{k-1}) + r_k$
4. แทนซ้ำจาก $k = 2$ ถึง 4 ได้ $s_4 - S_4 = (s_1 - S_1) + r_2 + r_3 + r_4 = r_2 + r_3 + r_4$

เอกลักษณ์นี้มีรูปแบบเดียวกับสมการ (1.5.4) ใน Atkinson 1989, §1.5, p. 30 ซึ่งเขียนความผิดของผลรวมเป็นผลรวมของความผิดแต่ละขั้น ตรวจด้วยการคำนวณ: `test_ex_2_1_4_identity` กับ seed 1 ถึง 5 โดยเทียบกับผลรวมที่แม่นตรงของเศษส่วน

เฉลย 2.1.5.

1. แบบ (ก) บวก 1 ก่อน: $1 + 2^{-24}$ อยู่กึ่งกลางระหว่าง 1 กับ $1 + 2^{-23}$ พอดี การปัดเข้าหาเลขคู่เลือก 1 ซึ่งส่วนเศษเป็นศูนย์ ดังนั้นได้ 1 ทุกครั้งที่บวก 2^{-24} เข้าไป ผลสุดท้ายคือ 1.0 ความผิดคือ $2^{-22} \approx 2.38 \times 10^{-7}$
2. แบบ (ข) บวก 2^{-24} ก่อน: ได้ $2^{-24} + 2^{-24} = 2^{-23}$ แล้วได้ 2^{-22} โดยไม่ผิด จากนั้นบวก 1 ได้ $1 + 2^{-22}$ ซึ่งเก็บได้ตรง ไม่มีความผิด
3. ในแบบ (ก) ทุกครั้งที่บวก 2^{-24} เข้ากับ 1 มันอยู่กึ่งกลางพอดีและถูกปัดกลับเป็น 1 ความผิดทั้งหมดจึงหายไป ส่วนแบบ (ข) ผลบวกย่อยโตจนถึง 2^{-22} ก่อนที่จะรวมกับ 1 จึงไม่หายไป

ตรวจด้วยการคำนวณ: `test_ex_2_1_5_tiny`.

เฉลย 2.1.6.

1. ขอบเขต $B(n) = u \sum_{k=2}^{n} k/60$ ประมาณได้ $u\, n^2/120$ ดังนั้นต้องการ $n^2 \approx 120 \cdot 10^{-3}/u = 0.12 \cdot 2^{24} \approx 2.01 \times 10^6$ จึงได้ n ประมาณ 1419 ตรวจด้วยโค้ดได้ $B(1418) < 10^{-3} \le B(1419) \approx 1.00085 \times 10^{-3}$ ดังนั้น $n = 1419$ เฟรม หรือ 23.65 วินาที
2. รันนาฬิกา 1419 เฟรม ค่าที่อ่านได้คือ 23.6498108 วินาที ความผิดจริงคือ 1.892×10^{-4} วินาที หรือประมาณ 0.19 มิลลิวินาที น้อยกว่าขอบเขตประมาณ 5.3 เท่า
3. ขอบเขตเป็นกรณีเลวร้ายสุดที่รับประกันได้ ไม่ได้บอกว่าความผิดจริงจะเป็นเท่าใด ถ้าต้องการรับประกันว่าความผิดไม่เกิน 1 มิลลิวินาที นาฬิกาต้องถูกรีเซ็ตหรือปรับใหม่ก่อนครบ 1419 เฟรม ซึ่งประมาณ 23 วินาที

ตรวจด้วยการคำนวณ: `test_ex_2_1_6_bound` และ `test_ex_2_1_6_observed`.

เฉลย 2.1.7.

1. จาก $\mathrm{fl}_{32}(\tfrac{1}{60}) = \tfrac{8{,}947{,}849}{2^{29}}$ จะได้

$$\mathrm{fl}_{32}\!\left(\tfrac{1}{60}\right) - \tfrac{1}{60} = \frac{60 \cdot 8{,}947{,}849 - 2^{29}}{60 \cdot 2^{29}} = \frac{536{,}870{,}940 - 536{,}870{,}912}{60 \cdot 2^{29}} = \frac{28}{60 \cdot 2^{29}} = \frac{7}{15 \cdot 2^{29}}.$$

   คูณ $216{,}000$ ได้ $\dfrac{216{,}000 \cdot 7}{15 \cdot 2^{29}} = \dfrac{100{,}800}{2^{29}} \approx 1.8775 \times 10^{-4}$ วินาที ซึ่งเป็นความผิดที่เหลือในการรวมแบบคาฮาน
2. ความคลาดเคลื่อนจากการเก็บ $\tfrac{1}{60}$ เป็นความแม่นคู่ต่อเฟรมมีขนาดประมาณ 2.3×10^{-19} วินาที คูณ $216{,}000$ ได้ประมาณ 5×10^{-14} วินาที ซึ่งเล็กจนไม่มีผลในทางปฏิบัติ (ตรวจด้วยการคำนวณ)

อัปเดตล่าสุด · Preprint · 2026-10-23

3. ขอบเขตความแม่นคู่คือ $2^{-53}\sum_{k=2}^{216{,}000} k/60 \approx 2^{-53}\cdot 23.17\cdot 2^{24} = 23.17\cdot 2^{-29} \approx 4.3\times 10^{-8}$ วินาที เมื่อเทียบกับความแม่นเดี่ยว ขอบเขตลดลง $2^{-24}/2^{-53} = 2^{29} \approx 5.4\times 10^{8}$ เท่า

ตรวจด้วยการคำนวณ: `test_ex_2_1_7_leftover` และ `test_ex_2_1_7_float64`.

2.2 ลบกันแล้วความแม่นหายไป

สถานการณ์สมมติ: หันกล้องไปหาเป้าหมาย

กล้องในเกมมีทิศที่มันมองอยู่ เรียกว่า a และมีทิศไปหาเป้าหมาย เรียกว่า b ทั้งสองเป็น**เวกเตอร์หนึ่งหน่วย** (unit vector) คือมีความยาวเท่ากับ 1 มุม θ ที่กล้องต้องหมุนหาได้จาก $\theta = \arccos(a \cdot b)$ โดย $a \cdot b$ คือ**ผลคูณจุด** (dot product) ของสองเวกเตอร์ และ $\arccos$ คือ**อาร์กโคไซน์** (arccos) ซึ่งเป็นฟังก์ชันย้อนกลับของโคไซน์ สูตรนี้ถูกต้องบนกระดาษ แต่ในเครื่องที่เก็บตัวเลขแบบความแม่นเดี่ยว เกิดเรื่องสามอย่าง อย่างแรก ถ้ากล้องหันตรงเป้าหมายพอดี มุมที่ถูกต้องคือ 0 แต่ผลคูณจุด $a \cdot a$ ของเวกเตอร์เดียวกันอาจได้ 1.0000001192 ซึ่งเกิน 1 ไปนิดเดียว $\arccos$ ของเลขที่เกิน 1 ไม่มีค่า เครื่องจึงตอบ NaN และกล้องหายไปจากจอ ตัวอย่างนี้ใช้เวกเตอร์แรกที่สุ่มด้วย seed 4 แล้วทำให้มีความยาวหนึ่ง คือ $(-0.36304104,\ -0.09731575,\ 0.92667735)$
อย่างที่สอง ถ้ากล้องหมุนไปเพียงประมาณ 1.0001×10^{-4} เรเดียน $\arccos$ ให้ 0 เป๊ะ ๆ อย่างที่สาม สูตร $\operatorname{atan2}(|a\times b|,\ a\cdot b)$ ให้มุมนี้ที่ผิดไปไม่ถึง 3×10^{-9} เรเดียน
สองสูตรเท่ากันบนกระดาษ แต่ในเครื่องไม่เท่ากัน ทำไม และเราจะรู้ได้อย่างไรว่าสูตรไหนปลอดภัย นี่คือคำถามของหัวข้อนี้ ตัวเลขทุกตัวในหัวข้อนี้มาจากโค้ดของเราเอง

ภาพรวมก่อนอ่าน

ลองนึกถึงการวัดความสูงของเพื่อนสองคนที่สูงเท่ากันเกือบพอดี แล้วเอาความสูงมาลบกัน ผลต่างจะเล็กมาก และความผิดเล็กน้อยของการวัดแต่ละครั้งอาจมีขนาดพอ ๆ กับผลต่างนั้นเอง ตัวเลขแต่ละตัวอาจแม่นดีแล้ว แต่ผลต่างที่เล็กมากจะไม่แม่นเลย
ในหัวข้อนี้เราจะดูว่าการลบแบบนี้ทำให้ความแม่นหายไปได้อย่างไร แล้ววัดว่าความผิดถูกขยายเท่าไรด้วย**อนุพันธ์** (derivative) ซึ่งบอกว่าค่าของฟังก์ชันเปลี่ยนไปเท่าไรเมื่อข้อมูลเปลี่ยนนิดเดียว เราจะใช้ตัวเลขนี้อธิบายทั้งกล้องที่หายไปและมุมเล็กที่ถูกปัดเป็นศูนย์

การลบแบบหักล้าง

การลบแบบหักล้าง (cancellation) คือการลบตัวเลขสองตัวที่ใกล้กันมาก ส่วนหน้าที่เหมือนกันจึงหักล้างกันไป เหลือเพียงผลต่างเล็ก ๆ ผลต่างที่ได้จึงเล็กเมื่อเทียบกับความผิดของตัวตั้งแต่ละตัว ความผิดของตัวตั้งแต่ละตัวยังอยู่ในผลต่าง ไม่ได้หายไป และค่าสัมบูรณ์ของความผิดในผลต่างไม่เกินผลรวมของค่าคลาดเคลื่อนสัมบูรณ์ของตัวตั้งสองตัว (ดูทฤษฎีบทด้านล่าง) เมื่อเอาความผิดขนาดนี้ไปหารด้วยผลต่างที่เล็กลง ค่าคลาดเคลื่อนสัมพัทธ์จึงโตขึ้นมาก แนวคิดนี้มาจาก Atkinson หัวข้อ 1.4 Atkinson 1989, §1.4, pp. 23--26 ส่วนตัวอย่างและตัวเลขในหัวข้อนี้เป็นของเราเอง

ลองดูตัวอย่างเล็ก ๆ ก่อน

ลองเอง: ลบค่าที่วัดได้

วัดได้ 5.1234 และ 5.1229 โดยแต่ละค่าปัดเป็นทศนิยมสี่ตำแหน่ง ค่าจริงของแต่ละตัวจึงห่างจากค่าที่วัดได้ไม่เกิน 0.00005 ผลต่างที่วัดได้คือ 0.0005 ลองเดาก่อนว่าผลต่างจริงอยู่ในช่วงไหน แล้วหาว่าค่าคลาดเคลื่อนสัมพัทธ์สูงสุดของ

ฉบับร่าง · Preprint · 2026-10-23

ผลต่างเป็นกี่เปอร์เซ็นต์ ทั้งที่แต่ละค่าที่วัดแม่นถึงประมาณ 0.001 เปอร์เซ็นต์

คำตอบคือ ผลต่างจริงอยู่ในช่วง $[0.0004,\ 0.0006]$ และค่าคลาดเคลื่อนสัมพัทธ์สูงสุดคือ $0.0001/0.0005 = 20$ เปอร์เซ็นต์ ค่าคลาดเคลื่อนสัมบูรณ์ของผลต่างไม่เกินผลรวมของความผิดของตัวตั้งสองตัว ซึ่งคือราว 0.0001 และยังเท่าเดิม แต่ผลต่างนั้นเล็กกว่าตัวตั้งราวหนึ่งหมื่นเท่า ค่าคลาดเคลื่อนสัมพัทธ์จึงใหญ่ขึ้นราว 2×10^4 เท่า

ความผิดส่งผ่านการคำนวณ

เพื่อวัดเป็นตัวเลข เราจะถามว่าความผิดของข้อมูลถูกขยายขึ้นเท่าไรเมื่อผ่านการลบหรือการคำนวณอื่น เริ่มจากการลบเอง

ทฤษฎีบท 2.2.1: ค่าคลาดเคลื่อนสัมพัทธ์ของการลบ

ให้ $x \neq y$ และให้ค่าที่เก็บไว้เป็น $\tilde{x} = x + e_1$ กับ $\tilde{y} = y + e_2$ โดย $|e_1| \le \rho|x|$ และ $|e_2| \le \rho|y|$ ถ้า $\tilde{x} \neq \tilde{y}$ แล้ว

$$\frac{|(\tilde{x} - \tilde{y}) - (x - y)|}{|x - y|} \le \frac{|x| + |y|}{|x - y|}\,\rho.$$

อ่านแบบเข้าใจง่าย

- ρ คือค่าคลาดเคลื่อนสัมพัทธ์ของข้อมูลแต่ละตัว ส่วน $\dfrac{|x| + |y|}{|x - y|}$ คือตัวขยายความผิดของการลบ ตัวนี้คือคอนดิชันนัมเบอร์ของการลบ ซึ่งนิยามทั่วไปในกรอบถัดไป
- ถ้า x กับ y ใกล้กันมาก $|x - y|$ เล็กมาก ตัวขยายนี้จึงใหญ่ ในกรอบ ``ลองเอง" ตัวขยายคือ $(5.1234 + 5.1229)/0.0005 \approx 2 \times 10^4$
- ผลลัพธ์ของการลบยังถูกปัดอีกครั้ง ซึ่งเพิ่มค่าคลาดเคลื่อนสัมพัทธ์อีกไม่เกิน u ตัวนี้มักเล็กเมื่อเทียบกับตัวขยาย

พิสูจน์

แนวคิด. ความผิดของผลต่างไม่เกินผลรวมของความผิดของตัวตั้งสองตัว แล้วหารด้วยขนาดของผลต่างจริง

1. **เขียนความผิดของผลต่าง.** $(\tilde{x} - \tilde{y}) - (x - y) = (x + e_1 - y - e_2) - (x - y) = e_1 - e_2$
2. **ใช้อสมการสามเหลี่ยม.** $|e_1 - e_2| \le |e_1| + |e_2| \le \rho(|x| + |y|)$
3. **หารด้วย** $|x - y|$**.** เมื่อหารทั้งสองข้างด้วย $|x - y| > 0$ ได้ขอบเขตที่ต้องการ

สรุป: ความผิดของผลต่างไม่เกินผลรวมของความผิดของตัวตั้ง และเมื่อหารด้วยผลต่างจริง ตัวขยายคือ $(|x| + |y|)/|x - y|$

ทฤษฎีบทนี้เป็นกรณีพิเศษของกฎทั่วไปสำหรับฟังก์ชันหนึ่งตัวแปร ถ้าฟังก์ชัน f หาอนุพันธ์ได้ ตัวขยายของมันหาได้จากอนุพันธ์

นิยาม 2.2.1: คอนดิชันนัมเบอร์ของฟังก์ชันหนึ่งตัวแปร

ให้ f หาอนุพันธ์ได้ที่ x โดย $x \neq 0$ และ $f(x) \neq 0$ นิยาม

$$\kappa_f(x) = \left| \frac{x\, f'(x)}{f(x)} \right|$$

ฉบับร่าง · Preprint · 2026-10-23

เรียก $\kappa_f(x)$ ว่า**คอนดิชันนัมเบอร์** (condition number) ของ f ที่ x

อ่านแบบเข้าใจง่าย

- $\kappa_f(x)$ บอกว่าค่าคลาดเคลื่อนสัมพัทธ์ของข้อมูลจะถูกคูณขึ้นกี่เท่าเมื่อผ่าน f เช่น $f(x) = x^2$ ได้ $\kappa = 2$ ถ้าข้อมูลผิด 1% ผลลัพธ์จะผิดราว 2%
- ถ้า κ ใหญ่ ฟังก์ชันตรงจุดนั้นไวต่อข้อมูลมาก ในหัวข้อ 2.3 เราจะใช้คำนี้กับปัญหาทั้งก้อน ไม่ใช่กับฟังก์ชันเดียว

ทฤษฎีบท 2.2.2: การส่งความผิดผ่านอนุพันธ์

ให้ f หาอนุพันธ์อันดับสองได้ใกล้ x โดย $x \neq 0$ และ $f(x) \neq 0$ ให้ $\tilde{x} = x(1+\delta)$ เมื่อ δ เล็กมาก แล้ว

$$\frac{|f(\tilde{x}) - f(x)|}{|f(x)|} = \kappa_f(x)\,|\delta| + O(\delta^2).$$

พิสูจน์

แนวคิด. การกระจายเทย์เลอร์รอบ x ให้เห็นว่าการเปลี่ยนค่าเล็กน้อยทำให้ผลเปลี่ยนไปเท่าไร

1. **การกระจายเทย์เลอร์.** จาก $\tilde{x} - x = x\delta$ ได้ $f(\tilde{x}) = f(x) + f'(x)\,x\delta + R$ โดย $|R| \le C\delta^2$ สำหรับค่าคงที่ C เมื่อ δ เล็ก เพราะ f หาอนุพันธ์อันดับสองได้
2. **หารด้วย** $f(x)$. $\dfrac{f(\tilde{x}) - f(x)}{f(x)} = \dfrac{x f'(x)}{f(x)}\,\delta + \dfrac{R}{f(x)}$ ดังนั้นขนาดของสองข้างต่างกันไม่เกิน $|R|/|f(x)|$
3. **ใช้อสมการสามเหลี่ยม.** ค่าสัมบูรณ์ของพจน์แรกคือ $\kappa_f(x)|\delta|$ จึงได้ $\left|\dfrac{|f(\tilde{x}) - f(x)|}{|f(x)|} - \kappa_f(x)|\delta|\right| \le \dfrac{C}{|f(x)|}\,\delta^2$ นี่คือความหมายของ $O(\delta^2)$

สรุป: การเปลี่ยนค่าเล็กน้อยทำให้ผลเปลี่ยนสัดส่วนกับ κ คูณ $|\delta|$ ส่วนที่เหลือเล็กกว่า δ^2

มุมจากผลคูณจุด

กลับไปที่กล้อง ผลคูณจุด $d = a \cdot b$ ถูกเก็บเป็นความแม่นเดี่ยว ความผิดของมันเป็นระดับหน่วยปัด $u = 2^{-24} \approx 6 \times 10^{-8}$ และมุมคือ $\theta = \arccos d$ ตัวขยายที่ต้องดูคือ κ ของ $\arccos$ ที่จุด $d = \cos\theta$ จาก $\arccos'(d) = -1/\sqrt{1-d^2}$ และ $\sqrt{1-\cos^2\theta} = \sin\theta$ เมื่อ $0 < \theta < \pi$ ได้

$$\kappa_{\arccos}(\cos\theta) = \frac{\cos\theta}{\theta\,\sin\theta} \approx \frac{1}{\theta^2} \qquad (\theta \text{ เล็ก}).$$

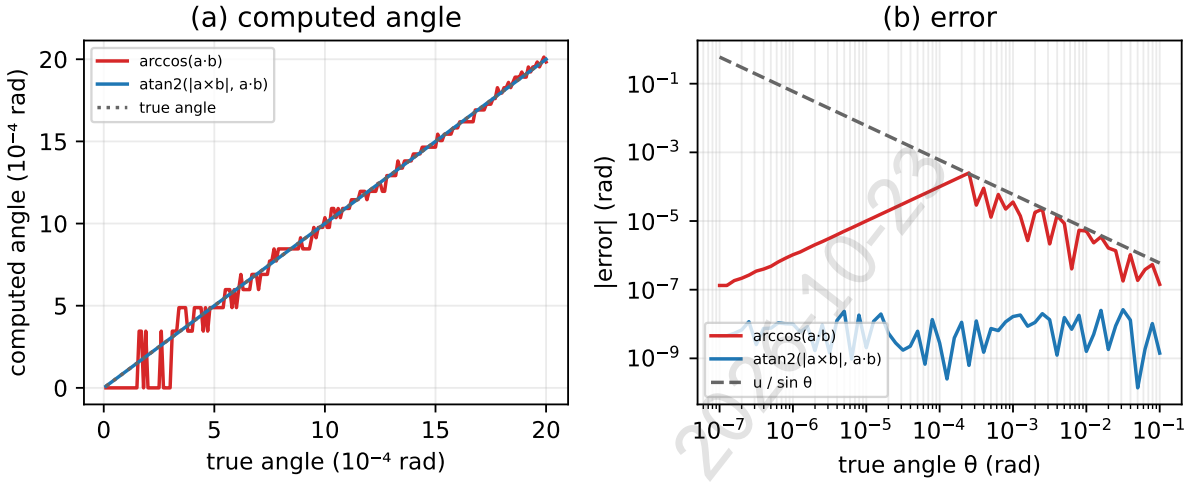
ตารางที่ 2.2 แสดงว่าความผิดของผลคูณจุดระดับ u ทำให้มุมผิดได้เท่าไร

ที่ $\theta = 10^{-4}$ ความผิดของผลคูณจุดเพียงหนึ่งหน่วยปัดทำให้มุมผิดได้ราวหกเท่าของตัวมันเอง และในเครื่องจริง ผลคูณจุดปัดเป็น 1 พอดี $\arccos 1 = 0$ จึงได้มุมศูนย์ นี่คือเหตุผลของการหยุดนิ่งในกล่องสถานการณ์ ส่วน NaN เกิดเพราะผลคูณจุดปัดแล้วกลายเป็น 1.0000001 ซึ่งอยู่นอกช่วงที่ $\arccos$ นิยามไว้

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 2.2: ตัวขยาย κ ของ arccos และค่าคลาดเคลื่อนสัมพัทธ์ที่เป็นไปได้ของมุม $u\kappa$ เมื่อผลคูณจุดคลาด $u = 2^{-24}$

มุมจริง θ (เรเดียน)	κ โดยประมาณ	$u\kappa$ โดยประมาณ
0.1	10^{2}	6×10^{-6}
0.01	10^{4}	6×10^{-4}
0.001	10^{6}	0.06
10^{-4}	10^{8}	6



รูปที่ 2.3: (ก) มุมที่คำนวณได้ของ arccos เทียบกับมุมจริง เมื่อมุมเล็ก arccos ติดอยู่ที่ศูนย์เป็นช่วง ๆ เส้นจุดคือมุมจริง (ข) ค่าคลาดเคลื่อนสัมบูรณ์เทียบกับมุมจริง แบบลอการิทึม เส้นประคือ $u/\sin\theta$ ความผิดของ arccos ตามเส้นนี้ และที่มุมเล็กมาก ๆ ความผิดเท่ากับมุมเอง ส่วน atan2 อยู่ในระดับ 10^{-8} หรือเล็กกว่าตลอดช่วง (สร้างจาก `numanal/ch02.py`)

สูตรที่เลี่ยงการหักล้าง

ปัญหาของ arccos คืออ่านมุมเล็กจากโคไซน์ ซึ่งเปลี่ยนน้อยมากใกล้จุด 1 สูตร `atan2` อ่านมุมจากไซน์ ซึ่งเปลี่ยนไปตามมุมโดยตรง ถ้า a และ b เป็นเวกเตอร์หนึ่งหน่วย $a \cdot b = \cos\theta$ และ $|a \times b| = \sin\theta$ ดังนั้น

$$\theta = \operatorname{atan2}\big(|a \times b|,\ a \cdot b\big).$$

ผลคูณไขว้ก็มีการลบเลขที่ใกล้กันอยู่ข้างใน แต่ความผิดที่เกิดขึ้นเป็นระดับ u แบบสัมบูรณ์ และสูตรนี้ไม่ได้ขยายความผิดตามตัวขยายที่โตขึ้นจากการลบ ทฤษฎีบทต่อไปนี้บอกเหตุผลนี้

ทฤษฎีบท 2.2.3: ความผิดของ atan2 บนวงกลมหนึ่งหน่วย

ให้ $x = \cos\theta$ และ $y = \sin\theta$ เมื่อ $0 < \theta < \pi/2$ และให้ค่าที่เก็บไว้เป็น $\tilde{x} = x + \varepsilon_x$ กับ $\tilde{y} = y + \varepsilon_y$ โดย ε_x และ ε_y เล็กมาก แล้ว

$$\operatorname{atan2}(\tilde{y}, \tilde{x}) - \theta = \cos\theta\,\varepsilon_y - \sin\theta\,\varepsilon_x + O(\varepsilon^2),$$

ดังนั้นความผิดของมุมไม่เกิน $|\varepsilon_x| + |\varepsilon_y|$ บวกพจน์ระดับ ε^2

พิสูจน์

แนวคิด. หาอนุพันธ์ย่อยของ `atan2` ที่จุดบนวงกลม แล้วใช้การกระจายเทย์เลอร์ลำดับแรก

ฉบับร่าง · Preprint · 2026-10-23

1. **อนุพันธ์ย่อย.** เมื่อ $x > 0$ จะได้ $\operatorname{atan2}(y, x) = \arctan(y/x)$ ดังนั้น
$$\frac{\partial \theta}{\partial y} = \frac{1/x}{1 + y^2/x^2} = \frac{x}{x^2 + y^2}, \qquad \frac{\partial \theta}{\partial x} = \frac{-y/x^2}{1 + y^2/x^2} = \frac{-y}{x^2 + y^2}$$
2. **ใช้จุดบนวงกลม.** เพราะ $x^2 + y^2 = 1$ จึงได้ $\partial\theta/\partial y = x = \cos\theta$ และ $\partial\theta/\partial x = -y = -\sin\theta$
3. **การกระจายเทย์เลอร์.** จากขั้น 2 ได้ $\operatorname{atan2}(\tilde{y}, \tilde{x}) - \theta = \cos\theta\,\varepsilon_y - \sin\theta\,\varepsilon_x + O(\varepsilon^2)$ และเพราะ $|\cos\theta| \le 1$ กับ $|\sin\theta| \le 1$ จึงได้ขอบเขตที่ต้องการ

สรุป: ความผิดของมุมเท่ากับความผิดของ x และ y ที่ถ่วงด้วยค่าไม่เกิน 1 โดยไม่มีตัวขยายที่โตตามมุม

ผลนี้อธิบายรูป 2.3 ได้ ความผิดของ $a \cdot b$ และ $|a \times b|$ จากการคำนวณในความแม่นเดี่ยวมีขนาดระดับ u แบบสัมบูรณ์ ไม่ว่ามุมจะใหญ่หรือเล็ก จึงได้ความผิดของมุมระดับ u ด้วย

ข้อควรระวัง: atan2 แม่นแบบสัมบูรณ์ ไม่ใช่สัมพัทธ์เสมอไป

ที่ $\theta = 10^{-6}$ สูตร `atan2` ให้ค่าภายใน 10^{-8} เรเดียน ฟังดูเล็กมาก แต่เทียบกับมุมจริงขนาด 10^{-6} นี่คือความผิดราว 1% ดังนั้นก่อนเลือกสูตร ให้ถามว่าเราต้องการความแม่นแบบสัมบูรณ์หรือสัมพัทธ์

เขียนสูตรใหม่ให้ไม่หักล้าง

สูตรรากของสมการกำลังสองเป็นตัวอย่างคลาสสิกของการลบแบบหักล้าง ลองแก้ $x^2 - 1000x + 1 = 0$ รากคือ $(1000 \pm \sqrt{999996})/2$ รากใหญ่ใกล้ 1000 ไม่มีปัญหา แต่รากเล็กใกล้ 0.001 ต้องลบ 1000 ด้วย $\sqrt{999996}$ ซึ่งใกล้กันมาก ในความแม่นเดี่ยวสูตรเดิมให้ 0.00100708 ทั้งที่รากจริงคือ 0.00100000100 ค่าคลาดเคลื่อนสัมพัทธ์จึงเป็น 0.71% สาเหตุคือ $\sqrt{999996}$ ถูกเก็บด้วยความผิดราว 1.4×10^{-5} ซึ่งหารสองแล้วเป็น 7×10^{-6} เทียบกับรากเล็กขนาด 10^{-3}

วิธีแก้คือไม่ลบ รากทั้งสองมีผลคูณเท่ากับ c เพราะ $(x - r_1)(x - r_2) = x^2 - (r_1 + r_2)x + r_1 r_2$ ดังนั้นถ้า r_1 เป็นรากใหญ่ รากเล็กคือ c/r_1 รากใหญ่ $(1000 + \sqrt{999996})/2 \approx 999.999$ คำนวณได้โดยการบวกเลขบวกสองตัว ไม่มีการหักล้าง ผลที่ได้คือ 0.00100000098 ซึ่งมีค่าคลาดเคลื่อนสัมพัทธ์ราว 2×10^{-8}

มุมเล็กในสูตรโคไซน์

ตัวอย่างเดียวกันเกิดกับ $1 - \cos x$ เมื่อ x เล็ก ในกล้องของเรา $1 - d$ คือผลต่างแบบเดียวกัน ถ้าเขียน $1 - \cos x$ ตรง ๆ ในความแม่นเดี่ยว จะได้ 4.768×10^{-7} เมื่อ $x = 10^{-3}$ ซึ่งผิดจาก 5.000×10^{-7} ราว 4.6% และได้ 0 เมื่อ $x = 10^{-4}$ ทั้งที่ค่าจริงคือ 5.0×10^{-9} ใช้เอกลักษณ์ $1 - \cos x = 2\sin^2(x/2)$ ซึ่งไม่มีการลบเลย ให้ 5.0000×10^{-9} ที่ $x = 10^{-4}$ โดยมีค่าคลาดเคลื่อนสัมพัทธ์ราว 5×10^{-9}

ทำไมจึงได้ศูนย์ที่ $x = 10^{-4}$ เพราะ $\cos x$ ที่เก็บเป็นความแม่นเดี่ยวจะถูกปัดเป็น 1 เมื่อ $1 - \cos x < 2^{-25}$ ซึ่งเกิดเมื่อ $x < 2^{-12} \approx 2.4 \times 10^{-4}$ ค่าที่ได้จึงเป็นศูนย์

ข้อควรระวัง: การตัดค่าแก้ NaN แต่ไม่แก้มุมเล็ก

บางคนแก้ NaN ด้วยการตัดผลคูณจุดให้อยู่ใน $[-1, 1]$ ก่อนเรียก arccos วิธีนี้ทำให้กล้องไม่หายไปแล้ว แต่ที่ $\theta \approx 10^{-4}$ ยังได้ 0 และที่ $\theta = 10^{-3}$ ยังได้ 1.0358×10^{-3} แทนที่จะเป็น 1.0000×10^{-3} ความผิดนี้มาจากสูตรเอง การตัดค่าไม่ได้แก้ให้

ฉบับร่าง · preprint · 2026-10-23

สรุปใจความ

- การลบเลขสองตัวที่ใกล้กันทำให้ค่าคลาดเคลื่อนสัมพัทธ์โต ตัวขยายคือ $(|x| + |y|)/|x - y|$
- คอนดิชันนัมเบอร์ $\kappa_f(x) = |xf'(x)/f(x)|$ บอกว่าค่าคลาดเคลื่อนสัมพัทธ์ของข้อมูลถูกขยายเป็นกี่เท่าเมื่อผ่าน f
- กล้องที่หายไปเกิดจาก $\arccos$ ที่อ่านมุมจากผลคูณจุดใกล้ 1 ตัวขยายของมันมีขนาดราว 10^{8} เมื่อมุมเป็น 10^{-4}
- สูตร $\operatorname{atan2}(|a \times b|,\, a \cdot b)$ ให้ค่าคลาดเคลื่อนสัมบูรณ์ระดับ u สำหรับทุกมุม เพราะอ่านค่าไซน์โดยตรง
- วิธีแก้คือเขียนสูตรใหม่ให้ไม่หักล้าง เช่นสูตรรากกำลังสอง และ $1 - \cos x = 2\sin^2(x/2)$ ส่วนการตัดค่าแก้ NaN ได้ แต่ไม่แก้มุมเล็ก

เรากลับไปที่คำถามของกล้อง ทำไมสูตรที่เท่ากันบนกระดาษให้คำตอบต่างกันในเครื่อง คำตอบคือการลบแบบหักล้างทำให้ความผิดถูกขยายตามตัวขยาย และการเลือกสูตรที่อ่านข้อมูลตรงกว่าช่วยลดตัวขยายนั้น

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 2.2.1 (⋆) วัดได้ $\tilde{x} = 2.4681$ และ $\tilde{y} = 2.4673$ โดยแต่ละค่าปัดเป็นทศนิยมสี่ตำแหน่ง จงหาช่วงที่ผลต่างจริง $x - y$ อยู่ ค่าคลาดเคลื่อนสัมพัทธ์สูงสุดของผลต่างที่วัดได้ $\tilde{x} - \tilde{y}$ เป็นกี่เปอร์เซ็นต์ และผลต่างนี้รายงานได้กี่หลักนัยสำคัญ

แบบฝึกหัด 2.2.2 (⋆) จงหาค่า $\kappa_f(x) = |xf'(x)/f(x)|$ ของฟังก์ชันต่อไปนี้

1. $f(x) = x^2$ ที่ $x = 3$
2. $f(x) = \sqrt{x}$ ที่ $x = 3$
3. $f(x) = 1/x$ ที่ $x = 3$
4. $f(x) = \arccos x$ ที่ $x = 0.99$

แล้วบอกว่าข้อใดขยายความผิดมากที่สุด และข้อนั้นบอกอะไรเกี่ยวกับกล้องที่หายไปของเรา

แบบฝึกหัด 2.2.3 (⋆⋆) ให้ a และ b เป็นเวกเตอร์หนึ่งหน่วยในสามมิติ ที่ทำมุม θ ต่อกัน

1. แสดงว่า $(a \cdot b)^2 + |a \times b|^2 = 1$ แล้วสรุปว่า $a \cdot b = \cos\theta$ และ $|a \times b| = \sin\theta$ เมื่อ $0 \le \theta \le \pi$
2. สำหรับฟังก์ชัน $\theta(x, y) = \operatorname{atan2}(y, x)$ เมื่อ $x > 0$ จงแสดงว่า $\partial\theta/\partial y = x/(x^2 + y^2)$ และ $\partial\theta/\partial x = -y/(x^2 + y^2)$
3. อธิบายด้วยคำพูดว่าทำไมมุมเล็ก θ ไม่ได้รับผลของการลบแบบเดียวกับ $\arccos$

ตัวช่วยคิด

ข้อ 1 ใช้เอกลักษณ์ของลากรานจ์ $|a|^2|b|^2 = (a \cdot b)^2 + |a \times b|^2$ ข้อ 2 เขียน $\theta = \arctan(y/x)$ แล้วหาอนุพันธ์ ข้อ 3 เทียบ $\sin\theta \approx \theta$ กับ $\cos\theta \approx 1 - \theta^2/2$

ฉบับร่าง · Preprint 2026-10-23

แบบฝึกหัด 2.2.4 (★★) สำหรับสมการ $x^2 - 10^k x + 1 = 0$ เมื่อ $k = 1, \dots, 6$ จงหารากเล็กด้วยสูตร $(-b - \sqrt{b^2 - 4c})/2$ ในความแม่นเดี่ยว แล้วหาค่าคลาดเคลื่อนสัมพัทธ์เทียบกับค่าจริง ตอบว่า k ใดให้ความผิดเกิน 1% จากนั้นอธิบายว่าทำไมบาง k ได้ค่าเป็นศูนย์ และใช้ผลคูณของรากหาอีกสูตรหนึ่ง แล้วตรวจว่าความผิดลดลงเหลือเท่าใด

แบบฝึกหัด 2.2.5 (★★)

1. แสดงว่า $1 - \cos x = 2\sin^2(x/2)$ โดยใช้เอกลักษณ์ $\cos x = 1 - 2\sin^2(x/2)$
2. สำหรับ $x = 2^{-k}$ เมื่อ $k = 8, \dots, 13$ จงทำนายว่าสูตร $1 - \cos x$ ในความแม่นเดี่ยวจะให้ศูนย์เมื่อใด โดยใช้ความจริงว่า $\cos x$ ที่เก็บเป็นความแม่นเดี่ยวจะถูกปัดเป็น 1 เมื่อ $1 - \cos x$ น้อยกว่า 2^{-25} จากนั้นตรวจด้วยโค้ด
3. อธิบายว่า $1 - d$ เมื่อ $d = \cos\theta$ เป็นการลบแบบเดียวกับที่ทำให้กล้องหายไปอย่างไร

แบบฝึกหัด 2.2.6 (★★★) กล้องหมุนเข้าหาเป้าหมายทีละ 10^{-4} เรเดียนต่อเฟรม เริ่มห่างจากเป้าหมาย 0.01 เรเดียน ดังนั้นที่เฟรม $k = 0, 1, \dots, 100$ มุมจริงคือ $\theta_k = 0.01 - k \cdot 10^{-4}$ ใช้ฟังก์ชัน `turned_pair` ในโค้ดเพื่อสร้างคู่เวกเตอร์ความแม่นเดี่ยวของแต่ละเฟรม แล้ว

1. หาเฟรม k แรกที่ $\arccos$ อ่านได้ 0 พอดี และมุมจริงที่เฟรมนั้น
2. ตรวจว่าตั้งแต่เฟรมนั้นจนถึงเฟรม 100 $\arccos$ อ่านได้ 0 ทุกเฟรมหรือไม่
3. ตรวจว่าค่าคลาดเคลื่อนสัมบูรณ์สูงสุดของ $\operatorname{atan2}$ ตลอดทั้งชุด น้อยกว่า $2u$ หรือไม่

ตัวช่วยคิด

ความผิดของ $\arccos$ มีขนาดราว $u/\sin\theta$ ซึ่งเท่ากับมุม θ เองเมื่อ $\theta^2 \approx u$ ลองดูว่าเฟรมไหนใกล้จุดนี้

แบบฝึกหัด 2.2.7 (★★★) ให้ $d = \cos\theta$ และ $\tilde{d} = d + \varepsilon$ เมื่อ $|\varepsilon| \le u$ ซึ่งเป็นความผิดของผลคูณจุดที่เก็บไว้

1. ใช้การกระจายเทย์เลอร์รอบ d แสดงว่า $\arccos\tilde{d} - \theta = -\varepsilon/\sin\theta + O(\varepsilon^2)$
2. สรุปว่าความผิดของ $\arccos$ มีขนาดประมาณ $u/\sin\theta$
3. ที่ $\theta = 0.1$ และ $\theta = 0.01$ จงคำนวณ $|\arccos(\cos\theta + u) - \theta|$ ด้วยเลขความแม่นคู่ แล้วเทียบกับ $u/\sin\theta$ ว่าต่างกันไม่เกิน 1% หรือไม่

ตัวช่วยคิด

อนุพันธ์ของ $\arccos$ คือ $-1/\sqrt{1 - d^2}$ และ $\sqrt{1 - \cos^2\theta} = \sin\theta$ เมื่อ $0 < \theta < \pi$

ฉบับร่าง · Preprint · 2026-10-22

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 2.2.1.

1. ผลต่างที่วัดได้คือ $2.4681 - 2.4673 = 0.0008$
2. แต่ละค่าปัดเป็นทศนิยมสี่ตำแหน่ง จึงคลาดไปไม่เกิน 0.00005 ผลต่างจริงจึงคลาดไปไม่เกิน $0.00005 + 0.00005 = 0.0001$ ดังนั้น $x - y$ อยู่ในช่วง $[0.0007,\ 0.0009]$
3. ค่าคลาดเคลื่อนสัมพัทธ์สูงสุดคือ $0.0001/0.0008 = 0.125$ หรือ 12.5 เปอร์เซ็นต์ ทั้งที่แต่ละค่าที่วัดมีค่าคลาดเคลื่อนสัมพัทธ์เพียงราว 0.002 เปอร์เซ็นต์
4. ผลต่างเป็น 0.0008 ± 0.0001 ซึ่งหลักที่ตำแหน่งทศนิยมที่สี่ยังไม่แน่นอน จึงรายงานได้เพียงหนึ่งหลักนัยสำคัญโดยประมาณ

ตรวจด้วยการคำนวณ: `test_ch02.py`, `test_ex_2_2_1_subtraction`.

เฉลย 2.2.2. ใช้สูตร $\kappa_f(x) = |x f'(x)/f(x)|$

1. $f(x) = x^2$, $f'(x) = 2x$: $\kappa = |x \cdot 2x / x^2| = 2$ ที่ทุก $x \neq 0$
2. $f(x) = \sqrt{x}$, $f'(x) = \dfrac{1}{2\sqrt{x}}$: $\kappa = \left| \dfrac{x/(2\sqrt{x})}{\sqrt{x}} \right| = \dfrac{1}{2}$
3. $f(x) = 1/x$, $f'(x) = -1/x^2$: $\kappa = \left| \dfrac{x \cdot (-1/x^2)}{1/x} \right| = 1$
4. $f(x) = \arccos x$, $f'(x) = -1/\sqrt{1-x^2}$: $\kappa = \dfrac{x}{\sqrt{1-x^2}\,\arccos x}$ ที่ $x = 0.99$ ได้ $\sqrt{1-0.99^2} = \sqrt{0.0199} \approx 0.14107$ และ $\arccos 0.99 \approx 0.14154$ ดังนั้น $\kappa \approx 0.99/(0.14107 \times 0.14154) \approx 49.58$

ข้อ 4 ขยายความผิดมากที่สุด คือประมาณ 50 เท่า ขณะที่ข้อ 1–3 ขยายไม่เกิน 2 เท่า เมื่อ d ใกล้ 1 ค่า arccos เปลี่ยนเร็วมากจนเกือบเป็นแนวตั้ง นี่คือต้นเหตุเดียวกับที่กล้องของเราหายไปเมื่อผลคูณจุดอยู่ใกล้ 1 มาก ตรวจด้วยการคำนวณ: `test_ch02.py`, `test_ex_2_2_2_kappa`.

เฉลย 2.2.3.

1. เอกลักษณ์ของลากรานจ์คือ $|a|^2|b|^2 = (a \cdot b)^2 + |a \times b|^2$ เมื่อ $|a| = |b| = 1$ จะได้ $(a \cdot b)^2 + |a \times b|^2 = 1$ นิยามมุม θ ระหว่างเวกเตอร์ให้ $a \cdot b = \cos\theta$ ดังนั้น $|a \times b|^2 = 1 - \cos^2\theta$ และเพราะ $|a \times b| \geq 0$ กับ $\sin\theta \geq 0$ เมื่อ $0 \leq \theta \leq \pi$ จึงได้ $|a \times b| = \sin\theta$
2. เมื่อ $x > 0$ จะได้ $\theta = \arctan(y/x)$ ดังนั้น
$$\frac{\partial\theta}{\partial y} = \frac{1/x}{1 + y^2/x^2} = \frac{x}{x^2 + y^2}, \qquad \frac{\partial\theta}{\partial x} = \frac{-y/x^2}{1 + y^2/x^2} = \frac{-y}{x^2 + y^2}$$
3. เมื่อ θ เล็ก $\sin\theta \approx \theta$ เปลี่ยนไปตามมุมโดยตรง ส่วน $\cos\theta \approx 1 - \theta^2/2$ เปลี่ยนน้อยมากจนต้องอ่านผลต่างของ 1 กับตัวเลขเล็กมาก สูตร atan2 จึงอ่านมุมจาก $|a \times b| \approx \theta$ ไม่ได้จากการลบ 1 ออกจากโคไซน์

ตรวจด้วยการคำนวณ: `test_ch02.py`, `test_ex_2_2_3_lagrange_identity_and_gradient` (สุ่มทดสอบ 1000 ครั้ง และเทียบอนุพันธ์เชิงตัวเลขที่จุด $(0.6,\ 0.8)$)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 2.2.4. ค่าจริงของรากเล็กคำนวณด้วยเลขทศนิยมหลายหลัก และค่าที่ได้จากสูตรเดิมในความแม่นเดี่ยวคือ

k	สูตรเดิม	ค่าจริง	ค่าคลาดเคลื่อนสัมพัทธ์
1	0.1010203	0.1010205	1.8×10^{-6}
2	0.0100021	0.0100010	1.1×10^{-4}
3	0.0010071	0.0010000	7.1×10^{-3}
4, 5, 6	0	10^{-k} โดยประมาณ	100%

1. ความผิดเกิน 1% เริ่มที่ $k = 4$ เพราะ $k = 3$ ให้ 0.71%
2. สำหรับ $k \geq 4$ ค่า $b^2 - 4c = 10^{2k} - 4$ ถูกปัดเป็น 10^{2k} พอดี เพราะช่องว่างของเลขความแม่นเดี่ยวที่ขนาดนั้นกว้างกว่า 4 ดังนั้น $\sqrt{b^2 - 4c}$ ได้ 10^k เป๊ะ และการลบ $10^k - 10^k$ ให้ศูนย์
3. ผลคูณของรากสองตัวของ $x^2 + bx + c$ เท่ากับ c จึงได้รากเล็กเป็น $c/$(รากใหญ่) โดยรากใหญ่คำนวณจาก $(-b + \sqrt{b^2 - 4c})/2$ ซึ่งเป็นการบวกเลขบวกสองตัว ไม่มีการหักล้าง
4. วิธีใหม่ให้ค่าคลาดเคลื่อนสัมพัทธ์น้อยกว่า 10^{-7} ทุกค่าของ k (ประมาณ 5×10^{-9} ถึง 3.5×10^{-8})

ตรวจด้วยการคำนวณ: `test_ch02.py`, `test_ex_2_2_4_quadratic_by_powers_of_ten`.

เฉลย 2.2.5.

1. จาก $\cos x = 1 - 2\sin^2(x/2)$ ได้ $1 - \cos x = 2\sin^2(x/2)$ ทันที
2. ค่า $\cos x$ ที่เก็บเป็นความแม่นเดี่ยวจะถูกปัดเป็น 1 เมื่อ $1 - \cos x < 2^{-25}$ (ครึ่งหนึ่งของช่องว่างใต้ 1 ซึ่งคือ 2^{-24}) และเพราะ $1 - \cos x \approx x^2/2$ เงื่อนไขนี้เทียบเท่ากับ $x < 2^{-12}$ ดังนั้นสำหรับ $x = 2^{-k}$ ผลจะเป็นศูนย์เมื่อ $k \geq 12$ โค้ดให้ 7.63×10^{-6}, 1.91×10^{-6}, 4.77×10^{-7}, 1.19×10^{-7} สำหรับ $k = 8, \ldots, 11$ และ 0 สำหรับ $k = 12, 13$ ตรงกับการทำนาย
3. ถ้า $d = \cos\theta$ แล้ว $1 - d \approx \theta^2/2$ คือผลต่างของ 1 กับเลขที่ใกล้ 1 มาก เป็นการลบแบบหักล้างตัวเดียวกับที่ทำให้มุมของกล้องเป็นศูนย์ ปัญหาไม่ได้อยู่ที่ตัวฟังก์ชันเอง แต่อยู่ที่การเขียนสูตรให้ลบเลขที่ใกล้กัน

ตรวจด้วยการคำนวณ: `test_ch02.py`, `test_ex_2_2_5_one_minus_cos_threshold`.

เฉลย 2.2.6. ใช้ `turned_pair` กับมุม θ_k แล้วเปรียบเทียบ $\arccos$ กับมุมที่คำนวณจากเวกเตอร์ที่เก็บไว้

1. เฟรม $k = 97$ มุมจริงคือ $\theta_{97} = 3 \times 10^{-4}$ และ $\arccos$ อ่านได้ 0 พอดี ที่เฟรม 96 มุมจริง 4×10^{-4} ยังอ่านได้ 3.45×10^{-4}
2. ใช่ ทั้งเฟรม 97 ถึง 100 อ่านได้ 0 ครบ ดังนั้นกล้องหยุดนิ่งห่างเป้าหมายประมาณ 3×10^{-4} เรเดียน ทั้งที่ยังไม่ถึงเป้า
3. ความผิดสูงสุดของ atan2 คือ 2.6×10^{-8} ซึ่งน้อยกว่า $2u \approx 1.2 \times 10^{-7}$

ตัวช่วยคิดบอกว่าความผิดของ $\arccos$ มีขนาด $u/\sin\theta$ ซึ่งเท่ากับมุมเมื่อ $\theta^2 \approx u$ นั่นคือ $\theta \approx \sqrt{u} = 2^{-12} \approx 2.4 \times 10^{-4}$ ใกล้กับเฟรมที่ $\arccos$ เริ่มติดศูนย์ ตรวจด้วยการคำนวณ: `test_ch02.py`, `test_ex_2_2_6_camera_sweep_sticks_at_three_ten_thousandths`.

สัมมนา. Preprint. 2026-10-23

เฉลย 2.2.7.

1. ให้ $f = \arccos$ จะได้ $f'(d) = -1/\sqrt{1-d^2} = -1/\sin\theta$ เมื่อ $d = \cos\theta$ การกระจายเทย์เลอร์:

$$\arccos(d+\varepsilon) = \arccos d + f'(d)\,\varepsilon + O(\varepsilon^2) = \theta - \frac{\varepsilon}{\sin\theta} + O(\varepsilon^2)$$

   ดังนั้น $\arccos\tilde{d} - \theta = -\varepsilon/\sin\theta + O(\varepsilon^2)$

2. จาก $|\varepsilon| \le u$ จะได้ความผิดประมาณ $|\varepsilon|/\sin\theta \le u/\sin\theta$ บวกเทอมระดับ u^2

3. เลขความแม่นคู่ให้ $|\arccos(\cos 0.1 + u) - 0.1| \approx 5.970 \times 10^{-7}$ เทียบกับ $u/\sin 0.1 \approx 5.970 \times 10^{-7}$ ต่างกันน้อยกว่า 0.01% และที่ $\theta = 0.01$ ได้ 5.962×10^{-6} เทียบกับ 5.961×10^{-6} ต่างกันราว 0.03% ทั้งสองกรณีต่ำกว่า 1%

ตรวจด้วยการคำนวณ: `test_ch02.py`, `test_ex_2_2_7_arccos_error_is_u_over_sin_theta`.

2.3 ปัญหาไม่ดีหรือวิธีไม่ดี

สถานการณ์สมมติ: คลิกที่พื้นไกลแล้วตัวละครไปผิดที่

ลองนึกถึงเกมที่ผู้เล่นคลิกเมาส์บนพื้นเพื่อสั่งให้ตัวละครเดินไปที่นั้น โปรแกรมต้องหาว่าลำแสงที่ออกจากกล้องผ่านตำแหน่งเมาส์ไปชนพื้นที่จุดไหน ลำแสงแบบนี้เรียกว่า **เรย์** (ray) คือเส้นตรงที่เริ่มจากจุดหนึ่งแล้วพุ่งไปทางหนึ่ง
เมื่อเมาส์อยู่ใกล้ขอบฟ้า จุดที่ลำแสงชนพื้นจะอยู่ไกลมาก ในกล้องสมมติของเรา เมาส์อยู่ต่ำกว่าขอบฟ้าหนึ่งพิกเซลทำให้จุดชนพื้นอยู่ไกล 900 เมตร แต่ถ้าเมาส์ลงไปอีกหนึ่งพิกเซล จุดชนพื้นจะกลายเป็น 450 เมตร ตัวละครจึงเดินไปคนละที่กับที่ผู้เล่นคลิก
คำถามคือ ความผิดนี้มาจากสูตรที่เราเขียน หรือมาจากตัวปัญหาเอง ถ้ามาจากสูตร เราเปลี่ยนสูตรได้ ถ้ามาจากตัวปัญหาการเปลี่ยนสูตรช่วยได้ไม่มาก ตัวเลขทั้งหมดในหัวข้อนี้มาจากโค้ดของเรา และเป็นสถานการณ์ที่เราสร้างขึ้นเอง ไม่มีการอ้างถึงเกมใด

ภาพรวมก่อนอ่าน

ในหัวข้อนี้เราจะแยกสองเรื่องให้ออกจากกัน เรื่องแรกคือ ข้อมูลที่เปลี่ยนนิดเดียวทำให้คำตอบเปลี่ยนมากเพียงใด เป็นสมบัติของ**ปัญหา** เรื่องที่สองคือ วิธีคำนวณของเราทำให้ผลผิดไปเท่าใด เป็นสมบัติของ**วิธี** เมื่อแยกสองเรื่องได้แล้ว เราจะเห็นว่าทำไมบางครั้งการเปลี่ยนสูตรช่วยได้ และบางครั้งช่วยไม่ได้เลย

ร่างเรียน · Preprint · 2026-10-23

ตารางที่ 2.3: ระยะที่ลำแสงชนพื้น และความไวต่อการขยับเมาส์หนึ่งพิกเซล (ตัวเลขจากโค้ดของเรา)

แถว y (พิกเซล)	D (เมตร)	ขยับ 1 พิกเซล D เปลี่ยนไป (เมตร)	สัดส่วน $\Delta D/D$	κ
1	900.00	450.00	50.0 %	941.5
2	450.00	150.00	33.3 %	470.2
10	90.00	8.182	9.1 %	93.3
100	9.000	0.08911	0.99 %	8.67
300	3.000	0.00997	0.33 %	2.77

จุดที่ลำแสงชนพื้น

ลองวาดภาพแบบง่าย กล้องสูง $h = 1.5$ เมตร มองขนานกับพื้น จอภาพมีความยาวโฟกัส $f = 600$ พิกเซล และขอบฟ้าอยู่กลางจอ เมาส์อยู่ที่แถว y พิกเซลต่ำกว่าขอบฟ้า

ลำแสงที่ผ่านแถวนี้ทำมุมกับแนวราบเท่ากับ $\arctan(y/f)$ เพื่อให้เขียนสูตรสะดวก เราวัดมุมจากแนวดิ่งแทน และเรียกมุมนี้ว่า **มุมเซนิท** (zenith angle) θ คือมุมระหว่างลำแสงกับแนวดิ่งที่ชี้ลงพื้น ดังนั้น

$$\theta = \frac{\pi}{2} - \arctan\frac{y}{f}.$$

เมื่อ θ เข้าใกล้ $\pi/2$ ลำแสงจะเกือบขนานกับพื้น และจุดที่ชนพื้นจะอยู่ไกลมาก ระยะแนวราบถึงจุดชนพื้นคือ

$$D(\theta) = h\tan\theta.$$

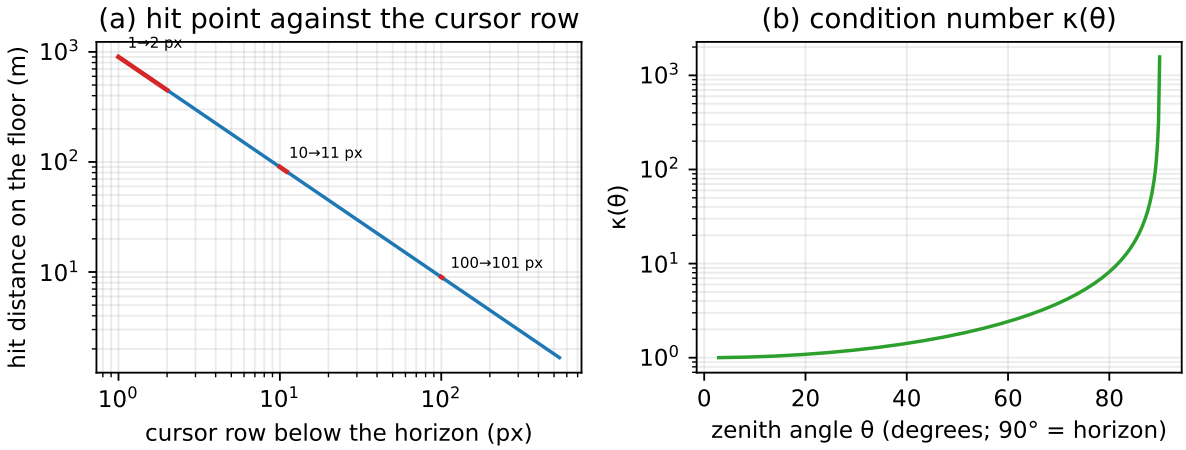
เมื่อแทน $\tan\theta = f/y$ จะได้สูตรที่ใช้ง่ายกว่า $D = hf/y$ ซึ่งบอกว่า เมื่อ y เล็ก D จะใหญ่มาก

ลองเอง: เดาก่อนดูตาราง

ที่แถว $y = 1$ ถ้าเมาส์ลงไปอีกหนึ่งพิกเซล ระยะ D จะเปลี่ยนไปกี่เปอร์เซ็นต์ ลองเดาก่อน แล้วค่อยเทียบกับตารางด้านล่าง ตารางนี้ใช้ $h = 1.5$ เมตร และ $f = 600$ พิกเซล

ตารางนี้บอกเรื่องหนึ่งชัดเจน สัดส่วนการเปลี่ยนของ D เท่ากับ $1/(y+1)$ พอดี เพราะ $D(y)/D(y+1) = (y+1)/y$ ดังนั้นที่แถวใกล้ขอบฟ้า การขยับเมาส์เพียงพิกเซลเดียวเปลี่ยนคำตอบไปครึ่งหนึ่ง ส่วนที่แถวไกลจากขอบฟ้า ผลเปลี่ยนน้อยมาก สดมภ์ κ เป็นตัวเลขที่เราจะนิยามในหัวข้อ ``ปัญหาหรือวิธี'' ข้างล่าง

ฉบับร่าง Preprint, 2026-10-23



รูปที่ 2.4: (ก) ระยะที่ลำแสงชนพื้นเทียบกับแถวของเมาส์ ทั้งสองแกนเป็นลอการิทึม เส้นสีแดงสั้น ๆ คือการขยับหนึ่งพิกเซลที่แถว 1, 10 และ 100 (ข) $\kappa(\theta)$ เทียบกับมุมเซนิทเป็นองศา ค่าโตขึ้นอย่างไม่มีขอบเขตเมื่อเข้าใกล้ 90 องศา (สร้างจาก `numanal/ch02.py`)

ตารางที่ 2.4: ช่วงระยะที่ครอบคลุมด้วยหนึ่งระดับของบัฟเฟอร์ 24 บิต ที่ $n = 0.1$ เมตร และ $f = 1000$ เมตร (ตัวเลขจากโค้ดของเรา)

ระยะ z (เมตร)	ช่วงต่อหนึ่งระดับ Δz
10	0.06 มิลลิเมตร
50	1.5 มิลลิเมตร
100	6 มิลลิเมตร
300	5.4 เซนติเมตร
500	15 เซนติเมตร
1000	60 เซนติเมตร

สถานการณ์สมมติ: อีกด้านหนึ่งของปัญหาเดียวกัน: ผิวสองผิวที่เครื่องแยกไม่ออก

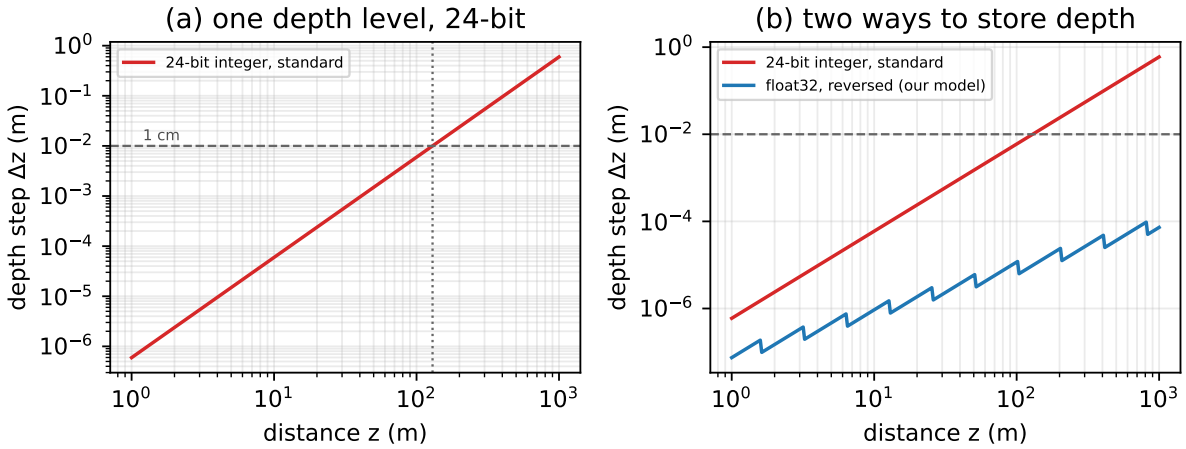
เมื่อกี้ปัญหาอยู่ที่การแปลงมุมเป็นระยะ ทีนี้ลองดูอีกด้านหนึ่ง สมมติว่ามีรอยสีบางบนกำแพง วางห่างจากผิวกำแพงเพียง 1 เซนติเมตร ตอนที่ผู้เล่นยืนใกล้ เราเห็นรอยสีกับกำแพงแยกกันชัดเจน แต่เมื่อผู้เล่นถอยออกไปไกล รอยสีกับกำแพงจะสลับกันกะพริบ ภาพแบบนี้เรียกว่า **ซี-ไฟต์ติง** (z-fighting) คือเครื่องไม่สามารถบอกได้ว่าผิวไหนอยู่หน้า เพราะทั้งสองผิวถูกเก็บเป็นค่าความลึกที่เหมือนกัน (สถานการณ์นี้เป็นสมมติ ตัวเลขมาจากโค้ดของเรา)

ระหว่างการวาดภาพ เครื่องเก็บค่าความลึกของแต่ละจุดในภาพไว้ใน **บัฟเฟอร์ความลึก** (depth buffer) ค่าที่เก็บไม่ใช่ระยะ z ตรง ๆ แต่เป็นค่า d ที่แปรผกผันกับ z บทความของ Reed อธิบายว่าการเก็บความลึกแบบนี้มีลักษณะเป็นส่วนกลับของระยะ Reed 2021 ในที่นี้เราใช้ระยะใกล้สุด $n = 0.1$ เมตร และระยะไกลสุด $f = 1000$ เมตร ค่า d ถูกเก็บเป็นจำนวนเต็ม 24 บิต ซึ่งมี 2^{24} ระดับ คือตั้งแต่ 0 ถึง $2^{24} - 1$ ระดับ และช่วงระหว่างระดับที่ปลายทั้งสองมี $2^{24} - 1$ ช่วง เราสมมติการเก็บแบบนี้เพื่อให้ตัวเลขชัดเจน ไม่ได้อ้างว่าการ์ดแสดงผลทุกตัวทำเช่นนี้

ปัญหาคือระดับที่อยู่ติดกันห่างกันเท่ากันในค่า d คือ $1/(2^{24} - 1)$ แต่ในระยะ z ความห่างนั้นไม่เท่ากัน ตารางต่อไปนี้แสดงว่า ระดับหนึ่งระดับครอบคลุมระยะ z เท่าไร

ช่วงต่อหนึ่งระดับโตตามระยะกำลังสอง ที่ระยะ 129.5 เมตร หนึ่งระดับมีขนาด 1 เซนติเมตรพอดี ดังนั้นรอยสีที่วางห่าง

ร่างเนื้อหา · Preprint · 2026-10-23



รูปที่ 2.5: (ก) ช่วงต่อหนึ่งระดับของบัฟเฟอร์ 24 บิต เทียบกับระยะ เส้นประคือ 1 เซนติเมตร (ข) เทียบกับ float32 แบบกลับด้าน (ของเรา) ซึ่งช่วงต่อหนึ่งระดับโตตามระยะ z แบบเชิงเส้นและเล็กกว่ามากที่ระยะไกล (สร้างจาก `numanal/ch02.py`)

กำแพง 1 เซนติเมตรเริ่มแยกไม่ออกจากกำแพงราว ๆ ระยะนี้ ถ้าเปลี่ยนระยะใกล้สุดเป็น 0.01 เมตร ระยะนั้นจะเป็น 41.0 เมตร และถ้าเป็น 1 เมตร จะเป็น 409.8 เมตร

ดูที่ผิวสองผิวที่ห่างกัน 1 เซนติเมตรโดยตรง ที่ $z = 500$ เมตร ทั้งสองผิวได้ระดับเดียวกันคือ $16{,}775{,}537$ ส่วนที่ $z = 100$ เมตร ผิวสองผิวได้ระดับต่างกันสองระดับ คือ $16{,}762{,}114$ กับ $16{,}762{,}116$ ที่ระยะ 100 เมตร ช่วงต่อหนึ่งระดับ 6 มิลลิเมตร ทำให้ 1 เซนติเมตรครอบคลุมเกือบสองระดับพอดี

บทความของ Reed รายงานการทดสอบการปัดเศษของเขาเอง โดยใช้ระยะใกล้สุด 0.1 ระยะไกลสุด $10{,}000$ และความลึก $10{,}000$ ค่าที่เว้นระยะเท่ากัน ถ้าใช้เมทริกซ์มุมมองรวมกับการฉายเป็นเมทริกซ์เดียว (precomposed view-projection matrix) และเก็บเป็น float32 แบบมาตรฐาน ราว 45 % ของความลึกที่ทดสอบแยกไม่ได้ และ 18 % สลับลำดับ ถ้าแยกเมทริกซ์มุมมองกับการฉายออกจากกัน ตัวเลขเป็น 77 % และ 0 % ส่วนแบบกลับด้านเป็น float32 ไม่มีจุดใดแยกไม่ได้ในทั้งสองกรณี Reed 2021 ตัวเลขเหล่านี้เป็นของบทความและใช้ได้กับการตั้งค่านั้นเท่านั้น เราไม่ได้ทดสอบซ้ำ และตัวเลขนี้ไม่ได้ถูกคำนวณในโค้ดของเรา

ฉบับร่าง · preprint · 2026-10-23

ปัญหาหรือวิธี

ในหัวข้อ 2.2 เราวัด κ ของสูตร $\arccos$ ตอนนี้เราวัดของ**ปัญหา**เองได้ โดยปัญหาคือการแปลงข้อมูลไปเป็นคำตอบ

นิยาม 2.3.1: คอนดิชันนัมเบอร์ของปัญหา

ให้ $z = P(x)$ เป็นปัญหาที่แปลงข้อมูล x เป็นคำตอบ z โดย P หาอนุพันธ์ได้ที่ x และ $x \neq 0$, $P(x) \neq 0$ นิยาม

$$\kappa_P(x) = \left| \frac{x\,P'(x)}{P(x)} \right|$$

เรียก $\kappa_P(x)$ ว่า **คอนดิชันนัมเบอร์** (condition number) ของปัญหาที่จุด x ถ้า $\kappa_P(x)$ ใหญ่ เรียกปัญหานั้นว่า **ปัญหาสภาวะเลว** (ill-conditioned problem) ที่จุดนั้น

อ่านแบบเข้าใจง่าย

- พูดง่าย ๆ คือ ปัญหาสภาวะเลวคือปัญหาที่ข้อมูลเปลี่ยนนิดเดียวแล้วคำตอบเปลี่ยนมาก
- $\kappa_P(x)$ บอกจำนวนเท่าที่ค่าคลาดเคลื่อนสัมพัทธ์ของข้อมูลถูกขยายเป็นค่าคลาดเคลื่อนสัมพัทธ์ของคำตอบ ถ้า $\kappa = 2$ และข้อมูลผิดหนึ่งเปอร์เซ็นต์ คำตอบจะผิดราวสองเปอร์เซ็นต์
- รูปนี้เป็นรูปเฉพาะจุดของนิยามที่ Atkinson ใช้ซึ่งดูการเปลี่ยนแปลงเล็ก ๆ ทั้งหมดของข้อมูล Atkinson 1989, §1.6, pp. 35--36 เราใช้ค่าอนุพันธ์แทนการหาค่าสูงสุดของการเปลี่ยนแปลงทั้งหมด
- ค่า κ ขึ้นกับปัญหาและจุดที่เราดู ไม่ขึ้นกับสูตรที่เราเลือกมาคำนวณ

ทฤษฎีบท 2.3.1: คอนดิชันนัมเบอร์ของลำแสง

สำหรับ $D(\theta) = h\tan\theta$ เมื่อ $0 < \theta < \pi/2$ จะได้

$$\kappa(\theta) = \frac{\theta}{\sin\theta\cos\theta} = \frac{2\theta}{\sin 2\theta} \;\geq\; 1, \qquad \lim_{\theta\to\pi/2} \kappa(\theta)\cos\theta = \frac{\pi}{2}.$$

พิสูจน์

แนวคิด. หาอนุพันธ์ของ D แล้วแทนลงในนิยาม จากนั้นใช้ขอบเขตของ $\sin$

1. **อนุพันธ์.** $D'(\theta) = h\sec^2\theta$
2. **แทนลงนิยาม.** ได้ $\kappa(\theta) = \left|\dfrac{\theta\, h\sec^2\theta}{h\tan\theta}\right| = \dfrac{\theta}{\sin\theta\cos\theta}$ เพราะ $\sec^2\theta/\tan\theta = 1/(\sin\theta\cos\theta)$ และทุกค่าเป็นบวกเมื่อ $0 < \theta < \pi/2$
3. **ขอบเขตล่าง.** เพราะ $\sin x \leq x$ เมื่อ $x \geq 0$ จึงได้ $\sin 2\theta \leq 2\theta$ ดังนั้น $2\theta/\sin 2\theta \geq 1$
4. **ลิมิต.** $\kappa(\theta)\cos\theta = \theta/\sin\theta$ และ $\sin\theta \to 1$ เมื่อ $\theta \to \pi/2$ จึงได้ $\pi/2$

สรุป: ตัวขยายของลำแสงมีค่าอย่างน้อยหนึ่ง และโตแบบ $(\pi/2)/\cos\theta$ เมื่อเข้าใกล้ขอบฟ้า

อ่านให้เห็นตัวเลข

ที่แถว $y = 1$ ได้ $\kappa = 941.5$ แปลว่าถ้ามุมผิดสัมพัทธ์ 0.1 เปอร์เซ็นต์ ระยะจะผิดราว 94 เปอร์เซ็นต์ ถ้าเทียบกับแถว $y = 300$ ที่ $\kappa = 2.77$ มุมผิดเท่าเดิมจะทำให้ระยะผิดเพียงราว 0.3 เปอร์เซ็นต์

นิยาม 2.3.2: ค่าคลาดเคลื่อนไปข้างหน้าและย้อนกลับ

ให้ $z = P(x)$ และให้วิธีคำนวณได้คำตอบ $\tilde{z}$ นิยาม

- **ค่าคลาดเคลื่อนไปข้างหน้า** (forward error) คือ $|\tilde{z} - z|$ ซึ่งบอกว่าคำตอบผิดไปเท่าไร
- **ค่าคลาดเคลื่อนย้อนกลับ** (backward error) คือ $|\tilde{x} - x|$ โดย $\tilde{x}$ เป็นข้อมูลที่ทำให้ $P(\tilde{x}) = \tilde{z}$ พอดี และเลือกค่าที่เล็กที่สุด ซึ่งบอกว่าคำตอบที่ได้เป็นคำตอบที่แม่นของข้อมูลที่เปลี่ยนไปเท่าไร

ฉบับร่าง Preprint 2026-10-23

ทฤษฎีบท 2.3.2: ความผิดไปข้างหน้าเท่ากับ κ คูณความผิดย้อนกลับ

ให้ P หาอนุพันธ์อันดับสองได้ใกล้ x โดย $x \neq 0$ และ $P(x) \neq 0$ ถ้า $\tilde{x} = x(1+\delta)$ คือข้อมูลที่ทำให้ $P(\tilde{x}) = \tilde{z}$ แล้ว

$$\frac{|\tilde{z} - z|}{|z|} = \kappa_P(x)\,|\delta| + O(\delta^2).$$

พิสูจน์

แนวคิด. เขียนความผิดไปข้างหน้าเป็นความผิดของ P ที่จุด $\tilde{x}$ แล้วใช้ทฤษฎีบทการส่งความผิดผ่านอนุพันธ์ของหัวข้อ 2.2

1. **เขียนใหม่.** เพราะ $\tilde{z} = P(\tilde{x})$ จึงได้ $|\tilde{z} - z|/|z| = |P(\tilde{x}) - P(x)|/|P(x)|$
2. **ใช้ทฤษฎีบทในหัวข้อ 2.2.** ทฤษฎีบท 2.2.2 กับ $\tilde{x} = x(1+\delta)$ ให้ค่า $\kappa_P(x)|\delta| + O(\delta^2)$ ทางขวา
3. **รวม.** สองขั้นนี้รวมกันได้สมการที่ต้องการ

สรุป: ความผิดไปข้างหน้าเท่ากับความผิดย้อนกลับคูณด้วย κ ถ้า κ ใหญ่ วิธีที่ย้อนกลับได้แม่นก็ยังให้คำตอบผิดไปไกลได้

ลองดูตัวอย่างจริงในความแม่นเดี่ยว ที่แถว $y = 1$ โค้ดของเราคำนวณ D ได้ 900.0236 เมตร ความผิดไปข้างหน้าจึงเป็น 0.0236 เมตร หรือค่าคลาดเคลื่อนสัมพัทธ์ 2.6×10^{-5} ถ้าเราหามุมย้อนกลับจาก D ที่คำนวณได้ จะได้มุมที่ต่างจากมุมจริงเพียง 4.4×10^{-8} เรเดียน ซึ่งมีขนาดราวหน่วยปัด และ $\kappa \times 4.4 \times 10^{-8}/\theta \approx 2.6 \times 10^{-5}$ ตรงกับความผิดไปข้างหน้าที่วัดได้ วิธีนี้แม่นในความหมายย้อนกลับ ความผิดไปข้างหน้าที่เหลือเป็นของปัญหาเอง

เกร็ดเพิ่มเติม

Atkinson เรียกวิธีคำนวณว่าเสถียร เมื่อความไวของคำตอบต่อข้อมูลมีขนาดราวเดียวกับความไวของปัญหาเดิม Atkinson 1989, §1.6, p. 37 ตัวอย่างข้างบนเข้าข่ายนี้ เพราะความผิดย้อนกลับมีขนาดเท่าหน่วยปัด ความไวที่เหลือจึงเป็นของปัญหาเอง

การรวมก็แม่นแบบย้อนกลับ

ตอนรวมนาฬิกาทีละเฟรม เราได้ความผิดแบบไปข้างหน้า คือ E_n ในหัวข้อ 2.1 ตอนนี้ลองดูมุมย้อนกลับของการรวม ซึ่งให้ผลที่เข้าใจง่ายกว่า

ทฤษฎีบท 2.3.3: การรวมแบบย้อนกลับ

ให้ $x_0, \dots, x_{n-1}$ และให้ $\hat{S}$ เป็นผลบวกที่คำนวณทีละตัวจากซ้ายไปขวาในความแม่นเดี่ยว ถ้าการบวกทุกครั้งอยู่ในช่วงเลขปกติ (ไม่เกิดโอเวอร์โฟลว์ และไม่เข้าสู่เลขซับนอร์มอล) จะมี η_k ที่ทำให้

$$\hat{S} = \sum_{k=0}^{n-1} x_k(1+\eta_k), \qquad |\eta_k| \le (1+u)^{n-1} - 1,$$

เมื่อ $u = 2^{-24}$ คือหน่วยปัด

ฉบับร่าง · Preprint 2026-10-23

พิสูจน์

แนวคิด. ดูว่าการปัดแต่ละครั้งคูณค่าเข้ากับข้อมูลที่อยู่ข้างหน้าเท่าไร แล้วนับจำนวนครั้งที่ข้อมูลแต่ละตัวถูกคูณ

1. **ตั้งค่า.** ให้ $S_0 = x_0$ และ $S_j = \mathrm{fl}(S_{j-1} + x_j) = (S_{j-1} + x_j)(1 + \varepsilon_j)$ เมื่อ $|\varepsilon_j| \le u$ เพราะการปัดเข้าค่าใกล้ที่สุด
2. **คลี่ความสัมพันธ์.** ทำซ้ำการแทนค่า จะได้ $S_{n-1} = \sum_{k} x_k \prod_{j=\max(k,1)}^{n-1} (1 + \varepsilon_j)$ เพราะ x_k ถูกคูณด้วยตัวปัดทุกตัวตั้งแต่ขั้นที่รวมเข้า x_k ไปจนถึงขั้นสุดท้าย
3. **นิยาม η_k.** ตั้ง $1 + \eta_k = \prod_{j=\max(k,1)}^{n-1}(1 + \varepsilon_j)$ ตัวคูณนี้มีอยู่ไม่เกิน $n - 1$ ตัว แต่ละตัวอยู่ระหว่าง $1 - u$ และ $1 + u$ จึงอยู่ในช่วง $[(1-u)^{n-1}, (1+u)^{n-1}]$ เมื่อดูผลต่างจาก 1 จะได้ $|\eta_k| \le (1+u)^{n-1} - 1$ เพราะ $(1+u)^m - 1 \ge 1 - (1-u)^m$

สรุป: การรวมแบบเลือกลำดับซ้ายไปขวาให้ผลเท่ากับผลบวกจริงของข้อมูลที่คูณด้วยตัวปรับเล็ก ๆ ต่อตัว และตัวปรับนั้นโตช้าตามจำนวนตัว

อ่านแบบเข้าใจง่าย

ความหมายคือ ผลบวกที่ได้ตอบคำถามของข้อมูลที่เปลี่ยนไปเล็กน้อยต่อแต่ละตัว ไม่ได้ตอบคำถามเดิมผิด สำหรับนาฬิกาหนึ่งชั่วโมงที่มี $216{,}000$ เฟรม ขอบเขตนี้มีค่าราว 1.3 เปอร์เซ็นต์ ซึ่งเป็นขอบเขตที่มากเกินจริงมาก แต่ก็ยังบอกได้ว่าความผิดย้อนกลับเล็กมาก

ปัญหาไม่ดีหรือวิธีไม่ดี: คำตอบ

ตอนนี้เรามีเครื่องมือแยกสองเรื่องออกจากกันแล้ว วิธีที่แม่นในความหมายย้อนกลับ เช่นการรวมทีละตัวหรือ $\mathrm{atan2}$ ให้ความผิดไปข้างหน้าเท่ากับ κ คูณหน่วยปัด ดังนั้นถ้าความผิดไปข้างหน้ายังใหญ่ สาเหตุคือปัญหาไวเอง ไม่ใช่วิธี

ยิ่งไปกว่านั้น ถ้าเซนเซอร์บอกแถวเมาส์ได้เพียงว่าอยู่ในช่วง $\pm\frac{1}{2}$ พิกเซล ระยะจริงจะอยู่ในช่วง $[600,\ 1800]$ เมตรเมื่อแถวอยู่ที่ 1 พิกเซล ไม่ว่าวิธีไหนจะตอบค่า c ออกมา เพราะ $|c - 600| + |c - 1800| \ge 1200$ ความผิดของคำตอบอย่างน้อยหนึ่งในสองปลายจึงต้องเป็น 600 เมตรหรือมากกว่า นี่คือขีดจำกัดของปัญหา ไม่ใช่ของวิธีที่เลือก

ข้อควรระวัง: อย่าสรุปว่าเปลี่ยนสูตรแก้ได้เสมอ

เปลี่ยนสูตรช่วยได้เมื่อวิธีไม่เสถียร เช่นรากเล็กของสมการกำลังสองหรือ $\arccos$ ในหัวข้อ 2.2 แต่เปลี่ยนสูตรไม่ช่วยเมื่อข้อมูลหยาบเกินไป ที่แถว 1 เมาส์ไม่มีข้อมูลละเอียดพอจะบอกว่าลำแสงอยู่ตรงไหน ทางแก้คือเพิ่มความละเอียดของข้อมูล ไม่ใช่เขียนสูตรใหม่

เกร็ดเพิ่มเติม

ปัญหาไม่ดีกับวิธีไม่ดีเป็นสองเรื่องที่ต่างกัน บทที่ 7 จะพบระบบเชิงเส้นที่ไวต่อข้อมูลในลักษณะคล้ายกันเมื่อเราพยายามทำภาพเบลอให้คมขึ้น บทที่ 12 จะพบอีกด้านหนึ่ง คือวิธีที่เลือกขั้นเวลาใหญ่เกินไปจนคำตอบระเบิด ซึ่งเป็นความผิดของวิธีที่เลือก ไม่ใช่ของปัญหา

ฉบับร่าง Preprint · 2026-10-23

สรุปใจความ

- ลำแสงที่ชนพื้นใกล้ขอบฟ้ามีระยะไวมาก หนึ่งพิกเซลที่แถว 1 เปลี่ยนระยะไปครึ่งหนึ่ง ตัวขยายคือ $\kappa(\theta) = \theta/(\sin\theta\cos\theta)$ ซึ่งโตแบบ $1/\cos\theta$
- **คอนดิชันนัมเบอร์ของปัญหา** บอกว่าข้อมูลที่ผิดเล็กน้อยถูกขยายเท่าไร และไม่ขึ้นกับสูตรที่เราเลือก
- วิธีที่แม่นในความหมายย้อนกลับให้ความผิดไปข้างหน้าเท่ากับ κ คูณความผิดย้อนกลับ ถ้าความผิดยังใหญ่ สาเหตุคือปัญหาไวเอง
- บัฟเฟอร์ความลึก 24 บิต ทำให้ผิวสองผิวที่ห่างกัน 1 เซนติเมตรแยกไม่ได้ที่ระยะ 500 เมตร ช่วงต่อหนึ่งระดับคือ $(f-n)z^2/(fn(2^{24}-1))$ ซึ่งถึง 1 เซนติเมตรที่ 129.5 เมตร

เรากลับไปที่คำถามของเมาส์ ความผิดนั้นไม่ได้มาจากสูตรที่แย่ แต่มาจากการที่ข้อมูลหนึ่งพิกเซลถูกขยายโดยความไวของปัญหาที่ใกล้ขอบฟ้า ผิวสองผิวบนกำแพงก็มาจากหลักเดียวกัน คือหน่วยการเก็บหนึ่งหน่วยคูณด้วยความไวที่โตตามระยะ

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 2.3.1 (⋆) ใช้ $h = 1.5$ เมตร และ $f = 600$ พิกเซล ตามกล้องในหัวข้อนี้

1. จงหาระยะ D ที่ลำแสงชนพื้นเมื่อเมาส์อยู่ที่แถว $y = 1,\ 2,\ 10,\ 100$
2. จงหาการเปลี่ยนแปลงของ D เมื่อเมาส์ขยับจากแถว 10 ไปแถว 11 และจากแถว 100 ไปแถว 101 แล้วเทียบสัดส่วน $\Delta D/D$ กับ $1/(y+1)$
3. จงหาแถวแรกที่การขยับหนึ่งพิกเซลทำให้ D เปลี่ยนน้อยกว่า 1 เมตร

แบบฝึกหัด 2.3.2 (⋆) ให้ $\kappa(\theta) = \theta/(\sin\theta\cos\theta)$ คือคอนดิชันนัมเบอร์ของลำแสง เมื่อ $0 < \theta < \pi/2$

1. จงหา $\kappa(\theta)$ ที่ $\theta = 60^\circ,\ 80^\circ$ และ 89°
2. แสดงว่า $\kappa(\theta) \geq 1$ โดยใช้ความจริงว่า $\sin x \leq x$ เมื่อ $x \geq 0$
3. แสดงว่า $\kappa(\theta)\cos\theta \to \pi/2$ เมื่อ $\theta \to \pi/2$ แล้วอธิบายด้วยคำพูดว่าทำไมจึงบอกว่า κ โตแบบ $1/\cos\theta$

แบบฝึกหัด 2.3.3 (⋆⋆) ให้ $D(\theta) = h\tan\theta$ และสมมติว่าวิธีคำนวณให้ค่า $\tilde{D} = D(\theta+\delta)$ เมื่อ δ เล็กมาก

1. แสดงว่า $|\tilde{D} - D|/|D| \approx \kappa(\theta)\,|\delta|/|\theta|$ โดยใช้ทฤษฎีบทในหัวข้อนี้
2. ที่ $\theta = 1.0$ และ $\delta = 10^{-6}$ จงตรวจด้วยโค้ดว่าอัตราส่วน $\dfrac{|\tilde{D}-D|/|D|}{|\delta|/|\theta|}$ ใกล้ $\kappa(1.0)$ ภายในหนึ่งส่วนหมื่นหรือไม่

ตัวช่วยคิด

ข้อ 1 ให้ $\tilde{x} = \theta(1+\delta')$ โดย $\delta' = \delta/\theta$ แล้วใช้ทฤษฎีบทความผิดไปข้างหน้าเท่ากับ κ คูณความผิดย้อนกลับ ข้อ 2 ใช้ค่า $\kappa(1.0) = 1/(\sin 1\cos 1)$

แบบฝึกหัด 2.3.4 (⋆⋆) ให้ $S_0 = x_0$ และ $S_j = \mathrm{fl}(S_{j-1} + x_j) = (S_{j-1}+x_j)(1+\varepsilon_j)$ เมื่อ $|\varepsilon_j| \leq u$ สำหรับ $j = 1, \ldots, n-1$

ฉบับร่าง . Preprint . 2026-10-23

1. แสดงโดยการเหนี่ยวนำว่า $S_{n-1} = \sum_{k} x_k \prod_{j=\max(k,1)}^{n-1}(1+\varepsilon_j)$
2. ใช้ฟังก์ชัน `summation_backward_weights` ในโค้ดกับข้อมูลสุ่ม 200 ตัวที่ได้จาก $\text{seed} = 1, 2, 3$ แล้วตรวจว่า $\sum_k x_k(1+\eta_k)$ เท่ากับผลรวมที่คำนวณได้ภายในความสัมพัทธ์ 10^{-12} และ $\max_k |\eta_k| \le (1+u)^{199} - 1$
3. ค่า $\max_k |\eta_k|$ ที่ได้เล็กกว่าขอบเขตมากหรือไม่ และทำไม

แบบฝึกหัด 2.3.5 (★★) ตั้ง $n = 0.1$ เมตร และ $f = 1000$ เมตร บัฟเฟอร์ความลึกเก็บเป็นจำนวนเต็ม 24 บิต

1. จงหาระยะ z ที่ช่วงต่อหนึ่งระดับเท่ากับ 1 เซนติเมตร โดยใช้สูตร $\Delta z = (f-n)z^2/(fn(2^{24}-1))$
2. ทำซ้ำข้อ 1 โดยเปลี่ยน f เป็น 10^4 เมตร ระยะเปลี่ยนไปกี่เปอร์เซ็นต์ และทำไมจึงเปลี่ยนน้อย
3. ที่ $f = 1000$ เมตร จงหาระยะเมื่อ $n = 0.01$ และเมื่อ $n = 1$ แล้วตรวจว่าระยะนั้นเป็นสัดส่วนกับ $\sqrt{n}$ หรือไม่

แบบฝึกหัด 2.3.6 (★★★) กลับไปที่สถานการณ์ผิวสองผิว ให้เก็บค่าความลึกแบบกลับด้าน $d = n/z$ เป็นจำนวนทศนิยมลอยตัว 32 บิต (แบบจำลองของเรา) โดยยังใช้ $n = 0.1$ เมตร

1. ด้วยโค้ดให้หาระยะ z โดยประมาณที่ช่วงต่อหนึ่งระดับเท่ากับ 1 เซนติเมตร แล้วเทียบกับ 129.5 เมตรของการเก็บแบบจำนวนเต็ม
2. อธิบายว่าทำไมช่วงต่อหนึ่งระดับของแบบทศนิยมลอยตัวโตตาม z เชิงเส้น (แนะนำ: ช่องว่างของทศนิยมลอยตัวเป็นกำลังของสองคูณตัวเลข)
3. ที่ $z = 500$ เมตร เลื่อนผิวที่สองออกจากผิวแรกทีละ 0.01 เมตร แล้วหาระยะเลื่อนแรกที่ผิวทั้งสองได้ระดับต่างกันในบัฟเฟอร์จำนวนเต็ม 24 บิต

แบบฝึกหัด 2.3.7 (★ ★ ★, โครงงานท้ายบท) ทำรายงานหนึ่งหน้าที่เชื่อมสามเรื่องของบทนี้เข้าด้วยกัน โดยใช้โค้ดของเราเป็นเครื่องมือ

1. สร้างตารางระยะ z ที่หนึ่งระดับของบัฟเฟอร์ 24 บิตครอบคลุม 1 เซนติเมตร สำหรับ $n \in \{0.01, 0.1, 1\}$ และ $f \in \{100, 1000, 10^4\}$ รวม 9 ค่า
2. เลือกหนึ่งกรณีในตารางแล้วตรวจว่า ผิวสองผิวที่ห่างกัน 1 เซนติเมตรที่ระยะ 100 เมตร ได้ระดับเดียวกันหรือไม่
3. สำหรับแต่ละเรื่องของบท คือ นาฬิกา (หัวข้อ 2.1) กล้องที่หันไปหาเป้าหมาย (หัวข้อ 2.2) และเมาส์กับผิวสองผิว (หัวข้อ 2.3) ให้เขียนว่าความผิดหลักมาจาก **สูตร** **ข้อมูลที่หยาบ** หรือ **ตัวปัญหาที่ไว** พร้อมตัวเลขหนึ่งค่าจากตารางหรือโค้ดที่สนับสนุนคำตอบ

ตัวช่วยคิด

ใช้ฟังก์ชัน `distance_for_step` และ `same_stored_level` ในโค้ด ข้อ 3 ให้ดูว่าความผิดนั้นเปลี่ยนไปเมื่อเปลี่ยนสูตรหรือไม่ ถ้าไม่เปลี่ยนแปลงแสดงว่าเป็นของข้อมูลหรือของปัญหา

ฉบับร่าง · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 2.3.1.

1. จาก $D = hf/y$ โดยมี $hf = 1.5 \times 600 = 900$ จึงได้ $D(1) = 900$, $D(2) = 450$, $D(10) = 90$, $D(100) = 9$ เมตร

2. ใช้ผลต่าง $D(y) - D(y+1) = 900/(y(y+1))$ จึงได้ ที่แถว $10 \to 11$ $900/(10 \cdot 11) = 8.182$ เมตร และสัดส่วน $8.182/90 = 1/11 \approx 9.1\%$ ที่แถว $100 \to 101$ ได้ $900/(100 \cdot 101) = 0.08911$ เมตร และสัดส่วน $0.08911/9 = 1/101 \approx 0.99\%$ ทั้งสองสอดคล้องกับ $1/(y+1)$

3. ต้องหาแถว y ที่ $900/(y(y+1)) < 1$ นั่นคือ $y(y+1) > 900$ ที่ $y = 29$ ได้ $29 \cdot 30 = 870 < 900$ จึงได้ $D(29) - D(30) = 900/870 \approx 1.034$ เมตร ที่ $y = 30$ ได้ $30 \cdot 31 = 930 > 900$ จึงได้ $D(30) - D(31) = 900/930 \approx 0.968$ เมตร ดังนั้นคำตอบคือ $y = 30$

เฉลย 2.3.2.

1. ที่ 60°: $\theta = \pi/3 \approx 1.04720$, $\sin\theta\cos\theta = 0.86603 \times 0.5 = 0.43301$ จึงได้ $\kappa \approx 2.418$ ที่ 80°: $\theta \approx 1.39626$, $\sin\theta\cos\theta \approx 0.98481 \times 0.17365 = 0.17101$ จึงได้ $\kappa \approx 8.165$ ที่ 89°: $\theta \approx 1.55334$, $\sin\theta\cos\theta \approx 0.017449$ จึงได้ $\kappa \approx 89.02$ โค้ดให้ 2.4184, 8.1648 และ 89.018 ตรงกัน

2. เขียน $\kappa = \theta/(\sin\theta\cos\theta) = 2\theta/\sin 2\theta$ เมื่อ $0 < \theta < \pi/2$ จะได้ $0 < 2\theta < \pi$ และ $\sin 2\theta > 0$ เพราะ $\sin x \le x$ สำหรับ $x \ge 0$ จึงได้ $\sin 2\theta \le 2\theta$ ดังนั้น $2\theta/\sin 2\theta \ge 1$

3. $\kappa(\theta)\cos\theta = \theta/\sin\theta$ และ $\sin\theta \to 1$ เมื่อ $\theta \to \pi/2$ จึงได้ลิมิต $\pi/2$ นั่นคือ $\kappa(\theta) \approx (\pi/2)/\cos\theta$ เมื่ออยู่ใกล้ขอบฟ้า ตัวหาร $\cos\theta$ เข้าใกล้ศูนย์ จึง κ โตแบบ $1/\cos\theta$

เฉลย 2.3.3.

1. ให้ $\tilde{x} = \theta(1+\delta')$ โดย $\delta' = \delta/\theta$ ตามนิยามของค่าคลาดเคลื่อนย้อนกลับ $\tilde{D} = D(\tilde{x})$ ทฤษฎีบท 2.3.2 ให้ $|\tilde{D} - D|/|D| = \kappa(\theta)|\delta'| + O(\delta'^2)$ ดังนั้น $|\tilde{D} - D|/|D| \approx \kappa(\theta)|\delta|/|\theta|$

2. ที่ $\theta = 1.0$ ได้ $\kappa(1.0) = 1/(\sin 1 \cos 1) = 1/(0.84147 \times 0.54030) \approx 2.1995$ โค้ดคำนวณ $\tilde{D} = h\tan(1 + 10^{-6})$ และ $D = h\tan 1$ แล้วได้อัตราส่วนใกล้ 2.1995 ภายในหนึ่งส่วนหมื่น ซึ่งยืนยันข้อ 1 (ตรวจด้วยโค้ดในชุดทดสอบ `test_ch02.py`)

เฉลย 2.3.4.

1. ฐาน $n = 1$: $S_0 = x_0$ ตรงตามสูตร ขั้นเหนี่ยวนำ สมมติว่าสูตรจริงสำหรับ S_{m-1} แล้วเขียน $S_m = (S_{m-1} + x_m)(1+\varepsilon_m)$ แทน S_{m-1} ที่เป็นผลรวมของ x_k คูณตัวปัดแล้ว จะได้ว่า x_k ทุกตัวที่ $k < m$ ถูกคูณด้วยตัวปัด ε_m เพิ่มเข้าไป และ x_m ถูกคูณด้วย $(1+\varepsilon_m)$ พอดี นั่นคือสูตรของข้อ 1 เมื่อ $n = m+1$

2. โค้ดให้ $\max_k |\eta_k| \approx 4.3 \times 10^{-7}$ สำหรับ seed $= 1$ ขณะที่ขอบเขต $(1+u)^{199} - 1 \approx 1.2 \times 10^{-5}$ ตรวจ seed อีกสองค่าให้ผลลัพธ์แบบเดียวกัน และผลรวม $\sum_k x_k(1+\eta_k)$ เท่ากับผลรวมที่คำนวณได้ภายในความสัมพัทธ์ 10^{-12}

3. ค่าจริงเล็กกว่าขอบเขตมาก เพราะความผิดของการปัดแต่ละครั้งมีเครื่องหมายสลับไปมา ทำให้ผลคูณของตัวปัดเกือบหักล้างกัน ขอบเขตจึงเป็นการคิดแบบเลวร้ายที่สุด

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 2.3.5.

1. $z^2 = 0.01 \cdot fn(2^{24}-1)/(f-n) = 0.01 \times 100 \times 16{,}777{,}215/999.9 \approx 16{,}778.9$ จึงได้ $z \approx 129.53$ เมตร

2. เมื่อ $f = 10^4$: $z^2 = 0.01 \times 1000 \times 16{,}777{,}215/9999.9 \approx 16{,}777.4$ จึงได้ $z \approx 129.53$ เมตร (เดิม 129.5275 เทียบกับ 129.5334) ต่างกันสัมพัทธ์ราว 4.5×10^{-5} (ไม่ถึงหนึ่งในหมื่นของระยะ) เพราะ $f/(f-n)$ ใกล้ 1 เมื่อ f มาก ทำให้ f แทบไม่มีผลต่อสูตร

3. เมื่อ $n = 0.01$: $z^2 = 0.01 \times 10 \times 16{,}777{,}215/999.99 \approx 1{,}677.7$ จึงได้ $z \approx 40.96$ เมตร เมื่อ $n = 1$: $z^2 = 0.01 \times 1000 \times 16{,}777{,}215/999 \approx 167{,}940$ จึงได้ $z \approx 409.8$ เมตร เมื่อ n เพิ่มจาก 0.01 ถึง 1 คือ 100 เท่า ระยะเพิ่มจาก 40.96 เป็น 409.8 คือราว 10 เท่า $= \sqrt{100}$ และการเพิ่ม n ทีละสิบเท่าให้ระยะเพิ่ม $\sqrt{10} \approx 3.16$ เท่า ($40.96 \to 129.5 \to 409.8$) ดังนั้น $z \propto \sqrt{n}$ จริง

เฉลย 2.3.6.

1. โค้ดให้ระยะแรกที่ช่วงต่อหนึ่งระดับถึง 0.01 เมตรประมาณ 9.38×10^4 เมตร หรือราว 94 กิโลเมตร เทียบกับ 129.5 เมตรของการเก็บแบบจำนวนเต็ม ระยะนี้ไกลกว่าราว 724 เท่า

2. ช่องว่างของทศนิยมลอยตัวเป็น 2^{e-23} เมื่อ $2^e \le d < 2^{e+1}$ จึงคิด $\Delta z = (\text{ช่องว่างของ } d) \cdot z^2/n$ ถ้าเขียน $2^e = \theta\, n/z$ โดย $\theta \in (1/2, 1]$ จะได้ $\Delta z = \theta\, z\, 2^{-23}$ เป็นเชิงเส้นในระยะ z ไม่ใช่กำลังสอง ค่า θ ทำให้ระยะจริงต่างจากการประมาณ $z\,2^{-23}$ ได้ไม่เกินสองเท่า จึงได้ระยะที่ 1 เซนติเมตรอยู่ในช่วง 8.4×10^4 ถึง 1.7×10^5 เมตร ซึ่งสอดคล้องกับ 9.38×10^4 ที่โค้ดให้

3. ที่ $z = 500$ เมตร ช่วงต่อหนึ่งระดับแบบจำนวนเต็มคือ 0.149 เมตร เลื่อนผิวที่สองทีละ 0.01 เมตร ระดับยังเท่าเดิมจนถึงการเลื่อน 0.05 เมตร และเปลี่ยนที่การเลื่อน 0.06 เมตร โค้ดให้ผลนี้ตรงกัน ระยะเลื่อนแรกที่ต่างคือ 0.06 เมตร

เฉลย 2.3.7.

1. ตารางระยะ z (เมตร) ที่หนึ่งระดับของบัฟเฟอร์ 24 บิตครอบคลุม 1 เซนติเมตร:

n (เมตร)	$f = 100$	$f = 1000$	$f = 10^4$
0.01	40.96	40.96	40.96
0.1	129.59	129.53	129.53
1	411.66	409.80	409.62

   ค่า n มีผลมากกว่า f มาก เพราะระยะเป็นสัดส่วนกับ $\sqrt{n}$ ส่วน f มีผลเพียงเล็กน้อยเมื่อ $f \gg n$

2. เลือก $n = 0.1$ และ $f = 1000$ ที่ระยะ 100 เมตร ผิวสองผิวที่ห่าง 1 เซนติเมตรได้ระดับ $16{,}762{,}114$ และ $16{,}762{,}116$ ต่างกันสองระดับ เพราะช่วงต่อหนึ่งระดับที่ 100 เมตรคือ 6 มิลลิเมตร และ 1 เซนติเมตรมีประมาณ 1.7 ระดับ ที่ $f = 100$ ช่วงต่อหนึ่งระดับต่างกันเล็กน้อย แต่ผลเหมือนกัน (โค้ด `same_stored_level` ให้ `False` ทั้งสองกรณี)

3. นาฬิกาในหัวข้อ 2.1: ความผิดใหญ่ (2.79 วินาที) มาจาก**สูตร** คือวิธีบวกตรงในความแม่นเดี่ยว วิธีคาฮานลดเหลือ 1.9×10^{-4} วินาที และส่วนที่เหลือมาจาก**ข้อมูลที่หยาบ** คือการเก็บ $1/60$ เอง กล้องในหัวข้อ 2.2: ความผิดหลักมาจาก**สูตร** $\arccos$ เพราะเปลี่ยนเป็น $\operatorname{atan2}$ แล้วมุม 10^{-4} เรเดียนได้ค่าที่ถูกต้อง ($\arccos$ ให้ 0) เมาส์ในหัวข้อ 2.3:

ฉบับร่าง · Preprint · 2026-10-23

ความไวอยู่ที่**ตัวปัญหา** เองเมื่อใกล้ขอบฟ้า ($\kappa = 941$ ที่แถว 1) ผิวสองผิว: ข้อมูลหยาบ (ระดับละ 1 เซนติเมตรเมื่อไกล) คูณด้วยความไวของการอ่านกลับซึ่งโตตามระยะ การเปลี่ยนชนิดการเก็บเป็นทศนิยมลอยตัวช่วยได้มาก (ข้อ 2.3.6) แต่ไม่ได้เปลี่ยนตัวปัญหาเลย

สรุปบทที่ 2 ในหน้าเดียว

สูตรที่ควรจำจากบทนี้คือ ความผิดของการรวมและการลบ คอนดิชันนัมเบอร์ของปัญหา และความสัมพันธ์ระหว่างความผิดไปข้างหน้ากับความผิดย้อนกลับ ตารางต่อไปนี้รวมสูตรและตัวเลขสำคัญ ทั้งหมดมาจากทฤษฎีบทในบทนี้หรือจากโค้ดของเรา

สูตรและตัวเลขสำคัญ

เรื่อง	สูตร	ตัวอย่างในบทนี้
ค่าคลาดเคลื่อนสัมพัทธ์	$\lvert\tilde{x} - x\rvert / \lvert x\rvert$	นาฬิกา 1200.4 เทียบ 1203.1: $0.22442\,\%$
ผลรวม n ตัว	$\lvert E_n\rvert \le u(1+u)^{n-2} \sum_k S_k$	หนึ่งชั่วโมง: ขอบเขต 23.5 วินาที
ตัวขยายของการลบ	$(\lvert x\rvert + \lvert y\rvert)/\lvert x - y\rvert$	ประมาณ 2×10^4
คอนดิชันนัมเบอร์	$\kappa = \lvert x\, f'(x)/f(x)\rvert$	$\arccos$ ที่ $\theta = 10^{-4}$: 10^8
ลำแสงชนพื้น	$\kappa(\theta) = \theta / (\sin\theta \cos\theta)$	แถว 1: 941.5
ความผิดไปข้างหน้า	$\approx \kappa$ คูณความผิดย้อนกลับ	2.6×10^{-5} ที่แถว 1
ช่วงต่อหนึ่งระดับ	$\Delta z = (f-n) z^2 / (f n (2^{24} - 1))$	1 เซนติเมตรที่ 129.5 เมตร

วิธีที่เทียบกันในบทนี้ (นาฬิกาหนึ่งชั่วโมง)

วิธี	ค่าใช้จ่ายต่อการบวกหนึ่งครั้ง	ความผิดที่ได้
บวกตรง	หนึ่งการบวก	2.79 วินาที
ตัดทิ้ง (chopping)	หนึ่งการบวกและการตัด	10.30 วินาที
รวมคู่ (pairwise)	ตามความลึกของต้นไม้ ประมาณ $\log_2 n$	ไม่เกิน 2.5×10^{-4} วินาที
คาฮาน (Kahan)	ราวสี่การบวกหรือลบ	1.9×10^{-4} วินาที (จากการเก็บ $1/60$)

อ่านต่อ. บทที่ 3 เริ่มจากคำถามที่บทนี้ทิ้งไว้ ถ้าเรารู้ค่าของตัวละครที่คีย์เฟรมไม่กี่จุด ค่าตรงกลางควรหาอย่างไร เส้นตรงระหว่างคีย์เฟรมให้การเคลื่อนไหวที่หยุดเป็นช่วง ๆ ส่วนพหุนามเส้นเดียวที่ผ่านทุกคีย์เฟรมให้ผลอย่างไร และทำไมเพิ่มจำนวนคีย์เฟรมจึงไม่ได้ทำให้ผลดีขึ้นเสมอไป

ฉบับร่าง Preprint · 2026-10-23

ภาค II

จากความต่อเนื่องสู่ข้อมูลไม่ต่อเนื่อง

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 3

เชื่อมจุด

สถานการณ์สมมติ: ตัวละครที่ผ่านสามคีย์เฟรม

คุณคงเคยเห็นแอนิเมชันที่คนวาดภาพสำคัญไว้ไม่กี่ภาพ แล้วคอมพิวเตอร์สร้างภาพที่อยู่ระหว่างกันจนการเคลื่อนไหวดูลื่นตา ภาพสำคัญแบบนี้เรียกว่า **คีย์เฟรม** (keyframe) คือเฟรมที่ผู้ออกแบบกำหนดเองและรู้ค่าแน่นอน ส่วนเฟรมที่เหลือโปรแกรมต้องคำนวณขึ้นมาเอง

ลองนึกถึงตัวละครหนึ่งตัวที่มีสามคีย์เฟรม เฟรมที่ 0 ตัวละครอยู่ที่ตำแหน่ง 0 เมตร เฟรมที่ 60 (วินาทีที่หนึ่ง) อยู่ที่ 3 เมตร และเฟรมที่ 120 (วินาทีที่สอง) กลับมาอยู่ที่ 1 เมตร ระหว่างคีย์ทั้งสามมี 118 เฟรมที่โปรแกรมต้องสร้างเอง (นี่เป็นสถานการณ์สมมติ ตัวเลขทั้งหมดมาจากโค้ดของเรา)

คำถามคือ โปรแกรมควรวาดตัวละครระหว่างคีย์อย่างไร วิธีแรกที่ง่ายที่สุดคือลากเส้นตรงระหว่างคีย์ทีละคู่ วิธีที่สองคือหาพหุนามเส้นเดียวที่ผ่านคีย์ทั้งสามพร้อมกัน ทั้งสองวิธีผ่านคีย์ทุกอันเหมือนกัน แต่การเคลื่อนที่ระหว่างคีย์ต่างกันมาก และถ้าเพิ่มคีย์ให้มากขึ้น ผลจะดีขึ้นเสมอหรือไม่

ลองดูอีกเรื่องที่คนทำแอนิเมชันเจอบ่อย คือการแก้คีย์ที่วางไปแล้ว สมมติว่าเราขยับคีย์ที่ 2 วินาทีจาก 3 เมตรเป็น 2.5 เมตร ตัวละครจะเปลี่ยนเฉพาะช่วงรอบ ๆ คีย์นั้น หรือเปลี่ยนไปไกลกว่านั้น

กับคีย์สามตัว สองวิธีต่างกันไม่มาก เพราะทั้งคู่เปลี่ยนตลอดช่วงระหว่างคีย์แรกกับคีย์สุดท้าย ตัวอย่างหกคีย์ด้านล่างแสดงความต่างที่ชัดกว่า

คำถามของบทนี้จึงมีสองส่วน คือ วิธีต่อจุดแบบไหนทำให้การเคลื่อนที่ดี และการแก้คีย์หนึ่งตัวจะทำให้อะไรเปลี่ยนไปบ้าง

ภาพรวมบท

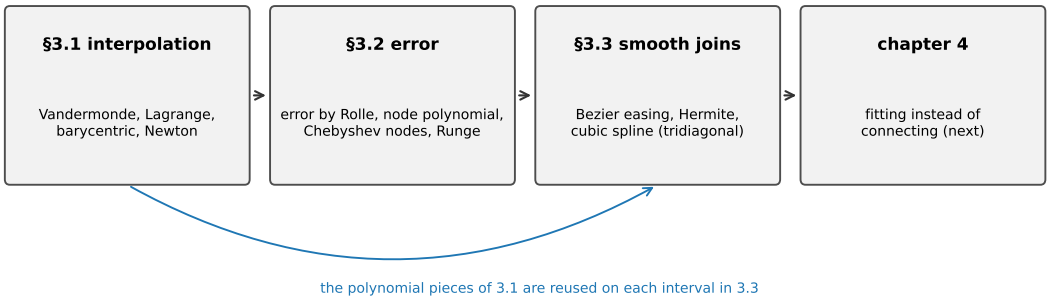
บทนี้ถามคำถามเดียวกันในหลายรูป คือ เมื่อเรารู้ค่าที่บางจุด เราจะเดาค่าระหว่างจุดนั้นอย่างไร และเดาได้ดีแค่ไหน

- ค่าระหว่างจุดหาได้โดยหาฟังก์ชันที่ผ่านจุดที่รู้ แล้วใช้ฟังก์ชันนั้นแทนค่าที่ไม่รู้ การทำเช่นนี้เรียกว่า**การประมาณค่าในช่วง** (interpolation) (หัวข้อ 3.1)
- ถ้าเพิ่มจุดมากขึ้น ค่าคลาดเคลื่อนระหว่างจุดอาจไม่ลดลงเสมอไป (หัวข้อ 3.2)
- เมื่อเรากำลังดูการเคลื่อนที่ ความเร็วต้องเรียบด้วย ไม่ใช่แค่ตำแหน่งที่ผ่านจุด (หัวข้อ 3.3)

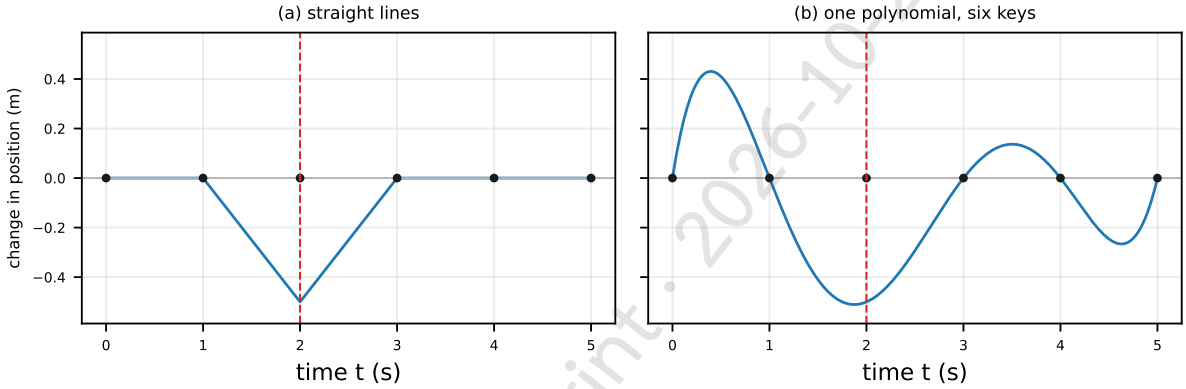
เราเริ่มจากคีย์เฟรมสามอันของตัวละคร แล้วจะเจออีกด้านหนึ่งของปัญหาเดียวกัน คือกล้องที่ต้องผ่านจุดเดิมแต่ความเร็วต้องเรียบ ซึ่งจะเจอในหัวข้อ 3.3

ภาพต่อไปนี้แสดงลำดับของหัวข้อในบทนี้ และสิ่งที่แต่ละหัวข้อใช้ต่อจากหัวข้อก่อนหน้า ถ้าอ่านตามลำดับจะเข้าใจง่ายที่สุด

ฉบับร่าง · preprint · 2026-10-23



รูปที่ 3.1: ลำดับของหัวข้อในบทที่ 3 กล่องแต่ละกล่องคือหัวข้อ ลูกศรสีดำคือการพึ่งพาตามลำดับ และลูกศรสีน้ำเงินคือการใช้พหุนามชิ้นของหัวข้อ 3.1 อีกครั้งในหัวข้อ 3.3 (สร้างจาก `numanal/ch03.py`)



รูปที่ 3.2: การเปลี่ยนของตำแหน่งเมื่อขยับคีย์ที่ 2 วินาที (เส้นประสีแดง) ลง 0.5 เมตร จุดดำคือคีย์ทั้งหก (ก) เส้นตรงเปลี่ยนเฉพาะสองช่วงข้างคีย์ (ข) พหุนามเส้นเดียวเปลี่ยนที่ทุกเวลาที่ไม่ใช่คีย์อื่น (สร้างจาก `numanal/ch03.py`)

ตารางนี้แสดงการเปลี่ยนของตำแหน่ง (เมตร) ที่บางเฟรม เมื่อขยับคีย์ที่เฟรม 120 ลง 0.5 เมตร ตัวอย่างนี้มีหกคีย์ที่เวลา $0, 1, \ldots, 5$ วินาที และตำแหน่ง $0, 1, 3, 2, 0, 1$ เมตร (ตัวเลขจากโค้ดของเรา)

เฟรม	เวลา (วินาที)	เส้นตรง (เมตร)	พหุนาม (เมตร)
30	0.5	0	$+0.41$
90	1.5	-0.25	-0.41
120	2.0	-0.50	-0.50
150	2.5	-0.25	-0.29

ที่เฟรม 30 คีย์ที่ขยับอยู่ห่างออกไปหนึ่งวินาทีครึ่ง เส้นตรงยังไม่เปลี่ยนเลย แต่พหุนามเปลี่ยนไปราว 0.41 เมตร ซึ่งเป็นราวสี่ในห้าของการขยับของคีย์เอง ผลนี้มาจากการที่พหุนามเส้นเดียวใช้ทุกคีย์พร้อมกัน ส่วนเส้นตรงใช้เฉพาะสองช่วงข้างคีย์นั้น ดูรูปที่ 3.2

ฉบับร่าง · preprint · 2026-10-23

3.1 ค่าตรงกลางอยู่ที่ไหน

สถานการณ์สมมติ: คีย์สองอันทับเฟรมเดียวกัน

สมมติว่าเรากำลังวางคีย์ของตัวละครบนไทม์ไลน์ แล้วเผลอลากคีย์ใหม่ไปทับคีย์เดิมที่เฟรม 60 พอดี ตอนนี้เฟรมนั้นมีสองค่า คือ 3 เมตรจากคีย์เดิม และ 2 เมตรจากคีย์ใหม่ ไม่มีเส้นทางใดที่ผ่านทั้งสองค่าพร้อมกันได้ ถ้าสองค่าเท่ากัน เส้นโค้งพหุนามที่ผ่านคีย์ทั้งหมดจะมีได้หลายเส้น และโปรแกรมไม่รู้ว่าควรเลือกเส้นไหน
เราจึงถามว่า คีย์สองอันควรทับกันได้หรือไม่ และถ้าทับกัน เส้นทางควรเป็นอย่างไร

ภาพรวมก่อนอ่าน

ลองนึกถึงการลากปากกาผ่านจุดหลายจุดบนกระดาษ ถ้าลากเป็นเส้นตรงทีละคู่ ปากกาจะหักมุมทุกจุดที่ผ่าน ถ้าอยากได้เส้นโค้งเส้นเดียวที่ผ่านทุกจุดโดยไม่มีรอยหัก เราต้องหาพหุนามที่เหมาะ หัวข้อนี้ตอบสามเรื่อง คือ พหุนามแบบนั้นมีอยู่จริงหรือไม่ มีได้แค่หนึ่งเดียวหรือไม่ และคำนวณได้อย่างไร

โจทย์การเชื่อมจุด

จุดบนแกนที่เรารู้ค่าเรียกว่า **โหนด** (node) ในตัวอย่างของเรา โหนดคือเวลา $t = 0, 1, 2$ วินาที และค่าที่รู้คือตำแหน่ง $y = 0, 3, 1$ เมตร คีย์เฟรมแต่ละอันคือโหนดหนึ่งโหนดพร้อมค่าของมัน

นิยาม 3.1.1: การประมาณค่าในช่วงด้วยพหุนาม

ให้โหนด $t_0, \ldots, t_n$ ต่างกันทั้งหมด และให้ค่าข้อมูล $y_0, \ldots, y_n$ **พหุนามประมาณค่าในช่วง** (interpolation polynomial) ของข้อมูลนี้คือพหุนาม p ที่มีดีกรีไม่เกิน n และ

$$p(t_i) = y_i, \qquad i = 0, 1, \ldots, n.$$

อ่านแบบเข้าใจง่าย

- ดีกรีไม่เกิน n หมายความว่าพหุนามมีพจน์สูงสุดเพียง t^n เราจึงมีตัวไม่ทราบค่า $n+1$ ตัว และมีเงื่อนไข $n+1$ ข้อพอดีกัน
- "ผ่านทุกจุด" หมายถึงผ่านพอดี ไม่ใช่แค่ใกล้จุด
- ตัวอย่างของเรามี $n = 2$ คือจุด $(0, 0)$, $(1, 3)$, $(2, 1)$ เราจึงหาพหุนามดีกรีไม่เกิน 2 ที่ผ่านทั้งสามจุด
- เส้นตรง $y = 2t$ ผ่าน $(0, 0)$ แต่ไม่ผ่าน $(1, 3)$ เพราะ $2 \cdot 1 = 2 \neq 3$ จึงไม่ใช่คำตอบ

เส้นตรงระหว่างคีย์

วิธีแรกคือลากเส้นตรงระหว่างคีย์ที่อยู่ติดกัน วิธีนี้เรียกว่า**การประมาณค่าในช่วงเชิงเส้น** (linear interpolation) บนช่วง $0 \le t \le 1$ เส้นตรงผ่าน $(0, 0)$ และ $(1, 3)$ จึงได้ $y = 3t$ บนช่วง $1 \le t \le 2$ เส้นตรงผ่าน $(1, 3)$ และ $(2, 1)$ จึงได้ $y = 3 - 2(t - 1)$ ผลที่ได้จึงเป็นเส้นตรงสองช่วงต่อกัน ไม่ใช่พหุนามเส้นเดียว

ความเร็วบนช่วงแรกคือ 3 เมตรต่อวินาที และบนช่วงที่สองคือ -2 เมตรต่อวินาที ที่คีย์กลาง ความเร็วจึงเปลี่ยนจาก 3 เป็น -2 ทันที ในหน่วยต่อเฟรม การเดินหนึ่งเฟรมคือ 0.05 เมตรก่อนถึงเฟรม 60 และ $-\frac{1}{30} \approx -0.0333$ เมตรหลังจากนั้น

ฉบับร่าง · Preprint · 2026-10-23

ข้อควรระวัง: ตำแหน่งต่อเนื่องไม่ได้แปลว่าการเคลื่อนที่เรียบ

เส้นตรงต่อคีย์ทำให้ตำแหน่งไม่กระโดด เพราะทุกคีย์ยังถูกผ่าน แต่ความเร็วกระโดดที่คีย์กลาง ตาจะเห็นว่าตัวละครหักมุมกะทันหัน ถ้าเราดูแค่ตำแหน่ง เราจะพลาดจุดนี้

พหุนามเส้นเดียวผ่านทุกคีย์

ลองหาพหุนาม $p(t) = a_0 + a_1 t + a_2 t^2$ ที่ผ่านคีย์ทั้งสาม เงื่อนไข $p(t_i) = y_i$ ให้ระบบสมการเชิงเส้นสามสมการที่มีตัวไม่ทราบค่าสามตัว เขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} 1 & t_0 & t_0^2 \\ 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

เมทริกซ์นี้เรียกว่า**เมทริกซ์แวนเดอร์มอนด์** (Vandermonde matrix) แถวของมันเก็บกำลังของโหนดแต่ละตัว คือ $1, t_i, t_i^2, \ldots$ ในกรณีทั่วไปได้ $V_{ij} = t_i^j$ ดังนั้นการหาสัมประสิทธิ์ของพหุนามก็คือการแก้ระบบเชิงเส้นนี้

กับคีย์ของเรา แถวแรกให้ $a_0 = 0$ แถวที่สองให้ $a_0 + a_1 + a_2 = 3$ และแถวที่สามให้ $a_0 + 2a_1 + 4a_2 = 1$ เมื่อแทน $a_0 = 0$ จะได้ $a_1 + a_2 = 3$ และ $2a_1 + 4a_2 = 1$ จากสมการแรก $a_1 = 3 - a_2$ แทนลงในสมการที่สองได้ $6 + 2a_2 = 1$ จึงได้ $a_2 = -\frac{5}{2}$ และ $a_1 = \frac{11}{2}$ นั่นคือ

$$p(t) = \tfrac{11}{2}\, t - \tfrac{5}{2}\, t^2.$$

ลองเอง: ยอดของพหุนาม

พหุนามนี้ไปถึงค่าสูงสุดที่เวลาไหน ค่านั้นคือเท่าไร ลองหาเวลาที่ $p'(t) = 0$ ก่อนคำนวณ เดาก่อนว่าค่าสูงสุดจะสูงกว่า 3 เมตรหรือไม่

ตัวอย่างนี้ยังไม่ตอบว่าระบบแบบนี้มีคำตอบเสมอหรือไม่ และมีคำตอบเดียวหรือเปล่า ทฤษฎีบทต่อไปนี้ตอบให้ครบทุกกรณี

ทฤษฎีบท 3.1.1: การมีอยู่และความเป็นหนึ่งเดียวของพหุนามประมาณค่า

ให้โหนด $t_0, \ldots, t_n$ ต่างกันทั้งหมด และให้ $y_0, \ldots, y_n$ เป็นจำนวนจริงใด ๆ จะมีพหุนาม p ที่มีดีกรีไม่เกิน n และ $p(t_i) = y_i$ ทุก i และพหุนามนี้มีเพียงตัวเดียว

อ่านแบบเข้าใจง่าย

ทฤษฎีบทนี้บอกว่า มีคำตอบเสมอ และมีคำตอบเดียว เงื่อนไข "โหนดต่างกันทั้งหมด" สำคัญมาก ถ้าโหนดสองตัวซ้ำกัน อาจไม่มีคำตอบเลย หรือมีได้มากกว่าหนึ่งตัว ดูตัวอย่างแรกได้ในแบบฝึกหัด 3.1.7 เมื่อมีคำตอบเดียวเสมอ เราจึงพูดถึง "พหุนามประมาณค่า" ได้โดยไม่กำกวม ไม่ว่าจะคำนวณด้วยวิธีใด ผลก็เป็นพหุนามเดียวกัน

ทฤษฎีบทนี้อยู่ใน Atkinson 1989, §3.1, Thm 3.1, p. 132 ซึ่งให้การพิสูจน์สามแบบ การพิสูจน์ด้านล่างเป็นของเราเอง

แนวคิด. ความเป็นหนึ่งเดียวมาจากการนับราก ส่วนการมีอยู่สร้างพหุนามจากพหุนามพื้นฐานที่เป็น 1 ที่โหนดของมันเอง และเป็น 0 ที่โหนดอื่น

1. **ความเป็นหนึ่งเดียว.** สมมติ p และ q ต่างก็เป็นคำตอบ ให้ $r = p - q$ ซึ่งมีดีกรีไม่เกิน n และ $r(t_i) = y_i - y_i = 0$ ที่โหนดทั้ง $n + 1$ ตัว พหุนามที่ไม่เป็นศูนย์และมีดีกรี d มีรากได้ไม่เกิน d ตัว เพราะทุกรากให้ตัวประกอบ $(t - t_i)$ หนึ่ง

ฉบับร่าง · Preprint 2026-10-23

ตัว ดังนั้น r ต้องเป็นพหุนามศูนย์ และ $p = q$

2. **การมีอยู่.** กำหนด $\ell_k(t) = \prod_{j \neq k} \dfrac{t - t_j}{t_k - t_j}$ ซึ่งมีดีกรี n เมื่อ $i \neq k$ ตัวประกอบ $(t_i - t_i)$ เป็นศูนย์จึงได้ $\ell_k(t_i) = 0$ ส่วน $\ell_k(t_k) = 1$ เพราะทุกตัวประกอบเป็น 1 นั่นคือ $\ell_k(t_i) = \delta_{ik}$ เมื่อ δ_{ik} เท่ากับ 1 ถ้า $i = k$ และเท่ากับ 0 ถ้า $i \neq k$ จากนั้น $p(t) = \sum_{k=0}^{n} y_k \ell_k(t)$ มีดีกรีไม่เกิน n และ $p(t_i) = y_i$

3. **ระบบมีคำตอบเดียว.** ถ้า $Vb = 0$ พหุนาม $\sum_j b_j t^j$ เป็นศูนย์ที่ทุกโหนด จึงเป็นพหุนามศูนย์ตามข้อ 1 และ $b = 0$ ดังนั้น V ไม่เอกฐาน ระบบ $Va = y$ จึงมีคำตอบเดียว

สรุป: มีพหุนามดีกรีไม่เกิน n ที่ผ่านทุกจุดเสมอ และมีเพียงตัวเดียว ดังนั้นคำว่า "พหุนามประมาณค่า" จึงมีความหมายชัดเจน

สูตรลากรานจ์

ตัวพหุนาม ℓ_k ในการพิสูจน์มีชื่อว่า**พหุนามลากรานจ์** (Lagrange polynomial) เมื่อรวมมันด้วยค่า y_k จะได้**รูปแบบลากรานจ์** (Lagrange form) ของพหุนามประมาณค่า

$$p(t) = \sum_{k=0}^{n} y_k \, \ell_k(t).$$

ลองนึกว่าแต่ละ ℓ_k เป็นสวิตช์ที่เปิดเฉพาะที่โหนดของมันเอง ที่โหนดแต่ละอันมีเพียงพจน์เดียวที่ไม่เป็นศูนย์ จึงได้ค่าตรงกับข้อมูลพอดี

สำหรับคีย์ของเรา $\ell_0(t) = \dfrac{(t-1)(t-2)}{2}$, $\ell_1(t) = -t(t-2)$ และ $\ell_2(t) = \dfrac{t(t-1)}{2}$ ดังนั้น $p(t) = 3\ell_1(t) + \ell_2(t)$ ซึ่งกระจายได้เป็น $\frac{11}{2}t - \frac{5}{2}t^2$ ตรงกับผลจากระบบสมการ รูปแบบนี้ไม่ต้องแก้ระบบ แต่ถ้าเพิ่มคีย์ใหม่ พหุนาม ℓ_k ทุกตัวต้องคำนวณใหม่

รูปแบบบารีเซนทริก

เมื่อต้องประเมินค่าที่จุดจำนวนมาก เช่นทุกเฟรม การคำนวณ ℓ_k ทุกตัวทุกครั้งทำให้ช้างานมาก รูปแบบบารีเซนทริกแก้ปัญหานี้โดยคำนวณน้ำหนักชุดเดียวครั้งเดียว แล้วใช้กับทุกจุดที่ต้องการ

ทฤษฎีบท 3.1.2: รูปแบบบารีเซนทริก

ให้โหนด $t_0, \ldots, t_n$ ต่างกันทั้งหมด และกำหนดน้ำหนัก $w_k = \prod_{j \neq k} \dfrac{1}{t_k - t_j}$ สำหรับ t ที่ไม่ใช่โหนดใด ๆ จะได้

$$p(t) = \frac{\displaystyle\sum_{k=0}^{n} \frac{w_k \, y_k}{t - t_k}}{\displaystyle\sum_{k=0}^{n} \frac{w_k}{t - t_k}}.$$

แนวคิด. รูปแบบลากรานจ์มี $\Omega(t) = \prod_j (t - t_j)$ คูณอยู่ทุกพจน์ เราหาค่าของ Ω ที่หายไปได้จากการใช้ข้อมูลที่เป็น 1 ทุกจุด

1. **เขียน ℓ_k ด้วย Ω.** เมื่อ t ไม่ใช่โหนด $\prod_{j \neq k}(t - t_j) = \Omega(t)/(t - t_k)$ ดังนั้น $\ell_k(t) = w_k \, \Omega(t)/(t - t_k)$

2. **แทนลงในรูปแบบลากรานจ์.** จะได้ $p(t) = \Omega(t) \sum_k w_k y_k/(t - t_k)$

3. **ใช้ข้อมูลที่เป็น 1 ทุกจุด.** พหุนามที่ผ่านค่า 1 ทุกจุดคือค่าคงที่ 1 ซึ่งมีดีกรีไม่เกิน n เมื่อใช้ความเป็นหนึ่งเดียว จะได้ $1 = \Omega(t) \sum_k w_k/(t - t_k)$

ฉบับร่าง Preprint · 2026-10-23

4. **หารกัน.** ตัวหารในข้อ 3 เท่ากับ $1/\Omega(t)$ ซึ่งไม่เป็นศูนย์ จึงหารสมการในข้อ 2 ด้วยข้อ 3 ได้สูตรที่ต้องการ

สรุป: เมื่อเรารู้น้ำหนัก w_k แล้ว ค่าที่จุดใดก็ได้คือผลหารของผลรวมสองชุดที่ใช้ตัวเลขชุดเดิม

สำหรับคีย์ของเรา $w_0 = \frac{1}{2}$, $w_1 = -1$, $w_2 = \frac{1}{2}$ ที่ $t = \frac{1}{2}$ ตัวเศษคือ $6 - \frac{1}{3} = \frac{17}{3}$ และตัวส่วนคือ $1 + 2 - \frac{1}{3} = \frac{8}{3}$ จึงได้ $p(\frac{1}{2}) = \frac{17}{8} = 2.125$ ซึ่งตรงกับรูปแบบลากรานจ์

เกร็ดเพิ่มเติม

Atkinson อธิบายรูปแบบลากรานจ์ของพหุนามประมาณค่าไว้ในบทที่ 3 Atkinson 1989, §3.1, (3.1.6), p. 134 ส่วนรูปแบบบารีเซนทริกข้างบนเป็นการพิสูจน์ของเราเอง ไม่ได้มาจากหนังสือเล่มนั้น

ผลต่างหารและสูตรของนิวตัน

รูปแบบลากรานจ์เพิ่มคีย์ใหม่ได้ยาก เพราะพหุนาม ℓ_k ทุกตัวต้องเปลี่ยน รูปแบบของนิวตันสร้างพหุนามทีละขั้น เพิ่มคีย์หนึ่งตัวจะเพิ่มพจน์หนึ่งพจน์โดยไม่แตะพจน์เดิม ตัวเลขที่ใช้สร้างพจน์เหล่านั้นเรียกว่าผลต่างหาร ซึ่งนิยามดังนี้

นิยาม 3.1.2: ผลต่างหาร

ให้ f เป็นฟังก์ชัน และให้โหนด $t_0, \ldots, t_k$ ต่างกันทั้งหมด **ผลต่างหาร** (divided difference) $f[t_0, \ldots, t_k]$ คือสัมประสิทธิ์ของ t^k ในพหุนามประมาณค่าที่ผ่านจุด $(t_i, f(t_i))$ ทั้ง $k+1$ จุด

อ่านแบบเข้าใจง่าย

- อันดับ 0: $f[t_0] = f(t_0)$ คือค่าข้อมูลเอง
- อันดับ 1: $f[t_0, t_1]$ คือความชันของเส้นตรงที่ผ่านสองคีย์ ของเราคือ $f[0,1] = \dfrac{3-0}{1-0} = 3$
- อันดับ 2 และสูงกว่า บอกว่าความชันเปลี่ยนเร็วแค่ไหนเมื่อเพิ่มโหนด

การคำนวณผลต่างหารไม่ต้องเริ่มจากพหุนาม มีสูตรเวียนซ้ำที่ใช้ผลต่างหารอันดับที่ต่ำกว่า

ทฤษฎีบท 3.1.3: การคำนวณซ้ำของผลต่างหาร

สำหรับโหนดต่างกันทั้งหมด $t_0, \ldots, t_k$ เมื่อ $k \geq 1$ จะได้

$$f[t_0, \ldots, t_k] = \frac{f[t_1, \ldots, t_k] - f[t_0, \ldots, t_{k-1}]}{t_k - t_0}.$$

นอกจากนี้ $f[t_0, \ldots, t_k]$ ไม่ขึ้นกับลำดับของโหนด

ทฤษฎีบทนี้อยู่ใน Atkinson 1989, §3.2, (3.2.7), p. 139 และ Iserles 2009, ภาคผนวก A.2.2.4, pp. 442--443 การพิสูจน์ด้านล่างเป็นของเราเอง

แนวคิด. สร้างพหุนามใหม่จากพหุนามของโหนดบางส่วนสองชุด แล้วใช้ความเป็นหนึ่งเดียวเทียบสัมประสิทธิ์ของ t^k

1. **เตรียมพหุนามสองตัว.** ให้ q เป็นพหุนามประมาณค่าที่โหนด $t_0, \ldots, t_{k-1}$ และ r ที่โหนด $t_1, \ldots, t_k$ ทั้งสองมีดีกรีไม่เกิน $k-1$ สัมประสิทธิ์ของ t^{k-1} ในทั้งสองตัวคือ $f[t_0, \ldots, t_{k-1}]$ และ $f[t_1, \ldots, t_k]$ ตามลำดับ

ฉบับร่าง · Preprint · 2026-1023

2. **สร้างพหุนามใหม่.** ให้ $s(t) = \dfrac{(t-t_0)\,r(t) - (t-t_k)\,q(t)}{t_k - t_0}$ ตรวจที่โหนดทีละตัว ถ้า $1 \le i \le k-1$ ทั้ง q และ r เท่ากับ $f(t_i)$ จึงได้ $s(t_i) = f(t_i)\dfrac{(t_i-t_0)-(t_i-t_k)}{t_k-t_0} = f(t_i)$ ที่ t_0 ได้ $s(t_0) = q(t_0) = f(t_0)$ เพราะ $t_0 - t_k = -(t_k - t_0)$ และที่ t_k ได้ $s(t_k) = r(t_k) = f(t_k)$

3. **ใช้ความเป็นหนึ่งเดียว.** s มีดีกรีไม่เกิน k และผ่านข้อมูลทุกจุด จึงเป็นพหุนามประมาณค่าที่โหนด $t_0, \ldots, t_k$ ทั้งหมด

4. **เทียบสัมประสิทธิ์ของ t^k.** ในพจน์ $(t-t_0)r(t)$ สัมประสิทธิ์ของ t^k คือสัมประสิทธิ์ของ t^{k-1} ใน r ซึ่งคือ $f[t_1, \ldots, t_k]$ ในพจน์ $(t-t_k)q(t)$ สัมประสิทธิ์ของ t^k คือ $f[t_0, \ldots, t_{k-1}]$ จึงได้สูตรในทฤษฎีบท

5. **ลำดับ.** พหุนามประมาณค่าขึ้นกับเซตของโหนดเท่านั้น ไม่ขึ้นกับลำดับที่เขียน ดังนั้นสัมประสิทธิ์ของ t^k ก็ไม่ขึ้นกับลำดับ

สรุป: ผลต่างหารอันดับ k คำนวณจากผลต่างหารอันดับ $k-1$ สองตัว และไม่ขึ้นกับลำดับโหนด

สำหรับคีย์ของเรา ตารางผลต่างหารคำนวณได้ดังนี้ แต่ละช่องคำนวณจากสองช่องที่อยู่ทางซ้ายด้วยสูตรเวียนซ้ำ

t_i	$f[t_i]$	$f[t_i, t_{i+1}]$	$f[t_i, t_{i+1}, t_{i+2}]$
0	0	3	$-\frac{5}{2}$
1	3	-2	
2	1		

เช่น $f[0,1,2] = \dfrac{f[1,2]-f[0,1]}{2-0} = \dfrac{-2-3}{2} = -\frac{5}{2}$ ซึ่งเป็นสัมประสิทธิ์ของ t^2 ในพหุนามที่เราหาไว้ก่อนหน้า

ข้อควรระวัง: ตัวหารเล็กทำให้ผลต่างหารเพี้ยน

ตัวหาร $t_k - t_0$ คือระยะระหว่างโหนด ถ้าโหนดสองตัวอยู่ใกล้กันมาก ตัวหารจะเล็ก และถ้าค่าข้อมูลมีค่าคลาดเคลื่อนเล็กน้อยอยู่แล้ว ผลต่างหารอาจเพี้ยนมากกว่าค่าคลาดเคลื่อนของข้อมูลหลายเท่า ตรงนี้คือการลบตัวเลขที่ใกล้กันแบบเดียวกับที่เราเห็นในบทที่ 2

ทฤษฎีบท 3.1.4: สูตรของนิวตัน

ให้โหนด $t_0, \ldots, t_n$ ต่างกันทั้งหมด พหุนามประมาณค่าคือ

$$p(t) = \sum_{k=0}^{n} f[t_0, \ldots, t_k]\,(t-t_0)(t-t_1)\cdots(t-t_{k-1}),$$

โดยผลคูณที่ว่างเมื่อ $k=0$ ให้ค่าเป็น 1

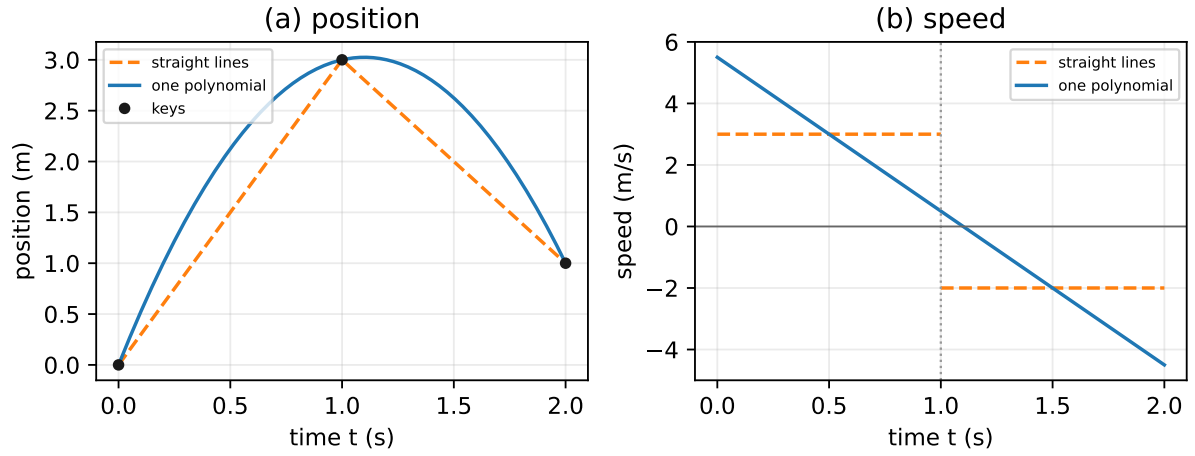
ทฤษฎีบทนี้คือสูตรที่ Atkinson 1989, §3.2, (3.2.9), p. 140 ให้ไว้ ดู Iserles 2009, ภาคผนวก A.2.2.4, p. 442 ด้วย การพิสูจน์ด้านล่างเป็นของเราเอง

แนวคิด. ผลต่าง $p_n - p_{n-1}$ เป็นศูนย์ที่โหนดเก่าทั้งหมด จึงมีตัวประกอบเป็นผลคูณของพจน์ใหม่ และสัมประสิทธิ์ของพจน์นั้นอ่านได้จาก t^n

1. **เริ่มที่ $n=0$.** $p_0(t) = y_0 = f[t_0]$ ซึ่งเป็นสูตรของทฤษฎีบท

2. **ผลต่างของพหุนามสองขั้น.** สมมติสูตรจริงถึง $n-1$ ให้ $D = p_n - p_{n-1}$ ซึ่งมีดีกรีไม่เกิน n และ $D(t_i) = 0$ สำหรับ $i = 0, \ldots, n-1$ เพราะทั้งสองผ่านข้อมูลที่โหนดเหล่านั้น

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 3.3: (ก) ตำแหน่งตามเวลา เส้นประเป็นเส้นตรงสองช่วง เส้นทึบเป็นพหุนามดีกรีสองเส้นเดียว จุดดำคือคีย์ทั้งสาม (ข) ความเร็ว เส้นตรงกระโดดที่ $t = 1$ จาก 3 เป็น -2 เมตรต่อวินาที ส่วนพหุนามเปลี่ยนอย่างต่อเนื่อง (สร้างจาก `numanal/ch03.py`)

3. **ตัวประกอบ.** D มี n ราก จึงหารด้วย $\prod_{j<n}(t - t_j)$ ลงตัว ผลหารมีดีกรีไม่เกิน 0 เป็นค่าคงที่ κ นั่นคือ $D(t) = \kappa \prod_{j<n}(t - t_j)$

4. **สัมประสิทธิ์ของ t^n.** p_{n-1} ไม่มีพจน์ t^n ดังนั้น κ คือสัมประสิทธิ์ของ t^n ใน p_n ซึ่งคือ $f[t_0, \ldots, t_n]$ จึงได้ $p_n = p_{n-1} + f[t_0, \ldots, t_n] \prod_{j<n}(t - t_j)$ ซึ่งเป็นสูตรที่ต้องการเมื่อใช้เหนี่ยวนำ

สรุป: พจน์ของนิวตันแต่ละพจน์ถูกเพิ่มเข้าไปโดยไม่แก้พจน์ก่อนหน้า

สำหรับคีย์ของเรา สัมประสิทธิ์ของสูตรของนิวตันคือ $f[0] = 0$, $f[0, 1] = 3$ และ $f[0, 1, 2] = -\frac{5}{2}$ จึงได้

$$p(t) = 0 + 3\,t - \tfrac{5}{2}\,t(t-1) = \tfrac{11}{2}\,t - \tfrac{5}{2}\,t^2$$

ซึ่งเป็นพหุนามเดียวกับที่ได้จากระบบเชิงเส้นและจากรูปแบบลากรานจ์ เพราะความเป็นหนึ่งเดียว

อ่านแบบเข้าใจง่าย

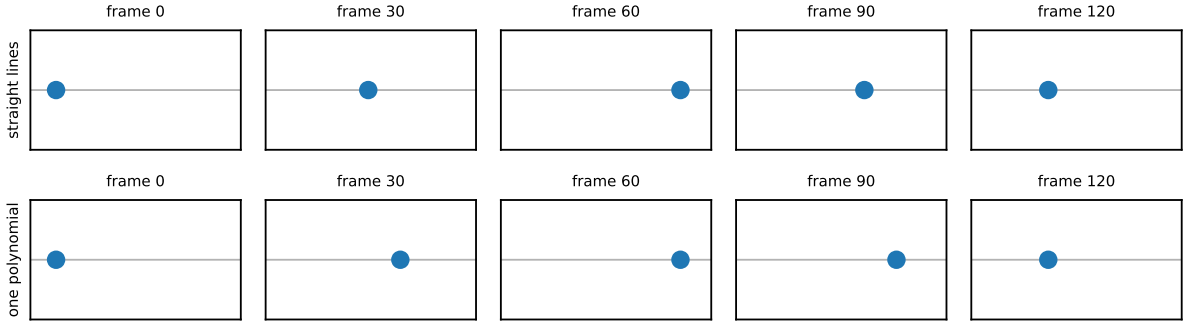
ถ้าเพิ่มคีย์ที่สี่ $(t, y) = (3, 4)$ สัมประสิทธิ์สามตัวแรกของรูปแบบนิวตัน (0, 3 และ $-\frac{5}{2}$) จะไม่เปลี่ยนเลย เราเพิ่มเพียงพจน์ $f[0, 1, 2, 3]\,t(t-1)(t-2)$ ส่วนรูปแบบลากรานจ์ต้องคำนวณพหุนาม ℓ_k ทุกตัวใหม่ ดูแบบฝึกหัด 3.1.6

ดูเฟรมที่เกิดขึ้นจริง

ตอนนี้เรามีพหุนามที่ผ่านคีย์ทั้งสาม และมีเส้นตรงสองช่วง ลองดูว่าเฟรมที่โปรแกรมสร้างขึ้นหน้าตาเป็นอย่างไร ที่เฟรม 30 ($t = 0.5$) เส้นตรงให้ตำแหน่ง 1.5 เมตร ส่วนพหุนามให้ 2.125 เมตร ที่เฟรม 90 ($t = 1.5$) เส้นตรงให้ 2.0 เมตร ส่วนพหุนามให้ 2.625 เมตร ทั้งสองแบบผ่านคีย์เดียวกัน แต่เฟรมระหว่างคีย์ต่างกันชัดเจน

ความเร็วก็ต่างกันด้วย ตอนเริ่มต้น พหุนามมีความเร็ว 5.5 เมตรต่อวินาที ขณะที่เส้นตรงมี 3 เมตรต่อวินาที ที่คีย์กลางพหุนามมีความเร็ว 0.5 เมตรต่อวินาที โดยไม่กระโดด แล้วความเร็วลดลงเรื่อย ๆ จนเป็น -4.5 เมตรต่อวินาทีที่วินาทีที่สอง ตัวเลขทั้งหมดมาจากโค้ดของเรา

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 3.4: แถบเฟรมที่ 0, 30, 60, 90 และ 120 ของตัวละคร แถวบนเป็นเส้นตรงสองช่วง แถวล่างเป็นพหุนาม เฟรม 60 เป็นคีย์กลาง ทั้งสองแถวตรงกัน (สร้างจาก `numanal/ch03.py`)

ลองเอง: ทายก่อนรัน

ก่อนรันโค้ดใน `numanal/ch03.py`ให้ทายตำแหน่งของตัวละครที่เฟรม 90 ในแบบเส้นตรงและแบบพหุนาม แล้วเปรียบเทียบกับผลที่โค้ดพิมพ์ออกมา เพราะเหตุใดที่เฟรม 60 ทั้งสองแบบจึงให้ตำแหน่งเดียวกัน

สรุปใจความ

- การประมาณค่าในช่วงคือการหาพหุนามที่ผ่านทุกจุดข้อมูล ถ้าโหนดมี $n+1$ จุดที่ต่างกัน พหุนามดีกรีไม่เกิน n มีอยู่และมีเพียงตัวเดียวเสมอ
- เส้นตรงต่อคีย์ทำให้ตำแหน่งต่อเนื่องแต่ความเร็วกระโดด พหุนามเส้นเดียวทำให้ความเร็วเรียบ แต่พหุนามอาจแกว่งเกินค่าของคีย์ได้ เช่นพหุนามของเราไปถึง 3.025 เมตรที่ $t = 1.1$ ทั้งที่คีย์ทุกอันอยู่ไม่เกิน 3 เมตร
- รูปแบบลากรานจ์และบารีเซนทริกให้พหุนามเดียวกัน แต่บารีเซนทริกประหยัดเมื่อประเมินค่าหลายจุด รูปแบบของนิวตันเพิ่มคีย์ได้ทีละพจน์
- คีย์สองอันที่ทับเฟรมเดียวกันทำให้โจทย์ไม่มีคำตอบ (ถ้าค่าต่างกัน) หรือมีคำตอบมากมาย (ถ้าค่าเท่ากัน) ทฤษฎีบทจึงต้องการโหนดต่างกันทั้งหมด (ดูแบบฝึกหัด 3.1.7)
- กลับไปที่คำถามของตอนต้น โปรแกรมควรวาดตัวละครด้วยเส้นตรงหรือพหุนาม คำตอบยังไม่จบ เพราะเรายังไม่รู้ว่าเพิ่มคีย์แล้วผลจะดีขึ้นหรือไม่ (หัวข้อ 3.2) และความเร็วที่เรียบยังสำคัญเพียงใด (หัวข้อ 3.3)

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 3.1.1 (⋆) ให้คีย์ $(t, y) = (0, 0),\ (1, 2),\ (2, 6)$

1. จงหาพหุนามดีกรีไม่เกิน 2 ที่ผ่านทั้งสามจุด โดยใช้รูปแบบลากรานจ์
2. จงหาความเร็วของพหุนามนั้นที่ $t = 0,\ 1,\ 2$
3. จงหาความเร็วของเส้นตรงสองช่วงที่ต่อคีย์ทั้งสาม และเทียบกับข้อ (ข)

แบบฝึกหัด 3.1.2 (⋆) ใช้ข้อมูลของแบบฝึกหัด 3.1.1

ฉบับร่าง · Preprint 2026-05-23

1. จงสร้างตารางผลต่างหารของข้อมูลนี้ จนได้ $f[t_0, t_1, t_2]$
2. จงเขียนพหุนามในรูปแบบของนิวตัน แล้วแสดงว่าสัมประสิทธิ์ของ t^2 เท่ากับ $f[t_0, t_1, t_2]$

แบบฝึกหัด 3.1.3 (⋆⋆) ใช้ข้อมูลของแบบฝึกหัด 3.1.1 จงหาค่าที่ $t = \frac{3}{2}$ สองวิธี คือด้วยรูปแบบบารีเซนทริก และด้วยรูปแบบลากรานจ์ แล้วเทียบผลลัพธ์ จากนั้นอธิบายว่าทำไมการคำนวณด้วยรูปแบบบารีเซนทริกจึงใช้น้ำหนัก w_k ชุดเดียวกันได้ทุกจุด

แบบฝึกหัด 3.1.4 (⋆⋆) ให้โหนด $t = 0,\ 1,\ 3$ และค่า $y = 0,\ 2,\ 0$

1. จงหาดีเทอร์มิแนนต์ของเมทริกซ์แวนเดอร์มอนด์ของโหนดเหล่านี้ด้วยมือ
2. จงแก้ระบบ $Va = y$ เพื่อหาสัมประสิทธิ์ของพหุนามดีกรีไม่เกิน 2 ที่ผ่านจุดเหล่านี้
3. จงหาเวลาที่พหุนามนั้นมีค่าสูงสุด และค่าสูงสุดนั้น

แบบฝึกหัด 3.1.5 (⋆⋆) ให้ข้อมูลเป็น $t = 2,\ 1,\ 0$ และ $y = 1,\ 3,\ 0$ ซึ่งเป็นลำดับย้อนกลับของตัวอย่างในหัวข้อนี้

1. จงคำนวณ $f[2, 1, 0]$ ด้วยตารางผลต่างหาร
2. จงอธิบายว่าทำไมค่านี้ต้องเท่ากับ $f[0, 1, 2] = -\frac{5}{2}$ เสมอ ไม่ว่าเรียงโหนดอย่างไร

แบบฝึกหัด 3.1.6 (⋆⋆⋆) เพิ่มคีย์ที่สี่ $(t, y) = (3, 4)$ ให้กับคีย์ของหัวข้อนี้ $(0, 0),\ (1, 3),\ (2, 1)$

1. จงหา $f[0, 1, 2, 3]$ ด้วยตารางผลต่างหาร
2. จงเขียนพหุนามดีกรีไม่เกิน 3 ในรูปแบบของนิวตัน แล้วแสดงว่าพจน์เดิมสามพจน์ไม่เปลี่ยน
3. จงคำนวณค่าของพหุนามดีกรีสามที่ $t = 0.5,\ 1.5,\ 2.5$ และเทียบกับพหุนามดีกรีสองเดิมที่จุดเดียวกัน
4. ในรูปแบบลากรานจ์ การเพิ่มคีย์ที่สี่ทำให้ต้องทำอะไรมากกว่าในรูปแบบของนิวตัน

แบบฝึกหัด 3.1.7 (⋆⋆) ลองดูคีย์สองตัวที่เวลาเดียวกัน $(t, y) = (0, 0),\ (1, 3),\ (1, 1)$

1. จงแสดงว่าดีเทอร์มิแนนต์ของเมทริกซ์แวนเดอร์มอนด์ของโหนดเหล่านี้เป็นศูนย์
2. จงอธิบายว่าทำไมไม่มีพหุนามใด ๆ ที่ผ่านคีย์ทั้งสามพร้อมกัน ไม่ว่าดีกรีจะเท่าไร
3. ในแอนิเมชัน ข้อนี้หมายความว่าอย่างไร และเหตุใดทฤษฎีบทจึงต้องการให้โหนดต่างกันทั้งหมด

ฉบับร่าง · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 3.1.1.

1. โหนดคือ $t_0 = 0, t_1 = 1, t_2 = 2$ และค่าคือ $y_0 = 0, y_1 = 2, y_2 = 6$ พหุนามลากรานจ์คือ

$$\ell_0(t) = \frac{(t-1)(t-2)}{2}, \qquad \ell_1(t) = \frac{t(t-2)}{-1}, \qquad \ell_2(t) = \frac{t(t-1)}{2}.$$

   แล้ว $p(t) = 0 \cdot \ell_0(t) + 2\ell_1(t) + 6\ell_2(t) = -2t(t-2) + 3t(t-1) = -2t^2 + 4t + 3t^2 - 3t$ จึงได้ $p(t) = t^2 + t$ ตรวจได้ว่า $p(1) = 2$ และ $p(2) = 6$

2. อนุพันธ์คือ $p'(t) = 2t + 1$ จึงได้ความเร็ว $p'(0) = 1, p'(1) = 3$ และ $p'(2) = 5$ มีหน่วยเป็น y ต่อวินาที

3. เส้นตรงช่วงแรกมีความชัน $(2-0)/(1-0) = 2$ และช่วงที่สองมีความชัน $(6-2)/(2-1) = 4$ ความเร็วของเส้นตรงจึงกระโดดจาก 2 เป็น 4 ที่คีย์กลาง ส่วนพหุนามความเร็วเปลี่ยนอย่างต่อเนื่องจาก 1 ที่ $t = 0$ ผ่าน 3 ที่ $t = 1$ ไปเป็น 5 ที่ $t = 2$ โดยไม่กระโดด

ตรวจด้วยโค้ดใน `test_ch03.py`(ชุดทดสอบของแบบฝึกหัด 3.1.1).

เฉลย 3.1.2.

1. เริ่มจาก $f[t_0] = 0, f[t_1] = 2, f[t_2] = 6$ อันดับ 1 คือ

$$f[t_0, t_1] = \frac{2-0}{1-0} = 2, \qquad f[t_1, t_2] = \frac{6-2}{2-1} = 4.$$

   อันดับ 2 คือ $f[t_0, t_1, t_2] = \dfrac{f[t_1, t_2] - f[t_0, t_1]}{t_2 - t_0} = \dfrac{4-2}{2-0} = 1$

2. พหุนามในรูปแบบของนิวตันคือ $p(t) = f[t_0] + f[t_0, t_1]\,(t - t_0) + f[t_0, t_1, t_2]\,(t - t_0)(t - t_1) = 0 + 2t + t(t-1) = t^2 + t$ สัมประสิทธิ์ของ t^2 คือ 1 ซึ่งเท่ากับ $f[t_0, t_1, t_2]$ ตามทฤษฎีบทสูตรของนิวตัน

ตรวจด้วยโค้ดใน `test_ch03.py`(ชุดทดสอบของแบบฝึกหัด 3.1.2).

เฉลย 3.1.3. น้ำหนักคือ $w_0 = \dfrac{1}{(0-1)(0-2)} = \dfrac{1}{2}, w_1 = \dfrac{1}{(1-0)(1-2)} = -1$ และ $w_2 = \dfrac{1}{(2-0)(2-1)} = \dfrac{1}{2}$ ที่ $t = \frac{3}{2}$

- รูปแบบบารีเซนทริก: ตัวส่วนคือ $\dfrac{1/2}{3/2} + \dfrac{-1}{1/2} + \dfrac{1/2}{-1/2} = \dfrac{1}{3} - 2 - 1 = -\dfrac{8}{3}$ และตัวเศษคือ $\dfrac{(1/2)\cdot 0}{3/2} + \dfrac{(-1)\cdot 2}{1/2} + \dfrac{(1/2)\cdot 6}{-1/2} = 0 - 4 - 6 = -10$ จึงได้ $p(\frac{3}{2}) = \dfrac{-10}{-8/3} = \dfrac{15}{4} = 3.75$
- รูปแบบลากรานจ์: $\ell_0(\frac{3}{2}) = \dfrac{(1/2)(-1/2)}{2} = -\dfrac{1}{8}, \ell_1(\frac{3}{2}) = -\frac{3}{2}\cdot(-\frac{1}{2}) = \frac{3}{4}$ และ $\ell_2(\frac{3}{2}) = \dfrac{(3/2)(1/2)}{2} = \dfrac{3}{8}$ จึงได้ $p(\frac{3}{2}) = 0 + 2\cdot\frac{3}{4} + 6\cdot\frac{3}{8} = \frac{3}{2} + \frac{9}{4} = \frac{15}{4}$ ตรงกัน

น้ำหนัก w_k ขึ้นกับโหนดเท่านั้น ไม่ขึ้นกับจุด t ที่เราต้องการค่า ดังนั้นเราคำนวณน้ำหนักครั้งเดียว แล้วใช้กับทุกจุดได้ ตรวจด้วยโค้ดใน `test_ch03.py`(ชุดทดสอบของแบบฝึกหัด 3.1.3).

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 3.1.4.

1. เมทริกซ์แวนเดอร์มอนด์คือ $V = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{bmatrix}$ กระจายตามแถวแรกได้ $\det V = 1 \cdot (1 \cdot 9 - 1 \cdot 3) = 6$ ซึ่งไม่เป็นศูนย์ จึงมีคำตอบเดียว

2. ระบบคือ $a_0 = 0$, $a_0 + a_1 + a_2 = 2$ และ $a_0 + 3a_1 + 9a_2 = 0$ จาก $a_0 = 0$ จะได้ $a_1 + a_2 = 2$ และ $3a_1 + 9a_2 = 0$ จึงได้ $a_1 = -3a_2$ แทนลงไปได้ $-2a_2 = 2$ นั่นคือ $a_2 = -1$ และ $a_1 = 3$ พหุนามคือ $p(t) = 3t - t^2$ ตรวจว่า $p(1) = 2$ และ $p(3) = 0$

3. อนุพันธ์คือ $p'(t) = 3 - 2t$ เป็นศูนย์ที่ $t = \frac{3}{2}$ ค่าสูงสุดคือ $p(\frac{3}{2}) = \frac{9}{2} - \frac{9}{4} = \frac{9}{4}$

ตรวจด้วยโค้ดใน `test_ch03.py`(ชุดทดสอบของแบบฝึกหัด 3.1.4) โค้ดใช้ `numpy` ตรวจดีเทอร์มิแนนต์ได้ 6 เป็นการตรวจแยกต่างหาก

เฉลย 3.1.5.

1. โหนดเรียงเป็น $t_0 = 2, t_1 = 1, t_2 = 0$ และค่า $f(t_0) = 1$, $f(t_1) = 3$, $f(t_2) = 0$

$$f[2,1] = \frac{3-1}{1-2} = -2, \qquad f[1,0] = \frac{0-3}{0-1} = 3.$$

   แล้ว $f[2,1,0] = \dfrac{f[1,0] - f[2,1]}{0-2} = \dfrac{3-(-2)}{-2} = -\frac{5}{2}$

2. พหุนามประมาณค่าขึ้นกับเซตของโหนด ไม่ขึ้นกับลำดับที่เขียน (ความเป็นหนึ่งเดียวในทฤษฎีบทการมีอยู่) และผลต่างหารคือสัมประสิทธิ์ของ t^2 ในพหุนามนั้น ดังนั้นค่าจึงเท่ากันเสมอ ทฤษฎีบทการคำนวณซ้ำก็บอกเช่นเดียวกัน

ตรวจด้วยโค้ดใน `test_ch03.py`(ชุดทดสอบของแบบฝึกหัด 3.1.5) โค้ดยังทดสอบการสลับลำดับแบบสุ่มด้วยเมล็ดสุ่มสามค่า

เฉลย 3.1.6.

1. ตารางผลต่างหารของข้อมูลสี่จุด $(0,0), (1,3), (2,1), (3,4)$ คือ

   - อันดับ 1: $f[0,1] = 3$, $f[1,2] = -2$, $f[2,3] = \dfrac{4-1}{3-2} = 3$
   - อันดับ 2: $f[0,1,2] = -\frac{5}{2}$, $f[1,2,3] = \dfrac{3-(-2)}{3-1} = \frac{5}{2}$
   - อันดับ 3: $f[0,1,2,3] = \dfrac{f[1,2,3] - f[0,1,2]}{3-0} = \dfrac{\frac{5}{2} + \frac{5}{2}}{3} = \frac{5}{3}$

2. รูปแบบของนิวตันคือ

$$p_3(t) = 0 + 3t - \tfrac{5}{2}\,t(t-1) + \tfrac{5}{3}\,t(t-1)(t-2).$$

   พจน์สามพจน์แรกคือ 0, $3t$ และ $-\frac{5}{2}t(t-1)$ ซึ่งเป็นพจน์เดิมของ p_2 ไม่เปลี่ยน เพิ่มมาเพียงพจน์ $\frac{5}{3}t(t-1)(t-2)$ เมื่อกระจายพหุนามจะได้ $p_3(t) = \frac{5}{3}t^3 - \frac{15}{2}t^2 + \frac{53}{6}t$ ตรวจว่า $p_3(1) = 3$, $p_3(2) = 1$ และ $p_3(3) = 4$

3. ค่าที่ $t = 0.5, 1.5, 2.5$ คือ $p_3 = 2.75, 2.00$ และ 1.25 ส่วน p_2 (พหุนามดีกรีสองเดิม) ให้ 2.125, 2.625 และ -1.875 ความต่างคือ $+0.625$, -0.625 และ $+3.125$ ตามลำดับ ทั้งสองพหุนามผ่านคีย์สามตัวแรกเหมือนกัน แต่ต่างกันเมื่อเพิ่มคีย์ที่สี่

4. ในรูปแบบลากรานจ์ พหุนาม ℓ_k ทุกตัวมีดีกรีเพิ่มขึ้นและตัวหารเปลี่ยน จึงต้องคำนวณใหม่ทั้งหมด ส่วนรูปแบบของนิวตันเพิ่มเพียงพจน์เดียว

ตรวจด้วยโค้ดใน `test_ch03.py`(ชุดทดสอบของแบบฝึกหัด 3.1.6).

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 3.1.7.

1. เมทริกซ์แวนเดอร์มอนด์คือ $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ แถวที่สองและแถวที่สามเหมือนกันทุกประการ จึงมีดีเทอร์มิแนนต์เป็นศูนย์

2. ถ้ามีพหุนาม p ที่ผ่านคีย์ทั้งสาม จะต้องได้ $p(1) = 3$ และ $p(1) = 1$ พร้อมกัน ซึ่งเป็นไปไม่ได้ เพราะ $p(1)$ เป็นตัวเลขเพียงตัวเดียว ข้อนี้ไม่ขึ้นกับดีกรีของพหุนาม

3. ในแอนิเมชัน หมายความว่าตัวละครต้องอยู่ที่สองตำแหน่งในเวลาเดียวกัน โปรแกรมจึงไม่มีคำตอบให้เลือก ทฤษฎีบทในหัวข้อนี้จึงต้องการให้โหนดต่างกันทั้งหมด เพื่อให้มีคำตอบเดียวเสมอ ตัวแก้ระบบในโค้ดจะหยุดพร้อมข้อผิดพลาดว่าเมทริกซ์เอกฐานเมื่อเรียกใช้กับข้อมูลนี้

ตรวจด้วยโค้ดใน `test_ch03.py`(ชุดทดสอบของแบบฝึกหัด 3.1.7).

3.2 ผิดไปเท่าไร และทำไมคีย์เฟรมมากไม่ได้แปลว่าดีกว่า

สถานการณ์สมมติ: เพิ่มคีย์เฟรมแล้วตัวละครไถลไปมา

สมมติว่าผู้ออกแบบต้องการให้ตัวละครเคลื่อนไปตามเส้นโค้งรูปกระดิ่งที่เรียบ เริ่มไกลจากจุดกลาง เข้ามาใกล้ แล้วถอยออกไป ช่วงเวลาที่สนใจคือ -5 ถึง 5 วินาที และตำแหน่งที่ต้องการคือ $F(t) = \dfrac{1}{1+t^2}$ เมตร ซึ่งสูงสุดที่ 1 เมตรตอน $t = 0$ และเหลือราว 0.04 เมตรที่ปลายทั้งสองข้าง (ฟังก์ชันนี้เราเลือกเองเพื่อใช้ในบทนี้ ไม่ได้มาจากเกมใด)
ถ้าเก็บคีย์เฟรม 5 คีย์ที่ห่างเท่า ๆ กัน ตัวละครยังเคลื่อนไปใกล้เส้นโค้งจริง แต่ถอยไปต่ำกว่าศูนย์ราว 0.38 เมตร เมื่อเก็บ 11 คีย์ ที่ $t = -4.7$ ตัวละครควรอยู่ที่ 0.04 เมตร แต่กลับไปอยู่ที่ราว 1.96 เมตร และเมื่อเก็บ 21 คีย์ ที่ $t \approx 4.9$ ตัวละครไปอยู่ที่ราว -60 เมตร ทั้งที่คีย์ทุกอันถูกต้องพอดี คำถามคือ ทำไมคีย์มากขึ้นจึงแย่ลง และเราจะวัดความคลาดเคลื่อนระหว่างคีย์ได้อย่างไร

ภาพรวมก่อนอ่าน

เราอยากรู้ว่าเส้นที่สร้างจากคีย์คลาดจากการเคลื่อนที่จริงเท่าไร คำถามนี้มีสองส่วน ส่วนแรกคือการเคลื่อนที่จริงเปลี่ยนแรงแค่ไหน ส่วนที่สองคือคีย์วางอยู่ตรงไหน ลองนึกถึงเชือกที่ผูกกับหมุดหลายตัว เชือกผ่านทุกหมุดพอดี แต่ระหว่างหมุดเชือกอาจโค้งออกไปได้มาก

ฉบับร่าง · Preprint · 2026-10-23

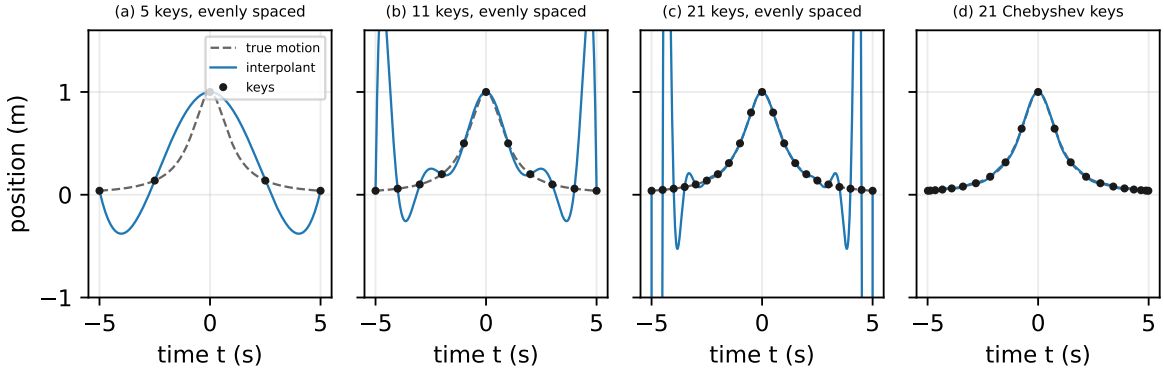
ค่าคลาดเคลื่อนอยู่ตรงไหน

ภาพแรกแสดงการประมาณค่าสี่แบบ เส้นประคือการเคลื่อนที่จริง จุดดำคือคีย์ และเส้นสีฟ้าคือพหุนามที่ผ่านทุกคีย์ ที่คีย์สองเส้นทับกันพอดี ระหว่างคีย์ เส้นสีฟ้าแยกออกจากเส้นประ และการแยกนี้รุนแรงขึ้นมากที่ปลายช่วงเมื่อคีย์ที่ห่างเท่ากันมีมากขึ้น

ค่าคลาดเคลื่อนที่เราต้องการเรียกว่า **ค่าคลาดเคลื่อนการประมาณค่า** (interpolation error) คือค่าจริงลบด้วยค่าที่พหุนาม p ให้ที่เวลา t ค่านี้เป็นศูนย์ที่คีย์ทุกอันเสมอ การดูเฉพาะที่คีย์จึงไม่บอกอะไร ตารางด้านล่างวัดค่าคลาดเคลื่อนบนตาราง 4001 จุดในช่วง $[-5, 5]$ คอลัมน์ที่สามคือเวลาที่ค่าคลาดเคลื่อนสูงสุด

ลองเอง: ทายก่อนดูตาราง

จากคีย์ห้าตัวไปเป็นสิบเอ็ดตัว ค่าคลาดเคลื่อนสูงสุดจะเพิ่มหรือลด ทายว่ามันเกิดที่กลางช่วงหรือที่ปลายช่วง แล้วค่อยเทียบกับตาราง



รูปที่ 3.5: การประมาณค่าด้วยคีย์ที่ห่างเท่ากัน 5, 11 และ 21 คีย์ และด้วยคีย์เชบีเชฟ 21 คีย์ เส้นประคือการเคลื่อนที่จริง เส้นทึบคือพหุนามที่ผ่านคีย์ ส่วนที่หลุดกรอบบนล่างถูกตัดออก (สร้างจาก `numanal/ch03.py`)

การวางคีย์	ค่าคลาดเคลื่อนสูงสุด (เมตร)	ที่เวลา t (วินาที)	ค่าสูงสุดของพหุนาม (เมตร)	ค่าต่ำสุดของพหุนาม (เมตร)
5 คีย์ ห่างเท่ากัน	0.438	3.975	1.000	−0.379
11 คีย์ ห่างเท่ากัน	1.916	−4.700	1.959	−0.257
21 คีย์ ห่างเท่ากัน	59.82	4.875	4.318	−59.78
21 คีย์ เชบีเชฟ	0.0153	−1.1075	1.000	0.0333

คำนวณด้วยมือกับสามคีย์

ลองทำตัวอย่างที่คำนวณได้ครบด้วยมือ ให้ $f(t) = t^3$ และโหนด $-1, 0, 1$ ค่าของ f ที่โหนดคือ $-1, 0, 1$ ซึ่งตรงกับเส้น t พอดี ดีกรีไม่เกิน 2 จึงได้พหุนามประมาณค่า $p(t) = t$ ที่ $t = \frac{1}{2}$ ค่าจริงคือ $\frac{1}{8}$ ส่วนพหุนามให้ $\frac{1}{2}$ ดังนั้นค่าคลาดเคลื่อนคือ $-\frac{3}{8}$

ค่านี้มาจากไหน ตัวประกอบ $\omega(t) = (t+1)\,t\,(t-1)$ เป็นศูนย์ที่โหนดทั้งสาม ที่ $t = \frac{1}{2}$ ได้ $\omega(\frac{1}{2}) = \frac{3}{2}\cdot\frac{1}{2}\cdot(-\frac{1}{2}) = -\frac{3}{8}$ ซึ่งตรงกับค่าคลาดเคลื่อนพอดี และ $f'''(t)/3! = 1$ เป็นค่าคงที่ ทฤษฎีบทต่อไปนี้บอกว่าโครงสร้างนี้มีเสมอ แม้ f จะไม่ใช่พหุนาม

ทฤษฎีบท 3.2.1: สูตรค่าคลาดเคลื่อนการประมาณค่า

ให้ f มีอนุพันธ์ต่อเนื่องถึงอันดับ $n+1$ บนช่วงปิด $[a, b]$ ให้โหนด $t_0, \ldots, t_n$ ต่างกันทั้งหมดและอยู่ใน $[a, b]$ และให้ p เป็นพหุนามประมาณค่าดีกรีไม่เกิน n ของ f ที่โหนดเหล่านี้ เมื่อ $t \in [a, b]$ จะมี ξ ซึ่งอยู่ในช่วงปิดที่เล็กที่สุดที่มีทั้ง t และโหนดทั้งหมด โดยที่

$$f(t) - p(t) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega(t), \qquad \omega(t) = \prod_{j=0}^{n}(t - t_j).$$

อ่านแบบเข้าใจง่าย

- **พหุนามโหนด** (node polynomial) $\omega(t)$ ขึ้นกับตำแหน่งของโหนดเท่านั้น ไม่ขึ้นกับค่า y_j ของข้อมูลเลย ถ้าเปลี่ยนค่าข้อมูลแต่คงโหนดไว้ ω ก็ยังเท่าเดิม
- ที่โหนดแต่ละตัว $\omega = 0$ จึงได้ค่าคลาดเคลื่อนเป็นศูนย์ตามที่เห็นในหัวข้อนี้
- ξ เป็นจุดที่เราไม่รู้ค่า ทฤษฎีบทรับประกันเพียงว่ามันอยู่ในช่วงที่กำหนด ซึ่งพอสำหรับการประเมินขนาดของค่า

ฉบับร่าง · Preprint · 2026-10-23

คลาดเคลื่อน

- เมื่อ $n = 0$ สูตรนี้คือ $f(t) - f(t_0) = f'(\xi)(t - t_0)$ ซึ่งเป็นทฤษฎีบทค่ามัชฌิมที่คุ้นเคยจากแคลคูลัส

แนวคิด. ดูฟังก์ชัน g ที่หักค่าคลาดเคลื่อนด้วย $K\omega$ ออกแล้วมีศูนย์ $n+2$ จุด จากนั้นทฤษฎีบทของโรลซ้ำ ๆ ทำให้อนุพันธ์อันดับ $n+1$ มีศูนย์

1. **กรณี t เป็นโหนด.** ทั้งสองข้างของสูตรเป็นศูนย์ เพราะ $f(t) = p(t)$ และ $\omega(t) = 0$ จึงเลือก ξ ในช่วงปิดที่มีโหนดทั้งหมดใดก็ได้ (ถ้า $n = 0$ ให้เลือก $\xi = t_0$)

2. **กรณี t ไม่ใช่โหนด.** ให้ I เป็นช่วงปิดที่เล็กที่สุดที่มี t และโหนดทั้งหมด เนื่องจาก $\omega(t) \neq 0$ กำหนด $K = \dfrac{f(t) - p(t)}{\omega(t)}$ และนิยาม $g(s) = f(s) - p(s) - K\,\omega(s)$ สำหรับ $s \in I$

3. **g มีศูนย์มากพอ.** ที่โหนดแต่ละตัว $f - p = 0$ และ $\omega = 0$ จึง $g(t_j) = 0$ ส่วน $g(t) = 0$ โดยนิยามของ K ดังนั้น g มีศูนย์ $n+2$ จุดที่ต่างกันใน I

4. **ใช้ทฤษฎีบทของโรลซ้ำ.** ทฤษฎีบทของโรลบอกว่า ถ้าฟังก์ชันเรียบบนช่วงและมีค่าเท่ากันที่ปลายทั้งสอง จะมีจุดในช่วงเปิดที่อนุพันธ์เป็นศูนย์ เมื่อนำไปใช้กับศูนย์ติดกันทุกคู่ของ g จะได้ g' มีศูนย์ $n+1$ จุดที่ต่างกันในส่วนในของ I ทำซ้ำจนได้ว่า $g^{(n+1)}$ มีศูนย์ ξ อย่างน้อยหนึ่งจุดในส่วนในของ I

5. **คำนวณ $g^{(n+1)}$.** p มีดีกรีไม่เกิน n จึง $p^{(n+1)} = 0$ และ ω เป็นพหุนามดีกรี $n+1$ ที่สัมประสิทธิ์ของ s^{n+1} เป็น 1 จึง $\omega^{(n+1)} = (n+1)!$ ดังนั้น $g^{(n+1)}(s) = f^{(n+1)}(s) - K\,(n+1)!$

6. **สรุป.** จาก $g^{(n+1)}(\xi) = 0$ ได้ $K = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}$ แทนนิยามของ K จะได้สูตรที่ต้องการ

สรุป: ค่าคลาดเคลื่อนเท่ากับอนุพันธ์อันดับ $n+1$ ที่จุดบางจุด หารด้วย $(n+1)!$ คูณกับ $\omega(t)$ ซึ่งขึ้นกับโหนดเท่านั้น

ทฤษฎีบท 3.2.2: ขอบเขตของค่าคลาดเคลื่อน

ภายใต้สมมติฐานของทฤษฎีบทข้างบนบนช่วง $[a, b]$ ให้ $M = \max_{t\in[a,b]} |f^{(n+1)}(t)|$ จะได้

$$\max_{t\in[a,b]} |f(t) - p(t)| \leq \frac{M}{(n+1)!} \max_{t\in[a,b]} |\omega(t)|.$$

ฉบับร่าง · Preprint · 2026-10-23

เหตุผลคือ ξ อยู่ใน $[a, b]$ จึง $|f^{(n+1)}(\xi)| \leq M$ แทนลงในสูตรแล้วหาค่าสูงสุดของทั้งสองข้าง ขอบเขตนี้แยกส่วนที่ขึ้นกับการเคลื่อนที่ (M) ออกจากส่วนที่ขึ้นกับการวางคีย์ $(\max|\omega|)$ ตอนนี้เราดูส่วนหลังก่อน

คีย์ควรวางที่ไหน

ความต่างระหว่างการวางคีย์สองแบบอยู่ที่ ω ใกล้ปลายช่วง เมื่อคีย์ห่างเท่ากัน ปลายช่วงมีคีย์อยู่ด้านเดียว จึงมีตัวประกอบ $(t - t_j)$ ที่ใหญ่หลายตัวพร้อมกัน ลองดูกรณีที่คำนวณได้ด้วยมือ คือ $n = 2$ บน $[-1, 1]$

- คีย์ห่างเท่ากัน $-1, 0, 1$: $\omega(t) = t^3 - t$ จุดวิกฤตคือ $t = \pm\frac{1}{\sqrt{3}}$ โดย $\omega(\frac{1}{\sqrt{3}}) = -\frac{2}{3\sqrt{3}} \approx -0.385$ และ $\omega(\pm 1) = 0$ จึงได้ $\max|\omega| = \frac{2}{3\sqrt{3}} \approx 0.385$

- คีย์เชบีเชฟ $-\frac{\sqrt{3}}{2}, 0, \frac{\sqrt{3}}{2}$: $\omega(t) = t^3 - \frac{3}{4}t$ จุดวิกฤตคือ $t = \pm\frac{1}{2}$ โดย $\omega(\frac{1}{2}) = -\frac{1}{4}$ และที่ปลาย $\omega(\pm 1) = \frac{1}{4}$ จึงได้ $\max|\omega| = \frac{1}{4}$

โหนดเชบีเชฟ (**โหนดเชบีเชฟ** (Chebyshev nodes)) คือโหนดที่เลือกจากมุมที่ห่างเท่ากันแล้วฉายลงบนแกนนอน จุดจึงกระจุกตัวใกล้ปลายช่วงมากกว่าตรงกลาง โหนดเหล่านี้คือศูนย์ของ**พหุนามเชบีเชฟ** (Chebyshev polynomial) $T_m(x) = \cos(m \arccos x)$ Iserles 2009, §10.5, p. 222 ซึ่งมีค่าอยู่ระหว่าง -1 ถึง 1 บนช่วง $[-1, 1]$

ทฤษฎีบท 3.2.3: โหนดเชบีเชฟและขนาดของ ω

ให้ $[a, b]$ และ $n \geq 0$ กำหนด

$$t_k = \frac{a+b}{2} + \frac{b-a}{2} \cos\frac{(2k+1)\pi}{2(n+1)}, \qquad k = 0, 1, \ldots, n.$$

โหนดเหล่านี้ต่างกันทั้งหมด และ

$$\max_{t\in[a,b]} |\omega(t)| = \left(\frac{b-a}{2}\right)^{n+1} 2^{-n}.$$

แนวคิด. โหนดเชบีเชฟคือศูนย์ของ T_{n+1} และ $2^{-n}T_{n+1}$ เป็นพหุนามที่กำลังสูงสุดเป็น 1 ซึ่งมีขนาดสูงสุดบน $[-1, 1]$ เท่ากับ 2^{-n}

1. **สมบัติของ T_m.** จากเอกลักษณ์ $\cos((m+1)\theta) + \cos((m-1)\theta) = 2\cos\theta\cos m\theta$ ได้สูตรเวียน $T_{m+1} = 2xT_m - T_{m-1}$ ดังนั้น T_m เป็นพหุนามดีกรี m และสัมประสิทธิ์ของ x^m คือ 2^{m-1} เมื่อ $m \geq 1$ (เริ่มจาก $T_1 = x$ แล้วเหนี่ยวนำ)

2. **ศูนย์ของ T_{n+1}.** $T_{n+1}(x) = 0$ เมื่อ $(n+1)\arccos x = \frac{(2k+1)\pi}{2}$ นั่นคือ $x = \cos\frac{(2k+1)\pi}{2(n+1)}$ เมื่อ $k = 0, \ldots, n$ มุมเหล่านี้อยู่ใน $(0, \pi)$ และต่างกัน จึงได้ศูนย์ $n+1$ จุดที่ต่างกัน เท่ากับดีกรี ไม่มีศูนย์อื่น

3. **รูปของ ω.** T_{n+1} มีศูนย์ต่างกัน $n+1$ จุดซึ่งคือโหนด และกำลังสูงสุดเป็น 2^n จึงได้ $T_{n+1}(x) = 2^n \prod_k (x - x_k)$ ดังนั้น $\omega(x) = 2^{-n}T_{n+1}(x)$ บน $[-1, 1]$

4. **ขนาดสูงสุด.** $|T_{n+1}(x)| \leq 1$ บน $[-1, 1]$ จึง $\max|\omega| \leq 2^{-n}$ และที่จุด $x = \cos\frac{k\pi}{n+1}$ เมื่อ $k = 0, \ldots, n+1$ ได้ $T_{n+1}(x) = (-1)^k$ จึงได้ค่าสูงสุดเท่ากับ 2^{-n} พอดี

5. **เปลี่ยนช่วง.** เขียน $t = \frac{a+b}{2} + rx$ เมื่อ $r = \frac{b-a}{2}$ จะได้ $\omega(t) = r^{n+1}\omega_{[-1,1]}(x)$ จึงได้ค่าสูงสุด $r^{n+1}2^{-n}$ บน $[a, b]$

สรุป: การวางคีย์เป็นศูนย์ของพหุนามเชบีเชฟทำให้ค่าสูงสุดของ $|\omega|$ คำนวณได้แน่นอน และเล็กกว่าการวางคีย์ห่างเท่ากันมากเมื่อ n ใหญ่

ลองเอง: โหนดเชบีเชฟบนช่วงตั้งแต่ 0 ถึง 4

บนช่วง $[0, 4]$ โหนดเชบีเชฟสามตัวคือ $2 - \sqrt{3}$, 2 และ $2 + \sqrt{3}$ ให้ $x = t - 2$ แล้วเขียน $\omega(t)$ เป็นพหุนามของ x จากนั้นหาค่าสูงสุดของ $|\omega|$ บน $[0, 4]$ โดยดูจุดวิกฤตและจุดปลาย

คำตอบ: $\omega = x^3 - 3x$ สำหรับ $x \in [-2, 2]$ ค่าสูงสุดของ $|\omega|$ คือ 2 ที่ $x = \pm 1$ และที่ปลาย $x = \pm 2$ ซึ่งตรงกับทฤษฎีบท เพราะ $r = 2$ และ $n = 2$ ได้ $2^3 \cdot 2^{-2} = 2$

ฉบับร่าง Preprint 2026-10-23

ทำไมจึงเรียกว่าเหมาะที่สุด คำตอบอยู่ในข้อความต่อไปนี้ สำหรับพหุนาม q ดีกรี $n+1$ ใด ๆ ที่สัมประสิทธิ์ของ t^{n+1} เป็น 1 บน $[-1,1]$ จะได้ $\max|q| \geq 2^{-n}$ เสมอ และ $2^{-n}T_{n+1}$ ทำได้เท่ากับค่านี้ เนื่องจาก ω ของโหนดใดก็ตามเป็นพหุนามแบบนี้ จึงไม่มีการวางโหนดใดให้ $\max|\omega|$ เล็กกว่าเชบีเชฟ

แนวคิด. ถ้า q เล็กกว่า 2^{-n} ทุกจุด ผลต่าง $2^{-n}T_{n+1} - q$ จะสลับเครื่องหมายที่จุดสุดขั้วของ T_{n+1} จนมีศูนย์มากกว่าดีกรี

1. **จุดสุดขั้ว.** ให้ $x_k = \cos\frac{k\pi}{n+1}$ สำหรับ $k = 0, \ldots, n+1$ คือจุดที่ T_{n+1} สลับค่าระหว่าง 1 กับ -1 จุดเหล่านี้เรียงลดลงจาก 1 ถึง -1 และ $T_{n+1}(x_k) = \cos k\pi = (-1)^k$

2. **ผลต่าง.** ให้ $D = 2^{-n}T_{n+1} - q$ ทั้งสองพจน์มีกำลัง t^{n+1} ที่สัมประสิทธิ์เป็น 1 จึงหักล้างกัน ทำให้ D มีดีกรีไม่เกิน n ที่ x_k ค่า $2^{-n}T_{n+1}(x_k) = 2^{-n}(-1)^k$ มีค่าสัมบูรณ์มากกว่า $|q(x_k)|$ ดังนั้น $D(x_k)$ มีเครื่องหมาย $(-1)^k$

3. **ศูนย์มากเกิน.** เครื่องหมายของ D สลับกันระหว่าง x_{k+1} กับ x_k ทุกคู่ ตามทฤษฎีบทค่ากลาง D มีศูนย์ในช่วง (x_{k+1}, x_k) อย่างน้อยหนึ่งจุด รวมเป็นศูนย์ $n+1$ จุดที่ต่างกัน

4. **สรุป.** D มีดีกรีไม่เกิน n แต่มีศูนย์ $n+1$ จุด จึง $D \equiv 0$ นั่นคือ $q = 2^{-n}T_{n+1}$ ซึ่งมีค่าสูงสุด 2^{-n} ขัดแย้งกับสมมติฐาน $\max|q| < 2^{-n}$

สรุป: ค่าสูงสุดของ $|\omega|$ บน $[-1,1]$ ไม่มีโหนดใดทำให้ต่ำกว่า 2^{-n} และโหนดเชบีเชฟทำได้เท่ากับค่านี้พอดี

การพิสูจน์ข้างบนเป็นของเราเอง ต้นฉบับ Atkinson กล่าวถึงชุดจุดแบบนี้ที่ให้การประมาณดีในหัวข้อ 4.7 Atkinson 1989, p. 159 (หัวข้อนั้นอยู่นอกฉบับนี้ ยังไม่ได้แปล)

ปรากฏการณ์รุงเง

ตัวอย่างของเราคือ $F(t) = \dfrac{1}{1+t^2}$ บน $[-5,5]$ ซึ่ง Atkinson ใช้เป็นตัวอย่างคลาสสิกเช่นกัน Atkinson 1989, §3.5, p. 158 ต้นฉบับเล่าว่า สำหรับ $3.64 < |x| < 5$ ค่าสูงสุดของ $|f(x) - p_n(x)|$ เมื่อ $n \geq k$ เป็นอนันต์สำหรับทุก k ดังนั้นพหุนามประมาณค่าบนคีย์ห่างเท่ากันไม่ลู่เข้าหาค่าจริงที่จุดเหล่านั้น ผลลัพธ์นี้เรียกว่า**ปรากฏการณ์รุงเง** (Runge phenomenon) ตามชื่อคาร์ล รุงเง ซึ่งงานของเขาในปี ค.ศ. 1901 เกี่ยวกับการประมาณค่าระหว่างลำดับค่าที่ห่างเท่ากัน Runge 1901

ตารางต่อไปนี้ขยายตารางในหัวข้อนี้ไปยังจำนวนคีย์ที่มากขึ้น ค่าทั้งหมดวัดบนตาราง 4001 จุดในช่วง $[-5,5]$

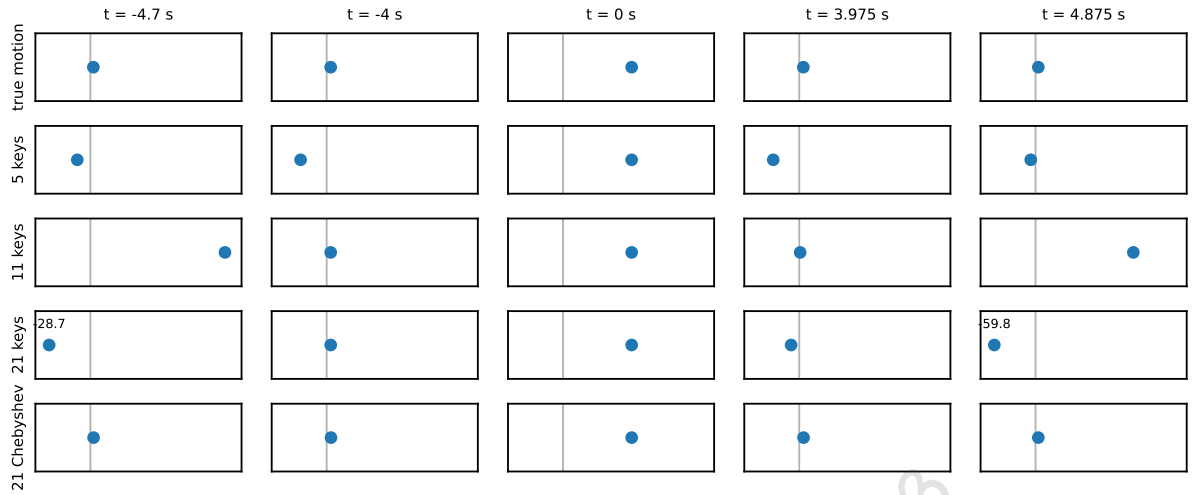
จำนวนคีย์	ห่างเท่ากัน (เมตร)	เชบีเชฟ (เมตร)
5	0.438	0.402
7	0.617	0.264
9	1.05	0.171
11	1.92	0.109
13	3.66	0.0692
17	14.4	0.0326
21	59.8	0.0153

ค่าคลาดเคลื่อนสูงสุดของคีย์ที่ห่างเท่ากันเพิ่มขึ้นทุกครั้งที่เพิ่มคีย์ ส่วนคีย์เชบีเชฟลดลงทุกครั้ง ตารางนี้เป็นตัวอย่างหนึ่งชุด จึงยังไม่พิสูจน์การลู่เข้าหรือการลู่ออก

รูปแถบเฟรมด้านล่างแสดงตำแหน่งของตัวละครที่ห้าเวลาที่ความคลาดเคลื่อนเด่น แถวบนคือการเคลื่อนที่จริง และแถวที่เหลือคือการประมาณค่าสี่แบบ ค่าที่หลุดจากกรอบเขียนไว้เหนือจุดที่ขอบ ที่ $t \approx -4.7$ คีย์ 21 ตัวที่ห่างเท่ากันให้ -28.7 เมตร ทั้งที่การเคลื่อนที่จริงอยู่ที่ราว 0.04 เมตร

ค่าคลาดเคลื่อนยังมีผลต่อความเร็วด้วย ที่ $t = -4.7$ การเคลื่อนที่จริงมีความเร็ว $+0.0176$ เมตรต่อวินาที คือเคลื่อนเข้าหาศูนย์ แต่ความเร็วของพหุนามคือ -0.595 เมตรต่อวินาทีสำหรับ 5 คีย์ และ -0.0114 เมตรต่อวินาทีสำหรับ 11 คีย์ ซึ่งชี้ไปทาง

ฉบับร่าง Preprint · 2026-10-23



รูปที่ 3.6: แถบเฟรมของตัวละครที่ $t = -4.7, -4.0, 0, 3.975$ และ 4.875 วินาที แถวบนคือการเคลื่อนที่จริง แถวถัดไปคือการประมาณค่าด้วยคีย์ห่างเท่ากัน 5, 11, 21 คีย์ และคีย์เชบีเชฟ 21 ตัว ค่าที่หลุดจากกรอบเขียนไว้เหนือจุด (สร้างจาก `numanal/ch03.py`)

ตรงข้าม ส่วน 21 คีย์ที่ห่างเท่ากันให้ 217.5 เมตรต่อวินาที (ตัวเลขนี้คำนวณจากการประมาณอนุพันธ์ด้วยผลต่างกลางในโค้ดของเรา) การประมาณค่าที่ตำแหน่งถูกต้องที่คีย์จึงไม่รับประกันความเร็วที่ถูกต้องระหว่างคีย์ หัวข้อ 3.3 จะกลับมาที่ความเร็วโดยตรง

เพื่อให้ได้ขอบเขตที่ชัดเจนสำหรับ F เราต้องรู้ขนาดของอนุพันธ์ของมัน

ทฤษฎีบท 3.2.4: อนุพันธ์ของ F มีขนาดไม่เกิน $k!$

สำหรับ $F(t) = \dfrac{1}{1+t^2}$ และจำนวนเต็ม $k \geq 0$ จะได้ $|F^{(k)}(t)| \leq k!$ สำหรับทุกจำนวนจริง t และเมื่อ k เป็นจำนวนคู่ ค่าเท่ากับ $k!$ ที่ $t = 0$

แนวคิด. แยกเศษส่วนย่อยให้เป็นผลต่างของ $\dfrac{1}{t \mp i}$ แล้วใช้สูตรอนุพันธ์ของ $(t-a)^{-1}$

1. **แยกเศษส่วนย่อย.** $F(t) = \dfrac{1}{2i}\left(\dfrac{1}{t-i} - \dfrac{1}{t+i}\right)$ เพราะ $\dfrac{1}{t-i} - \dfrac{1}{t+i} = \dfrac{2i}{t^2+1}$
2. **อนุพันธ์.** จาก $\dfrac{d^k}{dt^k}(t-a)^{-1} = (-1)^k k!\,(t-a)^{-(k+1)}$ ได้ $F^{(k)}(t) = \dfrac{(-1)^k k!}{2i}\Big((t-i)^{-(k+1)} - (t+i)^{-(k+1)}\Big)$
3. **ขอบเขต.** $|t \pm i| = \sqrt{1+t^2} \geq 1$ จึง $|F^{(k)}(t)| \leq \dfrac{k!}{2}\big(|t-i|^{-(k+1)} + |t+i|^{-(k+1)}\big) \leq k!$
4. **จุดที่เท่ากัน.** ที่ $t = 0$ จะได้ $(-i)^{-(k+1)} - i^{-(k+1)} = i^{k+1} - (-i)^{k+1}$ ซึ่งเท่ากับ $2i^{k+1}$ เมื่อ k คู่ (เพราะ $k+1$ เป็นเลขคี่) จึง $|F^{(k)}(0)| = k!$

สรุป: อนุพันธ์อันดับใดก็ตามของ F มีขนาดไม่เกิน $k!$ ทุกจุด และถึงค่านี้ที่ $t = 0$ เมื่อ k คู่

ฉบับร่าง · Preprint · 2026-10-23

ทฤษฎีบท 3.2.5: ขอบเขตค่าคลาดเคลื่อนสำหรับ F

ถ้า p เป็นพหุนามประมาณค่าดีกรีไม่เกิน n ของ F ที่โหนดต่างกัน $t_0, \dots, t_n$ บน $[a, b]$ แล้ว $\max_{t\in[a,b]} |F(t) - p(t)| \le \max_{t\in[a,b]} |\omega(t)|$

เหตุผลคือ จากทฤษฎีบทค่าคลาดเคลื่อนกับผลข้างบน $|F^{(n+1)}(\xi)|/(n+1)! \le 1$ สำหรับทุก ξ จึงได้ผลนี้ทันที

ข้อควรระวัง: ทฤษฎีบทต้องการความเรียบ

สูตรค่าคลาดเคลื่อนใช้ได้เมื่อ f มีอนุพันธ์อันดับ $n+1$ ต่อเนื่องทั้งช่วง ถ้าการเคลื่อนที่มีมุมหัก เช่น $|t|$ บน $[-1, 1]$ ทฤษฎีบทนี้ใช้ไม่ได้ เพราะ $|t|$ ไม่มีอนุพันธ์ที่ $t = 0$ แม้จะใช้โหนดเชบีเชฟ ค่าคลาดเคลื่อนสูงสุดก็ลดลงช้า คือ 0.123, 0.067, 0.035, 0.018 และ 0.0092 เมื่อ n (จำนวนช่วงระหว่างคีย์) เป็น 4, 8, 16, 32 และ 64 ซึ่งลดลงราวครึ่งหนึ่งทุกครั้งที่ n เป็นสองเท่า (ตัวเลขจากโค้ดของเรา) ส่วน F ในหัวข้อนี้ลดลงเร็วกว่านี้มากเมื่อเพิ่มคีย์

ข้อควรระวัง: ตรวจเฉพาะที่คีย์ได้ศูนย์เสมอ

ถ้าเทียบค่าที่คีย์กับค่าจริง ค่าคลาดเคลื่อนจะเป็นศูนย์ไม่ว่าการวางคีย์จะดีหรือแย่ ที่คีย์ทั้ง 21 ตัวของการวางแบบห่างเท่ากัน ค่าคลาดเคลื่อนเป็นศูนย์ ขณะที่ระหว่างคีย์สูงสุดถึง 59.8 เมตร โค้ดแสดงเรื่องนี้ในสองบรรทัด

```
err = interpolation_error(nodes, nodes)   # checks only at the keys
print(np.max(np.abs(err)))                # prints 0.0
```

ข้อควรระวัง: ขอบเขตอาจเกินจริงมาก และสูตรแวนเดอร์มอนด์อาจเพี้ยน

ขอบเขตนี้แทน $|f^{(n+1)}(\xi)|$ ด้วยค่าสูงสุดของมันบนช่วงทั้งช่วง จึงอาจหลวมมาก สำหรับคีย์เชบีเชฟ 21 ตัว ขอบเขต $\max|\omega|$ มีค่าราว 4.5×10^8 เมตร ทั้งที่ค่าคลาดเคลื่อนจริงสูงสุดราว 0.015 เมตร ขอบเขตนี้จึงใช้เปรียบเทียบการวางคีย์ได้ แต่ทำนายค่าจริงไม่ได้

อีกเรื่องคือการแก้ระบบ $Va = y$ ด้วยกำลัง $1, t, t^2, \dots$ เมื่อ 21 คีย์ห่างเท่ากันบน $[-5, 5]$ เมทริกซ์มี $\mathrm{cond}(V) \approx 3.1 \times 10^{15}$ ซึ่งอาจทำให้เสียความแม่นไปได้ถึงราว 15 หลัก จากที่เลขทศนิยมลอยตัวแบบ 64 บิตเก็บได้ราว 16 หลัก ผลลัพธ์ต่างจากรูปแบบบารีเซนทริกราว 1.2×10^{-6} เมตรบนตาราง 2001 จุด และที่คีย์เองยังมีค่าคลาดเคลื่อนราว 5×10^{-10} เมตร จึงควรใช้รูปแบบบารีเซนทริกที่ใช้น้ำหนักชุดเดียว

```
coeffs = monomial_coefficients(nodes, values)   # solves V a = y
p = np.polyval(coeffs[::-1], xs)                 # the risky route
```

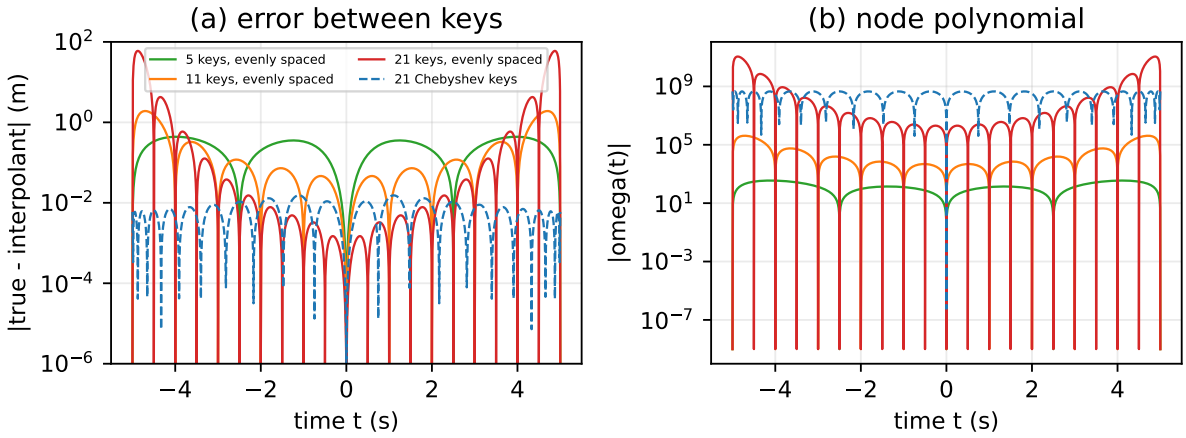
เลือกจำนวนคีย์ให้ได้ค่าคลาดเคลื่อนที่กำหนด

ขอบเขตที่เราเห็นข้างบนหลวมมาก แต่ก็ใช้เลือกจำนวนคีย์ได้เมื่อเรารู้ว่าต้องการค่าคลาดเคลื่อนไม่เกินเท่าไร ลองดูการแกว่งแบบนุ่ม ๆ ที่หนึ่งวินาที คือ $f(t) = \sin\frac{\pi t}{2}$ บน $[0, 1]$ ตัวละครเริ่มที่ 0 เมตร และไปถึง 1 เมตรที่ $t = 1$

ในกรณีนี้ $M = \max_{[0,1]} |f^{(n+1)}| = (\frac{\pi}{2})^{n+1}$ เพราะอนุพันธ์อันดับ k ของ f คือ $(\frac{\pi}{2})^k \sin(\frac{\pi t}{2} + \frac{k\pi}{2})$ และค่าสัมบูรณ์ของมันถึง 1 ที่ปลายช่วงข้างหนึ่งเสมอ จึงได้ขอบเขต

$$\max_{[0,1]} |f - p| \le \frac{(\pi/2)^{n+1}}{(n+1)!} \max_{[0,1]} |\omega|.$$

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 3.7: (ก) ค่าคลาดเคลื่อนระหว่างคีย์บนลอการิทึมฐาน 10 จุดที่คีย์ค่าคลาดเคลื่อนเป็นศูนย์ และค่าที่ต่ำกว่าขอบล่างถูกวาดที่ขอบ (ข) พหุนามโหนด ω บนช่วงเดียวกัน ค่าที่ต่ำกว่าขอบล่างถูกวาดที่ขอบเช่นกัน (สร้างจาก `numanal/ch03.py`)

สำหรับคีย์เชบีเชฟบน $[0, 1]$ ค่าสูงสุดของ $|\omega|$ คือ $2^{-(2n+1)}$ ตามทฤษฎีบทในหัวข้อนี้ ส่วนคีย์ห่างเท่ากันเราหาค่าสูงสุดของ $|\omega|$ ด้วยโค้ด ตารางต่อไปนี้เปรียบเทียบขอบเขตกับค่าที่วัดได้จริง (ตัวเลขจากโค้ดของเรา ดูรูปที่ 3.8 ประกอบ)

n (จำนวนคีย์ $n+1$)	ขอบเขต (ห่างเท่ากัน)	วัดได้ (ห่างเท่ากัน)	ขอบเขต (เชบีเชฟ)	วัดได้ (เชบีเชฟ)
3 (4 คีย์)	3.13×10^{-3}	2.39×10^{-3}	1.98×10^{-3}	1.56×10^{-3}
4 (5 คีย์)	2.83×10^{-4}	2.15×10^{-4}	1.56×10^{-4}	1.21×10^{-4}
5 (6 คีย์)	2.26×10^{-5}	1.71×10^{-5}	1.02×10^{-5}	7.80×10^{-6}
6 (7 คีย์)	1.60×10^{-6}	1.21×10^{-6}	5.72×10^{-7}	4.34×10^{-7}

จากตาราง ถ้าต้องการค่าคลาดเคลื่อนไม่เกิน 10^{-4} เมตร คีย์ห้าตัว ($n = 4$) ยังไม่พอทั้งสองแบบ เพราะค่าที่วัดได้ยังเกิน 10^{-4} ต้องใช้หกคีย์ ($n = 5$) ซึ่งทั้งขอบเขตและค่าที่วัดได้ต่ำกว่า 10^{-4} และขอบเขตของเชบีเชฟเล็กกว่าของคีย์ห่างเท่ากันอยู่ราว 1.6 ถึง 2.8 เท่าในทุกแถว

ลองเอง: ค่าคลาดเคลื่อนไม่เกิน 10^{-5} เมตร

ต้องใช้คีย์กี่ตัว ทายก่อนว่าขอบเขตของ $n = 5$ จะผ่านเกณฑ์นี้หรือไม่ แล้วดูค่าที่ $n = 6$ จากตาราง

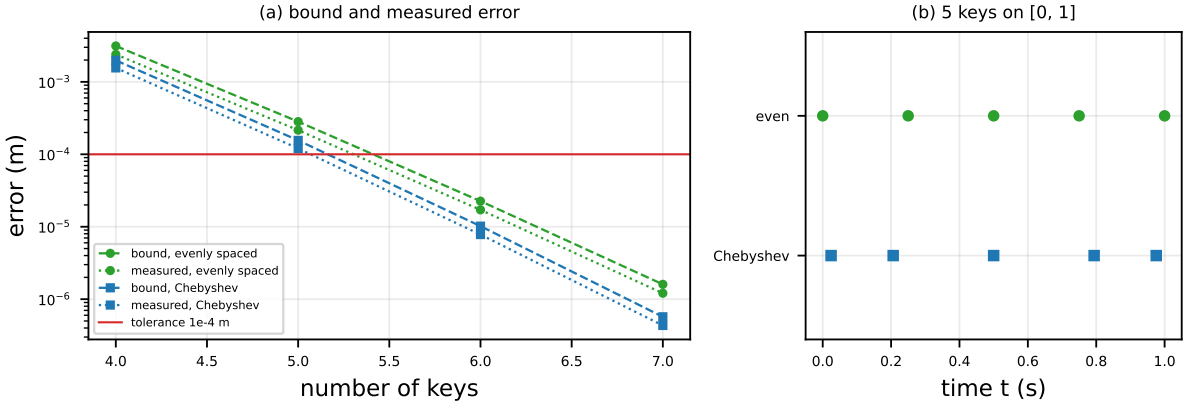
คำตอบคือ 7 คีย์ ทั้งสองแบบ ที่ $n = 5$ ขอบเขตของเชบีเชฟคือ 1.02×10^{-5} ซึ่งเกินเกณฑ์ไปนิดเดียว ส่วนที่ $n = 6$ ขอบเขตทั้งสองแบบต่ำกว่า 10^{-5} ตัวเลขที่ใกล้เกณฑ์แบบนี้ทำให้ความแม่นของ $\max |\omega|$ มีผลต่อคำตอบ ดูข้อควรระวังด้านล่าง

ข้อควรระวัง: ตรวจ $\max |\omega|$ บนตารางที่ตรงกับคีย์

ขอบเขตต้องใช้ค่าสูงสุดของ $|\omega|$ จริง ถ้าเราตรวจบนตารางจุดที่ตรงกับคีย์พอดี ค่าที่ได้จะผิด เช่น คีย์ 11 ตัวห่างเท่ากันบน $[0, 1]$ มี $\omega = 0$ ที่ทุกจุดของตาราง 11 จุด เพราะจุดตารางกับคีย์เป็นจุดเดียวกัน ค่าที่ได้คือ 0 แทนที่จะเป็น 4.17×10^{-6} ส่วนตาราง 21 จุดให้ 3.20×10^{-6} ต่ำกว่าค่าจริงราว 23% ควรตรวจบนตารางที่ละเอียดกว่า ดังโค้ดนี้

```
keys = np.linspace(0.0, 1.0, 11)          # 11 evenly spaced keys
grid = np.linspace(0.0, 1.0, 11)          # the check grid, the same points
print(np.max(np.abs(node_polynomial(keys, grid))))   # prints 0.0
```

สมบัติ · Preprint 2026-10-23



รูปที่ 3.8: (ก) ขอบเขต (เส้นประ) และค่าคลาดเคลื่อนที่วัดได้ (เส้นจุด) ของ $\sin \frac{\pi t}{2}$ บน $[0, 1]$ เมื่อใช้คีย์ห่างเท่ากัน (เขียว) และคีย์เชบีเชฟ (น้ำเงิน) เส้นแดงคือ 10^{-4} เมตร (ข) คีย์ห้าตัวแบบห่างเท่ากันและแบบเชบีเชฟบน $[0, 1]$ (สร้างจาก `numanal/ch03.py`)

สรุปใจความ

- ค่าคลาดเคลื่อนการประมาณค่าเท่ากับ $\dfrac{f^{(n+1)}(\xi)}{(n+1)!}\omega(t)$ สำหรับ ξ บางจุดในช่วง ส่วนแรกขึ้นกับความแรงของการเคลื่อนที่จริง ส่วนที่สองขึ้นกับตำแหน่งคีย์เท่านั้น
- ที่คีย์ค่าคลาดเคลื่อนเป็นศูนย์เสมอ จึงต้องตรวจระหว่างคีย์ทุกครั้ง
- คีย์ที่ห่างเท่ากันทำให้ $\max|\omega|$ ใหญ่มากที่ปลายช่วง โหนดเชบีเชฟให้ค่าสูงสุดเท่ากับ $(\frac{b-a}{2})^{n+1}2^{-n}$ ซึ่งเล็กกว่ามากเมื่อ n ใหญ่
- คีย์มากขึ้นไม่ได้ทำให้ผลดีขึ้นเสมอไป ความคลาดเคลื่อนสูงสุดของคีย์ห่างเท่ากันเพิ่มจาก 5 คีย์ไปเป็น 11 และ 21 คีย์ ส่วนคีย์เชบีเชฟ 21 ตัวให้ค่าเพียง 0.0153 เมตร
- ทฤษฎีบทต้องการความเรียบของการเคลื่อนที่ เมื่อมีมุมหัก เช่น $|t|$ ค่าคลาดเคลื่อนลดลงช้ามาก
- กลับไปที่ตัวละครที่โลลไปมา ปัญหาอยู่ที่วิธีวางคีย์ ไม่ใช่แค่จำนวนคีย์ และยังมีเรื่องความเร็วที่ต้องดูต่อ ซึ่งหัวข้อ 3.3 จะพิจารณา

อ่านต่อ. หัวข้อ 3.3 จะดูการต่อพหุนามดีกรีต่ำหลายช่วงให้เรียบ และดูว่าทำไมความเร็วที่เรียบสำคัญเมื่อเราให้กล้องเคลื่อนผ่านจุดที่กำหนด อ่านเพิ่มได้ใน Atkinson 1989, §3.5, pp. 154--159 และ Iserles 2009, ภาคผนวก A.2.2, pp. 442--443

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 3.2.1 (⋆) ให้ $f(t) = t^4$ และโหนด $0,\ 1,\ 2$

1. จงหาพหุนามดีกรีไม่เกิน 2 ที่ผ่านค่าของ f ที่โหนดทั้งสาม แล้วหาค่าคลาดเคลื่อน $f(3) - p(3)$
2. จงหา $\omega(3)$ และหา ξ ในช่วง $[0, 3]$ ที่ทำให้สูตรค่าคลาดเคลื่อนเป็นจริง

แบบฝึกหัด 3.2.2 (⋆) ให้โหนดห่างเท่ากัน $-2,\ 0,\ 2$ บนช่วง $[-2, 2]$ และให้โหนดเชบีเชฟสามตัวบนช่วงเดียวกัน

ฉบับร่าง · Preprint · 2026-10-25

1. จงแสดงว่า $\omega(t) = t^3 - 4t$ สำหรับโหนดห่างเท่ากัน แล้วหาค่าสูงสุดของ $|\omega|$ บน $[-2, 2]$ ด้วยมือ
2. จงแสดงว่าโหนดเชบีเชฟคือ $-\sqrt{3}$, 0, $\sqrt{3}$ และ $\omega(t) = t^3 - 3t$ แล้วหาค่าสูงสุดของ $|\omega|$ บน $[-2, 2]$ และเทียบกับสูตรในทฤษฎีบท

แบบฝึกหัด 3.2.3 (★★) ให้โหนดห่างเท่ากัน -1, $-\frac{1}{3}$, $\frac{1}{3}$, 1 บนช่วง $[-1, 1]$

1. จงแสดงว่า $\omega(x) = (x^2 - 1)(x^2 - \frac{1}{9})$
2. จงหาค่าสูงสุดของ $|\omega|$ บน $[-1, 1]$ โดยใส่ $u = x^2$ แล้วหาค่าสูงสุดของพหุนามกำลังสองใน $u \in [0, 1]$
3. โหนดเชบีเชฟสี่ตัวบน $[-1, 1]$ ให้ค่าสูงสุดเท่าใดตามทฤษฎีบท จงเทียบผลกับข้อ (ข)

แบบฝึกหัด 3.2.4 (★★) ให้ $f(t) = e^t$ บนช่วง $[0, 1]$ และโหนด 0, $\frac{1}{2}$, 1

1. จงแสดงว่า $\max_{[0,1]} |\omega| = \dfrac{\sqrt{3}}{36}$ (แนะนำ: หาจุดวิกฤตของ ω จากอนุพันธ์ของ ω ซึ่งเป็นสมการกำลังสอง)
2. จงใช้ทฤษฎีบทค่าคลาดเคลื่อนกับ $|f'''| \le e$ บน $[0, 1]$ หาขอบเขตของ $|e^t - p(t)|$ บน $[0, 1]$
3. ที่ $t = \frac{1}{4}$ จงคำนวณ $p(\frac{1}{4})$ ด้วยรูปแบบของนิวตัน แล้วตรวจว่าค่าคลาดเคลื่อนจริงไม่เกินขอบเขตนั้น

แบบฝึกหัด 3.2.5 (★)

1. เริ่มจาก $T_0 = 1$ และ $T_1 = x$ แล้วใช้สูตรเวียน $T_{m+1} = 2xT_m - T_{m-1}$ หา T_2 และ T_3
2. โหนดเชบีเชฟสามตัวบน $[-1, 1]$ คือ $\cos\frac{\pi}{6}$, $\cos\frac{\pi}{2}$, $\cos\frac{5\pi}{6}$ จงแสดงว่า $\omega(x) = \frac{1}{4}T_3(x)$

แบบฝึกหัด 3.2.6 (★★) ใช้ตาราง 4001 จุดบน $[-5, 5]$ ของหัวข้อนี้เป็นแนวทาง หาค่าคลาดเคลื่อนสูงสุดสำหรับ $n = 6$ และ $n = 8$ (คือ 7 และ 9 คีย์) ทั้งการวางแบบห่างเท่ากันและแบบเชบีเชฟ แล้วตอบ

1. ค่าคลาดเคลื่อนสูงสุดของการวางแบบห่างเท่ากันเพิ่มขึ้นตาม n หรือไม่ เทียบกับตารางในหัวข้อนี้
2. ที่ $n = 8$ การวางแบบเชบีเชฟให้ค่าคลาดเคลื่อนสูงสุดน้อยกว่าการวางแบบห่างเท่ากันกี่เท่า
3. ที่ $n = 4$ (5 คีย์) สองแบบต่างกันมากหรือไม่ ใช้ค่าสูงสุดของทั้งสองแบบตอบ

แบบฝึกหัด 3.2.7 (★★★) ใช้คีย์ห่างเท่ากันห้าตัว -5, $-\frac{5}{2}$, 0, $\frac{5}{2}$, 5 ของหัวข้อนี้

1. จงหา $\omega(6)$ และ $\omega(4.5)$
2. อธิบายว่าทำไมการประมาณค่าที่ $t = 6$ จึงเสี่ยงกว่าที่ $t = 4.5$ มาก แม้ $t = 4.5$ จะอยู่ใกล้ปลายช่วงแล้ว
3. จงใช้ทฤษฎีบทแสดงว่า $|F(6) - p(6)| \le |\omega(6)|$ แล้วอธิบายว่าเหตุใดขอบเขตนี้จึงหยาบมากในกรณีนี้

ตัวช่วยคิด

ข้อ (ข): ดูเครื่องหมายและขนาดของตัวประกอบ $t - t_j$ เมื่อ t อยู่นอกช่วงของคีย์ เทียบกับเมื่อ t อยู่ในช่วง ข้อ (ค): ใช้บทแทรกที่ว่า $|F^{(k)}(t)| \le k!$ สำหรับทุกจำนวนจริง t

ฟันเรา Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 3.2.1.

1. โหนดคือ $0, 1, 2$ และค่าของ f คือ $0, 1, 16$ เขียน $p(t) = a_0 + a_1 t + a_2 t^2$ จาก $p(0) = 0$ ได้ $a_0 = 0$ จาก $p(1) = 1$ ได้ $a_1 + a_2 = 1$ และจาก $p(2) = 16$ ได้ $2a_1 + 4a_2 = 16$ แทน $a_1 = 1 - a_2$ ลงในสมการหลัง ได้ $2 - 2a_2 + 4a_2 = 16$ จึง $a_2 = 7$ และ $a_1 = -6$ ดังนั้น $p(t) = 7t^2 - 6t$ เมื่อ $t = 3$ ได้ $p(3) = 63 - 18 = 45$ และ $f(3) = 81$ จึงได้ $f(3) - p(3) = 36$

2. $\omega(t) = t(t-1)(t-2)$ จึง $\omega(3) = 3 \cdot 2 \cdot 1 = 6$ อนุพันธ์อันดับสามคือ $f'''(t) = 24t$ จึง $f'''(\xi)/3! = 4\xi$ สูตรให้ $36 = 4\xi \cdot 6$ นั่นคือ $\xi = \frac{3}{2}$ ซึ่งอยู่ใน $[0, 3]$ ตามที่ต้องการ

ตัวเลขเหล่านี้ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

เฉลย 3.2.2.

1. $\omega(t) = (t+2)\,t\,(t-2) = t^3 - 4t$ จุดวิกฤตหาจาก $\omega'(t) = 3t^2 - 4 = 0$ ได้ $t = \pm\frac{2}{\sqrt{3}}$ ที่ $t = \frac{2}{\sqrt{3}}$ ได้ $\omega = \frac{8}{3\sqrt{3}} - \frac{8}{\sqrt{3}} = -\frac{16}{3\sqrt{3}}$ และที่ $t = -\frac{2}{\sqrt{3}}$ ได้ค่าตรงข้าม ที่ปลาย $t = \pm 2$ ได้ $\omega = 0$ จึงได้ $\max_{[-2,2]} |\omega| = \frac{16}{3\sqrt{3}} \approx 3.079$

2. โหนดเชบีเชฟคือ $2\cos\frac{\pi}{6} = \sqrt{3}$, $2\cos\frac{\pi}{2} = 0$ และ $2\cos\frac{5\pi}{6} = -\sqrt{3}$ จึง $\omega(t) = (t - \sqrt{3})\,t\,(t + \sqrt{3}) = t^3 - 3t$ จุดวิกฤต $3t^2 - 3 = 0$ คือ $t = \pm 1$ โดย $\omega(1) = -2$ และ $\omega(-1) = 2$ ที่ปลาย $\omega(\pm 2) = 2$ จึงได้ค่าสูงสุดเท่ากับ 2 ทฤษฎีบทให้ $r^{n+1}2^{-n}$ เมื่อ $r = 2$ และ $n = 2$ ได้ $2^3 \cdot 2^{-2} = 2$ ตรงกัน

ค่าที่ได้สำหรับโหนดห่างเท่ากันคือ 3.079 และสำหรับโหนดเชบีเชฟคือ 2 ซึ่งต่างกันราว 1.5 เท่า

เฉลย 3.2.3.

1. $\omega(x) = (x+1)(x+\frac{1}{3})(x-\frac{1}{3})(x-1)$ จับคู่ $(x+1)(x-1) = x^2 - 1$ และ $(x+\frac{1}{3})(x-\frac{1}{3}) = x^2 - \frac{1}{9}$ จึงได้ผลที่ต้องการ

2. ให้ $u = x^2$ เมื่อ $x \in [-1, 1]$ จะได้ $u \in [0, 1]$ และ $g(u) = (u - 1)(u - \frac{1}{9}) = u^2 - \frac{10}{9}u + \frac{1}{9}$ จุดวิกฤต $g'(u) = 2u - \frac{10}{9} = 0$ ได้ $u = \frac{5}{9} \in [0, 1]$ โดย $g(\frac{5}{9}) = \frac{25}{81} - \frac{50}{81} + \frac{9}{81} = -\frac{16}{81}$ ที่ปลาย $g(0) = \frac{9}{81}$ และ $g(1) = 0$ จึงได้ $\max_{[-1,1]} |\omega| = \frac{16}{81} \approx 0.1975$

3. ทฤษฎีบทให้ $2^{-3} = \frac{1}{8} = 0.125$ สำหรับโหนดเชบีเชฟ $n = 3$ บน $[-1, 1]$ อัตราส่วน $\dfrac{16/81}{1/8} = \dfrac{128}{81} \approx 1.58$ คือ การวางแบบห่างเท่ากันให้ค่าสูงกว่าราว 1.6 เท่า

เฉลย 3.2.4.

1. $\omega(t) = t(t-\frac{1}{2})(t-1) = t^3 - \frac{3}{2}t^2 + \frac{1}{2}t$ จึง $\omega'(t) = 3t^2 - 3t + \frac{1}{2}$ ซึ่งมีรากที่ $t = \dfrac{3 \pm \sqrt{3}}{6}$ ให้ $s = \dfrac{3 - \sqrt{3}}{6}$ จากสมการ $3s^2 - 3s + \frac{1}{2} = 0$ ได้ $s(s-1) = -\frac{1}{6}$ และ $s - \frac{1}{2} = -\dfrac{\sqrt{3}}{6}$ ดังนั้น $\omega(s) = -\frac{1}{6} \cdot \left(-\dfrac{\sqrt{3}}{6}\right) = \dfrac{\sqrt{3}}{36}$ ค่าที่ $t = 1 - s$ เป็นลบของค่านี้เพราะ $\omega(1 - t) = -\omega(t)$ จุดปลาย $\omega(0) = \omega(1) = 0$ จึงได้ $\max_{[0,1]} |\omega| = \dfrac{\sqrt{3}}{36} \approx 0.04811$

2. $f'''(t) = e^t$ และ $|f'''| \le e$ บน $[0, 1]$ จึงได้ $M/3! = e/6$ ทฤษฎีบทค่าคลาดเคลื่อนให้ $|e^t - p(t)| \le \dfrac{e}{6} \cdot \dfrac{\sqrt{3}}{36} = \dfrac{\sqrt{3}\,e}{216} \approx 0.02180$

สำเนา. Preprint. 2026-10-23

3. รูปแบบของนิวตันมีค่าสัมประสิทธิ์ $f[0]=1$, $f[0,\frac{1}{2}]=2(\sqrt{e}-1)\approx 1.297443$, $f[\frac{1}{2},1]=2(e-\sqrt{e})\approx 2.139121$ และ $f[0,\frac{1}{2},1]=f[\frac{1}{2},1]-f[0,\frac{1}{2}]\approx 0.841679$ จึง $p(t)=1+1.297443\,t+0.841679\,t(t-\frac{1}{2})$ เมื่อ $t=\frac{1}{4}$ ได้ $p(\frac{1}{4})=1+0.324361-0.052605=1.271756$ ค่าจริง $e^{1/4}\approx 1.284025$ จึงค่าคลาดเคลื่อนจริงราว 0.01227 ซึ่งไม่เกินขอบเขต 0.02180

ตัวเลขเหล่านี้ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

เฉลย 3.2.5.

1. $T_2=2x\cdot x-1=2x^2-1$ และ $T_3=2x(2x^2-1)-x=4x^3-3x$
2. โหนดคือ $\frac{\sqrt{3}}{2}, 0$ และ $-\frac{\sqrt{3}}{2}$ จึง $\omega(x)=x\left(x-\frac{\sqrt{3}}{2}\right)\left(x+\frac{\sqrt{3}}{2}\right)=x^3-\frac{3}{4}x$ ส่วน $\frac{1}{4}T_3(x)=\frac{1}{4}(4x^3-3x)=x^3-\frac{3}{4}x$ ทั้งสองเท่ากัน

เฉลย 3.2.6. ค่าต่อไปนี้มาจากโค้ดของเรา บนตาราง 4001 จุด

1. ค่าคลาดเคลื่อนสูงสุดของการวางแบบห่างเท่ากันคือ 0.44 ($n=4$), 0.62 ($n=6$), 1.05 ($n=8$), 1.9 ($n=10$) และ 60 ($n=20$) เมตร เพิ่มขึ้นตลอดเมื่อ n เพิ่ม ซึ่งสอดคล้องกับตารางในหัวข้อนี้
2. สำหรับ $n=6$ (7 คีย์) ได้ห่างเท่ากัน 0.617 เมตร และเชบีเชฟ 0.264 เมตร สำหรับ $n=8$ (9 คีย์) ได้ห่างเท่ากัน 1.045 เมตร และเชบีเชฟ 0.171 เมตร ที่ $n=8$ อัตราส่วนคือ $1.045/0.171\approx 6.1$ เท่า
3. ที่ $n=4$ การวางแบบห่างเท่ากันให้ 0.438 เมตร และแบบเชบีเชฟให้ 0.402 เมตร ต่างกันเพียงราว 8% ความต่างนี้โตขึ้นเมื่อ n เพิ่ม ดังตารางในหัวข้อนี้

ตัวเลขเหล่านี้ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

เฉลย 3.2.7.

1. $\omega(t)=(t+5)(t+\frac{5}{2})\,t\,(t-\frac{5}{2})(t-5)$ เมื่อ $t=6$ ได้ $11\cdot 8.5\cdot 6\cdot 3.5\cdot 1=1963.5$ เมื่อ $t=4.5$ ได้ $9.5\cdot 7\cdot 4.5\cdot 2\cdot(-0.5)=-299.25$
2. เมื่อ $t=6$ อยู่นอกช่วงทุกตัวประกอบ $t-t_j$ เป็นบวก และมีขนาดเพิ่มขึ้นเมื่อ t ห่างจากคีย์ ผลคูณจึงโตเร็ว ส่วนเมื่อ $t=4.5$ ตัวประกอบ $t-5$ เป็นลบและเล็ก (เท่ากับ -0.5) ทำให้ $|\omega|$ น้อยกว่าเกือบเจ็ดเท่า การทำนายนอกช่วงของคีย์จึงเสี่ยงมากกว่า
3. ทฤษฎีบทค่าคลาดเคลื่อนให้ $F(6)-p(6)=\dfrac{F^{(5)}(\xi)}{5!}\,\omega(6)$ และบทแทรกให้ $|F^{(5)}(\xi)|/5!\le 1$ จึง $|F(6)-p(6)|\le|\omega(6)|=1963.5$ อย่างไรก็ตาม ค่าจริงจากโค้ดคือ $|F(6)-p(6)|\approx 1.689$ เมตร (โดย $F(6)=\frac{1}{37}\approx 0.027$ และ $p(6)\approx 1.716$) ซึ่งหมายความว่า $|F^{(5)}(\xi)|/5!\approx 1.689/1963.5\approx 8.6\times 10^{-4}$ เล็กกว่า 1 มาก ขอบเขตนี้จึงหยาบเกินไปที่จะใช้ทำนายค่าจริง แต่ใช้เปรียบเทียบการวางคีย์ได้

ตัวเลขเหล่านี้ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

ฉบับร่าง Preprint 2026-10-23

3.3 ต่อเส้นให้เรียบ

สถานการณ์: เส้นโค้งที่บอกว่าการเปลี่ยนค่าควรเร็วแค่ไหน

ในมาตรฐานเว็บ CSS ของ W3C มีชุดฟังก์ชันชื่อ **การผ่อน** (easing) ซึ่งบอกว่า เมื่อค่าหนึ่งค่าเปลี่ยนจากค่าเริ่มไปค่าสุดท้ายในช่วงเวลาหนึ่ง ความคืบหน้าควรเป็นอย่างไรตลอดช่วงนั้น คีย์เวิร์ดที่รู้จักกันดีคือ `ease`, `ease-in`, `ease-out` และ `ease-in-out` สเปกไม่ได้เขียนเป็นสูตรภาษาคณิตศาสตร์ แต่กำหนดแต่ละคีย์เวิร์ดด้วยจุดควบคุมของเส้นโค้งเบซิเยร์ลูกบาศก์ เช่น `ease` เท่ากับ `cubic-bezier(0.25, 0.1, 0.25, 1)` Birtles, Jackson, and Rakow 2023, §2.2 สเปกนี้บอกนิยามเท่านั้น ไม่ได้บอกว่าซอฟต์แวร์ใดคำนวณอย่างไร

ลองนึกถึงการเปลี่ยนค่าจาก 0 ไป 1 ในเวลาหนึ่งวินาที ถ้าใช้เส้นตรง ค่าจะเพิ่มเท่า ๆ กันทุกช่วงเวลา แต่ถ้าใช้ `ease` ค่าจะเพิ่มเร็วในตอนแรก แล้วค่อย ๆ ช้าลงจนหยุดนิ่งที่ปลาย คำถามคือเส้นโค้งหนึ่งเส้นทำให้ความเร็วเปลี่ยนแบบนี้ได้อย่างไร และความเร็วนั้นเปลี่ยนที่ไหนบ้าง

ภาพรวมก่อนอ่าน

ในหัวข้อ 3.2 เราเห็นว่าพหุนามเส้นเดียวที่ผ่านทุกจุดอาจแกว่งมาก ทางออกหนึ่งคือต่อพหุนามดีกรีต่ำหลายช่วงเข้าด้วยกัน แต่ต่อแบบไหนจึงจะเรียบ ทั้งตำแหน่งและความเร็ว หัวข้อนี้มีสามขั้น ขั้นแรกคือเส้นโค้งเบซิเยร์ที่ใช้นิยามการผ่อน ขั้นที่สองคือการต่อด้วยพหุนามลูกบาศก์ที่รู้ทั้งค่าและความชันที่จุดปลาย (**แอร์มีต** (Hermite)) ขั้นที่สามคือ**สไปลน์ลูกบาศก์** (cubic spline) ซึ่งเรียบถึงอนุพันธ์อันดับสอง และหาได้จากระบบสมการสามแนวเฉียง

เส้นโค้งเบซิเยร์ลูกบาศก์

ลองนึกถึงจุดควบคุมเป็นมือจับสองอันที่ดึงเส้นโค้ง **เส้นโค้งเบซิเยร์** (Bézier curve) คือเส้นโค้งที่ถูกกำหนดด้วยจุดควบคุม เส้นโค้งนี้เริ่มที่ $P_0 = (0, 0)$ และจบที่ $P_3 = (1, 1)$ ส่วน $P_1 = (x_1, y_1)$ กับ $P_2 = (x_2, y_2)$ เป็นมือจับที่เราเลือกเอง เขียนเส้นโค้งด้วยพารามิเตอร์ $s \in [0, 1]$ ได้ดังนี้

$$x(s) = 3(1-s)^2 s\, x_1 + 3(1-s) s^2\, x_2 + s^3, \qquad y(s) = 3(1-s)^2 s\, y_1 + 3(1-s) s^2\, y_2 + s^3.$$

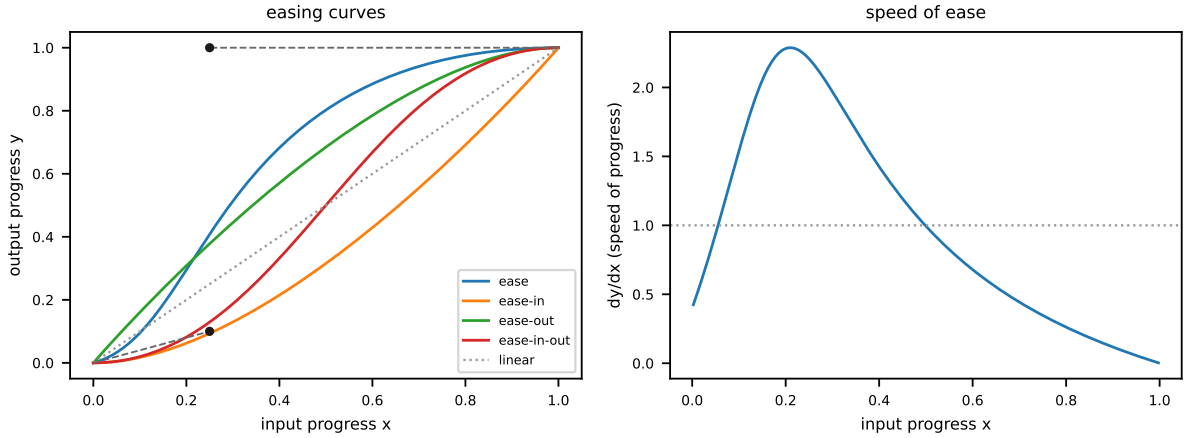
สังเกตว่า s ไม่ใช่เวลา มันเป็นตัวเลื่อนไปตามเส้นโค้ง แกนนอน x คือสัดส่วนของเวลาที่ผ่านไป และแกนตั้ง y คือสัดส่วนของค่าที่คืบหน้าไปแล้ว ดังนั้นถ้าเรารู้ x เราต้องหา s ที่ทำให้ $x(s) = x$ ก่อน แล้วจึงอ่าน $y(s)$ เราจะพิสูจน์ในไม่ช้าว่าสำหรับคีย์เวิร์ดทั้งสี่ มี s เพียงค่าเดียวเสมอ

อ่านแบบเข้าใจง่าย

- เส้นโค้งผ่าน $(0, 0)$ และ $(1, 1)$ เสมอ เพราะ $x(0) = y(0) = 0$ และ $x(1) = y(1) = 1$
- มือจับ P_1 บอกว่าเส้นโค้งออกตัวไปทางไหน ส่วน P_2 บอกว่าเส้นโค้งเข้ามาจบอย่างไร
- สเปกกำหนดให้ x_1, x_2 อยู่ในช่วง $[0, 1]$ ส่วน y_1, y_2 เลือกได้อิสระตามสเปก

รูปด้านซ้ายแสดงเส้นโค้งของคีย์เวิร์ดทั้งสี่ พร้อมเส้นตรง $y = x$ สำหรับเทียบ จุดสองจุดคือมือจับของ `ease` รูปด้านขวาแสดงความชัน dy/dx ของ `ease` ซึ่งเราจะอธิบายต่อไป

ฉบับร่าง Preprint · 2026-10-23



รูปที่ 3.9: ซ้าย: เส้นโค้งผ่อนของคีย์เวิร์ดทั้งสี่ เทียบกับเส้นตรง จุดเข้มคือมือจับ P_1 และ P_2 ของ `ease` ขวา: ความชัน dy/dx ของ `ease` ที่แต่ละค่า x ค่านี้คือความเร็วของความคืบหน้า (สร้างจาก `numanal/ch03.py`)

ลองเอง: จุดที่ $s = \frac{1}{2}$ ของ ease

ให้ $x_1 = x_2 = \frac{1}{4}$, $y_1 = \frac{1}{10}$ และ $y_2 = 1$ ลองคำนวณ $x(\frac{1}{2})$ และ $y(\frac{1}{2})$ ด้วยมือ โดยใช้ $x(\frac{1}{2}) = \frac{3}{8}x_1 + \frac{3}{8}x_2 + \frac{1}{8}$ เป็นตัวช่วย

คำตอบคือ $x(\frac{1}{2}) = \frac{3}{32} + \frac{3}{32} + \frac{4}{32} = \frac{10}{32} = 0.3125$ และ $y(\frac{1}{2}) = \frac{3}{8} \cdot \frac{1}{10} + \frac{3}{8} + \frac{1}{8} = 0.5375$ ดังนั้นเมื่อเวลาผ่านไปร้อยละ 31.25 ค่าจะคืบหน้าไปร้อยละ 53.75 ของทาง และเมื่อ $x = 0.5$ ค่าคืบหน้าไปราว 0.802 (ตัวเลขทั้งหมดในหัวข้อนี้ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`)

ต่อเส้นตรงแล้วความเร็วกระโดด

วิธีที่ง่ายที่สุดในการต่อจุดหลายจุดคือลากเส้นตรงระหว่างจุดที่อยู่ติดกัน ตำแหน่งจะต่อเนื่องเสมอ แต่ความเร็วบนแต่ละช่วงคงที่และเปลี่ยนทันทีที่ผ่านจุด เราเห็นเรื่องนี้แล้วในหัวข้อ 3.1 ที่คีย์กลาง ความเร็วเปลี่ยนจาก 3 เป็น -2 ทันที ตอนนี้ลองดูว่าเรื่องนี้ปรากฏในฉากจริงแบบใด

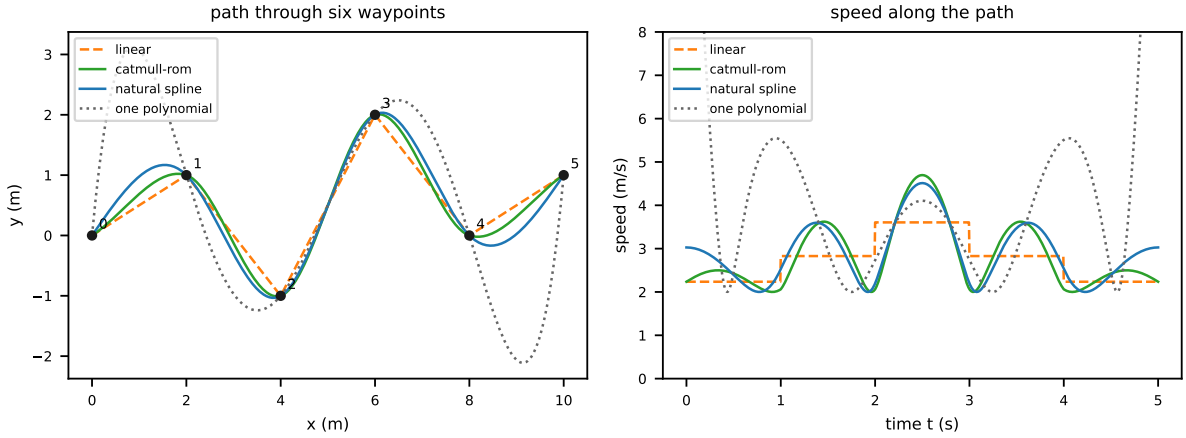
สถานการณ์สมมติ: อีกด้านหนึ่งของปัญหาเดียวกัน: กล้องในฉากตัดที่ต้องผ่านจุด

ในหัวข้อ 3.1 เราเห็นตัวละครที่ผ่านคีย์เฟรมโดยตำแหน่งถูกต้องทุกจุด ทีนี้ลองดูกล้องในฉากตัดที่ต้องผ่านจุดนำทาง (waypoint) **จุดนำทาง** (waypoint) หกจุดตามลำดับ คือจุดที่กล้องต้องผ่านพอดี ให้จุดแรกอยู่ที่วินาทีที่ 0 และจุดต่อ ๆ ไปอยู่ที่วินาทีที่ $1, 2, \ldots, 5$ ตำแหน่งในหน่วยเมตรคือ $(0,0)$, $(2,1)$, $(4,-1)$, $(6,2)$, $(8,0)$ และ $(10,1)$ ถ้าต่อจุดด้วยเส้นตรง กล้องผ่านทุกจุดแน่นอน แต่ความเร็วบนสี่ช่วงที่คั่นอยู่คือ $\sqrt{5} \approx 2.24$, $\sqrt{8} \approx 2.83$, $\sqrt{13} \approx 3.61$, $\sqrt{8}$ และ $\sqrt{5}$ เมตรต่อวินาที และ**ความเร็วกระโดด** (ความต่างของความเร็วทางซ้ายกับทางขวาของจุด) ที่จุดนำทางที่ 1 ถึง 4 คือ 3, 5, 5 และ 3 เมตรต่อวินาที ผู้ชมจะเห็นว่ากล้องเปลี่ยนความเร็วกะทันหันที่จุดนั้น ทั้งที่ตำแหน่งยังถูกต้อง

สาเหตุของเรื่องนี้คือรูปร่างของเส้นทางที่หักมุมที่จุดนำทาง ไม่ใช่การปัดเศษของตัวเลขในเครื่อง ตัวเลขทั้งหมดมาจากโค้ดของเรา และนี่เป็นสถานการณ์สมมติ เราไม่ได้บอกว่าระบบกล้องของเกมใดทำอย่างไร

ทางออกคือทำให้ความเร็วต่อเนื่องที่จุดนำทาง วิธีแรกคือ **แคทมัลล์–รอม** (Catmull–Rom) ซึ่งใช้ความชันจากจุดข้างเคียง

ร่างสำหรับ Preprint · 2026-10-23



รูปที่ 3.10: ซ้าย: กล้องผ่านจุดนำทางหกจุดด้วยเส้นตรง (เส้นประ) เส้นแคทมัลล์–รอม (เส้นทึบเขียว) และสไปลน์ลูกบาศก์ธรรมชาติ (เส้นทึบน้ำเงิน) พหุนามเส้นเดียว (เส้นจุดสีเทา) ขวา: ความเร็วตามเวลา เส้นตรงกระโดดที่ทุกจุด ส่วนอีกสองเส้นต่อเนื่อง และพหุนามเส้นเดียวที่ค่าความเร็วเกินแปดเมตรต่อวินาทีใกล้ $t = 5$ ถูกตัดออกจากกรอบ (สร้างจาก `numanal/ch03.py`)

และวิธีที่สองคือสไปลน์ลูกบาศก์ที่เราจะหาในหัวข้อย่อยสุดท้าย

ทำไมไม่ใช้พหุนามเส้นเดียวที่ผ่านทั้งหกจุดเลย คำตอบอยู่ในหัวข้อ 3.2 พหุนามดีกรีห้าที่ผ่านทุกจุดมีความเร็วต่อเนื่องอยู่แล้ว แต่บนเส้นทางกล้องนี้ค่า y ของพหุนามขึ้นไปถึง 3.11 และลงไปถึง -2.11 ทั้งที่จุดนำทางทั้งหกอยู่ระหว่าง -1 ถึง 2 ความเร็วที่ปลายช่วงสูงถึง 16.98 เมตรต่อวินาที (ตัวเลขจากโค้ดของเรา) กล้องจึงแกว่งออกนอกเส้นทางที่ผู้กำกับวางไว้ ทางที่ดีกว่าคือต่อพหุนามดีกรีต่ำหลายช่วง ซึ่งเป็นสิ่งที่เราจะทำในหัวข้อย่อยหลังจากนี้

กลับมาที่ `ease` ซึ่งเป็นคำถามแรกของหัวข้อนี้ คือความเร็วเปลี่ยนที่ไหน เพื่ออ่านความเร็วจากเส้นโค้ง เราต้องรู้ก่อนว่า y เป็นฟังก์ชันของ x ได้ ทฤษฎีบทต่อไปนี้บอกเรื่องนี้

ทฤษฎีบท 3.3.1: เส้นโค้งผ่อนเป็นฟังก์ชันของ x

ให้ $0 \le x_1 \le x_2 \le 1$ จะได้ $x'(s) > 0$ เมื่อ $0 < s < 1$ และ $x(0) = 0, x(1) = 1$ ดังนั้นสำหรับทุก $x \in [0,1]$ มี $s \in [0,1]$ เพียงค่าเดียวที่ทำให้ $x(s) = x$

แนวคิด. จัดรูปอนุพันธ์ของ $x(s)$ ให้เป็นผลรวมที่ทุกพจน์ไม่เป็นลบ

1. **อนุพันธ์.** หาอนุพันธ์ของ $x(s)$ ทีละพจน์แล้วจัดกลุ่มได้ $x'(s) = 3\big[(1-s)^2 d_0 + 2(1-s)s\,d_1 + s^2 d_2\big]$ โดย $d_0 = x_1$, $d_1 = x_2 - x_1$ และ $d_2 = 1 - x_2$

2. **เครื่องหมายของ d_k.** จากเงื่อนไข $0 \le x_1 \le x_2 \le 1$ ได้ $d_0, d_1, d_2 \ge 0$ และ $d_0 + d_1 + d_2 = 1$ ดังนั้นอย่างน้อยหนึ่งค่าเป็นบวก

3. **เครื่องหมายของ x'.** สัมประสิทธิ์ $(1-s)^2$, $2(1-s)s$ และ s^2 ไม่เป็นลบบน $[0,1]$ และเป็นบวกเมื่อ $0 < s < 1$ ดังนั้น $x'(s) \ge 0$ บน $[0,1]$ และเมื่อ $0 < s < 1$ พจน์ที่มี $d_k > 0$ เป็นบวก จึง $x'(s) > 0$

4. **มีค่าเดียว.** x เพิ่มขึ้นอย่างต่อเนื่องจาก $x(0) = 0$ ถึง $x(1) = 1$ ตามทฤษฎีบทค่ากลาง ทุกค่า $x \in [0,1]$ ถูกทำได้ และเพราะเพิ่มขึ้นอย่างเคร่งครัด จึงมีเพียงค่าเดียว

ฉบับร่าง Preprint · 2026-10-23

สรุป: ถ้ามือจับทั้งสองอยู่ในช่วง $[0,1]$ และเรียงตามลำดับ เวลาจะไปข้างหน้าเสมอ เราจึงหาพารามิเตอร์ s จากเวลา x ได้เพียงค่าเดียว

ความเร็วของการผ่อน

เพราะ y เป็นฟังก์ชันของ x ได้ ความชัน $dy/dx = y'(s)/x'(s)$ จึงบอกว่าเมื่อเวลาผ่านไปหนึ่งหน่วย ความคืบหน้าไปได้เร็วแค่ไหน นี่คือความเร็วของการผ่อนนั่นเอง ที่จุดเริ่มความชันเท่ากับ y_1/x_1 ซึ่งสำหรับ `ease` คือ $0.1/0.25 = 0.4$ ที่จุดจบความชันเท่ากับ $(1-y_2)/(1-x_2)$ ซึ่งเป็น 0 สำหรับ `ease` ระหว่างนั้น ความชันสูงสุดราว 2.29 ที่ $x \approx 0.21$ และที่ $x = 0.5$ มีค่าราว 0.99 ดังนั้น `ease` เริ่มด้วยความเร็ว 0.4 เร่งขึ้นจนราว 2.3 แล้วค่อยเบาลงจนหยุดที่ปลาย

ข้อควรระวัง: ความเร็วที่จุดเริ่มไม่จำเป็นต้องเป็นศูนย์

`ease` เริ่มด้วยความชัน 0.4 ไม่ใช่ 0 ถ้าเรานึกว่าการผ่อนต้องออกตัวจากการหยุดนิ่งเสมอ เราจะอ่านรูปผิด ส่วนที่หยุดนิ่งคือปลายด้านขวาของ `ease` ซึ่งมีความชัน 0 เท่านั้น

พหุนามลูกบาศก์จากค่าและความชัน

ถ้ารู้ค่าที่จุดสองจุดและความชันที่จุดเหล่านั้น เราจะหาพหุนามลูกบาศก์ได้หนึ่งเดียว ลองทำบนช่วง $[0,1]$ ก่อน

นิยาม 3.3.1: พหุนามแอร์มีตลูกบาศก์

บนช่วง $[0,1]$ ให้ p_0, m_0, p_1, m_1 เป็นจำนวนจริง พหุนามแอร์มีตลูกบาศก์ H คือ

$$H(s) = H_{00}(s)\,p_0 + H_{10}(s)\,m_0 + H_{01}(s)\,p_1 + H_{11}(s)\,m_1,$$

เมื่อ $H_{00} = 2s^3 - 3s^2 + 1$, $H_{10} = s^3 - 2s^2 + s$, $H_{01} = -2s^3 + 3s^2$ และ $H_{11} = s^3 - s^2$ (ฟังก์ชันพื้นฐานของแอร์มีต)

อ่านแบบเข้าใจง่าย

- ตรวจได้ทันทีว่า $H(0) = p_0$, $H(1) = p_1$, $H'(0) = m_0$ และ $H'(1) = m_1$ เพราะที่ $s = 0$ มีเพียง $H_{00} = 1$ และ $H'_{10} = 1$ ที่ไม่เป็นศูนย์ และที่ $s = 1$ มีเพียง $H_{01} = 1$ และ $H'_{11} = 1$
- ถ้า $p_0 = 0$, $m_0 = 1$, $p_1 = 2$ และ $m_1 = 0$ จะได้ $H(s) = -3s^3 + 4s^2 + s$ ที่ $s = \frac{1}{2}$ ได้ $H = \frac{9}{8}$ ซึ่งอยู่ระหว่างค่าปลาย 0 และ 2

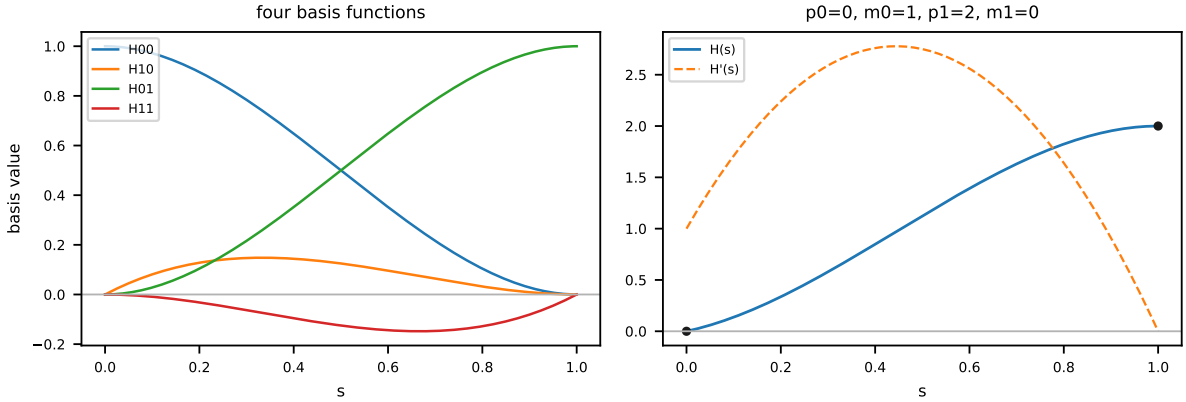
ทฤษฎีบท 3.3.2: ความมีหนึ่งเดียวของแอร์มีต

ถ้า p เป็นพหุนามดีกรีไม่เกิน 3 ที่ $p(0) = p_0$, $p'(0) = m_0$, $p(1) = p_1$ และ $p'(1) = m_1$ แล้ว $p = H$

แนวคิด. ผลต่างของพหุนามสองตัวที่เงื่อนไขเหมือนกันมีรากซ้ำที่ 0 และ 1 จึงต้องเป็นศูนย์

1. **ผลต่าง.** ให้ $D = p - H$ จะได้ $D(0) = D(1) = 0$ และ $D'(0) = D'(1) = 0$ เพราะเงื่อนไขทั้งสี่ตรงกัน
2. **รากซ้ำ.** ถ้า $D(0) = D'(0) = 0$ แล้ว s^2 หาร D ลงตัว และถ้า $D(1) = D'(1) = 0$ แล้ว $(s-1)^2$ หาร D ลงตัว

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 3.11: ซ้าย: ฟังก์ชันพื้นฐานทั้งสี่ของแอร์มีตบนช่วง $[0,1]$ ขวา: ตัวอย่าง $p_0=0, m_0=1, p_1=2, m_1=0$ โดย H เป็นเส้นทึบและ H' เป็นเส้นประ จุดคือค่าที่ปลายช่วง (สร้างจาก `numanal/ch03.py`)

3. **ดีกรี.** จึง $D=s^2(s-1)^2q(s)$ เมื่อ q เป็นพหุนาม ถ้า $q\neq 0$ แล้ว D มีดีกรีอย่างน้อยสี่ แต่ D มีดีกรีไม่เกินสาม ดังนั้น $q=0$ และ $D=0$ นั่นคือ $p=H$

สรุป: พหุนามลูกบาศก์ที่ทำให้ค่าและความชันที่สองปลายตรงกันมีเพียงหนึ่งเดียว

ถ้าช่วง $[t_k,t_{k+1}]$ ยาว $h_k\neq 1$ เราแปลงด้วย $s=(t-t_k)/h_k$ ความชันต่อหน่วย s จะเป็น h_k คูณความชันต่อหน่วย t ในวิธีแคทมัลล์–รอมที่ใช้ต่อไปนี้ ทุกช่วงยาวหนึ่งวินาที จึงไม่ต้องแปลง

แคทมัลล์–รอม: แทนเจนต์จากจุดข้างเคียง

วิธีแรกที่ใช้ได้กับจุดนำทางคือ ให้แต่ละจุด p_k มีความชัน m_k ที่คำนวณจากจุดข้างเคียง สูตรที่เราใช้คือ $m_k=\frac{1}{2}(p_{k+1}-p_{k-1})$ สำหรับจุดภายใน และที่จุดปลาย $m_0=p_1-p_0$ กับ $m_n=p_n-p_{n-1}$ จากนั้นช่วง k ที่อยู่ระหว่าง p_k กับ p_{k+1} ใช้พหุนามแอร์มีตลูกบาศก์ด้วยค่า p_k,m_k,p_{k+1},m_{k+1} วิธีนี้มีชื่อตามผู้เขียนสองคนที่ตีพิมพ์เรื่องสไปลน์แบบเฉพาะที่ในปี ค.ศ. 1974 Catmull and Rom 1974 ในที่นี้เราอ้างถึงเฉพาะชื่อและปีของบทความ โดยไม่ได้เปิดฉบับเต็ม และสูตรแทนเจนต์ที่ใช้ก็เป็นรูปแบบที่เราเลือกใช้ในบทนี้

ทฤษฎีบท 3.3.3: เส้นทางแคทมัลล์–รอมผ่านทุกจุดและความเร็วต่อเนื่อง

เส้นทางแคทมัลล์–รอมผ่านทุกจุดนำทาง และความเร็วทางซ้ายกับทางขวาของจุดภายใน p_k เท่ากันเสมอ (ทั้งคู่เท่ากับ m_k)

แนวคิด. ใช้สมบัติที่รู้แล้วของแอร์มีต คือค่าที่ปลายและความชันที่ปลายตรงตามที่กำหนด

1. **ผ่านจุด.** ช่วง k มี $H(0)=p_k$ และ $H(1)=p_{k+1}$ จึงผ่านทุกจุด
2. **ความเร็วทางขวาของ p_k.** ที่ปลายช่วง k ซึ่งเป็นช่วงแรกที่ p_k อยู่ที่ $s=0$ ความเร็วคือ $H'(0)=m_k$
3. **ความเร็วทางซ้ายของ p_k.** ช่วง $k-1$ จบที่ p_k ซึ่งเป็น $s=1$ ความเร็วคือ $H'(1)=m_k$
4. **สรุป.** ความเร็วทั้งสองข้างเท่ากับ m_k จึงต่อเนื่อง

ฉบับร่าง Preprint · 2026-10-23

สรุป: แทนเจนต์ที่ใช้ร่วมกันทำให้ความเร็วต่อเนื่อง แต่ความโค้งไม่ได้รับประกันว่าต่อเนื่อง

ที่กล้องของเรา ผลลัพธ์นี้หมายความว่า ความเร็วที่จุดนำทางที่ 1 เท่ากับ 2.06 เมตรต่อวินาทีทั้งสองข้าง และกระโดดเป็นศูนย์ที่จุดภายในทั้งสี่จุด (ดูรูปที่ 3.10) ความต่างระหว่างวิธีนี้กับสไปลน์คือ วิธีนี้ต้องการเพียงจุดข้างเคียง

ลองเอง: ตำแหน่งของกล้องที่ $t = \frac{1}{2}$ บนช่วงแรก

ช่วงแรกของเส้นทางกล้องผ่านจุดนำทางที่ 0 และ 1 ซึ่งมีค่า y เป็น 0 และ 1 ที่จุด 0 แทนเจนต์ของแคทมัลล์–รอมคือ 1 และที่จุด 1 คือ $-\frac{1}{2}$ ส่วนสไปลน์ธรรมชาติให้ $M_1 = -\frac{84}{11}$ จงหาค่า y ที่ $t = \frac{1}{2}$ ของทั้งสามวิธี (เส้นตรง แคทมัลล์–รอม และสไปลน์)

คำตอบคือ เส้นตรงให้ 0.5 แคทมัลล์–รอมให้ $\frac{11}{16} = 0.6875$ และสไปลน์ให้ $\frac{43}{44} \approx 0.977$ ตัวเลขของแคทมัลล์–รอมมาจาก $H_{10}(\frac{1}{2}) = \frac{1}{8}$, $H_{01}(\frac{1}{2}) = \frac{1}{2}$ และ $H_{11}(\frac{1}{2}) = -\frac{1}{8}$ ส่วนสไปลน์ใช้สูตรของช่วงแรกซึ่ง $y_0 = 0$ และ $C = 0$ (ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`) ทั้งสามวิธีผ่านจุดเดียวกันทุกจุด แต่ตำแหน่งที่ $t = \frac{1}{2}$ ต่างกันเกือบครึ่งเมตร

สไปลน์ลูกบาศก์: ความเรียบถึงความโค้ง

สไปลน์ลูกบาศก์คือฟังก์ชันที่เป็นพหุนามลูกบาศก์บนแต่ละช่วง และต่อกันอย่างเรียบจนถึงอนุพันธ์อันดับสอง เราจะเรียกความต่อเนื่องแบบนี้ว่าต่อเนื่องถึงอนุพันธ์อันดับสอง (C^2) ซึ่งเรียบกว่าแอร์มีตที่ต่อเนื่องเพียงอันดับหนึ่ง Atkinson 1989, §3.7, p. 164 ที่แต่ละจุดต่อ t_k เราเก็บตัวไม่รู้หนึ่งตัวคือ $M_k = s''(t_k)$ ซึ่งเป็นความโค้ง และ $h_k = t_{k+1} - t_k$

นิยาม 3.3.2: สไปลน์ลูกบาศก์ประมาณค่า

ให้ $t_0 < t_1 < \cdots < t_n$ และค่า $y_0, \ldots, y_n$ ฟังก์ชัน s เป็นสไปลน์ลูกบาศก์ประมาณค่าถ้า (ก) บนแต่ละช่วง $[t_k, t_{k+1}]$ เป็นพหุนามดีกรีไม่เกินสาม (ข) s, s' และ s'' ต่อเนื่องบน $[t_0, t_n]$ และ (ค) $s(t_k) = y_k$ ทุก k

อ่านแบบเข้าใจง่าย

- เงื่อนไข (ก) และ (ข) ทำให้ s เป็นพหุนามเป็นชิ้น ๆ ที่ต่อกันอย่างเรียบ ส่วนเงื่อนไข (ค) บังคับให้ผ่านทุกจุด
- ปลายด้านหนึ่งต้องมีเงื่อนไขเพิ่มเติม เพราะนับตัวไม่รู้แล้วมีมากกว่าสมการ ปลายที่ใช้กันบ่อยมีสองแบบ คือปลายธรรมชาติ $s''(t_0) = s''(t_n) = 0$ และปลายที่กำหนดความชัน $s'(t_0) = d_0$, $s'(t_n) = d_n$

การหาสไปลน์: ระบบสามแนวเฉียง

ลองดูวิธีหา s บนช่วง $[t_k, t_{k+1}]$ ก่อน บนช่วงนี้ s'' เป็นฟังก์ชันเชิงเส้นที่ค่าปลายคือ M_k และ M_{k+1} จึงเขียนได้เป็น $s''(t) = M_k \frac{t_{k+1} - t}{h_k} + M_{k+1} \frac{t - t_k}{h_k}$ ปริพันธ์สองครั้งและใช้เงื่อนไข $s(t_k) = y_k$, $s(t_{k+1}) = y_{k+1}$ จะได้

$$s(t) = \frac{M_k (t_{k+1} - t)^3 + M_{k+1} (t - t_k)^3}{6 h_k} + \left(\frac{y_k}{h_k} - \frac{M_k h_k}{6} \right) (t_{k+1} - t) + \left(\frac{y_{k+1}}{h_k} - \frac{M_{k+1} h_k}{6} \right) (t - t_k).$$

สูตรนี้มีค่าที่ M ตั้งอยู่เท่านั้น ส่วนการคำนวณ M_k ทำโดยบังคับให้อนุพันธ์ s' ต่อเนื่องที่จุดภายใน

ทฤษฎีบท 3.3.4: สไปลน์ลูกบาศก์มีหนึ่งเดียว

ให้ $t_0 < \cdots < t_n$ และค่า $y_0, \ldots, y_n$ ถ้าใช้ปลายธรรมชาติ หรือปลายที่กำหนดความชัน d_0, d_n จะมีสไปลน์ลูกบาศก์ประมาณค่าเพียงหนึ่งเดียวที่ทำตามเงื่อนไขปลายนั้น ระบบที่ใช้หาค่า M เป็นระบบสามแนวเฉียงที่เด่นทแยง

ฉบับร่าง Preprint · 2026-10-23

จึงแก้ได้เสมอ

แนวคิด. เขียนระบบเชิงเส้นสำหรับ $M_0, \dots, M_n$ แล้วแสดงว่าเมทริกซ์ไม่เป็นเอกฐาน

1. **สมการภายใน.** การบังคับให้ s' ที่ t_k จากช่วงซ้ายกับช่วงขวาเท่ากัน ให้สมการ
$$\frac{h_{k-1}}{6}M_{k-1} + \frac{h_{k-1}+h_k}{3}M_k + \frac{h_k}{6}M_{k+1} = \frac{y_{k+1}-y_k}{h_k} - \frac{y_k - y_{k-1}}{h_{k-1}}, \qquad k = 1, \dots, n-1.$$
2. **สมการปลาย.** ปลายธรรมชาติให้ $M_0 = 0$ และ $M_n = 0$ ปลายที่กำหนดความชันให้ $\frac{h_0}{3}M_0 + \frac{h_0}{6}M_1 = \frac{y_1-y_0}{h_0} - d_0$ และ $\frac{h_{n-1}}{6}M_{n-1} + \frac{h_{n-1}}{3}M_n = d_n - \frac{y_n - y_{n-1}}{h_{n-1}}$
3. **เด่นทแยง.** ในทุกแถว สมาชิกบนเส้นทแยงมุมมีค่าสัมบูรณ์อย่างน้อยเป็นสองเท่าของผลรวมสมาชิกที่เหลือในแถว เช่นแถวภายในคือ $\frac{h_{k-1}+h_k}{3}$ เทียบกับ $\frac{h_{k-1}+h_k}{6}$ และปลายธรรมชาติมี 1 เทียบกับ 0 เมทริกซ์จึงเด่นทแยงอย่างเคร่งครัด
4. **ไม่เป็นเอกฐาน.** ถ้า $Av = 0$ ให้ j เป็นตำแหน่งที่ $|v_j|$ มากที่สุด แถว j ให้ $|a_{jj}||v_j| = \left|\sum_{i\neq j} a_{ji}v_i\right| \le \sum_{i\neq j}|a_{ji}||v_j| \le \frac{1}{2}|a_{jj}||v_j|$ จึง $v_j = 0$ และทุก $v_i = 0$ ดังนั้นระบบมีคำตอบเดียวสำหรับทุกค่า y และ d
5. **สรุป.** เมื่อ M ได้แล้ว สูตรของ s ให้สไปลน์ที่ตรงกับเงื่อนไข (ก)–(ค) และ s' ต่อเนื่องเพราะสมการภายในถูกบังคับไว้ ค่าที่ M ต้องเป็นคำตอบของระบบ จึงมีเพียงหนึ่งเดียว

สรุป: การหาสไปลน์ลูกบาศก์คือแก้ระบบสามแนวเฉียงหนึ่งระบบ ซึ่งแก้ได้เร็วด้วยการกำจัดไปข้างหน้าและแทนกลับ และผลที่ได้มีเพียงหนึ่งเดียว

เมทริกซ์ของระบบแบบปลายกำหนดความชันมีสมบัติสมมาตร บวกแน่นอน และเด่นทแยง และ Atkinson บอกว่าการแก้ระบบนี้ใช้ราว $8n$ การดำเนินการ Atkinson 1989, §3.7, (3.7.22), p. 169

ลองเอง: สไปลน์ธรรมชาติบนจุด $0, 1, 0, 1$

ให้จุดอยู่ที่ $t = 0, 1, 2, 3$ และค่า $y = 0, 1, 0, 1$ ใช้ $h_k = 1$ และปลายธรรมชาติ จงหา M_1 และ M_2 จากสมการภายใน แล้วหาค่า $s(\frac{1}{2})$

คำตอบคือ $M = (0, -4, 4, 0)$ จากสมการ $\frac{2}{3}M_1 + \frac{1}{6}M_2 = -2$ และ $\frac{1}{6}M_1 + \frac{2}{3}M_2 = 2$ สูตรของช่วงแรกให้ $s(t) = -\frac{2}{3}t^3 + \frac{5}{3}t$ จึง $s(\frac{1}{2}) = \frac{3}{4}$ (ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`)

สไปลน์ของเส้นทางกล้องด้วยมือ

ลองทำกับเส้นทางกล้องของเรา ใช้เฉพาะพิกัด y ของจุดนำทางหกจุด คือ $0, 1, -1, 2, 0, 1$ ห่างกันหนึ่งวินาทีทุกช่วง และใช้ปลายธรรมชาติ $M_0 = M_5 = 0$ เมื่อคูณสมการภายในด้วย 6 ตัวขวามือจะเป็นหกเท่าของความต่างอันดับสองของ y ซึ่งได้ $-18, 30, -30$ และ 18 ตามลำดับ ระบบที่ได้คือ

$$4M_1 + M_2 = -18, \qquad M_1 + 4M_2 + M_3 = 30, \qquad M_2 + 4M_3 + M_4 = -30, \qquad M_3 + 4M_4 = 18.$$

คำตอบคือ $M_1 = -\frac{84}{11}$, $M_2 = \frac{138}{11}$, $M_3 = -\frac{138}{11}$ และ $M_4 = \frac{84}{11}$ ลองแทนลงในสมการแรกดู จะได้ $4\cdot(-\frac{84}{11}) + \frac{138}{11} = \frac{-336+138}{11} = -18$ ตรงตามที่ต้องการ สมการอื่นตรวจได้ในทำนองเดียวกัน

ค่า M บอกความโค้งที่จุดนำทาง ที่จุดที่ 2 ซึ่งเป็นค่าต่ำสุดในท้องถิ่นของ y ความโค้ง M_2 เป็นบวก และที่จุดที่ 3 ซึ่ง $y = 2$ เป็นค่าสูงสุดในท้องถิ่น M_3 เป็นลบ ค่าทั้งสองเกิดจากระบบเดียวกัน จึงมีผลต่อกันทุกจุด

สำหรับแคทมัลล์–รอม แทนเจนต์ของ y คือ $(1, -\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, 1)$ ซึ่งคำนวณได้ทีละจุดโดยไม่ต้องแก้ระบบ

ฉบับร่าง · Preprint · 2026-10-23

วิธี	ความเรียบ	ความเร็วต่ำสุด (m/s)	ความเร็วสูงสุด (m/s)	ความเร็วกระโดดสูงสุด (m/s)
เส้นตรงระหว่างจุด	C^0	2.236	3.606	5.000
แคทมัลล์–รอม	C^1	2.000	4.697	0
สไปลน์ลูกบาศก์ธรรมชาติ	C^2	2.000	4.513	0
พหุนามเส้นเดียวดีกรีห้า	C^∞	2.000	16.985	0

ตารางนี้วัดความเร็วจากตัวอย่างละเอียดบนแต่ละช่วง (ตัวเลขจากโค้ดของเรา) ในตัวอย่างนี้วิธีเรียบมีความเร็วสูงสุดมากกว่าเส้นตรง เพราะเส้นโค้งต้องเลี้ยวระหว่างจุดนำทาง ไม่ได้บอกว่าทุกเส้นทางจะเป็นเช่นนี้

ความคลาดเคลื่อนของสไปลน์

ให้ $f \in C^4[a,b]$ (มีอนุพันธ์ต่อเนื่องถึงอันดับสี่บนช่วงปิด) และให้ s เป็น**สไปลน์สมบูรณ์** (complete spline) คือสไปลน์ลูกบาศก์ที่ปลายทั้งสองกำหนดความชันเป็นความชันจริง $s'(a) = f'(a)$ และ $s'(b) = f'(b)$ ถ้าช่วงห่างที่ใหญ่ที่สุดคือ h ความคลาดเคลื่อนสูงสุดจะลดลงตามกำลังสี่ของ h โดยค่าคงที่ที่ Atkinson ให้ไว้ใช้ได้คือ $c_0 = \frac{5}{384}$

$$\max_{[a,b]} |f - s| \le \frac{5}{384}\, h^4 \, \max_{[a,b]} |f^{(4)}|.$$

Atkinson 1989, §3.7, Thm 3.4, (3.7.24), pp. 169--170 ผลนี้มาจากทฤษฎีของ de Boor ที่ Atkinson อ้างไว้ เราไม่ได้พิสูจน์ในหนังสือเล่มนี้ แต่ตรวจตัวเลขได้ ตารางต่อไปนี้เป็นการคำนวณของเราเองสำหรับ $\sin t$ บน $[0, \pi]$ โดยใช้ความชันที่ปลาย 1 และ -1 ซึ่งเป็นความชันจริง ที่ $n = 4$ ขอบเขตคือ $\frac{5}{384}(\frac{\pi}{4})^4 \approx 0.00495$ ซึ่งเกินค่าที่วัดได้ 1.124×10^{-3}

จำนวนช่วง n	ค่าคลาดเคลื่อนสูงสุด	อัตราส่วนกับแถวก่อนหน้า
4	1.124×10^{-3}	
8	6.32×10^{-5}	17.8
16	3.89×10^{-6}	16.3
32	2.41×10^{-7}	16.1

อัตราส่วนเข้าใกล้ $16 = 2^4$ ซึ่งตรงกับอัตรา $O(h^4)$ แต่ตัวเลขนี้เป็นการตรวจเชิงตัวเลขเท่านั้น ไม่ใช่การพิสูจน์ และผลนี้ใช้กับ $f \in C^4$ และปลายที่กำหนดความชันอย่างถูกต้องเท่านั้น

ข้อควรระวัง: สไปลน์ขยับทั้งเส้นเมื่อแก้จุดเดียว

สไปลน์แตกต่างจากแคทมัลล์–รอมตรงที่ความชันที่จุดหนึ่งขึ้นกับทุกจุด เมื่อเราย้ายจุดนำทางที่ 2 จาก $(4, -1)$ ไปเป็น $(4, -3)$ เส้นแคทมัลล์–รอมที่เวลา $t = 4.5$ ไม่เปลี่ยนเลย เพราะช่วงนั้นใช้เพียงจุด 3 ถึง 5 แต่สไปลน์ที่เวลาเดียวกันขยับราว 0.086 เมตร (ตัวเลขจากโค้ดของเรา) นี่คือเหตุผลที่สไปลน์เรียกว่าเป็นวิธีแบบทั่วโลก ส่วนแคทมัลล์–รอมเป็นวิธีเฉพาะที่

ข้อควรระวัง: ความเร็วต่อเนื่องต่อพารามิเตอร์ ไม่ใช่ต่อระยะทาง

ทุกวิธีในหัวข้อนี้ทำให้ความเร็วต่อเนื่องเมื่อเทียบกับ t แต่กล้องที่เดินด้วยความเร็วเท่ากันบนระยะทางต้องใช้ระยะทางตามเส้นโค้ง ซึ่งเป็นปริพันธ์ที่บางครั้งไม่มีสูตรปิด เรื่องนี้จะกลับมาในบทที่ 5 ความเร็วที่ต่อเนื่องจึงยังไม่ใช่ความเร็วคงที่

สัมมนา · Preprint · 2026-10-25

สรุปใจความ

- เส้นโค้งเบซิเยร์ของ `ease` ให้ความชัน 0.4 ที่จุดเริ่ม และ 0 ที่จุดจบ และ $x(s)$ เพิ่มขึ้นอย่างเคร่งครัด จึงมีค่าของ s เพียงค่าเดียวต่อ x
- ต่อเส้นตรงทำให้ตำแหน่งต่อเนื่องแต่ความเร็วกระโดด ในฉากกล้องกระโดดที่จุดนำทาง 3, 5, 5 และ 3 เมตรต่อวินาที
- แคทมัลล์–รอมใช้ความชันจากจุดข้างเคียง ทำให้ความเร็วต่อเนื่องและเป็นวิธีเฉพาะที่
- สไปลน์ลูกบาศก์เรียบถึงอนุพันธ์อันดับสอง หาได้จากระบบสามแนวเฉียงที่เด่นทแยง แต่แก้จุดเดียวขยับทั้งเส้น และมีความคลาดเคลื่อน $O(h^4)$ เมื่อ $f \in C^4$ และปลายกำหนดความชันถูกต้อง
- กลับไปที่กล้อง ความเร็วที่เรียบต้องการทั้งตำแหน่งที่ถูกต้องและความเร็วที่ต่อเนื่อง ซึ่งวิธีทั้งสามให้ผลต่างกันตามที่เห็นในรูป

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 3.3.1 (⋆) ให้ $x_1 = \frac{1}{3}$, $y_1 = 0$, $x_2 = \frac{2}{3}$ และ $y_2 = 1$

1. จงแสดงว่า $x(s) = s$ สำหรับทุก $s \in [0, 1]$
2. จงแสดงว่า $y(s) = 3s^2 - 2s^3$ แล้วสรุปว่า $y = 3x^2 - 2x^3$ (เส้นโค้งนี้เรียกว่า smoothstep)
3. จงหา dy/dx ที่ $x = \frac{1}{2}$

แบบฝึกหัด 3.3.2 (⋆) ให้ $p_0 = 0$, $m_0 = 1$, $p_1 = 2$ และ $m_1 = 0$ บนช่วง $[0, 1]$

1. จงแสดงว่า $H(s) = -3s^3 + 4s^2 + s$ โดยแทนค่าในนิยามของพหุนามแอร์มีตลูกบาศก์
2. จงหา $H(\frac{1}{2})$ และ $H'(s)$ แล้วหาค่าสูงสุดของ H บน $[0, 1]$ ว่ามีค่าเกิน 2 หรือไม่

แบบฝึกหัด 3.3.3 (⋆⋆) ให้จุด $t = 0, 1, 2, 3$ และค่า $y = 0, 1, 0, 1$ ใช้ปลายธรรมชาติและ $h_k = 1$

1. จงแสดงว่าสมการภายในให้ $\frac{2}{3}M_1 + \frac{1}{6}M_2 = -2$ และ $\frac{1}{6}M_1 + \frac{2}{3}M_2 = 2$
2. จงแก้ระบบ แล้วหา $s(\frac{1}{2})$ จากสูตรของช่วง $[0, 1]$
3. จงตรวจว่า s ที่ได้ต่อเนื่องที่ $t = 1$ ทั้งค่าและอนุพันธ์ โดยคำนวณจากสูตรของช่วง $[0, 1]$ และ $[1, 2]$

แบบฝึกหัด 3.3.4 (⋆⋆) ให้จุดนำทาง $p_0 = 0$, $p_1 = 2$ และ $p_2 = 0$ ที่เวลา $t = 0, 1, 2$

1. จงหาแทนเจนต์ m_0, m_1, m_2 ตามสูตรของแคทมัลล์–รอม
2. จงแสดงว่าความเร็วทางซ้ายและทางขวาของ p_1 เป็นศูนย์ทั้งคู่
3. จงหาความเร็วกระโดดที่ p_1 ของเส้นตรงที่ต่อจุดด้วยเส้นตรง

แบบฝึกหัด 3.3.5 (⋆⋆)

1. จงแสดงว่า $H_{00} + H_{01} = 1$ และ $H_{01} + H_{10} + H_{11} = s$

ฉบับร่าง · Preprint · 2026-10-23

2. ให้ $p_0 = a$, $p_1 = a + b$ และ $m_0 = m_1 = b$ จงแสดงว่า $H(s) = a + bs$ ทุก s (พหุนามแอร์มีตทำซ้ำเส้นตรงได้)
3. ถ้าจุดนำทางอยู่บนเส้นตรง $p_k = a + bk$ จงแสดงว่าแคทมัลล์–รอมให้เส้นตรงเดิม (ต้องดูแทนเจนต์ที่ปลายด้วย)

แบบฝึกหัด 3.3.6 ($\star\star\star$) ในฉากกล้องของหัวข้อนี้ ให้เลื่อนจุดนำทางที่ 2 จาก $(4, -1)$ ไปเป็น $(4, -3)$ โดยจุดอื่นไม่เปลี่ยน

1. จงหาความเร็วบนแต่ละช่วงของเส้นตรง และความเร็วกระโดดที่จุดนำทางที่ $1, 2, 3$ และ 4
2. จงอธิบายว่าทำไมความเร็วกระโดดที่จุด 2 จึงใหญ่ที่สุด และเปลี่ยนไปจากเดิมเท่าไร
3. ที่เวลา $t = 4.5$ วินาที แคทมัลล์–รอมจะเปลี่ยนหรือไม่ และสไปลน์ธรรมชาติเปลี่ยนไปเท่าใด (ตรวจด้วยโค้ด) เพราะเหตุใดผลจึงต่างกัน

แบบฝึกหัด 3.3.7 ($\star\star\star$, โครงงานท้ายบท) ให้เลื่อนจุดนำทางที่ 2 ไปที่ $(4, y)$ สำหรับ $y \in \{-1,\ 0,\ 1,\ \tfrac{3}{2},\ 2,\ 3\}$

1. จงแสดงว่าความเร็วกระโดดของเส้นตรงที่จุด 2 เท่ากับ $|3 - 2y|$ แล้วหาค่า y ที่ทำให้เป็นศูนย์ และอธิบายเหตุผลทางเรขาคณิต
2. ทำตารางความเร็วกระโดดที่จุด 2 ของเส้นตรงทั้งหกค่า y และบอกว่าแคทมัลล์–รอมกับสไปลน์ให้ค่าเท่าใด (ตรวจด้วยโค้ด)
3. เขียนคำแนะนำสั้น ๆ สามถึงสี่บรรทัดสำหรับผู้ออกแบบฉากกล้องว่าควรเลือกวิธีใด เมื่อต้องการให้แก้จุดเดียวได้โดยไม่กระทบส่วนอื่น และเมื่อต้องการความโค้งที่ต่อเนื่อง

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 3.3.1.

1. $x(s) = 3(1-s)^2 s \cdot \tfrac{1}{3} + 3(1-s)s^2 \cdot \tfrac{2}{3} + s^3 = s(1-s)^2 + 2s^2(1-s) + s^3 = s\big[(1-s)+s\big]^2 = s$
2. $y(s) = 3(1-s)^2 s \cdot 0 + 3(1-s)s^2 \cdot 1 + s^3 = 3s^2 - 3s^3 + s^3 = 3s^2 - 2s^3$ เพราะ $x = s$ จึงได้ $y = 3x^2 - 2x^3$
3. จาก $y'(s) = 6s - 6s^2$ และ $x'(s) = 1$ จึง $dy/dx = 6s - 6s^2$ ที่ $s = \tfrac{1}{2}$ ได้ $3 - \tfrac{3}{2} = \tfrac{3}{2}$

ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

เฉลย 3.3.2.

1. $H(s) = H_{10}(s) \cdot 1 + H_{01}(s) \cdot 2 = (s^3 - 2s^2 + s) + 2(-2s^3 + 3s^2) = -3s^3 + 4s^2 + s$
2. $H(\tfrac{1}{2}) = -\tfrac{3}{8} + 1 + \tfrac{1}{2} = \tfrac{9}{8}$ และ $H'(s) = -9s^2 + 8s + 1 = -(9s^2 - 8s - 1) = -(9s+1)(s-1)$ จุดวิกฤตคือ $s = 1$ และ $s = -\tfrac{1}{9}$ ซึ่งอยู่นอก $[0, 1]$ ค่าที่ปลายคือ $H(0) = 0$ และ $H(1) = 2$ ดังนั้นค่าสูงสุดบน $[0, 1]$ คือ 2 ที่ $s = 1$ ไม่มีค่าเกินสอง

ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

สมบูรณ์ · Preprint · 2026-10-23

เฉลย 3.3.3.

1. ใช้สมการภายในของ $k=1$ และ $k=2$ ตามสูตร $\frac{h_{k-1}}{6}M_{k-1}+\frac{h_{k-1}+h_k}{3}M_k+\frac{h_k}{6}M_{k+1}=\frac{y_{k+1}-y_k}{h_k}-\frac{y_k-y_{k-1}}{h_{k-1}}$ เมื่อ $h_k=1$ และ $M_0=M_3=0$ ได้ $\frac{2}{3}M_1+\frac{1}{6}M_2=(0-1)-(1-0)=-2$ และ $\frac{1}{6}M_1+\frac{2}{3}M_2=(1-0)-(0-1)=2$

2. คูณสมการด้วย 6 ได้ $4M_1+M_2=-12$ และ $M_1+4M_2=12$ จึง $M_1=-4$ และ $M_2=4$ ช่วง $[0,1]$ ให้ $s(t)=\frac{M_0(1-t)^3+M_1t^3}{6}+C(1-t)+Dt$ เมื่อ $C=\frac{y_0}{1}-0=0$ และ $D=y_1-\frac{M_1}{6}=1+\frac{4}{6}=\frac{5}{3}$ ดังนั้น $s(t)=-\frac{2}{3}t^3+\frac{5}{3}t$ และ $s(\frac{1}{2})=-\frac{1}{12}+\frac{5}{6}=\frac{3}{4}$

3. ช่วง $[1,2]$ ให้ $s(t)=\frac{M_1(2-t)^3+M_2(t-1)^3}{6}+C'(2-t)+D'(t-1)$ เมื่อ $C'=y_1-\frac{M_1}{6}=\frac{5}{3}$ และ $D'=y_2-\frac{M_2}{6}=-\frac{2}{3}$ ที่ $t=1$ ได้ $s=\frac{M_1}{6}+C'=-\frac{2}{3}+\frac{5}{3}=1$ ซึ่งตรงกับช่วงแรก ส่วนอนุพันธ์ของช่วงแรกที่ $t=1$ คือ $-2t^2+\frac{5}{3}=-\frac{1}{3}$ และของช่วงที่สองคือ $-\frac{M_1}{2}-C'+D'=2-\frac{5}{3}-\frac{2}{3}=-\frac{1}{3}$ เท่ากัน

ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

เฉลย 3.3.4.

1. แทนเจนต์ภายในคือ $m_1=\frac{1}{2}(p_2-p_0)=0$ ส่วนปลายคือ $m_0=p_1-p_0=2$ และ $m_2=p_2-p_1=-2$

2. ความเร็วทางซ้ายของ p_1 คือ $H'(1)$ ของช่วง $[0,1]$ ซึ่งจากนิยามได้ $H'(1)=m_1$ (เพราะที่ $s=1$ มีเพียง $H'_{11}(1)=1$ ที่ไม่เป็นศูนย์) จึงเท่ากับ 0 ส่วนความเร็วทางขวาคือ $H'(0)$ ของช่วง $[1,2]$ ซึ่งเท่ากับ $m_1=0$ เช่นกัน ทั้งสองข้างเป็นศูนย์

3. เส้นตรงที่ต่อจุดให้ความเร็วทางซ้ายของ p_1 เท่ากับ $p_1-p_0=2$ และทางขวาเท่ากับ $p_2-p_1=-2$ ความเร็วกระโดดคือ $|-2-2|=4$

ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

เฉลย 3.3.5.

1. $H_{00}+H_{01}=(2s^3-3s^2+1)+(-2s^3+3s^2)=1$ และ $H_{01}+H_{10}+H_{11}=(-2s^3+3s^2)+(s^3-2s^2+s)+(s^3-s^2)=s$

2. $H(s)=a\,H_{00}+(a+b)H_{01}+b\,H_{10}+b\,H_{11}=a(H_{00}+H_{01})+b(H_{01}+H_{10}+H_{11})=a+bs$ ตามข้อ 1

3. ที่จุดภายใน $m_k=\frac{1}{2}(p_{k+1}-p_{k-1})=\frac{1}{2}(2b)=b$ และที่ปลาย $m_0=p_1-p_0=b$ กับ $m_n=b$ ทุกแทนเจนต์เท่ากับ b ช่วง k จึงเป็น $a+b(k+s)$ ตามข้อ 2 และทั้งเส้นทางคือเส้นตรงเดิม

ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 3.3.6.

1. ตำแหน่งใหม่คือ $(0,0)$, $(2,1)$, $(4,-3)$, $(6,2)$, $(8,0)$, $(10,1)$ ความเร็วบนแต่ละช่วงคือ $v_1 = (2,1)$, $v_2 = (2,-4)$, $v_3 = (2,5)$, $v_4 = (2,-2)$, $v_5 = (2,1)$ ความเร็วกระโดดที่จุด $1,2,3,4$ คือ $|(0,-5)| = 5$, $|(0,9)| = 9$, $|(0,-7)| = 7$ และ $|(0,3)| = 3$ เมตรต่อวินาที

2. ความเร็วกระโดดที่จุด 2 คือ $|v_3 - v_2| = |3 - 2y|$ เมื่อ $y = -3$ ได้ 9 ก่อนเลื่อนจุดนี้ได้ 5 จึงเพิ่มขึ้น 4 เมตรต่อวินาที เพราะเมื่อจุดลงไปอีกสองเมตร ช่วงขาลงจาก 1 ถึง 2 ชันลง และช่วงขาขึ้นจาก 2 ถึง 3 ชันขึ้นมากขึ้น ความต่างจึงใหญ่ขึ้น

3. แคทมัลล์–รอมที่เวลา 4.5 ไม่เปลี่ยนเลย เพราะเวลานี้อยู่ในช่วงที่ 4 ซึ่งใช้เพียงจุด 3, 4, 5 และแทนเจนต์ $m_4 = \frac{1}{2}(p_5 - p_3)$ และ $m_5 = p_5 - p_4$ ไม่มีจุด 2 ส่วนสไปลน์ธรรมชาติที่เวลาเดียวกันเปลี่ยนราว 0.086 เมตร เพราะระบบของ M เชื่อมทุกจุดเข้าด้วยกัน (ตัวเลขจากโค้ดของเรา)

ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

เฉลย 3.3.7.

1. ความเร็วบนช่วงที่ $1 \to 2$ คือ $(2,\ y-1)$ และบนช่วง $2 \to 3$ คือ $(2,\ 2-y)$ ความเร็วกระโดดคือผลต่าง $(0,\ 3-2y)$ ขนาดของมันคือ $|3-2y|$ ค่าที่ทำให้เป็นศูนย์คือ $y = \frac{3}{2}$ ซึ่งทำให้จุด 1, 2, 3 อยู่บนเส้นตรงเดียวกันพอดี (ความชันทั้งสองช่วงเท่ากับ $\frac{1}{4}$) ทางตรงที่ผ่านจุดสามจุดไม่มีการหักมุม จึงไม่มีการกระโดด

2. ตารางความเร็วกระโดดของเส้นตรงที่จุด 2:

y	-1	0	1
ความเร็วกระโดด (m/s)	5	3	1
y	$\frac{3}{2}$	2	3
ความเร็วกระโดด (m/s)	0	1	3

   แคทมัลล์–รอมและสไปลน์ธรรมชาติให้ความเร็วกระโดดเป็นศูนย์ทุกค่า y (ตรวจด้วยโค้ด)

3. ตัวอย่างคำแนะนำ: ถ้าต้องการแก้จุดเดียวโดยไม่กระทบส่วนอื่น ให้เลือกแคทมัลล์–รอม เพราะช่วงที่ห่างจากจุดนั้นไม่เปลี่ยน ถ้าต้องการความโค้งต่อเนื่องและยอมให้จุดหนึ่งกระทบทั้งเส้น ให้เลือกสไปลน์ธรรมชาติ ทั้งสองวิธีให้ความเร็วต่อเนื่องเมื่อวัดเทียบกับเวลา

ตรวจด้วยโค้ดแล้ว ดู `test_ch03.py`

สรุปบทที่ 3 ในหน้าเดียว

บทนี้ถามคำถามเดียว คือเมื่อรู้ค่าที่บางจุด เราจะเดาระหว่างจุดได้อย่างไร และเดาได้แม่นแค่ไหน สูตรและตัวเลขในตารางนี้มาจากทฤษฎีบทในบทนี้หรือจากโค้ดของเรา

ฉบับร่าง · Preprint · 2026-10-23

สูตรและตัวเลขสำคัญ

เรื่อง	สูตร	ตัวอย่างในบทนี้
ค่าคลาดเคลื่อนการประมาณค่า	$\dfrac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega(t)$	เป็นศูนย์ที่โหนดเสมอ
โหนดเชบีเชฟ	$\max\lvert\omega\rvert = \left(\frac{b-a}{2}\right)^{n+1} 2^{-n}$	บน $[0,4]$ กับ $n=2$ ได้ 2
ปรากฏการณ์รุงเง	$F(t) = \dfrac{1}{1+t^2}$ บน $[-5,5]$	ห่างเท่ากัน 21 คีย์ 59.8 เมตร
เส้นโค้งผ่อน	$x(s) = 3(1-s)^2 s\,x_1 + 3(1-s)s^2 x_2 + s^3$	`ease` ที่ $x=0.3125$ ได้ $y=0.5375$
ความชันของการผ่อน	$dy/dx = y'(s)/x'(s)$	เริ่ม 0.4, จบ 0
แอร์มีตลูกบาศก์	$H = H_{00}p_0 + H_{10}m_0 + H_{01}p_1 + H_{11}m_1$	$H(\frac{1}{2}) = \frac{9}{8}$
แคทมัลล์–รอม	$m_k = \frac{1}{2}(p_{k+1} - p_{k-1})$ (ปลาย: ผลต่างเดียว)	ความเร็วต่อเนื่อง (C^1)
สไปลน์ลูกบาศก์	$\frac{h_{k-1}}{6}M_{k-1} + \frac{h_{k-1}+h_k}{3}M_k + \frac{h_k}{6}M_{k+1} = \cdots$	บน $0,1,0,1$: $M = (0,-4,4,0)$
ความคลาดเคลื่อนสไปลน์	$O(h^4)$ ($f \in C^4$, ปลายกำหนดความชัน)	อัตราส่วน ≈ 16
ความเร็วกระโดดของเส้นตรง	$\lvert v_k - v_{k-1}\rvert$	$3, 5, 5, 3$ เมตรต่อวินาที

วิธีที่เทียบกันในบทนี้

วิธี	ความเรียบ	ผลของการแก้จุดเดียว
เส้นตรงระหว่างจุด	ตำแหน่ง C^0 ความเร็วกระโดด	เฉพาะช่วงนั้น
แอร์มีตหรือแคทมัลล์–รอม	ความเร็ว C^1	ขยับเฉพาะช่วงใกล้เคียง
สไปลน์ลูกบาศก์	ความโค้ง C^2	ขยับทั้งเส้น

อ่านต่อ. บทที่ 4 เริ่มจากข้อมูลที่สั่นหรือมีสัญญาณรบกวน เช่นลายเส้นที่ลากด้วยนิ้วหรือค่าชีพจรที่วัดได้ ถ้าเราต่อจุดทุกจุดด้วยเส้นโค้ง เราจะได้สัญญาณรบกวนไปด้วย บทหน้าจึงถามว่า ถ้าไม่ผ่านทุกจุด แต่เลือกเส้นที่ใกล้ข้อมูลที่สุด เส้นนั้นควรเป็นอย่างไร และทำไมจึงใช้เป็นหลักในงานกราฟิกและการวิเคราะห์ข้อมูล

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 4

เส้นที่เหมาะกับข้อมูลที่สุด

สถานการณ์สมมติ: ลายเส้นที่สั่นเพราะมือ

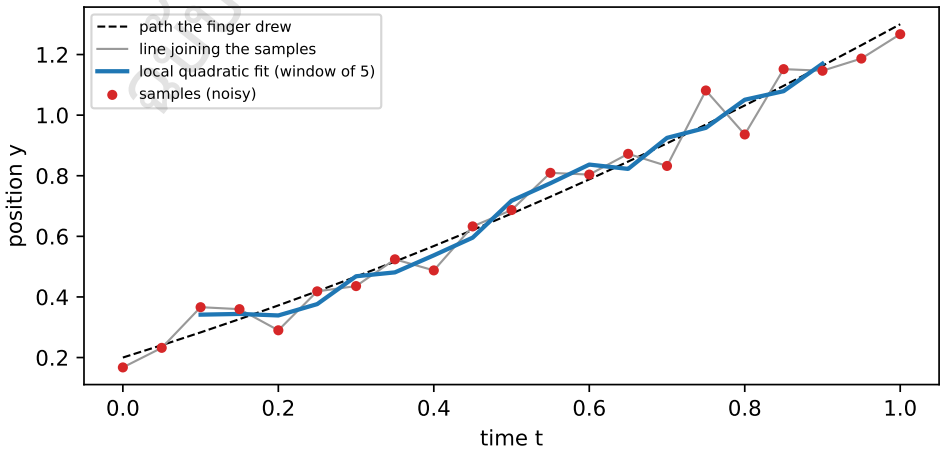
คุณคงเคยลากนิ้วบนจอแล้วเห็นเส้นที่ออกมาไม่เรียบ มือไม่นิ่งเสมอ และเซนเซอร์ที่บอกตำแหน่งนิ้วก็มีสัญญาณรบกวนเล็กน้อยด้วย สมมติว่าเราเก็บตำแหน่งนิ้วทุก 0.05 วินาทีตลอดหนึ่งวินาที ได้ 21 จุด ทางที่นิ้วลากจริงเป็นเส้นโค้งนุ่ม ๆ แต่ค่าที่เก็บได้ขึ้นลงรอบเส้นนั้นเล็กน้อย (ตัวเลขทั้งหมดมาจากโค้ดของเรา)

ถ้าเราต่อจุดเหล่านี้ด้วยเส้นตรงทีละคู่ เส้นที่ได้จะสั่นตามสัญญาณรบกวนทุกจุด เพราะการต่อจุดคือการผ่านทุกจุดพอดี บทที่ 3 ทำให้เราเห็นว่าทำได้ แต่จุดที่วัดได้ไม่ได้อยู่บนเส้นโค้งจริงพอดี การผ่านทุกจุดจึงเท่ากับเก็บความสั่นเข้าไปด้วย คำถามของบทนี้จึงเปลี่ยนไป จากการผ่านทุกจุด เป็นการหาเส้นที่อยู่ใกล้ข้อมูลทั้งหมดมากที่สุด เส้นแบบนั้นควรเป็นอย่างไร และคำนวณได้อย่างไร

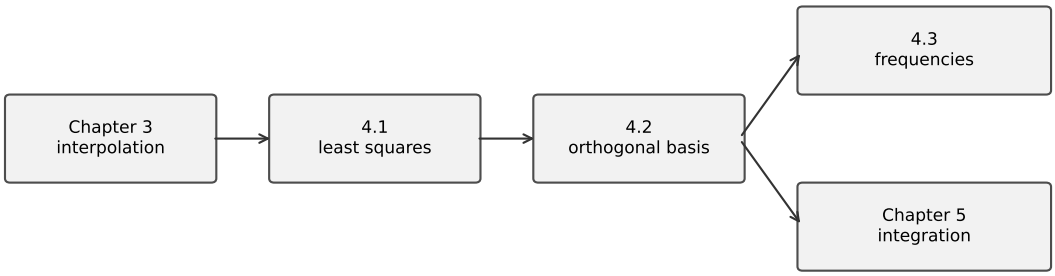
กราฟค่าชีพจรที่เซนเซอร์วัดทุกวินาทีก็มีปัญหาแบบเดียวกัน ค่าที่เห็นบนกราฟจะผันผวนรอบค่าที่แท้จริงเสมอ

ร่าง · Preprint · 2026-10-23

ลองดูรูปนี้ซึ่งเป็นลายเส้นที่เราสร้างขึ้นตามสถานการณ์ข้างบน จุดแดงคือค่าที่วัดได้ เส้นเทาที่ต่อจุดทุกจุดสั่นตามสัญญาณรบกวน การขึ้นลงรวมของเส้นนี้เท่ากับ 1.56 เท่าของการขึ้นลงของเส้นจริง (ตัวเลขจากโค้ดของเรา) ส่วนเส้นฟ้าคือหน้าต่างกำลังสองห้าจุดที่เราจะอธิบายในหัวข้อ 4.1 เส้นนี้เรียบกว่าและอยู่ใกล้เส้นจริงมากกว่าจุดที่วัดได้



รูปที่ 4.1: ลายเส้นที่วัดได้ (จุดแดง) เส้นเทาที่ต่อจุดทุกจุด เส้นฟ้าคือหน้าต่างกำลังสองห้าจุด และเส้นประคือเส้นทางที่นิ้วลากจริง (สร้างจาก `numanal/ch04.py`)



รูปที่ 4.2: ลำดับของหัวข้อในบทที่ 4 และการเชื่อมกับบทที่ 3 และบทที่ 5 (สร้างจาก `numanal/ch04.py`)

ภาพรวมบท

บทที่ 3 ปิดท้ายด้วยข้อสังเกตว่า การต่อจุดทุกจุดอาจเก็บความสั่นเข้าไปด้วย บทนี้ตอบคำถามนั้นจากสามมุม

- หัวข้อ 4.1: เมื่อข้อมูลมีสัญญาณรบกวน เราเลือกเส้นที่ทำให้ผลรวมกำลังสองของค่าตกค้างน้อยที่สุด นี่คือปัญหากำลังสองน้อยสุด และการทำให้เรียบด้วยหน้าต่างเลื่อน
- หัวข้อ 4.2: เมื่อเรารู้ฟังก์ชันทั้งหมด เช่น $\sin$ และต้องการพหุนามดีกรีต่ำที่ใกล้ฟังก์ชันนั้น การเลือกฐานเชิงตั้งฉากช่วยให้คำนวณแม่นขึ้นอย่างไร
- หัวข้อ 4.3: อีกด้านหนึ่งของปัญหาเดียวกัน ลายเส้นที่สั่นตามเวลาถูกมองเป็นผลรวมของความถี่หลายความถี่ คล้ายแถบของอีควอไลเซอร์ที่บอกว่ามีความถี่ใดอยู่ในเสียงบ้าง

เราเริ่มจากลายเส้นที่สั่น แล้วจะเจอ *อีกด้านหนึ่ง* ของปัญหาเดียวกันในหัวข้อ 4.3 คือเสียงที่ไม่ได้สั่นแบบสุ่มตามเวลา แต่ประกอบด้วยความถี่ที่ซ่อนอยู่

ภาพต่อไปนี้แสดงลำดับของหัวข้อในบทนี้ ลูกศรแสดงว่าหัวข้อหนึ่งใช้ผลของอีกหัวข้อหนึ่ง

4.1 เส้นตรงที่ดีที่สุด

สถานการณ์สมมติ: จุดสี่จุดที่ไม่อยู่บนเส้นเดียวกัน

สมมติว่าเราวัดตำแหน่งของวัตถุที่เคลื่อนที่เกือบเป็นเส้นตรงสี่ครั้ง ที่เวลา $t = 0, 1, 2, 3$ วินาที ได้ตำแหน่ง $y = 1.1,\ 2.9,\ 5.2,\ 6.8$ เมตร เส้นตรงไม่มีเส้นไหนผ่านทั้งสี่จุดพอดี เพราะจุดแรกกับจุดที่สองให้ความชัน 1.8 เมตรต่อวินาที แต่จุดที่สองกับจุดที่สามให้ความชัน 2.3 เมตรต่อวินาที เราจึงต้องเลือกเส้นหนึ่งเส้น และต้องบอกได้ว่าทำไมเส้นนั้นดีกว่าเส้นอื่น

ภาพรวมก่อนอ่าน

เส้นที่ดีควรอยู่ใกล้ทุกจุดพร้อมกัน เราวัดความห่างของแต่ละจุดจากเส้น ยกกำลังสอง แล้วรวมกัน เส้นที่ทำให้ผลรวมนี้น้อยที่สุดคือเส้นที่เราต้องการ คำตอบมาจากระบบสมการเชิงเส้นที่แก้ได้ด้วยมือ เมื่อมีตัวไม่ทราบค่าเพียงสองตัว

ฉบับร่าง · Preprint · 2026-10-23

ทำไมไม่ผ่านทุกจุด

พหุนามที่ผ่านทุกจุดพอดีเป็นสิ่งที่บทที่ 3 สร้างไว้แล้ว สำหรับจุดสี่จุดนี้ พหุนามดีกรีสามจะผ่านทั้งสี่จุด แต่พหุนามแบบนั้นจะหักเลี้ยวตามความสั่นของการวัด เมื่อข้อมูลมีความคลาดเคลื่อนเล็กน้อย การผ่านทุกจุดจึงเป็นการเก็บความคลาดเคลื่อนไว้ในผลลัพธ์ไม่ได้ลดมันลง

ดังนั้นเราต้องยอมให้เส้นไม่ผ่านจุดใดจุดหนึ่งเลย แต่ต้องมีเกณฑ์ที่บอกว่าเส้นไหนดี เกณฑ์นั้นคือความห่างรวมจากจุดทั้งหมด

นิยาม 4.1.1: ค่าตกค้างและปัญหากำลังสองน้อยสุด

ให้ข้อมูล (t_i, y_i) สำหรับ $i = 0, 1, \ldots, m-1$ และให้ตัวแบบเป็นเส้นตรง $p(t) = a_0 + a_1 t$ **ค่าตกค้าง** (residual) ของจุดที่ i คือ

$$r_i = y_i - p(t_i).$$

ปัญหากำลังสองน้อยสุด (least squares problem) คือการหา a_0 และ a_1 ที่ทำให้ผลรวมกำลังสองของค่าตกค้าง

$$S(a_0, a_1) = \sum_{i=0}^{m-1} r_i^2$$

มีค่าน้อยที่สุด

อ่านแบบเข้าใจง่าย

- ค่าตกค้างคือระยะที่เส้นพลาดที่จุดนั้น ถ้าเส้นผ่านจุดพอดี ค่านั้นเป็นศูนย์
- เราไม่รวม r_i ตรง ๆ เพราะค่าบวกกับค่าลบจะหักล้างกันได้ เส้นที่พลาดมากคนละทิศอาจได้ผลรวมเป็นศูนย์
- การยกกำลังสองทำให้จุดที่พลาดมากมีน้ำหนักมากขึ้นอย่างเร็ว และ S เป็นฟังก์ชันเรียบของ a_0 และ a_1 จึงหาจุดต่ำสุดได้ด้วยการตั้งอนุพันธ์เป็นศูนย์

ลองเอง: เส้น $y = 1 + 2t$

ลองใช้เส้น $y = 1 + 2t$ กับจุดสี่จุดข้างบน คำนวณค่าตกค้างทั้งสี่ตัว แล้วรวมกำลังสอง ผลที่ได้เท่าไร และเมื่อเทียบกับเส้นที่ดีที่สุดในข้อถัดไปแล้วได้ข้อสรุปอะไร

เส้น $y = 1 + 2t$ ให้ค่าตกค้าง $0.1, -0.1, 0.2, -0.2$ รวมกำลังสองได้ 0.10 ซึ่งมากกว่าเส้นที่ดีที่สุดที่เราจะหาต่อไป (ได้ 0.082) เส้นที่ดีที่สุดจึงต่างจาก $y = 1 + 2t$ เล็กน้อย

หาเส้นด้วยมือ

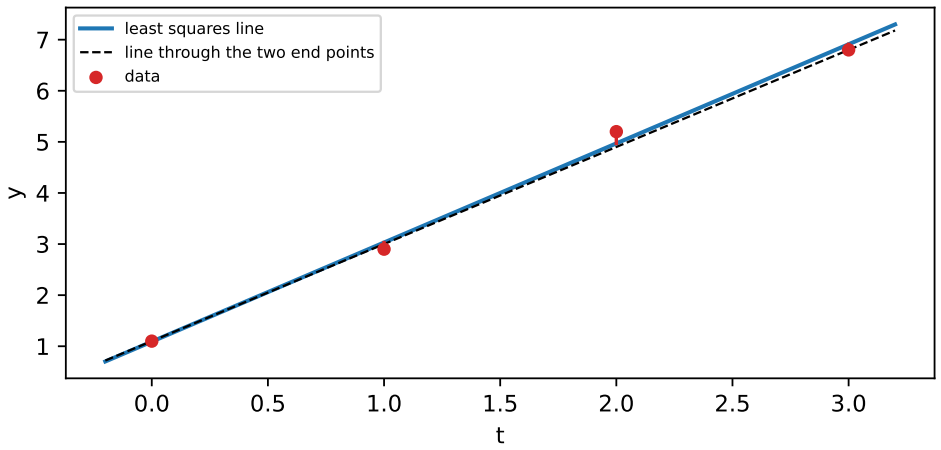
สำหรับเส้นตรง สูตรสั้น ๆ ได้จากการตั้งอนุพันธ์ของ S เป็นศูนย์ ซึ่งเราจะพิสูจน์ในทฤษฎีบทด้านล่าง ให้ $\bar{t}$ และ $\bar{y}$ เป็นค่าเฉลี่ยของ t_i และ y_i และให้

$$S_{tt} = \sum_i (t_i - \bar{t})^2, \qquad S_{ty} = \sum_i (t_i - \bar{t})(y_i - \bar{y}).$$

แล้วเส้นที่ดีที่สุดคือ $a_1 = S_{ty}/S_{tt}$ และ $a_0 = \bar{y} - a_1 \bar{t}$

สำหรับจุดสี่จุดของเรา $\bar{t} = 1.5$ และ $\bar{y} = 4.0$ ตารางต่อไปนี้คำนวณทีละจุด

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 4.3: จุดสี่จุด เส้นกำลังสองน้อยสุด (เส้นฟ้า) เส้นที่ผ่านจุดแรกและจุดสุดท้าย (เส้นประ) และค่าตกค้างของเส้นแรก (เส้นสีแดงแนวตั้ง) (สร้างจาก `numanal/ch04.py`)

i	t_i	$t_i - \bar{t}$	$y_i - \bar{y}$	$(t_i - \bar{t})(y_i - \bar{y})$
0	0	-1.5	-2.9	4.35
1	1	-0.5	-1.1	0.55
2	2	0.5	1.2	0.60
3	3	1.5	2.8	4.20

รวมคอลัมน์สุดท้ายได้ $S_{ty} = 9.7$ และ $S_{tt} = 2.25 + 0.25 + 0.25 + 2.25 = 5$ จึงได้

$$a_1 = \frac{9.7}{5} = 1.94, \qquad a_0 = 4.0 - 1.94 \cdot 1.5 = 1.09.$$

เส้นที่ดีที่สุดของเราจึงเป็น $p(t) = 1.09 + 1.94\,t$

ตารางต่อไปนี้ตรวจค่าตกค้างของเส้นนี้

t	y	$p(t)$	$r = y - p(t)$	r^2
0	1.1	1.09	0.01	0.0001
1	2.9	3.03	-0.13	0.0169
2	5.2	4.97	0.23	0.0529
3	6.8	6.91	-0.11	0.0121
			$\sum r = 0.00$	$\sum r^2 = 0.082$

ค่าตกค้างรวมได้ศูนย์ และ $\sum t_i r_i = 0 \cdot 0.01 + 1 \cdot (-0.13) + 2 \cdot 0.23 + 3 \cdot (-0.11) = 0$ เราจะเห็นว่าทำไมสองผลนี้ไม่บังเอิญ เส้นที่ผ่านจุดแรกและจุดสุดท้ายคือ $y = 1.1 + 1.9t$ ซึ่งให้ $\sum r^2 = 0.10$ เพราะพลาดที่จุดที่สองและสามมากกว่า

สมการปรกติ

ระบบสมการที่ให้เส้นดีที่สุดเรียกว่า **สมการปรกติ** (normal equations) ซึ่งเขียนได้ทั่วไปสำหรับตัวแบบที่เป็นผลรวมเชิงเส้นของฟังก์ชันพื้นฐานใด ๆ ให้เมทริกซ์ A มีแถวที่ i เป็น $(1, t_i)$ สำหรับเส้นตรง และให้ $b = (y_0, \ldots, y_{m-1})$ ค่าตกค้างคือ $r = b - Ax$ เมื่อ $x = (a_0, a_1)$

ฉบับร่าง · Preprint · 2026-10-23

นิยาม 4.1.2: คอลัมน์เป็นอิสระและเมทริกซ์ของตัวแบบ

เมทริกซ์ A ขนาด $m \times n$ มีคอลัมน์ที่**เป็นอิสระเชิงเส้น** (linearly independent) ถ้า $Av = 0$ เฉพาะเมื่อ $v = 0$ สำหรับเส้นตรงที่มีจุดต่างกันอย่างน้อยสองจุด คอลัมน์ 1 กับ t เป็นอิสระเชิงเส้นกัน

ทฤษฎีบท 4.1.1: สมการปรกติให้คำตอบของปัญหากำลังสองน้อยสุด

ให้ A เป็นเมทริกซ์ขนาด $m \times n$ ที่มีคอลัมน์เป็นอิสระเชิงเส้น และให้ $b \in \mathbb{R}^m$ จะได้ว่า

1. ระบบ $A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$ มีคำตอบเพียงตัวเดียว
2. เวกเตอร์ $x \in \mathbb{R}^n$ ทำให้ $\|b - Ax\|_2$ น้อยที่สุด ก็ต่อเมื่อ x เป็นคำตอบของระบบนี้
3. ที่คำตอบนั้น ค่าตกค้าง $r = b - Ax$ เชิงตั้งฉากกับทุกคอลัมน์ของ A นั่นคือ $A^{\mathsf{T}}r = 0$

อ่านแบบเข้าใจง่าย

- เงื่อนไขว่าคอลัมน์เป็นอิสระหมายความว่า ตัวแบบของเราไม่มี "ตัวซ้ำ" เช่น เส้นตรงจะใช้ทั้ง 1 และ t ได้ก็ต่อเมื่อ t ไม่ได้เป็นค่าคงที่คูณกับ 1
- ข้อ 3 บอกว่าที่คำตอบ ค่าตกค้างไม่เหลือส่วนที่ตัวแบบยังอธิบายได้ เราลดค่าตกค้างต่อไปอีกไม่ได้โดยการปรับ a_0 หรือ a_1
- สำหรับเส้นตรง ข้อ 3 เท่ากับ $\sum r_i = 0$ และ $\sum t_i r_i = 0$ ซึ่งเราเห็นแล้วในตารางข้างบน

แนวคิด. แยกเวกเตอร์ $b - Ax$ เป็นสองส่วนที่เชิงตั้งฉากกัน แล้วใช้ทฤษฎีพีทาโกรัสกับเวกเตอร์ของส่วนที่เชิงตั้งฉากนั้น

1. $A^{\mathsf{T}}A$ **ไม่เอกฐาน.** ถ้า $A^{\mathsf{T}}Av = 0$ แล้ว $\|Av\|_2^2 = v^{\mathsf{T}}A^{\mathsf{T}}Av = 0$ จึง $Av = 0$ และจากความเป็นอิสระของคอลัมน์ได้ $v = 0$ ดังนั้น $A^{\mathsf{T}}A$ ไม่เอกฐาน เมทริกซ์จึงมีอินเวอร์ส และระบบ $A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$ มีคำตอบเดียว x^* นี่คือข้อ 1

2. **ค่าตกค้างที่** x^* **เชิงตั้งฉาก.** ให้ $r^* = b - Ax^*$ จากสมการปรกติ $A^{\mathsf{T}}r^* = A^{\mathsf{T}}b - A^{\mathsf{T}}Ax^* = 0$ นี่คือข้อ 3 เมื่อ x^* เป็นคำตอบ

3. **การแยกสองส่วน.** สำหรับ $x \in \mathbb{R}^n$ ใด ๆ ให้ $d = x^* - x$ จะได้ $b - Ax = r^* + Ad$ และ

$$\|b - Ax\|_2^2 = \|r^*\|_2^2 + 2\, d^{\mathsf{T}}A^{\mathsf{T}}r^* + \|Ad\|_2^2 = \|r^*\|_2^2 + \|Ad\|_2^2,$$

   เพราะ $A^{\mathsf{T}}r^* = 0$ ทำให้พจน์ไขว้หายไป

4. x^* **เป็นตัวที่ดีที่สุดและเพียงตัวเดียว.** จากขั้น 3 ได้ $\|b - Ax\|_2 \geq \|r^*\|_2$ และเท่ากันก็ต่อเมื่อ $Ad = 0$ เมื่อคอลัมน์เป็นอิสระ $Ad = 0$ ก็ต่อเมื่อ $d = 0$ ดังนั้น $x = x^*$ นี่คือส่วน "ถ้า" ของข้อ 2 และแสดงว่า x^* เป็นตัวที่น้อยที่สุดเพียงตัวเดียว

5. **ส่วน "เฉพาะถ้า" ของข้อ 2.** ให้ x ทำให้น้อยที่สุด และให้ $z \in \mathbb{R}^n$ ใด ๆ จะได้ $g(s) = \|b - A(x + sz)\|_2^2 = \|b - Ax\|_2^2 - 2s\, z^{\mathsf{T}}A^{\mathsf{T}}(b - Ax) + s^2\|Az\|_2^2$ ฟังก์ชันนี้มีค่าต่ำสุดที่ $s = 0$ จึงต้องได้สัมประสิทธิ์ของ s เป็นศูนย์ มิฉะนั้นจะเลือก s เล็ก ๆ ที่มีเครื่องหมายเหมาะแล้วทำให้ $g(s) < g(0)$ ดังนั้น $z^{\mathsf{T}}A^{\mathsf{T}}(b - Ax) = 0$ สำหรับทุก z และได้ $A^{\mathsf{T}}(b - Ax) = 0$ หรือ $A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$ ซึ่งแปลว่า x เป็นคำตอบของสมการปรกติ

ฉบับร่าง · Preprint · 2026-10-23

สรุป: สมการปรกติคือเงื่อนไขที่แยกคำตอบที่ดีที่สุดออกมา และเพราะคอลัมน์เป็นอิสระ คำตอบนั้นมีเพียงตัวเดียว

ข้อควรระวัง: ถ้าคอลัมน์พึ่งพากัน

ถ้าคอลัมน์ของ A พึ่งพากัน เช่น ใช้ทั้ง 1 และ $2 \cdot 1$ เป็นฐาน แล้วมี $v \neq 0$ ที่ $Av = 0$ จะได้ว่า x กับ $x + v$ ให้ค่าตกค้างเท่ากันเสมอ นั่นคือ ตัวทำให้น้อยที่สุดไม่เป็นหนึ่งเดียว แม้จะยังมีค่าน้อยที่สุดอยู่ การแก้สมการปรกติในกรณีนี้จึงไม่ให้คำตอบเดียว

ลองตรวจสมการปรกติกับจุดสี่จุด ได้ $A^{\mathsf{T}}A = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}$ และ $A^{\mathsf{T}}b = \begin{bmatrix} 16 \\ 33.7 \end{bmatrix}$ เมื่อแทน $a_0 = 1.09$ และ $a_1 = 1.94$ ได้ $4(1.09) + 6(1.94) = 16.00$ และ $6(1.09) + 14(1.94) = 33.70$ ถูกต้องทั้งสองสมการ ตัวกำหนดของ $A^{\mathsf{T}}A$ คือ $4 \cdot 14 - 6 \cdot 6 = 20$ ไม่เป็นศูนย์ ซึ่งตรงกับข้อ 1 ของทฤษฎีบท

ลองเพิ่มดีกรี: เส้นโค้งกำลังสองด้วยมือ

ตัวแบบอาจเป็นพหุนามกำลังสอง $p(t) = a_0 + a_1 t + a_2 t^2$ ก็ได้ ขั้นตอนเหมือนเดิม เพียงแต่มีคอลัมน์ $1, t, t^2$ จึงได้สมการปรกติขนาด 3×3 สำหรับจุดสี่จุดเดิม

$$\begin{bmatrix} 4 & 6 & 14 \\ 6 & 14 & 36 \\ 14 & 36 & 98 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 16 \\ 33.7 \\ 84.9 \end{bmatrix}.$$

ตัวเลขในเมทริกซ์คือ $\sum t_i^{j+k}$ และตัวเลขทางขวาคือ $\sum t_i^j y_i$ เช่น $\sum t_i^2 y_i = 0 + 2.9 + 4(5.2) + 9(6.8) = 84.9$ ตัวเลขเหล่านี้คำนวณจาก $t = 0, 1, 2, 3$ และค่า y เดิม

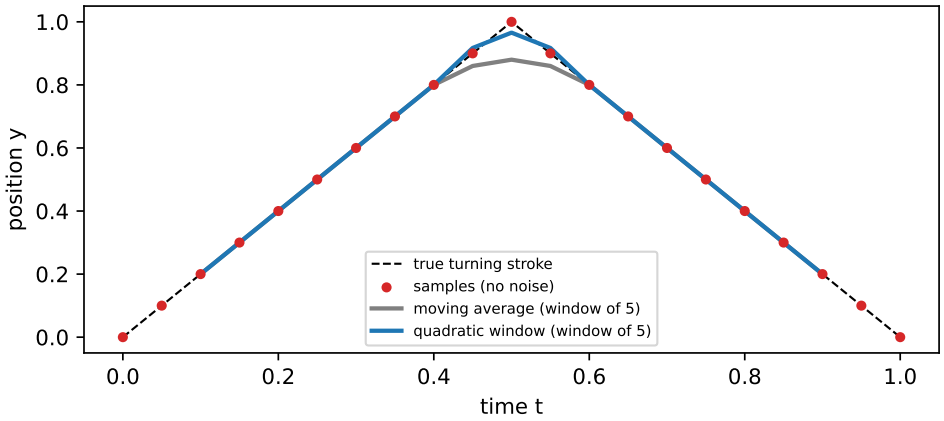
1. ลบ $\frac{6}{4}$ เท่าของแถวแรกออกจากแถวที่สอง ได้ $5a_1 + 15a_2 = 9.7$
2. ลบ $\frac{14}{4}$ เท่าของแถวแรกออกจากแถวที่สาม ได้ $15a_1 + 49a_2 = 28.9$
3. ลบ 3 เท่าของสมการที่ได้ในข้อ 1 ออกจากสมการที่ได้ในข้อ 2 ได้ $4a_2 = -0.2$ จึง $a_2 = -0.05$
4. แทน a_2 ในสมการข้อ 1 ได้ $5a_1 = 9.7 + 0.75$ จึง $a_1 = 2.09$ และจากแถวแรก $4a_0 = 16 - 6(2.09) + 0.7 = 4.16$ จึง $a_0 = 1.04$

ได้ $p(t) = 1.04 + 2.09t - 0.05t^2$ ค่าตกค้างคือ $0.06, -0.18, 0.18, -0.06$ และผลรวมกำลังสองเป็น 0.072 ซึ่งน้อยกว่า 0.082 ของเส้นตรง ผลนี้ไม่บังเอิญ เพราะเส้นตรงเป็นกรณีพิเศษของกำลังสอง (ตั้ง $a_2 = 0$) ตัวแบบที่กว้างกว่าจึงไม่มีทางให้ผลรวมที่น้อยที่สุดสูงกว่าตัวแบบที่แคบกว่า และดีกรีสามที่ผ่านจุดทั้งสี่พอดีให้ผลรวมเป็นศูนย์ ข้อนี้บอกว่า ผลรวมที่เล็กลงไม่ได้แปลว่าตัวแบบดีกว่าสำหรับข้อมูลใหม่

หน้าต่างเลื่อน: ทำให้ลายเส้นเรียบโดยไม่ลากเส้นเดียว

ถ้าลายเส้นยาวและโค้ง เส้นตรงเส้นเดียวจะไม่พอ เราจึงทำเป็นท้องถิ่น แนวคิดคือ ที่จุดหนึ่ง ดูจุดข้างเคียงห้าจุด คือสองจุดซ้ายและสองจุดขวา หาพหุนามกำลังสองที่ใกล้ห้าจุดนี้ที่สุดด้วยกำลังสองน้อยสุด แล้วใช้ค่าของพหุนามที่จุดกลางเป็นค่าที่เรียบ จากนั้นเลื่อนหน้าต่างไปจุดถัดไป นี่คือ **หน้าต่างเลื่อน** (sliding window) ซึ่งใช้แนวคิดเดียวกับการทำให้เรียบและการหาอนุพันธ์จากข้อมูลที่ตำแหน่งเท่ากัน (แนวคิดจาก Savitzky and Golay 1964)

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 4.4: ลายเส้นที่กลับทิศ (เส้นประ) ค่าเฉลี่ยเคลื่อนที่ (เส้นเทา) และหน้าต่างกำลังสองของห้าจุด (เส้นฟ้า) ยอดของเส้นฟ้าสูงกว่า (สร้างจาก `numanal/ch04.py`)

นิยาม 4.1.3: น้ำหนักของหน้าต่าง

ให้จุดในหน้าต่างอยู่ที่ตำแหน่งเท่ากันทั้งหมด **น้ำหนักของหน้าต่าง** (window weights) คือเลข w_k ที่ทำให้ค่าที่จุดกลางของพหุนามกำลังสองน้อยสุด เท่ากับ $\sum_k w_k y_k$ น้ำหนักเหล่านี้ขึ้นกับตำแหน่งของจุดในหน้าต่างเท่านั้น ไม่ขึ้นกับค่า y_k

ในทางปฏิบัติ น้ำหนักหาได้จากสมการปรกติของหน้าต่างนั้น คำตอบของสมการปรกติให้สัมประสิทธิ์ของพหุนาม และค่าคงที่ของพหุนามคือค่าที่จุดกลาง เมื่อคำนวณน้ำหนักทั้งห้าตัวออกมาจะได้

$$(w_{-2}, \ldots, w_2) = \left(-\tfrac{3}{35},\ \tfrac{12}{35},\ \tfrac{17}{35},\ \tfrac{12}{35},\ -\tfrac{3}{35}\right).$$

ถ้าเราใช้เส้นตรงแทนพหุนามกำลังสอง น้ำหนักทั้งห้าเท่ากับ $\frac{1}{5}$ ทุกตัว นั่นคือค่าเฉลี่ยเคลื่อนที่ (moving average) ที่ใครก็เคยใช้

ทำไมกำลังสองจึงดีกว่า ลองดูยอดแหลม ถ้าข้อมูลห้าจุดเป็น $0, 0, 1, 0, 0$ ค่าเฉลี่ยเคลื่อนที่ที่จุดกลางได้ $\frac{1}{5} = 0.2$ แต่หน้าต่างกำลังสองได้ $\frac{17}{35} \approx 0.486$ ยอดแหลมยังคงอยู่ได้มากกว่ามาก

ลายเส้นที่กลับทิศให้ภาพที่ชัดกว่า ลองดูลายเส้นรูปสามเหลี่ยมที่ยอดอยู่ที่ $t = 0.5$ สูง 1 โดยไม่มีสัญญาณรบกวน (ข้อมูลสร้างขึ้น) ค่าเฉลี่ยเคลื่อนที่ที่ยอดได้ 0.88 ส่วนหน้าต่างกำลังสองได้ $\frac{33.8}{35} \approx 0.966$ ดังรูปต่อไปนี้

ลองเอง: หน้าต่างกับพหุนามกำลังสอง

ตรวจว่าเมื่อ $y = (1, 4, 9, 16, 25)$ ค่าที่จุดกลางเท่ากับ 9 พอดี แล้วลองเปลี่ยนจุดกลางจาก 9 เป็น 10 ค่าที่จุดกลางจะเป็นเท่าไร (ใช้น้ำหนักข้างบน)

เมื่อ $y = (1, 4, 9, 16, 25)$ จะได้ $\frac{-3\cdot1+12\cdot4+17\cdot9+12\cdot16-3\cdot25}{35} = \frac{315}{35} = 9$ เพราะค่าทั้งห้าเป็นพหุนามกำลังสองของตำแหน่งอยู่แล้ว ส่วนเมื่อเปลี่ยนจุดกลางเป็น 10 ได้ $\frac{332}{35} \approx 9.486$ ค่าเปลี่ยนไปเพียงราว 0.49 และเกิดจากน้ำหนัก $\frac{17}{35}$ ของจุดกลางเท่านั้น

ข้อควรระวัง: หน้าต่างที่ปลายและหน้าต่างที่กว้างเกินไป

ที่ปลายของลายเส้นไม่มีจุดข้างซ้ายหรือขวาครบห้าจุด จึงคำนวณค่าเรียบไม่ได้ ในโค้ดของเราจุดเหล่านี้เก็บเป็นค่าว่าง (NaN) และหน้าต่างที่กว้างเกินไปจะทำให้มุมหักของลายเส้นหายไป เพราะหน้าต่างเห็นทั้งสองด้านของมุม จึงต้องเลือกความกว้างให้สั้นพอเทียบกับความโค้งที่เราต้องการเก็บไว้

ร่างต้น · preprint · 2026-10-23

เกร็ดเพิ่มเติม

ในระบบโต้ตอบ เช่นการลากเมาส์หรือนิ้วบนจอ มีวิธีทำให้เรียบอีกแบบที่ปรับความถี่ตัดของตัวกรองตามความเร็วของการเคลื่อนที่ บทความของ Casiez, Roussel และ Vogel เสนอ 1€ filter ซึ่งใช้ตัวกรองความถี่ต่ำอันดับหนึ่งและเพิ่มความถี่ตัดเมื่อสัญญาณเคลื่อนเร็วขึ้น เพื่อลดความหน่วง Casiez, Roussel, and Vogel 2012 เราไม่ลงรายละเอียดอัลกอริทึมนี้ เพราะหัวข้อนี้เน้นหน้าต่างกำลังสองน้อยสุด

ในโลกจริง: ความเข้มข้นของ CO_2 ที่มอนาโลอา

ตั้งแต่เดือนมีนาคม ค.ศ. 1958 มีการวัดความเข้มข้นของคาร์บอนไดออกไซด์ในอากาศที่หอสังเกตการณ์มอนาโลอา ฮาวาย บทความปี 1976 Keeling et al. 1976 เป็นเรื่องความแปรผันของ CO_2 ที่หอสังเกตการณ์นี้ ไฟล์ค่าเฉลี่ยรายเดือนของ NOAA ที่เราเก็บไว้มี 823 เดือน ตั้งแต่มีนาคม 1958 ถึงกันยายน 2026 หน่วยเป็น ppm (ส่วนในล้านส่วน) NOAA Global Monitoring Laboratory 2026 ไฟล์ระบุว่าการวัดช่วงธันวาคม 2022 ถึงกรกฎาคม 2023 ทำที่ Maunakea เพราะการวัดที่มอนาโลอาถูกระงับชั่วคราว

ถ้าลากเส้นตรงเส้นเดียวให้ทั้งชุดด้วยกำลังสองน้อยสุด ได้ความชันประมาณ 1.67 ppm ต่อปี แต่ถ้าดูเฉพาะปี 1958 ถึง 1979 ความชันประมาณ 1.01 ppm ต่อปี และตั้งแต่ปี 2016 ถึงกลางปี 2026 ความชันประมาณ 2.53 ppm ต่อปี ความชันเปลี่ยนไปมาก ตัวเลขทั้งหมดเป็นการคำนวณของเราจากไฟล์ (ดู `code/data/README.md`)

ข้อควรระวัง: ค่าตกค้างน้อยไม่ได้แปลว่าตัวแบบถูก

ค่าตกค้างรวมที่เล็กบอกเพียงว่าเส้นใกล้ข้อมูลที่เรามี ไม่ได้บอกว่าเส้นนั้นอธิบายการเคลื่อนที่จริงได้ดี ลายเส้นโค้งจะมีค่าตกค้างเล็กพอสมควรเมื่อใช้เส้นตรง แต่ความโค้งทั้งหมดหายไป และข้อมูลที่ยาวมากอาจมีความชันเปลี่ยนที่ไม่เห็นจากเส้นตรงเส้นเดียว เลือกตัวแบบให้ตรงกับสิ่งที่เรารู้เกี่ยวกับปรากฏการณ์ก่อนตรวจค่าตกค้าง

ข้อควรระวัง: จุดผิดปกติเดียวดึงเส้นได้มาก

เพราะเรายกกำลังสอง จุดที่พลาดมากมีอิทธิพลสูง ถ้าเปลี่ยนจุดที่ $t = 3$ จาก 6.8 เป็น 9.8 ซึ่งคลาดเคลื่อนเพียง 3 เมตร ความชันจะเปลี่ยนจาก 1.94 เป็น 2.84 เพิ่มขึ้นราว 46 เปอร์เซ็นต์ นอกจากนี้ สมการปรกติมีคอนดิชันนัมเบอร์เป็นกำลังสองของคอนดิชันนัมเบอร์ของ A ตัวอย่างเช่น พหุนามกำลังสามบนจุด $t = 0, \ldots, 10$ มีคอนดิชันนัมเบอร์ 1617 แต่ $A^{\mathsf{T}}A$ มีประมาณ 2.6×10^{6} การแก้ระบบนี้ด้วยทศนิยมจำกัดจึงเสียความแม่นมากกว่า บทที่ 8 จะแสดงวิธีที่หลีกเลี่ยงการยกกำลังสองนี้

สรุปใจความ

- เมื่อข้อมูลมีสัญญาณรบกวน เราไม่ควรผ่านทุกจุด แต่หาเส้นที่ทำให้ผลรวมกำลังสองของค่าตกค้างน้อยที่สุด
- คำตอบคือระบบสมการปรกติ $A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$ และถ้าคอลัมน์ของ A เป็นอิสระเชิงเส้น คำตอบมีเพียงตัวเดียว
- ที่คำตอบ ค่าตกค้างเชิงตั้งฉากกับทุกคอลัมน์ของตัวแบบ ในกรณีเส้นตรง ผลรวมของค่าตกค้างและผลรวมของ $t_i r_i$ เป็นศูนย์
- หน้าต่างเลื่อนกับพหุนามกำลังสองทำให้ลายเส้นเรียบโดยรักษายอดไว้มากกว่าค่าเฉลี่ยเคลื่อนที่ น้ำหนักของมันได้จากปัญหาเดียวกัน

ฉบับร่าง · Preprint · 2026-10-25

- กลับไปที่ลายเส้นที่สั่น: เส้นที่ไม่ผ่านทุกจุดแต่ใกล้ข้อมูลทั้งหมดที่สุดคือคำตอบ และถ้าจุดบางจุดผิดปกติมาก ต้องระวังเพราะมันดึงเส้นได้มาก

อ่านต่อ. ในต้นฉบับของ Atkinson ปัญหากำลังสองน้อยสุดถูกนิยามด้วยปริพันธ์บนช่วง และชี้ว่าระบบสมการปรกติของปัญหานั้นมีเมทริกซ์ฮิลเบิร์ตซึ่งไวต่อการเปลี่ยนแปลงเล็กน้อยมาก Atkinson 1989, §4.3, pp. 204–207 ส่วนพื้นฐานของเมทริกซ์บวกแน่นอนและผลคูณภายในอยู่ใน Iserles 2009, ภาคผนวก A.1.3, pp. 432–434 หัวข้อถัดไปถามว่า ถ้าต้องประมาณฟังก์ชันที่รู้ทั้งหมดด้วยพหุนามดีกรีต่ำ ควรเลือกฐานอย่างไร

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 4.1.1 (⋆) ให้จุด $(0,2),\ (1,3),\ (2,5)$

1. จงหาเส้นตรง $p(t)=a_0+a_1t$ ที่ทำให้ผลรวมกำลังสองของค่าตกค้างน้อยที่สุด โดยใช้ S_{tt} และ S_{ty}
2. จงหาค่าตกค้างของจุดทั้งสาม และผลรวมกำลังสองของค่าตกค้าง

แบบฝึกหัด 4.1.2 (⋆⋆) ให้ตัวแบบเป็นพหุนาม $p(t)=a_0+a_1t+\cdots+a_nt^n$ ซึ่งมีพจน์ค่าคงที่ จงแสดงว่า ที่ตัวที่ทำให้ผลรวมกำลังสองน้อยที่สุด ผลรวมของค่าตกค้าง $\sum_i r_i$ เป็นศูนย์

ตัวช่วยคิด

คอลัมน์แรกของ A เป็นเวกเตอร์ที่มีแต่ 1 ลองดูแถวแรกของ $A^{\mathsf{T}}r=0$

แบบฝึกหัด 4.1.3 (⋆⋆) ใช้น้ำหนักหน้าต่างเจ็ดจุดกำลังสอง $\frac{1}{21}(-2,\,3,\,6,\,7,\,6,\,3,\,-2)$ กับข้อมูล $y_k=t_k^2$ ที่ $t_k=0,1,\dots,6$

1. จงหาค่าที่จุดกลาง ($t=3$) ของหน้าต่างนี้
2. สมมติว่าค่าที่จุดกลางคลาดเคลื่อนเป็น $y_3=12$ แทน 9 จงหาค่าที่จุดกลางใหม่ และอธิบายว่าการเปลี่ยนค่านี้มีผลต่อผลลัพธ์เท่าไร
3. ถ้าค่า y_6 (ที่ตำแหน่งขวาสุด มีน้ำหนัก $-\frac{2}{21}$) เปลี่ยนไป 3 หน่วย ค่าที่จุดกลางจะเปลี่ยนเท่าไร เปรียบเทียบกับข้อ (ข)

แบบฝึกหัด 4.1.4 (⋆⋆) ให้จุดสี่จุดของหัวข้อ 4.1 และให้ y_i เปลี่ยนไปเป็น $y_i+\delta$ โดยจุดอื่นไม่เปลี่ยน

1. จงแสดงว่าความชันของเส้นกำลังสองน้อยสุดเปลี่ยนไป $(t_i-\bar t)\,\delta/S_{tt}$ (ใช้ $S_{ty}=\sum_j(t_j-\bar t)y_j$)
2. จงหา δ ที่ทำให้ความชันเปลี่ยนจาก 1.94 เป็น 2.04 เมื่อเปลี่ยนจุด $t_3=3$
3. จุดใดมีอิทธิพลต่อความชันมากที่สุด และทำไมจุดที่อยู่กลางข้อมูลจึงมีอิทธิพลน้อย

แบบฝึกหัด 4.1.5 (⋆⋆) ใช้ไฟล์ค่าเฉลี่ยรายเดือนของ $\mathrm{CO_2}$ ที่มอนาโลอาในโฟลเดอร์ `code/data` (ดูที่มาใน `code/data/README.md`)

1. จงหาความชันของเส้นกำลังสองน้อยสุดของข้อมูลทั้งหมด เป็น ppm ต่อปี
2. จงหาความชันของข้อมูลเฉพาะปี 1958 ถึง 1979 และของข้อมูลตั้งแต่ปี 2016 ถึงกลางปี 2026

ฉบับร่าง · Preprint · 2026-10-23

3. เส้นเดียวสำหรับทั้งชุดบอกอะไรเกี่ยวกับการเปลี่ยนแปลงของ CO_2 และไม่บอกอะไร

แบบฝึกหัด 4.1.6 (★★★) ให้ A เป็นเมทริกซ์ $m \times n$ ที่มีคอลัมน์เป็นอิสระเชิงเส้น จงแสดงว่า $A^{\mathsf{T}}A$ สมมาตร และ $v^{\mathsf{T}}A^{\mathsf{T}}Av > 0$ สำหรับทุก $v \neq 0$ จากนั้นอธิบายว่าเหตุใดค่าลักษณะเฉพาะทั้งหมดของ $A^{\mathsf{T}}A$ จึงเป็นบวก

ตัวช่วยคิด

เขียน $v^{\mathsf{T}}A^{\mathsf{T}}Av$ เป็น $\|Av\|_2^2$ แล้วใช้ความอิสระของคอลัมน์

แบบฝึกหัด 4.1.7 (★ ★ ★) ให้ A เป็นเมทริกซ์แวนเดอร์มอนด์ของพหุนามดีกรีไม่เกิน 5 บนจุด $t = 0, 1, \ldots, 10$ (คอลัมน์ $1, t, \ldots, t^5$)

1. จงคำนวณคอนดิชันนัมเบอร์ (condition number) ของ A และของ $A^{\mathsf{T}}A$ ด้วยโค้ด
2. จงตรวจว่า $\operatorname{cond}(A^{\mathsf{T}}A)$ มีค่าเป็นกำลังสองของ $\operatorname{cond}(A)$ โดยประมาณ
3. ถ้าเครื่องเก็บทศนิยมที่มีความแม่นสัมพัทธ์ราว 6×10^{-8} (ความแม่นเดี่ยว) ให้ขอบประมาณของความแม่นของคำตอบเป็นเท่าไรสำหรับทั้งสองระบบ และอ่านผลนี้ว่าอย่างไร

ตัวช่วยคิด

ขอบประมาณของความแม่นสัมพัทธ์ของคำตอบมีลำดับ $\operatorname{cond} \cdot u$ โดย u คือความแม่นสัมพัทธ์ของเครื่อง ตัวเลขนี้เป็นขอบคร่าว ๆ ไม่ใช่ค่าที่แน่นอน

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 4.1.1.

1. จุดทั้งสามมี $\bar{t} = \frac{0+1+2}{3} = 1$ และ $\bar{y} = \frac{2+3+5}{3} = \frac{10}{3}$ จึงได้ $S_{tt} = (-1)^2 + 0^2 + 1^2 = 2$ และ $S_{ty} = (-1)\big(2 - \frac{10}{3}\big) + 0 + 1 \cdot \big(5 - \frac{10}{3}\big) = \frac{4}{3} + \frac{5}{3} = 3$ ดังนั้น $a_1 = \frac{3}{2}$ และ $a_0 = \bar{y} - a_1\bar{t} = \frac{10}{3} - \frac{3}{2} = \frac{11}{6}$ เส้นคือ $p(t) = \frac{11}{6} + \frac{3}{2}t$ ตรวจได้จากสมการปรกติ $\begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} x = \begin{bmatrix} 10 \\ 13 \end{bmatrix}$ ซึ่งให้ $a_1 = \frac{3}{2}$ และ $a_0 = \frac{11}{6}$ เช่นกัน
2. ค่าตกค้างคือ $r_0 = 2 - \frac{11}{6} = \frac{1}{6}$, $r_1 = 3 - \frac{20}{6} = -\frac{1}{3}$ และ $r_2 = 5 - \frac{29}{6} = \frac{1}{6}$ ผลรวมเป็นศูนย์ตามที่ทฤษฎีบทให้ไว้ และ
$$S = \tfrac{1}{36} + \tfrac{1}{9} + \tfrac{1}{36} = \tfrac{6}{36} = \tfrac{1}{6}.$$

เฉลย 4.1.2. แนวคิดคือ คอลัมน์แรกของ A เป็นเวกเตอร์ $\mathbf{1}$ เพราะ a_0 คูณกับ 1 ทุกแถว ทฤษฎีบทให้ $A^{\mathsf{T}}r = 0$ ที่คำตอบ ดังนั้นแถวแรกของ $A^{\mathsf{T}}r = 0$ คือ $\mathbf{1}^{\mathsf{T}}r = \sum_i r_i = 0$ เท่านั้น ไม่ต้องใช้ดีกรีของพหุนามเลย การตรวจด้วยโค้ดกับข้อมูลสุ่มที่มีเมล็ดคงที่ให้ผลรวมที่เล็กกว่า 10^{-12} สำหรับดีกรี 1, 2 และ 3 (ผลรวมนี้เป็นศูนย์ในทางคณิตศาสตร์ ค่าที่เหลือเป็นความคลาดเคลื่อนของการคำนวณ)

ฉบับร่าง Preprint · 2026-10-23

เฉลย 4.1.3. น้ำหนัก $(-2,3,6,7,6,3,-2)/21$ รวมกันได้ $21/21 = 1$ จึงให้ค่าคงที่ได้ตรง

1. ค่าที่จุดกลางคือ
$$\frac{-2\cdot 0 + 3\cdot 1 + 6\cdot 4 + 7\cdot 9 + 6\cdot 16 + 3\cdot 25 - 2\cdot 36}{21} = \frac{189}{21} = 9.$$
ค่านี้ตรงกับ t^2 ที่ $t = 3$ พอดี เพราะ $y_k = t_k^2$ เป็นพหุนามกำลังสองและน้ำหนักทำให้ค่าพหุนามกำลังสองที่จุดกลางได้ตรง

2. เมื่อ y_3 เปลี่ยนจาก 9 เป็น 12 การเปลี่ยนแปลงคือ $\delta = 3$ และค่าที่จุดกลางเปลี่ยนไป $w_0\delta = \frac{7}{21}\cdot 3 = 1$ ได้ค่าใหม่ 10

3. เมื่อ y_6 เปลี่ยนไป 3 หน่วย น้ำหนักของจุดนี้คือ $-\frac{2}{21}$ ค่าที่จุดกลางเปลี่ยนไป $-\frac{2}{21}\cdot 3 = -\frac{2}{7} \approx -0.286$ เทียบกับข้อ (ข) จุดที่อยู่ใกล้จุดกลางมากกว่ามีน้ำหนักมากกว่า และจุดที่อยู่ไกลกว่ามีผลน้อยกว่า ในข้อ (ค) ยังมีเครื่องหมายลบด้วย

เฉลย 4.1.4.

1. เมื่อ y_i เปลี่ยนเป็น $y_i + \delta$ ค่าเฉลี่ย $\bar{y}$ เปลี่ยนเป็น $\bar{y} + \delta/m$ แต่ $\sum_j (t_j - \bar{t}) = 0$ จึงทำให้ $S_{ty} = \sum_j (t_j - \bar{t})y_j$ เปลี่ยนเพียง $(t_i - \bar{t})\delta$ และ S_{tt} ไม่เปลี่ยน ความชัน $a_1 = S_{ty}/S_{tt}$ จึงเปลี่ยนเป็น $(t_i - \bar{t})\delta/S_{tt}$

2. ที่ $t_3 = 3$ มี $t_3 - \bar{t} = 1.5$ และ $S_{tt} = 5$ ดังนั้น $0.3\,\delta = 0.10$ นั่นคือ $\delta = \frac{1}{3}$ จุดนั้นจึงควรอยู่ที่ประมาณ $y_3 = 6.8 + \frac{1}{3} \approx 7.13$ จึงจะได้ความชัน 2.04 ตรวจได้ทันที

3. จุดที่มี $|t_i - \bar{t}|$ มากที่สุดมีอิทธิพลมากที่สุด ในจุดสี่จุดนี้คือ $t = 0$ และ $t = 3$ ซึ่งห่างจาก $\bar{t}$ อยู่ 1.5 ส่วนจุดกลาง $t = 1.5$ ไม่มีจุดข้อมูล แต่จุด $t = 1$ หรือ $t = 2$ ซึ่งอยู่ใกล้ $\bar{t}$ มีอิทธิพลเพียง $0.5/5 = 0.1$ ต่อหน่วยของ δ จึงน้อยกว่าครึ่งหนึ่งของจุดที่ปลาย

เฉลย 4.1.5. เราคำนวณด้วยฟังก์ชัน `load_co2` และ `co2_trend` ในโมดูลของบท เพราะข้อมูลมี 823 จุด การทำด้วยมือจะยาวเกินไป ส่วนสูตรเหมือนกับในหัวข้อ 4.1 คือ $a_1 = S_{ty}/S_{tt}$ โดยใช้ทศนิยมปี

1. ข้อมูลทั้งหมด: $a_1 \approx 1.671$ ppm ต่อปี

2. ปี 1958 ถึง 1979: $a_1 \approx 1.007$ ppm ต่อปี และตั้งแต่ปี 2016 ถึงกลางปี 2026: $a_1 \approx 2.528$ ppm ต่อปี

3. เส้นเดียวให้อัตราเฉลี่ยทั้งช่วง 1.67 ppm ต่อปี แต่ไม่บอกว่าอัตรานี้เปลี่ยนไป ช่วงแรกเร็วราว 1.0 และช่วงหลังเร็วราว 2.5 ดังนั้นเส้นเดียวจึงซ่อนการเปลี่ยนแปลงของความชันไว้ ถ้าต้องการเห็นการเปลี่ยน ต้องดูหน้าต่างช่วงสั้นกว่า แบบเดียวกับหน้าต่างเลื่อนในหัวข้อนี้

เฉลย 4.1.6.

1. สมมาตร เพราะ $(A^{\mathsf{T}}A)^{\mathsf{T}} = A^{\mathsf{T}}(A^{\mathsf{T}})^{\mathsf{T}} = A^{\mathsf{T}}A$

2. สำหรับ $v \neq 0$ ได้ $v^{\mathsf{T}}A^{\mathsf{T}}Av = (Av)^{\mathsf{T}}(Av) = \|Av\|_2^2$ ถ้า $Av = 0$ จะขัดกับความเป็นอิสระของคอลัมน์เพราะ $v \neq 0$ ดังนั้น $Av \neq 0$ และ $\|Av\|_2^2 > 0$

3. ถ้า λ เป็นค่าลักษณะเฉพาะของ $A^{\mathsf{T}}A$ พร้อมเวกเตอร์ลักษณะเฉพาะ $v \neq 0$ จะได้ $\lambda\|v\|_2^2 = v^{\mathsf{T}}A^{\mathsf{T}}Av > 0$ จึง $\lambda > 0$ เมื่อค่าลักษณะเฉพาะทั้งหมดเป็นบวก ตัวกำหนดก็ไม่เป็นศูนย์ ซึ่งเป็นสิ่งที่ทฤษฎีบทในหัวข้อนี้ใช้

การตรวจด้วยโค้ดกับเมทริกซ์สุ่มขนาด 30×4 ด้วยเมล็ดคงที่ให้ค่าลักษณะเฉพาะบวกทั้งหมด

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 4.1.7.

1. ด้วยโค้ดของบทได้ $\mathrm{cond}(A) \approx 3.6558 \times 10^5$ และ $\mathrm{cond}(A^{\mathsf{T}}A) \approx 1.3365 \times 10^{11}$
2. $\mathrm{cond}(A)^2 = (3.6558 \times 10^5)^2 \approx 1.3365 \times 10^{11}$ ตรงกับ $\mathrm{cond}(A^{\mathsf{T}}A)$ เกือบทุกหลัก เหตุผลคือค่าเอกฐานของ $A^{\mathsf{T}}A$ เป็นกำลังสองของค่าเอกฐานของ A จึงได้ความสัมพันธ์นี้เป็นความจริงทั่วไป
3. ขอบประมาณความแม่นสัมพัทธ์ของคำตอบคือ $\mathrm{cond} \cdot u$ โดย $u \approx 6 \times 10^{-8}$
   - ระบบ A: $3.66 \times 10^5 \cdot 6 \times 10^{-8} \approx 0.02$ ยังพอใช้ได้
   - ระบบ $A^{\mathsf{T}}A$: $1.34 \times 10^{11} \cdot 6 \times 10^{-8} \approx 8 \times 10^3$ ซึ่งมากกว่า 1 มาก ขอบนี้ไม่ให้ความแม่นเหลืออยู่เลย

   การแก้สมการปรกติด้วยความแม่นเดี่ยวจึงเสี่ยงมาก ทั้งที่ระบบเดิมยังดี เหตุผลคือการยกกำลังสองทำให้คอนดิชันนัมเบอร์กลายเป็นกำลังสอง ขอบประมาณนี้เป็นตัวเลขคร่าว ๆ ไม่ใช่ค่าคลาดเคลื่อนที่แน่นอน

4.2 เลือกฐานเชิงตั้งฉาก

สถานการณ์สมมติ: ไซน์หนึ่งล้านครั้งต่อเฟรม

ลองนึกถึงฉากที่มีประกายไฟหรือฝุ่นฟุ้งขึ้นจากการระเบิด อนุภาคแต่ละตัวต้องถูกหมุนและเลื่อนตำแหน่งด้วยมุม ซึ่งต้องใช้ sin ของมุมนั้น สมมติว่าฉากหนึ่งต้องคำนวณ sin ราว 10^6 ครั้งต่อเฟรม (ตัวเลขนี้เป็นสมมติฐานของเรา ไม่ได้มาจากเกมใด)
การคำนวณ sin ทีละค่าต้องใช้ขั้นตอนมากกว่าการคูณและการบวกไม่กี่ครั้ง ดังนั้นเราอยากแทน sin ด้วยพหุนามดีกรีต่ำ ซึ่งคำนวณได้ด้วยการคูณและบวกไม่กี่ครั้ง มุมใด ๆ ดึงมาอยู่ในช่วง $[-\pi/2, \pi/2]$ ได้ด้วยสมบัติของ sin จึงพอประมาณเฉพาะช่วงนี้
คำถามของเราคือ พหุนามดีกรีต่ำพหุนามไหนแทน sin ได้ดีที่สุด คำตอบขึ้นกับว่า**วัดค่าคลาดเคลื่อนอย่างไร** ถ้าสนใจค่าคลาดเคลื่อนสัมบูรณ์ที่แย่ที่สุดในช่วง หรือค่าคลาดเคลื่อนรวมทั้งช่วง พหุนามที่ดีที่สุดอาจไม่ใช่พหุนามเดียวกัน

ภาพรวมก่อนอ่าน

หัวข้อนี้ใช้ความคิดเดียวกับหัวข้อ 4.1 แต่เปลี่ยนจากจุดข้อมูลจำนวนจำกัดเป็นฟังก์ชันทั้งช่วง ผลรวมจึงกลายเป็นปริพันธ์

- ก่อนเลือกพหุนาม เราต้องตกลงก่อนว่า "ผิดน้อย" หมายถึงอะไร
- เทย์เลอร์ที่จุดเดียวแม่นมากที่จุดนั้น แต่ค่าคลาดเคลื่อนโตขึ้นเร็วที่ปลายช่วง
- ถ้าเลือกฐานของพหุนามให้เชิงตั้งฉากกัน สมการปรกติจากหัวข้อ 4.1 จะเป็นเมทริกซ์แนวทแยง แก้ได้ทันที
- ถ้าต้องการค่าคลาดเคลื่อนสัมบูรณ์สูงสุดน้อยที่สุด ค่าคลาดเคลื่อนจะแกว่งขึ้นลงเท่า ๆ กันหลายจุด

ฉบับร่าง · Preprint · 2026-10-23

วัดค่าคลาดเคลื่อนอย่างไร

เพื่อให้ช่วงเป็น $[-1,1]$ เราใช้มุม $\theta = \pi t/2$ เมื่อ $t \in [-1,1]$ มุม θ จะอยู่ใน $[-\pi/2, \pi/2]$ เราจึงประมาณฟังก์ชัน

$$f(t) = \sin\Big(\frac{\pi t}{2}\Big), \qquad -1 \le t \le 1,$$

ด้วยพหุนาม $p(t)$ แล้วได้ $\sin\theta \approx p(2\theta/\pi)$ ค่าคลาดเคลื่อนคือ $e(t) = f(t) - p(t)$

นิยาม 4.2.1: ค่าคลาดเคลื่อนสองแบบ

สำหรับฟังก์ชัน e ที่ต่อเนื่องบน $[-1,1]$ ให้

$$\|e\|_\infty = \max_{-1\le t\le 1} |e(t)|, \qquad \|e\|_2 = \Big(\int_{-1}^{1} e(t)^2\,dt\Big)^{1/2}.$$

นอร์มค่าสูงสุด (maximum norm) คือค่าคลาดเคลื่อนสัมบูรณ์ที่แย่ที่สุดในช่วง ส่วน **นอร์ม** L_2 (L2 norm) คือรากของค่าคลาดเคลื่อนกำลังสองรวมทั้งช่วง

อ่านแบบเข้าใจง่าย

- นอร์มค่าสูงสุดตอบว่า "ตรงจุดที่ผิดที่สุด ผิดเท่าไร" ซึ่งสำคัญเมื่อค่าคลาดเคลื่อนนั้นมองเห็นได้บนจอ
- นอร์ม L_2 ตอบว่า "ทั้งช่วงผิดรวมกันเท่าไร" จุดที่ผิดมากจะมีผลมาก เพราะยกกำลังสองก่อนรวม
- สองคำตอบนี้อาจไม่ตรงกัน เราจึงต้องเลือกให้ชัดว่าต้องการแบบไหน

ลองเอง: เทย์เลอร์ตรงจุด 0

พหุนามเทย์เลอร์ดีกรีสามของ $\sin(\pi t/2)$ ที่ $t=0$ คือ $p_T(t) = \frac{\pi}{2}t - \frac{1}{6}\big(\frac{\pi}{2}\big)^3 t^3$ จงคำนวณ $e(t)$ ที่ $t=\frac{1}{2}$ และที่ $t=1$ แล้วดูว่าค่าคลาดเคลื่อนที่ $t=1$ มากกว่าที่ $t=\frac{1}{2}$ กี่เท่า

ที่ $t=\frac{1}{2}$ ได้ $e \approx 0.00245$ และที่ $t=1$ ได้ $e \approx 0.07517$ ค่าคลาดเคลื่อนที่ปลายจึงมากกว่าราว 30.6 เท่า (ตัวเลขนี้มาจากโค้ดของเรา) พจน์ที่เราตัดทิ้งจากเทย์เลอร์มีขนาดประมาณ t^{n+1} ซึ่งเล็กมากใกล้ 0 แต่โตเร็วเมื่อ $|t|$ เข้าใกล้ 1 Atkinson สรุปไว้ว่า เทย์เลอร์เป็นการประมาณแบบสม่ำเสมอ (uniform) ที่ไม่ดีเสมอ Atkinson 1989, §4.6, p. 224

ฐานเชิงตั้งฉาก: ทำให้สมการปรกติง่ายขึ้น

ในหัวข้อ 4.1 สมการปรกติมีเมทริกซ์ $A^{\mathsf{T}}A$ ซึ่งสำหรับจุดข้อมูลเป็นผลรวม และสำหรับฟังก์ชันทั้งช่วงเป็นปริพันธ์ เมื่อฐานเป็น $1, t, t^2, \ldots$ เมทริกซ์นี้มีคอนดิชันนัมเบอร์โตเร็วเมื่อดีกรีเพิ่ม เพราะฟังก์ชันเหล่านี้เกือบขึ้นต่อกันเชิงเส้น Atkinson 1989, §4.4, p. 207 Hesthaven และ Warburton ชี้ว่าเมทริกซ์ที่สร้างจากฐานเอกนามคล้าย**เมทริกซ์ฮิลเบิร์ต** (Hilbert matrix) (สมาชิก $1/(i+j-1)$) ซึ่งไม่เสถียรมากเมื่อขนาดใหญ่ และแนะนำให้สร้างฐานตั้งฉากด้วยกระบวนการแกรม–ชมิดท์ Hesthaven and Warburton 2008, §3.1, pp. 44–45

ฉบับร่าง · Preprint · 2026-10-23

นิยาม 4.2.2: ผลคูณภายในและพหุนามเชิงตั้งฉาก

ให้ $w(t) > 0$ บน $(-1, 1)$ เรียก w ว่า **ฟังก์ชันถ่วงน้ำหนัก** (weight function) และนิยามสำหรับฟังก์ชันต่อเนื่อง f, g

$$(f, g) = \int_{-1}^{1} f(t)\, g(t)\, w(t)\, dt.$$

นี่คือ **ผลคูณภายใน** (inner product) บนฟังก์ชัน เราเรียก f กับ g ว่า **เชิงตั้งฉาก** (orthogonal) เมื่อ $(f, g) = 0$ ลำดับพหุนาม $\varphi_0, \varphi_1, \ldots$ ที่ดีกรีของ φ_n เท่ากับ n และ $(\varphi_j, \varphi_k) = 0$ เมื่อ $j \neq k$ เรียกว่า **พหุนามเชิงตั้งฉาก** (orthogonal polynomials)

อ่านแบบเข้าใจง่าย

- ผลคูณภายในนี้เป็นผลคูณจุดของหัวข้อ 4.1 ที่รวมเป็นปริพันธ์แทนผลรวม เมื่อ $w = 1$ เป็นการรวมธรรมดาบนช่วง
- ถ้า f ต่อเนื่องและ $w > 0$ แล้ว $(f, f) = 0$ ก็ต่อเมื่อ $f = 0$ ทั้งช่วง ดังนั้น $\|f\|_2 = \sqrt{(f, f)}$ จึงเป็นขนาดที่ใช้ได้จริง
- ความเชิงตั้งฉากทำให้สัมประสิทธิ์ของพหุนามตัวหนึ่งไม่ขึ้นกับตัวอื่น การเพิ่มพหุนามตัวใหม่ไม่เปลี่ยนสัมประสิทธิ์เดิม

สองตระกูลที่ใช้บ่อย

เมื่อ $w = 1$ พหุนามเชิงตั้งฉากที่ได้เรียกว่า **พหุนามเลอฌ็องดร์** (Legendre polynomials) ตัวแรกคือ

$$P_0 = 1, \quad P_1 = t, \quad P_2 = \tfrac{1}{2}(3t^2 - 1), \quad P_3 = \tfrac{1}{2}(5t^3 - 3t).$$

ตัวต่อไปหาได้จากสูตรเวียนเกิด

$$(k+1)P_{k+1}(t) = (2k+1)\, t\, P_k(t) - k\, P_{k-1}(t),$$

ซึ่งเรียกว่า **ความสัมพันธ์เวียนเกิดสามพจน์** (three-term recurrence) Atkinson ให้สูตรนี้และค่า $(P_k, P_k) = 2/(2k+1)$ โดยไม่แสดงการหาสัมประสิทธิ์ทั้งหมด Atkinson 1989, §4.4, pp. 210–215 ในหัวข้อนี้เราใช้ช่วงอ้างอิง $[-1, 1]$ ซึ่งทำให้สูตรของพหุนามเลอฌ็องดร์และเชบีเชฟเขียนได้สะดวก Atkinson 1989, §4.4, pp. 210–215

เมื่อ $w(t) = 1/\sqrt{1 - t^2}$ พหุนามเชิงตั้งฉากที่ได้เรียกว่า **พหุนามเชบีเชฟ** (Chebyshev polynomials) นิยามง่ายที่สุดด้วยสูตรเวียนเกิดข้างล่าง และเราจะเห็นว่ามันคือโคไซน์เมื่อแทน $t = \cos\varphi$

n	$T_n(t)$
0	1
1	t
2	$2t^2 - 1$
3	$4t^3 - 3t$

ร่างบ่าง · Preprint · 2026-10-23

ทฤษฎีบท 4.2.1: พหุนามเชบีเชฟ

ให้ $T_0 = 1, T_1 = t$ และ $T_{n+1} = 2tT_n - T_{n-1}$ เมื่อ $n \geq 1$ จะได้ว่า

1. T_n เป็นพหุนามดีกรี n และ $T_n(\cos\varphi) = \cos n\varphi$ สำหรับทุก $\varphi \in \mathbb{R}$
2. $\displaystyle\int_{-1}^{1} \frac{T_m(t)\,T_n(t)}{\sqrt{1-t^2}}\,dt$ เท่ากับ 0 เมื่อ $m \neq n$ เท่ากับ π เมื่อ $m = n = 0$ และเท่ากับ $\pi/2$ เมื่อ $m = n \geq 1$

แนวคิด. ใช้สูตรผลบวกของโคไซน์เพื่อให้ได้สูตรเวียนเกิด แล้วแทน $t = \cos\varphi$ ซึ่งทำให้ตัวส่วน $\sqrt{1-t^2}$ หายไป ผลคูณของโคไซน์กลายเป็นผลบวกที่หาปริพันธ์ได้ทันที

1. **สูตรผลบวก.** จาก $\cos(A+B) + \cos(A-B) = 2\cos A\cos B$ ได้ $\cos\big((n+1)\varphi\big) + \cos\big((n-1)\varphi\big) = 2\cos\varphi\cos n\varphi$
2. **ข้อ 1 ของทฤษฎีบท.** ที่ $n = 0, 1$ ตรวจได้ตรง ๆ ว่า $T_0(\cos\varphi) = \cos 0$ และ $T_1(\cos\varphi) = \cos\varphi$ สมมติว่า $T_{n-1}(\cos\varphi) = \cos(n-1)\varphi$ และ $T_n(\cos\varphi) = \cos n\varphi$ แล้ว $T_{n+1}(\cos\varphi) = 2\cos\varphi\cos n\varphi - \cos(n-1)\varphi = \cos(n+1)\varphi$ ตามขั้น 1 เรื่องดีกรี: $2tT_n$ มีดีกรี $n+1$ และสัมประสิทธิ์นำไม่เป็นศูนย์ ส่วน T_{n-1} มีดีกรีน้อยกว่า จึงได้ T_{n+1} ดีกรี $n+1$
3. **การแทนค่า.** เมื่อ $t = \cos\varphi$ และ $\varphi \in [0,\pi]$ ได้ $dt = -\sin\varphi\,d\varphi$ และ $\sqrt{1-t^2} = \sin\varphi \geq 0$ จึง $\int_{-1}^{1} g(t)/\sqrt{1-t^2}\,dt = \int_0^{\pi} g(\cos\varphi)\,d\varphi$ ดังนั้นปริพันธ์ในข้อ 2 เท่ากับ $\int_0^{\pi} \cos m\varphi\cos n\varphi\,d\varphi$
4. **ผลคูณเป็นผลบวก.** $\cos m\varphi\cos n\varphi = \frac{1}{2}\big[\cos(m-n)\varphi + \cos(m+n)\varphi\big]$ และ $\int_0^{\pi}\cos k\varphi\,d\varphi$ เท่ากับ π เมื่อ $k = 0$ และเท่ากับ 0 เมื่อ k เป็นจำนวนเต็มที่ไม่เป็นศูนย์ เพราะ $\sin k\pi = 0$
5. **กรณี.** ถ้า $m \neq n$ ทั้ง $m-n$ และ $m+n$ ไม่เป็นศูนย์ จึงได้ 0 ถ้า $m = n = 0$ ได้ $\int_0^{\pi} 1\,d\varphi = \pi$ ถ้า $m = n \geq 1$ ได้ $\frac{1}{2}(\pi + 0) = \frac{\pi}{2}$

สรุป: สูตรเวียนเกิดให้พหุนามเชบีเชฟที่เป็นโคไซน์เมื่อแทน $t = \cos\varphi$ และการแทนนี้ทำให้ความเชิงตั้งฉากกลายเป็นปริพันธ์ของโคไซน์

ค่าสัมประสิทธิ์ทีละตัว

เมื่อมีฐานเชิงตั้งฉาก การหาพหุนามที่ใกล้ฟังก์ชันที่สุดในนอร์ม L_2 ไม่ต้องแก้ระบบใหญ่ ค่าสัมประสิทธิ์แต่ละตัวคำนวณได้จากปริพันธ์กับฟังก์ชันนั้นเอง

ทฤษฎีบท 4.2.2: การประมาณกำลังสองน้อยสุดด้วยฐานเชิงตั้งฉาก

ให้ $w > 0$ บน $(-1,1)$ ให้ f ต่อเนื่องบน $[-1,1]$ และให้ $\varphi_0, \ldots, \varphi_n$ เป็นพหุนามเชิงตั้งฉากที่ $(\varphi_k, \varphi_k) > 0$ ทุก k จะมีพหุนาม p^* ดีกรีไม่เกิน n ตัวเดียวที่ทำให้ $\|f - p\|_2$ น้อยที่สุดในบรรดาพหุนามดีกรีไม่เกิน n และ

$$p^* = \sum_{k=0}^{n} c_k \varphi_k, \qquad c_k = \frac{(f, \varphi_k)}{(\varphi_k, \varphi_k)}.$$

แนวคิด. ส่วนที่เหลือ $f - p^*$ ตั้งฉากกับทุกฐาน จึงตั้งฉากกับทุกพหุนามในช่วงนั้น ทำให้เกิดสามเหลี่ยมมุมฉากแบบพีทาโกรัส

ฉบับร่าง · Preprint 2026-10-23

1. **ส่วนที่เหลือตั้งฉาก.** สำหรับทุก j ได้ $(f - p^*, \varphi_j) = (f, \varphi_j) - \sum_k c_k(\varphi_k, \varphi_j) = (f, \varphi_j) - c_j(\varphi_j, \varphi_j) = 0$
2. **พีทาโกรัส.** ฐานทั้งหมดเป็นอิสระเชิงเส้น (ถ้า $\sum d_k\varphi_k = 0$ เอาผลคูณภายในกับ φ_j จะได้ $d_j = 0$) และมี $n+1$ ตัว จึงเป็นฐานของพหุนามดีกรีไม่เกิน n ทุกพหุนาม q เขียนได้เป็น $q = \sum d_k\varphi_k$ ดังนั้น $p^* - q$ อยู่ในช่วงนี้และ $(f - p^*, p^* - q) = 0$ จากขั้น 1 จึงได้ $\|f - q\|_2^2 = \|f - p^*\|_2^2 + \|p^* - q\|_2^2$
3. **น้อยสุดและมีตัวเดียว.** จากขั้น 2 ได้ $\|f - q\|_2 \geq \|f - p^*\|_2$ เสมอ เท่ากันก็ต่อเมื่อ $\|p^* - q\|_2 = 0$ ซึ่งเมื่อ $w > 0$ บังคับให้ $p^* - q = 0$ บนช่วง จึงเป็นศูนย์ทั้งหมดเพราะเป็นพหุนาม ดังนั้น $q = p^*$

สรุป: ฐานเชิงตั้งฉากทำให้การฉายบนพหุนามกลายเป็นการคำนวณค่าสัมประสิทธิ์ทีละตัว และให้คำตอบเดียวที่ดีที่สุดในนอร์ม L_2

ใช้ทฤษฎีบทนี้กับ $f(t) = \sin(\pi t/2)$, $w = 1$ และดีกรีสาม ค่าสัมประสิทธิ์คือ $c_1 = \dfrac{(f, P_1)}{(P_1, P_1)} = \dfrac{12}{\pi^2}$ และ $c_3 = \dfrac{(f, P_3)}{(P_3, P_3)} = \dfrac{168}{\pi^2} - \dfrac{1680}{\pi^4}$ (ค่าปริพันธ์ตรวจด้วยโค้ดของเรา) ส่วน c_0 และ c_2 เป็นศูนย์เพราะ f เป็นฟังก์ชันคี่ (ดูแบบฝึกหัด 4.2.6) จึงได้

$$p^*(t) = 1.5532\,t - 0.5622\,t^3$$

ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดของพหุนามนี้อยู่ที่ปลายช่วง (ดูตารางต่อไป) เพราะทฤษฎีบทนี้วัดค่าคลาดเคลื่อนทั้งช่วงด้วยนอร์ม L_2 ไม่ได้วัดจุดที่แย่ที่สุด

สำหรับฐานเอกนาม $1, t, \ldots, t^n$ เมทริกซ์สมการปรกติคือเมทริกซ์ที่สมาชิกเป็น $2/(i+j+1)$ เมื่อ $i+j$ เป็นเลขคู่ และเป็นศูนย์เมื่อ $i+j$ เป็นเลขคี่ ตารางต่อไปนี้เทียบคอนดิชันนัมเบอร์ของเมทริกซ์นี้กับของฐานเลอฌ็องดร์

n	ฐาน $1, t, \ldots, t^n$	ฐาน $P_0, \ldots, P_n$
1	3.0	3
2	14.13	5
3	67.60	7
4	358.3	9
5	1870	11

ค่าของคอลัมน์ขวาเป็นอัตราส่วนของสมาชิกแนวทแยงที่ใหญ่ที่สุดกับเล็กที่สุด ฐาน P_k ยังไม่ได้ทำให้มีความยาวหนึ่ง ถ้าปรับจะได้คอนดิชันนัมเบอร์เท่ากับ 1 ที่ทุกดีกรี ค่าของฐานเลอฌ็องดร์โตเชิงเส้นตาม $2n+1$ ขณะที่ฐานเอกนามโตราวห้าเท่าในทุกดีกรี

ข้อควรระวัง: ฐานดีไม่ได้แปลว่าพหุนามดี

ฐานเชิงตั้งฉากทำให้สมการปรกติง่ายและเสถียรกว่า แต่ไม่ได้ลดค่าคลาดเคลื่อนของพหุนามเอง ถ้าดีกรีต่ำเกินไป ค่าคลาดเคลื่อนยังมากอยู่ การเลือกฐานจึงเป็นเรื่องการคำนวณ ส่วนการเลือกดีกรีและนอร์มเป็นเรื่องอีกอย่าง

ตัวอย่าง · Preprint · 2026-10-22

อีกเกณฑ์หนึ่ง: ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดน้อยที่สุด

กำลังสองน้อยสุดทำให้ค่าคลาดเคลื่อนรวมทั้งช่วงน้อยที่สุด แต่ค่าคลาดเคลื่อนที่ปลายอาจยังใหญ่ ถ้าเป็นภาพบนจอ จุดเดียวที่ผิดมากก็มองเห็นได้ทันที ทางเลือกคือเลือกพหุนามที่ทำให้ $\|e\|_\infty$ เล็กที่สุด พหุนามนี้เรียกว่า **มินิแมกซ์** (minimax) ทฤษฎีบทของไวเออร์สตราส (Weierstrass) รับประกันว่าค่าคลาดเคลื่อนนี้เข้าใกล้ศูนย์เมื่อดีกรีเพิ่มขึ้น สำหรับฟังก์ชันต่อเนื่องบนช่วงปิด Atkinson 1989, §4.1, Thm 4.1, p. 198 (เราไม่พิสูจน์ที่นี่)

เงื่อนไขสำคัญเรียกว่า **การแกว่งเท่ากัน** (equioscillation): ค่าคลาดเคลื่อนสลับบวกกับลบและมีขนาดเท่ากันที่จุดสำคัญหลายจุด

ทฤษฎีบท 4.2.3: ทฤษฎีบทการแกว่งเท่ากันของเชบีเชฟ (อ้างจากต้นฉบับ)

ให้ f ต่อเนื่องบน $[-1,1]$ และ $n \geq 0$ จะมีพหุนาม p^* ดีกรีไม่เกิน n เพียงตัวเดียวที่ทำให้ $\|f-p\|_\infty$ น้อยที่สุด และถ้า $E = \|f-p^*\|_\infty$ จะมีจุด $-1 \leq x_0 < x_1 < \cdots < x_{n+1} \leq 1$ อย่างน้อย $n+2$ จุด ที่ $f(x_j) - p^*(x_j) = \sigma(-1)^j E$ สำหรับ $j = 0, \ldots, n+1$ โดยที่ $\sigma = \pm 1$ เป็นค่าคงที่

ทฤษฎีบทนี้เป็นผลที่ Atkinson อ้างไว้ โดยการพิสูจน์ยาวกว่าที่จะแสดงที่นี่ (ดู Atkinson 1989, §4.6, Thm 4.10, p. 224 และแหล่งที่เขาอ้าง)

แนวคิด. ทำไมการสลับเครื่องหมายที่ $n+2$ จุดจึงเพียงพอ: ถ้ามีพหุนาม q ดีกรีไม่เกิน n ที่มีค่าคลาดเคลื่อนสูงสุดเล็กกว่า ผลต่าง $p^* - q$ จะสลับเครื่องหมายที่จุดเดียวกันทั้ง $n+2$ จุด จึงมีศูนย์อย่างน้อย $n+1$ จุด ทั้งที่ดีกรีไม่เกิน n ทำให้ผลต่างนั้นเป็นศูนย์ ซึ่งขัดกับสมมติฐาน (แนวคิดนี้มาจาก Atkinson Thm 4.9 ซึ่งมีการพิสูจน์)

อนุกรมเชบีเชฟเป็นตัวประมาณที่ใกล้มินิแมกซ์

Atkinson แสดงว่า ถ้าเขียน $f(\cos\varphi)$ เป็นอนุกรมโคไซน์ $\sum c_j \cos(j\varphi)$ แล้วตัดที่ดีกรี n จะได้พหุนามที่เรียกว่า **อนุกรมเชบีเชฟ** (Chebyshev series) ที่ตัดแล้ว โดยค่าสัมประสิทธิ์คือ

$$c_j = \frac{2}{\pi}\int_0^{\pi} f(\cos\varphi)\cos(j\varphi)\,d\varphi$$

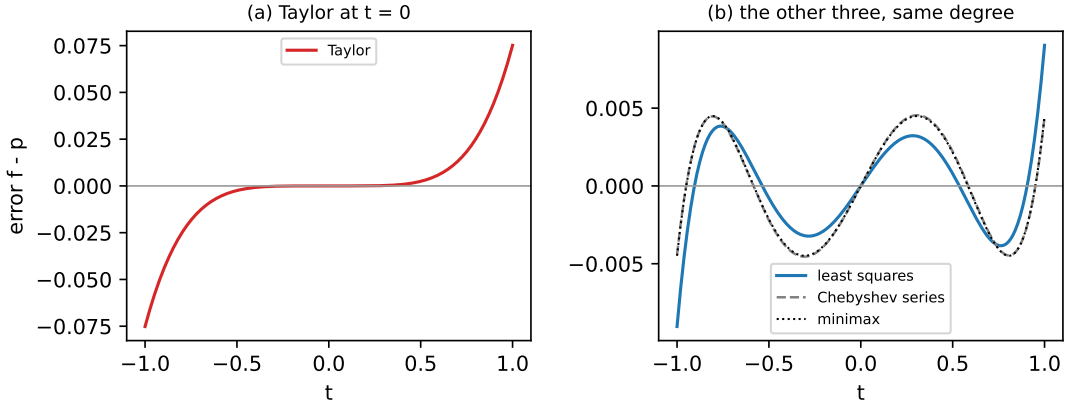
Atkinson 1989, §4.5, p. 219 ปริพันธ์นี้มีฟังก์ชันที่เป็นคาบและเรียบในตัวแปร φ กฎสี่เหลี่ยมคางหมูบนจุดที่เว้นระยะเท่ากันจึงแม่นมาก (รายละเอียดในบทที่ 5) โค้ดของเราใช้ 64 ช่วง Atkinson พบว่า สำหรับ e^x ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดของวิธีนี้ใกล้กับมินิแมกซ์มาก Atkinson 1989, §4.5, p. 221 เราตรวจกับ $\sin(\pi t/2)$ ของเรา

ตารางต่อไปนี้แสดงค่าคลาดเคลื่อนสูงสุดของพหุนามแต่ละดีกรี และนอร์ม L_2 ของดีกรีสาม ทุกตัวคำนวณด้วยโค้ดของเรา

วิธี	ดีกรี 1	ดีกรี 3	ดีกรี 5	L_2 ดีกรี 3
เทย์เลอร์	0.5708	0.07517	0.004525	0.03234
กำลังสองน้อยสุด (เลอฌ็องดร์)	0.2159	0.009037	0.0001608	0.003922
อนุกรมเชบีเชฟ	0.1404	0.004546	0.0000682	0.004492
มินิแมกซ์	0.1382	0.004492	0.0000677	0.004468

แต่ละวิธีชนะในนอร์มของตัวเอง กำลังสองน้อยสุดมี L_2 น้อยที่สุด และมินิแมกซ์มีค่าคลาดเคลื่อนสูงสุดน้อยที่สุด ที่ดีกรีสาม อนุกรมเชบีเชฟให้ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดต่างจากมินิแมกซ์เพียงราว 1.2% มินิแมกซ์ของดีกรีสามมีค่าคลาดเคลื่อนสัมบูรณ์เท่ากับ 0.0044917 ที่หกจุด $t = -1,\ -0.8056,\ -0.3057,\ 0.3057,\ 0.8056,\ 1$ โดยสลับเครื่องหมาย ซึ่งเป็นภาพของการแกว่งเท่ากันในทฤษฎีบทข้างบน (จุดเหล่านี้มาจากการคำนวณของเรา)

ร่างปริญญา Preprint 2026-10-23



รูปที่ 4.5: ค่าคลาดเคลื่อน $f - p$ ของพหุนามดีกรีสาม (ก) เทย์เลอร์ตรงจุด 0 ค่าคลาดเคลื่อนที่ปลายใหญ่กว่าอย่างเห็นได้ชัด (ข) สามวิธีที่เหลือ เส้นมินิแมกซ์สลับเครื่องหมายหกจุดด้วยขนาดเท่ากัน (สร้างจาก `numanal/ch04_approx.py`)

เราคำนวณมินิแมกซ์ด้วย**วิธีแลกเปลี่ยนจุด** (exchange method) ที่เขียนเอง เริ่มจากจุดอ้างอิง $n+2$ จุด แก้ระบบเชิงเส้นให้ $p(x_j) + (-1)^j E = f(x_j)$ แล้วเลื่อนจุดอ้างอิงไปยังจุดที่ค่าคลาดเคลื่อนสลับเครื่องหมายและใหญ่ที่สุด ทำซ้ำจนค่า E ตรงกับค่าคลาดเคลื่อนสัมบูรณ์สูงสุดบนกริดละเอียด วิธีนี้เป็นแนวทางที่รู้จักกันในชื่อวิธีของเรมีซ (Remez) ซึ่งเราเขียนขึ้นใหม่เอง

ลองเอง: กลับไปที่ฉาก

ถ้าเกณฑ์คือค่าคลาดเคลื่อนสัมบูรณ์สูงสุดไม่เกิน 0.005 ต่อจุด วิธีใดและดีกรีเท่าไรใช้ได้ จงดูจากตารางข้างบน

ถ้าเกณฑ์คือ 0.005 ดีกรีสามใช้ได้กับอนุกรมเชบีเชฟและมินิแมกซ์ (ค่าคลาดเคลื่อนสัมบูรณ์สูงสุด 0.00455 และ 0.00449) ส่วนกำลังสองน้อยสุดต้องขยับไปดีกรีห้า เพราะดีกรีสามให้ค่าคลาดเคลื่อนสัมบูรณ์สูงสุด 0.00904 และเทย์เลอร์ก็ต้องดีกรีห้าเช่นกัน ค่าสัมประสิทธิ์คำนวณครั้งเดียวตอนเตรียมโปรแกรม ไม่ได้คำนวณทุกเฟรม เมื่อประเมินพหุนามด้วยวิธีฮอร์เนอร์ ดีกรี n ใช้การคูณ n ครั้งต่อหนึ่งจุด ดังนั้นดีกรีสามใช้การคูณน้อยกว่าดีกรีห้าในทุกการประเมิน (แบบฝึกหัด 4.2.7)

สรุปใจความ

- ก่อนเลือกพหุนามต้องบอกว่าค่าคลาดเคลื่อนวัดอย่างไร นอร์มค่าสูงสุดกับนอร์ม L_2 ให้คำตอบต่างกัน
- เทย์เลอร์ตรงจุดเดียวแม่นที่จุดนั้น แต่ค่าคลาดเคลื่อนโตเร็วที่ปลายช่วง
- ฐานเชิงตั้งฉากทำให้สัมประสิทธิ์คิดได้ทีละตัว และให้พหุนามที่ดีที่สุดในนอร์ม L_2
- มินิแมกซ์ให้ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดน้อยที่สุด และค่าคลาดเคลื่อนสลับเครื่องหมายที่จุดจำนวนหนึ่ง
- อนุกรมเชบีเชฟใกล้มินิแมกซ์มาก ที่ดีกรีสามต่างกันเพียงราว 1.2% กลับไปที่ฉาก: ดีกรีสามพอสำหรับเกณฑ์ 0.005 ด้วยนอร์มค่าสูงสุด

อ่านต่อ. หัวข้อถัดไปถามว่า คลื่นที่ซ่อนอยู่ในเสียงและภาพเป็นความถี่ใดบ้าง ฟังก์ชัน $\sin$ และ $\cos$ บนช่วงคาบก็เชิงตั้งฉากกันเช่นกัน หัวข้อ 4.3 ใช้ความตั้งฉากนี้ ส่วนบทที่ 5 จะใช้รากของพหุนามเชิงตั้งฉากเป็นจุดของกฎเกาส์ Atkinson 1989, §4.4, Thm 4.4, p. 213

ฉบับร่าง Preprint 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 4.2.1 (⋆) ให้ $P_2(t) = \frac{1}{2}(3t^2 - 1)$ และ $w = 1$ บน $[-1, 1]$ จงแสดงว่า $(P_2, 1) = 0$ และ $(P_2, t) = 0$ จากนั้นหา (P_2, P_2)

แบบฝึกหัด 4.2.2 (⋆) จากสูตรเวียนเกิด $T_{n+1} = 2tT_n - T_{n-1}$ จงหา T_4 แล้วตรวจว่า $T_4(\cos\varphi) = \cos 4\varphi$ ที่ $\varphi = \pi/3$ และที่ $\varphi = \pi/4$

แบบฝึกหัด 4.2.3 (⋆⋆) ให้ $f(t) = t^4$ บน $[-1, 1]$ และ $w = 1$

1. จงหาพหุนามดีกรีไม่เกินสองที่ทำให้ $\|f - p\|_2$ น้อยที่สุด
2. จงหาค่าคลาดเคลื่อนสูงสุด $\|f - p\|_\infty$ ของพหุนามนั้น
3. พหุนาม $t^2 - \frac{1}{8}$ มีค่าคลาดเคลื่อนสูงสุดเล็กกว่าหรือไม่ ให้ใช้ $t^4 - (t^2 - \frac{1}{8}) = \frac{1}{8}T_4(t)$ และดูค่าที่ $t = \cos(k\pi/4)$

แบบฝึกหัด 4.2.4 (⋆⋆) ให้ G เป็นเมทริกซ์สมการปรกติบน $[-1, 1]$ โดย $G_{jk} = (\varphi_j, \varphi_k)$ จงคำนวณคอนดิชันนัมเบอร์ของ G เมื่อ $\varphi_k = t^k$ และเมื่อ $\varphi_k = P_k$ สำหรับ $n = 1, \ldots, 5$ แล้วสรุปเป็นตาราง จากนั้นอธิบายว่าเหตุใดฐานเลอฌ็องดร์จึงทำให้การแก้สมการปรกติง่ายกว่า

> **ตัวช่วยคิด**
>
> เมทริกซ์ของเอกนามมีสมาชิกเป็นศูนย์เมื่อ $j + k$ เป็นเลขคี่ ลองแยกดูดัชนีคู่กับคี่ และสำหรับ P_k ลองดูว่าสมาชิกนอกแนวทแยงหายไปหรือไม่

แบบฝึกหัด 4.2.5 (⋆⋆) ให้ $p_T(t) = \frac{\pi}{2}t - \frac{1}{6}\left(\frac{\pi}{2}\right)^3 t^3$ และ $e(t) = \sin(\pi t/2) - p_T(t)$

1. จงตรวจว่า $|e(0.6)| < 0.01 < |e(0.7)|$
2. ใช้วิธีแบ่งครึ่งช่วงสองสามรอบ หาค่า t โดยประมาณที่ $|e(t)| = 0.01$ เป็นครั้งแรกใน $[0, 1]$
3. ตอบว่าพหุนามเทย์เลอร์นี้ให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 0.01 สำหรับ $|t|$ ไม่เกินประมาณเท่าใด

แบบฝึกหัด 4.2.6 (⋆⋆⋆) ให้ f เป็นฟังก์ชันคี่บน $[-1, 1]$ และ $w = 1$

1. จงแสดงโดยอุปนัยจากสูตรเวียนเกิดของ P_k ว่า $P_k(-t) = (-1)^k P_k(t)$
2. จงแสดงว่า $(f, P_k) = 0$ เมื่อ k เป็นเลขคู่
3. จงอธิบายว่าทำไมพหุนามกำลังสองน้อยสุดดีกรีสามของ $\sin(\pi t/2)$ จึงมีเฉพาะพจน์ t และ t^3

> **ตัวช่วยคิด**
>
> ในข้อ (ข) ลองแทน t ด้วย $-t$ ในปริพันธ์ $\int_{-1}^{1} f(t)P_k(t)\,dt$ แล้วดูว่าปริพันธ์เท่ากับลบของตัวมันเองหรือไม่

แบบฝึกหัด 4.2.7 (⋆⋆⋆, โครงงานของบทนี้) ในฉากเดิม ค่าคลาดเคลื่อนที่ยอมรับได้คือ $\|e\|_\infty \le \varepsilon$ สำหรับทุกอนุภาคในทุกเฟรม ให้ใช้ตารางค่าคลาดเคลื่อนของหัวข้อ 4.2

1. สำหรับ $\varepsilon = 0.01$ และ $\varepsilon = 0.005$ จงหาดีกรีที่เล็กที่สุดในกลุ่ม $\{1, 3, 5\}$ ที่แต่ละวิธีผ่านเกณฑ์
2. ถ้าฉากต้องประเมินพหุนาม 10^6 ครั้งต่อเฟรม การเลือกดีกรีสามกับดีกรีห้าต่างกันเท่าไรในจำนวนการคูณต่อเฟรม (ใช้การประเมินแบบฮอร์เนอร์)

ฉบับร่าง Preprint · 2026-10-23

3. ถ้าเกณฑ์เปลี่ยนเป็นนอร์ม L_2 ของดีกรีสาม คำตอบจะเปลี่ยนอย่างไร และเพราะเหตุใด

4. เลือกวิธีและดีกรีสำหรับฉากเดิมที่ $\varepsilon = 0.005$ และอธิบายว่าค่าสัมประสิทธิ์ของพหุนามคำนวณเมื่อไร

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 4.2.1.

1. $(P_2, 1) = \int_{-1}^{1} \frac{1}{2}(3t^2-1)\,dt = \frac{1}{2}\big(3\cdot\frac{2}{3} - 2\big) = \frac{1}{2}(2-2) = 0$

2. $(P_2, t) = \int_{-1}^{1} \frac{1}{2}(3t^3 - t)\,dt = 0$ เพราะฟังก์ชันในปริพันธ์เป็นฟังก์ชันคี่ และช่วง $[-1,1]$ สมมาตร

3. $(P_2, P_2) = \int_{-1}^{1} \frac{1}{4}(9t^4 - 6t^2 + 1)\,dt = \frac{1}{4}\big(9\cdot\frac{2}{5} - 6\cdot\frac{2}{3} + 2\big) = \frac{1}{4}\cdot\frac{8}{5} = \frac{2}{5}$ ซึ่งตรงกับ $2/(2k+1)$ เมื่อ $k=2$

เฉลย 4.2.2. จากสูตรเวียนเกิดกับ $T_0 = 1$ และ $T_1 = t$:

$$T_2 = 2t\cdot t - 1 = 2t^2 - 1, \qquad T_3 = 2t(2t^2-1) - t = 4t^3 - 3t,$$

$$T_4 = 2t(4t^3 - 3t) - (2t^2 - 1) = 8t^4 - 8t^2 + 1.$$

ที่ $\varphi = \pi/3$ มี $t = \cos\varphi = \frac{1}{2}$ จึง $T_4 = \frac{8}{16} - \frac{8}{4} + 1 = -\frac{1}{2}$ และ $\cos(4\pi/3) = -\frac{1}{2}$ ตรงกัน ที่ $\varphi = \pi/4$ มี $t = \frac{\sqrt{2}}{2}$ จึง $t^2 = \frac{1}{2}$, $t^4 = \frac{1}{4}$ และ $T_4 = 2 - 4 + 1 = -1$ ขณะที่ $\cos\pi = -1$ ตรงกัน

เฉลย 4.2.3.

1. ใช้ทฤษฎีบทของหัวข้อนี้กับ $w = 1$ และฐาน P_0, P_1, P_2 ค่าสัมประสิทธิ์คือ $c_k = (f, P_k)/(P_k, P_k)$

   - $(f, P_0) = \int_{-1}^{1} t^4\,dt = \frac{2}{5}$ จึง $c_0 = \frac{2/5}{2} = \frac{1}{5}$
   - $(f, P_1) = 0$ เพราะ $t^4\cdot t$ เป็นฟังก์ชันคี่ จึง $c_1 = 0$
   - $(f, P_2) = \int_{-1}^{1} \frac{1}{2}t^4(3t^2-1)\,dt = \frac{1}{2}\big(3\cdot\frac{2}{7} - \frac{2}{5}\big) = \frac{1}{2}\cdot\frac{16}{35} = \frac{8}{35}$ จึง $c_2 = \frac{8/35}{2/5} = \frac{4}{7}$

   ดังนั้น $p(t) = \frac{1}{5} + \frac{4}{7}\cdot\frac{1}{2}(3t^2-1) = \frac{6}{7}t^2 - \frac{3}{35}$

2. ค่าคลาดเคลื่อนคือ $e(t) = t^4 - \frac{6}{7}t^2 + \frac{3}{35}$ ที่ $t = \pm 1$ ได้ $1 - \frac{6}{7} + \frac{3}{35} = \frac{8}{35} \approx 0.2286$ ที่จุดวิกฤต $t^2 = \frac{3}{7}$ ได้ $e = -\frac{24}{245} \approx -0.098$ และที่ $t = 0$ ได้ $\frac{3}{35} \approx 0.086$ ค่าสูงสุดของ $|e|$ จึงเป็น $\frac{8}{35}$ ที่ $t = \pm 1$

3. $t^4 - (t^2 - \frac{1}{8}) = t^4 - t^2 + \frac{1}{8} = \frac{1}{8}(8t^4 - 8t^2 + 1) = \frac{1}{8}T_4(t)$ เมื่อ $|T_4| \le 1$ จึงมี $|e| \le \frac{1}{8}$ และที่ $t = \cos(k\pi/4)$ มี $T_4 = \cos k\pi = (-1)^k$ สำหรับ $k = 0, \dots, 4$ ดังนั้นค่าคลาดเคลื่อนผลัดกันที่ $\pm\frac{1}{8}$ ห้าจุด ทฤษฎีบทการแกว่งเท่ากันต้องการเพียงสี่จุดสำหรับดีกรีสอง จึงสรุปได้ว่า $t^2 - \frac{1}{8}$ เป็นมินิแมกซ์ และค่าคลาดเคลื่อนสูงสุดของมันคือ $\frac{1}{8} = 0.125$ ซึ่งเล็กกว่า 0.2286 ของกำลังสองน้อยสุด

ข้อนี้แสดงว่าพหุนามที่ดีที่สุดในนอร์ม L_2 ไม่จำเป็นต้องดีที่สุดในนอร์มค่าสูงสุด (ตรวจค่าตัวเลขด้วยโค้ดของเรา)

ตัวอย่าง · Preprint · 2026-10-23

เฉลย 4.2.4. ตารางต่อไปนี้คำนวณด้วย `numpy.linalg.cond` บนเมทริกซ์ G ที่คำนวณด้วยปริพันธ์ตรง ๆ (ตรวจด้วยโค้ดของเรา)

n	ฐาน $1, t, \ldots, t^n$	ฐาน $P_0, \ldots, P_n$
1	3.0	3
2	14.129	5
3	67.60	7
4	358.3	9
5	1870	11

สำหรับ $n = 2$ คิดด้วยมือได้: G แยกเป็นบล็อกของ $\{1, t^2\}$ คือ $\begin{bmatrix} 2 & 2/3 \\ 2/3 & 2/5 \end{bmatrix}$ กับสมาชิก $2/3$ ของ t ค่าลักษณะเฉพาะของบล็อกแรกคือ $\frac{6}{5} \pm \frac{2\sqrt{61}}{15}$ ซึ่งเป็น 2.2414 และ 0.1586 ค่า $2/3$ อยู่ระหว่างสองค่านี้ จึงได้คอนดิชันนัมเบอร์

$$\frac{9 + \sqrt{61}}{9 - \sqrt{61}} = \frac{71 + 9\sqrt{61}}{10} \approx 14.129$$

สำหรับฐาน P_k เมทริกซ์ G เป็นแนวทแยงด้วย $G_{kk} = 2/(2k+1)$ คอนดิชันนัมเบอร์คืออัตราส่วนของค่าใหญ่สุดกับเล็กสุด $= \frac{2}{2/(2n+1)} = 2n + 1$ การแก้สมการปรกติจึงเป็นการหารทีละตัว $c_k = (f, P_k)/G_{kk}$ ไม่มีการกำจัดตัวแปรและไม่มีการขยายค่าคลาดเคลื่อนข้ามตัวแปร ฐานยังไม่ได้ปรับให้มีความยาวหนึ่ง ถ้าปรับจะได้คอนดิชันนัมเบอร์เท่ากับ 1 พอดี

เฉลย 4.2.5.

1. $e(0.6) = \sin 54^\circ - 0.942478 + 0.645964 \cdot 0.216 = 0.809017 - 0.942478 + 0.139528 \approx 0.006067$ ซึ่งน้อยกว่า 0.01 และ $e(0.7) = \sin 63^\circ - 1.099557 + 0.645964 \cdot 0.343 \approx 0.013015$ ซึ่งมากกว่า 0.01

2. ที่ $t = 0.65$ ได้ $e \approx 0.00902$ และที่ $t = 0.675$ ได้ $e \approx 0.01087$ จึงอยู่ระหว่าง 0.65 กับ 0.675 ค่าประมาณประมาณ 0.66 (โค้ดของเราให้ 0.6637)

3. เพราะ $\sin(\pi t/2)$ และ $p_T(t)$ เป็นฟังก์ชันคี่ ฟังก์ชัน e จึงเป็นฟังก์ชันคี่ และ $|e(-t)| = |e(t)|$ ดังนั้นพหุนามเทย์เลอร์ให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 0.01 สำหรับ $|t|$ ไม่เกินประมาณ 0.66 และเกินนั้นค่าคลาดเคลื่อนสัมบูรณ์จะเกิน 0.01 ทั้งสองข้าง

เฉลย 4.2.6.

1. ฐาน: $P_0 = 1$ เป็นคู่ และ $P_1 = t$ เป็นคี่ สมมติว่าข้อความจริงถึง $k-1$ และ k แล้ว

$$\begin{aligned}(k+1)P_{k+1}(-t) &= (2k+1)(-t)P_k(-t) - kP_{k-1}(-t) \\ &= (2k+1)(-t)(-1)^k P_k(t) - k(-1)^{k-1}P_{k-1}(t) \\ &= (-1)^{k+1}\big[(2k+1)tP_k(t) - kP_{k-1}(t)\big] = (-1)^{k+1}(k+1)P_{k+1}(t)\end{aligned}$$

   ดังนั้นข้อความจริงถึง $k+1$ ด้วย

2. ให้ $g(t) = f(t)P_k(t)$ เมื่อ k เป็นเลขคู่ จากข้อ 1 $P_k(-t) = P_k(t)$ และ $f(-t) = -f(t)$ จึง $g(-t) = -g(t)$ ฟังก์ชัน g เป็นฟังก์ชันคี่ และปริพันธ์บนช่วงสมมาตรของฟังก์ชันคี่เป็นศูนย์ ดังนั้น $(f, P_k) = 0$

3. จากข้อ 2 สัมประสิทธิ์ c_0 และ c_2 เป็นศูนย์ จึงเหลือ $c_1P_1 + c_3P_3$ ซึ่งเป็นพหุนามคี่ ที่มีเฉพาะพจน์ t และ t^3 โดยไม่ต้องคำนวณปริพันธ์ของ c_0 และ c_2 เลย

ฉบับร่าง Preprint · 2026-10-23

เฉลย 4.2.7.

1. จากตารางของหัวข้อ 4.2 (ค่าคลาดเคลื่อนสูงสุด)

   - $\varepsilon = 0.01$: เทย์เลอร์ดีกรีสามได้ 0.0752 (ไม่ผ่าน) ดีกรีห้าได้ 0.00452 (ผ่าน) จึงใช้ดีกรี 5; กำลังสองน้อยสุดดีกรีหนึ่งได้ 0.216 (ไม่ผ่าน) ดีกรีสามได้ 0.00904 (ผ่าน) จึงใช้ดีกรี 3; อนุกรมเชบีเชฟและมินิแมกซ์ดีกรีสามผ่าน (0.00455 และ 0.00449) จึงใช้ดีกรี 3
   - $\varepsilon = 0.005$: เทย์เลอร์ใช้ดีกรี 5 (0.00452 ผ่าน); กำลังสองน้อยสุดใช้ดีกรี 5 เพราะดีกรีสามได้ 0.00904 (ไม่ผ่าน) และดีกรีห้าได้ 0.000161; อนุกรมเชบีเชฟและมินิแมกซ์ใช้ดีกรี 3

2. การประเมินพหุนามดีกรี n ด้วยฮอร์เนอร์ใช้การคูณ n ครั้งและการบวก n ครั้ง ดีกรีสามจึงใช้การคูณ 3 ครั้ง และดีกรีห้าใช้ 5 ครั้ง เมื่อประเมิน 10^6 ครั้งต่อเฟรม ต่างกัน 2×10^6 การคูณต่อเฟรม (นับจากจำนวนขั้นตอน ไม่ได้วัดเวลาของเครื่องจริง)

3. กำลังสองน้อยสุดชนะในนอร์ม L_2 เพราะทฤษฎีบทในหัวข้อนี้ให้พหุนามที่ทำให้ $\|f - p\|_2$ น้อยที่สุด ค่าของดีกรีสามคือ 0.00392 น้อยกว่า 0.00447 ของมินิแมกซ์ แต่ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดของพหุนามเดียวกันคือ 0.00904 ซึ่งเกินเกณฑ์ 0.005 คำตอบจึงขึ้นกับว่าเราวัดค่าคลาดเคลื่อนแบบใด

4. เลือกมินิแมกซ์ (หรืออนุกรมเชบีเชฟ) ดีกรีสาม เพราะค่าคลาดเคลื่อนสัมบูรณ์สูงสุดคือ 0.00449 ซึ่งผ่าน 0.005 ค่าสัมประสิทธิ์ $t \mapsto 1.5481t - 0.5526t^3$ คำนวณครั้งเดียวตอนเตรียมโปรแกรม ไม่ได้คำนวณทุกเฟรม ทุกเฟรมจึงเหลือเพียงการประเมินพหุนามสามดีกรี

4.3 คลื่นที่ซ่อนอยู่ในเสียงและภาพ

สถานการณ์สมมติ: อีกด้านหนึ่งของปัญหาเดียวกัน

ลองนึกถึงเพลงที่เปิดในหน้าเว็บ แล้วมีแถบแนวตั้งหลายแถบขึ้นลงตามเสียง คล้ายอีควอไลเซอร์บนเครื่องเสียง (**อีควอไลเซอร์** (equaliser) คือแถบแสดงว่าเสียงในแต่ละช่วงความถี่ดังแค่ไหน) ไมโครโฟนเก็บเสียงเป็นค่าหนึ่งค่าในทุกช่วงเวลาสั้น ๆ ไม่มีแท่งใดอยู่ในสัญญาณนั้นให้เห็นตรง ๆ

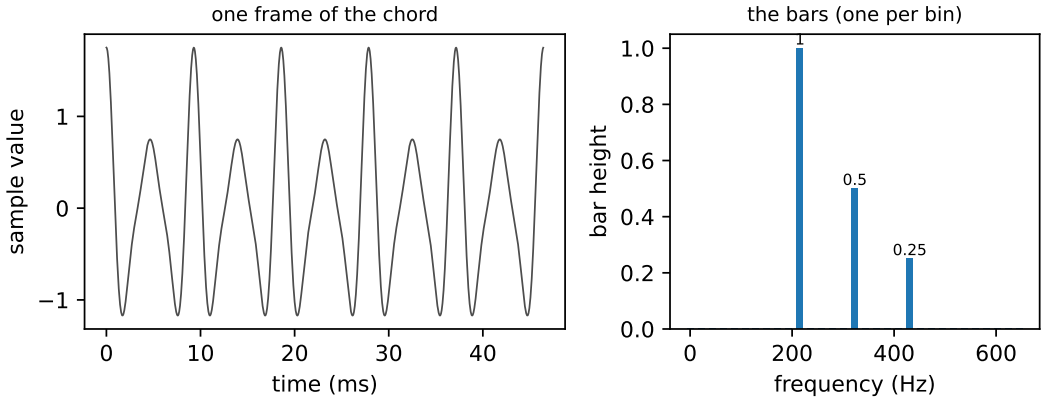
หัวข้อ 4.1 และ 4.2 จัดการกับค่าที่เปลี่ยนตามเวลาและมีสัญญาณรบกวน คราวนี้ลองดูอีกด้าน เสียงชุดเดียวกันอาจเป็นผลรวมของการสั่นหลายความถี่ และแท่งแต่ละแท่งบอกว่าความถี่หนึ่งมีอยู่มากแค่ไหน

ข้อกำหนด Web Audio API ของ W3C ระบุว่า โหนด `AnalyserNode` ให้การวิเคราะห์ความถี่และเวลาแบบเรียลไทม์ ค่าเริ่มต้นของ `fftSize` คือ 2048 และจำนวน**ช่องความถี่** (frequency bin) `frequencyBinCount` เท่ากับครึ่งหนึ่งของค่านี้ ช่องหนึ่งแทนตำแหน่งความถี่หนึ่งตำแหน่งในข้อมูลที่วิเคราะห์ Adenot and Choi 2026, §1.8.2 ก่อนแปลงฟูเรียร์ ข้อกำหนดคูณข้อมูลเวลาด้วย**หน้าต่างเบลแมน** (Blackman window) (น้ำหนักใกล้ศูนย์บริเวณปลายช่วง เพื่อลดการรั่ว) และทำให้ค่าเรียบตามเวลาก่อนแปลงเป็นเดซิเบล Adenot and Choi 2026, §1.8.6 ข้อกำหนดไม่ได้บอกว่าช่องแต่ละช่องตรงกับความถี่ใด ตัวเลขความถี่ในหัวข้อนี้จึงเป็นการคำนวณของเราเอง

เพื่อให้เห็นภาพ เราสร้างเสียงประสานจากสามโทนบริสุทธิ์ โดยสมมติว่าสุ่มที่ $44{,}100$ ครั้งต่อวินาที และใช้ 2048 จุด โทนทั้งสามอยู่ตรงช่องความถี่พอดีที่ 10, 15 และ 20 ด้วยความดัง 1, 0.5 และ 0.25

คำถามของหัวข้อนี้คือ เมื่อดูเฉพาะรูปคลื่นที่ยุ่งเหยิง แท่งเหล่านี้มาจากไหน สูงเท่าไร และคำนวณได้อย่างไรให้เร็วพอจะทำทุกเฟรม

ร่างสำหรับ Preprint · 2026-10-23



รูปที่ 4.6: ซ้าย: รูปคลื่นของเสียงประสานหนึ่งเฟรม (2048 จุด) ขวา: แท่งของแต่ละช่องความถี่ ความสูง 1, 0.5 และ 0.25 ตรงกับความดังของโทนทั้งสาม (สร้างจาก `numanal/ch04_spectrum.py`)

ภาพรวมก่อนอ่าน

สัญญาณที่**เป็นคาบ** (periodic) คือค่าที่ซ้ำเดิมทุกช่วงเวลาเท่า ๆ กัน สัญญาณแบบนี้เขียนเป็นผลรวมของคลื่นที่ความถี่เป็นจำนวนเท่าของความถี่พื้นฐานได้ ถ้าเราสุ่ม N จุดในหนึ่งคาบ เราต้องหาค่าสัมประสิทธิ์เพียง N ตัว ความคิดหลักคือคลื่นแต่ละความถี่ตั้งฉากกันบนจุดเหล่านี้ จึงหาค่าสัมประสิทธิ์ทีละตัวได้ด้วยผลคูณภายในแบบเดียวกับหัวข้อ 4.2

คลื่นฐานบนจุดที่เว้นระยะเท่ากัน

ให้ N เป็นจำนวนจุดที่ $j = 0, 1, \ldots, N-1$ และให้ $\omega = e^{2\pi i/N}$ จำนวนนี้อยู่บนวงกลมหนึ่งหน่วยในระนาบเชิงซ้อน และเป็น**รากปฐมฐานของหนึ่งอันดับ** N (primitive root of unity)ที่ยกกำลัง N แล้วได้ 1 แต่ยกกำลังที่น้อยกว่า N ไม่ได้ 1 คลื่นความถี่ k บนจุดเหล่านี้คือลำดับ $\omega^{0\cdot k}, \omega^{1\cdot k}, \ldots, \omega^{(N-1)k}$ เมื่อ k โตขึ้น คลื่นจะหมุนเร็วขึ้น และเมื่อ k ถึง N คลื่นกลับมาเหมือนเดิม จึงมีคลื่นที่ต่างกันเพียง N ลูก

ทฤษฎีบท 4.3.1: ผลรวมของรากปฐมฐาน

ให้ $N \geq 1$ และ m เป็นจำนวนเต็ม จะได้

$$\sum_{j=0}^{N-1} \omega^{jm} = \begin{cases} N, & \text{ถ้า } N \text{ หาร } m \text{ ลงตัว,} \\ 0, & \text{ถ้า } N \text{ หาร } m \text{ ไม่ลงตัว.} \end{cases}$$

แนวคิด. ถ้า N หาร m ลงตัว ทุกพจน์เท่ากับ 1 ถ้าไม่ลงตัว ผลรวมเป็นอนุกรมเรขาคณิตที่มีอัตราส่วนเป็นรากของ 1

1. **กรณีลงตัว.** ถ้า $m = Nr$ แล้ว $\omega^{jm} = (\omega^N)^{jr} = 1$ ทุก j ผลรวมจึงเป็น N
2. **กรณีไม่ลงตัว.** ให้ $z = \omega^m$ จะได้ $z^N = (\omega^N)^m = 1$ และ $z \neq 1$ เพราะ $z = e^{2\pi i m/N}$ เป็น 1 ก็ต่อเมื่อ m/N เป็นจำนวนเต็ม
3. **ผลรวม.** $\sum_{j=0}^{N-1} z^j = \dfrac{z^N - 1}{z - 1} = 0$ เพราะ $z^N = 1$

สรุป: ผลรวมของรากปฐมฐานเป็นศูนย์ เว้นแต่ทุกพจน์เท่ากัน ซึ่งเกิดเมื่อ N หาร m ลงตัว

ฉบับร่าง · Preprint · 2026-10-23

ค่าสัมประสิทธิ์และการกลับ

นิยาม 4.3.1: ผลการแปลงฟูเรียร์แบบไม่ต่อเนื่อง

ให้ $x_0, \dots, x_{N-1}$ เป็นค่าตัวอย่างของสัญญาณหนึ่งคาบ **ผลการแปลงฟูเรียร์แบบไม่ต่อเนื่อง** (discrete Fourier transform) (DFT) คือการหาค่า

$$c_k = \frac{1}{N} \sum_{j=0}^{N-1} x_j \, \omega^{-jk}, \qquad k = 0, 1, \dots, N-1,$$

เรียก c_k ว่า ค่าสัมประสิทธิ์ของช่องความถี่ k

อ่านแบบเข้าใจง่าย

- c_k บอกว่าในสัญญาณมีคลื่นความถี่ k ผสมอยู่มากแค่ไหน และคำนวณจากทุกจุดพร้อมกัน
- c_k เป็นจำนวนเชิงซ้อน ขนาดของมันบอกความดัง ส่วนมุมบอกเฟสหรือตำแหน่งเริ่มของคลื่น
- สูตรให้ค่าซ้ำทุก N ช่อง คือ $c_{k+N} = c_k$ ช่องที่อยู่เลยครึ่งทางไปแล้ว จึงหมายถึงความถี่ที่ติดลบ ซึ่งมีผลต่อการทับซ้อนที่เราจะเห็นข้างล่าง

ทฤษฎีบท 4.3.2: การกลับของผลการแปลงฟูเรียร์แบบไม่ต่อเนื่อง

ถ้า c_k นิยามตามนิยามข้างบน จะได้ $x_j = \sum_{k=0}^{N-1} c_k \, \omega^{jk}$ สำหรับทุก $j = 0, \dots, N-1$

แนวคิด. แทนสูตรของ c_k ลงไป แล้วสลับลำดับการรวม ผลรวมทางขวาจะกลายเป็นผลรวมของรากปฐมฐานที่ทำให้เหลือพจน์เดียว

1. **แทนค่า.** $\displaystyle\sum_k c_k \omega^{jk} = \frac{1}{N} \sum_{l=0}^{N-1} x_l \sum_{k=0}^{N-1} \omega^{(j-l)k}$ เพราะ $\omega^{jk}\omega^{-lk} = \omega^{(j-l)k}$
2. **ใช้ทฤษฎีบทผลรวม.** ผลรวมภายในเท่ากับ N ถ้า N หาร $j-l$ ลงตัว และเท่ากับ 0 ถ้าไม่ลงตัว
3. **เหลือพจน์เดียว.** เพราะ $j, l \in \{0, \dots, N-1\}$ จึง $N \mid (j-l)$ ก็ต่อเมื่อ $l = j$ ดังนั้นผลรวมเท่ากับ $\frac{1}{N} \cdot N \cdot x_j = x_j$

สรุป: การแปลงไปและกลับได้ค่าเดิม เพราะคลื่นของแต่ละความถี่ตั้งฉากกับคลื่นของความถี่อื่นบนจุดเหล่านี้

ความเกี่ยวกับหัวข้อ 4.2. ให้ $(u, v) = \sum_j u_j \overline{v_j}$ เป็นผลคูณภายในบนจุดทั้ง N และให้ $e_k = (\omega^{jk})_j$ ทฤษฎีบทผลรวมบอกว่า $(e_k, e_l) = 0$ เมื่อ $k \neq l$ และ $(e_k, e_k) = N$ ดังนั้น $c_k = (x, e_k)/(e_k, e_k)$ เป็นสูตรสัมประสิทธิ์เชิงตั้งฉากแบบเดียวกับหัวข้อ 4.2 เพียงแต่ปริพันธ์กลายเป็นผลรวมบนจุดที่เว้นระยะเท่ากัน และเป็นเวอร์ชันเชิงซ้อน

ลองเอง: สองจุด

ให้ $N = 2$ และ $x = (3, 1)$ จงหา c_0 และ c_1 เมื่อ $\omega = -1$ แล้วตรวจว่า $x_0 = c_0 + c_1$ และ $x_1 = c_0 + c_1\omega$

คำตอบ: $c_0 = \frac{1}{2}(3+1) = 2$ และ $c_1 = \frac{1}{2}\big(3 + 1 \cdot (-1)\big) = 1$ เมื่อแทนกลับได้ $x_0 = 2 + 1 = 3$ และ $x_1 = 2 + 1 \cdot (-1) = 1$ ตรงกับค่าเดิม

ฉบับร่าง Preprint 2026-10-23

จากค่าสัมประสิทธิ์ถึงแท่งของอีควอไลเซอร์

สำหรับสัญญาณจริง $c_{N-k} = \overline{c_k}$ (ดูแบบฝึกหัด 4.3.2) โดย $\overline{c_k}$ คือ**สังยุค** (complex conjugate)ของ c_k ซึ่งมีส่วนจินตภาพกลับเครื่องหมาย ดังนั้นความถี่ k กับ $N-k$ เป็นคู่กัน ให้ข้อมูลเดียวกัน ลองดูคลื่นโคไซน์ $x_j = A\cos(2\pi kj/N + \varphi)$ เมื่อ $0 < k < N/2$ เขียนได้เป็น $\frac{A}{2}\left(e^{i\varphi}\omega^{kj} + e^{-i\varphi}\omega^{(N-k)j}\right)$ จากทฤษฎีบทการกลับจะได้ $c_k = \frac{A}{2}e^{i\varphi}$ จึงได้

$$\text{ความสูงของแท่งที่ } k = 2|c_k| = A.$$

นี่คือเหตุผลที่แท่งสูงเท่าความดังของโทน ไม่ว่ามุมเฟสจะเป็นเท่าไร

ผลนี้ตรงกับเสียงประสานของเรา แท่งที่ช่อง 10, 15 และ 20 สูง 1, 0.5 และ 0.25 พอดี และช่องอื่นทั้งหมดต่ำกว่า 4×10^{-13} ซึ่งเป็นเพียงความคลาดเคลื่อนของการคำนวณบนเครื่อง ช่องความถี่ k ตรงกับความถี่ $k\, f_s/N$ ที่ $f_s = 44{,}100$ Hz และ $N = 2048$ จึงห่างกันราว 21.53 Hz ซึ่งเราคำนวณเองทั้งหมด (ข้อกำหนดไม่ได้ให้ความถี่ของแต่ละช่อง)

แท่งเหล่านี้มีอีกสมบัติหนึ่งที่สำคัญ คือพลังงานของสัญญาณไม่หายไปไหน ทฤษฎีบทถัดไปบอกว่าผลรวมกำลังสองของสัญญาณเท่ากับผลรวมของแท่งทุกช่อง

ทฤษฎีบท 4.3.3: เอกลักษณ์ของพาร์เซวัล

ให้ c_k เป็นค่าสัมประสิทธิ์ของ $x_0, \ldots, x_{N-1}$ จะได้

$$\sum_{j=0}^{N-1} |x_j|^2 = N \sum_{k=0}^{N-1} |c_k|^2.$$

บทพิสูจน์ (แบบเต็ม). ใช้การกลับ $x_j = \sum_k c_k \omega^{jk}$ และได้ $\overline{x_j} = \sum_l \overline{c_l}\, \omega^{-jl}$ เมื่อคูณสองสูตรนี้แล้วรวมตาม j จะได้ $\sum_j |x_j|^2 = \sum_k \sum_l c_k \overline{c_l} \sum_j \omega^{j(k-l)}$ ผลรวมภายในเท่ากับ N เมื่อ $k = l$ และเท่ากับ 0 เมื่อ $k \neq l$ ตามทฤษฎีบทผลรวมของรากปฐมฐาน จึงได้ $N \sum_k |c_k|^2$

สำหรับเสียงประสานของเรา ค่าเฉลี่ยกำลัง $\frac{1}{N}\sum_j x_j^2$ เท่ากับ 0.65625 ซึ่งเท่ากับ $\frac{1}{2}(1^2 + 0.5^2 + 0.25^2)$ เพราะโทนแต่ละโทนที่มีแอมพลิจูด A ให้กำลังเฉลี่ย $\frac{1}{2}A^2$ (ตรวจด้วยโค้ดของเรา) พลังงานจึงแบ่งให้ทั้งสามแท่งตามความดังยกกำลังสอง

ลองเอง: แท่งที่สูงที่สุดรับพลังงานเท่าไร

จากเสียงประสานนี้ แท่งที่ช่อง 10 รับกำลังเฉลี่ยเท่าไรจากค่าเฉลี่ยกำลังรวม 0.65625 และคิดเป็นกี่เปอร์เซ็นต์

คำตอบ: โทนที่ช่อง 10 มีแอมพลิจูด 1 จึงให้กำลังเฉลี่ย $\frac{1}{2}$ คิดเป็น $\frac{0.5}{0.65625} \approx 76\%$ ของทั้งหมด ส่วนอีกราว 24% อยู่ในสองโทนที่เหลือ

ข้อควรระวัง: อย่าอ่านแท่งจากผลดิบ

ผลดิบ $Y_k = \sum_j x_j \omega^{-jk}$ ยังไม่ได้หารด้วย N และมีขนาด $N|c_k|$ สำหรับเสียงประสาน $|Y_{10}|$ เท่ากับ 1024 ไม่ใช่ 1 ถ้าอ่านผลดิบเป็นแท่ง แท่งจะสูงเกินจริง $N/2$ เท่า แท่งที่ถูกคือ $2|Y_k|/N$ (ตัวเลข 1024 นี้ตรวจด้วยโค้ดของเรา)

ฉบับร่าง preprint 2026-10-23

การประมาณค่าเชิงตรีโกณมิติ

เราเห็นแล้วว่า N ค่าของสัญญาณคาบอยู่บนคาบ 2π ที่โหนด $t_j = 2\pi j/N$ ถ้าเราอยากได้ฟังก์ชันเรียบที่ผ่านทุกโหนด พหุนามที่บทที่ 3 ใช้ไม่มีเงื่อนไขว่าต่อกันอย่างไรที่ปลายของคาบ จึงเลือกผลรวมของไซน์และโคไซน์แทน Atkinson 1989, §3.8, pp. 176–177 ผลรวมแบบนี้เรียกว่า **พหุนามตรีโกณมิติ** (trigonometric polynomial) และการหาพหุนามชนิดนี้ที่ผ่านทุกโหนดเรียกว่า **การประมาณค่าเชิงตรีโกณมิติ** (trigonometric interpolation)

นิยาม 4.3.2: พหุนามตรีโกณมิติของข้อมูลจริง

ให้ $x_0, \dots, x_{N-1}$ เป็นจำนวนจริง โดย N เป็นจำนวนคู่ และให้ c_k เป็น DFT ของ x นิยาม

$$p(t) = \operatorname{Re} c_0 + \sum_{k=1}^{N/2-1} \left(2 \operatorname{Re} c_k \cos kt - 2 \operatorname{Im} c_k \sin kt\right) + \operatorname{Re} c_{N/2} \cos \tfrac{N}{2} t.$$

อ่านแบบเข้าใจง่าย

- p เป็นผลรวมของคลื่นที่ความถี่ตั้งแต่ 0 ถึง $N/2$ จึงเรียบและมีคาบ 2π
- ค่าสัมประสิทธิ์ทั้งหมดคำนวณจากข้อมูลแล้ว ไม่มีตัวแปรอิสระให้ปรับ ต่างจากเส้นตรงของหัวข้อ 4.1
- ทฤษฎีบทถัดไปบอกว่าเส้นโค้งนี้ผ่านทุกโหนดจริง ๆ ซึ่งเป็นความต่างจากการประมาณแบบกำลังสองน้อยสุด

ทฤษฎีบท 4.3.4: การประมาณค่าที่โหนด

ให้ $x_0, \dots, x_{N-1}$ เป็นจำนวนจริง N เป็นจำนวนคู่ และ $t_j = 2\pi j/N$ จะได้ $p(t_j) = x_j$ สำหรับทุก $j = 0, \dots, N-1$

แนวคิด. ใช้การกลับของ DFT แล้วจับคู่พจน์ k กับ $N-k$ เพราะค่าสัมประสิทธิ์ของข้อมูลจริงเป็นคู่สังยุคกัน

1. **ที่โหนด.** $\omega^{jk} = e^{ikt_j}$ จึง $x_j = \sum_{k=0}^{N-1} c_k e^{ikt_j}$
2. **จับคู่.** สำหรับ $1 \le k \le N/2 - 1$ ใช้ $c_{N-k} = \overline{c_k}$ (แบบฝึกหัด 4.3.2) ได้ $c_k e^{ikt_j} + c_{N-k} e^{-ikt_j} = 2\operatorname{Re}(c_k e^{ikt_j}) = 2 \operatorname{Re} c_k \cos kt_j - 2 \operatorname{Im} c_k \sin kt_j$
3. **พจน์ที่เหลือ.** c_0 เป็นจำนวนจริง และ $c_{N/2} = \overline{c_{N/2}}$ จึงเป็นจำนวนจริง นอกจากนี้ $e^{i(N/2)t_j} = (-1)^j = \cos\left(\frac{N}{2} t_j\right)$ ดังนั้นพจน์สองตัวนี้ตรงกับนิยามของ p

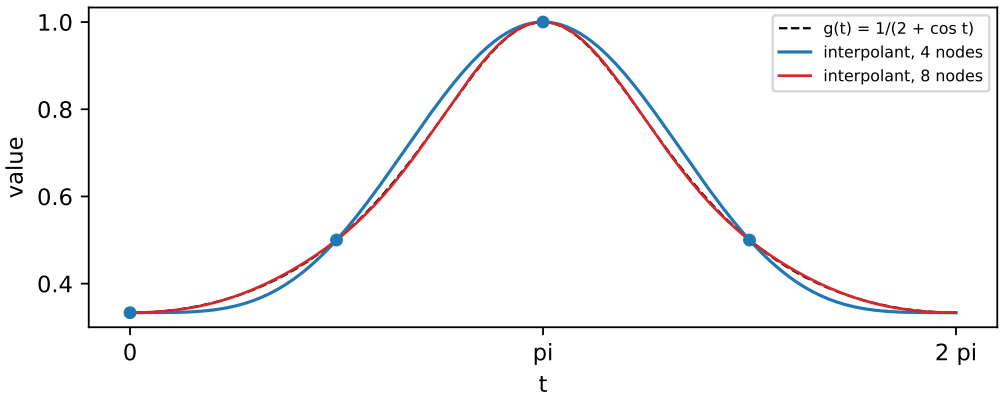
สรุป: ข้อมูลจริงที่สุ่มบนคาบมีพหุนามตรีโกณมิติผ่านทุกโหนด และสัมประสิทธิ์หาได้ทันทีจาก DFT

เสียงประสานของเราเป็นพหุนามตรีโกณมิติอยู่แล้ว จึงได้ว่า $p(0.3) = -0.85535$ ที่จุดที่ไม่ใช่โหนด ซึ่งตรงกับค่าจริงของเสียงประสาน (ตรวจด้วยโค้ดของเรา)

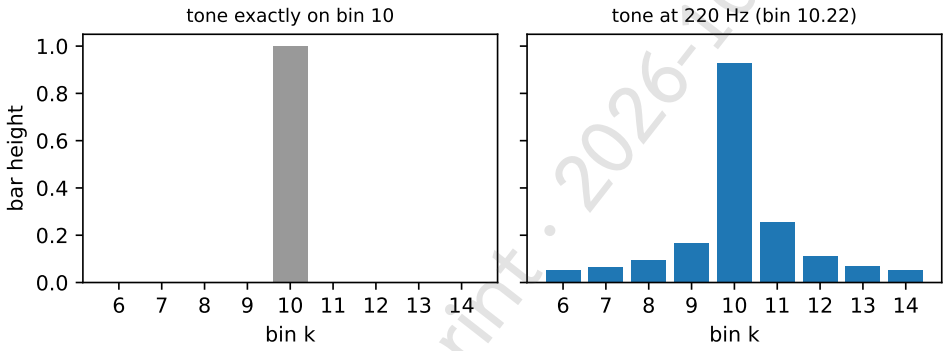
ลองกับฟังก์ชันที่ไม่ได้เป็นพหุนามตรีโกณมิติ เช่น $g(t) = 1/(2 + \cos t)$ ซึ่งเป็นคาบและเรียบ ตารางนี้วัดค่าคลาดเคลื่อนสัมบูรณ์สูงสุดระหว่างโหนดบนกริดละเอียด (ตรวจด้วยโค้ดของเรา)

จำนวนโหนด N	4	8	16	32
ค่าคลาดเคลื่อนสัมบูรณ์สูงสุด	0.0463	3.40×10^{-3}	1.77×10^{-5}	4.70×10^{-10}

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 4.7: ฟังก์ชัน $g(t) = 1/(2 + \cos t)$ (เส้นประ) พหุนามตรีโกณมิติที่ผ่านสี่โหนด (เส้นฟ้า พร้อมจุดโหนด) และที่ผ่านแปดโหนด (เส้นแดง) เส้นแดงแทบทับเส้นประ (สร้างจาก `numanal/ch04_spectrum.py`)



รูปที่ 4.8: ซ้าย: โทนที่ตรงช่อง 10 พอดี มีแท่งเดียว ขวา: โทนที่ 220 Hz ตกอยู่ระหว่างช่อง แท่งกระจายไปช่อง 9 ถึง 12 (สร้างจาก `numanal/ch04_spectrum.py`)

ที่โหนดค่าคลาดเคลื่อนเป็นศูนย์ทั้งหมด (ไม่เกิน 3×10^{-15} จากการปัดเศษ) แต่ระหว่างโหนดค่าคลาดเคลื่อนลดลงเร็วมาก เมื่อ N เป็นสองเท่า ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดลดราว 14 เท่า 190 เท่า และ 38,000 เท่า ตามลำดับ การลู่เข้าเร็วแบบนี้ใช้กับฟังก์ชันที่เป็นคาบและเรียบเท่านั้น ส่วนทฤษฎีบทที่ผ่านโหนดใช้ได้กับข้อมูลจริงทุกชุดที่สุ่มบนคาบ

ถ้าความถี่ไม่ตรงช่อง

ในเสียงประสานเราเลือกโทนให้ตรงช่องพอดี ในชีวิตจริงโน้ตไม่ได้ตรงช่องเสมอไป ลองดูโทน 220 Hz ซึ่งอยู่ที่ช่อง $220 \cdot 2048/44100 \approx 10.22$ ของเรา แท่งที่ได้ (ตรวจด้วยโค้ดของเรา) คือ

ช่อง k	8	9	10	11	12	13
แท่ง	0.0936	0.1675	0.9269	0.2538	0.1106	0.0705

พลังงานของโทนนี้กระจายไปหลายช่องข้างเคียง เรียกปรากฏการณ์นี้ว่า **การรั่วของสเปกตรัม** (spectral leakage) (ยังไม่พบคำนี้ในฐานศัพท์ของเล่มนี้ จึงเป็นคำที่เราใช้เอง)

ฉบับร่าง · Preprint · 2026-10-23

ข้อควรระวัง: การรั่วและความละเอียดของช่อง

ตารางข้างบนไม่ได้ใช้หน้าต่างใด ๆ เลย ข้อกำหนด Web Audio ใช้หน้าต่างเบลแมนก่อนแปลง จึงให้ตัวเลขอื่น ค่าที่เห็นในเบราว์เซอร์จึงไม่เท่ากับค่าของเราบนช่องเดียวกัน แต่หลักการเหมือนกัน คือโทนที่ไม่ตรงช่องจะรั่ว และความกว้างของช่อง f_s/N กำหนดความละเอียดของการแยกความถี่ ถ้าสองโทนต่างกันเพียงเศษของช่อง แท่งที่ได้อาจรวมกันเป็นแท่งเดียวได้

ลองดูตัวอย่างที่ความละเอียดสำคัญ ให้โทนสองโทนที่ 1000 Hz กับ 1010 Hz ซึ่งห่างกันราว 0.46 ช่อง และโทน 1000 Hz กับ 1040 Hz ซึ่งห่างกันราว 1.86 ช่อง แท่งของสองกรณี (ตรวจด้วยโค้ดของเรา) คือ

ช่อง k	45	46	47	48	49
1000 และ 1010 Hz	0.224	0.726	1.188	0.230	0.138
1000 และ 1040 Hz	0.285	0.808	0.395	0.684	0.480

กรณีแรกเห็นเป็นยอดเดียวที่ช่อง 47 ไม่มีหุบให้แยกสองโทน กรณีหลังมียอดสองยอดที่ช่อง 46 และ 48 คั่นด้วยหุบที่ช่อง 47 ดังนั้นความกว้างของช่องบอกได้ว่าความถี่ที่ใกล้กันเกินราวหนึ่งช่องมักแยกไม่ออกในแท่งเดียว

ความถี่สูงที่ปลอมตัวเป็นความถี่ต่ำ

เมื่อเราสุ่มที่อัตรา f_s ความถี่ที่สูงที่สุดที่แยกได้โดยไม่สับสนคือครึ่งหนึ่งของ f_s เรียกว่า **ความถี่ไนควิสต์** (Nyquist frequency) ซึ่งคือ $f_s/2$ ความถี่ที่สูงกว่านี้จะกลายเป็นความถี่ต่ำที่ดูเหมือนกันทุกจุด เรียกว่า **การทับซ้อนความถี่** (aliasing)

ทฤษฎีบท 4.3.5: การทับซ้อนของความถี่

ให้ $f_s > 0$ เป็นอัตราการสุ่ม และให้ f เป็นจำนวนจริงและ m เป็นจำนวนเต็ม จะได้

$$\cos\!\left(\frac{2\pi(f + m f_s)\, j}{f_s}\right) = \cos\!\left(\frac{2\pi f\, j}{f_s}\right) \qquad \text{สำหรับทุกจำนวนเต็ม } j.$$

ดังนั้นโทนความถี่ f กับโทนความถี่ $f + m f_s$ ให้ค่าตัวอย่างเหมือนกันทุกจุด

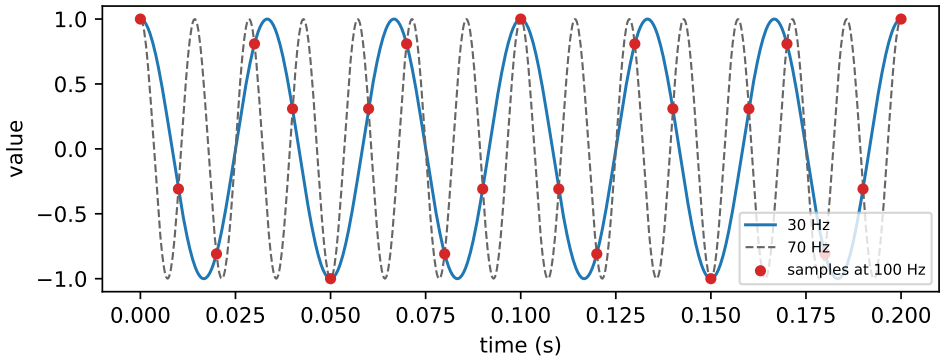
บทพิสูจน์ (แบบเต็ม). $\dfrac{2\pi(f + m f_s)j}{f_s} = \dfrac{2\pi f j}{f_s} + 2\pi m j$ และ $2\pi m j$ เป็นจำนวนเท่าของ 2π ค่าของ cos จึงไม่เปลี่ยน

ตัวอย่าง ที่ $f_s = 100$ Hz โทน 70 Hz กับโทน 30 Hz ให้ค่าเหมือนกันทุกจุด เพราะ $70 = 100 - 30$ จึงเป็นกรณี $m = -1$ ตัวเลขนี้ตรวจด้วยโค้ดของเรา: ค่าต่างกันไม่เกิน 10^{-12} และแท่งสูงสุดของการแปลง 100 จุดอยู่ที่ช่อง 30 รูปถัดไปแสดงจุดตัวอย่างเดียวกันของสองโทน

ข้อควรระวัง: แท่งที่สูงกว่าไนควิสต์ไม่ใช่ความถี่จริง

แท่งที่ช่อง 30 ไม่บอกว่ามีโทน 30 Hz หรือ 70 Hz อยู่จริง เราต้องคัดความถี่ที่สูงกว่า $f_s/2$ ออกก่อนสุ่ม หรือเลือก f_s ให้สูงพอ จึงจะให้แถบอีควอไลเซอร์ที่ซื่อตรง

ร่างสำหรับ Preprint 2026-10-23



รูปที่ 4.9: โทน 30 Hz (เส้นทึบ) และ 70 Hz (เส้นประ) ถูกสุ่มที่ 100 Hz จุดแดงอยู่บนเส้นทั้งสองพร้อมกัน จึงแยกไม่ออกจากจุดข้อมูลเพียงอย่างเดียว (สร้างจาก `numanal/ch04_spectrum.py`)

เร็วได้อย่างไร: จาก N^2 สู่ $N \log N$

การคำนวณตรงต้องหาผลคูณ $x_j \omega^{-jk}$ ทุกคู่ (j,k) จึงเป็น N^2 การคูณเชิงซ้อน (Atkinson นับแบบเดียวกันที่ m^2 Atkinson 1989, §3.8, p. 181) ที่ $N = 2048$ คือ $4{,}194{,}304$ การคูณต่อเฟรม ซึ่งหนักเกินกว่าจะทำทุกเฟรม ความคิดของ**ผลการแปลงฟูเรียร์แบบเร็ว** (fast Fourier transform) (FFT) คือแบ่งจุดเป็นคู่กับคี่ แล้วรวมผลของสองครึ่งที่เล็กลงเป็นครึ่งละครึ่ง

ทฤษฎีบท 4.3.6: การแบ่งครึ่งแบบคู่และคี่

ให้ $N = 2M$ และให้ $Y_k = \sum_{j=0}^{N-1} x_j \omega^{-jk}$ (ไม่มี $\frac{1}{N}$) ให้ $E_k = \sum_{j=0}^{M-1} x_{2j}\, \omega_M^{-jk}$ และ $O_k = \sum_{j=0}^{M-1} x_{2j+1}\, \omega_M^{-jk}$ โดย $\omega_M = e^{2\pi i/M}$ จะได้สำหรับ $k = 0, \dots, M-1$

$$Y_k = E_k + \omega^{-k} O_k, \qquad Y_{k+M} = E_k - \omega^{-k} O_k.$$

แนวคิด. แยกผลรวมตามดัชนีคู่และคี่ ดัชนีคู่ให้ $\omega^2 = \omega_M$ และ $\omega^M = -1$ ทำให้ครึ่งหลังของผลลัพธ์ใช้ค่าชุดเดียวกับครึ่งแรก

1. **แยกคู่และคี่.** จาก $Y_k = \sum_j x_{2j} \omega^{-2jk} + \sum_j x_{2j+1} \omega^{-(2j+1)k}$ และ $\omega^2 = \omega_M$ ได้ $Y_k = E_k + \omega^{-k} O_k$ สำหรับ $k = 0, \dots, M-1$
2. **ครึ่งหลัง.** E และ O มีคาบ M ในดัชนี k เพราะ $\omega_M^M = 1$ และ $\omega^{-(k+M)} = \omega^{-k}\omega^{-M} = -\omega^{-k}$ จึง $Y_{k+M} = E_k - \omega^{-k} O_k$

สรุป: จากผลของสองครึ่งขนาด M เราได้ผลขนาด N โดยใช้การคูณเพิ่มเพียง M ครั้งคือ $\omega^{-k} O_k$

ลองทำตามขั้นนี้ด้วยมือกับ $N = 4$ และ $x = (1, 2, 3, 4)$ ดัชนีคู่ $(x_0, x_2) = (1, 3)$ ให้ $E_0 = 4$ และ $E_1 = -2$ ดัชนีคี่ $(x_1, x_3) = (2, 4)$ ให้ $O_0 = 6$ และ $O_1 = -2$ ที่ $N = 4$ จะได้ $\omega^{-1} = -i$ จึง

$$Y_0 = E_0 + O_0 = 10, \qquad Y_1 = E_1 + (-i) O_1 = -2 + 2i,$$

$$Y_2 = E_0 - O_0 = -2, \qquad Y_3 = E_1 - (-i) O_1 = -2 - 2i.$$

ตรวจกับผลรวมตรง $Y_1 = 1 + 2(-i) + 3(-1) + 4i = -2 + 2i$ ตรงกัน การคูณด้วย**ตัวคูณหมุน** (twiddle factor) ซึ่งคือเลข ω^{-k} ที่ใช้หมุนผลของครึ่งคี่ก่อนรวมกับครึ่งคู่ มีสองครั้งที่ระดับบน ได้แก่ $(-i)O_0$ และ $(-i)O_1$ และอีกสองครั้งในการแบ่งย่อยขนาด 2 รวมเป็น $\frac{N}{2} \log_2 N = 4$ ครั้ง ตามทฤษฎีบทนับการคูณข้างล่าง

ฉบับร่าง · Preprint · 2026-10-23

ลองขนาด $N = 8$ กับข้อมูลที่ซ้ำทุก 4 จุด คือ $x = (1, 2, 3, 4, 1, 2, 3, 4)$ ขั้นแรกแยกดัชนีคู่ $(1, 3, 1, 3)$ และดัชนีคี่ $(2, 4, 2, 4)$ ไปทำ DFT ขนาด 4 ได้ E และ O แล้วใช้ตัวคูณหมุน ω^{-k} ที่ $\omega = e^{2\pi i/8}$ รวมเป็น Y_k และ Y_{k+4}

k	E_k	O_k	Y_k	Y_{k+4}
0	8	12	20	-4
1	0	0	0	0
2	-4	-4	$-4+4i$	$-4-4i$
3	0	0	0	0

ที่ $k = 2$ ตัวคูณหมุนคือ $\omega^{-2} = -i$ จึงได้ $Y_2 = -4+(-i)(-4) = -4+4i$ และ $Y_6 = -4-(-i)(-4) = -4-4i$ ข้อมูลซ้ำทุก 4 จุด จึงมีพลังงานเฉพาะช่องคู่ ซึ่งตรงกับแถวที่เป็นศูนย์ในตาราง

ทฤษฎีบท 4.3.7: จำนวนการคูณของ FFT

ถ้า $N = 2^m$ และ $T(N)$ เป็นจำนวนการคูณด้วยตัวคูณหมุนของ FFT จะได้ $T(1) = 0$, $T(N) = 2T(N/2) + N/2$ และ $T(N) = \frac{N}{2}\log_2 N$

บทพิสูจน์ (แบบเต็ม). ให้ $T(2^m) = m2^{m-1}$ เราตรวจด้วยการเหนี่ยวนำ $m = 0$ ได้ $T(1) = 0$ สมมติว่าจริงถึง $m-1$ จะได้ $T(2^m) = 2(m-1)2^{m-2} + 2^{m-1} = (m-1)2^{m-1} + 2^{m-1} = m2^{m-1}$ ซึ่งเป็นสูตรที่ต้องการ

ตารางนับการคูณ (ค่าสุดท้ายคำนวณด้วยสูตรและตรวจด้วยโค้ด)

N	ตรง N^2	FFT $\frac{N}{2}\log_2 N$	อัตราส่วน
8	64	12	5.3
64	4,096	192	21.3
1,024	1,048,576	5,120	204.8
2,048	4,194,304	11,264	372.4

ตารางต่อไปนี้แจกแจงที่ $N = 8$ ว่าการคูณมาจากระดับใด ระดับ 0 คือการแบ่งครั้งแรก

ระดับ	จำนวนส่วนย่อย	ขนาดแต่ละส่วน	การคูณด้วยตัวคูณหมุน
0	1	8	4
1	2	4	$2 \times 2 = 4$
2	4	2	$4 \times 1 = 4$
รวม			$12 = \frac{8}{2} \cdot 3$

ที่ $N = 2048$ FFT ใช้การคูณน้อยกว่าแบบตรงราว 372 เท่า Iserles 2009, §10.2–10.3, p. 214 ระบุว่า FFT ใช้ $O(N \log_2 N)$ การดำเนินการ ซึ่งสอดคล้องกับจำนวนนี้

เกร็ดเพิ่มเติม

บทความปี ค.ศ. 1965 ของ Cooley และ Tukey เสนออัลกอริทึมคำนวณอนุกรมฟูเรียร์เชิงซ้อน และกล่าวถึงประโยชน์ของขนาด $N = 2^m$ บนคอมพิวเตอร์เลขฐานสอง Cooley and Tukey 1965 วิธีที่เขียนในหัวข้อนี้เป็นการแบ่งครึ่งรัศมีสอง (radix-2) ซึ่งเป็นกรณีพิเศษของการแบ่ง $m = pq$ ที่ Atkinson อธิบายไว้ Atkinson 1989, §3.8, pp. 181–182

ฉบับร่าง · preprint · 2026-10-23

ภาพสองมิติ

ภาพเป็นตารางของค่า การแปลงสองมิติทำได้โดยแปลงทีละแถวแล้วทีละคอลัมน์ เพราะผลรวมสองชั้นสลับลำดับได้ ผลที่ได้เท่ากับการแปลงสองมิติเต็มรูปแบบ ซึ่งเราตรวจแล้วกับภาพสุ่มขนาด 8×8 ด้วยโค้ดของเรา บทนี้ไม่ได้ทำการบีบอัดภาพ เพียงแสดงว่าความคิดของการแปลงเดียวกันขยายไปถึงภาพได้

กลับไปที่แถบอีควอไลเซอร์

อีควอไลเซอร์ที่ง่ายที่สุดเลือกเฉพาะช่องที่แท่งสูงกว่าเกณฑ์ ที่เกณฑ์ 0.1 สัญญาณเสียงประสานรวมกับสัญญาณรบกวนที่ค่าเบี่ยงเบนมาตรฐาน 0.05 ให้ช่อง 10, 15 และ 20 พอดี ส่วนสัญญาณรบกวนอย่างเดียวให้แท่งสูงสุดเพียงราว 0.0067 (ตรวจด้วยโค้ดของเรา) ข้อกำหนดทำให้ค่าเรียบตามเวลา แท่งจึงไม่ได้ตอบสนองทันทีต่อการเปลี่ยนของเสียง

สรุปใจความ

- สัญญาณที่เป็นคาบแยกเป็นคลื่นความถี่ทีละช่องได้ด้วย DFT และกลับมาได้ครบ เพราะคลื่นของความถี่ต่างกันตั้งฉากกันบนจุดที่เว้นระยะเท่ากัน
- ความสูงของแท่งที่ช่อง k คือ $2|c_k|$ ซึ่งเท่ากับความดังของโทนนั้นเมื่อโทนตรงช่องพอดี และพลังงานของสัญญาณถูกแบ่งให้ทุกช่องโดยไม่หายไป (พาร์เซวัล)
- ข้อมูลจริงที่สุ่มบนคาบผ่านพหุนามตรีโกณมิติที่ทุกโหนด และสัมประสิทธิ์หาได้จาก DFT ทันที
- โทนที่ไม่ตรงช่องรั่วไปช่องข้างเคียง และความกว้างช่อง f_s/N กำหนดความละเอียดของการแยกความถี่
- ความถี่สูงกว่าไนควิสต์ $f_s/2$ ปรากฏเป็นความถี่ต่ำ (การทับซ้อน) แท่งจึงไม่บอกความจริงถ้าไม่คัดออกก่อน
- FFT แบ่งครึ่งซ้ำ ใช้ $\frac{N}{2}\log_2 N$ การคูณแทน N^2 จึงทำทุกเฟรมได้ กลับไปที่แถบอีควอไลเซอร์ แถบคือคำตอบของคำถามว่าความถี่ใดอยู่ในเสียง

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 4.3.1 (⋆) ให้ $N = 4$ และ $x = (1, 0, -1, 0)$ เมื่อ $\omega = e^{2\pi i/4} = i$ จงหา $c_k = \frac{1}{4}\sum_{j=0}^{3} x_j \omega^{-jk}$ สำหรับ $k = 0, 1, 2, 3$ แล้วตรวจว่า $x_j = \sum_{k=0}^{3} c_k \omega^{jk}$ ที่ $j = 0, 1, 2, 3$

แบบฝึกหัด 4.3.2 (⋆⋆) ให้ $x_0, \dots, x_{N-1}$ เป็นจำนวนจริงทุกตัว จงแสดงว่า $c_{N-k} = \overline{c_k}$ สำหรับ $1 \le k \le N-1$ จากนั้นหาแท่งที่ $k = 3$ ของสัญญาณ $x_j = 2\cos(2\pi \cdot 3j/8)$ เมื่อ $N = 8$

ตัวช่วยคิด

ใช้ $\omega^{-j(N-k)} = \omega^{jk}$ และลองดูว่าการเอาเครื่องหมายลบกลับเข้าไปในเลขชี้กำลังทำให้ได้อะไร

แบบฝึกหัด 4.3.3 (⋆) ที่อัตราการสุ่ม $44{,}100$ Hz และ $N = 2048$ จงหา

1. ความกว้างของช่องความถี่ (ระยะห่างระหว่างช่องติดกัน)
2. ความถี่ของช่อง $k = 46$ และ $k = 47$
3. โทน 1000 Hz อยู่ใกล้ช่องใดมากกว่ากัน

แบบฝึกหัด 4.3.4 (⋆⋆) ที่อัตราการสุ่ม 8000 Hz เราสุ่มโคไซน์ความถี่ 5000 Hz ที่ $N = 16$ จุด $j = 0, \dots, 15$ โดยใช้ $\cos(2\pi \cdot 5000\, j/8000)$

1. แสดงว่าค่าที่ได้เหมือนกับค่าของโคไซน์ความถี่ 3000 Hz ที่จุดเดียวกันทุกจุด
2. แท่งสูงสุดของการแปลงปรากฏที่ช่องใด และความถี่ของช่องนั้นเท่าไร

แบบฝึกหัด 4.3.5 (⋆⋆) จงนับจำนวนการคูณเชิงซ้อนที่ต้องใช้ในการคำนวณ DFT ของ $N = 16$ จุด ทั้งด้วยสูตรตรง (ทุกคู่ (j, k)) และด้วยสูตรเวียนเกิดของ FFT โดยให้จำนวนการคูณของ FFT ที่ขนาด n คือ $T(n)$ ซึ่ง $T(1) = 0$ และ $T(n) = 2T(n/2) + n/2$ แล้วเปรียบเทียบสองจำนวนนี้

แบบฝึกหัด 4.3.6 (⋆ ⋆ ⋆) ที่อัตราการสุ่ม $44{,}100$ Hz จงแสดงว่าไม่มี N ที่เป็นกำลังของ 2 ซึ่งทำให้โทน 220 Hz ตรงกับช่องความถี่พอดี

ตัวช่วยคิด

โทนตรงช่องพอดีเมื่อ $220N/44100$ เป็นจำนวนเต็ม ลองหารทั้งสองข้างด้วย 20 แล้วดูตัวประกอบเฉพาะของ 2205 เทียบกับของ N

แบบฝึกหัด 4.3.7 (⋆ ⋆ ⋆) กลับไปที่แถบของอีควอไลเซอร์ สมมติว่าแอปเปลี่ยนไปใช้ $N = 512$ แทน 2048 ที่อัตราการสุ่ม $44{,}100$ Hz จงหา

1. ความกว้างของช่องความถี่
2. ช่องที่โทน 430.66 Hz ตรงพอดี
3. โทน 215.33 Hz อยู่ที่ตำแหน่งกี่ช่อง และผลนี้ทำให้เกิดอะไรขึ้นกับแท่ง
4. ช่วงเวลาที่ใช้วิเคราะห์ N/f_s เทียบกับของ $N = 2048$ เป็นกี่เท่า

แบบฝึกหัด 4.3.8 (⋆ ⋆ ⋆) ใช้ฟังก์ชันคัดช่องที่มากับโค้ดของบทนี้ (`equaliser_keep`) กับสัญญาณ x คือเสียงประสานสามโทนของหัวข้อนี้บวกสัญญาณรบกวนแบบปกติที่ค่าเบี่ยงเบนมาตรฐาน 0.05 ขนาด 2048 จุด จากตัวสร้างเลขสุ่มที่มีเมล็ด 7

1. เกณฑ์ 0.1 และ 0.6 เลือกช่องใดบ้าง
2. เหตุใดโทนที่ช่อง 15 จึงหายไปเมื่อเกณฑ์เป็น 0.6
3. หากสัญญาณมีแต่สัญญาณรบกวน (ไม่มีเสียงประสาน) แท่งสูงสุดควรมีขนาดเท่าใด เทียบกับเกณฑ์ 0.1

ฉบับร่าง · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 4.3.1. ที่ $\omega = i$ จะได้ $\omega^{-jk} = (-i)^{jk}$ และ $(-i)^0 = 1$, $(-i)^1 = -i$, $(-i)^2 = -1$, $(-i)^3 = i$ ดังนั้น

1. $c_0 = \frac{1}{4}(x_0 + x_1 + x_2 + x_3) = \frac{1}{4}(1 + 0 - 1 + 0) = 0$
2. $c_1 = \frac{1}{4}\sum_j x_j(-i)^j = \frac{1}{4}\big(1 \cdot 1 + (-1)(-1)\big) = \frac{1}{4}(1 + 1) = \frac{1}{2}$ เพราะ $j = 2$ ให้ $(-i)^2 = -1$
3. $c_2 = \frac{1}{4}\sum_j x_j(-i)^{2j} = \frac{1}{4}\big(1 \cdot 1 + (-1) \cdot (-i)^4\big) = \frac{1}{4}(1 - 1) = 0$ เพราะ $(-i)^4 = 1$
4. $c_3 = \frac{1}{4}\sum_j x_j(-i)^{3j} = \frac{1}{4}\big(1 + (-1)(-i)^6\big) = \frac{1}{4}\big(1 - (-1)\big) = \frac{1}{2}$ เพราะ $(-i)^6 = -1$

จึงได้ $c = (0, \frac{1}{2}, 0, \frac{1}{2})$ ตรวจกลับ: $j = 0$ ได้ $c_1 + c_3 = 1$ ตรงกับ x_0; $j = 1$ ได้ $\frac{1}{2}i + \frac{1}{2}(-i) = 0$ ตรงกับ x_1; $j = 2$ ได้ $\frac{1}{2}(-1) + \frac{1}{2}(-1) = -1$ ตรงกับ x_2; $j = 3$ ได้ 0 ตรงกับ x_3

เฉลย 4.3.2. เพราะ $\omega^N = 1$ จึง $\omega^{-j(N-k)} = \omega^{-jN}\omega^{jk} = \omega^{jk}$ ดังนั้น

$$c_{N-k} = \frac{1}{N}\sum_j x_j \omega^{-j(N-k)} = \frac{1}{N}\sum_j x_j \omega^{jk}.$$

เนื่องจาก x_j เป็นจำนวนจริง ได้ $\overline{c_k} = \frac{1}{N}\sum_j x_j\,\overline{\omega^{-jk}} = \frac{1}{N}\sum_j x_j \omega^{jk}$ ซึ่งเท่ากับ c_{N-k} พอดี

สำหรับ $x_j = 2\cos(2\pi \cdot 3j/8)$ ที่ $N = 8$: เขียน $2\cos\theta = e^{i\theta} + e^{-i\theta}$ จะได้ $x_j = \omega^{3j} + \omega^{-3j}$ เมื่อ $\omega = e^{2\pi i/8}$ จากการกลับของ DFT ได้ $c_3 = 1$ และ $c_5 = 1$ (เพราะ $5 = 8 - 3$) แท่งที่ $k = 3$ มีความสูง $2|c_3| = 2$ ซึ่งเท่ากับแอมพลิจูด $A = 2$ ตรงกับหัวข้อนี้

เฉลย 4.3.3.

1. ความกว้างของช่องคือ $f_s/N = 44100/2048 \approx 21.53$ Hz
2. ช่อง $k = 46$ อยู่ที่ $46 \times 21.53 \approx 990.5$ Hz และช่อง $k = 47$ อยู่ที่ $47 \times 21.53 \approx 1012.0$ Hz
3. $1000/21.53 \approx 46.44$ จึงอยู่ใกล้ช่อง 46 มากกว่าช่อง 47 (ระยะ 9.5 Hz เทียบกับ 12.0 Hz)

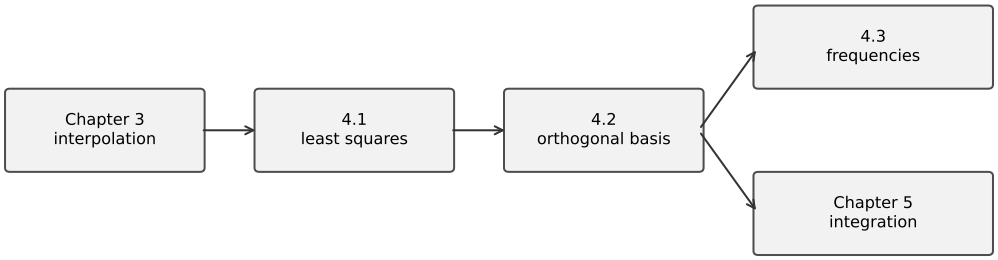
เฉลย 4.3.4.

1. เพราะ $5000 = 8000 - 3000$ จึง $\cos(2\pi \cdot 5000j/8000) = \cos\big(2\pi j - 2\pi \cdot 3000j/8000\big) = \cos(2\pi \cdot 3000j/8000)$ สำหรับทุก j จึงได้ค่าเหมือนกันทุกจุด
2. เนื่องจากค่าเหมือนโทน 3000 Hz ความกว้างของช่องคือ $8000/16 = 500$ Hz และ $3000/500 = 6$ จึงตรงช่อง $k = 6$ พอดี แท่งสูงสุดจึงปรากฏที่ช่อง 6 และความถี่ของช่องนั้นคือ 3000 Hz ไม่ใช่ 5000 Hz (ผลนี้ตรวจด้วยโค้ดของเรา)

เฉลย 4.3.5. ผลคูณตรงมี $N^2 = 16^2 = 256$ ครั้ง สูตรเวียนเกิดให้ $T(1) = 0$, $T(2) = 2 \cdot 0 + 1 = 1$, $T(4) = 2 \cdot 1 + 2 = 4$, $T(8) = 2 \cdot 4 + 4 = 12$ และ $T(16) = 2 \cdot 12 + 8 = 32$ ตรงกับ $\frac{N}{2}\log_2 N = 8 \cdot 4 = 32$ จึงใช้การคูณ 32 ครั้ง น้อยกว่าแบบตรงราว 8 เท่า (อัตราส่วน $256/32 = 8$)

เฉลย 4.3.6. โทนตรงช่องพอดีเมื่อ $220N/44100$ เป็นจำนวนเต็ม ซึ่งเทียบเท่ากับ $44100 \mid 220N$ หารทั้งสองข้างด้วย 20 ได้ $2205 \mid 11N$ ตัวประกอบเฉพาะของ 2205 คือ $3^2 \cdot 5 \cdot 7^2$ ซึ่งไม่มี 11 จึงได้ $2205 \mid N$ แต่ $N = 2^m$ ไม่มีตัวประกอบ 3 เลย จึงหารด้วย 2205 ไม่ลงสำหรับทุก $m \geq 0$ (การตรวจด้วยโค้ดถึง $m = 20$ ได้ผลเช่นเดียวกัน)

ฉบับร่าง Preprint · 2026-10-23



รูปที่ 4.10: ลำดับหัวข้อของบทที่ 4 อีกครั้ง หัวข้อ 4.3 ใช้ความตั้งฉากจากหัวข้อ 4.2 และส่งต่อไปบทที่ 5 ผ่านกฎบนจุดที่เว้นระยะเท่ากัน (สร้างจาก `numanal/ch04.py`)

เฉลย 4.3.7.

1. $44100/512 \approx 86.13$ Hz
2. $430.66/86.13 = 5$ จึงตรงช่อง $k = 5$ พอดี
3. $215.33/86.13 = 2.5$ จึงอยู่กึ่งกลางระหว่างช่อง 2 กับ 3 ไม่ตรงช่อง ผลคือแท่งจะกระจายไปยังช่องข้างเคียงทั้งสองด้าน เหมือนการรั่วของสเปกตรัมในหัวข้อนี้
4. N/f_s ของ 512 คือ $512/44100 \approx 11.6$ ms และของ 2048 คือ ≈ 46.4 ms ดังนั้นช่วงเวลาเป็นหนึ่งในสี่ ช่องจึงกว้างขึ้นสี่เท่า และความละเอียดของความถี่ลดลงเท่ากัน

เฉลย 4.3.8.

1. เกณฑ์ 0.1 เลือกช่อง 10, 15 และ 20 (ความถี่ 215.33, 322.99 และ 430.66 Hz) เกณฑ์ 0.6 เลือกเฉพาะช่อง 10 เพราะแท่งที่ช่อง 10 สูง 1 ส่วนแท่งอีกสองแท่งสูง 0.5 และ 0.25
2. แท่งที่ช่อง 15 มีความสูงราว 0.5 ซึ่งน้อยกว่าเกณฑ์ 0.6 จึงถูกตัดทิ้ง ไม่ได้หายเพราะสัญญาณรบกวน
3. จากตัวสร้างเลขสุ่มที่มีเมล็ด 7 วัดได้ว่าแท่งสูงสุดของสัญญาณรบกวนอย่างเดียวคือ 0.0067 ซึ่งต่ำกว่าเกณฑ์ 0.1 มาก (ตรวจด้วยโค้ดของเรา) ดังนั้นเกณฑ์ 0.1 แยกเสียงจริงออกจากสัญญาณรบกวนได้

ฉบับร่าง · Preprint · 2026-10-23

สรุปบทที่ 4 ในหน้าเดียว

บทนี้ถามว่า เมื่อข้อมูลมีสัญญาณรบกวน หรือเมื่อเรารู้ฟังก์ชันทั้งหมด เราจะหาเส้นหรือพหุนามที่ใกล้ที่สุดได้อย่างไร และความถี่ใดซ่อนอยู่ในสัญญาณ สูตรและตัวเลขในตารางมาจากทฤษฎีบทในบทนี้หรือจากโค้ดของเรา

สูตรและตัวเลขสำคัญ

เรื่อง	สูตร	ตัวอย่างในบทนี้
กำลังสองน้อยสุด	$A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$	เส้น $1.09 + 1.94\,t$
หน้าต่างกำลังสองห้าจุด	$(-3, 12, 17, 12, -3)/35$	ยอด 0.966
สัมประสิทธิ์ของฐานเชิงตั้งฉาก	$c_k = (f, \varphi_k)/(\varphi_k, \varphi_k)$	$p^* = 1.5532\,t - 0.5622\,t^3$
มินิแมกซ์ดีกรีสาม	$E = 0.0044917$	หกจุดสลับเครื่องหมาย
ผลการแปลงฟูเรียร์แบบไม่ต่อเนื่อง	$c_k = \frac{1}{N}\sum_j x_j \omega^{-jk}$	แท่ง 1, $\frac{1}{2}$, $\frac{1}{4}$
แท่งของอีควอไลเซอร์	$2\lvert c_k\rvert = A$	ช่อง 10, 15, 20
พาร์เซวัล	$\sum_j \lvert x_j\rvert^2 = N\sum_k \lvert c_k\rvert^2$	เสียงประสาน $\frac{1}{N}\sum_j x_j^2 = 0.65625$
การประมาณค่าเชิงตรีโกณมิติ	ผ่านทุกโหนด $t_j = 2\pi j/N$	$\max\lvert g - p\rvert = 4.7 \times 10^{-10}$ ที่ $N = 32$
ความกว้างของช่อง	f_s/N	21.53 Hz ที่ $N = 2048$
ความถี่ไนควิสต์	$f_s/2$	$22{,}050$ Hz ที่ $44{,}100$ Hz
การทับซ้อน	$f + m f_s$ ให้ค่าเหมือน f	70 Hz กับ 30 Hz ที่ 100 Hz
FFT	$\frac{N}{2}\log_2 N$ การคูณ	$11{,}264$ แทน $4{,}194{,}304$ ที่ $N = 2048$

วิธีที่เทียบกันในบทนี้

วิธี	ใช้เมื่อ	ข้อควรระวัง
กำลังสองน้อยสุด (4.1)	ข้อมูลมีสัญญาณรบกวน	สมการปรกติมีคอนดิชันนัมเบอร์เป็นกำลังสอง
เทย์เลอร์ (4.2)	รู้ฟังก์ชันที่จุดเดียว	ค่าคลาดเคลื่อนโตเร็วที่ปลายช่วง
เลอฌ็องดร์และเชบีเชฟ (4.2)	ต้องการค่าคลาดเคลื่อนรวมหรือสูงสุดน้อย	ฐานดีไม่ได้แปลว่าพหุนามดี
DFT ตรง (4.3)	สัญญาณเป็นคาบ	N^2 การคูณ
FFT (4.3)	สัญญาณเป็นคาบ ยาวมาก	ต้องการ N เป็นกำลังของ 2

อ่านต่อ. บทที่ 5 ถามว่า จากข้อมูลไม่กี่จุด เราจะหาอนุพันธ์และพื้นที่ใต้กราฟได้อย่างไร ในหัวข้อ 4.3 เราเห็นแล้วว่า สำหรับฟังก์ชันคาบ การใช้จุดที่เว้นระยะเท่ากันให้ผลแม่นมาก Atkinson 1989, §3.8, p. 179 บทที่ 5 จะเทียบกับกฎที่เลือกจุดพิเศษเอง เช่นกฎเกาส์ซึ่งใช้รากของพหุนามเชิงตั้งฉาก Atkinson 1989, §4.4, Thm 4.4, p. 213

ฉบับร่าง · preprint · 2026-10-02

บทที่ 5

อนุพันธ์และพื้นที่จากข้อมูลไม่กี่จุด

สถานการณ์สมมติ: รถผีที่วิ่งซ้ำ

สมมติว่าเกมแข่งรถของเรามีโหมดรถผี คือการเล่นรอบก่อนของเราถูกเล่นซ้ำ และมีรถอีกคันวิ่งตามเส้นทางเดิมให้เห็นบนจอ รถผีคันนี้วิ่งบนทางตรงยาว และเกมนับระยะเป็นเมตรจากจุดเริ่มของการเล่นซ้ำ

ระบบของเราไม่ได้เก็บสูตรการวิ่งของรถผี มันเก็บแค่ **ตำแหน่ง** ของรถทุกเฟรม ซึ่งมี 60 เฟรมต่อวินาที แต่หน้าจอต้องแสดง **ความเร็ว** ของรถผี และ **ระยะที่รถผีวิ่งไปแล้ว** ด้วย คำถามแรกคือ จากตำแหน่งที่เก็บไว้ เราจะได้ความเร็วออกมาเท่าไร

ลองดูตัวอย่างนี้ รถผีวิ่งจริงด้วยความเร็วระหว่าง 15 ถึง 25 เมตรต่อวินาที ถ้าเริ่มนับจากจุดกำเนิดของโลก ความเร็วที่อ่านได้จะใกล้ความเร็วจริงมาก แต่ถ้ารถผีเริ่มไกลจากจุดกำเนิดราว 10^6 เมตร และตำแหน่งเก็บด้วย**เลขความแม่นเดี่ยว** (single precision) ความเร็วที่อ่านได้จะเป็นได้แค่ $15, 18.75, 22.5$ หรือ 26.25 เมตรต่อวินาที ค่าเหล่านี้กระโดดเป็นขั้น ทั้งที่รถวิ่งเปลี่ยนความเร็วอย่างช้า ๆ (ตัวเลขทั้งหมดมาจากโค้ดของเรา รถผีนี้เป็นการสร้างขึ้นเพื่อการเรียนและไม่ได้บอกอะไรเกี่ยวกับเกมใด)

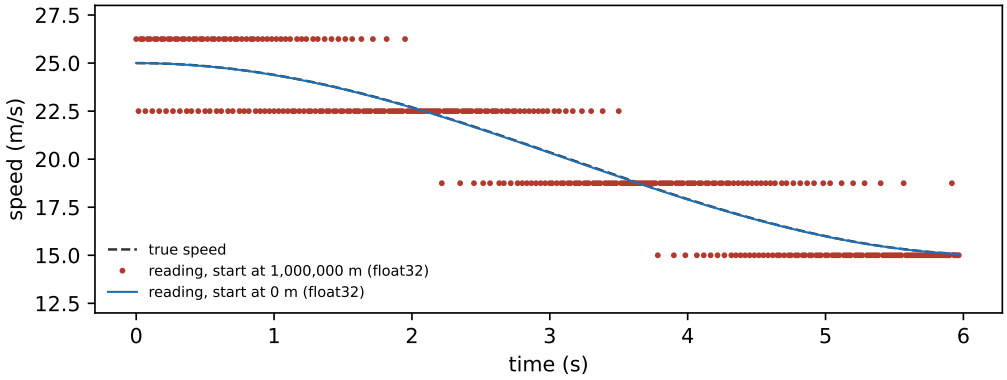
คำถามของบทนี้คือ เราหาอนุพันธ์ (ความเร็ว) และพื้นที่ (ระยะทาง) จากข้อมูลที่เป็นจุดไม่กี่จุดได้อย่างไร และความผิดมาจากไหนบ้าง

ภาพรวมบท

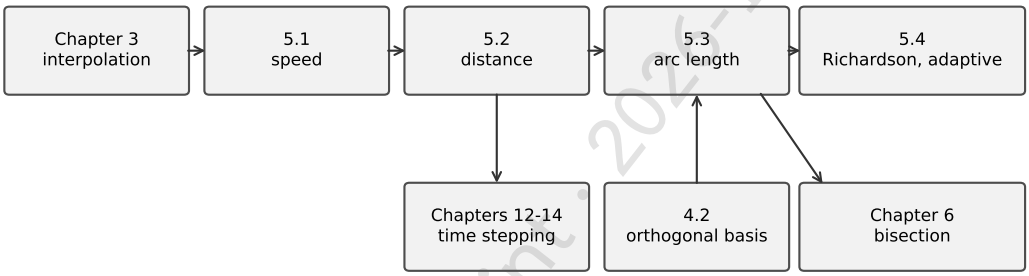
บทที่ 4 ทำให้เรามีเส้นที่ดีที่สุดผ่านข้อมูลที่มีสัญญาณรบกวน ในบทนี้ข้อมูลเป็นตำแหน่งที่เก็บมาแล้ว และเราจะถามเรื่องที่สัมพันธ์กันสองอย่าง

- หัวข้อ 5.1: จากตำแหน่งไปหาความเร็ว คือการหา **อนุพันธ์** (derivative) ของตำแหน่ง ซึ่งเป็นความชันของกราฟตำแหน่งกับเวลา เราประมาณมันด้วยผลต่างของจุดที่ใกล้กัน และต้องต่อรองระหว่างค่าคลาดเคลื่อนจากการประมาณกับค่าคลาดเคลื่อนจากการเก็บตัวเลข
- หัวข้อ 5.2: จากความเร็วไปหาระยะทาง คือการหาพื้นที่ใต้กราฟความเร็วจากค่าที่จุดไม่กี่จุด ด้วยกฎเชิงสี่เหลี่ยมคางหมูและกฎของซิมป์สัน
- หัวข้อ 5.3: อีกด้านหนึ่งของเรื่องเดียวกัน กล้องที่วิ่งด้วยความเร็วคงที่ตามเส้นโค้งต้องรู้ว่าระยะตามเส้นโค้งเท่าไรจึงจะวางกล้องถูกที่ และระยะนั้นคือปริพันธ์ที่ไม่มีสูตรปิด
- หัวข้อ 5.4: รู้ได้อย่างไรว่าคำตอบแม่นพอ คือใช้ผลต่างของการคำนวณสองระดับเป็นตัวประมาณค่าคลาดเคลื่อน แล้วแบ่งช่วงเฉพาะที่ที่ความเร็วเปลี่ยนเร็ว

ฉบับร่าง · Preprint · 2026-05-23



รูปที่ 5.1: ความเร็วที่อ่านได้จากตำแหน่งที่เก็บทุกเฟรม เส้นประคือความเร็วจริง ซึ่งซ้อนอยู่กับเส้นฟ้าเกือบทั้งช่วง จุดแดงคือการอ่านจากรถผีที่เริ่มไกลจากจุดกำเนิด 10^6 เมตร (สร้างจาก `numanal/ch05.py`)



รูปที่ 5.2: ลำดับของหัวข้อในบทที่ 5 และการเชื่อมกับบทที่ 3, บทที่ 4 และบทที่ 12--14 (สร้างจาก `numanal/ch05.py`)

เราเริ่มจากรถผีที่วิ่งบนทางตรง ซึ่งเป็นการหาความเร็วจากตำแหน่ง แล้วจะเจอ*อีกด้านหนึ่ง*ในหัวข้อ 5.3 คือการหาตำแหน่งจากระยะตามเส้นโค้ง

ภาพต่อไปนี้แสดงลำดับของหัวข้อในบทนี้ และการเชื่อมกับบทที่ 4 และบทที่ 12--14

ฉบับร่าง · Preprint · 2026-10-23

5.1 ความเร็วจากตำแหน่ง

สถานการณ์สมมติ: ความเร็วที่ไม่ได้ถูกเก็บไว้

ไฟล์รถผีมีแค่ตำแหน่งที่เฟรม $0, 1, 2, \ldots$ ไม่มีค่าความเร็วอยู่ในไฟล์เลย เมื่อผู้เล่นมองความเร็วบนจอ โปรแกรมต้องคำนวณจากตำแหน่งสองเฟรมที่อยู่ติดกัน คำถามคือการคำนวณนั้นให้ค่าใกล้ความเร็วจริงแค่ไหน และทำไมการใช้เฟรมที่ถี่กว่าจึงไม่ได้ทำให้แม่นขึ้นเสมอไป

ภาพรวมก่อนอ่าน

ความเร็วคือความชันของกราฟตำแหน่งที่เวลาหนึ่ง ความชันนั้นคือความชันของ*เส้นสัมผัส*ที่จุดนั้น แต่เรามีแค่จุดไม่กี่จุด จึงใช้ความชันของ*เส้นตัด*ที่ลากผ่านจุดสองจุดที่อยู่ใกล้กันแทน เส้นตัดนี้ใกล้เส้นสัมผัสเมื่อจุดทั้งสองใกล้กัน แต่เมื่อจุดใกล้

กันมาก ตัวเลขจะเริ่มมีปัญหาของมันเอง

จากเส้นตัดสู่เส้นสัมผัส

ให้ $X(t)$ เป็นตำแหน่งของรถผีที่เวลา t วินาที ความเร็วที่เวลา t คือ $X'(t)$ ซึ่งเป็นความชันของกราฟ X ที่จุด t เราไม่มีสูตรของ X ที่โปรแกรมรู้ มีแต่ค่าที่เก็บไว้ที่เวลา $t_k = k/60$ ดังนั้นเราต้องประมาณ $X'(t)$ จากค่าเหล่านี้ การหาสูตรผลต่างจากพหุนามที่ประมาณฟังก์ชันเป็นแนวคิดในตำรา Atkinson 1989, §5.7, pp. 315--317 ส่วนตัวดำเนินการผลต่างไปข้างหน้า ถอยหลัง และกลาง มีนิยามในตำรา Iserles 2009, §8.1, pp. 139--140

ความคิดที่ง่ายที่สุดคือ ลากเส้นตรงผ่านจุด $(t, X(t))$ กับจุดที่อยู่ใกล้กันอีกจุดหนึ่ง แล้วใช้ความชันของเส้นนั้น ความชันนี้คือระยะที่เปลี่ยนไปหารด้วยเวลาที่ใช้ ซึ่งเป็นเรื่องเดียวกับความเร็วเฉลี่ยในช่วงสั้น ๆ ระหว่างสองเวลานั้น

นิยาม 5.1.1: ผลหารเชิงผลต่างสองแบบ

ให้ X เป็นฟังก์ชันของเวลา และให้ $h > 0$ เป็น**ขนาดขั้น** (step size) ของเวลา

- **ผลหารเชิงผลต่างไปข้างหน้า** (forward difference quotient) คือ

$$D_+X(t) = \frac{X(t+h) - X(t)}{h}.$$

- **ผลหารเชิงผลต่างกลาง** (central difference quotient) คือ

$$D_cX(t) = \frac{X(t+h) - X(t-h)}{2h}.$$

อ่านแบบเข้าใจง่าย

- ผลหารไปข้างหน้าใช้จุดที่ t กับจุดที่อยู่ข้างหน้าหนึ่งขั้น ผลหารกลางใช้จุดสองฝั่งของ t ที่ห่างจาก t เท่ากัน
- ทั้งสองเป็นความชันของเส้นตัด ตัวเศษคือระยะที่เปลี่ยนไป ตัวส่วนคือเวลาที่ใช้ไป
- ผลหารกลางได้ความชันของเส้นตัดที่ *อยู่กลางช่วง* จึงน่าจะตรงกว่า ซึ่งเราจะพิสูจน์ในทฤษฎีบท 5.1.1
- ถ้า $h = 1/60$ วินาที ผลหารไปข้างหน้าที่เฟรม k คือ $60\,(X_{k+1} - X_k)$ ตัวเลขนี้คือ *ความเร็วที่อ่านได้* ของบทนี้

ลองเอง: ลูกบาศก์ที่จุด 1

ให้ $X(t) = t^3$ เราจะประมาณ $X'(1) = 3$ ลองคิดก่อนดูผลข้างล่าง

- ไปข้างหน้า $h = 0.1$: $\dfrac{1.1^3 - 1^3}{0.1} = \dfrac{0.331}{0.1} = 3.31$ ค่าคลาดเคลื่อนคือ 0.31
- กลาง $h = 0.1$: $\dfrac{1.1^3 - 0.9^3}{0.2} = \dfrac{1.331 - 0.729}{0.2} = 3.01$ ค่าคลาดเคลื่อนคือ 0.01

ทำไมผลกลางแม่นกว่ามากขนาดนี้ ลองตรวจดูว่าค่าคลาดเคลื่อนไปข้างหน้า 0.31 อยู่ในช่วงที่ทฤษฎีบทข้างล่างทำนายไว้หรือไม่ (คำใบ้: $X''(t) = 6t$)

ฉบับร่าง · preprint · 2026-10-23

ความผิดแบบแรกคือการตัดปลาย

ความผิดนี้ไม่ได้มาจากเครื่องคอมพิวเตอร์ มาจากการแทนความชันของเส้นโค้งด้วยความชันของเส้นตัด ความผิดนี้เกิดจากการตัดพจน์ที่มี h ยกกำลังสูงกว่าทิ้งไป เราจึงเรียกมันว่า **ค่าคลาดเคลื่อนตัดปลาย** (truncation error)

ทฤษฎีบท 5.1.1: ค่าคลาดเคลื่อนตัดปลายของผลหารทั้งสองแบบ

1. ถ้า X'' ต่อเนื่องบน $[t, t+h]$ จะมี $\xi \in (t, t+h)$ ที่ทำให้
$$D_+X(t) - X'(t) = \frac{h}{2}X''(\xi).$$
2. ถ้า X''' ต่อเนื่องบน $[t-h, t+h]$ จะมี $\xi \in (t-h, t+h)$ ที่ทำให้
$$D_cX(t) - X'(t) = \frac{h^2}{6}X'''(\xi).$$

อ่านแบบเข้าใจง่าย

- ข้อ 1 บอกว่าค่าคลาดเคลื่อนไปข้างหน้าเป็นสัดส่วนกับ h ถ้าลด h ลงครึ่งหนึ่ง ค่าคลาดเคลื่อนก็ลดลงครึ่งหนึ่ง เราเขียนว่า $O(h)$ เพื่อบอกลำดับนี้
- ข้อ 2 บอกว่าค่าคลาดเคลื่อนกลางเป็นสัดส่วนกับ h^2 ถ้าลด h ลงครึ่งหนึ่ง ค่าคลาดเคลื่อนจะลดเป็นหนึ่งในสี่ เขียนว่า $O(h^2)$
- เราไม่รู้ค่า ξ แต่รู้ว่ามันอยู่ในช่วงไหน และนั่นพอสำหรับการประมาณขอบเขตของค่าคลาดเคลื่อนสัมบูรณ์

พิสูจน์. **แนวคิด.** ใช้ทฤษฎีบทของเทย์เลอร์ที่มีพจน์เศษ แล้วหักล้างพจน์ที่ไม่ต้องการ

1. (ข้อ 1) เพราะ X'' ต่อเนื่อง ทฤษฎีบทของเทย์เลอร์ให้ $X(t+h) = X(t) + hX'(t) + \frac{h^2}{2}X''(\xi)$ สำหรับบาง $\xi \in (t, t+h)$
2. ย้ายข้างและหารด้วย h จะได้ $D_+X(t) - X'(t) = \frac{h}{2}X''(\xi)$ ซึ่งเป็นข้อ 1
3. (ข้อ 2) เพราะ X''' ต่อเนื่อง เขียนเทย์เลอร์ที่จุด $t+h$ และ $t-h$ ได้ว่า
$$X(t \pm h) = X(t) \pm hX'(t) + \frac{h^2}{2}X''(t) \pm \frac{h^3}{6}X'''(\xi_\pm),$$
โดย $\xi_+ \in (t, t+h)$ และ $\xi_- \in (t-h, t)$
4. ลบสมการของ $t-h$ ออกจากสมการของ $t+h$ พจน์ $X(t)$ และ $\frac{h^2}{2}X''(t)$ หักล้างกันหมด จึงได้ $X(t+h) - X(t-h) = 2hX'(t) + \frac{h^3}{6}\big(X'''(\xi_+) + X'''(\xi_-)\big)$
5. หารด้วย $2h$ จะได้ $D_cX(t) - X'(t) = \frac{h^2}{12}\big(X'''(\xi_+) + X'''(\xi_-)\big)$
6. ค่าเฉลี่ยของ $X'''(\xi_+)$ กับ $X'''(\xi_-)$ อยู่ระหว่างค่าต่ำสุดกับค่าสูงสุดของ X''' บนช่วงระหว่าง ξ_- กับ ξ_+ เมื่อ X''' ต่อเนื่อง ทฤษฎีบทค่ากลางให้มี ξ ในช่วงนั้นที่ $X'''(\xi)$ เท่ากับค่าเฉลี่ยนี้ จึงได้ $\frac{h^2}{12} \cdot 2X'''(\xi) = \frac{h^2}{6}X'''(\xi)$
7. จุด $\xi_\pm$ อยู่ใน $(t-h, t+h)$ ดังนั้น ξ อยู่ในช่วงนั้นด้วย ซึ่งเป็นข้อ 2

ฉบับร่าง Preprint 2026-10-23

สรุป: ผลหารทั้งสองต่างจากอนุพันธ์จริงด้วยพจน์ที่มี h ยกกำลังหนึ่งและสอง และสัมประสิทธิ์คือความโค้งของเส้นโค้งที่ระดับสูงกว่า □

ข้อควรระวัง: ค่าคลาดเคลื่อนตัดปลายไม่ใช่ค่าคลาดเคลื่อนทั้งหมด

ทฤษฎีบทข้างบนบอกว่า h เล็กลงจะลดค่าคลาดเคลื่อนตัดปลาย แต่ไม่ได้บอกว่าควรเลือก h เล็กที่สุดเท่าที่ทำได้ เพราะตัวเลขที่เก็บไว้มีค่าคลาดเคลื่อนของมันเอง และการหารด้วย h เล็กมากจะขยายค่าคลาดเคลื่อนนั้น ดูตัวอย่างในหัวข้อถัดไป

ที่ 10^6 เมตรด้วยเลขความแม่นเดี่ยว ถ้า $h = 2^{-20}$ วินาที ตัวเลขสองค่าที่เก็บไว้เป็นค่าเดียวกัน ผลหารจึงให้ 0 ทั้งที่รถวิ่งด้วยความเร็วราว 24.4 เมตรต่อวินาที ค่าคลาดเคลื่อนตัดปลายเล็กมากในกรณีนี้ แต่ค่าคลาดเคลื่อนรวมเป็น -24.4

ความผิดแบบที่สองคือตัวเลขที่เก็บไว้ไม่แม่นเป๊ะ

เลขจุดลอยตัวที่เก็บตำแหน่งไว้ไม่สามารถแทนทุกจำนวนจริงได้ ค่าที่เก็บไว้จึงต่างจากตำแหน่งจริงเล็กน้อย เราเขียนแทนความต่างนี้ด้วย ε และกำหนดขอบเขต

$$|\varepsilon| \le e.$$

สำหรับเลขความแม่นเดี่ยว ตัวเลขที่เก็บได้ที่อยู่ใกล้กันห่างกันเป็นช่องว่างคงที่ในแต่ละช่วงขนาด จึงใช้ครึ่งหนึ่งของช่องว่างนั้นเป็น e ได้ (ดูบทที่ 1) ที่ 10^6 เมตร ช่องว่างคือ 0.0625 เมตร และ $e = 0.03125$ เมตร ส่วนที่จุดกำเนิด ช่องว่างเล็กกว่านี้มาก การวิเคราะห์ค่าคลาดเคลื่อนของค่าที่วัดได้ในลักษณะนี้คล้ายกับที่ Atkinson 1989, §5.7, pp. 318--319 ทำกับอนุพันธ์อันดับสอง ส่วนการพิสูจน์สำหรับอนุพันธ์อันดับหนึ่งในบทนี้เป็นของเราเอง

อ่านแบบเข้าใจง่าย

- เราจะคิดเฉพาะค่าคลาดเคลื่อนจากการเก็บตัวเลขกับค่าคลาดเคลื่อนตัดปลาย และสมมติว่าการลบและการหารไม่เพิ่มค่าคลาดเคลื่อนอีก ในโค้ดของเราผลต่างและการหารทำด้วย 64 บิต จึงทำให้ค่าคลาดเคลื่อนจากขั้นนี้เล็กมาก
- การลบค่าสองค่าที่อยู่ใกล้กันมากมักแม่นพอดี ค่าคลาดเคลื่อนที่เหลืออยู่ในค่าที่เก็บไว้ จะถูกขยายเมื่อหารด้วย h เล็ก

ข้อควรระวัง: การลบแบบหักล้างกับตัวหารเล็ก

บทที่ 2 เรียกการลบจำนวนสองจำนวนที่ใกล้กันมากว่า **การลบแบบหักล้าง** (cancellation) ซึ่งทำให้หลักที่แม่นหายไป ในบทนี้ตัวเศษของผลหารมีค่าคลาดเคลื่อนเล็กน้อย แต่ตัวหาร h ที่เล็กมากจะขยายค่าคลาดเคลื่อนนั้น จึงอย่าลืมดูทั้งสองส่วนพร้อมกัน ห้ามตัดสินใจจากตัวเศษอย่างเดียว อย่าสับสนกับ **ค่าคลาดเคลื่อนไปข้างหน้า** (forward error) ในบทที่ 2 ซึ่งเป็นชื่อของค่าคลาดเคลื่อนของผลลัพธ์ทั้งหมด

ลองเอง: ทำไมอ่านได้ 26.25 ไม่ใช่ 25

ที่เฟรม 0 รถผีอยู่ที่ 10^6 เมตรพอดี ที่เฟรม 1 ระยะที่รถวิ่งไปจริงคือ $X(1/60) - X(0) \approx 0.416667$ เมตร ลองหาตัวเลขที่เก็บได้ (ผลคูณของ 0.0625) ที่อยู่ใกล้ $10^6 + 0.416667$ ที่สุด แล้วคำนวณ $60 \times (\text{ตัวเลขนั้น} - 10^6)$ คุณจะได้ค่า 26.25 ซึ่งผิดจากความเร็วจริง 25 เมตรต่อวินาทีราว 1.25 ค่าคลาดเคลื่อนสัมบูรณ์นี้น้อยกว่าขอบเขตที่เราจะคำนวณข้างล่าง (3.77) หรือไม่

ร่าง · Preprint · 2026-10-23

ทฤษฎีบท 5.1.2: ค่าคลาดเคลื่อนจากการเก็บตัวเลข

ให้ค่าที่เก็บไว้คือ $\tilde{X} = X + \varepsilon$ โดย $|\varepsilon| \le e$ สำหรับทุกค่าที่ใช้ในผลหาร จะได้

$$\left|\tilde{D}_+X(t) - D_+X(t)\right| \le \frac{2e}{h}, \qquad \left|\tilde{D}_cX(t) - D_cX(t)\right| \le \frac{e}{h}.$$

พิสูจน์. **แนวคิด.** ค่าคลาดเคลื่อนของการเก็บแต่ละค่าถูกหารด้วย h และผลหารใช้ค่าสองค่า

1. ผลหารไปข้างหน้าที่ใช้ค่าที่เก็บ $\tilde{X}(t+h)$ และ $\tilde{X}(t)$ คือ $\tilde{D}_+X(t) = \dfrac{X(t+h)+\varepsilon_1 - X(t) - \varepsilon_0}{h}$ ซึ่งแยกได้เป็น $D_+X(t) + \dfrac{\varepsilon_1 - \varepsilon_0}{h}$
2. ใช้อสมการสามเหลี่ยม $|\varepsilon_1 - \varepsilon_0| \le |\varepsilon_1| + |\varepsilon_0| \le 2e$ จึงได้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน $2e/h$
3. ผลหารกลางมีตัวส่วน $2h$ และมีค่าที่เก็บสองค่าเช่นกัน ดังนั้น $|\tilde{D}_cX - D_cX| = \dfrac{|\varepsilon_+ - \varepsilon_-|}{2h} \le \dfrac{2e}{2h} = \dfrac{e}{h}$

สรุป: ค่าคลาดเคลื่อนจากการเก็บตัวเลขถูกขยายด้วยตัวหาร h ผลหารไปข้างหน้าขยายเป็นสองเท่าเพราะมีค่าที่เก็บสองค่า แต่ผลหารกลางมีตัวส่วนเป็นสองเท่าด้วย จึงเสมอกัน □

ทฤษฎีบท 5.1.3: ขอบเขตรวมของค่าคลาดเคลื่อนสัมบูรณ์ของผลหาร

สมมติว่า $|X''| \le M_2$ บน $[t, t+h]$ และ $|X'''| \le M_3$ บน $[t-h, t+h]$ และค่าที่เก็บไว้มีค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน e เมื่อ $\tilde{D}$ คือผลหารจากค่าที่เก็บ จะได้

$$\left|\tilde{D}_+X(t) - X'(t)\right| \le \frac{M_2}{2}\,h + \frac{2e}{h}, \qquad \left|\tilde{D}_cX(t) - X'(t)\right| \le \frac{M_3}{6}\,h^2 + \frac{e}{h}.$$

พิสูจน์. **แนวคิด.** แยกค่าคลาดเคลื่อนเป็นสองส่วน แล้วใช้อสมการสามเหลี่ยม

1. เขียน $\tilde{D}_+X(t) - X'(t) = \big(D_+X(t) - X'(t)\big) + \big(\tilde{D}_+X(t) - D_+X(t)\big)$
2. ส่วนแรกมีขนาดไม่เกิน $\frac{M_2}{2}h$ ตามทฤษฎีบท 5.1.1 และส่วนหลังไม่เกิน $\frac{2e}{h}$ ตามทฤษฎีบท 5.1.2
3. ใช้อสมการสามเหลี่ยมได้ข้อความแรก ข้อความที่สองได้ด้วยวิธีเดียวกันโดยใช้ $\frac{M_3}{6}h^2$ และ $\frac{e}{h}$

สรุป: ค่าคลาดเคลื่อนรวมเป็นผลบวกของค่าคลาดเคลื่อนตัดปลายกับค่าคลาดเคลื่อนจากการเก็บตัวเลข ซึ่งดึงกันคนละทิศเมื่อ h เปลี่ยน □

อ่านแบบเข้าใจง่าย

- ส่วนแรก $\frac{M_2}{2}h$ เพิ่มขึ้นเมื่อ h ใหญ่ขึ้น ส่วนที่สอง $\frac{2e}{h}$ เพิ่มขึ้นเมื่อ h เล็กลง ทั้งสองดึงกันจึงมีค่า h ที่ดีที่สุด
- ขอบเขตนี้เป็นกรณีเลวร้ายที่สุด ค่าคลาดเคลื่อนสัมบูรณ์จริงบางครั้งน้อยกว่ามาก แต่ไม่เคยเกินขอบเขต

ขนาดขั้นที่ดีที่สุด

ถ้าเราเลือก h ให้ขอบเขตรวมเล็กที่สุด จะได้ค่าที่ดีที่สุดแบบเดียวกันทุกครั้ง ตัวอย่างเช่นสำหรับผลหารไปข้างหน้า

ฉบับร่าง · Preprint · 2026-10-29

ทฤษฎีบท 5.1.4: ขนาดขั้นที่ดีที่สุด

1. ให้ $M_2, e > 0$ ฟังก์ชัน $g(h) = \frac{M_2}{2}h + \frac{2e}{h}$ บน $(0,\infty)$ มีค่าต่ำสุดเพียงจุดเดียวที่ $h^* = 2\sqrt{e/M_2}$ และค่าต่ำสุดคือ $2\sqrt{M_2 e}$
2. ให้ $M_3, e > 0$ ฟังก์ชัน $k(h) = \frac{M_3}{6}h^2 + \frac{e}{h}$ บน $(0,\infty)$ มีค่าต่ำสุดเพียงจุดเดียวที่ $h^* = (3e/M_3)^{1/3}$ และค่าต่ำสุดคือ $\frac{3^{2/3}}{2}(M_3 e^2)^{1/3}$

พิสูจน์. **แนวคิด.** ลองหาจุดที่อนุพันธ์ของ g และ k เป็นศูนย์ แล้วดูเครื่องหมายก่อนและหลังจุดนั้น

1. (ข้อ 1) $g'(h) = \frac{M_2}{2} - \frac{2e}{h^2}$ ซึ่งเป็นลบเมื่อ $h < h^*$ และเป็นบวกเมื่อ $h > h^*$ โดยที่ h^* เป็นคำตอบเดียวของ $g'(h) = 0$ ในช่วง $(0,\infty)$ เพราะ $h^2 = 4e/M_2$
2. ที่ h^* ได้ $\frac{M_2}{2}h^* = \sqrt{M_2 e}$ และ $\frac{2e}{h^*} = \sqrt{M_2 e}$ รวมกันได้ $g(h^*) = 2\sqrt{M_2 e}$
3. (ข้อ 2) $k'(h) = \frac{M_3}{3}h - \frac{e}{h^2}$ ซึ่งเป็นลบก่อนจุดที่ $h^3 = 3e/M_3$ และเป็นบวกหลังจุดนั้น ดังนั้น $h^* = (3e/M_3)^{1/3}$
4. ที่ h^* จาก $M_3 h^{*3} = 3e$ จะได้ $\frac{M_3}{6}h^{*2} = \frac{e}{2h^*}$ จึงได้ $k(h^*) = \frac{e}{2h^*} + \frac{e}{h^*} = \frac{3e}{2h^*}$ แทน h^* ได้ $\frac{3^{2/3}}{2}(M_3 e^2)^{1/3}$

สรุป: ค่า h ที่ดีที่สุดขึ้นกับค่าคลาดเคลื่อนจากการเก็บตัวเลข e และขนาดของอนุพันธ์อันดับสูงกว่า ไม่ได้ขึ้นกับจุดที่เราจะหาความเร็ว □

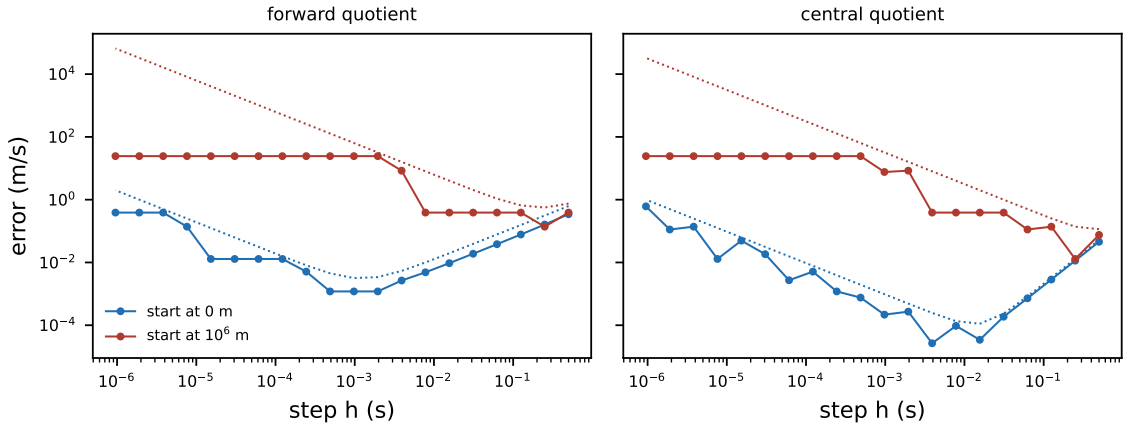
อ่านแบบเข้าใจง่าย

- ผลหารไปข้างหน้ามี h^* ตามสัดส่วน $\sqrt{e}$ ส่วนผลหารกลางมีตามสัดส่วน $e^{1/3}$ ที่ค่า e เท่ากัน ผลหารกลางจึงเลือกขนาดขั้นที่ใหญ่กว่าและได้ขอบเขตเล็กกว่า
- ถ้า e มาจากการปัดของเลขจุดลอยตัวที่ขนาดประมาณ $|X|$ เราได้ $e \approx u|X|$ เมื่อ u คือหน่วยปัด ดังนั้นขนาดขั้นที่ดีที่สุดของผลหารไปข้างหน้าอยู่ในระดับ $\sqrt{u}$

รูปนี้แสดงรูปตัววี (V) ของค่าคลาดเคลื่อน เมื่อ h ใหญ่ค่าคลาดเคลื่อนตัดปลายเด่น เมื่อ h เล็กค่าคลาดเคลื่อนจากการเก็บตัวเลขเด่น ที่จุดกำเนิด ขนาดขั้นที่ดีที่สุดของผลหารไปข้างหน้าอยู่ราว 0.0012 วินาที และขอบเขตที่ดีที่สุดอยู่ราว 0.0031 เมตรต่อวินาที ที่ 10^6 เมตร ขนาดขั้นที่ดีที่สุดคือ $\sqrt{5}/10 \approx 0.2236$ วินาที และขอบเขตที่ดีที่สุดคือ $\sqrt{5}/4 \approx 0.559$ เมตรต่อวินาที อัตราส่วนของขอบเขตทั้งสองคือราว 181 ซึ่งเท่ากับรากที่สองของ 2^{15} เพราะช่องว่างของเลขที่ 10^6 มากกว่าที่จุดกำเนิด 2^{15} เท่า

ดังนั้นเมื่อรถผีอยู่ไกลจากจุดกำเนิด ทั้งขนาดขั้นที่ดีที่สุดและความแม่นที่ดีที่สุดเปลี่ยนไป ขนาดขั้นที่ดีที่สุดเป็นราว 0.22 วินาที ซึ่งราว 13 เท่าของเฟรมละ $1/60$ วินาที เพราะที่เฟรมเล็กนั้นค่าคลาดเคลื่อนจากการเก็บตัวเลขเด่นมาก ผลหารกลางที่ 10^6 เมตรให้ขอบเขตที่ดีที่สุด 0.111 เมตรต่อวินาที ที่ขนาดขั้นราว 0.42 วินาที ส่วนขอบเขตที่ใช้เฟรมติดกัน $1/60$ วินาทีคือ 3.77 เมตรต่อวินาทีสำหรับผลหารไปข้างหน้า และ 1.88 เมตรต่อวินาทีสำหรับผลหารกลาง (ตัวเลขทั้งหมดมาจากโค้ดของเรา)

ฉบับร่าง Preprint · 2026-10-25



รูปที่ 5.3: ค่าคลาดเคลื่อนสัมบูรณ์ที่วัดได้ของผลหารที่ $t = 1$ เมื่อขนาดขั้นลดลงจาก $\frac{1}{2}$ ถึง 2^{-20} ซ้ายคือผลหารไปข้างหน้า ขวาคือผลหารกลาง จุดคือค่าที่วัดได้ เส้นประคือขอบเขตจากทฤษฎีบท 5.1.3 สีฟ้าคือรถผีที่เริ่มที่จุดกำเนิด และสีแดงคือรถผีที่เริ่มที่ 10^6 เมตร (สร้างจาก `numanal/ch05.py`)

ข้อควรระวัง: ขนาดขั้นที่ดีที่สุดคือค่าของความแม่นเท่านั้น

ขนาดขั้นที่ดีที่สุดในทฤษฎีบท 5.1.4 ทำให้*ขอบเขต*เล็กที่สุด ไม่ได้รับประกันว่าค่าที่ได้จะตรงกับความเร็วทุกช่วงเวลา เพราะผลหารที่ใช้ขั้นกว้างจะได้ความเร็วเฉลี่ยในช่วงนั้น ความเร็วของรถผีเปลี่ยนไปในช่วง 0.4 วินาทีด้วย เราจึงต้องรู้ว่าต้องการความเร็วที่จุดหรือความเร็วเฉลี่ยในช่วง

ลองเอง: ขนาดขั้นพอดีกับเฟรมหรือไม่

ใช้ค่า $e = 0.03125$ เมตร และ $M_2 = 2.5$ คำนวณขอบเขตของผลหารไปข้างหน้าเมื่อ $h = 1/60$ ได้ 3.77 เมตรต่อวินาที แล้วลองคำนวณเมื่อ $h = 0.2236$ วินาที ว่าขอบเขตเป็นเท่าไร และอัตราส่วนของ h ทั้งสองเป็นเท่าไร (คำใบ้: ขอบเขตคือ $\frac{M_2}{2}h + \frac{2e}{h}$)

เกร็ดเพิ่มเติม

ในเลขความแม่นคู่ (64 บิต) ที่ค่าประมาณ 40 เมตร ช่องว่างคือ $2^{-47} \approx 7.1 \times 10^{-15}$ ดังนั้น $e \approx 3.55 \times 10^{-15}$ และขนาดขั้นที่ดีที่สุดของผลหารไปข้างหน้าเป็น 7.5×10^{-8} วินาที ซึ่งราว 7 เท่าของ $\sqrt{u} \approx 10^{-8}$ ความแม่นที่ได้ดีกว่ามาก แต่ยังมีค่า h ที่ดีที่สุดเสมอ ไม่ใช่ h ที่เล็กที่สุด

สรุปใจความ

- ความเร็วคือความชันของกราฟตำแหน่ง เราประมาณด้วยผลหารเชิงผลต่างจากค่าที่เก็บไว้ ผลหารไปข้างหน้ามีค่าคลาดเคลื่อนตัดปลายเป็น $\frac{M_2}{2}h$ ส่วนผลหารกลางเป็น $\frac{M_3}{6}h^2$
- ค่าที่เก็บไว้มีค่าคลาดเคลื่อน e และการหารด้วย h ขยายค่าคลาดเคลื่อนนั้นเป็น $2e/h$ หรือ e/h
- ขนาดขั้นที่ดีที่สุดคือ $2\sqrt{e/M_2}$ สำหรับผลหารไปข้างหน้า และ $(3e/M_3)^{1/3}$ สำหรับผลหารกลาง ไม่ใช่ขนาดที่เล็กที่สุดที่ทำได้

ร่างต้น · Preprint 2026-10-23

- รถผีที่อยู่ไกลจากจุดกำเนิดมี e ใหญ่กว่า จึงอ่านความเร็วได้เป็นขั้นและขอบเขตที่ดีที่สุดแย่ลงตามรากที่สองของ e

อ่านต่อ. ตำรา Atkinson 1989, §5.7, pp. 315--320 มีสูตรผลต่างที่ใช้จุดมากกว่าสองจุด และดูผลของค่าคลาดเคลื่อนในค่าที่วัดแบบสุ่ม ส่วน Iserles 2009, §8.1, pp. 139--140 มีการนิยามตัวดำเนินการผลต่างอย่างเป็นระบบ

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 5.1.1 (⋆) ให้ $X(t) = t^4$ และให้ $t = 2$ และ $h = 0.1$

1. จงคำนวณ $D_+X(2)$ และ $D_cX(2)$ และหาค่าคลาดเคลื่อนของแต่ละค่าเทียบกับ $X'(2)$
2. จงตรวจว่าค่าคลาดเคลื่อนไปข้างหน้าอยู่ในช่วง $\left[\frac{h}{2}\min X'',\ \frac{h}{2}\max X''\right]$ บน $[2, 2.1]$ ตามทฤษฎีบท 5.1.1 หรือไม่
3. จงแสดงว่าค่าคลาดเคลื่อนกลางเท่ากับ $4th^2$ พอดี แล้วอธิบายว่าทำไมจึงเท่ากันพอดี

แบบฝึกหัด 5.1.2 (⋆) ให้ $X(t) = \sin t$ ใกล้ $t = 1$ สมมติว่าค่าที่เก็บไว้มีค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน $e = 5 \times 10^{-10}$ และ $|X''| \le M_2 = 1$

1. จงหาขนาดขั้น h^* ที่ทำให้ขอบเขตของผลหารไปข้างหน้าเล็กที่สุด และค่าของขอบเขตนั้น
2. จงคำนวณขอบเขตที่ $h = 10^{-6}$ และที่ $h = 10^{-4}$ ค่าใดให้ขอบเขตเล็กกว่า และเหตุใด

แบบฝึกหัด 5.1.3 (⋆⋆) ให้ค่าที่เก็บไว้ของรถผีที่เฟรม $0, 1, 2, 3$ คือ $x_0 = 10^6$, $x_1 = 10^6 + 0.4375$, $x_2 = 10^6 + 0.8125$ และ $x_3 = 10^6 + 1.25$ เมตร

1. จงแสดงว่าค่าเฉลี่ยของผลหารไปข้างหน้า N ค่าแรก เท่ากับ $(x_N - x_0)/(Nh)$ เสมอ (ตั้งแต่ตรงนี้ไม่มีค่าคลาดเคลื่อนจากการเก็บ)
2. จงหาค่าเฉลี่ยของผลหารไปข้างหน้าสามค่าแรก และตรวจว่าได้ 25 เมตรต่อวินาที
3. ค่าเฉลี่ยนี้ยังมีค่าคลาดเคลื่อนตัดปลายอยู่หรือไม่ จงหาขอบเขตของค่าคลาดเคลื่อนตัดปลายเมื่อ $N = 3$ โดยใช้ $M_2 = 2.5$ และเทียบกับขอบเขตค่าคลาดเคลื่อนจากการเก็บตัวเลขที่ $2e/(Nh)$

แบบฝึกหัด 5.1.4 (⋆⋆) ต้องการให้ขอบเขตของผลหารไปข้างหน้าที่ดีที่สุดที่ 10^6 เมตร (ดังนั้น $2^{19} \le x < 2^{20}$) มีค่าไม่เกิน 0.05 เมตรต่อวินาที เมื่อ $M_2 = 2.5$

1. จงหาค่า e สูงสุดที่ยอมได้
2. ถ้าเลขจุดลอยตัวมี k บิตหลังจุดในส่วนเศษ ช่องว่างที่ 2^{19} คือ 2^{19-k} จงหาค่า k ที่น้อยที่สุดที่พอ และตรวจว่าเลขความแม่นเดี่ยว ($k = 23$) พอหรือไม่

แบบฝึกหัด 5.1.5 (⋆⋆) จงแสดงว่าขอบเขต $2e/h$ ของค่าคลาดเคลื่อนจากการเก็บตัวเลข สามารถเกิดขึ้นได้จริง โดยเลือกค่าที่เก็บไว้ $\tilde{X}(t) = X(t) - e$ และ $\tilde{X}(t+h) = X(t+h) + e$ แล้วหาค่าคลาดเคลื่อนของผลหารไปข้างหน้า จากนั้นอธิบายว่าเหตุใดความผิดแบบนี้จึงน่ากังวลสำหรับรถผีที่เฟรมติดกันที่ 10^6 เมตร

แบบฝึกหัด 5.1.6 (⋆⋆⋆) ที่ $t = 1$ เลือก $h = 10^{-8}$ ให้ $t_1 = 1 + h$ ในเลขความแม่นคู่ และให้ $h_{\text{จริง}} = t_1 - 1$

1. จงแสดงว่า $|h_{\text{จริง}} - h| \le 2^{-53}$ เพราะการปัดของ $1 + h$

ฉบับร่าง · Preprint · 2026-10-23

2. ถ้าหารด้วย h (ขนาดขั้นที่ตั้งไว้) แทนที่จะหารด้วย $h_{\text{จริง}}$ จงหาค่าคลาดเคลื่อนเพิ่มขึ้นโดยประมาณถ้าความเร็วเท่ากับ 24.4 เมตรต่อวินาที

3. จงอธิบายว่าทำไมการหารด้วยขนาดขั้นจริงจึงแก้ปัญหานี้ และยังมีค่าคลาดเคลื่อนอะไรอยู่

แบบฝึกหัด 5.1.7 ($\star\star\star$) (โครงงานเล็ก) ใช้รถผีที่ 10^6 เมตรของหัวข้อนี้ คือ $e = \frac{1}{32}$ เมตร และ $M_3 = \frac{5}{4}$ เมตรต่อวินาที^{3} ถ้าเปรียบเทียบตำแหน่งที่ m เฟรมก่อนและหลังเฟรมที่ต้องการ จะได้ผลหารกลางช่วงกว้าง $\dfrac{X_{k+m} - X_{k-m}}{2mh}$ โดย $h = 1/60$

1. จงแสดงว่าขอบเขตค่าคลาดเคลื่อนสัมบูรณ์ของผลหารกว้างนี้คือ $\dfrac{M_3}{6}(mh)^2 + \dfrac{e}{mh}$

2. จงหาจำนวนเต็ม m ที่เล็กที่สุดที่ทำให้ขอบเขตไม่เกิน 0.5 เมตรต่อวินาที และคำนวณขอบเขตที่ได้

3. ช่วงเวลาที่ผลหารกว้างนี้ใช้คือกี่วินาที และความเร็วที่ได้เป็นความเร็วที่จุดใด จงอธิบายว่าทำไมจึงยอมรับได้หรือไม่ได้ในการแสดงบนจอ

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 5.1.1.

1. $X'(t) = 4t^3$ จึง $X'(2) = 32$ ผลหารไปข้างหน้า $D_+X(2) = \dfrac{2.1^4 - 2^4}{0.1} = \dfrac{19.4481 - 16}{0.1} = 34.481$ ค่าคลาดเคลื่อนคือ 2.481 ผลหารกลาง $D_cX(2) = \dfrac{2.1^4 - 1.9^4}{0.2} = \dfrac{19.4481 - 13.0321}{0.2} = 32.08$ ค่าคลาดเคลื่อนคือ 0.08

2. $X''(t) = 12t^2$ บน $[2, 2.1]$ มีค่าอยู่ระหว่าง 48 ถึง $12 \cdot 2.1^2 = 52.92$ ทฤษฎีบท 5.1.1 บอกว่าค่าคลาดเคลื่อนคือ $\frac{h}{2}X''(\xi)$ ซึ่งอยู่ระหว่าง $0.05 \cdot 48 = 2.4$ ถึง $0.05 \cdot 52.92 = 2.646$ ค่า 2.481 อยู่ในช่วงนี้

3. $X'''(t) = 24t$ เป็นฟังก์ชันเชิงเส้น ผลต่างสมมาตรให้ $(t+h)^4 - (t-h)^4 = 8t^3h + 8th^3$ เมื่อหารด้วย $2h$ ได้ $4t^3 + 4th^2$ ดังนั้นค่าคลาดเคลื่อนคือ $4th^2 = 8 \cdot 0.01 = 0.08$ พอดี เหตุผลคือค่าเฉลี่ยของ X''' ที่สองปลายตรงกับ $X'''(t)$ เพราะ X''' เป็นเชิงเส้น ทำให้ ξ ในทฤษฎีบทเป็นจุด t เอง ($\frac{h^2}{6} \cdot 24 \cdot 2 = 0.08$)

เฉลย 5.1.2.

1. ขอบเขตคือ $g(h) = \frac{M_2}{2}h + \frac{2e}{h} = \frac{1}{2}h + \frac{10^{-9}}{h}$ จากทฤษฎีบท 5.1.4 ได้ $h^* = 2\sqrt{e/M_2} = 2\sqrt{5 \times 10^{-10}} \approx 4.4721 \times 10^{-5}$ และค่าต่ำสุด $2\sqrt{M_2 e} = 2\sqrt{5 \times 10^{-10}} \approx 4.4721 \times 10^{-5}$ เท่ากันพอดีเพราะ $M_2 = 1$

2. ที่ $h = 10^{-6}$: $g = 5 \times 10^{-7} + 10^{-3} \approx 1.0005 \times 10^{-3}$ ที่ $h = 10^{-4}$: $g = 5 \times 10^{-5} + 10^{-5} = 6 \times 10^{-5}$ ค่าที่ 10^{-4} ให้ขอบเขตเล็กกว่า เพราะที่ 10^{-6} ค่าคลาดเคลื่อนจากการเก็บตัวเลขเป็น $2e/h = 10^{-3}$ ซึ่งใหญ่กว่าค่าคลาดเคลื่อนตัดปลายมาก

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 5.1.3.

1. ผลหารไปข้างหน้า N ค่าแรกคือ $\frac{x_{k+1}-x_k}{h}$ สำหรับ $k=0,\ldots,N-1$ ผลรวมของค่าเหล่านี้คือ $\frac{1}{h}\sum_{k=0}^{N-1}(x_{k+1}-x_k)$ ซึ่งพจน์กลางหักล้างกันทีละคู่ เหลือ $\frac{x_N-x_0}{h}$ เมื่อหารด้วย N จะได้ $\frac{x_N-x_0}{Nh}$

2. ค่าที่เก็บไว้คือ $x_0=10^6$, $x_1=10^6+0.4375$, $x_2=10^6+0.8125$, $x_3=10^6+1.25$ ผลหารไปข้างหน้าคือ $60\times 0.4375=26.25$, $60\times 0.375=22.5$, $60\times 0.4375=26.25$ ค่าเฉลี่ยคือ $\frac{75}{3}=25.0$ เมตรต่อวินาที ตรงกับความเร็วจริงที่เฟรมเริ่มต้น

3. ผลหารเฉลี่ยคือ *ความชันของเส้นตัด* ที่ผ่านจุดเริ่มและจุดสุดท้ายของช่วง Nh ดังนั้นค่าคลาดเคลื่อนตัดปลายยังมีอยู่ ตามทฤษฎีบท 5.1.1 เป็น $\frac{Nh}{2}X''(\xi)$ จึงมีขอบเขต $\frac{M_2Nh}{2}=\frac{2.5\cdot 0.05}{2}=0.0625$ เมตรต่อวินาที ส่วนค่าคลาดเคลื่อนจากการเก็บตัวเลขมีขอบเขต $\frac{2e}{Nh}=\frac{0.0625}{0.05}=1.25$ เมตรต่อวินาที การหาค่าเฉลี่ยช่วยลดส่วนหลังจาก 3.75 เหลือ 1.25 แต่เพิ่มส่วนแรกจาก 0.0208 เป็น 0.0625

เฉลย 5.1.4.

1. ต้องการ $2\sqrt{M_2e}\le 0.05$ จึง $M_2e\le 0.000625$ และเมื่อ $M_2=2.5$ ได้ $e\le 0.00025$ เมตร

2. ที่ $2^{19}\le x<2^{20}$ ช่องว่างคือ 2^{19-k} ครึ่งหนึ่งคือ $e=2^{18-k}$ ต้องการ $2^{18-k}\le 0.00025$ ค่า $2^{-12}\approx 0.000244$ ใช้ได้ แต่ $2^{-11}\approx 0.000488$ ใช้ไม่ได้ จึงต้องมี $18-k\le -12$ นั่นคือ $k\ge 30$ ดังนั้น $k=30$ น้อยที่สุด เลขความแม่นเดี่ยวมี $k=23$ จึงได้ $e=2^{-5}=0.03125$ และขอบเขตคือ $2\sqrt{2.5\cdot 0.03125}\approx 0.559$ เมตรต่อวินาที ไม่พอ เลขความแม่นคู่ ($k=52$) พอ

เฉลย 5.1.5. ให้ $\tilde{X}(t)=X(t)-e$ และ $\tilde{X}(t+h)=X(t+h)+e$ ซึ่งมีค่าคลาดเคลื่อนเท่ากับ e ทั้งสองค่า จึงอยู่ในขอบเขตของการเก็บ ผลหารไปข้างหน้าที่ได้คือ

$$\frac{\tilde{X}(t+h)-\tilde{X}(t)}{h}=\frac{X(t+h)-X(t)+2e}{h}=D_+X(t)+\frac{2e}{h}.$$

ค่าคลาดเคลื่อนจากการเก็บตัวเลขเท่ากับ $2e/h$ พอดี ขอบเขตนี้จึงเกิดขึ้นได้จริง สำหรับรถผีที่ 10^6 เมตรกับเฟรมติดกัน $e=0.03125$ และ $h=1/60$ ได้ $2e/h=3.75$ เมตรต่อวินาที ซึ่งใหญ่มากเมื่อเทียบกับความเร็ว 15 ถึง 25 เมตรต่อวินาที (ในการจำลองของเรา ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดที่วัดได้คือ 3.67 ซึ่งอยู่ใต้ขอบเขต 3.77 รวมค่าคลาดเคลื่อนตัดปลาย)

เฉลย 5.1.6.

1. ที่ $1+h$ ค่าที่อยู่ระหว่าง 1 กับ 2 มีช่องว่าง 2^{-52} การปัดไปยังจำนวนที่ใกล้ที่สุดทำให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกินครึ่งหนึ่งของช่องว่าง คือ 2^{-53} จึง $|h_{\text{จริง}}-h|\le 2^{-53}\approx 1.1\times 10^{-16}$ (ค่า $h=10^{-8}$ ไม่ใช่ผลคูณของ 2^{-52})

2. ผลหารที่คำนวณด้วย h คือ $\dfrac{X(t_1)-X(t_0)}{h}\approx X'(1)\cdot\dfrac{h_{\text{จริง}}}{h}=X'(1)\Big(1+\dfrac{h_{\text{จริง}}-h}{h}\Big)$ ค่าคลาดเคลื่อนเพิ่มประมาณ $X'(1)\cdot\dfrac{2^{-53}}{10^{-8}}\approx 24.4\times 1.1\times 10^{-8}\approx 2.7\times 10^{-7}$ เมตรต่อวินาที ซึ่งเล็กมาก แต่ความผิดนี้ไม่ได้เกิดจากการเก็บตัวเลข

3. การหารด้วย $h_{\text{จริง}}$ ทำให้ตัวคูณนั้นหายไป เพราะตัวเศษและตัวส่วนสอดคล้องกัน ยังมีค่าคลาดเคลื่อนจากการเก็บตัวเลขของทั้งสองค่า และค่าคลาดเคลื่อนตัดปลายที่ใช้ $h_{\text{จริง}}$ ซึ่งเล็กมากจนไม่ต่างจาก h ในทางปฏิบัติ

ฉบับร่าง. Preprint. 2026-10-23

เฉลย 5.1.7.

1. ผลหารกลางที่มีจุดห่าง $H = mh$ จากจุดกลาง คือ $\frac{X(t+H)-X(t-H)}{2H}$ ซึ่งเป็นผลหารกลางตามนิยามที่ขั้น H ทฤษฎีบท 5.1.1 และ 5.1.2 ให้ค่าคลาดเคลื่อนตัดปลาย $\frac{M_3}{6}H^2$ และค่าคลาดเคลื่อนจากการเก็บ $\frac{2e}{2H} = \frac{e}{H}$ จึงได้ขอบเขต $\frac{M_3}{6}(mh)^2 + \frac{e}{mh}$

2. คำนวณทีละค่า ($M_3 = 1.25$, $h = 1/60$, $e = 1/32$):

   - $m = 1$: $H = 1/60$, ขอบเขต ≈ 1.875
   - $m = 2$: $H = 1/30$, ขอบเขต ≈ 0.9377
   - $m = 3$: $H = 1/20$, ขอบเขต ≈ 0.6255 ยังเกิน 0.5
   - $m = 4$: $H = 1/15$, ขอบเขต ≈ 0.46968 ไม่เกิน 0.5

   จึงได้ $m = 4$ ขอบเขตประมาณ 0.4697 เมตรต่อวินาที

3. ผลหารใช้ตำแหน่งที่เฟรม $k - 4$ ถึง $k + 4$ ห่างกัน $8/60 \approx 0.133$ วินาที ผลลัพธ์เป็นความเร็วที่เวลาเฟรม k เอง แต่เป็นค่าที่เฉลี่ยทั้งช่วง ยอมรับได้เมื่อความเร่งเล็ก เพราะความเร่งของรถผีไม่เกิน 2.5 เมตรต่อวินาทีกำลังสอง และการแสดงบนจอช้ากว่าเฟรมไม่กี่เฟรม แต่ถ้าความเร็วเปลี่ยนเร็วเกินไป ค่านี้จะตามไม่ทันและควรเลือกช่วงที่สั้นลง

5.2 ระยะทางจากความเร็ว

สถานการณ์สมมติ: ระยะที่ต้องแสดงบนจอ

ในโหมดรถผี จอแสดงระยะที่รถวิ่งไปแล้วด้วย ระบบรู้ความเร็วของรถผีได้ทุกเวลา แต่ถ้าเก็บความเร็วทุกเฟรมตลอดการแข่งยาว ๆ ไฟล์การเล่นซ้ำจะใหญ่เกินไป สมมติว่าเราจึงเก็บความเร็วเพียงบางเวลา แล้วให้ระบบประมาณระยะจากค่าเหล่านั้น

ระยะที่รถวิ่งไปคือพื้นที่ใต้กราฟความเร็ว ถ้ามีความเร็วทุกเวลา เราก็หาพื้นที่นั้นได้ แต่ถ้ามีแค่บางจุด จะหาพื้นที่จากจุดเหล่านั้นอย่างไร และต้องเก็บถี่แค่ไหนจึงจะคลาดจากระยะจริงไม่เกินหนึ่งเซนติเมตร (ตัวเลขทั้งหมดมาจากโค้ดของเรา)

ภาพรวมก่อนอ่าน

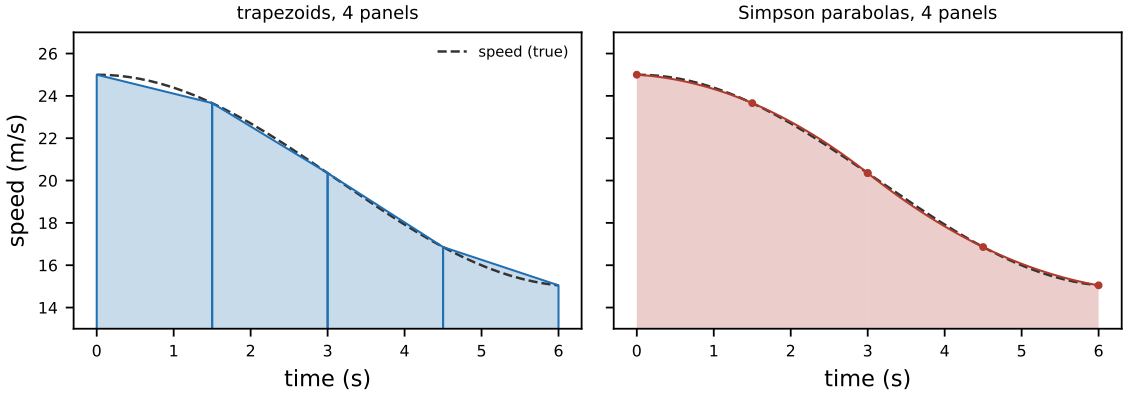
ลองนึกว่าเรารู้ค่าของกราฟที่จุดไม่กี่จุด แต่ไม่มีสูตรของกราฟ วิธีที่ทำได้คือ

- แทนกราฟระหว่างจุดด้วยรูปที่หาพื้นที่ได้ง่าย เช่นเส้นตรงซึ่งได้รูปสี่เหลี่ยมคางหมู หรือพาราโบลาซึ่งได้กฎของซิมป์สัน
- รวมพื้นที่ของรูปเหล่านั้น ผลที่ได้คือผลบวกถ่วงน้ำหนักของค่าที่จุด
- ถามต่อว่า รูปนั้นตรงกับกราฟจริงแค่ไหน และค่าคลาดเคลื่อนจะลดลงเร็วแค่ไหนเมื่อจุดถี่ขึ้น

จากจุดที่เก็บไว้สู่ผลรวม

การประมาณพื้นที่จากค่าที่จุดไม่กี่จุดเรียกว่า **ควอดราเจอร์** (quadrature) คือการเขียนปริพันธ์เป็นผลรวมของค่าฟังก์ชันที่จุดที่เลือก ให้ $I(f) = \int_a^b f(t)\,dt$ เป็นค่าจริง

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 5.4: ซ้าย: กฎเชิงสี่เหลี่ยมคางหมูสี่ช่วงใต้ความเร็วของรถผี ขวา: กฎของซิมป์สันสี่ช่วง คือสองพาราโบลา จุดคือค่าที่เก็บไว้ ทั้งสองกฎใช้ค่าที่จุดห้าจุดเท่ากัน (สร้างจาก `numanal/ch05_quad.py`)

นิยาม 5.2.1: สูตรควอดราเจอร์

สูตรควอดราเจอร์ที่ใช้จุด $x_0, \ldots, x_n$ ในช่วง $[a, b]$ คือ

$$Q(f) = \sum_{j=0}^{n} w_j \, f(x_j).$$

จุด x_j เรียกว่า **โหนด** (node) ของสูตร และตัวเลข w_j เรียกว่า **น้ำหนักของสูตร** (quadrature weight) ค่าคลาดเคลื่อนของสูตรคือ $E(f) = I(f) - Q(f)$

อ่านแบบเข้าใจง่าย

- โหนดคือจุดที่เราเลือกและเก็บค่าไว้ น้ำหนักบอกว่าค่าที่จุดนั้นมีส่วนในผลรวมมากแค่ไหน
- สูตรนี้ใช้แค่ค่า $f(x_j)$ ไม่ต้องรู้สูตรของ f ซึ่งตรงกับสถานการณ์ของเรา คือรู้ความเร็วเฉพาะบางเวลา
- สูตรที่ดีคือสูตรที่ให้ $E(f)$ เล็ก สำหรับฟังก์ชันที่เรียบพอ ความหมายของ "เรียบพอ" จะชัดขึ้นเมื่อถึงหัวข้อระดับความแม่นยำข้างล่าง

ฉบับร่าง · Preprint · 2026-10-23

กฎเชิงสี่เหลี่ยมคางหมู

ความคิดแรกคือแทนกราฟระหว่างจุดสองจุดด้วยเส้นตรง พื้นที่ใต้เส้นตรงระหว่าง a กับ b เป็นรูปสี่เหลี่ยมคางหมู มีความสูงสองด้านเป็น $f(a)$ และ $f(b)$ และกว้าง $b - a$ พื้นที่ของมันคือ $\frac{b-a}{2}\big(f(a) + f(b)\big)$ นี่คือ **กฎเชิงสี่เหลี่ยมคางหมู** (trapezoidal rule) หนึ่งช่วง

ถ้าช่วงกว้างเกินไป เราแบ่งเป็น n ช่วงย่อยที่กว้างเท่ากัน ใช้กฎหนึ่งช่วงกับแต่ละช่วง แล้วรวมผล วิธีนี้เรียกว่า **กฎแบบประกอบ** (composite rule) ซึ่งก็คือการนำกฎของช่วงเดียวไปใช้ซ้ำกับแต่ละช่วงย่อย

นิยาม 5.2.2: กฎเชิงสี่เหลี่ยมคางหมูแบบประกอบ

ให้ $n \geq 1$ $h = \dfrac{b-a}{n}$ และ $x_j = a + jh$ กฎเชิงสี่เหลี่ยมคางหมูแบบประกอบ n ช่วงคือ

$$T_n(f) = h\Big[\tfrac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \tfrac{1}{2}f(x_n)\Big].$$

เมื่อ $n = 1$ จะได้ $T_1(f) = \dfrac{b-a}{2}\big(f(a) + f(b)\big)$

อ่านแบบเข้าใจง่าย

- จุดภายในแต่ละจุดเป็นปลายของช่วงสองช่วงที่อยู่ติดกัน จึงมีน้ำหนัก 1 ส่วนจุดปลายทั้งสองมีน้ำหนัก $\frac{1}{2}$ เพราะนับพื้นที่ของช่วงนั้นเพียงครึ่งเดียว
- เมื่อ n เพิ่มขึ้น h เล็กลง และเส้นตรงในแต่ละช่วงจะใกล้เส้นโค้งมากขึ้น

ลองเอง: พื้นที่ใต้ x^2 บน $[0, 2]$

ค่าจริงคือ $\int_0^2 x^2\,dx = \frac{8}{3}$ ลองคำนวณเอง

- หนึ่งช่วง ($n = 1$, $h = 2$): $T_1 = \dfrac{2}{2}(0 + 4) = 4$ ค่าคลาดเคลื่อนคือ $\frac{8}{3} - 4 = -\frac{4}{3}$
- สองช่วง ($n = 2$, $h = 1$): $T_2 = 1 \cdot \big(0 + 1 + \frac{1}{2}\cdot 4\big) = 3$ ค่าคลาดเคลื่อนคือ $\frac{8}{3} - 3 = -\frac{1}{3}$

เมื่อ n เป็นสองเท่า ค่าคลาดเคลื่อนลดลงสี่เท่า ทำไมจึงเป็นเช่นนั้น (คำใบ้: $f''(x) = 2$ และทฤษฎีบทข้างล่างมีตัว h^2)

ความผิดของกฎนี้มาจากการแทนความโค้งด้วยเส้นตรง ความโค้งคือ f'' เราจึงคาดว่าค่าคลาดเคลื่อนขึ้นกับ f'' และขนาดของช่วง

ทฤษฎีบท 5.2.1: ค่าคลาดเคลื่อนของกฎเชิงสี่เหลี่ยมคางหมู

ให้ $f \in C^2[a, b]$ จะได้ว่า

1. (หนึ่งช่วง) มี $\xi \in [a, b]$ ที่ทำให้ $I(f) - T_1(f) = -\dfrac{(b-a)^3}{12} f''(\xi)$
2. (แบบประกอบ n ช่วง) มี $\xi \in [a, b]$ ที่ทำให้ $I(f) - T_n(f) = -\dfrac{(b-a)h^2}{12} f''(\xi)$ เมื่อ $h = \frac{b-a}{n}$

พิสูจน์. **แนวคิด.** เขียนฟังก์ชันเป็นเส้นตรงบวกเศษ แล้วพิสูจน์ว่าเศษทำให้เกิดค่าคลาดเคลื่อนเป็นปริพันธ์ของ f'' คูณด้วยฟังก์ชันที่มีเครื่องหมายคงที่

1. ทฤษฎีบทของเทย์เลอร์ที่จุด a ให้ $f(x) = f(a) + f'(a)(x-a) + \int_a^b (x-t)_+ f''(t)\,dt$ สำหรับ $x \in [a, b]$ โดย $u_+ = \max(u, 0)$ เพราะ $\int_a^x (x-t) f''(t)\,dt = \int_a^b (x-t)_+ f''(t)\,dt$
2. กฎ I และ T_1 ตรงกับเส้นตรงทุกเส้น (ตรวจโดยตรง) ดังนั้น $E = I - T_1$ เป็นศูนย์กับพหุนามดีกรีไม่เกินหนึ่ง เมื่อนำข้อ 1 มาแทน จะได้ $E(f) = \int_a^b K(t) f''(t)\,dt$ โดย $K(t) = E_x\big[(x-t)_+\big]$ หมายถึงนำ E ไปใช้กับตัวแปร x ขณะที่ t คงที่ (การสลับลำดับปริพันธ์ทำได้เพราะฟังก์ชันต่อเนื่องบนช่วงปิด)

ฉบับร่าง · Preprint · 2026-10-23

3. สำหรับ $t \in [a,b]$ ได้ $\int_a^b (x-t)_+\,dx = \dfrac{(b-t)^2}{2}$ และ T_1 ของ $(x-t)_+$ คือ $\dfrac{b-a}{2}\big[(a-t)_+ + (b-t)_+\big] = \dfrac{b-a}{2}(b-t)$ เพราะ $(a-t)_+ = 0$ จึงได้

$$K(t) = \frac{(b-t)^2}{2} - \frac{(b-a)(b-t)}{2} = -\frac{(t-a)(b-t)}{2}.$$

4. บน $[a,b]$ ทั้ง $t-a$ และ $b-t$ ไม่เป็นลบ ดังนั้น $K \le 0$ และไม่เปลี่ยนเครื่องหมาย เพราะ f'' ต่อเนื่อง ทฤษฎีบทค่ามัชฌิมของปริพันธ์ให้มี $\xi \in [a,b]$ ที่ $\int_a^b K f'' = f''(\xi)\int_a^b K$

5. เพราะ $\int_a^b (t-a)(b-t)\,dt = \frac{(b-a)^3}{6}$ จึงได้ $\int_a^b K = -\frac{(b-a)^3}{12}$ ซึ่งคือข้อ 1

6. ข้อ 2 ใช้ข้อ 1 กับแต่ละช่วง $[x_{j-1}, x_j]$ ได้ค่าคลาดเคลื่อนรวม $-\frac{h^3}{12}\sum_{j=1}^{n} f''(\xi_j) = -\frac{nh^3}{12}\bar{f}$ โดย $\bar{f}$ คือค่าเฉลี่ยของ $f''(\xi_j)$ ซึ่งอยู่ระหว่างค่าต่ำสุดกับค่าสูงสุดของ f'' บน $[a,b]$ ทฤษฎีบทค่ากลางให้ ξ ที่ $f''(\xi) = \bar{f}$ และ $nh^3 = (b-a)h^2$

สรุป: ค่าคลาดเคลื่อนของกฎเชิงสี่เหลี่ยมคางหมูเป็นสัดส่วนกับ h^2 และขึ้นกับความโค้ง f'' กฎนี้แม่นกับเส้นตรงทุกเส้น แต่ไม่แม่นกับ x^2 □

กฎของซิมป์สัน

ถ้าใช้พาราโบลาแทนเส้นตรง ต้องใช้จุดสามจุดคือปลายทั้งสองและจุดกลาง $c = \frac{a+b}{2}$ พื้นที่ใต้พาราโบลาที่ผ่านจุดทั้งสามเป็น $\frac{b-a}{6}\big(f(a) + 4f(c) + f(b)\big)$ สูตรนี้คือ **กฎของซิมป์สัน** (Simpson's rule) หนึ่งช่วง

นิยาม 5.2.3: กฎของซิมป์สันแบบประกอบ

ให้ n เป็นจำนวนคู่ $h = \dfrac{b-a}{n}$ และ $x_j = a + jh$ กฎของซิมป์สันแบบประกอบคือ

$$S_n(f) = \frac{h}{3}\Big[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)\Big].$$

เมื่อ $n = 2$ จะได้ $S_2(f) = \dfrac{b-a}{6}\big(f(a) + 4f(c) + f(b)\big)$

อ่านแบบเข้าใจง่าย

- ทุกคู่ของช่วงย่อยทำพาราโบลาหนึ่งเส้น จุดกลางของแต่ละคู่มีน้ำหนัก 4 จุดที่ใช้ร่วมกันระหว่างสองคู่มีน้ำหนัก 2 และจุดปลายมีน้ำหนัก 1
- จำนวนช่วงต้องเป็นคู่ เพราะพาราโบลาหนึ่งเส้นใช้สองช่วง

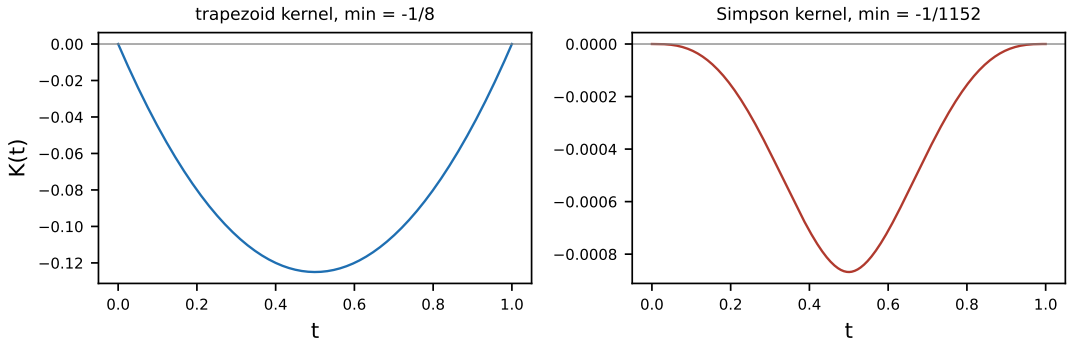
ลองเอง: ซิมป์สันกับ x^3 และ x^4

บน $[0,2]$ ใช้ $n = 2$ ($h = 1$) จะได้ $S_2(f) = \frac{1}{3}\big(f(0) + 4f(1) + f(2)\big)$

- $f = x^3$: $S_2 = \frac{1}{3}(0 + 4 + 8) = 4$ ตรงกับ $\int_0^2 x^3\,dx = 4$ พอดี
- $f = x^4$: $S_2 = \frac{1}{3}(0 + 4 + 16) = \frac{20}{3}$ แต่ $\int_0^2 x^4\,dx = \frac{32}{5}$ ค่าคลาดเคลื่อนคือ $-\frac{4}{15}$

กฎที่สร้างจากพาราโบลาตรงกับ x^3 ได้ ทั้งที่พาราโบลาตรงกับ x^3 ไม่ได้ทั่วไป ลองคิดว่าเพราะอะไรก่อนอ่านหัวข้อถัดไป

ร่างหนังสือ · Preprint · 2026-10-23



รูปที่ 5.5: เคอร์เนลเปอาโนของกฎหนึ่งช่วงบน $[0,1]$ ซ้ายคือกฎเชิงสี่เหลี่ยมคางหมู ขวาคือกฎของซิมป์สัน แต่ละกราฟมีแกนตั้งของตัวเอง (สร้างจาก `numanal/ch05_quad.py`)

ระดับความแม่นยำ: ทำไมซิมป์สันได้ดีกรีสาม

เราต้องการวัดว่าสูตรแม่นกับพหุนามดีกรีสูงสุดเท่าไร

นิยาม 5.2.4: ระดับความแม่นยำ

สูตร Q มี **ระดับความแม่นยำ** (degree of precision) เท่ากับ m ถ้า $E(p)=0$ สำหรับทุกพหุนาม p ที่มีดีกรีไม่เกิน m และมีพหุนามดีกรี $m+1$ อย่างน้อยหนึ่งตัวที่ $E(p)\neq 0$

อ่านแบบเข้าใจง่าย

- สูตรที่มีระดับ m ตรงกับ $1, x, \ldots, x^m$ ทุกตัว และผิดกับพหุนามดีกรี $m+1$ บางตัว
- กฎเชิงสี่เหลี่ยมคางหมูมีระดับ 1 กฎของซิมป์สันมีระดับ 3
- เหตุผลคือความสมมาตร ลองดู $f(x)=(x-c)^3$ บน $[a,b]$ ค่าที่จุด a, c, b คือ $-h^3, 0, h^3$ เมื่อ $h=\frac{b-a}{2}$ ซึ่งหักล้างกันพอดี จึงได้ $S_2=0$ และปริพันธ์ของ f ก็เป็นศูนย์ด้วย

ค่าคลาดเคลื่อนของกฎของซิมป์สันจึงเล็กกว่าที่เราอาจคาดไว้ ดังทฤษฎีบทข้างล่าง

ก่อนพิสูจน์ ลองดูว่าค่าคลาดเคลื่อนของกฎหนึ่งช่วงกระจายตัวอย่างไรตามตำแหน่งของความโค้ง ฟังก์ชัน $K(t)$ ที่ปรากฏในการพิสูจน์เรียกว่า **เคอร์เนลเปอาโน** (Peano kernel) ค่า $K(t)$ บอกว่าความโค้งใกล้จุด t มีน้ำหนักเท่าไรในค่าคลาดเคลื่อนรวม ภาพต่อไปนี้แสดงเคอร์เนลของกฎทั้งสองบนช่วง $[0,1]$ ซึ่งไม่เปลี่ยนเครื่องหมายในทั้งสองกรณี

ทฤษฎีบท 5.2.2: ค่าคลาดเคลื่อนของกฎของซิมป์สัน

ให้ $f\in C^4[a,b]$ จะได้ว่า

1. (หนึ่งช่วง) มี $\xi\in[a,b]$ ที่ทำให้ $I(f)-S_2(f)=-\dfrac{(b-a)^5}{2880}f^{(4)}(\xi)$
2. (แบบประกอบ n ช่วง, n เป็นจำนวนคู่) มี $\xi\in[a,b]$ ที่ทำให้ $I(f)-S_n(f)=-\dfrac{(b-a)h^4}{180}f^{(4)}(\xi)$ เมื่อ $h=\frac{b-a}{n}$

พิสูจน์. **แนวคิด.** เหมือนกับกฎเชิงสี่เหลี่ยมคางหมู แต่ใช้เทย์เลอร์ถึงพจน์ที่สาม และใช้ความจริงว่ากฎนี้ตรงกับพหุนามดีกรีไม่เกิน

สำนักกด Preprint · 2026-10-23

สาม

1. ในข้อนี้ให้ $h = \frac{b-a}{2}$ เป็นครึ่งความกว้างของช่วงเดียว และ $c = a + h$ เทย์เลอร์ที่จุด a ให้ $f(x) = p_3(x) + \frac{1}{6}\int_a^b (x-t)_+^3 f^{(4)}(t)\,dt$ โดย p_3 เป็นพหุนามดีกรีไม่เกินสาม

2. กฎของซิมป์สันตรงกับ p_3 ทุกตัว (ดูแบบฝึกหัด 5.2.3) จึงได้ $E(f) = \int_a^b K(t) f^{(4)}(t)\,dt$ โดย $K(t) = \frac{1}{6} E_x\big[(x-t)_+^3\big]$

3. ทำนองเดียวกับข้อ 3 ของทฤษฎีบทก่อน $\int_a^b (x-t)_+^3\,dx = \frac{(b-t)^4}{4}$ และซิมป์สันของ $(x-t)_+^3$ คือ $\frac{h}{3}\big[4(c-t)_+^3 + (b-t)^3\big]$ เพราะค่าที่ a เป็นศูนย์ จึงได้ $6K(t) = \frac{(b-t)^4}{4} - \frac{h}{3}\big[4(c-t)_+^3 + (b-t)^3\big]$ เมื่อ $t \in [a,b]$ แยกเป็นสองกรณี ถ้า $t \in [c,b]$ จะได้ $6K(t) = \frac{(b-t)^3}{12}\big(3(b-t) - 4h\big)$ และ $3(b-t) - 4h = b + 2a - 3t$ จึงได้

$$K(t) = \frac{(b-t)^3}{72}(b + 2a - 3t) \qquad (c \le t \le b).$$

ถ้า $t \in [a,c]$ ให้ $s = c - t$ จะได้ $b - t = h + s$ และ $6K(t) = \frac{(h+s)^4}{4} - \frac{h}{3}\big[4s^3 + (h+s)^3\big] = \frac{(s-h)^3(3s+h)}{12}$ (ตรวจโดยกระจายพหุนาม) เมื่อ $s - h = a - t$ และ $3s + h = a + 2b - 3t$ จึงได้

$$K(t) = \frac{(t-a)^3}{72}(3t - a - 2b) \qquad (a \le t \le c).$$

4. ตรวจเครื่องหมาย: บน $[a,c]$ ตัวประกอบ $3t - a - 2b$ เพิ่มขึ้นตาม t และมีค่า $-h$ ที่ $t = c$ จึงไม่เกิน $-h < 0$ บนช่วงนี้ บน $[c,b]$ ตัวประกอบ $b + 2a - 3t$ ลดลงตาม t และมีค่า $-h$ ที่ $t = c$ จึงไม่เกิน $-h < 0$ ดังนั้น $K \le 0$ บน $[a,b]$ ทั้งสองส่วนเข้ากันได้เพราะ K ต่อเนื่องที่ c (ค่าเป็น $-h^4/72$ จากทั้งสองด้าน)

5. เนื่องจาก K ไม่เปลี่ยนเครื่องหมาย ทฤษฎีบทค่ามัชฌิมของปริพันธ์ให้มี $\xi \in [a,b]$ ที่ $E(f) = f^{(4)}(\xi)\int_a^b K$ สำหรับ $\int_a^b K$ ใช้ $f(x) = x^4/24$ ซึ่งมี $f^{(4)} = 1$ จึงได้ $\int_a^b K = \frac{1}{24}E(x^4)$ เพราะค่านี้ไม่เปลี่ยนเมื่อเลื่อนช่วง ใช้ช่วง $[-h,h]$ ที่มี $c = 0$ จะได้ $\int_{-h}^{h} x^4\,dx = \frac{2h^5}{5}$ และ $S_2(x^4) = \frac{h}{3}(h^4 + 0 + h^4) = \frac{2h^5}{3}$ จึงได้ $E(x^4) = -\frac{4h^5}{15}$ และ $\int_a^b K = -\frac{h^5}{90} = -\frac{(b-a)^5}{2880}$ ได้ข้อ 1

6. ข้อ 2 ใช้ข้อ 1 กับ $\frac{n}{2}$ คู่ของช่วงย่อย แต่ละคู่กว้าง $2h$ ได้ค่าคลาดเคลื่อนรวม $-\frac{h^5}{90}\cdot\frac{n}{2}\,\bar{f}$ เมื่อ $\bar{f}$ คือค่าเฉลี่ยของ $f^{(4)}(\xi_j)$ ทฤษฎีบทค่ากลางให้ ξ ที่ $f^{(4)}(\xi) = \bar{f}$ และ $\frac{n}{2}h^5 = \frac{(b-a)h^4}{2}$ จึงได้ $-\frac{(b-a)h^4}{180} f^{(4)}(\xi)$

สรุป: ค่าคลาดเคลื่อนของซิมป์สันเป็นสัดส่วนกับ h^4 ตัวคูณเป็น $\frac{1}{90}$ ของช่วงกว้าง และขึ้นกับ $f^{(4)}$ เท่านั้น $\square$

ลองเอง: ซิมป์สันสี่ช่วงกับ x^4 บน $[0,2]$

ที่นี่ $h = \frac{1}{2}$ และจุดคือ $0, \frac{1}{2}, 1, \frac{3}{2}, 2$ ค่า x^4 ที่จุดเหล่านี้คือ $0, \frac{1}{16}, 1, \frac{81}{16}, 16$ จึงได้

$$S_4 = \frac{1}{6}\Big(0 + 4\cdot\tfrac{1}{16} + 2\cdot 1 + 4\cdot\tfrac{81}{16} + 16\Big) = \frac{1}{6}\cdot\frac{77}{2} = \frac{77}{12}$$

ค่าจริงคือ $\frac{32}{5}$ จึงได้ค่าคลาดเคลื่อน $\frac{32}{5} - \frac{77}{12} = -\frac{1}{60}$ เทียบกับ $n = 2$ ที่ได้ $-\frac{4}{15}$ ค่าคลาดเคลื่อนลดลง 16 เท่า ตรงกับทฤษฎีบท เพราะ $-\frac{(b-a)h^4}{180}\cdot 24 = -\frac{2\cdot\frac{1}{16}}{180}\cdot 24 = -\frac{1}{60}$

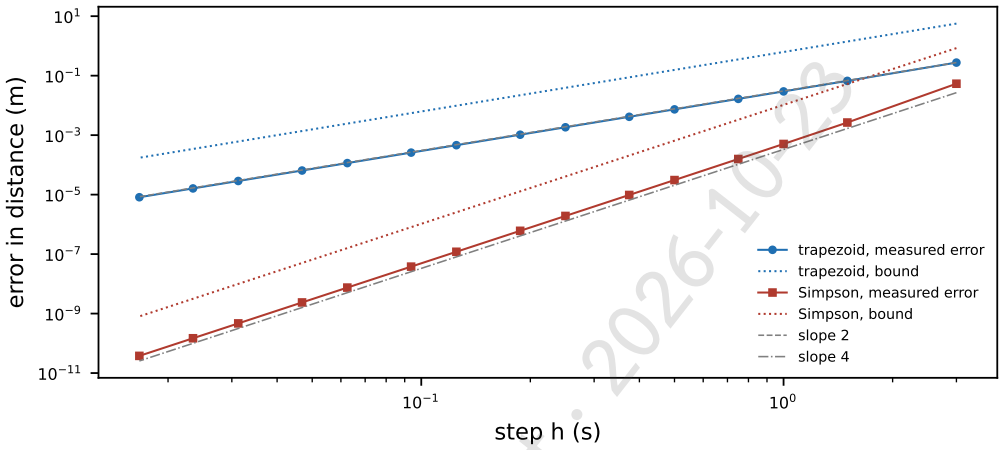
เกร็ดเพิ่มเติม

ค่าคงที่ $\frac{1}{90}$ และ $\frac{3}{80}$ ที่ใช้ในตารางข้างล่าง ตรงกับตาราง 5.8 ของตำรา Atkinson 1989, §5.2, น. 266 เราคำนวณค่าเหล่านี้ใหม่ด้วยเศษส่วนที่แน่นอนในโค้ด

ร่างปริญญา · Preprint 2026-10-23

ตารางที่ 5.1: ระยะของรถผีใน 6 วินาทีแรก จากกฎเชิงสี่เหลี่ยมคางหมูและกฎของซิมป์สันแบบประกอบ

n	h (วินาที)	T_n (เมตร)	E_T (เมตร)	S_n (เมตร)	E_S (เมตร)
2	3	121.13611	2.751×10^{-1}	121.46478	-5.358×10^{-2}
4	1.5	121.34442	6.678×10^{-2}	121.41386	-2.657×10^{-3}
8	0.75	121.39462	1.658×10^{-2}	121.41136	-1.577×10^{-4}
16	0.375	121.40706	4.137×10^{-3}	121.41121	-9.731×10^{-6}
48	0.125	121.41074	4.594×10^{-4}	121.41120	-1.197×10^{-7}
360	$1/60$	121.41119	8.167×10^{-6}	121.41120	-3.782×10^{-11}



รูปที่ 5.6: ค่าคลาดเคลื่อนของระยะรถผีเทียบกับขนาดขั้น h บนสเกลลอการิทึมสองแกน เส้นประคือขอบเขตจากทฤษฎีบท เส้นเทาคือความชัน 2 และ 4 (สร้างจาก `numanal/ch05_quad.py`)

ดูค่าคลาดเคลื่อนจากรถผี

ตารางต่อไปนี้ให้ระยะของรถผีใน 6 วินาทีแรก ค่าจริงคือ $120 + 10\sin 3 \approx 121.41120$ เมตร ระยะ T_n และ S_n คำนวณจากความเร็วที่จุด $0, h, 2h, \ldots, 6$ และค่าคลาดเคลื่อนเป็น $E = I - Q$ ตัวเลขทั้งหมดมาจาก `numanal/ch05_quad.py`

จากตารางนี้ เมื่อ n จาก 8 เป็น 16 ค่าคลาดเคลื่อนของกฎเชิงสี่เหลี่ยมคางหมูลดลงประมาณ 4 เท่า ส่วนค่าคลาดเคลื่อนของซิมป์สันลดลงประมาณ 16 เท่า ตรงกับอันดับ h^2 และ h^4 ที่ทฤษฎีบทข้างบนบอก เราเห็นอันดับเดียวกันบนกราฟสเกลลอการิทึม

ขอบเขตที่รับประกันกับค่าคลาดเคลื่อนจริง

ทฤษฎีบทให้ขอบเขตที่ใช้ได้กับทุกช่วง ในรถผี $b - a = 6$ และ $\max|f''| = 1.25$ กับ $\max|f^{(4)}| = \frac{5}{16}$ ทั้งสองค่ามากที่สุดที่ $t = 0$ จึงได้

$$|E_T| \le \frac{6\,h^2}{12} \cdot 1.25 = \frac{22.5}{n^2}, \qquad |E_S| \le \frac{6\,h^4}{180} \cdot \frac{5}{16} = \frac{13.5}{n^4}, \qquad h = \frac{6}{n}.$$

ถ้าต้องการค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 0.01 เมตร (หนึ่งเซนติเมตร) จะต้องมี $n^2 \ge 2250$ สำหรับกฎเชิงสี่เหลี่ยมคางหมู นั่นคือ $n \ge 48$ และ $n^4 \ge 1350$ สำหรับซิมป์สัน นั่นคือ $n \ge 6.07$ แต่ n ต้องเป็นจำนวนคู่ จึงได้ $n = 8$

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 5.2: สูตรนิวตัน–โคตส์สี่แบบแรก เมื่อ $w_j = h\,c_j$ และ $h = \frac{b-a}{n}$

n	น้ำหนัก c_j	ระดับ	ค่าคลาดเคลื่อน $E(f)$
1	$\frac{1}{2},\ \frac{1}{2}$	1	$-\frac{1}{12}h^3 f''(\xi)$
2	$\frac{1}{3},\ \frac{4}{3},\ \frac{1}{3}$	3	$-\frac{1}{90}h^5 f^{(4)}(\xi)$
3	$\frac{3}{8},\ \frac{9}{8},\ \frac{9}{8},\ \frac{3}{8}$	3	$-\frac{3}{80}h^5 f^{(4)}(\xi)$
4	$\frac{14}{45},\ \frac{64}{45},\ \frac{8}{15},\ \frac{64}{45},\ \frac{14}{45}$	5	$-\frac{8}{945}h^7 f^{(6)}(\xi)$

ข้อควรระวัง: ขอบเขตที่รับประกันไม่ใช่ค่าคลาดเคลื่อนที่เกิดจริง

ขอบเขต ข้าง บน ใช้ เลือก จำนวน ช่วง ได้ อย่าง ปลอดภัย แต่ ค่า จริง มัก น้อย กว่า มาก จาก การ คำนวณ ของ เรา กฎเชิงสี่เหลี่ยมคางหมูใช้เพียง $n = 11$ ช่วงก็พอ (ที่ $n = 10$ ค่าคลาดเคลื่อนสัมบูรณ์คือ 0.0106 เมตร เกินหนึ่งเซนติเมตรเล็กน้อย) และซิมป์สันใช้ $n = 4$ ช่วงก็พอ ขอบเขตกว้างกว่าเพราะ $\max|f''|$ บรรลุที่ $t = 0$ เท่านั้น ไม่ได้บรรลุตลอดช่วง

สูตรนิวตัน–โคตส์: เพิ่มจุดแล้วดีขึ้นเสมอหรือไม่

เราสร้างสูตรทั่วไปได้โดยเลือก $n+1$ จุดที่เว้นระยะเท่ากัน แทนฟังก์ชันด้วยพหุนามที่ผ่านจุดเหล่านั้นในรูปแบบลากรานจ์ (ที่ทำในบทที่ 3) แล้วหาปริพันธ์ของพหุนามนั้น น้ำหนัก w_j คือปริพันธ์ของพหุนามลากรานจ์ ℓ_j ที่เป็น 1 ที่โหนด x_j และเป็น 0 ที่โหนดอื่น สูตรนี้เรียกว่า **วิธีนิวตัน–โคตส์** (Newton–Cotes rule)

ตารางต่อไปนี้ให้สูตรสี่แบบแรก โดยเขียนน้ำหนักเป็น $w_j = h\,c_j$ เมื่อ $h = \frac{b-a}{n}$ และ c_j เป็นเศษส่วนที่แน่นอน

สำหรับ n คู่ ตำรากล่าวว่าสูตรมีระดับ $n+1$ และค่าคลาดเคลื่อนเป็น $C_n h^{n+3} f^{(n+2)}(\xi)$ ซึ่งเป็นอันดับเดียวกับที่เราเห็นในตาราง (ทฤษฎีบทนี้เราไม่พิสูจน์ในบทนี้ ดู Atkinson 1989, §5.2, Thm 5.1, pp. 264–265) ค่า C_n ของตำราคือ $\frac{1}{(n+2)!}\int_0^n p^2(p-1)\cdots(p-n)\,dp$ และเราตรวจว่าตรงกับค่าในตารางสำหรับ $n \le 10$ ด้วยการคำนวณ

ข้อควรระวัง: จุดมากขึ้นไม่ได้แปลว่าดีขึ้นเสมอ

ลองใช้ฟังก์ชัน $f(t) = \frac{1}{1+t^2}$ บน $[-5, 5]$ ซึ่งเป็นฟังก์ชันเดียวกับหัวข้อ 3.2 แต่กว้างขึ้น ค่าจริงคือ $2\arctan 5 \approx 2.746802$ เมื่อใช้วิธีนิวตัน–โคตส์บนช่วงทั้งหมดด้วย n ต่าง ๆ จะได้

n	$Q_n(f)$	$E = I - Q_n$
2	6.794872	-4.048070
4	2.374005	0.372796
6	3.870449	-1.123647
8	1.500489	1.246313
10	4.673301	-1.926499
12	-0.312937	3.059738

ค่าคลาดเคลื่อนสลับเครื่องหมายและโตขึ้น ทั้งที่กฎแบบประกอบ T_{16} ของฟังก์ชันเดียวกันผิดเพียง 0.00069 สาเหตุคือน้ำหนักของสูตรนิวตัน–โคตส์เริ่มเป็นลบที่ $n = 8$ และผลรวมของค่าสัมบูรณ์ของน้ำหนักคือ 11.6 ที่ $n = 8$ และ 30.6 ที่ $n = 10$ เมื่อน้ำหนักทั้งหมดเป็นบวก ผลรวมจะเท่ากับ n เสมอ ตำรากล่าวว่า ถ้าสูตรลู่เข้าสำหรับทุกฟังก์ชันต่อเนื่อง ผลรวมของค่าสัมบูรณ์ของน้ำหนักต้องถูกจำกัด Atkinson 1989, §5.2, Thm 5.2 ซึ่งเป็นเหตุผลว่าทำไมเราจึงใช้กฎแบบประกอบแทนสูตรดีกรีสูง

ฉบับร่าง Preprint · 2026-10-23

ค่าคลาดเคลื่อนเชิงกำกับและสูตรออยเลอร์–แมคลอริน

ทฤษฎีบทข้างบนให้ขอบเขตที่ขึ้นกับ f'' ทั้งช่วง แต่สูตรเชิงกำกับ (asymptotic) บอกเพิ่มว่าค่าคลาดเคลื่อนนั้นขึ้นกับค่าที่ปลายช่วงเท่านั้นในพจน์แรก

$$T_n(f) - I(f) = \frac{h^2}{12}\big(f'(b) - f'(a)\big) - \frac{h^4}{720}\big(f'''(b) - f'''(a)\big) + \cdots$$

นี่คือสูตรออยเลอร์–แมคลอริน **สูตรออยเลอร์-แมคลอริน** (Euler--Maclaurin formula) ตำราให้เฉพาะพจน์แรกเป็นค่าคลาดเคลื่อนเชิงกำกับ Atkinson 1989, §5.1, pp. 254--255 และนำไปสู่การแก้ไขค่าคลาดเคลื่อนด้วย $f'(a)$ และ $f'(b)$ พจน์ที่สองเราตรวจด้วยตัวอย่าง x^4 ในแบบฝึกหัด 5.2.5 ส่วนการใช้สูตรเต็มเพื่อเร่งการลู่เข้าอยู่ในหัวข้อ 5.4 Atkinson 1989, §5.4

ผลที่ตามมาคือ ถ้า f เป็นคาบ 4π และเราหาปริพันธ์ตลอดหนึ่งคาบ ปลายทั้งสองมีค่าอนุพันธ์เท่ากัน จึงทุกพจน์หายไป กฎเชิงสี่เหลี่ยมคางหมูจะแม่นเป๊ะทุกจำนวนช่วง $n \geq 2$ (ตำราชี้ว่าเป็นกรณีที่ลู่เข้าดีมาก Atkinson 1989, §5.1, น. 262--263) ความเร็วของรถผีคือ $20 + 5\cos(t/2)$ ซึ่งมีคาบ 4π ดังนั้นบนช่วง $[0, 4\pi]$ ผลของ T_n คือ 80π เสมอ (ตัวเลขนี้ตรวจด้วยโค้ดของเรา) นี่ไม่ได้บอกว่ากฎจะแม่นกับช่วงที่ไม่ครบคาบ

ข้อควรระวัง: เมื่อความเร็วหักมุม

ทฤษฎีบทข้างบนต้องการให้ f'' (หรือ $f^{(4)}$) ต่อเนื่อง ถ้าความเร็วมีมุมหัก เช่นรถเบรกทันทีที่ $t = 1.3$ วินาที อันดับที่ได้จะไม่เป็น 2 หรือ 4 อีกต่อไป ตัวอย่างของเรา $g(t) = 20 + |t - 1.3|$ บน $[0, 2]$ ให้ผลดังนี้ (ตัวเลขจากการคำนวณของเรา) เมื่อ n เพิ่มเป็นสองเท่า ค่าคลาดเคลื่อนของซิมป์สันลดลงเพียง 4 เท่า ไม่ใช่ 16 เท่า เช่น จาก $n = 16$ ไป 32 ได้อัตราส่วน 4.00 ค่าคลาดเคลื่อนนี้จึงเป็นอันดับ h^2 ส่วนกฎเชิงสี่เหลี่ยมคางหมูมีอัตราส่วนสลับระหว่าง 2.67 กับ 6 ขึ้นกับตำแหน่งของมุมเทียบกับจุดที่เลือก แบบฝึกหัด 5.2.6 อธิบายเหตุผลของทั้งสองกรณี

สรุปใจความ

- ระยะคือพื้นที่ใต้กราฟความเร็ว เราหาค่าประมาณด้วยควอดราเจอร์ คือผลรวมของน้ำหนักคูณกับค่าที่โหนด
- กฎเชิงสี่เหลี่ยมคางหมูตรงกับเส้นตรง ค่าคลาดเคลื่อนเป็น h^2 กฎของซิมป์สันตรงกับพหุนามดีกรีสาม ค่าคลาดเคลื่อนเป็น h^4
- เพิ่มจำนวนจุดในสูตรนิวตัน–โคตส์ไม่ได้ทำให้ดีขึ้นเสมอไป น้ำหนักจะเป็นลบ และค่าคลาดเคลื่อนสลับเครื่องหมายและโตขึ้น จึงใช้กฎแบบประกอบ
- กลับไปที่รถผี: เพื่อรับประกันค่าคลาดเคลื่อนสัมบูรณ์ไม่เกินหนึ่งเซนติเมตร กฎเชิงสี่เหลี่ยมคางหมูต้องใช้ 48 ช่วง และซิมป์สันต้องใช้ 8 ช่วง แต่ค่าจริงใช้เพียง 11 และ 4 ช่วง

อ่านต่อ. หัวข้อนี้มาจาก Atkinson 1989, §5.1--5.2, pp. 251--270 ทฤษฎีบทเคอร์เนลเปอาโนที่เราใช้เป็นแนวทางในการพิสูจน์ปรากฏในภาคผนวก Iserles 2009, A.2.2.6--A.2.2.7, pp. 443--444 ซึ่งเป็นการนิยามแบบทั่วไปโดยไม่มีบทพิสูจน์ในตำรานั้น หัวข้อ 5.3 จะเลือกจุดให้ดีที่สุดแทนการเว้นระยะเท่ากัน

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 5.2.1 (⋆) ให้ $f(x) = x^2, x^3$ และ x^4 บน $[0, 2]$ สำหรับแต่ละฟังก์ชัน จงหา $T_1(f)$ และ $S_2(f)$ แล้วหาค่าคลาดเคลื่อน $I(f) - T_1(f)$ และ $I(f) - S_2(f)$ อธิบายว่าเหตุใดกฎของซิมป์สันให้ค่าตรงเป๊ะสำหรับ x^2 และ x^3 แต่ไม่ตรงสำหรับ x^4

ตัวอย่าง Preprint 2026-10-23

แบบฝึกหัด 5.2.2 (⋆) ความเร็วของรถผีคือ $20 + 5\cos(t/2)$ สมมติว่าเราอ่านความเร็วได้เฉพาะที่เวลา $t = 0, 2, 4, 6$ วินาที

1. จงหาความเร็วทั้งสี่ค่า (ทศนิยมห้าตำแหน่ง) แล้วคำนวณ T_3 ของระยะบน $[0, 6]$
2. ระยะจริงคือ $120 + 10\sin 3$ เมตร จงหาค่าคลาดเคลื่อน $I - T_3$ และบอกว่าเป็นบวกหรือลบ
3. (ใบ้: ค่าเฉลี่ยของ f'' บน $[0, 6]$ คือ $\frac{f'(6)-f'(0)}{6}$ ใช้เครื่องหมายของค่านี้ตอบข้อ 2)

แบบฝึกหัด 5.2.3 (⋆⋆) ให้ $c = \frac{a+b}{2}$ จงแสดงว่า $S_2(p) = \int_a^b p(x)\,dx$ สำหรับทุกพหุนาม p ที่มีดีกรีไม่เกินสาม (แนะนำ: ตรวจกรณี $p(x) = (x-c)^k$ สำหรับ $k = 0, 1, 2, 3$ แล้วใช้ความเป็นเชิงเส้น)

แบบฝึกหัด 5.2.4 (⋆⋆) กฎจุดกึ่งกลางคือ $M(f) = (b-a)f(c)$ เมื่อ $c = \frac{a+b}{2}$

1. จง ทำ ตาม ขั้น ตอน ของ ทฤษฎีบท ค่า คลาด เคลื่อน ของ กฎ เชิง สี่เหลี่ยม คางหมู เพื่อ แสดง ว่า $I(f) - M(f) = \int_a^b K(t)f''(t)\,dt$ เมื่อ $K(t) = E_x\big[(x-t)_+\big]$ และ $E = I - M$
2. จงแสดงว่า $K(t) \geq 0$ บน $[a, b]$ และ $\int_a^b K = \frac{(b-a)^3}{24}$ แล้วสรุปว่ามี $\xi \in [a, b]$ ที่ $I(f) - M(f) = \frac{(b-a)^3}{24} f''(\xi)$
3. ใช้ $f(x) = x^4$ บน $[\frac{1}{2}, 2]$ ตรวจว่าค่าคลาดเคลื่อนจริงอยู่ในช่วงที่ข้อ 2 ทำนายไว้

แบบฝึกหัด 5.2.5 (⋆⋆) ให้ $f(x) = x^4$ บน $[0, 1]$ และให้ $h = \frac{1}{n}$

1. จงหา $T_2(f)$ และตรวจว่า $T_2(f) - I(f) = \frac{13}{160}$
2. จงแสดงว่า $T_n(f) - I(f) = \frac{h^2}{3} - \frac{h^4}{30}$ สำหรับทุก n (ใช้ $\sum_{k=1}^{n-1} k^4 = \frac{(n-1)n(2n-1)(3n^2-3n-1)}{30}$)
3. เปรียบเทียบกับสูตรออยเลอร์–แมคลอริน $\frac{h^2}{12}\big(f'(1) - f'(0)\big) - \frac{h^4}{720}\big(f'''(1) - f'''(0)\big)$ จงตรวจว่าตรงกันหรือไม่ และอธิบายว่าเหตุใดพจน์ถัดไปทั้งหมดเป็นศูนย์

แบบฝึกหัด 5.2.6 (⋆ ⋆ ⋆) จุดหักมุมทำให้อันดับของสูตรลดลง ให้ $h > 0$

1. จงแสดงว่าสำหรับ $f(t) = |t|$ บน $[-h, h]$ กฎของซิมป์สันหนึ่งช่วงมีค่าคลาดเคลื่อน $I(f) - S_2(f) = \frac{h^2}{3}$
2. จงแสดงว่าสำหรับ $f(t) = |t - h|$ บน $[-h, h]$ กฎของซิมป์สันหนึ่งช่วงแม่นพอดี
3. ให้ $g(t) = 20 + |t - 1.3|$ บน $[0, 2]$ คำนวณ S_{16} และ S_{32} จงหาอัตราส่วน $|I - S_{16}|/|I - S_{32}|$ แล้วอธิบายว่าทำไมค่าคลาดเคลื่อนของซิมป์สันจึงเป็นอันดับ h^2 ไม่ใช่ h^4 (ใบ้: ช่วงย่อยที่ไม่มีจุดหักมุมมีค่าคลาดเคลื่อนศูนย์ เพราะ g เป็นเส้นตรงบนช่วงนั้น)

แบบฝึกหัด 5.2.7 (⋆ ⋆ ⋆) ระบบรถผีเก็บความเร็วบน $[0, 6]$ ที่เวลาเว้นระยะเท่ากัน รวม $n + 1$ ค่า และหน้าจอคำนวณระยะด้วย T_n หรือ S_n ตอบคำถามต่อไปนี้

1. จงหาจำนวนช่วง n ที่น้อยที่สุดที่รับประกันค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 0.01 เมตร สำหรับกฎเชิงสี่เหลี่ยมคางหมูและกฎของซิมป์สัน โดยใช้ขอบเขตจากทฤษฎีบท (ซิมป์สันต้องเป็นจำนวนคู่)
2. จงหาจำนวนช่วงที่น้อยที่สุดที่ค่าคลาดเคลื่อนสัมบูรณ์จริงไม่เกิน 0.01 เมตร จากการคำนวณ
3. อธิบายว่าเหตุใดคำตอบข้อ 1 และข้อ 2 ต่างกัน และถ้าอยากเก็บจำนวนค่าความเร็วให้น้อยที่สุด ควรเลือกกฎใด
4. ถ้าเก็บความเร็วทุกเฟรม ($n = 360$) ค่าคลาดเคลื่อนของซิมป์สันเป็นเท่าไร

ร่างแรก · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 5.2.1.

1. x^2: $T_1 = \frac{2-0}{2}(0^2+2^2) = 4$ และ $S_2 = \frac{2}{6}(0+4\cdot 1+4) = \frac{8}{3}$ เพราะ $I = \int_0^2 x^2\,dx = \frac{8}{3}$ ค่าคลาดเคลื่อนคือ $I - T_1 = -\frac{4}{3}$ และ $I - S_2 = 0$

2. x^3: $T_1 = 1\cdot(0+8) = 8$ และ $S_2 = \frac{1}{3}(0+4\cdot 1+8) = 4$ เพราะ $I = 4$ ค่าคลาดเคลื่อนคือ $I - T_1 = -4$ และ $I - S_2 = 0$

3. x^4: $T_1 = 1\cdot(0+16) = 16$ และ $S_2 = \frac{1}{3}(0+4\cdot 1+16) = \frac{20}{3}$ เพราะ $I = \frac{32}{5}$ ค่าคลาดเคลื่อนคือ $I - T_1 = -\frac{48}{5}$ และ $I - S_2 = -\frac{4}{15}$

กฎของซิมป์สันตรงกับทุกพหุนามดีกรีไม่เกินสาม (แบบฝึกหัด 5.2.3) จึงตรงกับ x^2 และ x^3 ส่วน x^4 มีดีกรีสี่ ทฤษฎีบทของซิมป์สันให้ $I - S_2 = -\frac{(b-a)^5}{2880}f^{(4)}(\xi) = -\frac{32}{2880}\cdot 24 = -\frac{4}{15}$ พอดี เพราะ $f^{(4)} = 24$ คงที่

เฉลย 5.2.2.

1. $v(0) = 20+5\cos 0 = 25$, $v(2) = 20+5\cos 1 \approx 22.70151$, $v(4) = 20+5\cos 2 \approx 17.91927$ และ $v(6) = 20+5\cos 3 \approx 15.05004$ ใช้ $n = 3$ จึง $h = 2$ และ

$$T_3 = 2\left[\tfrac{25+15.05004}{2} + 22.70151 + 17.91927\right] = 2(20.02502 + 40.62078) \approx 121.2916$$

2. ระยะจริงคือ $120 + 10\sin 3 \approx 121.41120$ ดังนั้น $I - T_3 \approx 0.11961$ เป็นบวก หมายความว่าค่าที่อ่านได้ต่ำกว่าระยะจริง

3. $f''(t) = -\frac{5}{4}\cos\frac{t}{2}$ และค่าเฉลี่ยของ f'' บน $[0,6]$ คือ $\frac{f'(6)-f'(0)}{6} = \frac{-2.5\sin 3 - 0}{6} \approx -0.0588$ ซึ่งเป็นลบ ทฤษฎีบทค่าคลาดเคลื่อนของกฎเชิงสี่เหลี่ยมคางหมูให้ $I - T_3 = -\frac{(b-a)h^2}{12}f''(\xi)$ โดย $f''(\xi)$ มีค่าใกล้ค่าเฉลี่ยนี้ จึงได้ค่าคลาดเคลื่อนเป็นบวก ซึ่งสอดคล้องกับข้อ 2

เฉลย 5.2.3. ให้ $h = \frac{b-a}{2}$ จะได้ $a = c-h$ และ $b = c+h$ ทุกพหุนามดีกรีไม่เกินสามเขียนเป็น $\sum_{k=0}^{3}\alpha_k(x-c)^k$ ได้ และทั้ง S_2 และปริพันธ์เป็นเชิงเส้นในฟังก์ชัน จึงพอตรวจ $k = 0,1,2,3$

- $k=0$: $S_2(1) = \frac{2h}{6}(1+4+1) = 2h = \int_a^b 1\,dx$
- $k=1$: $f(a) = -h$, $f(c) = 0$, $f(b) = h$ จึง $S_2 = \frac{2h}{6}(-h+0+h) = 0$ และ $\int_a^b (x-c)\,dx = 0$
- $k=2$: $f(a) = h^2$, $f(c) = 0$, $f(b) = h^2$ จึง $S_2 = \frac{2h}{6}\cdot 2h^2 = \frac{2h^3}{3}$ และ $\int_a^b (x-c)^2\,dx = \frac{2h^3}{3}$
- $k=3$: $f(a) = -h^3$, $f(c) = 0$, $f(b) = h^3$ จึง $S_2 = 0$ และ $\int_a^b (x-c)^3\,dx = 0$ เพราะฟังก์ชันเป็นคี่รอบ c

ทุกกรณีตรงกัน จึงได้ $S_2(p) = \int_a^b p$ สำหรับทุก p ดีกรีไม่เกินสาม

สัมมนา · Preprint · 2026-10-23

เฉลย 5.2.4.

1. ตามขั้นตอนของทฤษฎีบท เขียน $f(x)=f(a)+f'(a)(x-a)+\int_a^b (x-t)_+ f''(t)\,dt$ กฎ I และ M ตรงกับ 1 และ x ทั้งคู่ ($M(1)=b-a$ และ $M(x)=(b-a)c=\frac{b^2-a^2}{2}$) จึงได้ $E(f)=\int_a^b K(t)f''(t)\,dt$ โดย $K(t)=\int_a^b (x-t)_+\,dx-(b-a)(c-t)_+$ และ $\int_a^b (x-t)_+\,dx=\frac{(b-t)^2}{2}$ จึงได้ $K(t)=\frac{(b-t)^2}{2}-(b-a)(c-t)_+$

2. ให้ $h=\frac{b-a}{2}$ บน $[c,b]$ ได้ $K(t)=\frac{(b-t)^2}{2}\ge 0$ บน $[a,c]$ ให้ $u=c-t\in[0,h]$ จะได้ $b-t=h+u$ และ $K=\frac{(h+u)^2}{2}-2hu=\frac{(h-u)^2}{2}=\frac{(t-a)^2}{2}\ge 0$ ดังนั้น $K\ge 0$ บน $[a,b]$ และ
$$\int_a^b K=2\int_0^h \frac{s^2}{2}\,ds=\frac{h^3}{3}=\frac{(b-a)^3}{24}$$
เมื่อ K ไม่เปลี่ยนเครื่องหมายและ f'' ต่อเนื่อง ทฤษฎีบทค่ามัชฌิมของปริพันธ์ให้ $\xi\in[a,b]$ ที่ $E(f)=f''(\xi)\int_a^b K$

3. $f=x^4$, $f''=12x^2$ บน $[\frac12,2]$ ค่า $\frac{(b-a)^3}{24}=\frac{3.375}{24}=0.140625$ และ $f''\in[3,48]$ จึงทำนายว่าค่าคลาดเคลื่อนอยู่ใน $[0.421875,\ 6.75]$ ค่าจริงคือ $I=\frac{2^5-(1/2)^5}{5}=\frac{1023}{160}=6.39375$ และ $M=1.5\cdot(1.25)^4=3.662109375$ จึง $I-M=2.731640625$ ซึ่งอยู่ในช่วงนั้น

เฉลย 5.2.5.

1. $h=\frac12$: $T_2=\frac12\big[\frac02+\frac{1}{16}+\frac12\big]=\frac12\cdot\frac{9}{16}=\frac{9}{32}$ และ $I=\frac15$ จึง $T_2-I=\frac{9}{32}-\frac15=\frac{45-32}{160}=\frac{13}{160}$

2. จากนิยามของกฎเชิงสี่เหลี่ยมคางหมู เมื่อ $h=1/n$ เรามี
$$T_n=h\big[\frac12+\sum_{k=1}^{n-1}(kh)^4\big]=\frac{h}{2}+h^5\sum_{k=1}^{n-1}k^4.$$
ใช้สูตรผลรวมที่โจทย์ให้ จะได้
$$h^5\sum k^4=\frac{(n-1)n(2n-1)(3n^2-3n-1)}{30\,n^5}.$$
เมื่อกระจายตัวเศษ $(n-1)(2n-1)(3n^2-3n-1)=6n^4-15n^3+10n^2-1$ จึงได้
$$h^5\sum_{k=1}^{n-1}k^4=\frac{6n^4-15n^3+10n^2-1}{30n^4}=\frac15-\frac{1}{2n}+\frac{1}{3n^2}-\frac{1}{30n^4}.$$
รวมกับ $\frac{h}{2}=\frac{1}{2n}$ ได้ $T_n=\frac15+\frac{h^2}{3}-\frac{h^4}{30}$ ดังนั้น $T_n-I=\frac{h^2}{3}-\frac{h^4}{30}$ ตรวจ $n=2$: $\frac{1}{12}-\frac{1}{480}=\frac{13}{160}$

3. สูตรออยเลอร์–แมคลอริน: $f'(1)-f'(0)=4$ จึงพจน์แรกคือ $\frac{h^2}{12}\cdot 4=\frac{h^2}{3}$ และ $f'''(x)=24x$ จึง $f'''(1)-f'''(0)=24$ พจน์ที่สองคือ $-\frac{h^4}{720}\cdot 24=-\frac{h^4}{30}$ ตรงกับข้อ 2 พจน์ถัดไปทุกตัวมีรูป $f^{(2k-1)}(1)-f^{(2k-1)}(0)$ โดย $2k-1\ge 5$ และ $f^{(5)}\equiv 0$ สำหรับ x^4 จึงเป็นศูนย์

สัมมนา Preprint · 2026-10-23

เฉลย 5.2.6.

1. $I = \int_{-h}^{h} |t|\,dt = h^2$ และ $S_2 = \frac{2h}{6}\big(|-h| + 4|0| + |h|\big) = \frac{h}{3}\cdot 2h = \frac{2h^2}{3}$ ดังนั้น $I - S_2 = \frac{h^2}{3}$

2. บน $[-h, h]$ ได้ $|t - h| = h - t$ ซึ่งเป็นเส้นตรง ค่าที่จุดคือ $f(-h) = 2h$, $f(0) = h$, $f(h) = 0$ จึง $S_2 = \frac{2h}{6}(2h + 4h + 0) = 2h^2$ และ $I = \int_{-h}^{h}(h - t)\,dt = 2h^2$ ดังนั้น $I - S_2 = 0$

3. ให้ $g(t) = 20 + |t - 1.3|$ ช่วงย่อยของ S_n ที่ไม่มี 1.3 ให้ g เป็นเส้นตรง ซึ่งซิมป์สันแม่นพอดี (แบบฝึกหัด 5.2.3) ค่าคลาดเคลื่อนรวมจึงเท่ากับค่าคลาดเคลื่อนของช่วงเดียวที่มี 1.3 ถ้าช่วงนั้นมีครึ่งความกว้าง h และมุมอยู่ห่างจากจุดกลาง δ (โดย $|\delta| \le h$) จะได้ $I = h^2 + \delta^2$ และ $S_2 = \frac{2h^2}{3} + \frac{4h|\delta|}{3}$ จึง

$$I - S_2 = \frac{h^2}{3} + \delta^2 - \frac{4h|\delta|}{3}$$

ซึ่งเป็น $O(h^2)$ ไม่ใช่ $O(h^4)$ สำหรับ $n = 16$ ได้ $h = \frac{1}{8}$ และ $\delta = -\frac{3}{40}$ จึง $I - S_{16} = -\frac{1}{600}$ สำหรับ $n = 32$ ได้ $h = \frac{1}{16}$ และ $\delta = -\frac{1}{80}$ จึง $I - S_{32} = \frac{1}{2400}$ อัตราส่วน $|I - S_{16}|/|I - S_{32}| = 4$ พอดี แต่นี่เป็นผลของตำแหน่งมุม 1.3 บนกริดนี้ เพราะค่าคลาดเคลื่อนเป็นฟังก์ชันของ δ/h ซึ่งเปลี่ยนไปตามกริด อัตราส่วนจึงไม่ได้เป็น 16 อย่างที่อันดับ h^4 จะให้

เฉลย 5.2.7.

1. ขอบเขตของทฤษฎีบทคือ $|E_T| \le \frac{6h^2}{12}\cdot 1.25 = \frac{22.5}{n^2}$ และ $|E_S| \le \frac{6h^4}{180}\cdot\frac{5}{16} = \frac{13.5}{n^4}$ โดย $h = \frac{6}{n}$ เมื่อ $\frac{22.5}{n^2} \le 0.01$ ได้ $n^2 \ge 2250$ จึง $n \ge 47.4$ ได้ $n = 48$ เมื่อ $\frac{13.5}{n^4} \le 0.01$ ได้ $n^4 \ge 1350$ จึง $n \ge 6.07$ แต่ n ต้องเป็นจำนวนคู่ ที่ $n = 6$ ขอบเขตคือ 0.0104 เกิน จึงได้ $n = 8$ คำตอบคือ 48 ช่วงสำหรับกฎเชิงสี่เหลี่ยมคางหมู และ 8 ช่วงสำหรับกฎของซิมป์สัน

2. จากการคำนวณ: กฎเชิงสี่เหลี่ยมคางหมูที่ $n = 10$ มี $|E_T| = 0.01060$ เกิน ที่ $n = 11$ มี $|E_T| = 0.008758$ จึงได้ $n = 11$ สำหรับกฎของซิมป์สัน $n = 2$ มี $|E_S| = 0.0536$ เกิน ที่ $n = 4$ มี $|E_S| = 0.00266$ จึงได้ $n = 4$

3. ขอบเขตต้องรับประกันทุกช่วงของ $[0, 6]$ จึงใช้ค่าสูงสุดของ $|f''|$ และ $|f^{(4)}|$ ซึ่งบรรลุเพียงที่ $t = 0$ ส่วนค่าจริงขึ้นกับค่าเฉลี่ยทั้งช่วง เกณฑ์รับประกันจึงต้องใช้จำนวนช่วงมากกว่า สำหรับกฎเชิงสี่เหลี่ยมคางหมูคือ 48 เทียบกับ 11 ช่วง ถ้าต้องการเก็บค่าน้อยที่สุดโดยยังผิดไม่เกินหนึ่งเซนติเมตร ควรใช้กฎของซิมป์สัน 4 ช่วง คือเก็บ 5 ค่า เทียบกับกฎเชิงสี่เหลี่ยมคางหมูที่ต้องเก็บ 12 ค่า

4. ที่ $n = 360$ ได้ $I - S_{360} \approx -3.782 \times 10^{-11}$ เมตร ซึ่งเล็กกว่าที่จอใด ๆ จะแสดงมาก ดังนั้นการเก็บความเร็วทุกเฟรมให้ความแม่นเกินจำเป็นสำหรับระยะนี้

5.3 เลือกจุดให้ฉลาด

สถานการณ์สมมติ: คำถามที่แพง

ลองนึกถึงโหมดรถผีอีกครั้ง คราวนี้ระบบต้องบอกระยะที่รถวิ่งไปในหกวินาทีแรก ความเร็วของรถผีไม่ได้ถูกเก็บไว้ทุกเวลา ระบบต้องถามความเร็วจากระบบฟิสิกส์ของเกมทุกครั้งที่ต้องการ และแต่ละคำถามให้ระบบคำนวณแรงและการชนของทั้งฉาก จึงแพงมาก สมมติว่าเราอนุญาตให้ถามได้เพียง **สามครั้ง**
ในหัวข้อ 5.2 เราถามที่เวลาที่เว้นระยะเท่ากันเสมอ เช่นที่ 0, 3 และ 6 วินาที ซึ่งให้ค่าคลาดเคลื่อน 0.275 เมตร คำถามของหัวข้อนี้คือ ถ้าเลือกเวลาให้ดีกว่านั้นได้ ค่าคลาดเคลื่อนจะเล็กลงแค่ไหน และเรารู้ได้อย่างไรว่าเลือกดีที่สุดแล้ว (ตัวเลข

ฉบับร่าง Preprint · 2026-10-23

ทั้งหมดมาจากโค้ดของเรา)

ภาพรวมก่อนอ่าน

ในหัวข้อ 5.2 จุดที่ใช้คำนวณถูกบังคับให้เว้นระยะเท่ากัน แล้วน้ำหนักของแต่ละจุดจึงถูกหาตามมา หัวข้อนี้เลือก*ทั้งจุดและน้ำหนัก*ไปพร้อมกัน ลองนับดูก่อน ถ้ามี N จุดและน้ำหนักอีก N ตัว เราเลือกตัวแปรได้ $2N$ ตัว และอาจบังคับให้สูตรแม่นกับ $1, x, \ldots, x^{2N-1}$ ซึ่งเป็นเงื่อนไข $2N$ ข้อ ทฤษฎีบทข้างล่างจะบอกว่าทำได้จริงเท่าไร

เลือกทั้งจุดและน้ำหนัก

สูตรที่เลือกทั้งจุดและน้ำหนักแบบนี้เรียกว่า**ควอดราเจอร์แบบเกาส์** (Gaussian quadrature) ซึ่งต่อไปเราจะเรียกสั้น ๆ ว่า*สูตรเกาส์* ลองทำกับ $N = 2$ บนช่วง $[-1, 1]$ ก่อน ให้จุดสองจุดสมมาตรกันที่ $\pm c$ และให้น้ำหนักเท่ากันที่ w ทั้งคู่ เงื่อนไขแรกคือสูตรต้องให้ค่าเท่ากับปริพันธ์ของ 1 ซึ่งคือ 2 ดังนั้น $2w = 2$ จึง $w = 1$ เงื่อนไขที่สองคือต้องแม่นกับ x^2 ซึ่งปริพันธ์คือ $\frac{2}{3}$ ดังนั้น $2c^2 = \frac{2}{3}$ จึง $c = \frac{1}{\sqrt{3}}$ สูตรที่ได้คือ

$$Q(f) = f\left(-\tfrac{1}{\sqrt{3}}\right) + f\left(\tfrac{1}{\sqrt{3}}\right).$$

ตรวจต่อว่าแม่นกับ x และ x^3 ด้วย ทั้งสองเป็นฟังก์ชันคี่ ค่าที่จุดสมมาตรจึงหักล้างกันพอดี และปริพันธ์บนช่วงสมมาตรก็เป็นศูนย์เช่นกัน สูตรนี้จึงแม่นกับพหุนามดีกรีไม่เกินสาม เท่ากับกฎของซิมป์สันที่ใช้สามจุด แต่ใช้เพียงสองจุด

ลองเอง: สูตรสองจุดกับ x^4

ให้ $f(x) = x^4$ คำนวณ $Q(f)$ ด้วยจุด $\pm\frac{1}{\sqrt{3}}$ แล้วเทียบกับ $\int_{-1}^{1} x^4\,dx$ ค่าคลาดเคลื่อนที่ได้เป็นศูนย์หรือไม่ ผลนี้บอกอะไรเกี่ยวกับดีกรีของสูตรสองจุด (คำตอบอยู่ในแบบฝึกหัด 5.3.2)

ในรถผี: สามจุดบนช่วง $[0, 6]$

ช่วง $[0, 6]$ ไม่ใช่ $[-1, 1]$ เราเลื่อนจุดกึ่งกลางไปที่ 3 และยืดด้วยครึ่งความกว้าง 3 จุดของสูตรสามจุดจึงเป็น $3 - 3\sqrt{3/5}$, 3 และ $3 + 3\sqrt{3/5}$ หรือประมาณ 0.676, 3 และ 5.324 วินาที น้ำหนักคือ $3 \cdot \frac{5}{9} = \frac{5}{3}$, $3 \cdot \frac{8}{9} = \frac{8}{3}$ และ $\frac{5}{3}$ ซึ่งรวมกันได้ 6 พอดี

ความเร็วของรถผีที่จุดทั้งสามคือ 24.717, 20.354 และ 15.564 เมตรต่อวินาที (จากโค้ดของเรา) ผลรวมถ่วงน้ำหนักคือ

$$\tfrac{5}{3} \cdot 24.717 + \tfrac{8}{3} \cdot 20.354 + \tfrac{5}{3} \cdot 15.564 \approx 121.4119$$

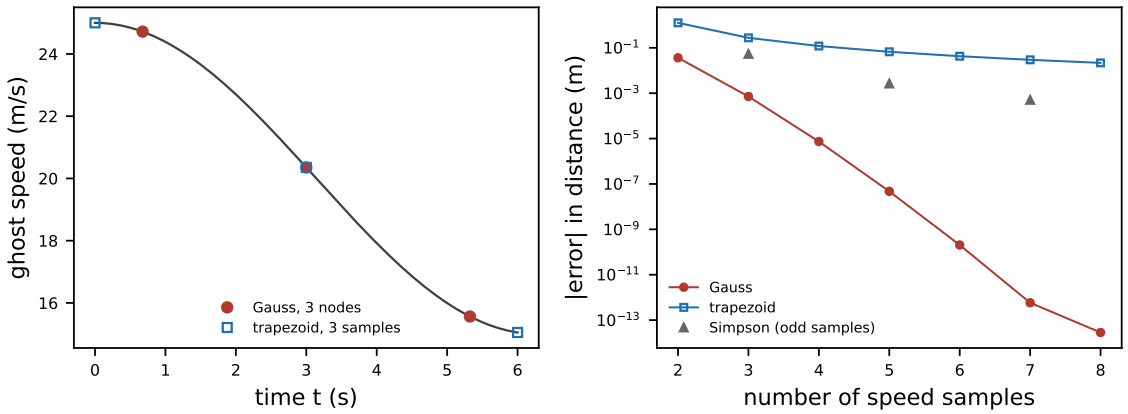
ค่าจริงคือ 121.41120 เมตร ดังนั้นค่าคลาดเคลื่อนคือ -0.00072 เมตร หรือราว 0.7 มิลลิเมตร

ลองเอง: สามจุดของรถผี

ใช้เครื่องคิดเลขหาความเร็ว $20 + 5\cos(t/2)$ ที่ $t = 0.676$, 3 และ 5.324 แล้วคูณด้วยน้ำหนัก $\frac{5}{3}, \frac{8}{3}, \frac{5}{3}$ และบวกกัน ผลควรใกล้ 121.41 เมตร จากนั้นลองทำแบบเดียวกันกับจุด 0, 3 และ 6 ด้วยกฎเชิงสี่เหลี่ยมคางหมูสองช่วง ผลควรเป็น 121.136 เมตร

เมื่อเทียบกับจำนวนตัวอย่างสามตัวเท่ากัน ค่าคลาดเคลื่อนของกฎเชิงสี่เหลี่ยมคางหมูคือ 0.275 เมตร และของซิมป์สันคือ -0.054 เมตร (จากหัวข้อ 5.2) สูตรเกาส์เล็กกว่าเชิงสี่เหลี่ยมคางหมูราว 380 เท่า และเล็กกว่าซิมป์สันราว 75 เท่า ภาพทางขวาของรูปข้างล่างแสดงว่าเมื่อเพิ่มตัวอย่าง ค่าคลาดเคลื่อนของเกาส์ลดลงเร็วมาก ที่สี่ตัวอย่างเหลือราว 7×10^{-6} เมตร ที่ห้าตัวอย่างเหลือราว 5×10^{-8} เมตร ขณะที่กฎเชิงสี่เหลี่ยมคางหมูที่ห้าตัวอย่างยังผิดราว 0.067 เมตร

ฉบับร่าง · Preprint · 2026-10-23



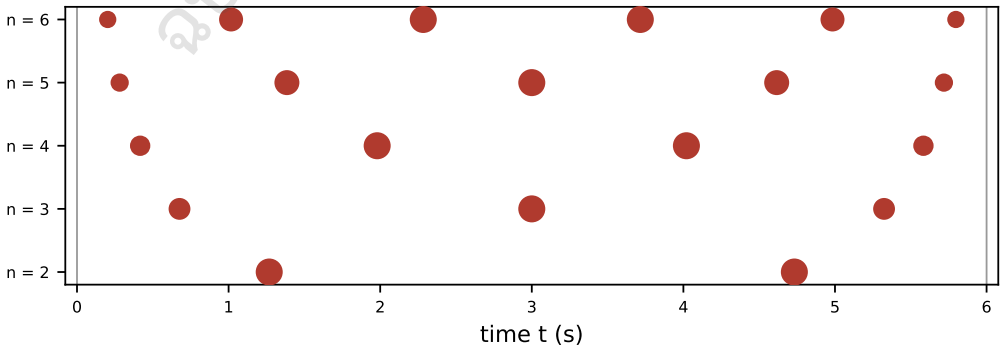
รูปที่ 5.7: ซ้าย: ความเร็วของรถผีบน $[0, 6]$ จุดแดงคือจุดเกาส์สามจุด จุดกลวงคือสามตัวอย่างที่เว้นระยะเท่ากันของกฎเชิงสี่เหลี่ยมคางหมู ขวา: ค่าคลาดเคลื่อนของระยะเทียบกับจำนวนตัวอย่างที่ใช้ บนสเกลลอการิทึม สามเหลี่ยมคือซิมป์สันที่จำนวนตัวอย่างเป็นเลขคี่ (สร้างจาก `numanal/ch05_gauss.py`)

ตารางที่ 5.3: จำนวนตัวอย่างความเร็วที่น้อยที่สุดที่ทำให้ค่าคลาดเคลื่อนสัมบูรณ์ของระยะรถผีบน $[0, 6]$ ไม่เกินค่าที่กำหนด (ตัวเลขจากโค้ดของเรา)

วิธี	ไม่เกิน 1 เซนติเมตร	ไม่เกิน 1 มิลลิเมตร
สูตรเกาส์	3	3
ซิมป์สันแบบประกอบ	5	7
เชิงสี่เหลี่ยมคางหมูแบบประกอบ	12	34

จำนวนตัวอย่างที่ต้องใช้ขึ้นกับความแม่นที่ต้องการ ตารางข้างล่างเทียบจำนวนตัวอย่างที่น้อยที่สุดของแต่ละวิธีให้ค่าคลาดเคลื่อนสัมบูรณ์ของระยะรถผีไม่เกินหนึ่งเซนติเมตรและหนึ่งมิลลิเมตร เมื่อเป้าหมายเล็กลง ความแตกต่างระหว่างสูตรเกาส์กับสูตรแบบประกอบมากขึ้น

รูปข้างล่างแสดงโหนดของสูตรเกาส์ตั้งแต่สองถึงหกจุดบนช่วงเดียวกัน เมื่อ n เพิ่มขึ้น โหนดจะกระจุกใกล้ปลายช่วงมากขึ้น และเมื่อ n เป็นเลขคี่ จุด 3 ซึ่งเป็นจุดกึ่งกลางจะเป็นโหนดเสมอ เพราะ P_n เป็นฟังก์ชันคี่



รูปที่ 5.8: ตำแหน่งโหนดของสูตรเกาส์ $n = 2$ ถึง 6 บนช่วง $[0, 6]$ วินาที ขนาดจุดแสดงน้ำหนักของโหนดนั้น (สร้างจาก `numanal/ch05_gauss.py`)

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 5.4: สูตรเกาส์ n จุดกับความยาวของเส้นทางกล้อง ค่าอ้างอิงคือ 4.50230 เมตร (ตัวเลขจากโค้ดของเรา)

n	ความยาวที่ประมาณได้ (เมตร)	ค่าคลาดเคลื่อน (มิลลิเมตร)
2	4.48047	21.84
3	4.49641	5.89
4	4.50055	1.75
5	4.50171	0.593
6	4.50208	0.221

ข้อควรระวัง: จุดของสูตรเกาส์ไม่ซ้อนกันและไม่ตรงกับเฟรม

สูตรสามจุดกับห้าจุดบน $[0, 6]$ มีจุดร่วมกันเพียงจุดกลาง 3 ดังนั้นตัวอย่างที่เก็บไว้สำหรับสูตรสามจุดนำมาใช้ซ้ำในสูตรห้าจุดแทบไม่ได้ ตำราชี้ว่าจุดของแต่ละสูตรต่างจากสูตรก่อนหน้า Atkinson 1989, §5.3, p. 282 นอกจากนี้จุดเกาส์มักไม่ตรงกับเวลาของเฟรม $1/60$ วินาที และไม่รวมปลายช่วง ระบบที่เก็บค่าเฉพาะที่เฟรมจึงใช้สูตรนี้ตรง ๆ ไม่ได้ ต้องถามความเร็วที่เวลาที่เราเลือกเองได้

สถานการณ์สมมติ: อีกด้านหนึ่งของปัญหาเดียวกัน: กล้องที่วิ่งด้วยความเร็วคงที่

ในฉากคัตซีน ผู้ออกแบบวางกล้องไว้บนเส้นโค้งเบซิเยร์ลูกบาศก์ที่กำหนดด้วยจุดควบคุม แบบเดียวกับหัวข้อ 3.3 แล้วให้กล้องเคลื่อนตามเส้นโค้งตลอดสามวินาที โดยต้องการให้กล้องวิ่งด้วยความเร็วคงที่ นั่นคือเวลาหนึ่งในสี่ของการเคลื่อนที่ต้องผ่านระยะหนึ่งในสี่ของเส้นทางพอดี

วิธีที่ง่ายที่สุดคือแบ่งเวลาเท่า ๆ กัน แล้วให้ตัวแปรของเส้นโค้งเดินเท่า ๆ กัน แต่ตัวแปรนี้ไม่ใช่ระยะ ตรงที่เส้นโค้งโค้งมาก ตัวแปรเดินเร็วและกล้องก็เร็วตาม ในตัวอย่างของเรา ความเร็วของกล้องจะเปลี่ยนจากราว 0.30 เมตรต่อวินาทีตอนเริ่มไปถึงราว 1.97 เมตรต่อวินาทีตอนกลางทาง ซึ่งห่างกันราว 6.6 เท่า (ตัวเลขทั้งหมดมาจากโค้ดของเรา และเป็นการสร้างขึ้นเพื่อการเรียน ไม่เกี่ยวกับเกมใด)

ฉบับร่าง · Preprint 2016-10-23

ต้นเหตุของความเร็วที่ไม่คงที่คือตัวแปรของเส้นโค้งไม่ได้วัดระยะ ไม่ใช่ความผิดจากการคำนวณ ระยะตามเส้นโค้งจากจุดเริ่มถึงพารามิเตอร์ s คือ

$$\ell(s) = \int_0^s |B'(\sigma)|\, d\sigma,$$

เมื่อ B คือเส้นโค้งเบซิเยร์ $\ell(s)$ คือ**ความยาวส่วนโค้ง** (arc length) จากจุดเริ่มถึงพารามิเตอร์ s โดย $|B'(\sigma)|$ คือความเร็วของกล้องเทียบกับพารามิเตอร์ ปริพันธ์นี้เป็นรากที่สองของพหุนามดีกรีสี่ ซึ่งโดยทั่วไปไม่มีสูตรในรูปที่เราคุ้นเคย เราจึงต้องประมาณมันด้วยสูตรเกาส์ ซึ่งเป็นเครื่องมือเดียวกับที่ใช้กับรถผี

เส้นทางทั้งหมดยาว $L = \ell(1) = 4.50230$ เมตร ตารางข้างล่างแสดงว่าสูตรเกาส์ที่มีจำนวนจุดต่างกันให้ความยาวและค่าคลาดเคลื่อนเท่าใด ค่าอ้างอิงที่ใช้เทียบคำนวณด้วยสูตรเกาส์แปดจุดบน 64 ช่วงย่อย ซึ่งได้ค่าเดิมถึงทศนิยมสิบสามตำแหน่งเมื่อเพิ่มเป็น 128 ช่วงย่อย

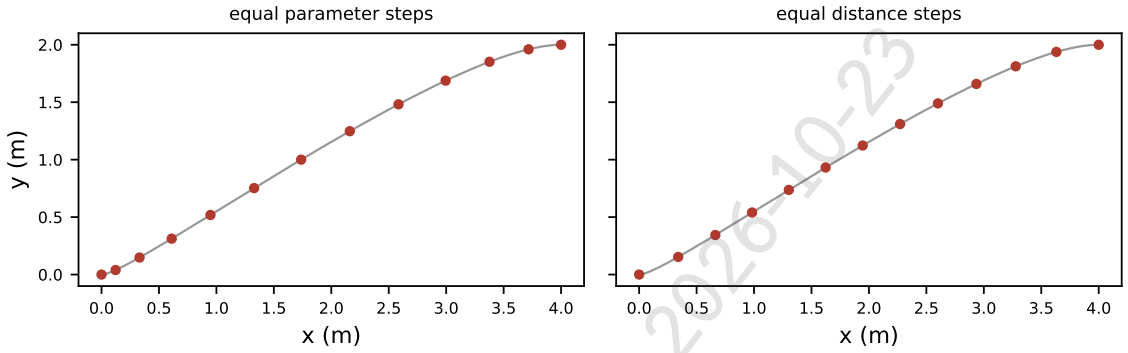
สามจุดให้ค่าคลาดเคลื่อนราว 5.9 มิลลิเมตร และห้าจุดให้ราว 0.6 มิลลิเมตร ค่าคลาดเคลื่อนลดลงเร็วเมื่อจำนวนจุดเพิ่ม แต่ที่สองจุดยังผิดเกินหนึ่งเซนติเมตรมาก ค่าคลาดเคลื่อนของกล้องเป็นค่าคลาดเคลื่อนเชิงตัวเลขที่เราควบคุมได้ด้วยจำนวนจุด ส่วนความเร็วที่ไม่คงที่ของกล้องแบบตัวแปรเท่ากันเป็นเรื่องของการเลือกตัวแปร

ถ้ากล้องต้องการระยะ $kL/4$ ที่เวลา $3k/4$ วินาที เราต้องหาตัวแปร s ที่ทำให้ $\ell(s) = kL/4$ ตารางข้างล่างเทียบกับกล้องที่ใช้ตัวแปร $s = k/4$ ในเวลาเดียวกัน

ในช่วงเวลาสามส่วนสี่วินาทีแต่ละช่วง กล้องที่ใช้ตัวแปรเท่ากันเดินได้ 0.689, 1.320, 1.434 และ 1.059 เมตร ขณะที่กล้อง

ตารางที่ 5.5: ตำแหน่งกล้องที่เวลาเท่ากันเมื่อกล้องวิ่งเร็วคงที่กับเมื่อตัวแปรเดินเท่ากัน (เส้นทางยาว $L = 4.50230$ เมตร ใช้เวลาสามวินาที)

เวลา τ (วินาที)	ระยะที่ควรได้ $kL/4$ (เมตร)	ระยะที่ได้เมื่อ $s = k/4$ (เมตร)	s ที่ถูกต้อง
0.00	0.00000	0.00000	0.00000
0.75	1.12558	0.68928	0.34137
1.50	2.25115	2.00906	0.54117
2.25	3.37673	3.44305	0.73754
3.00	4.50230	4.50230	1.00000

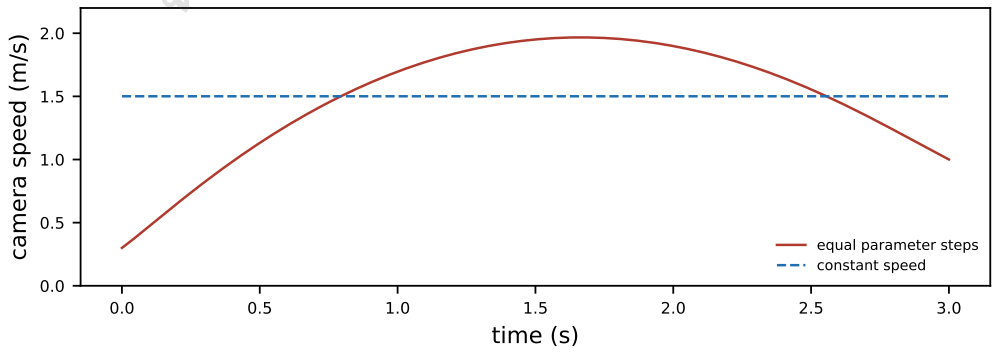


รูปที่ 5.9: เส้นโค้งเบซิเยร์ลูกบาศก์ของกล้อง (เส้นเทา) และตำแหน่งกล้องทุกหนึ่งในสิบสองของเวลา ซ้าย: ตัวแปร s เท่ากันทุกช่วง จุดแน่นตอนเริ่มและห่างตอนกลาง ขวา: ระยะตามเส้นโค้งเท่ากันทุกช่วง (สร้างจาก `numanal/ch05_gauss.py`)

ที่วิ่งคงที่ต้องได้ 1.126 เมตรทุกช่วง รูปข้างล่างแสดงตำแหน่งเดียวกันบนเส้นโค้ง

รูปข้างล่างแสดงความเร็วของกล้องตามเวลาทั้งสองแบบ ความเร็วของตัวแปรเท่ากันเริ่มที่ 0.30 เมตรต่อวินาที ไปถึงค่าสูงสุดราว 1.97 เมตรต่อวินาทีตรงกลางทาง แล้วลดลงเหลือ 1.00 เมตรต่อวินาทีตอนจบ ส่วนความเร็วคงที่เท่ากับ $L/3 = 1.50$ เมตรต่อวินาทีตลอด

การหาตัวแปร s ที่ระยะ d เป็นการหารากของสมการ $\ell(s) - d = 0$ ในโค้ดของเราใช้**วิธีแบ่งครึ่ง** (bisection method) ซึ่งหารากโดยแบ่งช่วงที่ครอบรากเป็นครึ่งหนึ่งซ้ำ ๆ บทที่ 6 จะอธิบายวิธีนี้และวิธีที่เร็วกว่า ตัวแปร s ที่ได้มีค่าเดียวเสมอเพราะ



รูปที่ 5.10: ความเร็วของกล้องตามเวลาในสามวินาที เส้นทึบคือกล้องที่ตัวแปรเท่ากันทุกช่วงเวลา เส้นประคือกล้องที่วิ่งด้วยความเร็วคงที่ (สร้างจาก `numanal/ch05_gauss.py`)

ฉบับร่าง Preprint · 2026-10-23

ตารางที่ 5.6: การแบ่งครึ่งหกครั้งแรกเพื่อหา s ที่ทำให้ $\ell(s) = L/2$ (ตัวเลขจากโค้ดของเรา)

ครั้งที่	ขอบล่าง	ขอบบน	จุดกลาง	ℓ(จุดกลาง) $- L/2$
1	0	1	0.5	-0.24209
2	0.5	1	0.75	1.19190
3	0.5	0.75	0.625	0.49278
4	0.5	0.625	0.5625	0.12585
5	0.5	0.5625	0.53125	-0.05850
6	0.53125	0.5625	0.546875	0.03365

ℓ เพิ่มอย่างเคร่งครัด (แบบฝึกหัด 5.3.7) ดังนั้นกล้องที่วิ่งคงที่ต้องการทั้งความยาวตามเส้นโค้งและการหารากสมการ ซึ่งเป็นการกลับจากระยะไปหาตำแหน่งที่บทที่ 5 สัญญาไว้ในภาพรวม

ตารางข้างล่างแสดงการแบ่งครึ่งหกครั้งแรกสำหรับระยะครึ่งทาง $d = L/2 = 2.25115$ เมตร แต่ละแถวแสดงช่วงที่ยังครอบคำตอบอยู่ก่อนแบ่ง ค่าที่จุดกลางบอกว่าความยาวที่ได้ยังน้อยหรือเกินระยะที่ต้องการ หลังหกครั้งความยาวของช่วงที่ครอบคำตอบเหลือ $\frac{1}{64}$ ของช่วงเริ่มต้น คำตอบจริงคือ $s \approx 0.54117$

ทำไมสามจุดแม่นขนาดนี้

ก่อนพิสูจน์ เราต้องเลือกโหนดให้ถูก โหนดเหล่านั้นเกี่ยวข้องกับพหุนามพิเศษที่ตั้งฉากกับพหุนามดีกรีต่ำกว่า เราเรียกพหุนามพวกนี้ว่า **พหุนามเชิงตั้งฉาก** (orthogonal polynomials) ซึ่งเราพบแล้วในบทที่ 4

นิยาม 5.3.1: พหุนามเชิงตั้งฉากบน $[-1, 1]$

พหุนาม p_N ที่มีดีกรี N เรียกว่าพหุนามเชิงตั้งฉาก ถ้า $p_N \neq 0$ และ

$$\int_{-1}^{1} p_N(x)\, q(x)\, dx = 0 \qquad \text{สำหรับทุกพหุนาม } q \text{ ที่มีดีกรีไม่เกิน } N-1.$$

อ่านแบบเข้าใจง่าย

- ความตั้งฉากในที่นี้หมายความว่า ปริพันธ์ของผลคูณ $p_N q$ เป็นศูนย์ เหมือนเวกเตอร์สองตัวที่ผลคูณภายในเป็นศูนย์
- ตัวอย่างคือพหุนามเลอฌ็องดร์ $P_0 = 1$, $P_1 = x$ และ $P_2 = \frac{1}{2}(3x^2 - 1)$ เราเลือกสเกลให้ $P_N(1) = 1$ เลอฌ็องดร์ที่เราเห็นในบทที่ 4 คือกรณีนี้
- x^2 ตัวเดียวไม่ตั้งฉากกับ 1 เพราะ $\int_{-1}^{1} x^2\, dx = \frac{2}{3} \neq 0$ การลบ $\frac{1}{3}$ ออกทำให้ตั้งฉากกับ 1 ได้ (ตรวจแล้วในแบบฝึกหัด 5.3.3)

ทฤษฎีบท 5.3.1: รากของพหุนามเชิงตั้งฉาก

รากของ P_N มี N ตัว ต่างกันทั้งหมด และอยู่ในช่วง $(-1, 1)$ Iserles 2009, §3.1, Lemma 3.2, p. 36

พิสูจน์. **แนวคิด.** ถ้า P_N เปลี่ยนเครื่องหมายน้อยกว่า N ครั้ง จะหาพหุนามดีกรีต่ำกว่าที่ทำให้ผลคูณไม่เปลี่ยนเครื่องหมาย ซึ่งขัดกับความตั้งฉาก

ฉบับร่าง · Preprint 2026-10-23

สรุป: เราอ้างผลนี้จากตำรา ไม่พิสูจน์ซ้ำในบทนี้ เพราะเป็นส่วนของทฤษฎีพหุนามเชิงตั้งฉาก □

ตอนนี้เราพร้อมบอกทฤษฎีบทของสูตรเกาส์ ให้ $x_1, \dots, x_N$ เป็นรากของ P_N และให้ $\ell_j(x) = \prod_{k \neq j} \frac{x - x_k}{x_j - x_k}$ เป็นพหุนามลากรานจ์ที่เป็น 1 ที่ x_j และเป็น 0 ที่โหนดอื่น

ทฤษฎีบท 5.3.2: สูตรเกาส์บน $[-1, 1]$

ให้ $w_j = \int_{-1}^{1} \ell_j(x)\,dx$ และ $Q_N(f) = \sum_{j=1}^{N} w_j f(x_j)$ จะได้ว่า

1. $Q_N(p) = \int_{-1}^{1} p(x)\,dx$ สำหรับทุกพหุนาม p ที่มีดีกรีไม่เกิน $2N - 1$
2. สำหรับโหนด $c_1, \dots, c_N$ และน้ำหนัก $b_1, \dots, b_N$ ใด ๆ มีพหุนามดีกรี $2N$ ที่กฎไม่แม่น ดังนั้นระดับความแม่นยำสูงสุดของสูตร N จุดคือ $2N - 1$
3. น้ำหนัก w_j เป็นบวกทุกตัว

พิสูจน์. **แนวคิด.** แบ่งพหุนามด้วย $\omega(x) = \prod_j (x - x_j)$ แล้วใช้ความตั้งฉาก ส่วนข้อที่สองใช้พหุนามกำลังสองของ ω เป็นตัวอย่างค้าน และข้อที่สามใช้ข้อหนึ่ง

1. ข้อ 1 เพราะ $P_N = k_N\,\omega$ เมื่อ $k_N \neq 0$ คือสัมประสิทธิ์ของ x^N จึง $\int q\,\omega = 0$ สำหรับทุกพหุนาม q ที่มีดีกรีไม่เกิน $N - 1$ เมื่อ p มีดีกรีไม่เกิน $2N - 1$ การหารพหุนามให้ $p = q\,\omega + r$ โดย $\deg q \leq N - 1$ และ $\deg r \leq N - 1$
2. ปริพันธ์ของ p เท่ากับ $\int q\omega + \int r = \int r$
3. เพราะ $\omega(x_j) = 0$ จึง $p(x_j) = r(x_j)$ และ $Q_N(p) = \sum_j w_j r(x_j)$
4. ความแม่นของ Q_N บนพหุนามดีกรีไม่เกิน $N - 1$: เพราะ $r = \sum_j r(x_j)\ell_j$ เมื่อ $\deg r \leq N - 1$ (การประมาณด้วยพหุนามลากรานจ์ที่ตรงทุกจุด) นำไปหาปริพันธ์จะได้ $\int r = \sum_j r(x_j) w_j = Q_N(r)$
5. รวมขั้นที่สามและสี่ได้ $Q_N(p) = Q_N(r) = \int r = \int p$
6. ข้อ 2 ให้ $\nu(x) = \prod_{j=1}^{N} (x - c_j)^2$ มีดีกรี $2N$ และ $\nu \geq 0$ ไม่เป็นศูนย์เกือบทุกที่ ดังนั้น $\int \nu > 0$ แต่ $\sum_j b_j \nu(c_j) = 0$ เพราะ $\nu(c_j) = 0$ ทุก j จึงไม่มีสูตรใดแม่นกับ ν
7. ข้อ 3 ใช้ข้อ 1 กับ ℓ_j^2 ซึ่งมีดีกรี $2N - 2$ จะได้ $\int \ell_j^2 = Q_N(\ell_j^2) = \sum_k w_k \ell_j(x_k)^2 = w_j$ เพราะ $\ell_j(x_k) = 0$ เมื่อ $k \neq j$ และ $\ell_j(x_j) = 1$ ส่วน $\int \ell_j^2 > 0$ เพราะ $\ell_j^2 \geq 0$ และไม่เป็นศูนย์ จึงได้ $w_j > 0$

สรุป: สูตร N จุดแม่นถึงดีกรี $2N - 1$ เพราะโหนดเป็นรากของพหุนามที่ตั้งฉากกับพหุนามดีกรีต่ำ ไม่มีสูตรใดทำได้ดีกว่า และน้ำหนักเป็นบวก □

เกร็ดเพิ่มเติม

ตาราอิเซิร์ลส์นับ*อันดับ* (order) ของสูตรเป็น p ถ้าแม่นกับ P_{p-1} สูตรเกาส์ ν จุดจึงมีอันดับ 2ν ตรงกับทฤษฎีบท 3.3 Iserles 2009, §3.1, Thm 3.3, pp. 36--37 ส่วนตารา Atkinson ใช้ระดับความแม่นยำ $2n - 1$ เมื่อมี n จุด Atkinson 1989, §5.3, Thm 5.3, p. 272 ถ้าอ่านหนังสือเล่มอื่นที่นับจุดเป็น $n + 1$ จะได้ตัวเลข $2n + 1$ ซึ่งเป็นสิ่งเดียวกัน

ฉบับร่าง Preprint · 2026-10-23

หาโหนดด้วยค่าลักษณะเฉพาะ

การหาโหนดเป็นรากของ P_N ด้วยนิวตันต้องเดาค่าเริ่มต้นให้ดี และยังต้องหาน้ำหนักแยกต่างหาก วิธีที่เป็นระบบกว่าคือเปลี่ยนปัญหาเป็นการหา**ค่าลักษณะเฉพาะ** (eigenvalue) ของ**เมทริกซ์จาโคบี** (Jacobi matrix) ซึ่งเป็นเมทริกซ์สมมาตรที่สมาชิกบนแนวทแยงหลักเป็นศูนย์ทั้งหมด และสมาชิกที่ไม่เป็นศูนย์อยู่เฉพาะบนสองเส้นที่ติดกับแนวทแยงหลัก จึงเป็น**เมทริกซ์สามแนวเฉียง** (tridiagonal matrix) สมาชิกบนเส้นข้างเคียงคือ $\frac{k}{\sqrt{4k^2-1}}$ สำหรับ $k = 1, \dots, N-1$ ซึ่งมาจากสัมประสิทธิ์ของการเวียนเกิดของพหุนามเลอฌ็องดร์

ค่าลักษณะเฉพาะของเมทริกซ์นี้คือโหนด และน้ำหนักที่คู่กับค่านั้นคือ 2 คูณกำลังสองของสมาชิกแรกของเวกเตอร์ลักษณะเฉพาะหนึ่งหน่วย วิธีนี้เรียกว่า**อัลกอริทึมโกลับ–เวลช์** (Golub–Welsch algorithm) ซึ่งตำราอ้างถึง Atkinson 1989, §5.3, p. 282 ในโค้ดของเรา (`numanal/ch05_gauss.py`) ใช้ตัวแก้ค่าลักษณะเฉพาะของเมทริกซ์สมมาตรใน numpy แล้วตรวจกับการหารากของ P_N ด้วยนิวตันสำหรับ $N \le 10$ ได้ตรงกันภายใน 10^{-13} (ทดสอบใน `test_ch05_gauss.py`)

ลองทำด้วยมือกับ $N = 2$ เมทริกซ์จาโคบีคือ

$$J = \begin{pmatrix} 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & 0 \end{pmatrix},$$

ค่าลักษณะเฉพาะคือ $\pm\frac{1}{\sqrt{3}}$ และเวกเตอร์ลักษณะเฉพาะหนึ่งหน่วยคือ $\frac{1}{\sqrt{2}}(1, \pm 1)$ ดังนั้นโหนดคือ $\pm\frac{1}{\sqrt{3}}$ และน้ำหนักคือ $2 \cdot \frac{1}{2} = 1$ ทั้งคู่ ตรงกับที่เราได้จากเงื่อนไขการแม่นยำในตอนต้นของหัวข้อ

ค่าคลาดเคลื่อนของสูตรเกาส์

ทฤษฎีบทต่อไปนี้บอกค่าคลาดเคลื่อนในรูปอนุพันธ์ลำดับสูงของฟังก์ชัน การพิสูจน์ใช้พหุนามเฮอร์ไมต์ที่ตรงกับ f และ f' ที่โหนด (ดู Atkinson 1989, §5.3, Thm 5.3, pp. 272–273) และตำรายังให้รูปของค่าคงที่สำหรับช่วง $[-1, 1]$ ไว้ที่ Atkinson 1989, §5.3, (5.3.29), p. 276 เราไม่พิสูจน์ซ้ำในบทนี้ และรูปบนช่วงทั่วไปได้จากการเปลี่ยนตัวแปรในตำรา (5.3.30)

ทฤษฎีบท 5.3.3: ค่าคลาดเคลื่อนของสูตรเกาส์ (อ้างจากตำรา)

ให้ $f \in C^{2N}[a, b]$ และให้ Q_N เป็นสูตรเกาส์ N จุดที่ย้ายมาบน $[a, b]$ จะมี $\xi \in (a, b)$ ที่ทำให้

$$\int_a^b f(x)\,dx - Q_N(f) = C_N\,(b-a)^{2N+1} f^{(2N)}(\xi), \qquad C_N = \frac{(N!)^4}{(2N+1)\big[(2N)!\big]^3}.$$

อ่านแบบเข้าใจง่าย

- ค่าคลาดเคลื่อนขึ้นกับอนุพันธ์อันดับ $2N$ ของ f ถ้า f เป็นพหุนามดีกรีไม่เกิน $2N-1$ อนุพันธ์นั้นเป็นศูนย์และค่าคลาดเคลื่อนเป็นศูนย์ตามทฤษฎีบท 5.3.2
- สำหรับ $N = 2$ บน $[-1, 1]$ ได้ $C_2 = \frac{1}{4320}$ และ $(b-a)^5 = 32$ ซึ่งตรงกับแบบฝึกหัด 5.3.2

สำหรับรถผีด้วยสามจุด $b - a = 6$ และ $\max|f^{(6)}| \le \frac{5}{64}$ ได้ขอบเขต 0.0108 เมตร ซึ่งกว้างกว่าค่าคลาดเคลื่อนสัมบูรณ์จริง 0.00072 เมตรราว 15 เท่า ซึ่งเป็นเรื่องปกติของขอบเขตกรณีเลวร้ายที่สุด

ฉบับร่าง · preprint · 2026-10-23

ตารางที่ 5.7: ค่าคลาดเคลื่อนของระยะเมื่อความเร็วมีมุมที่ $t = 1.3$ บน $[0, 2]$ เมื่อใช้จำนวนตัวอย่างเท่ากัน (ตัวเลขจากโค้ดของเรา)

จำนวนตัวอย่าง	เกาส์ (เมตร)	ซิมป์สัน (เมตร)
3	-3.733×10^{-2}	2.333×10^{-2}
5	-2.552×10^{-2}	-1.000×10^{-2}
9	9.238×10^{-3}	6.667×10^{-3}
17	-1.366×10^{-3}	-1.667×10^{-3}
33	-1.533×10^{-4}	4.167×10^{-4}

ลองเอง: ขอบเขตของสูตรสี่จุด

ใช้ทฤษฎีบท 5.3.3 กับ $N = 4$ บนช่วง $[0, 6]$ โดยคำนวณ $C_4 = \frac{(4!)^4}{9\,(8!)^3}$ และใช้ $\max |f^{(8)}| \le \frac{5}{4^4}$ แล้วเทียบกับค่าคลาดเคลื่อนสัมบูรณ์จริงของสูตรสี่จุด ซึ่งอยู่ในตารางค่าคลาดเคลื่อนของหัวข้อนี้ ขอบเขตของคุณกว้างกว่าค่าคลาดเคลื่อนสัมบูรณ์จริงกี่เท่า

ขอบเขตของสูตรสี่จุดคือ 1.1×10^{-4} เมตร ซึ่งกว้างกว่าค่าคลาดเคลื่อนสัมบูรณ์จริง 7.4×10^{-6} เมตรราว 15 เท่า เหมือนกรณีสามจุด

เกร็ดเพิ่มเติม

ถ้าไม่อยากคำนวณขอบเขต เราประมาณค่าคลาดเคลื่อนของสูตรเกาส์ด้วยสูตรที่แม่นกว่าได้ ตามทางปฏิบัติที่ตำราอ้างถึงคือ $I - I_n \approx I_m - I_n$ สำหรับ $m > n$ Atkinson 1989, §5.3, (5.3.44), p. 282 คู่สูตรที่ Kronrod ศึกษาเพิ่มจุดใหม่ $m = n + 1$ จุดเข้าไปในสูตรเกาส์ n จุด ได้สูตรที่แม่นถึงดีกรี $3n + 1$ และให้ตัวประมาณค่าคลาดเคลื่อนที่ถูกกว่าการใช้สูตรเกาส์ที่มี $2n + 1$ จุด ตามที่ตำรารายงาน Atkinson 1989, §5.3, pp. 283--284 ส่วนลำดับ Patterson ใช้จำนวนจุด $3, 7, 15, \ldots$ และระดับความแม่นยำ $5, 11, 23, \ldots$ โดยทุกจุดอยู่ในช่วงและน้ำหนักเป็นบวกตามที่ตำรารายงาน เราไม่ได้ทำซ้ำตัวอย่างของตำรา และการเลือกจำนวนจุดอย่างปรับตัวได้นี้เป็นแนวทางของหัวข้อ 5.4

เมื่อความเร็วมีมุม

ทฤษฎีบทต้องการให้ความเร็วเรียบ เพื่อดูว่าเรื่องนี้สำคัญแค่ไหน เราใช้ความเร็วที่สร้างขึ้นแบบในหัวข้อ 5.2 คือ $20 + |t - 1.3|$ บน $[0, 2]$ ซึ่งมีมุมที่ $t = 1.3$ ค่าจริงของปริพันธ์คือ $40 + \frac{1.3^2 + 0.7^2}{2} = 41.09$ เมตร ตารางข้างล่างเทียบสูตรเกาส์กับซิมป์สันที่จำนวนตัวอย่างเท่ากัน

ที่สามถึงเก้าตัวอย่าง ซิมป์สันแม่นกว่า ที่สิบเจ็ดตัวอย่างเกาส์เริ่มแม่นกว่าเล็กน้อย และที่สามสิบสามตัวอย่างค่าคลาดเคลื่อนของเกาส์เหลือ 1.5×10^{-4} เมตร ขณะที่ซิมป์สันยังเหลือ 4.2×10^{-4} เมตร ตรงกันข้ามกับรถผีที่ความเร็วเรียบ ซึ่งสูตรเกาส์เจ็ดจุดให้ค่าคลาดเคลื่อนสัมบูรณ์น้อยกว่า 10^{-12} เมตร ความแตกต่างนี้มาจากความเรียบของความเร็วที่ต่างกัน

ข้อควรระวัง: ตรวจความเรียบก่อนเลือกสูตร

อย่าเลือกสูตรเกาส์เพียงเพราะมันมีอันดับสูง ก่อนใช้ ให้ตรวจว่า $f^{(2N)}$ ต่อเนื่องบนช่วงปิด และให้ดูตัวอย่างมุมหักข้างบน อัตราการลู่เข้าของสูตรเกาส์ขึ้นกับความเรียบของ f Atkinson 1989, §5.3, p. 278

ฉบับร่าง · Preprint · 2026-10-23

สรุปใจความ

- ถ้าเลือกทั้งจุดและน้ำหนักได้ สูตร N จุดแม่นกับพหุนามดีกรี $2N-1$ และไม่มีสูตร N จุดใดทำได้ดีกว่า นี่คือคำตอบของคำถามว่าเลือกดีที่สุดแล้วหรือไม่
- โหนดของสูตรเกาส์คือรากของพหุนามเชิงตั้งฉาก P_N น้ำหนักเป็นบวกทุกตัว และคำนวณได้จากค่าลักษณะเฉพาะของเมทริกซ์ที่เป็นสามแนวเฉียง
- รถผีที่ถามความเร็วเพียงสามครั้ง สูตรเกาส์ให้ค่าคลาดเคลื่อน 0.00072 เมตร ซึ่งเล็กกว่ากฎเชิงสี่เหลี่ยมคางหมูราว 380 เท่า และเล็กกว่าซิมป์สันราว 75 เท่า
- กล้องที่วิ่งด้วยความเร็วคงที่ต้องการความยาวตามเส้นโค้ง ซึ่งเป็นปริพันธ์ที่ไม่มีสูตรปิด และการหาตำแหน่งจากระยะซึ่งเป็นการหารากสมการ ตัวแปรของเส้นโค้งไม่ใช่ระยะ จึงต้องแปลงทุกครั้ง

อ่านต่อ. หัวข้อนี้มาจาก Atkinson 1989, §5.3, pp. 270–284 และ Iserles 2009, §3.1, pp. 33–37 ส่วนการปรับจำนวนช่องหรือจุดให้ได้ค่าคลาดเคลื่อนที่ต้องการโดยไม่ต้องเดาเอง อยู่ในหัวข้อ 5.4 และวิธีแบ่งครึ่งกับการหารากสมการทั่วไปอยู่ในบทที่ 6

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 5.3.1 (⋆) ให้ $f(x) = x^3$ บน $[1, 3]$ จงใช้สูตรเกาส์สองจุดหาปริพันธ์ของ f โดยหาโหนดและน้ำหนักจากช่วง $[-1, 1]$ ก่อน แล้วตรวจว่าค่าที่ได้ตรงกับค่าจริงหรือไม่

แบบฝึกหัด 5.3.2 (⋆) ให้ $f(x) = x^4$ บน $[-1, 1]$ จงหาค่าคลาดเคลื่อนของสูตรเกาส์สองจุด แล้วเทียบกับทฤษฎีบทค่าคลาดเคลื่อน โดยใช้ว่า $f^{(4)}$ เป็นค่าคงที่

แบบฝึกหัด 5.3.3 (⋆⋆) จงตรวจว่า $p(x) = x^2 - \tfrac{1}{3}$ ตั้งฉากกับ 1 และ x บน $[-1, 1]$ แล้วใช้ข้อนี้หาโหนดของสูตรเกาส์สองจุดโดยไม่ใช้โค้ดหรือตารางของตำรา

แบบฝึกหัด 5.3.4 (⋆⋆) สูตรหนึ่งจุด $Q(f) = b\, f(c)$ บน $[-1, 1]$ จงหาค่า b และ c ที่ทำให้สูตรแม่นกับ 1 และ x แล้วแสดงว่าสูตรนี้ไม่แม่นกับ x^2 ดังนั้นระดับความแม่นยำของสูตรนี้คือ 1

แบบฝึกหัด 5.3.5 (⋆⋆, เมื่อสูตรเกาส์ไม่เป๊ะ) เส้นทางกล้องเส้นตรงบนแกน x มีจุดควบคุม $P_0 = (0, 0)$, $P_1 = (0.1, 0)$, $P_2 = (2.9, 0)$ และ $P_3 = (3, 0)$ (หน่วยเป็นเมตร) ให้ $B(s)$ เป็นเส้นโค้งเบซิเยร์ลูกบาศก์ของจุดควบคุมนี้ เมื่อ $s \in [0, 1]$

1. จงแสดงว่า $|B'(s)| = 0.3 + 16.2\, s(1 - s)$ บน $[0, 1]$ และเป็นบวกทุก s (คำใบ้: B' มีเฉพาะส่วนแนวนอน)
2. จงหาความยาวของเส้นทางด้วยสูตรเกาส์สองจุดบน $[0, 1]$ และตรวจว่าได้ 3 เมตรพอดี
3. เส้นทางของหัวข้อ 5.3 มี $y'(s) \neq 0$ จงอธิบายว่าทำไม $|B'(s)| = \sqrt{x'(s)^2 + y'(s)^2}$ โดยทั่วไปจึงไม่เป็นพหุนาม และจงตรวจด้วยโค้ดว่าสูตรเกาส์สองจุดให้ค่าคลาดเคลื่อนไม่เป็นศูนย์ (คำใบ้: ดูรากที่สองของผลรวม)

แบบฝึกหัด 5.3.6 (⋆ ⋆ ⋆, คำนวณและอธิบาย) ใช้ระยะรถผีบน $[0, 6]$ จากหัวข้อ 5.2 และ 5.3 จงหาจำนวนตัวอย่างความเร็วที่น้อยที่สุดที่ทำให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 0.01 เมตร สำหรับแต่ละวิธีต่อไปนี้: สูตรเกาส์, สูตรซิมป์สันแบบประกอบ และสูตรเชิงสี่เหลี่ยมคางหมูแบบประกอบ (นับจำนวนตัวอย่างจากจำนวนจุดที่ใช้ ไม่ใช่จำนวนช่วง) แล้วอธิบายว่าทำไมสามวิธีต่างกันมาก

แบบฝึกหัด 5.3.7 (⋆ ⋆ ⋆, กลับจากระยะไปหาตำแหน่ง) ให้ $\ell(s) = \int_0^s |B'(\sigma)|\, d\sigma$ เป็นความยาวของเส้นทางกล้องจากจุดเริ่มถึงพารามิเตอร์ s (คำใบ้: ทฤษฎีบทมูลฐานของแคลคูลัสและทฤษฎีบทค่ากลาง)

1. จงแสดงว่า ℓ เพิ่มขึ้นอย่างเคร่งครัดบน $[0, 1]$ เมื่อ $|B'(s)| > 0$ ทุก s

ฉบับร่าง · Preprint · 2026-10-23

2. จงอธิบายว่าทำไมระยะ d แต่ละค่าในช่วง $[0, \ell(1)]$ จึงมีพารามิเตอร์ s เพียงค่าเดียว

3. สมมติว่าเราใช้สูตรเกาส์สามจุดหาความยาวทั้งเส้นทางแทน $\ell(1)$ และแบ่งเส้นทางที่ระยะเท่า ๆ กันตามค่าที่ได้ จุดสุดท้ายจะคลาดจากปลายทางจริงกี่เมตร (ใช้ค่าในหัวข้อ 5.3)

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 5.3.1.

1. บน $[-1, 1]$ สูตรเกาส์สองจุดมีโหนด $\pm\frac{1}{\sqrt{3}}$ และน้ำหนัก 1 ทั้งคู่

2. เปลี่ยนช่วงเป็น $[1, 3]$ ด้วย $x = 2 + \xi$ โดยจุดกึ่งกลางคือ 2 และครึ่งความกว้างคือ 1 จึงได้โหนด $2 \pm \frac{1}{\sqrt{3}}$ และน้ำหนัก $1 \cdot 1 = 1$ ทั้งคู่

3. ให้ $u = \frac{1}{\sqrt{3}}$ จะได้ $Q(f) = (2-u)^3 + (2+u)^3$ ใช้ $(2 \pm u)^3 = 8 \pm 12u + 6u^2 \pm u^3$ แล้วบวกกัน จะได้ $Q(f) = 16 + 12u^2 = 16 + 4 = 20$

4. ค่าจริงคือ $\int_1^3 x^3\,dx = \frac{3^4 - 1^4}{4} = 20$ ตรงกันพอดี

ตรวจ: x^3 มีดีกรี $3 = 2 \cdot 2 - 1$ จึงอยู่ในขอบเขตที่ทฤษฎีบท 5.3.2 ข้อ 1 รับรองว่าเป๊ะ (ทดสอบในไฟล์ ชื่อ test_ex_5_3_1)

เฉลย 5.3.2.

1. สูตรเกาส์สองจุดบน $[-1, 1]$ ให้ $Q(x^4) = \left(\frac{1}{\sqrt{3}}\right)^4 + \left(-\frac{1}{\sqrt{3}}\right)^4 = \frac{1}{9} + \frac{1}{9} = \frac{2}{9}$

2. ค่าจริงคือ $I = \int_{-1}^{1} x^4\,dx = \frac{2}{5}$

3. ค่าคลาดเคลื่อนคือ $I - Q = \frac{2}{5} - \frac{2}{9} = \frac{18-10}{45} = \frac{8}{45}$

4. ทฤษฎีบทค่าคลาดเคลื่อนให้ $I - Q = C_2 (b-a)^5 f^{(4)}(\xi)$ โดย $C_2 = \frac{(2!)^4}{5\,(4!)^3} = \frac{16}{5 \cdot 13824} = \frac{1}{4320}$ และ $(b-a)^5 = 2^5 = 32$ เมื่อ $f^{(4)}(x) = 24$ ทุก x จะได้ $\frac{32 \cdot 24}{4320} = \frac{768}{4320} = \frac{8}{45}$ ตรงกับข้อ 3 พอดี เพราะ $f^{(4)}$ เป็นค่าคงที่ ค่าในทฤษฎีบทจึงไม่ใช่แค่ขอบเขต แต่เป็นค่าจริง

เฉลย 5.3.3.

1. $\int_{-1}^{1} \left(x^2 - \frac{1}{3}\right) dx = \frac{2}{3} - \frac{1}{3} \cdot 2 = 0$

2. $\int_{-1}^{1} \left(x^2 - \frac{1}{3}\right) x\,dx = \int_{-1}^{1} x^3\,dx - \frac{1}{3}\int_{-1}^{1} x\,dx = 0 - 0 = 0$ เพราะเป็นฟังก์ชันคี่บนช่วงที่สมมาตรรอบศูนย์

3. พหุนามทุกตัวที่มีดีกรีไม่เกินหนึ่งเป็นผลรวมของ 1 และ x ดังนั้น p ตั้งฉากกับทุกพหุนามดีกรีไม่เกินหนึ่ง และ p มีดีกรี 2 กับสัมประสิทธิ์ของ x^2 เท่ากับ 1 จึงเป็นพหุนามเชิงตั้งฉากดีกรีสองแบบมอนิก ซึ่งมีได้เพียงชุดเดียว (ถ้ามีสองชุด ผลต่างของทั้งสองจะตั้งฉากกับตัวเองจนเป็นศูนย์)

4. รากของ p คือ $x^2 = \frac{1}{3}$ จึง $x = \pm\frac{1}{\sqrt{3}}$ ตรงกับโหนดของสูตรเกาส์สองจุดในแบบฝึกหัด 5.3.2 และในตารางของตำราพอดี

ตรวจ: ทดสอบในไฟล์ ชื่อ test_ex_5_3_3 (ผลต่างของปริพันธ์ทั้งสองเป็นศูนย์ และโหนดยกกำลังสองเท่ากับ $\frac{1}{3}$)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 5.3.4.

1. ให้ $f = 1$: $Q(1) = b$ และ $I(1) = 2$ ดังนั้น $b = 2$
2. ให้ $f = x$: $Q(x) = bc$ และ $I(x) = 0$ เมื่อ $b = 2$ จะได้ $2c = 0$ จึง $c = 0$
3. ตรวจ x^2 ด้วย $b = 2$ และ $c = 0$: $Q(x^2) = 2 \cdot 0^2 = 0$ แต่ $I(x^2) = \frac{2}{3}$ ไม่เท่ากัน
4. ดังนั้นสูตรนี้แม่นกับ 1 และ x แต่ไม่แม่นกับ x^2 ระดับความแม่นยำจึงเป็น 1 ซึ่งเท่ากับ $2 \cdot 1 - 1$ ตรงกับทฤษฎีบท 5.3.2 ข้อ 2 ว่าสูตรที่มี N จุดไม่สามารถแม่นเกิน $2N - 1$ ได้

สูตรนี้คือจุดกึ่งกลาง $Q(f) = 2f(0)$ ซึ่งเป็นสูตรหนึ่งจุดที่ดีที่สุด (ทดสอบในไฟล์ ชื่อ test_ex_5_3_4)

เฉลย 5.3.5.

1. จากสูตรของ B' ใช้ $P_1 - P_0 = (0.1, 0)$, $P_2 - P_1 = (2.8, 0)$ และ $P_3 - P_2 = (0.1, 0)$ จะได้

$$x'(s) = 3(1-s)^2(0.1) + 6(1-s)s(2.8) + 3s^2(0.1) = 0.3(1-s)^2 + 16.8\,s(1-s) + 0.3s^2.$$

   เพราะ $(1-s)^2 + s^2 = 1 - 2s(1-s)$ จึงได้ $0.3\big(1 - 2s(1-s)\big) + 16.8s(1-s) = 0.3 + 16.2\,s(1-s)$ และ $y'(s) = 0$ ดังนั้น $|B'(s)| = x'(s)$ ซึ่งเป็นบวกทุก $s \in [0,1]$ เพราะ $s(1-s) \geq 0$

2. x' เป็นพหุนามดีกรีสอง ดังนั้นสูตรเกาส์สองจุดแม่นถึงดีกรี 3 ตรวจโดยตรง: โหนดของ $[0,1]$ คือ $\frac{1}{2} \pm \frac{1}{2\sqrt{3}}$ และน้ำหนักแต่ละตัวคือ $\frac{1}{2}$ ที่โหนดทั้งสอง $s(1-s) = \frac{1}{4} - \frac{1}{12} = \frac{1}{6}$ จึง $x' = 0.3 + 16.2 \cdot \frac{1}{6} = 3$ ดังนั้น $Q = \frac{1}{2}(3+3) = 3$ เมตร ตรงกับ $\ell(1) = \int_0^1 x'(s)\,ds = 3$ พอดี

3. ถ้า $y' \neq 0$ ผลรวม $(x')^2 + (y')^2$ เป็นพหุนามดีกรีสี่ และรากที่สองของพหุนามนี้โดยทั่วไปไม่เป็นพหุนาม ในเส้นทางของหัวข้อ 5.3 สูตรเกาส์สองจุดให้ค่าคลาดเคลื่อนราว 0.0218 เมตร ซึ่งไม่เป็นศูนย์ (ตรวจด้วยโค้ดแล้ว) เพิ่มจุดทำให้ค่าคลาดเคลื่อนลดลงเร็ว แต่ไม่ได้ทำให้เป็นศูนย์

ทดสอบในไฟล์ ชื่อ test_ex_5_3_5 (ทั้งสองข้อ 2 และ 3 มีในไฟล์เดียวกัน)

เฉลย 5.3.6.

1. ค่าจริงของระยะคือ $120 + 10\sin 3 \approx 121.41120$ เมตร
2. สูตรเกาส์ N จุด: $N = 1$ ให้ค่าคลาดเคลื่อน -0.71092 เมตร, $N = 2$ ให้ 0.03637 เมตร, $N = 3$ ให้ -0.000716 เมตร เป็นครั้งแรกที่ไม่เกิน 0.01 ดังนั้นใช้ 3 ตัวอย่าง
3. สูตรซิมป์สันแบบประกอบ: $n = 2$ (3 ตัวอย่าง) ให้ -0.05358 เมตร, $n = 4$ (5 ตัวอย่าง) ให้ -0.00266 เมตร ดังนั้นใช้ 5 ตัวอย่าง
4. สูตรเชิงสี่เหลี่ยมคางหมูแบบประกอบ: $n = 10$ (11 ตัวอย่าง) ให้ 0.01060 เมตร ยังเกิน ส่วน $n = 11$ (12 ตัวอย่าง) ให้ 0.008758 เมตร ดังนั้นใช้ 12 ตัวอย่าง
5. ทำไมต่างกัน: ค่าคลาดเคลื่อนของสูตรเกาส์ตามทฤษฎีบทเป็นสัดส่วนกับ $\max|f^{(2N)}|$ ซึ่งเล็กลงเร็วเมื่อ N เพิ่ม (สำหรับรถผีคือ $5/4^N$) ส่วนสูตรแบบประกอบค่าคลาดเคลื่อนลดตามกำลังสองหรือสี่ของขนาดช่วงเท่านั้น จึงเสียตัวอย่างมากกว่ามาก อีกข้อหนึ่งคือโหนดของสูตรเกาส์ไม่ซ้อนกันจึงนำตัวอย่างเดิมมาใช้ซ้ำแทบไม่ได้

ทดสอบในไฟล์ ชื่อ test_ex_5_3_6 (นับเฉพาะจำนวนตัวอย่างที่น้อยที่สุดสำหรับทั้งสามวิธี และค่าคลาดเคลื่อนของสูตรเกาส์ตามที่ปรากฏในข้อ 2)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 5.3.7.

1. เพราะ $|B'|$ ต่อเนื่อง ทฤษฎีบทมูลฐานของแคลคูลัสให้ $\ell'(s) = |B'(s)|$ ถ้า $0 \le s_1 < s_2 \le 1$ จะได้ $\ell(s_2) - \ell(s_1) = \int_{s_1}^{s_2} |B'(\sigma)|\,d\sigma > 0$ เพราะปริพันธ์ของฟังก์ชันบวกบนช่วงที่มีความยาวบวกเป็นบวก
2. ฟังก์ชัน ℓ ต่อเนื่องบน $[0,1]$ โดยมี $\ell(0) = 0$ ทฤษฎีบทค่ากลางให้มี s อย่างน้อยหนึ่งค่าที่ $\ell(s) = d$ สำหรับแต่ละ $d \in [0, \ell(1)]$ และข้อ 1 บอกว่าไม่มีค่า s อื่นอีก เพราะ ℓ เพิ่มอย่างเคร่งครัดจึงไม่ผ่านค่า d เดิมสองครั้ง
3. สูตรเกาส์สามจุดให้ $\tilde{L} = 4.49641$ เมตร และค่าจริง $L = 4.50230$ เมตร ค่าคลาดเคลื่อนคือ $L - \tilde{L} = 0.00589$ เมตร (5.9 มิลลิเมตร) ถ้าแบ่งระยะที่ $\frac{k}{4}\tilde{L}$ กล้องที่ตำแหน่ง $k/4$ จะผิดไป $\frac{k}{4} \cdot 0.00589$ เมตร และจุดสุดท้าย ($k = 4$) จะคลาดจากปลายทางจริงอยู่ 0.00589 เมตร

ทดสอบในไฟล์ ชื่อ test_ex_5_3_7 (ค่าผกผันตรงตามระยะที่ต้องการ และค่าคลาดเคลื่อนที่ปลายคือ 0.00589 เมตร)

5.4 รู้ได้อย่างไรว่าแม่นพอ

สถานการณ์สมมติ: รู้ได้อย่างไรว่าคำตอบพอแล้ว

ในหัวข้อ 5.2 เราเห็นว่าขอบเขตที่รับประกันต้องใช้ 48 ช่วงกับกฎเชิงสี่เหลี่ยมคางหมู และ 8 ช่วงกับกฎของซิมป์สันเพื่อให้ผิดไม่เกินหนึ่งเซนติเมตร ทั้งที่ค่าจริงใช้เพียง 11 และ 4 ช่วง ขอบเขตนั้นต้องรู้อนุพันธ์อันดับสูงของความเร็ว ซึ่งระบบของเรารู้ล่วงหน้าไม่ได้

สมมติว่าถนนของรถผีมีจุดหนึ่งที่รถกระชากความเร็วขึ้นราว 6 เมตรต่อวินาที เกิดที่เวลาประมาณ 4.2 วินาที และกระชากนั้นกินเวลาเพียงราวหนึ่งในสิบวินาที (สร้างขึ้นเพื่อการเรียน) ระบบถามความเร็วได้ทีละจุด และต้องตอบว่าระยะถูกภายในหนึ่งเซนติเมตรหรือยัง โดยไม่รู้สูตรของความเร็ว

คำถามของหัวข้อนี้คือ ระบบจะรู้ได้อย่างไรว่าผลที่ได้แม่นพอ และควรถามความเร็วที่จุดไหนต่อไป

ภาพรวมก่อนอ่าน

ทุกวิธีในหัวข้อนี้ใช้ข้อเท็จจริงข้อเดียว คือถ้าคำนวณด้วยขนาดขั้นสองขนาด ผลต่างของสองค่านั้นบอกได้โดยประมาณว่าค่าที่ละเอียดกว่ายังผิดอยู่เท่าไร และเราไม่ต้องรู้คำตอบจริง ความคิดนี้มีสามรูปแบบ

- **การประมาณค่านอกช่วงแบบริชาร์ดสัน** (Richardson extrapolation) คือการรวมผลสองระดับด้วยผลต่างของมัน เพื่อหักค่าคลาดเคลื่อนส่วนนำออกจากค่าที่ละเอียดกว่า
- **โรมเบิร์ก** (Romberg integration) คือการทำการรวมแบบนี้ซ้ำเป็นตาราง โดยใช้จุดเดิมซ้ำ
- **ควอดราเจอร์แบบอะแดปทีฟ** (adaptive quadrature) คือการใช้ผลต่างนี้ตัดสินว่าช่วงไหนต้องแบ่งละเอียดขึ้น ส่วนช่วงที่เรียบเราปล่อยไว้

ใช้จุดเดิมซ้ำได้

ก่อนจะเทียบสองระดับ เราต้องคำนวณระดับที่ละเอียดกว่าโดยไม่ทิ้งงานเดิม กฎเชิงสี่เหลี่ยมคางหมูทำได้พอดี ถ้า T_n คือกฎที่แบ่งช่วงเป็น n ช่วงเท่า ๆ กันด้วยขนาดขั้น $h = \frac{b-a}{n}$ เมื่อแบ่งละเอียดเป็นสองเท่า จุดเดิมทั้งหมดยังอยู่ เราเพิ่มเฉพาะจุดกลางของทุกช่วง

$$T_{2n} = \tfrac{1}{2}T_n + \tfrac{h}{2}\sum_{j=1}^{n} f\Big(a + (2j-1)\tfrac{h}{2}\Big).$$

ฉบับร่าง · Preprint · 2026-10-23

ครึ่งหนึ่งของ T_n เดิมถูกใช้ต่อ และผลรวมมีเฉพาะจุดใหม่ ต่างจากหัวข้อ 5.3 ที่สูตรเกาส์สามจุดกับห้าจุดมีจุดร่วมกันเพียงจุดกลาง

ลองเอง: x^2 บน $[0,2]$

ที่ $n=1$ ได้ $T_1=\frac{2}{2}(0+4)=4$ ขนาดขั้นของ T_2 คือ $\frac{h}{2}=1$ และจุดใหม่คือ $x=1$ ตรวจว่า $T_2=\frac{1}{2}\cdot 4+1\cdot 1=3$ ตรงกับค่าในหัวข้อ 5.2 โดยไม่ต้องหาค่า f ที่ 0 กับ 2 ซ้ำ

สองคำตอบบอกค่าคลาดเคลื่อนได้

ค่าคลาดเคลื่อนของกฎเชิงสี่เหลี่ยมคางหมูไม่ได้มีเพียงพจน์ h^2 จากสูตรออยเลอร์–แมคลอริน ที่เห็นในหัวข้อ 5.2 ค่าคลาดเคลื่อนเป็นผลรวมของกำลังคู่ของ h และสัมประสิทธิ์ขึ้นกับอนุพันธ์ที่ปลายช่วงเท่านั้น ถ้า f เรียบพอ เรามีทฤษฎีบทนี้

ทฤษฎีบท 5.4.1: รูปของค่าคลาดเคลื่อนของกฎเชิงสี่เหลี่ยมคางหมู

ให้ $f\in C^{2m+2}[a,b]$ และ $n\ge 1$ จะมีค่าคงที่ $c_1,\dots,c_m$ ซึ่งไม่ขึ้นกับ n และค่าคงที่ C ที่ทำให้

$$T_n - I = c_1 h^2 + c_2 h^4 + \cdots + c_m h^{2m} + F_n, \qquad |F_n| \le C h^{2m+2}, \qquad h=\frac{b-a}{n}.$$

ค่า c_j ขึ้นกับ $f^{(2j-1)}$ ที่จุดปลาย a และ b เท่านั้น

หมายเหตุ. ทฤษฎีบทนี้คือทฤษฎีบทออยเลอร์–แมคลอริน เราไม่พิสูจน์ซ้ำในบทนี้ ตำราพิสูจน์ไว้ในส่วนของสูตรออยเลอร์–แมคลอริน Atkinson 1989, §5.4, Thm 5.5, p. 285 และใช้ผลนี้กับการรวมแบบ Richardson ที่หน้า 294--295 ของตำรา Atkinson 1989, §5.4, pp. 294--295

อ่านแบบเข้าใจง่าย

- c_1h^2 คือพจน์ที่ใหญ่ที่สุดของค่าคลาดเคลื่อน ถ้าขนาดขั้นลดลงครึ่งหนึ่ง พจน์นี้ลดลงเป็นหนึ่งในสี่
- ถ้า f เรียบพอ พจน์ที่เหลือจะเล็กมากเมื่อ h เล็ก ทำให้เราหักพจน์ h^2 ออกได้

ทฤษฎีบทต่อไปนี้คือเหตุผลของวิธีริชาร์ดสัน เราเขียน T_n และ T_{2n} ด้วยทฤษฎีบทนี้ แล้วรวมกันให้พจน์ h^2 หักล้างกัน

ทฤษฎีบท 5.4.2: การรวมสองระดับที่หักพจน์ h^2 ออก

ให้ $f\in C^6[a,b]$ และ $T_n - I = c_1h^2 + c_2h^4 + O(h^6)$ โดย $h=\frac{b-a}{n}$ จะได้

$$R_n = \frac{4T_{2n}-T_n}{3} = I - \frac{c_2}{4}h^4 + O(h^6).$$

ดังนั้นพจน์ h^2 หายไป และ R_n มีค่าคลาดเคลื่อนเป็น h^4

พิสูจน์. **แนวคิด.** เขียน T_{2n} ด้วยขนาดขั้น $h/2$ แล้วคูณและลบให้พจน์ h^2 หักล้างกัน

1. เพราะ $(h/2)^2=h^2/4$ และ $(h/2)^4=h^4/16$ จึงได้ $T_{2n}-I=\frac{c_1}{4}h^2+\frac{c_2}{16}h^4+O(h^6)$
2. คูณด้วย 4 จะได้ $4T_{2n}-4I=c_1h^2+\frac{c_2}{4}h^4+O(h^6)$
3. ลบ $T_n-I=c_1h^2+c_2h^4+O(h^6)$ จะได้ $4T_{2n}-T_n-3I=-\frac{3}{4}c_2h^4+O(h^6)$ เพราะพจน์ c_1h^2 ตัดกันพอดี

ร่าง · Preprint · 2026-10-23

4. หารด้วย 3 จะได้ $R_n - I = -\frac{c_2}{4}h^4 + O(h^6)$

สรุป: การรวมด้วยน้ำหนัก 4 และ -1 แล้วหารด้วย 3 ทำให้พจน์ h^2 หายไปและเหลือ h^4 □

ทฤษฎีบทนี้มีอีกความหมายหนึ่ง $R_n = \frac{4T_{2n}-T_n}{3}$ คือกฎของซิมป์สันแบบประกอบ S_{2n} จากหัวข้อ 5.2 นั่นเอง กฎของซิมป์สันจึงเป็นการรวมแบบนี้ที่ใช้จุดชุดเดียวกันกับ T_n และ T_{2n}

ส่วนที่ใช้ประมาณค่าคลาดเคลื่อนคือผลต่าง จากการเขียน $T_n - I \approx c_1 h^2$ และ $T_{2n} - I \approx \frac{c_1}{4}h^2$ ได้ว่า $T_{2n} - T_n \approx -\frac{3}{4}c_1 h^2$ จึงได้

$$I - T_{2n} \approx \frac{T_{2n} - T_n}{3}.$$

ตัวประมาณนี้ผิดเพียงราว $\frac{c_2}{4}h^4$ ซึ่งเล็กกว่าค่าคลาดเคลื่อนของ T_{2n} เอง ($\approx \frac{c_1}{4}h^2$) ราว h^2 เท่า จึงใช้ประมาณค่าคลาดเคลื่อนได้ดีเมื่อ h เล็ก

ลองดูตัวเลขของรถผี $T_8 = 121.39462$ และ $T_{16} = 121.40706$ ผลต่างหารสามได้ 0.0041465 เมตร เป็นตัวประมาณของค่าคลาดเคลื่อนของ T_{16} ค่าคลาดเคลื่อนจริงคือ 0.0041368 เมตร ตัวประมาณผิดไปราว 0.2 เปอร์เซ็นต์ของค่าคลาดเคลื่อนนั้น และ $\frac{4T_{16}-T_8}{3} = 121.41121$ ตรงกับ S_{16} ในตารางหัวข้อ 5.2 (ตัวเลขทั้งหมดมาจากโค้ดของเรา)

ลองเอง: e^x บน $[0,1]$

ค่าจริงคือ $e - 1 \approx 1.718282$ คำนวณ T_1 และ T_2 ด้วยมือ แล้วคำนวณ $\frac{T_2-T_1}{3}$ เทียบกับ $I - T_2$ (คำตอบ: $T_1 \approx 1.859141$, $T_2 \approx 1.753931$, ตัวประมาณ -0.035070 และค่าคลาดเคลื่อนจริง -0.035649)

ข้อควรระวัง: เมื่อความเร็วมีมุม การรวมนี้ใช้ไม่ได้

ทฤษฎีบท 5.4.1 ต้องการให้ f เรียบพอ เมื่อความเร็วมีมุมหัก สมมติฐานนี้ไม่เป็นจริง ใช้ความเร็วหักมุมของหัวข้อ 5.2 คือ $20 + |t - 1.3|$ บน $[0, 2]$ คำนวณด้วยมือได้ $T_4 = 41.15$ และ $T_8 = 41.10$ ค่าคลาดเคลื่อนจริงของ T_4 และ T_8 คือ 0.06 และ 0.01 เมตร ตัวประมาณ $\frac{T_8-T_4}{3} = -0.0167$ จึงใหญ่กว่าค่าคลาดเคลื่อนจริงราว $\frac{5}{3}$ เท่า และที่ $n = 8$ ถึง 16 อัตราส่วนกลับเป็น $\frac{5}{9}$ ส่วนซิมป์สันที่ $n = 4$ ให้ตัวประมาณ $\frac{S_8-S_4}{15} = -0.0011$ ซึ่งมีเครื่องหมายตรงข้ามกับค่าคลาดเคลื่อนจริง $+0.0067$

สาเหตุคือ $T_n - I$ เท่ากับ $h^2\theta(1-\theta)$ พอดี เมื่อ θ คือส่วนทศนิยมของ $1.3/h$ ตำแหน่งของมุมเทียบกับตารางเปลี่ยนไปเมื่อ n เปลี่ยน จึงไม่มีพจน์ h^2 คงที่ให้ใช้ (เราตรวจสูตรนี้ด้วยโค้ดทุกค่า n ตั้งแต่ 2 ถึง 200 ความต่างไม่เกิน 10^{-12} การพิสูจน์ทั่วไปไม่อยู่ในบทนี้)

ตารางโรมเบิร์ก: ทำการรวมซ้ำ

ทำการรวมแบบเดียวกันซ้ำได้หลายชั้น โดยแต่ละชั้นใช้ผลจากชั้นก่อนหน้า ผลที่ได้เรียงเป็นตาราง

นิยาม 5.4.1: ตารางโรมเบิร์ก

ให้ $R_{0,0} = T_1$ และสำหรับ $k \geq 1$ ให้

$$R_{k,0} = T_{2^k}, \qquad R_{k,j} = R_{k,j-1} + \frac{R_{k,j-1} - R_{k-1,j-1}}{4^j - 1} \qquad (1 \leq j \leq k).$$

คอลัมน์ $j = 1$ คือกฎของซิมป์สัน S_{2^k} และคอลัมน์ $j = 2$ คือ **กฎของบูล** (Boole's rule) แบบประกอบ ค่าบนแนวทแยง $R_{k,k}$ คือผลของโรมเบิร์ก

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 5.8: ตารางโรมเบิร์กของรถผีบน $[0, 6]$ เฉพาะแนวทแยง (ตัวเลขจากโค้ดของเรา)

k	จุดตัวอย่าง	$R_{k,k}$ (เมตร)	$I - R_{k,k}$ (เมตร)
0	2	120.150113	1.2611
1	3	121.464782	-5.358×10^{-2}
2	5	121.410462	7.378×10^{-4}
3	9	121.411203	-2.611×10^{-6}
4	17	121.411200	2.322×10^{-9}

อ่านแบบเข้าใจง่าย

- ทุกช่องคำนวณจากสองช่องที่อยู่ทางซ้ายของมัน ช่องทางขวาจึงควรแม่นกว่า
- การคำนวณถึงแถว k ใช้จุดทั้งหมด $2^k + 1$ จุด และทุกช่องก่อนหน้าใช้จุดชุดนี้ซ้ำ ไม่ต้องคำนวณฟังก์ชันใหม่

แนวทแยงของตารางโรมเบิร์กลู่เข้าเร็วมาก ที่ $k = 3$ เราได้ค่าคลาดเคลื่อนเพียง 2.6×10^{-6} เมตรด้วยจุด 9 จุด ตารางนี้ให้เกณฑ์หยุดที่ตรงไปตรงมา คือหยุดเมื่อ $|R_{k,k} - R_{k-1,k-1}| \le \varepsilon$ ตำราเรียกเกณฑ์นี้ว่าง่ายและระมัดระวังมาก Atkinson 1989, §5.4, p. 298 ที่ $\varepsilon = 0.01$ เมตร ผลต่างแถวแรกที่ผ่านคือ $k = 3$ ทั้งที่ $R_{2,2}$ ซึ่งใช้ 5 จุดก็ผิดไม่ถึงหนึ่งเซนติเมตรแล้ว เกณฑ์นี้ไม่รู้ว่า $R_{2,2}$ ผิดน้อยแค่ไหน จึงต้องการจุดเพิ่ม

ปรับจุดตามความโค้ง: ควอดราเจอร์แบบอะแดปทีฟ

Romberg ใช้จุดกระจายเท่ากันทั้งช่วง ถ้าความเร็วเปลี่ยนเร็วเพียงตรงจุดเดียว เราต้องแบ่งทั้งช่วงให้ละเอียดตามจุดนั้น ซึ่งสิ้นเปลือง หัวข้อ 5.2 ให้คำใบ้ว่าควรทำอย่างไร ค่าคลาดเคลื่อนของกฎของซิมป์สันหนึ่งช่วงกว้าง ℓ มีขนาด $\frac{\ell^5}{2880}\left|f^{(4)}(\xi)\right|$ สำหรับจุด ξ บางจุด ถ้าช่วงยาวเป็นสองเท่า ขนาดค่าคลาดเคลื่อนเพิ่มเป็น 32 เท่า ดังนั้นควรแบ่งเฉพาะช่วงที่ผิดมาก และปล่อยช่วงที่เรียบไว้

นิยาม 5.4.2: ควอดราเจอร์แบบอะแดปทีฟ

ให้ f บนช่วง $[a, b]$ และ **ความคลาดเคลื่อนยินยอม** (tolerance) $\varepsilon > 0$ ทำดังนี้ สำหรับช่วง $[a, b]$ ที่กำลังพิจารณา

1. ให้ $m = \frac{a+b}{2}$ และคำนวณ $S_1 = \frac{b-a}{6}\big(f(a) + 4f(m) + f(b)\big)$
2. เพิ่มจุด $\frac{a+m}{2}$ และ $\frac{m+b}{2}$ แล้วคำนวณ S_2 เป็นผลรวมของซิมป์สันสองช่วงครึ่ง
3. ถ้า $\frac{|S_2 - S_1|}{15} \le \varepsilon$ รับช่วงนี้ ค่าที่ใช้คือ $S_2 + \frac{S_2 - S_1}{15}$
4. ถ้าไม่ผ่าน แบ่ง $[a, b]$ เป็น $[a, m]$ และ $[m, b]$ แล้วทำข้อ 1 ถึง 4 กับแต่ละช่วงด้วยความคลาดเคลื่อนยินยอม $\varepsilon/2$

อ่านแบบเข้าใจง่าย

- เลข $15 = 2^4 - 1$ มาจากค่าคลาดเคลื่อนของซิมป์สันที่เป็น h^4 ตามแนวคิดของการรวมในหัวข้อนี้ จึงเทียบ S_2 กับ S_1 แล้วหารด้วย 15
- ค่า $S_2 + \frac{S_2 - S_1}{15}$ คือ **กฎของบูล** (Boole's rule) บนช่วงนั้น ตำรา Atkinson 1989, §5.4, p. 295 เรียกการรวม

ร่างปรับปรุง · Preprint · 2026-10-23

แบบประกอบของกฎนี้ว่ากฎของบูลแบบประกอบ

- ตำรา Atkinson 1989, §5.5, p. 301 ใช้เกณฑ์ $|S_2 - S_1| \le \varepsilon$ ซึ่งเข้มกว่าโดยไม่หาร 15 และรับค่า S_2 เราใช้เกณฑ์ที่หาร 15 เพราะมันคือตัวประมาณค่าคลาดเคลื่อนจากหัวข้อก่อนโดยตรง
- ทุกช่วงที่ถูกทดสอบเพิ่มจุดใหม่สองจุด ถ้าทดสอบ N ช่วง จำนวนจุดที่ไม่ซ้ำคือ $3 + 2N$

ความคลาดเคลื่อนยินยอมถูกแบ่งครึ่งทุกครั้งที่แบ่งช่วง ทฤษฎีบทต่อไปนี้บอกว่าการแบ่งแบบนี้ทำให้ความคลาดเคลื่อนรวมเท่ากับที่ตั้งไว้พอดี ไม่มากไม่น้อย

ทฤษฎีบท 5.4.3: ความคลาดเคลื่อนยินยอมรวมกันเป็น ε พอดี

ในการแบ่งช่วงของนิยามข้างบน ให้ $\mathcal{L}$ เป็นเซตของช่วงสุดท้ายที่รับค่า τ_J เป็นความคลาดเคลื่อนยินยอมของช่วง J และ $|J|$ เป็นความยาวของช่วง จะได้

$$\sum_{J \in \mathcal{L}} \tau_J = \varepsilon \qquad \text{และ} \qquad \sum_{J \in \mathcal{L}} |J| = b - a.$$

พิสูจน์. **แนวคิด.** การแบ่งครึ่งแต่ละครั้งแบ่งทั้งความยาวและความคลาดเคลื่อนยินยอมเป็นสองฝั่งที่เท่ากัน จึงรวมกลับเป็นค่าเดิมเสมอ

1. ถ้าไม่มีการแบ่งเลย $\mathcal{L} = \{[a, b]\}$ และ $\tau = \varepsilon$ ทั้งสองข้อความเป็นจริง
2. สมมติว่าข้อความเป็นจริงทุกกรณีที่มีการแบ่งน้อยกว่า N ครั้ง และพิจารณาการแบ่งที่มี N ครั้ง การแบ่งครั้งแรกแยก $[a, b]$ เป็นสองฝั่ง $[a, m]$ และ $[m, b]$ ที่มีความคลาดเคลื่อนยินยอม $\varepsilon/2$ ทั้งคู่
3. แต่ละฝั่งถูกแบ่งน้อยกว่า N ครั้ง ตามสมมติฐานผลรวมของความคลาดเคลื่อนยินยอมที่ใบของแต่ละฝั่งเป็น $\varepsilon/2$ รวมกันได้ ε
4. ความยาวของสองฝั่งรวมกันเป็น $\frac{b-a}{2} + \frac{b-a}{2} = b - a$ จึงเป็นข้อที่สองของทฤษฎีบท

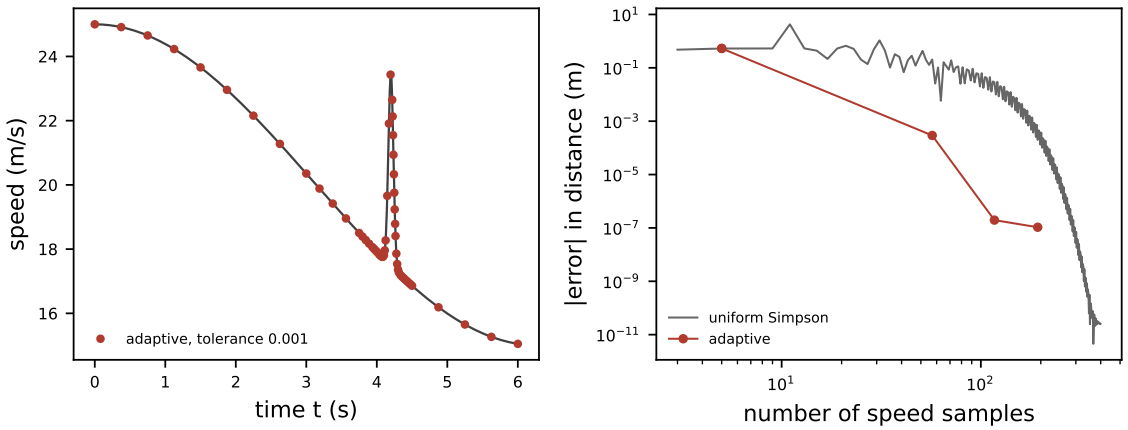
สรุป: ผลรวมของความคลาดเคลื่อนยินยอมที่ใบทั้งหมดเท่ากับ ε ทุกครั้ง ดังนั้นถ้าทุกช่วงสุดท้ายผ่านการทดสอบ ผลรวมของตัวประมาณค่าคลาดเคลื่อนทั้งหมดไม่เกิน ε □

ข้อสรุปนี้มีเงื่อนไขสำคัญ ตัวประมาณค่าคลาดเคลื่อนของแต่ละช่วงจะใกล้ค่าคลาดเคลื่อนจริงเมื่อฟังก์ชันเรียบพอในช่วงนั้น และเกณฑ์ไม่เห็นสิ่งที่จุดตัวอย่างไม่เห็น ดูข้อควรระวังข้างล่าง

ตารางข้างล่างเทียบจำนวนตัวอย่างที่ทำให้ค่าคลาดเคลื่อนสัมบูรณ์ของระยะไม่เกินเป้าหมาย ระหว่างซิมป์สันแบ่งเท่า ๆ กัน กับอะแดปทีฟที่ตั้งความคลาดเคลื่อนยินยอมเท่ากับเป้าหมาย ซิมป์สันแบ่งเท่า ๆ กันนับจากจำนวนช่วงที่เล็กที่สุดที่ผลจริงผ่านเป้า ซึ่งระบบไม่รู้ค่าจริง ดังนั้นตัวเลขนี้เป็นขอบล่างของวิธีแบ่งเท่า ๆ กัน

เมื่อกระชากกว้าง (0.25 วินาที) ซิมป์สันแบ่งเท่า ๆ กันทำได้ดีกว่า เพราะช่วงที่ห่างเท่ากันก็พอจับกระชากได้ทั้งหมด เมื่อกระชากแคบ (0.05 วินาที) อะแดปทีฟใช้จุดราวหนึ่งในสามของซิมป์สันที่เป้า 10^{-3} เมตร เพราะจุดของมันกระจุกอยู่ที่กระชาก อย่าคาดว่าอะแดปทีฟเร็วกว่าเสมอ

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 5.11: ซ้าย: ความเร็วของรถผีที่มีกระชาก $t = 4.2$ วินาที (กว้าง 0.05 วินาที) จุดแดงคือจุดที่อะแดปทีฟใช้ที่ความคลาดเคลื่อนยินยอม 10^{-3} เมตร ขวา: ค่าคลาดเคลื่อนเทียบกับจำนวนตัวอย่างบนสเกลลอการิทึม เส้นเทาคือซิมป์สันแบ่งเท่า ๆ กัน จุดแดงคืออะแดปทีฟที่ความคลาดเคลื่อนยินยอม 10^{-1} ถึง 10^{-5} (สร้างจาก `numanal/ch05_adapt.py`)

ตารางที่ 5.9: จำนวนตัวอย่างความเร็วที่ทำให้ค่าคลาดเคลื่อนสัมบูรณ์ของระยะบน $[0, 6]$ ไม่เกินเป้า เมื่อกระชากอยู่ที่ 4.2 วินาที (ตัวเลขจากโค้ดของเรา)

กว้างของกระชาก (วินาที)	เป้า (เมตร)	ซิมป์สันแบ่งเท่า ๆ กัน	อะแดปทีฟ
0.25	10^{-3}	43	61
	10^{-4}	49	101
0.10	10^{-3}	93	57
	10^{-4}	109	101
0.05	10^{-3}	169	57
	10^{-4}	209	117

ฉบับร่าง · Preprint · 2026-10-23

ข้อควรระวัง: ตัวอย่างที่ไม่เห็นกระชากจะไม่ถูกตรวจ

ที่ความคลาดเคลื่อนยินยอม $\varepsilon = 10^{-2}$ เมตร (หนึ่งเซนติเมตร) การทดสอบแรกผ่านทันที เพราะจุดห้าจุดแรก 0, 1.5, 3, 4.5 และ 6 วินาที ให้ความเร็วที่เรียบเหมือนไม่มีกระชากเลย ระบบจึงหยุดด้วยจุดเพียงห้าจุดและผิดไป 0.53 เมตร สำหรับกระชากกว้าง 0.05 วินาที และ 1.06 เมตร สำหรับกว้าง 0.10 วินาที

ความผิดนี้ไม่ได้มาจากสูตรซิมป์สัน แต่มาจากการเลือกจุด เหตุการณ์ที่สั้นกว่าระยะห่างของจุดตัวอย่างจะหายไปจากการตรวจ ซึ่งเป็นความผิดแบบเดียวกับที่บทที่ 12 จะพบอีกรูปแบบหนึ่ง ไม่มีวิธีใดรับประกันได้จากจุดตัวอย่างเพียงอย่างเดียว (ดูแบบฝึกหัด 5.4.6)

กลับไปที่หนึ่งเซนติเมตร

ตารางข้างล่างสรุปคำตอบของคำถามในสถานการณ์ ทั้งสามวิธีใช้จุดน้อยมาก และระบบได้ตัวประมาณของตัวเองจากจุดเดียวกัน

วิธี	จุดตัวอย่าง	ค่าคลาดเคลื่อนสัมบูรณ์จริง (มิลลิเมตร)	ค่าที่ระบบใช้ตัดสิน (มิลลิเมตร)
ซิมป์สัน S_4	5	-2.66	ตัวประมาณ $\frac{S_4-S_2}{15}=-3.39$
อะแดปทีฟ (กฎของบูลบนช่วงเดียว)	5	$+0.74$	$\frac{\lvert S_2-S_1\rvert}{15}=3.39\le 10$ ผ่าน
โรมเบิร์ก $R_{3,3}$	9	-0.0026	$\lvert R_{3,3}-R_{2,2}\rvert=0.74\le 10$ ผ่าน

ตารางนี้ตอบคำถามว่า ระบบรู้ได้อย่างไรว่าแม่นพอ คำตอบคือผลต่างของการคำนวณสองระดับ จากจุดที่มีอยู่แล้ว โดยไม่ต้องรู้คำตอบจริงและไม่ต้องรู้ค่าอนุพันธ์ ความเรียบของรถผีทำให้ห้าจุดพอ แต่ความเรียบนั้นเป็นสมมติฐานที่ระบบต้องตรวจเอง

เกร็ดเพิ่มเติม

ความคิดของการเทียบสองผลเพื่อประมาณค่าคลาดเคลื่อนระหว่างขั้นยังใช้กับสมการเชิงอนุพันธ์ด้วย ตำราอิเซิร์ลส์เรียกว่า **กลวิธีมิลน์** (Milne device) ซึ่งรันสูตรสองสูตรบนข้อมูลเดียวกันแล้วดูผลต่างของคำตอบเพื่อประมาณค่าคลาดเคลื่อนของขั้นนั้น Iserles 2009, §6.2, pp. 107--108 เราจะใช้ความคิดนี้เมื่อถึงการเลือกขนาดก้าวเองในบทที่ 13

สรุปใจความ

- ผลต่างของการคำนวณสองระดับ $\frac{T_{2n}-T_n}{3}$ หรือ $\frac{S_{2n}-S_n}{15}$ ประมาณค่าคลาดเคลื่อนของค่าที่ละเอียดกว่า โดยไม่ต้องรู้คำตอบจริง
- ตารางโรมเบิร์กใช้จุดซ้ำและเพิ่มความแม่นได้เร็ว แต่เกณฑ์หยุดของมันระมัดระวังมาก
- อะแดปทีฟแบ่งเฉพาะช่วงที่ผลต่างใหญ่ ความคลาดเคลื่อนยินยอมรวมกันเป็น ε พอดี แต่เห็นได้เฉพาะสิ่งที่จุดตัวอย่างเห็น
- กลับไปที่รถผี: ความเร็วที่เรียบต้องการห้าจุดเท่านั้น ส่วนกระชากแคบต้องตั้งค่าอย่างระวัง และไม่มีวิธีใดที่รับประกันได้จากจุดตัวอย่างเพียงอย่างเดียว

อ่านต่อ. ตำรา Atkinson 1989, §5.4, pp. 284--299 มีการพิสูจน์สูตรออยเลอร์–แมคลอรินและการรวมแบบริชาร์ดสัน และ Atkinson 1989, §5.5, pp. 299--305 อธิบายควอดราเจอร์อัตโนมัติและการแบ่งช่วงแบบอะแดปทีฟ ส่วน Iserles 2009, §6.2, pp. 107--108 มีกลวิธีมิลน์

ฉบับร่าง · Preprint · 2026-10-25

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 5.4.1 (⋆) ให้ $f(x)=\frac{1}{1+x}$ บน $[0,1]$ จงคำนวณ T_1 และ T_2 ด้วยมือ แล้วหาตัวประมาณ $\frac{T_2-T_1}{3}$ ของ $I-T_2$ เทียบกับค่าคลาดเคลื่อนจริง $\ln 2-T_2$ จากนั้นคำนวณ $S_2=\frac{4T_2-T_1}{3}$ และค่าคลาดเคลื่อนของ S_2

แบบฝึกหัด 5.4.2 (⋆) บน $[0,1]$ ให้ S_4 เป็นกฎของซิมป์สันที่ใช้ 4 ช่วงย่อย และ S_2 เป็นกฎของซิมป์สันที่ใช้ 2 ช่วงย่อย จงหาสัมประสิทธิ์ของ $f(0)$, $f(\frac{1}{4})$, $f(\frac{1}{2})$, $f(\frac{3}{4})$ และ $f(1)$ ในผลรวม $\frac{16S_4-S_2}{15}$ แล้วตรวจว่าตรงกับกฎของบูล $\frac{2h}{45}(7,32,12,32,7)$ เมื่อ $h=\frac{1}{4}$

แบบฝึกหัด 5.4.3 (⋆⋆) บน $[0,1]$ ให้ Q เป็นกฎของบูลที่ได้จากแบบฝึกหัด 5.4.2 จงตรวจด้วยมือว่า $Q(x^5)=\int_0^1 x^5\,dx$ แล้วหา $Q(x^6)$ และเทียบกับ $\int_0^1 x^6\,dx$ ว่าค่าคลาดเคลื่อนเป็นเท่าใด

แบบฝึกหัด 5.4.4 (⋆) ตารางโรมเบิร์กถึงแถว $k=6$ ต้องคำนวณฟังก์ชันที่กี่จุด และถ้าใช้กฎเกาส์สามจุดกับกฎเกาส์ห้าจุดบนช่วงเดียวกัน จะมีจุดที่ต่างกันทั้งหมดกี่จุด จงอธิบายว่าทำไมสองวิธีนี้ต่างกัน (ดูหัวข้อ 5.3 เรื่องจุดของสูตรเกาส์)

แบบฝึกหัด 5.4.5 (★★★, มุมหักและการรวมสองระดับ) ใช้ความเร็ว $20 + |t - 1.3|$ บน $[0, 2]$ ซึ่งเป็นความเร็วหักมุมจากหัวข้อ 5.2 และ 5.4 ค่าจริงของปริพันธ์คือ 41.09

1. จงคำนวณ T_4 และ T_8 ด้วยมือ แล้วหา $T_4 - 41.09$ และ $T_8 - 41.09$
2. จงหาตัวประมาณ $\frac{T_8 - T_4}{3}$ ของ $I - T_8$ และหาอัตราส่วนระหว่างตัวประมาณกับค่าคลาดเคลื่อนจริง
3. จงคำนวณ θ ซึ่งเป็นส่วนทศนิยมของ $1.3/h$ เมื่อ $h = \frac{2}{4}$ และ $h = \frac{2}{8}$ แล้วตรวจว่า $T_n - I = h^2 \theta (1 - \theta)$ ตรงกับข้อ 1
4. อธิบายว่าทำไมอัตราส่วนในข้อ 2 ไม่ใช่ 1 ถ้าอัตราส่วนของค่าคลาดเคลื่อนสองระดับเป็น 6 และ $\frac{8}{3}$ ตามลำดับ

ตัวช่วยคิด

ข้อ 3: ถ้า $h = \frac{1}{2}$ จะได้ $\frac{1.3}{h} = 2.6$ ดังนั้น θ คือส่วนทศนิยม 0.6 ข้อ 4: ถ้า $E_n = T_n - I$ ได้ว่า $\frac{E_n - E_{2n}}{3E_{2n}} = \frac{1}{3}\left(\frac{E_n}{E_{2n}} - 1\right)$ ซึ่งเท่ากับ 1 ก็ต่อเมื่อ $\frac{E_n}{E_{2n}} = 4$

แบบฝึกหัด 5.4.6 (★★★, โครงงานรถผี) ให้ความเร็วของรถผีบน $[0, 6]$ เป็น $20 + 5\cos(t/2)$ บวกกระชาก $6\,e^{-((t-4.2)/w)^2}$ เมตรต่อวินาที เมื่อ $w = 0.05$ วินาที จงทำดังนี้

1. ใช้ควอดราเจอร์แบบอะแดปทีฟกับช่วง $[0, 6]$ ด้วย $\varepsilon = 10^{-2}$ และ $\varepsilon = 10^{-3}$ รายงานจำนวนตัวอย่างที่ใช้และค่าคลาดเคลื่อนจริงของแต่ละกรณี
2. ตั้งช่วงเริ่มต้นสี่ช่วงเท่ากัน แล้วใช้ความคลาดเคลื่อนยินยอม $\varepsilon/4$ กับแต่ละช่วง ที่ $\varepsilon = 10^{-2}$ รายงานจำนวนจุดที่ไม่ซ้ำและค่าคลาดเคลื่อนจริง
3. ลองกระชากกว้าง $w = 0.01$ วินาทีที่เวลา $t = 2.1$ วินาที (ความสูง 6 เท่าเดิม) ตรวจว่าการเริ่มด้วยสี่ช่วงและแปดช่วงยังพลาดกระชากนี้หรือไม่
4. เขียนย่อหน้าสั้น ๆ ถึงระบบของรถผีว่าควรตั้ง ε เท่าใด ควรเริ่มด้วยกี่ช่วง และรู้ได้อย่างไรว่าไม่มีรอยกระชากที่พลาดไป

ตัวช่วยคิด

ใช้ฟังก์ชัน `adaptive_for_target` และ `composite_adaptive` ในโค้ดของบทนี้ ตรวจผลของข้อ 3 ด้วยการเปลี่ยนค่าเวลาของกระชากแล้วดูว่าช่วงที่ตั้งไว้ยังมีจุดใกล้กระชากหรือไม่ ข้อ 4 ไม่มีคำตอบเดียว ให้มีเหตุผลรองรับจากข้อ 1 ถึง 3

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 5.4.1.

1. $T_1 = \frac{1-0}{2}\big(f(0)+f(1)\big) = \frac{1}{2}\big(1+\frac{1}{2}\big) = \frac{3}{4}$
2. T_2 ใช้ขนาดขั้น $h=\frac{1}{2}$ และจุด $0, \frac{1}{2}, 1$: $T_2 = \frac{1}{2}\big[\frac{1}{2}f(0) + f(\frac{1}{2}) + \frac{1}{2}f(1)\big] = \frac{1}{2}\big[\frac{1}{2}+\frac{2}{3}+\frac{1}{4}\big] = \frac{1}{2}\cdot\frac{17}{12} = \frac{17}{24} \approx 0.708333$
3. ตัวประมาณ $\frac{T_2-T_1}{3} = \frac{\frac{17}{24}-\frac{18}{24}}{3} = -\frac{1}{72} \approx -0.013889$
4. ค่าคลาดเคลื่อนจริง $\ln 2 - T_2 \approx 0.693147 - 0.708333 = -0.015186$ ตัวประมาณอยู่ใกล้ค่าคลาดเคลื่อนจริงราว 91 เปอร์เซ็นต์
5. $S_2 = \frac{4T_2-T_1}{3} = \frac{\frac{17}{6}-\frac{3}{4}}{3} = \frac{\frac{25}{12}}{3} = \frac{25}{36} \approx 0.694444$
6. ค่าคลาดเคลื่อนของ S_2 คือ $\ln 2 - \frac{25}{36} \approx -0.001297$ ซึ่งเล็กกว่าค่าคลาดเคลื่อนของ T_2 ราว 12 เท่า

ตรวจ: ทดสอบในไฟล์ ชื่อ test_ex_5_4_1 (ค่า $\frac{3}{4}$, $\frac{17}{24}$, $-\frac{1}{72}$ และ $\frac{25}{36}$)

เฉลย 5.4.2.

1. กฎของซิมป์สันสี่ช่วงย่อยบน $[0,1]$ มี $h=\frac{1}{4}$ จึงได้ $S_4 = \frac{h}{3}\big(f_0+4f_1+2f_2+4f_3+f_4\big) = \frac{1}{12}\big(f_0+4f_1+2f_2+4f_3+f_4\big)$ เมื่อ $f_j = f(\frac{j}{4})$
2. กฎของซิมป์สันสองช่วงย่อยมี $h=\frac{1}{2}$ และจุด $0, \frac{1}{2}, 1$ จึงได้ $S_2 = \frac{1}{6}\big(f_0+4f_2+f_4\big)$
3. คูณ S_4 ด้วย 16 จะได้สัมประสิทธิ์ $\frac{16}{12}=\frac{4}{3}$ ของ $(1,4,2,4,1)$ และ S_2 มีสัมประสิทธิ์ $\frac{1}{6}$ ของ $(1,0,4,0,1)$
4. หาร $16S_4 - S_2$ ด้วย 15 ได้สัมประสิทธิ์ดังนี้ ที่ f_0: $\frac{\frac{4}{3}-\frac{1}{6}}{15} = \frac{\frac{7}{6}}{15} = \frac{7}{90}$ ที่ f_1: $\frac{\frac{16}{3}}{15} = \frac{16}{45}$ ที่ f_2: $\frac{\frac{8}{3}-\frac{2}{3}}{15} = \frac{2}{15}$ ที่ f_3: $\frac{16}{45}$ ที่ f_4: $\frac{7}{90}$
5. กฎของบูลเมื่อ $h=\frac{1}{4}$ คือ $\frac{2h}{45} = \frac{1}{90}$ คูณกับ $(7,32,12,32,7)$ จะได้ $(\frac{7}{90}, \frac{32}{90}, \frac{12}{90}, \frac{32}{90}, \frac{7}{90})$ ซึ่งเท่ากับข้อ 4 เพราะ $\frac{32}{90}=\frac{16}{45}$ และ $\frac{12}{90}=\frac{2}{15}$

ตรวจ: สัมประสิทธิ์รวมกันได้ 1 (กฎแม่นกับ 1) ทดสอบในไฟล์ ชื่อ test_ex_5_4_2 (ทั้งเศษส่วนตรงกันพอดี)

เฉลย 5.4.3.

1. กฎของบูลใช้จุด $0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$ ค่า x^5 ที่จุดเหล่านี้คือ $0, \frac{1}{1024}, \frac{1}{32}, \frac{243}{1024}, 1$ จึงได้ $Q(x^5) = \frac{1}{90}\big(0+\frac{32}{1024}+\frac{12}{32}+\frac{32\cdot 243}{1024}+7\big) = \frac{1}{90}\big(\frac{1}{32}+\frac{12}{32}+\frac{243}{32}+7\big) = \frac{1}{90}(8+7) = \frac{1}{6}$ ซึ่งเท่ากับ $\int_0^1 x^5\,dx = \frac{1}{6}$ พอดี
2. สำหรับ x^6 ค่าที่จุดคือ $0, \frac{1}{4096}, \frac{1}{64}, \frac{729}{4096}, 1$ จึงได้ $Q(x^6) = \frac{1}{90}\big(\frac{32}{4096}+\frac{12}{64}+\frac{32\cdot 729}{4096}+7\big) = \frac{1}{90}\cdot\frac{825}{64} = \frac{55}{384} \approx 0.143229$
3. ค่าจริงคือ $\frac{1}{7} \approx 0.142857$ ค่าคลาดเคลื่อนคือ $\frac{1}{7} - \frac{55}{384} = \frac{384-385}{2688} = -\frac{1}{2688} \approx -3.72\times 10^{-4}$

ตรวจ: ทดสอบในไฟล์ ชื่อ test_ex_5_4_3 (ค่า $\frac{55}{384}$ และค่าคลาดเคลื่อน $-\frac{1}{2688}$ ตรงกับการคำนวณ)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 5.4.4.

1. ตารางโรมเบิร์กถึงแถว k ใช้กฎที่มี 2^k ช่วงย่อยในคอลัมน์แรก จุดของทุกระดับอยู่บนตารางเดียวกันที่แบ่ง $[0,1]$ เป็น 2^k ช่วง จึงมี 2^k+1 จุด ดังนั้นแถว $k=6$ ใช้ $2^6+1=65$ จุด

2. กฎเกาส์สามจุดกับห้าจุดบน $[0,1]$ มีจุดกลาง $\frac{1}{2}$ ร่วมกัน ส่วนจุดอื่นต่างกัน จึงมีจุดที่ต่างกันทั้งหมด $3+5-1=7$ จุด

3. เมื่อเพิ่มจากสามจุดเป็นห้าจุด เราต้องคำนวณฟังก์ชันใหม่สี่จุด (ห้าจุดลบจุดกลาง) ไม่ใช่สองจุด เพราะโหนดของกฎเกาส์ไม่ซ้อนกัน ส่วนโรมเบิร์กซ้อนกันเพราะทุกระดับเพิ่มเพียงจุดกลางของช่วงที่มีอยู่

ตรวจ: ทดสอบในไฟล์ ชื่อ test_ex_5_4_4 (ผลนับ $3, 5, 9, 17, 33, 65$ และ 7 จุดสำหรับเกาส์)

เฉลย 5.4.5.

1. บน $[0,2]$ ความเร็วที่จุดของ T_4 (ขนาดขั้น 0.5) คือ $21.3,\ 20.8,\ 20.3,\ 20.2,\ 20.7$ จึงได้ $T_4 = 0.5\big(\frac{21.3}{2} + 20.8 + 20.3 + 20.2 + \frac{20.7}{2}\big) = 0.5 \cdot 82.3 = 41.15$ ค่าคลาดเคลื่อนคือ $41.15 - 41.09 = 0.06$

2. ที่ T_8 (ขนาดขั้น 0.25) ความเร็วที่จุดคือ $21.3, 21.05, 20.8, 20.55, 20.3, 20.05, 20.2, 20.45, 20.7$ จึงได้ $T_8 = 0.25\big(\frac{21.3+20.7}{2} + 143.4\big) = 0.25 \cdot 164.4 = 41.10$ ค่าคลาดเคลื่อนคือ $41.10 - 41.09 = 0.01$

3. ตัวประมาณ $\frac{T_8 - T_4}{3} = \frac{41.10 - 41.15}{3} \approx -0.016667$ และค่าคลาดเคลื่อนจริงของ T_8 คือ $I - T_8 = -0.01$ อัตราส่วนคือ $\frac{-0.016667}{-0.01} = \frac{5}{3}$

4. เมื่อ $h = \frac{2}{4} = \frac{1}{2}$ จะได้ $\frac{1.3}{h} = 2.6$ จึง $\theta = 0.6$ และ $h^2\theta(1-\theta) = \frac{1}{4} \cdot 0.24 = 0.06$ ตรงกับข้อ 1 เมื่อ $h = \frac{2}{8} = \frac{1}{4}$ จะได้ $\frac{1.3}{h} = 5.2$ จึง $\theta = 0.2$ และ $h^2\theta(1-\theta) = \frac{1}{16} \cdot 0.16 = 0.01$ ตรงกับข้อ 1

5. ให้ $E_n = T_n - I$ ตัวประมาณในข้อ 2 คือ $\frac{E_8 - E_4}{3}$ เมื่อเทียบกับค่าคลาดเคลื่อนจริง $-E_8$ จึงได้อัตราส่วน $\frac{E_4 - E_8}{3E_8} = \frac{1}{3}\big(\frac{E_4}{E_8} - 1\big) = \frac{1}{3}(6-1) = \frac{5}{3}$ ซึ่งไม่เท่ากับ 1 เพราะ $\frac{E_4}{E_8} = 6$ ไม่ใช่ 4 อัตราส่วนค่าคลาดเคลื่อนที่ $n = 8$ ถึง 16 คือ $\frac{8}{3}$ (ดูจาก $\theta = 0.4$ ที่ $n = 16$) ทำให้อัตราส่วนของตัวประมาณกลับเป็น $\frac{1}{3}\big(\frac{8}{3} - 1\big) = \frac{5}{9}$ เหตุผลคือ $\theta(1-\theta)$ เปลี่ยนไปตามตำแหน่งของมุมเทียบกับตาราง ไม่ใช่ h^2 คงที่

ตรวจ: ทดสอบในไฟล์ ชื่อ test_ex_5_4_5 (ค่า 41.15 และ 41.10 ตรงกับการคำนวณด้วยมือ และอัตราส่วน $\frac{5}{3}$, $\frac{5}{9}$ สลับกัน)

เฉลย 5.4.6.

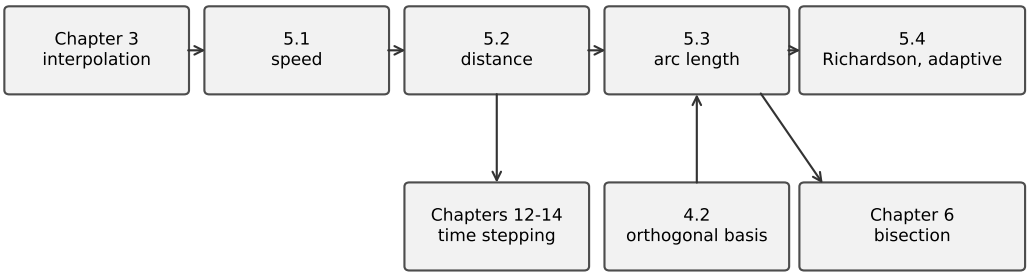
1. ที่ $\varepsilon = 10^{-2}$ การทดสอบแรกบน $[0,6]$ ได้ $\frac{|S_2 - S_1|}{15} \approx 0.0034 \le 0.01$ จึงรับทันทีด้วยจุดห้าจุด และค่าคลาดเคลื่อนจริงคือ $+0.532$ เมตร (ผิดกว่าหนึ่งเซนติเมตรมาก) ที่ $\varepsilon = 10^{-3}$ จำนวนจุดคือ 57 และค่าคลาดเคลื่อนจริงคือ -2.9×10^{-4} เมตร ซึ่งอยู่ในเป้า

2. ตั้งช่วงเริ่มต้นสี่ช่วงแล้วใช้ $\varepsilon/4 = 0.0025$ กับแต่ละช่วง จำนวนจุดที่ไม่ซ้ำคือ 45 และค่าคลาดเคลื่อนจริงคือ -4.5×10^{-4} เมตร (ผิดไม่ถึงครึ่งมิลลิเมตร) การตั้งช่วงเริ่มทำให้มีจุดใกล้กระชากมากขึ้น เช่นจุด $t = 4.125$ อยู่ห่างจากศูนย์กลางกระชากเพียง 1.5 เท่าของความกว้าง จึงเห็นกระชากได้

3. เมื่อ $w = 0.01$ และ $t = 2.1$ การเริ่มด้วยสี่ช่วงให้จุด 17 จุด และค่าคลาดเคลื่อน 0.106 เมตร การเริ่มด้วยแปดช่วงให้จุด 33 จุด และค่าคลาดเคลื่อน 0.106 เมตร ทั้งสองพลาด เพราะกระชากแคบกว่าระยะห่างของจุดทั้งสองแบบ (การเริ่มด้วยแปดช่วงจับกระชากกว้าง 0.05 ที่ $t = 2.1$ ได้ แต่กระชาก 0.01 จับไม่ได้)

4. ระบบควรเริ่มด้วยกี่ช่วงขึ้นกับความกว้างของกระชากที่คาดว่าจะเกิด เราตั้ง ε เป็นค่าคลาดเคลื่อนสัมบูรณ์ที่ยอมรับได้ของผลสุดท้าย ไม่ใช่เครื่องมือตรวจกระชากที่หายไป ไม่มีวิธีที่รับประกันได้จากจุดตัวอย่างเพียงอย่างเดียว ถ้าต้องการความมั่นใจ ต้องรู้ล่วงหน้าว่ากระชากที่แคบที่สุดกว้างเท่าใด แล้วเลือกช่วงเริ่มต้นให้ถี่กว่านั้น

สัมมนา · Preprint · 2026-10-23



รูปที่ 5.12: แผนผังบทที่ 5 ฉบับสมบูรณ์ ถึงหัวข้อ 5.4 แล้ว (สร้างจาก `numanal/ch05_adapt.py`)

ตรวจ: ทดสอบในไฟล์ ชื่อ test_ex_5_4_6 (ทั้งสองการนับจุด 45 และ 37 และค่าคลาดเคลื่อนของการเริ่มด้วยสองช่วงที่พลาดกระชาก)

สรุปท้ายบทที่ 5

ภาพต่อไปนี้คือแผนผังของบทที่ 5 ฉบับสมบูรณ์ เห็นลำดับของหัวข้อ 5.1 ถึง 5.4 และการเชื่อมกับบทที่ 3, บทที่ 4, บทที่ 6 และบทที่ 12--14

วิธี	จุดที่ใช้	ลำดับค่าคลาดเคลื่อน	เหมาะเมื่อ
กฎเชิงสี่เหลี่ยมคางหมู T_n	$n+1$	h^2	ใช้จุดซ้ำได้ เมื่อแบ่งละเอียดขึ้น
กฎของซิมป์สัน S_n	$n+1$ (n คู่)	h^4	ฟังก์ชันเรียบพอ
เกาส์ N จุด	N (เลือกเอง)	h^{2N}	เลือกจุดได้ แต่จุดไม่ซ้อนกัน
โรมเบิร์ก ระดับ k	2^k+1	h^{2k+2}	เรียบพอ ต้องการตัวประมาณค่าคลาดเคลื่อน
อะแดปทีฟ	$3+2N$	ปรับตามความเรียบเฉพาะที่	มีรอยหัก หรือกระชากที่แคบ

สูตรที่ใช้บ่อย.

- $T_n - I \approx c_1 h^2$ และ $I - T_{2n} \approx \frac{T_{2n}-T_n}{3}$ ส่วน $S_{2n} = \frac{4T_{2n}-T_n}{3}$ และ $I - S_{2n} \approx \frac{S_{2n}-S_n}{15}$
- $R_{k,j} = R_{k,j-1} + \frac{R_{k,j-1}-R_{k-1,j-1}}{4^j-1}$ และหยุดเมื่อ $|R_{k,k} - R_{k-1,k-1}| \le \varepsilon$
- อะแดปทีฟรับช่วงเมื่อ $\frac{|S_2-S_1|}{15} \le \tau$ ความคลาดเคลื่อนยินยอมรวมของทุกช่วงสุดท้ายเท่ากับ ε

อ่านต่อ. บทที่ 6 หนึ่งสมการ หนึ่งตัวไม่รู้ จะเริ่มจากวิธีล้อมคำตอบไว้ในกรงและวิธีแบ่งครึ่ง (bisection) ซึ่งเป็นวิธีที่หัวข้อ 5.3 ใช้หาตัวแปรของเส้นโค้งที่ระยะที่ต้องการ และบทนั้นจะบอกว่าทำไมการแบ่งครึ่งจึงรับประกันค่าคลาดเคลื่อน จากนั้นจะถามว่าวิธีที่เร็วกว่าอย่างวิธีของนิวตันมีเงื่อนไขอะไรบ้าง

ฉบับร่าง · Preprint · 2026-10-23

ภาค III

การแก้สมการ

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 6

หนึ่งสมการ หนึ่งตัวไม่รู้

สถานการณ์สมมติ: กระสุนกับผนังโค้ง

ลองนึกถึงฉากในเกมยิงปืนที่เราเคยเห็นบนจอ ผู้เล่นยิงกระสุนไปเป็นเส้นตรง แล้วเกมต้องรู้ว่ากระสุนไปโดนผิววัตถุตรงไหนก่อน เราเรียกจุดนั้นว่า*จุดชน* ฉากแบบนี้ต้องหาจุดชนทุกครั้งที่ยิง

ลองดูสองกรณี ทั้งสองใช้กระสุนเส้นเดียวกันที่เริ่มจากจุด $(0, 2)$ แล้ววิ่งไปทางขวาตามเส้น $y = 2 - \frac{x}{4}$ กรณีแรก วัตถุเป็นลูกบอลกลม จุดชนหาได้ด้วยสูตรสมการกำลังสองที่เราเคยใช้ตั้งแต่มัธยม กรณีที่สอง วัตถุเป็นผนังคลื่น $y = \sin x$ ที่โค้งขึ้นลงไปมา เราไม่มีสูตรที่บอกจุดชนได้ตรง ๆ ต้องหาด้วยการคำนวณ

คำถามของบทนี้คือ เมื่อไม่มีสูตร เราจะหาจุดชนได้อย่างไร และเรามั่นใจได้แค่ไหนว่าคำตอบที่ได้ใกล้จุดชนจริง (ตัวเลขทั้งหมดในบทนี้มาจากโค้งของเรา และเป็นสถานการณ์สมมติ ไม่ได้บอกอะไรเกี่ยวกับเกมใด)

บทที่ 5 ทิ้งคำถามแบบเดียวกันไว้กับกล้องที่วิ่งด้วยความเร็วคงที่: เราต้องหาตัวแปร s ที่ทำให้ความยาวตามเส้นโค้งถึงระยะที่ต้องการ ซึ่งเป็นการหารากของสมการ $\ell(s) - d = 0$ (สถานการณ์นี้ก็เป็นสถานการณ์สมมติ)

ภาพรวมบท

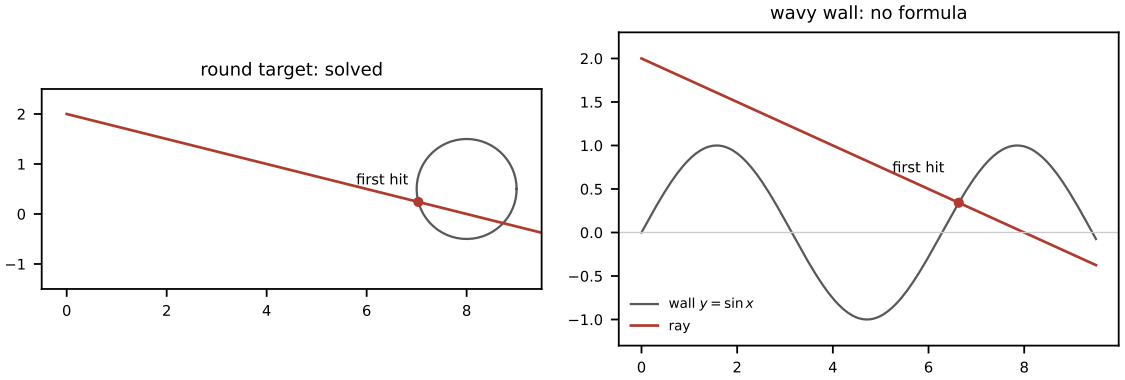
เราเริ่มจากกระสุนที่ไปชนผนังโค้ง แล้วจะเจออีกด้านหนึ่งของปัญหาเดียวกัน ตามลำดับนี้

- หัวข้อ 6.1: ล้อมคำตอบไว้ในกรง เราหารากของสมการด้วยการแบ่งช่วงที่รู้ว่ามีคำตอบ และด้วยการทำซ้ำของฟังก์ชัน $x = g(x)$ เมื่อการทำซ้ำนั้นหดระยะห่างระหว่างจุดลงเรื่อย ๆ
- หัวข้อ 6.2: แทนด้วยเส้นสัมผัส วิธีของนิวตันใช้เส้นสัมผัสเพื่อขยับคำตอบเข้าใกล้ราก และเราจะเห็นอีกด้านหนึ่งคือการหาค่า $1/\sqrt{x}$ ที่ใช้ทำให้เวกเตอร์มีความยาวหนึ่งหน่วย โดยเดาหนึ่งครั้งแล้วปรับด้วยเส้นสัมผัสหนึ่งรอบ
- หัวข้อ 6.3: เมื่อไม่รู้จะเริ่มตรงไหน **ค่าเดาตั้งต้น** (initial guess) คือค่าแรกที่ใช้เดาไว้ก่อนเริ่มทำซ้ำ ถ้าค่านี้ไกลเกินไปอาจทำให้วิธีของนิวตันพาเราไปผิดจุด

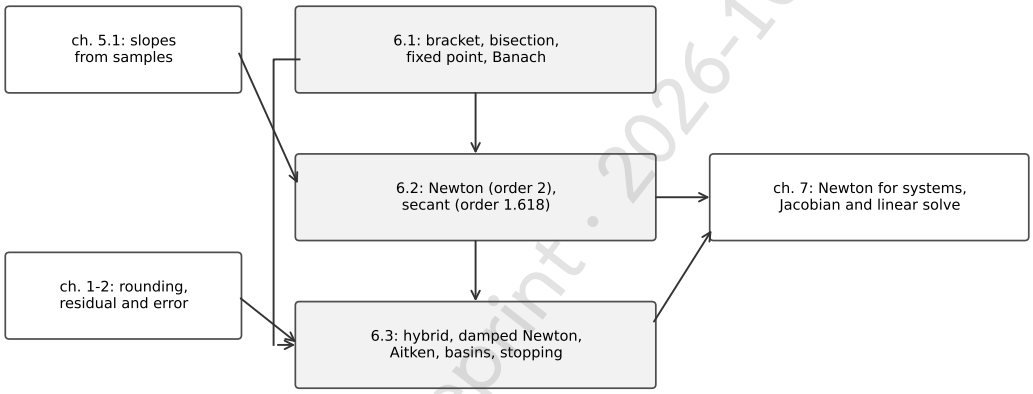
เราเริ่มจากหัวข้อ 6.1 โดยให้กระสุนชนผนังโค้งเป็นตัวอย่างตลอด ภาพต่อไปนี้แสดงกระสุนเส้นเดียวกันที่ไปชนลูกบอลกลมทางซ้าย และผนังคลื่นทางขวา

ภาพรวมของลำดับหัวข้อในบทนี้แสดงไว้ในรูปต่อไปนี้

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 6.1: กระสุนเส้นเดียวกัน ทางซ้ายชนลูกบอลกลม (รัศมี 1 ที่จุด $(8, 0.5)$) จุดชนหาได้จากสมการกำลังสอง ทางขวาชนผนัง $y = \sin x$ จุดชนหาไม่ได้ด้วยสูตร (สร้างจาก `numanal/ch06.py`)



รูปที่ 6.2: ลำดับของหัวข้อในบทที่ 6 และการเชื่อมกับบทที่ 5, บทที่ 7 และบทที่ 1--2 (สร้างจาก `numanal/ch06.py`)

ฉบับร่าง · Preprint · 2026-10-23

6.1 ล้อมคำตอบไว้ในกรง

สถานการณ์สมมติ: กรงที่รู้ว่าอยู่ข้างใน

เมื่อกระสุนอยู่ที่ $x = 0$ มันอยู่เหนือผนัง และที่ $x = 8$ มันอยู่ใต้ผนัง เรารู้จากตรงนี้แล้วว่าจุดชนต้องอยู่ระหว่างสองจุดนั้น แต่ยังไม่รู้ว่าอยู่ตรงไหน คำถามคือ เราจะทำให้ช่วงนั้นแคบลงเรื่อย ๆ ได้อย่างไร และแต่ละครั้งที่แคบลง เรารู้ได้ไหมว่าคำตอบห่างจากจุดที่เราเลือกไม่เกินเท่าไร

ภาพรวมก่อนอ่าน

ความคิดของหัวข้อนี้คือ ถ้ามีจุดหนึ่งที่อยู่นอกผนังและอีกจุดหนึ่งที่อยู่ในผนัง แล้วเราเลือกจุดกึ่งกลาง จุดนั้นจะอยู่ฝั่งเดียวกับจุดนอก หรือฝั่งเดียวกับจุดใน เราก็ทิ้งครึ่งหนึ่งที่ไม่มีจุดชนได้ทุกครั้ง ทำซ้ำไปช่วงจะเหลือครึ่งหนึ่งทุกรอบ จุดชนจึงถูก "ล้อม" ไว้ในกรงที่เล็กลงเรื่อย ๆ

ช่วงที่ล้อมคำตอบ

เราเริ่มจากฟังก์ชันที่บอกระยะของกระสุนเหนือผนัง

$$\operatorname{gap}(x) = \left(2 - \tfrac{x}{4}\right) - \sin x. \tag{6.1}$$

ถ้า $\operatorname{gap}(x) > 0$ กระสุนอยู่เหนือผนัง (เรียกว่า*ด้านนอก*) ถ้า $\operatorname{gap}(x) < 0$ กระสุนอยู่ใต้ผนัง (*ด้านใน*) จุดชนคือจุดที่ $\operatorname{gap}(x) = 0$ นั่นคือ จุดที่หาคำตอบของสมการ $\operatorname{gap}(x) = 0$ ซึ่งเรียกว่า **ราก** (root) ของฟังก์ชัน

เราคำนวณได้ว่า $\operatorname{gap}(0) = 2 > 0$ และ $\operatorname{gap}(8) \approx -0.989 < 0$ ดังนั้นต้องมีรากอยู่ในช่วง $[0, 8]$

นิยาม 6.1.1: ช่วงที่ล้อมคำตอบ

ให้ f เป็นฟังก์ชันต่อเนื่องบน $[a, b]$ ถ้า $f(a)\,f(b) < 0$ เราเรียก $[a, b]$ ว่า **ช่วงล้อม** (bracket) ของ f (ช่วงที่ล้อมคำตอบ คำนี้เราตั้งขึ้นเองในหนังสือเล่มนี้)

อ่านแบบเข้าใจง่าย. ช่วงล้อมคือช่วงที่ปลายสองข้างอยู่คนละฝั่งของเส้นศูนย์ ปลายข้างหนึ่งเป็นบวก อีกข้างเป็นลบ ทฤษฎีบทค่ากลางของฟังก์ชันต่อเนื่องบอกว่า กราฟต้องข้ามศูนย์ในช่วงนี้อย่างน้อยหนึ่งจุด เราจึงมั่นใจได้ว่ามีรากอยู่ข้างใน ตัวอย่างคือ $[0, 8]$ ของผนังของเรา ตัวอย่างที่ไม่ใช่ช่วงล้อมคือ $[9, 12]$ มีรากอยู่สองราก แต่ค่าที่ปลายทั้งสองเป็นลบ ช่วงล้อมจึงไม่บอกว่ามีรากเพียงรากเดียว และไม่บอกว่ามีรากอยู่ที่ไหนแน่ ๆ

วิธีแบ่งครึ่ง

สถานการณ์สมมติ: เลือกจุดกลางแล้วทิ้งครึ่งหนึ่ง

ถ้าเราเลือกจุดกลางของ $[0, 8]$ คือ $x = 4$ จะได้ $\operatorname{gap}(4) \approx 1.757 > 0$ กระสุนที่ $x = 4$ ยังอยู่เหนือผนัง ดังนั้นจุดชนต้องอยู่ใน $[4, 8]$ ครึ่งหนึ่งของช่วงหายไปแล้ว โดยไม่ต้องรู้ค่าของผนังที่จุดชนเลย

ลำดับการทำงานของ **วิธีแบ่งครึ่ง** (bisection method) เขียนเป็นขั้นตอนได้ดังนี้

1. เริ่มจากช่วงล้อม $[a_0, b_0]$ เช่น $[0, 8]$
2. หาจุดกลาง $c_n = \dfrac{a_n + b_n}{2}$
3. ถ้า $f(a_n)$ กับ $f(c_n)$ มีเครื่องหมายต่างกัน (หรือ $f(c_n) = 0$) ให้เก็บช่วง $[a_n, c_n]$ มิฉะนั้นเก็บ $[c_n, b_n]$ เป็นช่วงใหม่ ช่วงที่เก็บยังล้อมรากเสมอ
4. ทำซ้ำข้อ 2--3 จนได้ความแม่นที่ต้องการ

ค่า c_n คือ*คำตอบที่* n ของวิธีนี้ เราจะเห็นในตารางที่ 6.1 ว่าค่า c_n เข้าใกล้จุดชนขึ้นทุกรอบ และช่วงที่ล้อมรากเล็กลงครึ่งหนึ่งทุกรอบ

ลองเอง

เริ่มจาก $[0, 8]$ และรู้ค่า $\operatorname{gap}(4) \approx 1.757$, $\operatorname{gap}(6) \approx 0.779$, $\operatorname{gap}(7) \approx -0.407$ จงหาว่าหลังสองขั้นช่วงล้อมคือช่วงใด (ขั้นแรกได้ $[4, 8]$) และจุดกึ่งกลางถัดไปคืออะไร แล้วเทียบคำตอบกับตารางที่ 6.1

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 6.1: วิธีแบ่งครึ่งบน $[0,8]$ สำหรับผนัง $y=\sin x$ (ตัวเลขจากโค้ดของเรา)

n	c_n	$\lvert c_n-\alpha\rvert$	ขอบเขต $\frac{8}{2^{n+1}}$
0	4	2.632181	4
1	6	0.632181	2
2	7	0.367819	1
3	6.5	0.132181	0.5
4	6.75	0.117819	0.25
5	6.625	0.007181	0.125

ทฤษฎีบท 6.1.1: ความแม่นของวิธีแบ่งครึ่ง

ให้ f ต่อเนื่องบน $[a,b]$ และ $f(a)\,f(b)<0$ ให้ c_n เป็นคำตอบที่ n ของวิธีแบ่งครึ่ง จะมีรากหนึ่งราก $\alpha\in[a,b]$ ของ f ที่ทำให้ทุก $n\geq 0$ จะได้

$$\lvert c_n-\alpha\rvert\leq\frac{b-a}{2^{n+1}}.$$

พิสูจน์. **แนวคิด.** ช่วงที่เก็บแต่ละชั้นซ้อนอยู่ในชั้นก่อน ความยาวลดลงครึ่งหนึ่ง และปลายทั้งสองไม่เคยมีเครื่องหมายเดียวกัน

1. **ช่วงซ้อนกัน.** ให้ $I_n=[a_n,b_n]$ คือช่วงที่เก็บที่ขั้น n โดยวิธีการ ช่วง I_{n+1} เป็นครึ่งหนึ่งของ I_n ความยาวจึงเป็น $(b-a)/2^n$ และ $I_0\supseteq I_1\supseteq\cdots$
2. **ปลายยังล้อมรากเสมอ.** เราแสดงว่า $f(a_n)f(b_n)\leq 0$ ทุก n โดยอุปนัย ที่ $n=0$ จริงตามสมมติฐาน ถ้า $f(a_n)f(c_n)\leq 0$ ปลายใหม่คือ (a_n,c_n) และ $f(a_n)f(c_n)\leq 0$ ตามที่ต้องการ ถ้าไม่ใช่ แปลว่า $f(c_n)$ มีเครื่องหมายเดียวกับ $f(a_n)$ จึงได้ $f(c_n)f(b_n)\leq 0$ เพราะ $f(a_n)f(b_n)\leq 0$ (ปลายใหม่คือ (c_n,b_n))
3. **จุดเดียวที่ซ้อนกันทั้งหมด.** ช่วง I_n เป็นปิดและซ้อนกัน และความยาวไปที่ศูนย์ ดังนั้นมีจุด α จุดเดียวที่อยู่ในทุก I_n ปลาย a_n และ b_n ต่างลู่เข้าหา α จึงได้ $f(a_n)\to f(\alpha)$ และ $f(b_n)\to f(\alpha)$
4. α **เป็นราก.** จากข้อ 2 เมื่อ $n\to\infty$ จะได้ $f(\alpha)^2\leq 0$ จึง $f(\alpha)=0$
5. **ขอบเขต.** $\alpha\in I_n$ และ c_n เป็นจุดกึ่งกลางของ I_n ดังนั้น $\lvert c_n-\alpha\rvert\leq\frac{1}{2}(b_n-a_n)=\frac{b-a}{2^{n+1}}$

สรุป: ความยาวของช่วงที่ล้อมรากลดลงครึ่งหนึ่งทุกรอบ และจุดกึ่งกลางอยู่ห่างจากรากไม่เกินครึ่งความยาวนั้น □

อ่านแบบเข้าใจง่าย

ทฤษฎีบทนี้บอกสองอย่าง อย่างแรกคือ วิธีจะลู่เข้าเสมอ ขอแค่ f ต่อเนื่องและปลายสองข้างอยู่คนละฝั่งของศูนย์ อย่างที่สองคือ เราได้ *ขอบเขตค่าคลาดเคลื่อนสัมบูรณ์ที่รับประกัน* ก่อนจะรู้คำตอบจริงด้วยซ้ำ เพราะ $(b-a)/2^{n+1}$ คำนวณได้จากช่วงเริ่มต้นกับ n เท่านั้น

ลองใช้กับผนังของเรา ตารางที่ 6.1 ให้ค่า c_n สำหรับ $n=0,\ldots,5$ บนช่วง $[0,8]$ โดยมีรากจริง $\alpha\approx 6.632181$ (ตัวเลขจากโค้ดของเรา)

ตารางนี้ให้ข้อสังเกตสองอย่าง อย่างแรก ค่าคลาดเคลื่อนไม่ได้ลดทุกรอบ เช่น ที่ $n=6$ ค่า $c_6=6.6875$ ไกลจากรากกว่า c_5 (ค่าคลาดเคลื่อน 0.0553 เทียบกับ 0.0072) อย่างที่สอง ขอบเขตเป็น*การรับประกัน* ที่ลดลงครึ่งหนึ่งทุกรอบ มันบอกว่าค่าคลาดเคลื่อนสัมบูรณ์ไม่เกินเท่าไร ไม่ได้บอกว่าแต่ละรอบจะดีขึ้นแน่นอน

ฉบับร่าง Preprint · 2026-10-23

เกร็ดเพิ่มเติม

ถ้าต้องการความแม่น ε เราหาจำนวนรอบที่น้อยที่สุดจาก $(b-a)/2^{n+1} \le \varepsilon$ บนช่วง $[0,8]$ กับ $\varepsilon = 10^{-6}$ ได้ $n = 22$ รอบ ตัวเลขนี้เป็นการรับประกัน และโค้ดของเราทดสอบว่า c_{22} อยู่ห่างจากรากไม่เกิน 10^{-6} จริง อย่างไรก็ตามค่าที่ได้ดีกว่านี้ในบางกรณี เราจะเห็นวิธีที่เร็วกว่าในหัวข้อ "ทำซ้ำ $x = g(x)$" ด้านล่าง

ข้อควรระวัง: ช่วงที่มีรากหลายตัว

ทฤษฎีบทบอกว่า วิธีจะหาราก α ซึ่ง*รากหนึ่งราก* ในช่วงเท่านั้น มันไม่รับประกันว่าได้รากแรกที่กระสุนชน ลองดูสามตัวอย่าง

- บน $[0,8]$ gap เปลี่ยนเครื่องหมายเพียงครั้งเดียวที่ $x \approx 6.632$ จึงรู้ว่านั่นคือจุดชนแรก
- บน $[0,12]$ gap เปลี่ยนเครื่องหมายสามครั้ง ที่ $x \approx 6.632,\ 9.928$ และ 11.501 ในการคำนวณของเรา วิธีแบ่งครึ่งให้ 6.632 แต่ทฤษฎีบทไม่รับประกันว่าได้รากแรก
- บน $[6.5,10]$ ค่าที่ปลายทั้งสองเป็นบวก ($\text{gap}(10) \approx 0.044$) แม้จะมีรากสองราก วิธีแบ่งครึ่งจึงเริ่มไม่ได้ เพราะเงื่อนไข $f(a)f(b) < 0$ ไม่เป็นจริง

ก่อนเชื่อผลลัพธ์ ให้ตรวจก่อนว่าช่วงที่ใช้มีการเปลี่ยนเครื่องหมายเพียงครั้งเดียว วิธีง่ายคือดูค่าบนตารางจุดที่ห่างกันพอสมควร ดังที่โค้ดของเราทำ

อีกทาง: ทำซ้ำ $x = g(x)$

วิธีแบ่งครึ่งใช้ค่า gap อย่างเดียว ไม่ได้ใช้อนุพันธ์ของมัน อีกวิธีคือเขียนสมการใหม่ให้อยู่ในรูป $x = g(x)$ แล้วทำซ้ำ $x_{n+1} = g(x_n)$

นิยาม 6.1.2: จุดตรึง

จุด α เรียกว่า **จุดตรึง** (fixed point) ของ g ถ้า $g(\alpha) = \alpha$ การทำซ้ำ $x_{n+1} = g(x_n)$ เริ่มจาก x_0 เรียกว่า *การทำซ้ำแบบจุดตรึง*

อ่านแบบเข้าใจง่าย. จุดตรึงคือจุดที่ g ส่งกลับมาที่ตัวเอง เมื่อเริ่มใกล้มัน การทำซ้ำอาจพาเราเข้าไปหามัน หรืออาจพาเราออกไป ขึ้นอยู่กับว่า g "ดึงเข้า" หรือ "ผลักออก" ตัวอย่างเช่น $g(x) = x + \frac{1}{2}\,\text{gap}(x)$ มีจุดตรึงที่เดียวกับรากของ gap เพราะ $g(x) = x$ ก็ต่อเมื่อ $\text{gap}(x) = 0$ จุดตรึงนี้คือ $\alpha \approx 6.632181$ เหมือนจุดชนของผนัง แต่การทำซ้ำอาจลู่เข้าหรือไม่ก็ได้ ตัวอย่างอีกแบบคือ $\tilde{g}(x) = 8 - 4\sin x$ มีจุดตรึงที่เดียวกัน เพราะ $x = 8 - 4\sin x$ เทียบเท่ากับ $\sin x = 2 - \frac{x}{4}$ แต่เมื่อเริ่มที่ $x_0 = 6$ จะได้ $\tilde{g}(6) \approx 9.118$ ซึ่งออกนอกช่วง $[6,7]$ และการทำซ้ำกระโดดไปมาไม่เข้าหาราก (ดูแบบฝึกหัด 6.1.6)

นิยาม 6.1.3: การหดตัว

ให้ $[a,b]$ เป็นช่วงปิด ฟังก์ชัน g บน $[a,b]$ เป็น **การหดตัว** (contraction) ถ้ามี L ที่ $0 \le L < 1$ ซึ่งทำให้

$$|g(x) - g(y)| \le L\,|x - y| \qquad \text{สำหรับทุก } x, y \in [a,b].$$

ค่า L เรียกว่า**ค่าคงตัวการหดตัว** (contraction constant) การหดตัวคือฟังก์ชัน**ลิพชิทซ์** (Lipschitz) ที่มีค่าคงที่น้อยกว่า 1

ฉบับร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย. การหดตัวบอกว่า ระยะระหว่างภาพของสองจุดใด ๆ ไม่ยาวกว่าระยะระหว่างสองจุดเดิม และยาวน้อยกว่าเป็น L เท่า ตัวอย่างเช่น ถ้า $L = \frac{1}{2}$ ทุกการทำซ้ำจะทำให้ระยะห่างระหว่างจุดสองจุดเหลือไม่เกินครึ่งหนึ่ง ไม่ว่าจุดนั้นจะอยู่ตรงไหนในช่วง

ทฤษฎีบทสำคัญของหัวข้อนี้คือ การหดตัวทำให้การทำซ้ำลู่เข้าสู่จุดตรึงเพียงจุดเดียว และเราประเมินค่าคลาดเคลื่อนได้ทั้งก่อนและหลังการคำนวณ

ทฤษฎีบท 6.1.2: ทฤษฎีบทจุดตรึง (รูปแบบของบานัค)

ให้ $[a,b]$ เป็นช่วงปิด และให้ $g : [a,b] \to [a,b]$ เป็นการหดตัวด้วยค่าคงที่ L โดยที่ $0 \le L < 1$ จะได้ว่า

1. มีจุดตรึง $\alpha \in [a,b]$ จุดเดียวของ g ในช่วง $[a,b]$
2. สำหรับ $x_0 \in [a,b]$ ใด ๆ ลำดับ $x_{n+1} = g(x_n)$ อยู่ใน $[a,b]$ และ $x_n \to \alpha$
3. สำหรับ $n \ge 1$

$$|x_n - \alpha| \le \frac{L^n}{1-L}\,|x_1 - x_0| \qquad \text{และ} \qquad |x_n - \alpha| \le \frac{L}{1-L}\,|x_n - x_{n-1}|.$$

พิสูจน์. **แนวคิด.** ระยะระหว่างสองรอบติดกันหดลงด้วย L ทุกรอบ จึงรวมกันเป็นอนุกรมเรขาคณิตที่ลู่เข้า แล้วลำดับเป็นลำดับโคชี

1. **อยู่ในช่วง.** ถ้า $x_n \in [a,b]$ แล้ว $x_{n+1} = g(x_n) \in [a,b]$ เพราะ g ส่ง $[a,b]$ เข้าไปในตัวมันเอง ดังนั้นทุก x_n อยู่ใน $[a,b]$ โดยอุปนัย
2. **ระยะติดกันหดลง.** จาก $|x_{k+1} - x_k| = |g(x_k) - g(x_{k-1})| \le L|x_k - x_{k-1}|$ โดยอุปนัยได้ $|x_{k+1} - x_k| \le L^k|x_1 - x_0|$
3. **เป็นลำดับโคชี.** สำหรับ $m > n$ ใช้อสมการสามเหลี่ยมกับข้อ 2 จะได้

$$|x_m - x_n| \le \sum_{k=n}^{m-1} L^k|x_1 - x_0| \le \frac{L^n}{1-L}|x_1 - x_0|.$$

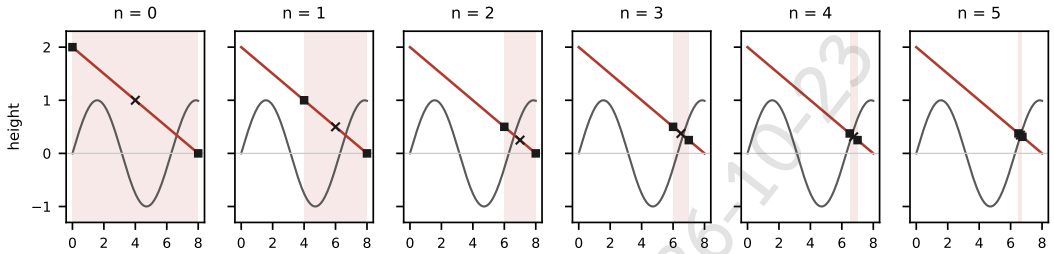
   ค่านี้ไปที่ 0 เมื่อ $n \to \infty$ เพราะ $0 \le L < 1$ ลำดับจึงเป็นลำดับโคชี และเพราะจำนวนจริงสมบูรณ์ มีลิมิต α ซึ่งอยู่ใน $[a,b]$ เพราะช่วงเป็นปิด
4. α **เป็นจุดตรึง.** g ต่อเนื่อง (เพราะเป็นการหดตัว) ดังนั้น $\alpha = \lim x_{n+1} = \lim g(x_n) = g(\alpha)$
5. **ความเป็นเอกลักษณ์.** ถ้า $\beta \in [a,b]$ ก็เป็นจุดตรึง จะได้ $|\alpha - \beta| = |g(\alpha) - g(\beta)| \le L|\alpha - \beta|$ จึง $(1-L)|\alpha - \beta| \le 0$ และ $\alpha = \beta$
6. **ขอบเขตแรก.** ในข้อ 3 ให้ $m \to \infty$ จะได้ $|\alpha - x_n| \le \frac{L^n}{1-L}|x_1 - x_0|$
7. **ขอบเขตที่สอง.** จากข้อ 2 โดยใช้การหดตัว $k-n+1$ ครั้งจะได้ $|x_{k+1} - x_k| \le L^{k-n+1}|x_n - x_{n-1}|$ รวมกันสำหรับ $k = n, \dots, m-1$ แล้วให้ $m \to \infty$ จะได้ $|\alpha - x_n| \le \frac{L}{1-L}|x_n - x_{n-1}|$

สรุป: การหดตัวทำให้ระยะระหว่างรอบติดกันลดลงเรื่อย ๆ ทำให้ลำดับโคชีและลู่เข้า ความเป็นเอกลักษณ์ตามจากการหดตัวเอง □

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 6.2: การทำซ้ำ $x_{n+1} = g(x_n)$ สำหรับ $g(x) = x + \frac{1}{2}\,\mathrm{gap}(x)$ เริ่มที่ $x_0 = 6$ (ตัวเลขจากโค้ดของเรา)

n	x_n	$\lvert x_n - \alpha\rvert$	อัตราส่วน $\dfrac{\lvert x_n-\alpha\rvert}{\lvert x_{n-1}-\alpha\rvert}$
0	6	0.632181	
1	6.389708	0.242474	0.3836
2	6.537834	0.094348	0.3891
3	6.594652	0.037530	0.3978
4	6.617093	0.015089	0.4020
5	6.626088	0.006094	0.4039



รูปที่ 6.3: กรงที่เล็กลง: ค่า $n = 0$ ถึง 5 ของวิธีแบ่งครึ่ง บนช่วง $[0, 8]$ แสดงกระสุน ผนัง และช่วงล้อม (แถบสีแดงอ่อน) จุดสี่เหลี่ยมคือปลายของช่วง กากบาทคือจุดกึ่งกลาง (สร้างจาก `numanal/ch06.py`)

ค่า L มักหาได้ง่ายจากอนุพันธ์ ถ้า g หาอนุพันธ์ได้ต่อเนื่องบน $[a, b]$ และ $g([a, b]) \subseteq [a, b]$ และ $\max_{[a,b]} |g'| < 1$ แล้ว g เป็นการหดตัวด้วย $L = \max_{[a,b]} |g'|$ เพราะทฤษฎีบทค่าเฉลี่ยให้ $g(x) - g(y) = g'(\xi)(x - y)$ สำหรับ ξ อยู่ระหว่าง x กับ y

ข้อควรระวัง: ต้องส่งช่วงเข้าไปในตัวมันเอง

เงื่อนไข $g([a, b]) \subseteq [a, b]$ ไม่ควรถูกข้าม ถ้า g ส่งจุดในช่วงออกไปนอกช่วง การทำซ้ำอาจหลุดจากช่วงที่เรากำหนดไว้ และทฤษฎีบทไม่ได้บอกอะไรต่อ ในกรณีของผนัง $\tilde{g}(6) \approx 9.118$ อยู่นอก $[6, 7]$ จึงใช้ทฤษฎีบทบนช่วงนั้นไม่ได้

การลู่เข้าของผนังของเรา

ให้ $g(x) = x + \frac{1}{2}\,\mathrm{gap}(x)$ เราได้ $g'(x) = \frac{7}{8} - \frac{1}{2}\cos x$ บน $[6, 7]$ ค่า $\cos x$ อยู่ในช่วง $[\cos 7, 1]$ ดังนั้น g' มีค่าอยู่ระหว่าง 0.375 ถึง 0.498 และ $g(6) \approx 6.389708$, $g(7) \approx 6.796507$ จึงส่ง $[6, 7]$ เข้าไปใน $[6, 7]$ ได้ และเป็นการหดตัวด้วย

$$L = \frac{7}{8} - \frac{1}{2}\cos 7 \approx 0.498049.$$

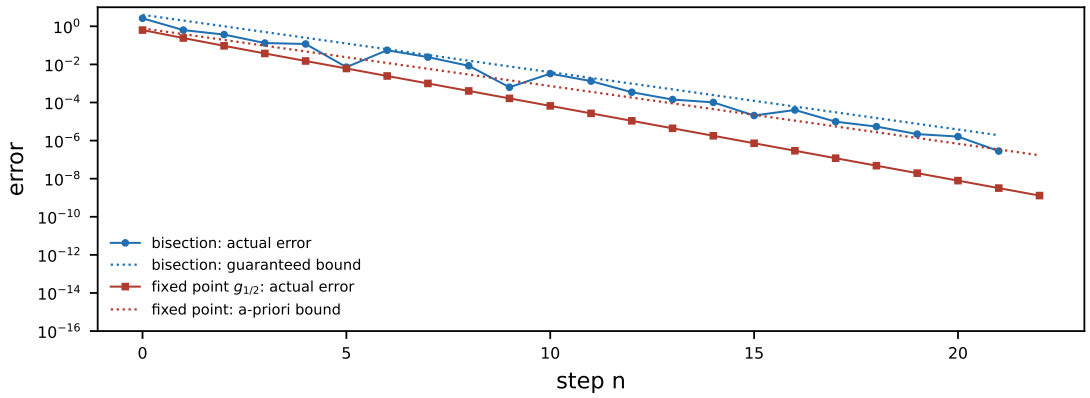
(เลือกช่วง $[6, 7]$ เพราะมีรากเพียงรากเดียวและอนุพันธ์อยู่ในช่วงที่ต้องการ)

ตารางที่ 6.2 เริ่มจาก $x_0 = 6$ ที่อยู่ในช่วง ค่า x_n เข้าใกล้ α และอัตราส่วนของค่าคลาดเคลื่อนแต่ละรอบเข้าใกล้ $g'(\alpha) \approx 0.40514$ ซึ่งน้อยกว่า L ที่ทฤษฎีบทให้ไว้

ความเร็วในการลู่เข้า

เมื่อเราได้วิธีสองวิธี เราต้องเปรียบเทียบความเร็ว ใช้นิยามต่อไปนี้

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 6.4: ค่าคลาดเคลื่อนเทียบกับจำนวนรอบบนแกนลอการิทึม: วิธีแบ่งครึ่งกับวิธีทำซ้ำ $g_{1/2}$ เส้นประคือขอบเขตที่รับประกัน เส้นทึบคือค่าคลาดเคลื่อนจริง (สร้างจาก `numanal/ch06.py`)

นิยาม 6.1.4: อันดับและอัตราการลู่เข้าเชิงเส้น

ให้ $x_n \to \alpha$

1. ลำดับนี้มี **อันดับของการลู่เข้า** (order of convergence) $p \geq 1$ ถ้ามี $C \geq 0$ และ N ที่ทำให้ $|x_{n+1} - \alpha| \leq C\,|x_n - \alpha|^p$ สำหรับ $n \geq N$ (นิยามตามตำรา Atkinson 1989, §2.0, p. 56)
2. ถ้า $p = 1$ และมี K กับ $\rho \in (0,1)$ ที่ทำให้ $|x_n - \alpha| \leq K\rho^n$ สำหรับทุก n เราว่าลำดับ**ลู่เข้าแบบเชิงเส้น** (linear convergence)ด้วย**อัตราการลู่เข้า** (rate of convergence) ρ

อ่านแบบเข้าใจง่าย. อันดับบอกว่าค่าคลาดเคลื่อนในรอบหนึ่งเป็นกี่กำลังของค่าคลาดเคลื่อนรอบก่อน อัตราเชิงเส้นบอกว่าค่าคลาดเคลื่อนลดลงด้วยตัวคูณคงที่โดยประมาณ แต่ไม่ได้บอกว่าทุกรอบลดลงด้วยตัวคูณนั้นพอดี ดูตารางของวิธีแบ่งครึ่ง: ค่าคลาดเคลื่อนขึ้นลงบ้าง

จากทฤษฎีบทที่เราได้ วิธีแบ่งครึ่งลู่เข้าแบบเชิงเส้นด้วยอัตรา $\frac{1}{2}$ โดยมี $K = \frac{b-a}{2}$ เพราะ $|c_n - \alpha| \leq \frac{b-a}{2}\left(\frac{1}{2}\right)^n$ และวิธีทำซ้ำ g ที่หดตัวด้วย L ลู่เข้าแบบเชิงเส้นด้วยอัตรา L โดยมี $K = |x_1 - x_0|/(1-L)$ ที่ได้จากทฤษฎีบท

ผลลัพธ์ที่น่าสนใจของตัวอย่างนี้คือ การทำซ้ำ $g_{1/2}$ จาก $x_0 = 6$ ต้องการเพียง 15 รอบ เพื่อให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-6} จริง ในขณะที่วิธีแบ่งครึ่งต้องการ 22 รอบเพื่อรับประกันค่าคลาดเคลื่อนนั้น ส่วนขอบเขตก่อนการคำนวณของทฤษฎีบทให้ 20 รอบ ส่วนนี้ยังมีช่องว่างเพราะ $L \approx 0.498$ มากกว่าอัตราจริง 0.405 ทำให้ขอบเขตที่รับประกันเข้มน้อยกว่าความจริง เรื่องนี้เป็นแบบฝึกหัด 6.1.7

สรุปใจความ

- วิธีแบ่งครึ่งหา "ช่วงที่ล้อม" ของราก: ช่วงที่มีค่าเป็นบวกและลบที่ปลาย ความยาวลดลงครึ่งหนึ่งทุกรอบ และได้ขอบเขตค่าคลาดเคลื่อนสัมบูรณ์ $(b-a)/2^{n+1}$ โดยไม่ต้องรู้ค่าราก
- วิธีทำซ้ำ $x = g(x)$ เร็วกว่าเมื่อ g หดระยะห่างได้จริงและส่งช่วงเข้าไปในตัวมันเอง ขอบเขต $L^n/(1-L)$ บอกความเร็วล่วงหน้า
- ช่วงล้อมบอกเพียงว่า "มีราก" อย่างน้อยหนึ่งรากข้างใน หากมีหลายรากต้องระวังว่าเราได้รากไหน

Preprint · 2026-10-23

- วิธีแบ่งครึ่งใช้เพียงเครื่องหมายของฟังก์ชัน ส่วนการทำซ้ำใช้ค่าของ g ทั้งสองวิธียังไม่ใช้อนุพันธ์ วิธีที่ใช้อนุพันธ์ (หัวข้อ 6.2) มักเร็วกว่ามากเมื่อเริ่มใกล้ราก

อ่านต่อ. บทที่ 2 ของ Atkinson อธิบายการหารากด้วยการแบ่งครึ่งและการทำซ้ำอย่างละเอียด Atkinson 1989, §2.1, pp. 56--58 และ Atkinson 1989, §2.5, pp. 76--83 โดยมี Lemma 2.5 และทฤษฎีบท 2.6 ที่ใกล้กับทฤษฎีบทของเรา สำหรับรูปแบบเวกเตอร์ที่มีลูกบอลรอบค่าเดาตั้งต้น ดู Iserles 2009, §7.1, pp. 123--127

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 6.1.1 (⋆) กระสุนออกจากจุด $(0,1)$ วิ่งตามเส้น $y = 1 - \frac{x}{2}$ ไปทางขวา และไปชนลูกบอลที่มีรัศมี 1 และจุดศูนย์กลาง $(4,0)$ (ในภาพสองมิติ)

1. จงแทน y ในสมการวงกลม แล้วจัดรูปเป็นสมการกำลังสองในตัวแปร x
2. จงหาจุดตัดทั้งสองจุด และบอกว่ากระสุนไปถึงจุดใดก่อน
3. จงตรวจว่าจุดแรกอยู่บนวงกลมจริง

แบบฝึกหัด 6.1.2 (⋆) ให้ $f(x) = x^3 - 2$ บน $[1,2]$

1. จงทำวิธีแบ่งครึ่งสามรอบแรก หาค่า c_0, c_1, c_2, c_3 และช่วงล้อมหลังแต่ละรอบ
2. จงหาขอบเขตค่าคลาดเคลื่อนสัมบูรณ์ $\dfrac{b-a}{2^{n+1}}$ ของ $c_0, \dots, c_3$ และตรวจว่า $|c_3 - 2^{1/3}|$ ไม่เกินขอบเขตนั้น
3. จงหาจำนวนรอบที่น้อยที่สุดที่ทำให้รับประกันค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-3}

แบบฝึกหัด 6.1.3 (⋆⋆) ให้ $g(x) = \cos x$ บน $[0,1]$ และให้ $x_0 = 0$

1. จงแสดงว่า g ส่ง $[0,1]$ เข้าในตัวมันเอง และ $|g'(x)| \le \sin 1$ บน $[0,1]$ จึงเป็นการหดตัวด้วย $L = \sin 1$
2. จงหาจำนวนรอบ n ที่น้อยที่สุดซึ่งทำให้ขอบเขต $\dfrac{L^n}{1-L}|x_1 - x_0|$ ไม่เกิน 10^{-4}
3. จงคำนวณการทำซ้ำจริง และหารอบแรกที่ $|x_n - \alpha| \le 10^{-4}$ จงเทียบกับข้อ 2 และอธิบายว่าทำไมต่างกัน

แบบฝึกหัด 6.1.4 (⋆⋆) ให้ $g(x) = x + \frac{1}{2}\operatorname{gap}(x)$ และ $x_0 = 6$ ตามตัวอย่างของผนังในหัวข้อ 6.1

1. จงหาค่า $L = \frac{7}{8} - \frac{1}{2}\cos 7$ และตรวจว่า $L < \frac{1}{2}$
2. จงหาจำนวนรอบ n ที่น้อยที่สุดซึ่งทำให้ขอบเขตหลังการคำนวณ $\dfrac{L}{1-L}|x_n - x_{n-1}|$ ไม่เกิน 10^{-8} และจำนวนรอบที่น้อยที่สุดสำหรับขอบเขตก่อนการคำนวณ $\dfrac{L^n}{1-L}|x_1 - x_0|$
3. จงหาจำนวนรอบจริงที่ทำให้ $|x_n - \alpha| \le 10^{-8}$ และอธิบายว่าขอบเขตหลังการคำนวณดีกว่าขอบเขตก่อนการคำนวณในกรณีนี้อย่างไร

แบบฝึกหัด 6.1.5 (⋆⋆) ให้ $g_2(x) = x + 2\operatorname{gap}(x)$ จุดตรึงของ g_2 คือรากเดียวกับ α ของผนัง

1. จงแสดงว่า $g_2'(x) = \frac{1}{2} - 2\cos x$ และบน $[6,7]$ ค่า $|g_2'(x)|$ อยู่ระหว่าง 1 กับ 1.5

ฉบับร่าง · Preprint · 2026-10-23

2. จงคำนวณ $x_1, \dots, x_5$ จาก $x_0 = 6.5$ และบอกว่าลำดับหลุดจาก $[6, 7]$ หรือไม่
3. จงหา $g_2'(\alpha)$ โดยประมาณ และอธิบายว่าทำไมทฤษฎีบทจุดตรึงใช้ไม่ได้กับ g_2 แม้ g_2 มีจุดตรึงเดียวกับ g

แบบฝึกหัด 6.1.6 (⋆) ให้ $\tilde{g}(x) = 8 - 4\sin x$

1. จงแสดงว่า $\tilde{g}(x) = x$ เทียบเท่ากับ $\mathrm{gap}(x) = 0$
2. จงคำนวณ $\tilde{g}(6)$ และ $\tilde{g}(7)$ และบอกว่าทฤษฎีบทจุดตรึงใช้ได้กับ $[6, 7]$ หรือไม่ เพราะเหตุใด
3. จงคำนวณ $x_1, \dots, x_6$ จาก $x_0 = 6$ และอธิบายว่าลำดับเข้าใกล้ α หรือไม่

แบบฝึกหัด 6.1.7 (⋆ ⋆ ⋆)

ตัวช่วยคิด

เขียน $x_{n+1} - \alpha = g(x_n) - g(\alpha)$ แล้วใช้ทฤษฎีบทค่าเฉลี่ยกับช่วงระหว่าง x_n และ α ต้องมั่นใจว่าจุดที่ได้จากทฤษฎีบทค่าเฉลี่ยเข้าใกล้ α ด้วย

ให้ g หาอนุพันธ์ได้ต่อเนื่องใกล้จุดตรึง α และ $|g'(\alpha)| < 1$ และให้ $x_{n+1} = g(x_n)$ ลู่เข้าหา α โดยที่ $x_n \neq \alpha$ ทุก n

1. จงแสดงว่า $\dfrac{x_{n+1} - \alpha}{x_n - \alpha} \to g'(\alpha)$
2. สำหรับผนังของเรา จงคำนวณ $g'(\alpha) = \frac{7}{8} - \frac{1}{2}\cos\alpha$ และเทียบกับอัตราส่วนค่าคลาดเคลื่อนในตารางของหัวข้อ 6.1
3. จงอธิบายว่าทำไมอัตราส่วนจริงจึงน้อยกว่า $L \approx 0.498$ ที่ใช้ในทฤษฎีบท

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 6.1.1.

1. แทน $y = 1 - \frac{x}{2}$ ในสมการ $(x-4)^2 + y^2 = 1$ จะได้ $(x-4)^2 + \left(1 - \frac{x}{2}\right)^2 = 1$ กระจายได้ $x^2 - 8x + 16 + 1 - x + \frac{x^2}{4} = 1$ รวมพจน์แล้วได้ $\frac{5}{4}x^2 - 9x + 16 = 0$ หรือ $5x^2 - 36x + 64 = 0$
2. ดิสคริมิแนนต์คือ $36^2 - 4 \cdot 5 \cdot 64 = 1296 - 1280 = 16$ จึงได้ $x = \frac{36 \pm 4}{10}$ คือ $x = 3.2$ หรือ $x = 4$ เมื่อ $x = 3.2$ ได้ $y = 1 - 1.6 = -0.6$ และเมื่อ $x = 4$ ได้ $y = -1$ กระสุนวิ่งไปทางขวา จึงไปถึง $x = 3.2$ ก่อน จุดชนแรกคือ $(3.2, -0.6)$
3. $(3.2 - 4)^2 + (-0.6)^2 = 0.64 + 0.36 = 1$ ดังนั้นจุดแรกอยู่บนวงกลมจริง

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 6.1.2.

1. $f(1) = -1$ และ $f(2) = 6 > 0$ จึงเริ่มจาก $[1, 2]$ ได้
   - รอบ 0: $c_0 = 1.5$ และ $f(1.5) = 1.375 > 0$ ต่างเครื่องหมายกับ $f(1)$ จึงได้ช่วงล้อม $[1, 1.5]$
   - รอบ 1: $c_1 = 1.25$ และ $f(1.25) = -0.046875 < 0$ เครื่องหมายเดียวกับ $f(1)$ จึงได้ช่วงล้อม $[1.25, 1.5]$
   - รอบ 2: $c_2 = 1.375$ และ $f(1.375) = 0.599609 > 0$ ต่างจาก $f(1.25)$ จึงได้ช่วงล้อม $[1.25, 1.375]$
   - รอบ 3: $c_3 = 1.3125$ (และ $f(1.3125) = 0.260986 > 0$)
2. ขอบเขต $\frac{b-a}{2^{n+1}}$ ใช้ $b - a = 1$ ได้ $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$ สำหรับ $c_0, \ldots, c_3$ และ $|c_3 - 2^{1/3}| = |1.3125 - 1.259921| = 0.052579 \le \frac{1}{16} = 0.0625$ ตามที่ทฤษฎีบทให้
3. ต้องการ $\frac{1}{2^{n+1}} \le 10^{-3}$ คือ $2^{n+1} \ge 1000$ เนื่องจาก $2^9 = 512 < 1000 \le 1024 = 2^{10}$ จึงต้อง $n + 1 \ge 10$ นั่นคือ $n = 9$ (รอบที่ $n = 8$ ไม่พอ เพราะขอบเขตเป็น $1/512 \approx 0.00195$)

เฉลย 6.1.3.

1. บน $[0, 1]$ ฟังก์ชัน $\cos$ ลดลง ดังนั้น $\cos([0, 1]) = [\cos 1, \cos 0] = [0.540\ldots, 1] \subseteq [0, 1]$ อนุพันธ์คือ $g'(x) = -\sin x$ และ $\sin x$ เพิ่มขึ้นบน $[0, 1]$ จึง $|g'(x)| \le \sin 1 \approx 0.841471 < 1$ ดังนั้นเป็นการหดตัวด้วย $L = \sin 1$ (จากข้อความหลังทฤษฎีบท)
2. $x_1 = \cos 0 = 1$ จึง $|x_1 - x_0| = 1$ ต้องการ $\frac{L^n}{1-L} \le 10^{-4}$ คือ $L^n \le 1.58529 \times 10^{-5}$ เมื่อหาลอการิทึม $n \ge \frac{\ln(1.58529 \times 10^{-5})}{\ln L} = \frac{-11.0522}{-0.172604} \approx 64.03$ จึงได้ $n = 65$ (ตรวจว่า $L^{64} \approx 1.59 \times 10^{-5}$ ยังไม่พอ)
3. จากการคำนวณ (โค้ดของเรา) $\alpha \approx 0.739085$ และรอบแรกที่ $|x_n - \alpha| \le 10^{-4}$ คือ $n = 21$ ต่างจากขอบเขตก่อนการคำนวณมาก เพราะขอบเขตใช้ค่า $L = \sin 1$ ซึ่งเป็นค่าที่แย่ที่สุดบนทั้งช่วง ส่วนที่ใกล้จุดตรึงจริงอัตราคือ $|g'(\alpha)| = \sin \alpha \approx 0.674$ ซึ่งน้อยกว่า L

ตรวจด้วยการคำนวณ: test_ex_6_1_3 ในไฟล์ทดสอบ.

เฉลย 6.1.4.

1. $g'(x) = \frac{7}{8} - \frac{1}{2}\cos x$ ค่า $\cos x$ บน $[6, 7]$ น้อยสุดที่ $x = 7$ คือ $\cos 7 \approx 0.753902$ จึง $\max|g'| = \frac{7}{8} - \frac{1}{2}\cos 7 \approx 0.498049$ ซึ่งน้อยกว่า $\frac{1}{2}$
2. จาก $x_0 = 6$ และ $x_1 = 6.389708$ ได้ $|x_1 - x_0| = 0.389708$ ขอบเขตก่อนการคำนวณ $\frac{L^n}{1-L} \cdot 0.389708 \le 10^{-8}$ ให้ $n = 27$ สำหรับขอบเขตหลังการคำนวณ $\frac{L}{1-L}|x_n - x_{n-1}|$ ตรวจทีละรอบได้ $n = 20$ ให้ 1.15×10^{-8} (ยังไม่พอ) และ $n = 21$ ให้ 4.67×10^{-9} จึงได้ $n = 21$
3. ค่าคลาดเคลื่อนสัมบูรณ์จริง $|x_n - \alpha|$ เป็น 1.95×10^{-8} ที่ $n = 19$, 7.91×10^{-9} ที่ $n = 20$ ดังนั้นรอบจริงคือ $n = 20$ ขอบเขตหลังการคำนวณได้ $n = 21$ ใกล้ความจริงมาก เพราะใช้ความต่างของสองรอบล่าสุดซึ่งสะท้อนอัตราจริง 0.405 ส่วนขอบเขตก่อนการคำนวณใช้ $L = 0.498$ จึงต้องมากกว่า 27 รอบ

ตรวจด้วยการคำนวณ: test_ex_6_1_4 ในไฟล์ทดสอบ.

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 6.1.5.

1. $g_2(x) = x + 2\,\mathrm{gap}(x) = x + 2\big(2 - \frac{x}{4} - \sin x\big)$ จึง $g_2'(x) = 1 + 2\big(-\frac{1}{4} - \cos x\big) = \frac{1}{2} - 2\cos x$ บน $[6,7]$ ค่า $\cos x$ อยู่ใน $[0.753902, 1]$ ดังนั้น $g_2'(x) \in [-1.5,\ -1.007804]$ และ $|g_2'(x)| \in [1.007804,\ 1.5]$

2. จาก $x_0 = 6.5$ ได้ $\mathrm{gap}(6.5) = 0.159880$ จึง $x_1 = 6.5 + 2(0.159880) = 6.819760$ จากนั้น $x_2 = 6.387490$, $x_3 = 6.985514$, $x_4 = 6.200763$ และ $x_5 = 7.265039$ ค่า x_5 อยู่นอก $[6,7]$ ดังนั้นลำดับหลุดออกจากช่วง

3. จาก $\cos\alpha \approx 0.939716$ ได้ $g_2'(\alpha) = \frac{1}{2} - 2(0.939716) \approx -1.37943$ ค่าสัมบูรณ์มากกว่า 1 ดังนั้นจุดตรึงเดียวกับ g แต่ g_2 ผลักเข้าหาจุดตรึงไม่ได้ ทฤษฎีบทจุดตรึงต้องการ $|g'| < 1$ ซึ่งไม่เป็นจริงบนช่วงใด ๆ ที่มี α ใน $[6,7]$

ตรวจด้วยการคำนวณ: test_ex_6_1_5 ในไฟล์ทดสอบ.

เฉลย 6.1.6.

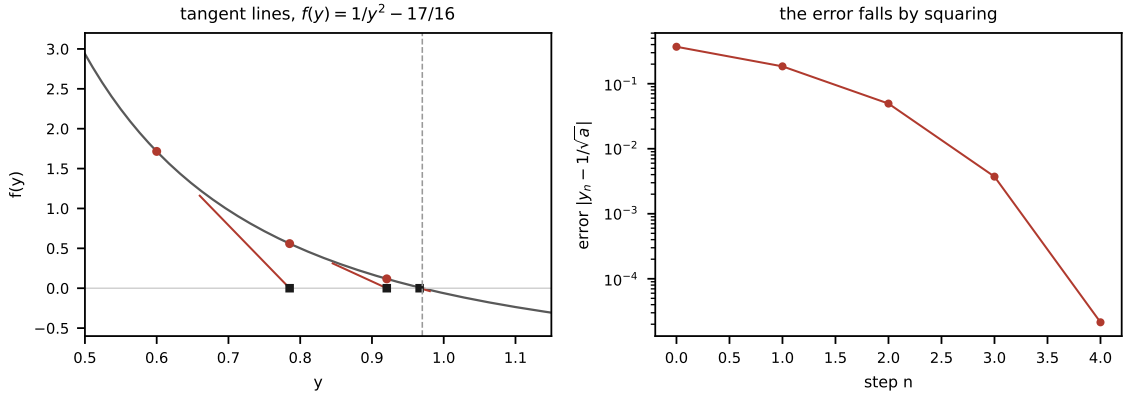
1. $\tilde{g}(x) = x$ ก็ต่อเมื่อ $8 - 4\sin x = x$ ซึ่งเท่ากับ $4\sin x = 8 - x$ หรือ $\sin x = 2 - \frac{x}{4}$ และเท่ากับ $\mathrm{gap}(x) = 0$ ตามนิยาม

2. $\tilde{g}(6) = 8 - 4\sin 6 = 8 + 1.117662 = 9.117662$ และ $\tilde{g}(7) = 8 - 4\sin 7 = 8 - 2.627946 = 5.372054$ ภาพของ $[6,7]$ จึงไม่อยู่ใน $[6,7]$ ทฤษฎีบทจุดตรึงจึงใช้ไม่ได้บนช่วงนี้ นอกจากนี้ $\tilde{g}'(x) = -4\cos x$ มีค่าสัมบูรณ์ 4 ที่ $x = 2\pi \approx 6.283$ ซึ่งอยู่ใน $[6,7]$

3. จาก $x_0 = 6$ ได้ $x_1 = 9.117662$, $x_2 = 6.790757$, $x_3 = 6.055775$, $x_4 = 8.901822$, $x_5 = 6.002227$ และ $x_6 = 9.109107$ ลำดับกระโดดไปมาระหว่างค่าประมาณ 6 กับค่าประมาณ 9 ไม่เข้าใกล้ α

ตรวจด้วยการคำนวณ: test_ex_6_1_6 ในไฟล์ทดสอบ.

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 6.5: ทางซ้าย: เส้นสัมผัสของ $f(y) = 1/y^2 - \frac{17}{16}$ ที่ $y_0 = 0.6$ และ y_1, y_2 จุดกลมคือ $(y_n, f(y_n))$ จุดสี่เหลี่ยมคือ y_{n+1} บนแกนนอน เส้นประคือ $1/\sqrt{17/16}$ ทางขวา: ค่าคลาดเคลื่อนของ y_n สำหรับ $n = 0$ ถึง 4 บนแกนลอการิทึม (สร้างจาก `numanal/ch06.py`)

เฉลย 6.1.7.

1. เนื่องจาก $x_{n+1} - \alpha = g(x_n) - g(\alpha)$ ทฤษฎีบทค่าเฉลี่ยให้จำนวน ξ_n ระหว่าง x_n กับ α ที่ทำให้ $x_{n+1} - \alpha = g'(\xi_n)(x_n - \alpha)$ เมื่อหารด้วย $x_n - \alpha \neq 0$ ได้ $\dfrac{x_{n+1} - \alpha}{x_n - \alpha} = g'(\xi_n)$ เพราะ $|\xi_n - \alpha| \le |x_n - \alpha| \to 0$ และ g' ต่อเนื่องที่ α จึงได้ $g'(\xi_n) \to g'(\alpha)$

2. $\cos\alpha \approx 0.939716$ จึง $g'(\alpha) = \frac{7}{8} - \frac{1}{2}(0.939716) \approx 0.405142$ ซึ่งตรงกับอัตราส่วนในตารางที่เข้าใกล้ 0.4051 เมื่อ n มาก

3. ทุกค่า x_n อยู่ระหว่าง 6 กับ α และ ξ_n อยู่ในช่วงนี้ด้วย บนช่วง $[6, \alpha]$ ค่า g' อยู่ระหว่าง 0.375 ถึง 0.405 (ค่าต่ำสุดที่ $x = 2\pi$) ค่าสูงสุด $L \approx 0.498$ เกิดที่ $x = 7$ ซึ่งการทำซ้ำไม่เคยเข้าไปถึง ขอบเขต L จึงเผื่อมากกว่าอัตราที่เกิดจริง

ตรวจด้วยการคำนวณ: test_ex_6_1_7 ในไฟล์ทดสอบ.

ฉบับร่าง · Preprint · 2026-10-25

6.2 แทนด้วยเส้นสัมผัส

สถานการณ์สมมติ: ทำให้ทิศทางยาวหนึ่งหน่วย

กระสุนของเรามีทิศทางเป็นเวกเตอร์ $(1, -\frac{1}{4})$ ความยาวกำลังสองของมันคือ $1 + \frac{1}{16} = \frac{17}{16}$ ซึ่งไม่เท่ากับหนึ่ง ถ้าโปรแกรมต้องการทิศทางเดียวกันที่ยาวหนึ่งหน่วย ต้องคูณเวกเตอร์ด้วย $\dfrac{1}{\sqrt{17/16}}$ ซึ่งก็คือ $1/\sqrt{x}$ เมื่อ $x = \frac{17}{16}$ ลองนึกว่าเราเขียนโปรแกรมฉากนี้เอง ทิศทางทุกทิศต้องถูกปรับความยาวให้เป็นหนึ่งก่อนนำไปคำนวณการชนหรือการสะท้อน ถ้าต้องคำนวณเองโดยไม่มีคำสั่งรากที่สอง เราจะหา $1/\sqrt{x}$ ได้อย่างไร และรู้ได้อย่างไรว่าคำตอบแม่นพอ

ภาพรวมก่อนอ่าน

ความคิดของหัวข้อนี้คือ เริ่มจากค่าเดาตั้งต้น y_0 แล้วลากเส้นสัมผัสของกราฟ ณ จุดนั้น จุดที่เส้นสัมผัสตัดแกนนอนมักใกล้คำตอบกว่าเดิม ทำซ้ำแบบเดิมจุดจะเข้าหาคำตอบ สิ่งที่ยังไม่รู้คือเข้าได้เร็วแค่ไหน รูปต่อไปนี้แสดงทั้งสองเรื่อง

ตารางที่ 6.3: วิธีนิวตันบนผนัง $\mathrm{gap}(x) = 0$ เริ่มที่ $x_0 = 6$ (ตัวเลขจากโค้ดของเรา)

n	x_n	$\lvert x_n - \alpha\rvert$	หลักที่ถูกต้อง
0	6	0.632181	0.2
1	6.644054	0.011873	1.9
2	6.632161	2.08×10^{-5}	4.7
3	6.632181	6.2×10^{-11}	10.2

จากเส้นสัมผัสสู่จุดถัดไป

เส้นสัมผัสของกราฟ f ที่จุด x_n คือเส้นตรงที่แตะกราฟที่จุดนั้นและมีความชันเท่ากับ $f'(x_n)$ เส้นนี้คือ $f(x_n) + f'(x_n)(x - x_n)$ ถ้าหาจุดที่เส้นนี้ตัดแกนนอน จะได้จุดใหม่ที่เราเรียกว่า x_{n+1}

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{6.2}$$

การทำซ้ำสูตรนี้เรียกว่า **วิธีนิวตัน** (Newton's method) ส่วน $f(x_n)/f'(x_n)$ คือระยะที่เราเลื่อนจาก x_n ถ้ากราฟชันมาก ระยะที่เลื่อนจะสั้น ถ้ากราฟแบนมาก ระยะที่เลื่อนจะยาว

ลองกับผนังของเรา

ผนังของเรามี $\mathrm{gap}(x)$ จากหัวข้อ 6.1 และอนุพันธ์คือ $\mathrm{gap}'(x) = -\frac{1}{4} - \cos x$ เริ่มจาก $x_0 = 6$ ได้ $\mathrm{gap}(6) \approx 0.779415$ และ $\mathrm{gap}'(6) \approx -1.210170$ ดังนั้น

$$x_1 = 6 - \frac{0.779415}{-1.210170} \approx 6.644054.$$

ตารางที่ 6.3 ทำซ้ำต่อ โดยใช้จำนวนหลักทศนิยมที่ถูกต้อง $-\log_{10}\lvert x_n - \alpha\rvert$ เป็นตัวบอกความแม่น ถ้าค่าคลาดเคลื่อนสัมบูรณ์คือราว 10^{-d} เราก็มีราว d หลักที่ถูกต้อง

ดูคอลัมน์สุดท้าย รอบแรกเพิ่มหลักที่ถูกต้องประมาณ 1.7 หลัก รอบที่สองเพิ่มประมาณ 2.8 หลัก และรอบที่สามเพิ่มประมาณ 5.5 หลัก นี่คือผลของ*ความเร็วกำลังสอง* ที่เราจะพิสูจน์ด้านล่าง ในตัวอย่างนี้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-6} เมื่อผ่านสามรอบ ส่วนวิธีแบ่งครึ่งต้องการ 22 รอบเพื่อ*รับประกัน*ค่าคลาดเคลื่อนนั้น และการทำซ้ำ $g_{1/2}$ ต้องการ 15 รอบ ตัวเลข 3 ที่นี่เป็นจำนวนรอบจริงของตัวอย่างนี้ ยังไม่มีการรับประกันในแบบเดียวกับหัวข้อ 6.1 (เรื่องนี้เป็นหัวข้อ 6.3)

ลองเอง

เริ่มจาก $x_0 = 6.5$ โดย $\mathrm{gap}(6.5) \approx 0.159880$ และ $\mathrm{gap}'(6.5) \approx -1.226588$ จงหา x_1 ด้วยมือ แล้วดูว่า x_1 อยู่ใกล้ $\alpha \approx 6.632181$ กว่าการเริ่มจาก $x_0 = 6$ หรือไม่ (คำตอบอยู่ในแบบฝึกหัด 6.2.2)

ใช้เส้นสัมผัสหา $1/\sqrt{a}$

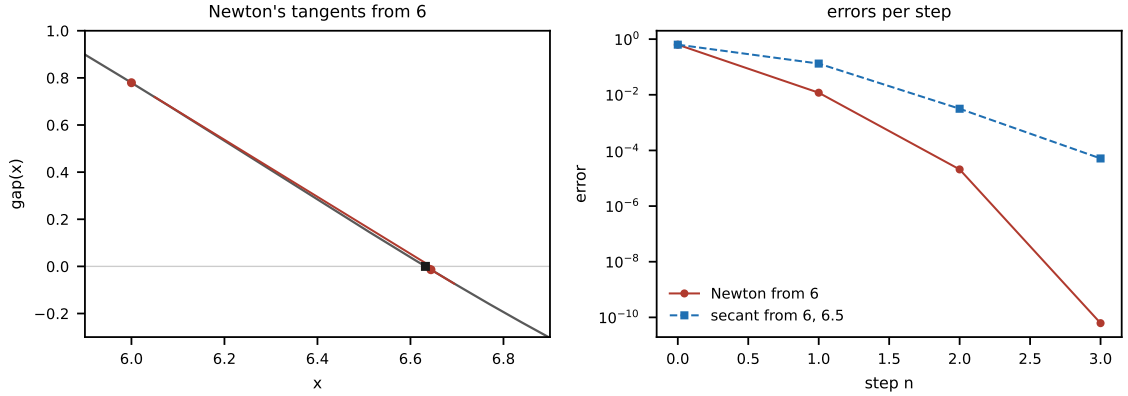
กลับไปที่ความยาวของกระสุน ให้ $a > 0$ และ $y = 1/\sqrt{a}$ จะได้ $y > 0$ และ $1/y^2 = a$ นั่นคือ y เป็นรากบวกของ

$$f(y) = \frac{1}{y^2} - a.$$

ความชันคือ $f'(y) = -2/y^3$ เมื่อแทนในสูตร (6.2) จะได้

$$y_{n+1} = y_n - \frac{1/y_n^2 - a}{-2/y_n^3} = \frac{y_n\bigl(3 - a\,y_n^2\bigr)}{2}.$$

Preprint 2026-10-23



รูปที่ 6.6: ทางซ้าย: เส้นสัมผัสของ gap ที่ $x_0 = 6$ และ x_1 (จุดกลม) จุดสี่เหลี่ยมคือราก ทางขวา: ค่าคลาดเคลื่อนของวิธีนิวตันจาก 6 และวิธีเซแคนต์จาก 6 กับ 6.5 บนแกนลอการิทึม (สร้างจาก `numanal/ch06.py`)

ตารางที่ 6.4: วิธีนิวตันหา $1/\sqrt{17/16} \approx 0.9701425$ จากค่าเดาตั้งต้น $y_0 = 0.6$ (ตัวเลขจากโค้ดของเรา)

n	y_n	$\lvert y_n - 1/\sqrt{a}\rvert$	หลักที่ถูกต้อง
0	0.600000	0.370143	0.4
1	0.785250	0.184893	0.7
2	0.920644	0.049498	1.3
3	0.966419	0.003724	2.4
4	0.970121	2.14×10^{-5}	4.7
5	0.970142	7.09×10^{-10}	9.1

$$y_{n+1} = \frac{y_n\left(3 - a\,y_n^2\right)}{2} \tag{6.3}$$

สูตรนี้ไม่มีรากที่สองและไม่มีการหาร มีแต่การคูณ การบวก และการหารด้วย 2 จึงเป็นการคำนวณที่ง่ายมากเมื่อเทียบกับการหารากที่สองโดยตรง

เริ่มจาก $y_0 = 0.6$ และ $a = \frac{17}{16}$ ได้ $y_1 = \dfrac{0.6\,(3 - \frac{17}{16} \cdot 0.36)}{2} = 0.78525$ ตารางที่ 6.4 ทำต่อให้ครบหกค่า

สามรอบแรกค่าคลาดเคลื่อนลดลงช้า เพราะ y_0 ยังไกลจากคำตอบ เส้นสัมผัสที่เอียงมากจึงพาเราไปได้ไม่ตรงนัก ตั้งแต่รอบที่สามเป็นต้นไป จำนวนหลักที่ถูกต้องเพิ่มเกือบเป็นสองเท่าทุกรอบ ซึ่งตรงกับความเร็วกำลังสองที่จะพิสูจน์ด้านล่าง

ถ้าเริ่มที่ $y_0 = 1$ แทน จำนวนหลักที่ถูกต้องของรอบที่ $0, 1, 2, 3$ คือ $1.5,\ 2.9,\ 5.5,\ 10.9$ (ตัวเลขจากโค้ดของเรา) จุดเริ่มที่อยู่ใกล้คำตอบจึงทำให้เร็วขึ้นได้

ลองเอง

เริ่มจาก $y_0 = 1$ และ $a = \frac{17}{16}$ จงหา y_1 ด้วยมือ แล้วดูว่าค่าคลาดเคลื่อนเท่าไร (คำตอบอยู่ในแบบฝึกหัด 6.2.3)

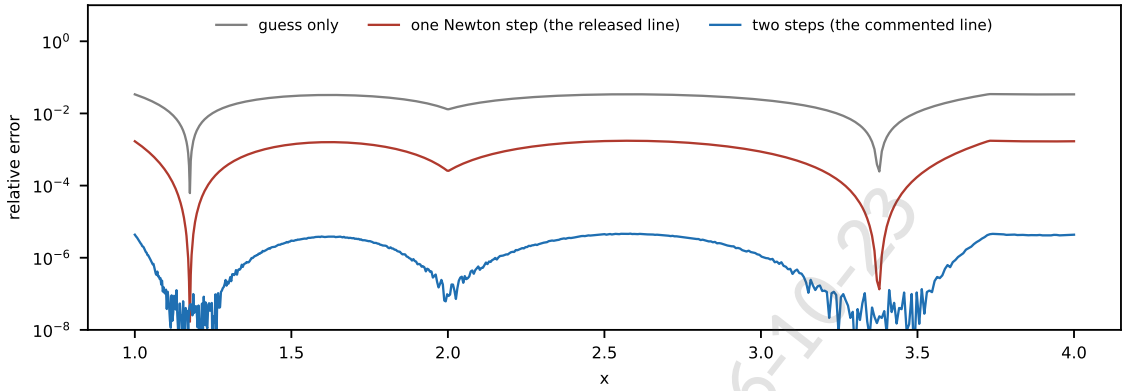
ข้อควรระวัง: ความชันเกือบเป็นศูนย์

สูตร (6.2) หารด้วย $f'(x_n)$ ถ้า $f'(x_n)$ เกือบเป็นศูนย์ ระยะที่เลื่อนจะใหญ่มาก และจุดใหม่อาจไปไกลจากรากเดิม สำหรับผนังของเรา $\mathrm{gap}'(x) = 0$ เมื่อ $\cos x = -\frac{1}{4}$ ซึ่งเกิดที่ $x \approx 4.4597$ ถ้าเริ่มที่ $x_0 = 4.4$ จะได้ $\mathrm{gap}(4.4) \approx 1.851602$ และ $\mathrm{gap}'(4.4) \approx 0.057333$ ดังนั้น $x_1 \approx 4.4 - \frac{1.851602}{0.057333} \approx -27.9$ ซึ่งอยู่ห่างจากราก $\alpha \approx 6.632$

ฉบับร่าง Preprint 2026-10-23

ตารางที่ 6.5: ฟังก์ชันแบบเดียวกับโค้ดต้นฉบับบน $x = \frac{17}{16}$ (ตัวเลขจากโค้ดของเรา)

ขั้นตอน	y	ค่าคลาดเคลื่อนสัมพัทธ์	หลักที่ถูกต้อง
ค่าเดาตั้งต้นจากบิต	0.950590	-2.02×10^{-2}	1.7
หนึ่งรอบ (บรรทัดที่ใช้)	0.969555	-6.05×10^{-4}	3.2
สองรอบ (บรรทัดที่เป็นความเห็น)	0.970142	-5.70×10^{-7}	6.2



รูปที่ 6.7: ค่าคลาดเคลื่อนสัมพัทธ์ของฟังก์ชันแบบเดียวกับ `Q_rsqrt` บน $[1, 4]$: ค่าเดาตั้งต้นอย่างเดียว หนึ่งรอบ และสองรอบ (สร้างจาก `numanal/ch06.py`)

มาก ทฤษฎีบทในหัวข้อนี้ไม่ได้บอกว่าต้องเริ่มใกล้แค่ไหนจึงจะไม่ก้าวหลุด เรื่องนี้เป็นหัวข้อ 6.3

ซอร์สที่เดาจากบิต

อีกด้านหนึ่งของปัญหาเดียวกัน: เดาหนึ่งครั้งแล้วปรับหนึ่งรอบ

ในตารางที่แล้วเราต้องทำหลายรอบกว่าจะได้ความแม่นสูง อีกด้านหนึ่งคือโค้ดที่ยอมให้ค่าคลาดเคลื่อนสัมพัทธ์ราวสองในพันเพื่อแลกกับการทำเพียงหนึ่งรอบ ตัวอย่างจริงคือฟังก์ชัน `Q_rsqrt` ในไฟล์ `q_math.c` ของซอร์สโค้ด Quake III Arena ที่ id Software เผยแพร่ Software 1999 ฟังก์ชันนี้รับตัวเลขทศนิยม 32 บิต แล้วคืนค่าประมาณของ $1/\sqrt{x}$ ฟังก์ชันทำสามขั้น ขั้นแรกสร้างค่าเดาตั้งต้นจากบิตของ x คือเลื่อนบิตของ x ไปทางขวาหนึ่งตำแหน่ง ลบออกจากค่าคงที่ `0x5f3759df` แล้วอ่านผลลัพธ์กลับเป็นเลขทศนิยม ขั้นที่สองคือหนึ่งรอบของวิธีนิวตันสำหรับ $f(y) = 1/y^2 - x$ ซึ่งเขียนได้เป็น $y(\frac{3}{2} - \frac{x}{2}y^2)$ นอกจากนี้ ในโค้ดมีบรรทัดของรอบที่สองไว้เป็นความเห็น ซึ่งหมายความว่าเอาออกได้

ความเห็นในโค้ดต้นฉบับบอกเพียงว่าเป็นการปรับบิตโดยตรงของตัวเลขทศนิยม และไม่ได้อธิบายที่มาของค่าคงที่นั้น

ในไฟล์เดียวกันมีฟังก์ชัน `VectorNormalizeFast` ที่ใช้ค่าที่ได้เพื่อทำให้เวกเตอร์ยาวหนึ่งหน่วย (ฟังก์ชันนี้อยู่ในบล็อกเงื่อนไขของตัวคอมไพล์ LCC) ซอร์สนี้บอกเพียงว่ามีฟังก์ชันนี้ ไม่ได้บอกว่าโปรแกรมที่ทำงานจริงเรียกมันบ่อยแค่ไหน ตัวเลขต่อไปนี้มาจากการเขียนใหม่ของเราเอง

เราเขียนฟังก์ชันนี้ใหม่ด้วยจำนวนเต็ม 32 บิตและเลขทศนิยม 32 บิต ให้การปัดเศษตรงกับการคำนวณเลขทศนิยมเดี่ยว ตารางที่ 6.5 ใช้ $x = \frac{17}{16}$ ซึ่งเป็นความยาวกำลังสองของกระสุนเรา

บนตารางจุด 3001 จุดในช่วง $[1, 4]$ ค่าคลาดเคลื่อนสัมพัทธ์สูงสุดคือประมาณ 3.4% ก่อนทำรอบ ประมาณ 0.18% หลังหนึ่งรอบ และประมาณ 4.7×10^{-6} หลังสองรอบ ค่าเหล่านี้เป็นค่าสูงสุดของตารางจุด ไม่ใช่ค่าสูงสุดที่แน่นอน

ฉบับร่าง Preprint · 2026-10-23

ค่าคลาดเคลื่อนหลังหนึ่งรอบไม่ได้เกิดแบบสุ่ม มันตามกฎโดยประมาณ $\varepsilon_{n+1} \approx -\frac{3}{2}\varepsilon_n^2$ เมื่อ ε_n คือค่าคลาดเคลื่อนสัมพัทธ์ของ y_n ที่มาของค่า $\frac{3}{2}$ คือ ให้ $ay_n^2 = 1 + \delta_n$ จากสูตร (6.3) จะได้ $\delta_{n+1} = -\frac{3}{4}\delta_n^2 + \frac{1}{4}\delta_n^3$ และเพราะ $\delta_n \approx 2\varepsilon_n$ จึงได้ $\varepsilon_{n+1} \approx -\frac{3}{2}\varepsilon_n^2$ ตรวจได้ที่ $x = 4$: $\varepsilon_0 \approx -0.033785$ ให้ $-\frac{3}{2}\varepsilon_0^2 \approx -1.71 \times 10^{-3}$ ซึ่งใกล้ค่าที่วัดได้ -1.69×10^{-3} (ต่างประมาณ 1%)

ค่าคลาดเคลื่อนนี้ไปถึงกระสุนโดยตรง เพราะเวกเตอร์หลังปรับคือ $u = y \cdot (1, -\frac{1}{4})$ ซึ่งยาว $|u| = y\sqrt{17/16}$ ดังนั้นความยาวหลังปรับเท่ากับ $1 + \varepsilon$ เมื่อ ε คือค่าคลาดเคลื่อนสัมพัทธ์ของ y หลังหนึ่งรอบ ความยาวเท่ากับ 0.999395 และหลังสองรอบเท่ากับ 0.99999943 (ตัวเลขจากโค้ดของเรา) หมายความว่าการปรับหนึ่งรอบยังทิ้งค่าคลาดเคลื่อนสัมพัทธ์ราว 0.06% ไว้ในความยาวของกระสุน

นี่คือหน้าตาอีกด้านของเส้นสัมผัสเดิม แทนที่จะเดินหลายรอบจากค่าเดาตั้งต้นหยาบ โค้ดใช้ค่าเดาตั้งต้นจากบิตที่ใกล้คำตอบพอ แล้วให้หนึ่งรอบทำงาน สาเหตุจริงของความแม่นคือความเร็วกำลังสอง ไม่ใช่โชคที่ค่าเดาตั้งต้นดี

ความเร็วกำลังสอง

ตัวเลขทุกชุดในหัวข้อนี้บอกเรื่องเดียวกัน ค่าคลาดเคลื่อนรอบถัดไปขึ้นกับ*กำลังสอง*ของค่าคลาดเคลื่อนรอบนี้ ทฤษฎีบทต่อไปนี้พูดเรื่องนี้อย่างแม่นยำ

นิยาม 6.2.1: การลู่เข้าแบบกำลังสอง

ลำดับ $x_n \to \alpha$ เรียกว่าลู่เข้า**แบบกำลังสอง** (quadratic convergence) ถ้ามีค่าคงที่ $C \geq 0$ และ N ที่ทำให้ $|x_{n+1} - \alpha| \leq C\,|x_n - \alpha|^2$ สำหรับทุก $n \geq N$ นี่คืออันดับ $p = 2$ ในนิยามของหัวข้อ 6.1

อ่านแบบเข้าใจง่าย. ถ้าค่าคลาดเคลื่อนปัจจุบันคือ 10^{-3} และ $C = 1$ รอบถัดไปจะผิดไม่เกิน 10^{-6} และรอบถัดไปอีกไม่เกิน 10^{-12} จำนวนหลักที่ถูกต้องเพิ่มเป็นสองเท่าทุกรอบ แต่กฎนี้ใช้ได้เมื่อค่าคลาดเคลื่อนเริ่มเล็กพอแล้วเท่านั้น

ทฤษฎีบท 6.2.1: ความลู่เข้าแบบกำลังสองของวิธีนิวตันในท้องถิ่น

ให้ f มีอนุพันธ์อันดับสองต่อเนื่องบนช่วงปิด $I = [\alpha - r, \alpha + r]$ โดย $f(\alpha) = 0$ และ $f'(x) \neq 0$ ทุก $x \in I$ ให้ $m = \min_I |f'| > 0$ และ $M = \max_I |f''|$ กำหนด $C = M/(2m)$ ถ้า $0 < \delta \leq r$ และ $C\delta < 1$ แล้ว สำหรับ x_0 ที่ทำให้ $|x_0 - \alpha| \leq \delta$ ลำดับนิวตัน $x_{n+1} = x_n - f(x_n)/f'(x_n)$ มีสมบัติต่อไปนี้

1. $|x_n - \alpha| \leq \delta$ และ $f'(x_n) \neq 0$ ทุก $n \geq 0$
2. $|x_{n+1} - \alpha| \leq C\,|x_n - \alpha|^2$ ทุก $n \geq 0$
3. $C|x_n - \alpha| \leq (C|x_0 - \alpha|)^{2^n} \leq (C\delta)^{2^n}$ ดังนั้น $x_n \to \alpha$ แบบกำลังสอง

พิสูจน์. **แนวคิด.** เขียน $x_{n+1} - \alpha$ เป็นกำลังสองของค่าคลาดเคลื่อนรอบก่อนด้วยสูตรของเทย์เลอร์ที่มีเศษเหลือแบบลากรานจ์ แล้วใช้ขอบเขตของ f' และ f'' บน I

1. **สูตรของเทย์เลอร์.** ทั้ง x_n และ α อยู่ใน I เมื่อ (1) เป็นจริงถึงรอบที่ n ดังนั้นมี ξ_n อยู่ระหว่างสองจุดนี้ที่ทำให้ $0 = f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \frac{1}{2}f''(\xi_n)(\alpha - x_n)^2$ จึงได้ $f(x_n) = f'(x_n)(x_n - \alpha) - \frac{1}{2}f''(\xi_n)(\alpha - x_n)^2$
2. **ค่าคลาดเคลื่อนรอบถัดไป.** แทนในสูตรนิวตันจะได้

$$x_{n+1} - \alpha = (x_n - \alpha) - \frac{f(x_n)}{f'(x_n)} = \frac{\frac{1}{2}f''(\xi_n)\,(\alpha - x_n)^2}{f'(x_n)}.$$

ฉบับร่าง Preprint 2026-10-23

ตารางที่ 6.6: นิวตันสำหรับ x^2-2 จาก $x_0=1.5$: ตัวคูณ e_{n+1}/e_n^2 เทียบกับ $1/(2x_n)$ (ตัวเลขจากโค้ดของเรา)

n	x_n	$\lvert x_n-\sqrt{2}\rvert$	e_{n+1}/e_n^2	$1/(2x_n)$
0	1.5	8.58×10^{-2}	0.333333	0.333333
1	1.4166667	2.45×10^{-3}	0.352941	0.352941
2	1.4142157	2.12×10^{-6}	(ดูข้อความ)	0.353553

3. **ขอบเขต.** เพราะ ξ_n และ x_n อยู่ใน I เราได้ $\lvert f''(\xi_n)\rvert\le M$ และ $\lvert f'(x_n)\rvert\ge m>0$ ดังนั้น $\lvert x_{n+1}-\alpha\rvert\le \frac{M}{2m}\lvert x_n-\alpha\rvert^2 = C\lvert x_n-\alpha\rvert^2$ ซึ่งคือ (2) เมื่อ (1) ถึงรอบ n แล้ว

4. **รักษาช่วง.** สมมติ $\lvert x_n-\alpha\rvert\le\delta$ จากข้อ 3 และ $C\delta<1$ จะได้ $\lvert x_{n+1}-\alpha\rvert\le C\delta\,\lvert x_n-\alpha\rvert\le\lvert x_n-\alpha\rvert\le\delta$ เริ่มจาก $\lvert x_0-\alpha\rvert\le\delta$ โดยอุปนัยได้ (1) ทุก n

5. **อัตรา.** ให้ $\eta_n=C\lvert x_n-\alpha\rvert\ge0$ จาก (2) จะได้ $\eta_{n+1}\le\eta_n^2$ โดยอุปนัย $\eta_n\le\eta_0^{2^n}$ เพราะ $\eta_{n+1}\le\eta_n^2\le(\eta_0^{2^n})^2$ และ $\eta_0=C\lvert x_0-\alpha\rvert\le C\delta<1$ ดังนั้น $\eta_n\to0$ และได้ (3)

สรุป: ทุกค่า x_n ยังอยู่ในช่วงที่เราควบคุมได้ และค่าคลาดเคลื่อนรอบถัดไปไม่เกินค่าคงที่คูณกำลังสองของค่าคลาดเคลื่อนรอบนี้ □

อ่านแบบเข้าใจง่าย

ทฤษฎีบทนี้บอกว่า ถ้าเริ่มใกล้รากพอ (ภายในระยะ δ ที่ $C\delta<1$) และ f' ไม่เป็นศูนย์ที่ราก วิธีนิวตันจะลู่เข้าเร็วมาก ตัวเลข δ ขึ้นกับ C ซึ่งขึ้นกับ f'' และ f' ใกล้ราก ไม่ได้บอกว่าต้องเริ่มใกล้แค่ไหนในทางปฏิบัติ หัวข้อ 6.3 จะพูดเรื่องนี้

ตัวอย่างที่คำนวณได้ทุกขั้น. ให้ $f(x)=x^2-2$ และ $\alpha=\sqrt{2}$ สูตรนิวตันคือ $x_{n+1}=\frac{x_n^2+2}{2x_n}$ เมื่อกระจายจะได้

$$x_{n+1}-\sqrt{2}=\frac{(x_n-\sqrt{2})^2}{2x_n}.$$

ดังนั้นค่าคลาดเคลื่อนรอบถัดไปคือ $\frac{1}{2x_n}$ คูณกำลังสองของค่าคลาดเคลื่อนรอบนี้ เมื่อ $x_n\to\sqrt{2}$ ตัวคูณนี้เข้าใกล้ $\frac{1}{2\sqrt{2}}\approx0.35355$ ซึ่งเป็นค่า $\frac{f''(\alpha)}{2f'(\alpha)}$ ที่ทฤษฎีบทให้ไว้ในรูปของ C

แถวที่ 0 และ 1 คอลัมน์ที่สี่กับคอลัมน์ที่ห้าเท่ากันตามที่ตัวอย่างบอก แถวที่ 2 ไม่คำนวณคอลัมน์ที่สี่ เพราะค่าคลาดเคลื่อน e_3 อยู่ใกล้ระดับการปัดเศษของเลขทศนิยม จึงให้อ่านค่าจากคอลัมน์ที่ห้าแทน

รากซ้ำ

ข้อควรระวัง: รากซ้ำทำให้ความเร็วกำลังสองหาย

ทฤษฎีบทต้องการ $f'(\alpha)\neq0$ ลองดู $f(x)=(x-1)^2$ ซึ่งมีรากซ้ำที่ 1 สูตรนิวตันคือ $x_{n+1}=x_n-\frac{(x_n-1)^2}{2(x_n-1)}=\frac{x_n+1}{2}$ ดังนั้น $x_{n+1}-1=\frac12(x_n-1)$ ค่าคลาดเคลื่อนลดลงครึ่งหนึ่งทุกรอบ เป็นการลู่เข้าแบบเชิงเส้น ไม่ใช่แบบกำลังสอง จากค่าเดาตั้งต้น 3 ได้ $3,2,1.5,1.25,\ldots$ ค่าคลาดเคลื่อนเป็น $2,1,\frac12,\frac14,\ldots$

เกร็ดเพิ่มเติม

ถ้า α เป็นราก**ที่มีภาวะรากซ้ำ** (multiplicity) p (คือ α ซ้ำกัน p ครั้ง เขียนได้เป็น $f(x)=(x-\alpha)^p g(x)$ โดย $g(\alpha)\neq0$) วิธีนิวตันลู่เข้าแบบเชิงเส้นด้วยอัตรา $\frac{p-1}{p}$ Atkinson 1989, §2.7, pp. 87--91 ตัวอย่างข้างบนเป็นกรณี

ฉบับร่าง · Preprint, 2026-10-23

ตารางที่ 6.7: วิธีเซแคนต์บนผนัง เริ่มที่ $x_0 = 6$ และ $x_1 = 6.5$ (ตัวเลขจากโค้ดของเรา)

n	x_n	$\lvert x_n - \alpha\rvert$	หลักที่ถูกต้อง
2	6.6290321664	3.15×10^{-3}	2.5
3	6.6321300016	5.14×10^{-5}	4.3
4	6.6321814257	2.32×10^{-8}	7.6
5	6.6321814489	1.71×10^{-13}	12.8

ตารางที่ 6.8: จำนวนรอบบนผนังเพื่อให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-6} (ตัวเลขจากโค้ดของเรา)

วิธี	จุดเริ่ม	รอบ	การรับประกัน
วิธีแบ่งครึ่ง	ช่วง $[0, 8]$	22	รับประกันจาก $(b-a)/2^{n+1}$
การทำซ้ำ $g_{1/2}$	$x_0 = 6$	15	จริง (ขอบเขตก่อนการคำนวณ 20)
วิธีนิวตัน	$x_0 = 6$	3	จริง
วิธีเซแคนต์	$x_0 = 6, x_1 = 6.5$	3	จริง

$p = 2$ ที่อัตรา $\frac{1}{2}$ ข้อพิสูจน์ทั่วไปอยู่ในหนังสืออ้างอิง

ถ้าไม่มีอนุพันธ์

ในบางโจทย์เราหาอนุพันธ์เป็นสูตรไม่ได้ เราจึงแทน $f'(x_n)$ ด้วย**ผลหารเชิงผลต่าง** (difference quotient) ของจุดสองจุดล่าสุด ได้สูตรของ**วิธีเซแคนต์** (secant method)

$$x_{n+1} = x_n - f(x_n)\,\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}.$$

เส้นตรงที่ผ่านสองจุดนี้เป็นเส้นตัดกราฟ ไม่ใช่เส้นสัมผัส วิธีนี้ไม่ต้องหาอนุพันธ์ แต่ลำดับของมันคือ $p = \frac{1+\sqrt{5}}{2} \approx 1.618$ ซึ่งต่ำกว่า 2 ของวิธีนิวตัน Atkinson 1989, §2.3, p. 69 เราใช้ผลนี้โดยไม่พิสูจน์ซ้ำ

สำหรับผนังของเรา เริ่มจาก $x_0 = 6$ และ $x_1 = 6.5$ วิธีเซแคนต์ให้ค่าตามตารางที่ 6.7

หลักที่ถูกต้องเพิ่มขึ้นประมาณ 1.7 เท่าในแต่ละรอบ ซึ่งสอดคล้องกับอันดับ 1.618 โดยประมาณ (อัตราส่วนของหลักไม่ตรงกับอันดับเป๊ะ เพราะมีค่าคงที่ปนอยู่) วิธีนิวตันได้อันดับ 2 แต่ต้องคำนวณ f' ทุกรอบ ส่วนวิธีเซแคนต์ต้องการค่า f ใหม่เพียงหนึ่งค่าต่อรอบ

ตารางที่ 6.8 สรุปจำนวนรอบที่แต่ละวิธีใช้บนผนังเพื่อให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-6} โดยเริ่มจากจุดเริ่มของวิธีนั้น

ตารางนี้ไม่ได้บอกว่าวิธีไหนดีที่สุดในทุกกรณี วิธีแบ่งครึ่งทนต่อการเริ่มต้นแบบไม่รู้ได้ดี แต่ช้า ส่วนวิธีที่ใช้อนุพันธ์หรือความชันเร็วมาก แต่ต้องเริ่มดี

สรุปใจความ

- วิธีนิวตันใช้เส้นสัมผัสชี้จากค่าเดาตั้งต้นไปจุดถัดไป สำหรับ $1/\sqrt{a}$ สูตร $y_{n+1} = y_n(3 - a y_n^2)/2$ ไม่มีรากที่สองและไม่มีการหาร
- ถ้าเริ่มใกล้ราก ที่ $f' \neq 0$ ค่าคลาดเคลื่อนรอบถัดไปไม่เกินค่าคงที่คูณกำลังสองของค่าคลาดเคลื่อนรอบนี้ หลักที่ถูกต้องเพิ่มเกือบเป็นสองเท่า แต่ต้องเริ่มใกล้พอ
- ฟังก์ชันแบบเดียวกับโค้ดของ id Software ทำค่าเดาตั้งต้นจากบิต (ผิดสูงสุดราว 3% บนช่วง $[1, 4]$) แล้วหนึ่ง

ฉบับร่าง Preprint · 2026-10-23

รอบ เหลือค่าคลาดเคลื่อนสัมพัทธ์สูงสุดราว 0.2% ค่าเดาตั้งต้นที่ดีจึงทำให้หนึ่งรอบพอใช้

- รากซ้ำทำให้ความเร็วกำลังสองหาย และวิธีเซแคนต์ไม่ต้องใช้ f' แต่ลำดับเป็น 1.618
- กลับไปที่กระสุน: $1/\sqrt{17/16} \approx 0.970143$ หาได้ทั้งด้วยวิธีนิวตันเต็มและด้วยรอบเดียวจากค่าเดาตั้งต้นที่ได้จากบิต

ข้อควรระวัง: การอ่านบิตผ่านตัวชี้

ในโค้ดต้นฉบับ บิตของค่าทศนิยมถูกอ่านผ่านตัวชี้ชนิด `long` ภาษา C มาตรฐานไม่รับรองการอ่านหน่วยความจำด้วยชนิดข้อมูลต่างจากชนิดที่เก็บ และขนาดของ `long` ขึ้นกับระบบ ในการทดลองของเราใช้จำนวนเต็ม 32 บิตที่ตรงกับบิตของ float32 ซึ่งเป็นความหมายที่โค้ดตั้งใจบนระบบที่ `long` มี 32 บิต

อ่านต่อ. บท 2 ของ Atkinson อธิบายวิธีนิวตันและวิธีเซแคนต์อย่างละเอียด Atkinson 1989, §2.2--2.3, pp. 58--72 และรากซ้ำที่ Atkinson 1989, §2.7, pp. 87--91 ส่วนรูปแบบเวกเตอร์ของวิธีนิวตัน (Newton--Raphson) สำหรับระบบสมการ ดู Iserles 2009, §7.2, pp. 127--129 หัวข้อ 6.3 จะบอกว่าค่าเดาตั้งต้นที่ไกลเกินไปทำให้นิวตันพาเราไปผิดจุดได้อย่างไร

ฉบับร่าง · Preprint · 2026-05-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 6.2.1 (⋆) ให้ $f(x) = x^2 - 2$ และ $x_0 = 1$

1. จงเขียนสูตรนิวตันของ f ให้ชัด แล้วหา x_1, x_2, x_3
2. จงหาค่าคลาดเคลื่อนสัมบูรณ์ $|x_n - \sqrt{2}|$ สำหรับ $n = 1, 2, 3$ และจำนวนหลักที่ถูกต้องของแต่ละรอบ
3. จงอธิบายว่าจำนวนหลักที่ถูกต้องเพิ่มประมาณเท่าไรในรอบที่ 2 และรอบที่ 3

แบบฝึกหัด 6.2.2 (⋆⋆) ให้ gap เป็นฟังก์ชันของผนังในหัวข้อ 6.1 และ $x_0 = 6.5$

1. จงหาเส้นสัมผัสของ gap ที่ $x_0 = 6.5$ โดยใช้ $\mathrm{gap}(6.5) \approx 0.159880$ และ $\mathrm{gap}'(6.5) \approx -1.226588$ (ตรวจได้จาก $\mathrm{gap}'(x) = -\frac{1}{4} - \cos x$)
2. จงหา x_1 และ x_2 ด้วยวิธีนิวตัน แล้วเทียบกับ $\alpha \approx 6.632181$
3. จงอธิบายว่าทำไมการเริ่มจาก 6.5 ให้ค่าคลาดเคลื่อน $x_1 - \alpha$ เล็กกว่าการเริ่มจาก 6 ในหัวข้อ 6.2 (ดูระยะห่างจาก α ของจุดเริ่ม และความชันที่จุดนั้น)

แบบฝึกหัด 6.2.3 (⋆⋆) ให้ $a > 0$ และ $f(y) = \dfrac{1}{y^2} - a$ สำหรับ $y > 0$

1. จงหา $f'(y)$ แล้วแสดงว่าสูตรนิวตันของ f คือ $y_{n+1} = \dfrac{y_n(3 - a y_n^2)}{2}$
2. จงคำนวณ y_1 เมื่อ $a = \frac{17}{16}$ และ $y_0 = 1$ แล้วเทียบกับ $1/\sqrt{17/16} \approx 0.9701425$
3. ให้ $u_n = a y_n^2$ และ $\delta_n = u_n - 1$ จงแสดงว่า $\delta_{n+1} = -\frac{3}{4}\delta_n^2 + \frac{1}{4}\delta_n^3$

แบบฝึกหัด 6.2.4 (⋆⋆) ให้ $f(x) = x^2 - 2$ และ $\alpha = \sqrt{2}$

1. จงแสดงว่า $x_{n+1} - \alpha = \dfrac{(x_n - \alpha)^2}{2x_n}$ ทุก n เมื่อ $x_n \neq 0$
2. จากค่าเดาตั้งต้น $x_0 = 1.5$ จงหา $\dfrac{x_1 - \alpha}{(x_0 - \alpha)^2}$ และ $\dfrac{x_2 - \alpha}{(x_1 - \alpha)^2}$ แล้วเทียบกับ $\dfrac{1}{2\sqrt{2}} \approx 0.353553$
3. จงอธิบายว่าทำไม $\dfrac{1}{2x_n}$ เข้าใกล้ $\dfrac{1}{2\sqrt{2}}$ เมื่อ n มาก

แบบฝึกหัด 6.2.5 (★★) ให้ $f(x) = (x-1)^2$ ซึ่งมีรากซ้ำที่ $\alpha = 1$

1. จงแสดงว่า $f'(1) = 0$ จึงไม่เข้าเงื่อนไข $f'(\alpha) \neq 0$ ของทฤษฎีบทในหัวข้อนี้
2. จงแสดงว่าสูตรนิวตันคือ $x_{n+1} = \dfrac{x_n + 1}{2}$ และ $x_{n+1} - 1 = \frac{1}{2}(x_n - 1)$ เมื่อ $x_n \neq 1$
3. จากค่าเดาตั้งต้น $x_0 = 3$ จงหา x_1, x_2, x_3 และอัตราส่วนค่าคลาดเคลื่อนของแต่ละรอบ

แบบฝึกหัด 6.2.6 (★★★)

ตัวช่วยคิด

ในสูตรนิวตัน แทน $f'(x_n)$ ด้วยความชันของเส้นตรงที่ผ่าน $(x_{n-1}, f(x_{n-1}))$ และ $(x_n, f(x_n))$

ให้ $g = \text{gap}$ เริ่มจาก $x_0 = 6$ และ $x_1 = 6.5$

1. จงแสดงว่าสูตรเซแคนต์ $x_{n+1} = x_n - g(x_n)\dfrac{x_n - x_{n-1}}{g(x_n) - g(x_{n-1})}$ มาจากวิธีนิวตัน
2. จงหา x_2 และ x_3 และคำนวณค่าคลาดเคลื่อน $e_n = x_n - \alpha$ สำหรับ $n \leq 3$
3. จงคำนวณ $K_3 = e_4/(e_3 e_2)$ และ $K_4 = e_5/(e_4 e_3)$ โดยใช้จุด x_4, x_5 จากโค้ดของเรา แล้วสังเกตว่าทั้งสองค่าใกล้ -0.14

แบบฝึกหัด 6.2.7 (★★) ให้ $x = 4$ และให้ฟังก์ชันแบบ `Q_rsqrt` ทำกับ x นี้ (ดูหัวข้อ 6.2 ส่วน "อีกด้านหนึ่งของปัญหาเดียวกัน")

1. จงเขียน 4 เป็นบิตของเลขทศนิยม 32 บิต ซึ่งได้ `0x40800000` จากนั้นคำนวณ `0x5f3759df` ลบด้วยบิตของ 4 ที่เลื่อนขวาหนึ่งตำแหน่ง ได้ `0x3ef759df` จงหาค่าเดาตั้งต้น y_0 ที่บิตนี้แทน (เลขชี้กำลังคือ 125 จึงได้ 2^{-2} และส่วนเศษคือ $\texttt{0x7759df}/2^{23}$)
2. จงคำนวณ $y_1 = y_0\,(1.5 - 2y_0^2)$ และเทียบกับ $1/\sqrt{4} = 0.5$
3. จงหาค่าคลาดเคลื่อนสัมพัทธ์ ε_0 และ ε_1 แล้วตรวจว่า $\varepsilon_1 \approx -\frac{3}{2}\varepsilon_0^2$

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 6.2.8 (⋆⋆⋆)

ตัวช่วยคิด

ค่าคลาดเคลื่อนสัมพัทธ์หลังหนึ่งรอบประมาณ $\frac{3}{2}$ คูณกำลังสองของค่าคลาดเคลื่อนสัมพัทธ์ของค่าเดาตั้งต้น (ดูหัวข้อ 6.2 ส่วน "อีกด้านหนึ่งของปัญหาเดียวกัน")

ให้ $M' =$ `0x5f3659df` ซึ่งต่างจากค่าคงที่ของโค้ดต้นฉบับ `0x5f3759df` อยู่ 2^{16} ใช้ฟังก์ชันแบบ `Q_rsqrt` โดยเปลี่ยนค่าคงที่เป็น M'

1. จงหาค่าคลาดเคลื่อนสัมพัทธ์ของค่าเดาตั้งต้น และของหนึ่งรอบที่ $x = 4$ เมื่อใช้ M' แล้วเทียบกับค่าเดิม
2. จงหาค่าคลาดเคลื่อนสัมพัทธ์สูงสุดบนตารางจุด $[1, 4]$ หลังหนึ่งรอบของทั้งสองค่าคงที่ แล้วบอกว่าค่าใดแม่นกว่า
3. จงอธิบายด้วยกฎ $\varepsilon_{n+1} \approx -\frac{3}{2}\varepsilon_n^2$ ว่าทำไมค่าคงที่ที่ทำให้ค่าเดาตั้งต้นผิดมากขึ้นจึงทำให้ผลหนึ่งรอบผิดมากขึ้นด้วย

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 6.2.1.

1. $f'(x) = 2x$ จึงได้ $x_{n+1} = x_n - \dfrac{x_n^2 - 2}{2x_n}$ จาก $x_0 = 1$: $x_1 = 1 - \frac{1-2}{2} = \frac{3}{2}$ และ $x_2 = \frac{3}{2} - \frac{\frac{9}{4}-2}{3} = \frac{17}{12} \approx 1.4166667$ และ $x_3 = \frac{17}{12} - \frac{1/144}{17/6} = \frac{577}{408} \approx 1.4142157$
2. $\sqrt{2} \approx 1.4142136$ จึง $|x_1 - \sqrt{2}| \approx 0.0857864$ (หลักที่ถูก ≈ 1.07), $|x_2 - \sqrt{2}| \approx 0.0024531$ (หลักที่ถูก ≈ 2.61) และ $|x_3 - \sqrt{2}| \approx 2.12 \times 10^{-6}$ (หลักที่ถูก ≈ 5.67)
3. รอบที่ 2 เพิ่มประมาณ $2.61 - 1.07 \approx 1.5$ หลัก และรอบที่ 3 เพิ่มประมาณ $5.67 - 2.61 \approx 3.1$ หลัก ซึ่งเพิ่มมากขึ้นตามความเร็วกำลังสอง
4. ตรวจ x_2 ทีละขั้นได้ว่า $x_1^2 = \frac{9}{4}$ จึง $x_1^2 - 2 = \frac{1}{4}$ และ $\dfrac{x_1^2 - 2}{2x_1} = \dfrac{1/4}{3} = \dfrac{1}{12}$ ดังนั้น $x_2 = \frac{3}{2} - \frac{1}{12} = \frac{17}{12}$ ซึ่งตรงกับข้อ 1

ตรวจด้วยการคำนวณ: test_ex_6_2_1 ในไฟล์ทดสอบ.

เฉลย 6.2.2.

1. เส้นสัมผัสที่ $x_0 = 6.5$ เป็นเส้นตรง $\ell(x) = \mathrm{gap}(6.5) + \mathrm{gap}'(6.5)(x - 6.5) = 0.159880 - 1.226588\,(x - 6.5)$ ความชันเป็นลบ จึงลงจากจุด $(6.5, 0.159880)$ ไปตัดแกนนอนทางขวา
2. ตั้ง $\ell(x_1) = 0$ จะได้ $x_1 = 6.5 + \dfrac{0.159880}{1.226588}$ และ $\dfrac{0.159880}{1.226588} \approx 0.130345$ จึง $x_1 \approx 6.630345$ ค่าคลาดเคลื่อน $x_1 - \alpha \approx -1.836 \times 10^{-3}$ รอบที่สองใช้ $\mathrm{gap}(x_1) \approx 0.0021850$ และ $\mathrm{gap}'(x_1) \approx -1.190343$ จึงได้ $x_2 = x_1 - \dfrac{0.0021850}{-1.190343} \approx 6.6321810$ ค่าคลาดเคลื่อน $x_2 - \alpha \approx -4.83 \times 10^{-7}$ ตามโค้ดของเรา
3. การเริ่มจาก 6.5 อยู่ใกล้ α มากกว่า (ระยะ 0.132 แทน 0.632) ค่าคลาดเคลื่อนของรอบแรกจึงประมาณเป็นกำลังสองของระยะนี้ซึ่งเล็กกว่ามาก ความชันของผนังที่จุดนั้นก็ใกล้ค่าที่ราก จึงไม่ทำให้เส้นสัมผัสไปไกลจากรากนัก

ตรวจด้วยการคำนวณ: test_ex_6_2_2 ในไฟล์ทดสอบ.

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 6.2.3.

1. $f'(y) = -\dfrac{2}{y^3}$ จึง $y_{n+1} = y_n - \dfrac{1/y_n^2 - a}{-2/y_n^3} = y_n + \dfrac{y_n}{2} - \dfrac{ay_n^3}{2} = \dfrac{y_n(3 - ay_n^2)}{2}$

2. เมื่อ $a = \frac{17}{16}$ และ $y_0 = 1$ จะได้ $y_1 = \dfrac{1 \cdot (3 - \frac{17}{16})}{2} = \dfrac{31}{32} = 0.96875$ เทียบกับ 0.9701425 ค่าคลาดเคลื่อนคือ -1.39×10^{-3}

3. ให้ $u = ay^2$ จะได้ $u_{n+1} = u_n(3 - u_n)^2/4$ เมื่อ $u_n = 1 + \delta_n$ จะได้ $(1+\delta)(2-\delta)^2/4 = (1+\delta)(4 - 4\delta + \delta^2)/4 = 1 - \frac{3}{4}\delta^2 + \frac{1}{4}\delta^3$ ดังนั้น $\delta_{n+1} = -\frac{3}{4}\delta_n^2 + \frac{1}{4}\delta_n^3$

4. ตรวจรอบที่สองด้วยตัวเลข: จาก $y_1 = 0.96875$ ได้ $y_1^2 \approx 0.9384766$ และ $ay_1^2 \approx 0.9971314$ จึง $y_2 = \dfrac{0.96875\,(3 - 0.9971314)}{2} \approx 0.9701395$ ซึ่งมีค่าคลาดเคลื่อน $\approx -3.0 \times 10^{-6}$ ตรงกับตารางที่ 6.4

ตรวจด้วยการคำนวณ: test_ex_6_2_3 ในไฟล์ทดสอบ.

เฉลย 6.2.4.

1. $x_{n+1} = x_n - \dfrac{x_n^2 - 2}{2x_n} = \dfrac{x_n^2 + 2}{2x_n}$ ดังนั้น $x_{n+1} - \sqrt{2} = \dfrac{x_n^2 - 2\sqrt{2}x_n + 2}{2x_n} = \dfrac{(x_n - \sqrt{2})^2}{2x_n}$ เพราะ $(x - \sqrt{2})^2 = x^2 - 2\sqrt{2}x + 2$

2. จาก $x_0 = 1.5$: $\dfrac{x_1 - \alpha}{(x_0 - \alpha)^2} = \dfrac{1}{2 \cdot 1.5} = \dfrac{1}{3}$ (ตรงกับ $0.0024531/0.0857864^2 \approx 0.3333$) และ $\dfrac{x_2 - \alpha}{(x_1 - \alpha)^2} = \dfrac{1}{2x_1} = \dfrac{1}{2 \cdot \frac{17}{12}} = \dfrac{6}{17} \approx 0.352941$ เทียบกับ $\frac{1}{2\sqrt{2}} \approx 0.353553$ ต่างกันน้อยกว่า 0.2%

3. เมื่อ n มาก $x_n \to \sqrt{2}$ ดังนั้น $2x_n \to 2\sqrt{2}$ และ $\frac{1}{2x_n} \to \frac{1}{2\sqrt{2}}$ นี่คือค่า $\frac{f''(\alpha)}{2f'(\alpha)}$ สำหรับ $f(x) = x^2 - 2$ ซึ่งเป็นค่าคงที่ของทฤษฎีบท

4. ตรวจค่าแรกด้วยมือ: $x_1 - \alpha = 1.4166667 - 1.4142136 = 0.0024531$ และ $\dfrac{(x_0 - \alpha)^2}{2x_0} = \dfrac{0.0857864^2}{3} = \dfrac{0.0073593}{3} = 0.0024531$ ตรงกัน

ตรวจด้วยการคำนวณ: test_ex_6_2_4 ในไฟล์ทดสอบ.

เฉลย 6.2.5.

1. $f'(x) = 2(x-1)$ จึง $f'(1) = 0$ ทฤษฎีบทในหัวข้อนี้ต้องการ $f' \neq 0$ บนช่วงที่มีราก ดังนั้นไม่เข้าเงื่อนไข

2. $x_{n+1} = x_n - \dfrac{(x_n - 1)^2}{2(x_n - 1)} = x_n - \dfrac{x_n - 1}{2} = \dfrac{x_n + 1}{2}$ และ $x_{n+1} - 1 = \dfrac{x_n - 1}{2}$ เมื่อ $x_n \neq 1$

3. $x_0 = 3$ ได้ $x_1 = 2$, $x_2 = \frac{3}{2}$, $x_3 = \frac{5}{4}$ ค่าคลาดเคลื่อนคือ $2, 1, \frac{1}{2}, \frac{1}{4}$ อัตราส่วนค่าคลาดเคลื่อนของแต่ละรอบคือ $\frac{1}{2}$ ทุกรอบ ซึ่งเป็นการลู่เข้าแบบเชิงเส้น

4. ตรวจด้วยสูตรข้อ 2: $x_1 = \frac{3+1}{2} = 2$, $x_2 = \frac{2+1}{2} = \frac{3}{2}$ และ $x_3 = \frac{\frac{3}{2}+1}{2} = \frac{5}{4}$ ตรงกับการคำนวณข้อ 3

ตรวจด้วยการคำนวณ: test_ex_6_2_5 ในไฟล์ทดสอบ.

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 6.2.6.

1. แทน $f'(x_n)$ ด้วยความชันของเส้นตรงผ่านสองจุด $\dfrac{g(x_n)-g(x_{n-1})}{x_n-x_{n-1}}$ จะได้ $x_{n+1} = x_n - g(x_n)\cdot \dfrac{x_n-x_{n-1}}{g(x_n)-g(x_{n-1})}$ ซึ่งคือสูตรเซแคนต์ เหตุผลคือ ผลหารเชิงผลต่างเป็นความชันของเส้นตรงที่ตัดกราฟที่จุดสองจุด ซึ่งประมาณความชันของเส้นสัมผัสที่จุดใดจุดหนึ่งในช่วงนั้น

2. จาก $x_0 = 6$, $x_1 = 6.5$ ได้ $\operatorname{gap}(6) \approx 0.779415$ และ $\operatorname{gap}(6.5) \approx 0.159880$ จึงได้

$$x_2 = 6.5 - \frac{0.159880\cdot(6.5-6)}{0.159880-0.779415} = 6.5 + 0.129032 \approx 6.6290322$$

ค่าคลาดเคลื่อน $e_2 \approx -3.149\times10^{-3}$ ต่อไป $\operatorname{gap}(x_2) \approx 0.0037484$ และ $\operatorname{gap}(x_1) \approx 0.159880$ จึง $x_3 = x_2 - \dfrac{0.0037484\cdot(x_2-x_1)}{0.0037484-0.159880} \approx 6.6321300$ ค่าคลาดเคลื่อน $e_3 \approx -5.145\times10^{-5}$

3. จากโค้ดของเรา $x_4 \approx 6.6321814$ ค่าคลาดเคลื่อน $e_4 \approx -2.321\times10^{-8}$ และ x_5 ค่าคลาดเคลื่อน $e_5 \approx -1.714\times10^{-13}$ จึงได้ $K_3 = \dfrac{e_4}{e_3e_2} \approx -0.1432$ และ $K_4 = \dfrac{e_5}{e_4e_3} \approx -0.1436$ สองค่านี้ใกล้กัน บ่งว่า $e_{n+1} \approx K\,e_ne_{n-1}$ โดยค่า K ควรเข้าใกล้ $\dfrac{f''(\alpha)}{2f'(\alpha)}$ ซึ่งเป็นสมการค่าคลาดเคลื่อนของวิธีเซแคนต์ที่ Atkinson ให้ไว้ (สมการ 2.3.11 หน้า 69) สำหรับผนังของเราค่านี้คือ $\dfrac{\sin\alpha}{2\operatorname{gap}'(\alpha)} \approx -0.14371$ ซึ่งเป็นค่าที่ K_3 และ K_4 เข้าใกล้ (ดูการตรวจในโค้ด)

ตรวจด้วยการคำนวณ: test_ex_6_2_6 ในไฟล์ทดสอบ.

เฉลย 6.2.7.

1. $4 = 2^2$ เลขชี้กำลังที่เก็บคือ $2+127 = 129$ ซึ่งเป็น `0x81` ส่วนเศษเป็นศูนย์ บิตจึงเป็น `0x40800000` เมื่อเลื่อนบิตนี้ไปทางขวาหนึ่งตำแหน่งจะได้ `0x20400000` จากนั้น `0x5f3759df` ลบด้วย `0x20400000` ได้ `0x3ef759df` เลขชี้กำลังของบิตนี้คือ `0x7d` ซึ่งเท่ากับ 125 จึงได้ $2^{125-127} = \frac{1}{4}$ ส่วนเศษคือ $7821791/8388608 \approx 0.932430$ ดังนั้น $y_0 = \frac{1}{4}(1+0.932430) \approx 0.483108$ (ค่าจากโค้ดของเราคือ 0.4831075)

2. $y_0^2 \approx 0.2333928$ จึง $1.5 - 2y_0^2 \approx 1.0332143$ และ $y_1 \approx 0.483108\times1.0332143 \approx 0.499154$ เทียบกับ $\frac{1}{2}$ ค่าคลาดเคลื่อนคือ -8.5×10^{-4}

3. $\varepsilon_0 = \dfrac{y_0-\frac{1}{2}}{\frac{1}{2}} \approx -0.033785$ และ $\varepsilon_1 = \dfrac{y_1-\frac{1}{2}}{\frac{1}{2}} \approx -0.0016928$ ส่วน $-\frac{3}{2}\varepsilon_0^2 \approx -1.712\times10^{-3}$ ซึ่งใกล้ ε_1 ต่างประมาณ 1% จึงสอดคล้องกับกฎ $\varepsilon_{n+1} \approx -\frac{3}{2}\varepsilon_n^2$ ในหัวข้อ 6.2

ตรวจด้วยการคำนวณ: test_ex_6_2_7 ในไฟล์ทดสอบ.

เฉลย 6.2.8.

1. เมื่อ $x = 4$ บิตของ 4 คือ `0x40800000` เลื่อนขวาหนึ่งตำแหน่งได้ `0x20400000` จากนั้น `0x5f3659df` $-$ `0x20400000` $=$ `0x3ef659df` เลขชี้กำลังยังเป็น 125 จึงได้ $\frac{1}{4}$ และส่วนเศษคือ `0x7659df` $= 7756255$ ส่วนเศษเป็น $7756255/8388608 \approx 0.924618$ ดังนั้น $y_0' = \frac{1}{4}(1+0.924618) \approx 0.481154$ ซึ่งต่างจาก $y_0 \approx 0.483108$ อยู่ -0.001954 และมีค่าคลาดเคลื่อนสัมพัทธ์ -3.769×10^{-2} (เดิม -3.378×10^{-2}) หลังหนึ่งรอบ $y_1' = y_0'(1.5-2y_0'^2) \approx 0.498948$ มีค่าคลาดเคลื่อนสัมพัทธ์ -2.104×10^{-3} (เดิม -1.693×10^{-3})

2. บนตารางจุด 3001 จุดใน $[1,4]$ ค่าคลาดเคลื่อนสัมพัทธ์สูงสุดของค่าเดาตั้งต้นเปลี่ยนจาก 3.437×10^{-2} เป็น 3.842×10^{-2} และหลังหนึ่งรอบเปลี่ยนจาก 1.752×10^{-3} เป็น 2.186×10^{-3} ค่าคงที่เดิมจึงแม่นกว่า

ฉบับร่าง preprint · 2026-10-23

3. ค่าเดาตั้งต้นที่ผิดมากขึ้นทำให้ ε_0 ใหญ่ขึ้น และกฎ $\varepsilon_1 \approx -\frac{3}{2}\varepsilon_0^2$ บอกว่าค่าคลาดเคลื่อนสัมพัทธ์หลังหนึ่งรอบเป็นกำลังสองของค่านี้ ที่ $\varepsilon_0 \approx 0.0384$ ได้ $\frac{3}{2}(0.0384)^2 \approx 2.2 \times 10^{-3}$ ซึ่งใกล้ค่าคลาดเคลื่อนสัมพัทธ์สูงสุดหลังหนึ่งรอบที่วัดได้ 2.19×10^{-3} (ค่าสูงสุดทั้งสองเกิดที่จุดต่างกันบนตาราง จึงตรงกันโดยประมาณ)

ตรวจด้วยการคำนวณ: test_ex_6_2_8 ในไฟล์ทดสอบ.

6.3 เมื่อไม่รู้จะเริ่มตรงไหน

สถานการณ์สมมติ: ค่าเดาตั้งต้นจากเฟรมก่อน

สมมติว่าเกมเก็บจุดชนของเฟรมก่อนไว้ แล้วใช้เป็นค่าเดาตั้งต้นของการทำซ้ำในเฟรมนี้ ถ้าผนังขยับไม่มาก จุดชนใหม่ควรอยู่ใกล้จุดเดิม และวิธีของนิวตันเพียงไม่กี่รอบก็พอ ปัญหาคือ ค่าเดาตั้งต้นบางค่าอยู่ใกล้จุดที่ผนังแบนมาก ก้าวแรกของนิวตันจึงพากระสุนไปไกลมาก

ลองให้ค่าเดาตั้งต้นเป็น $x_0 = 4.4$ บนผนังของเรา ก้าวแรกของนิวตันจะไปที่ $x_1 \approx -27.9$ ซึ่งอยู่ข้างหลังจุดเริ่มของกระสุนที่ $x = 0$ ผลสุดท้ายคือวิธีไปลงที่ $x \approx 9.928$ ซึ่งเป็นจุดชนที่สอง ไม่ใช่จุดชนแรกที่ 6.632

คำถามคือ ค่าเดาตั้งต้นแบบไหนพาเราไปถึงจุดชนแรก เรารู้ได้อย่างไรว่าควรหยุดเมื่อไร และมีวิธีที่ไม่ปล่อยให้ก้าวเดียวพากระสุนหลุดไปหรือไม่ (ตัวเลขทั้งหมดในหัวข้อนี้มาจากโค้ดของเรา และเป็นสถานการณ์สมมติ)

ภาพรวมก่อนอ่าน

ค่าเดาตั้งต้นกำหนดว่าวิธีจะไปถึงรากไหน หรือไม่ถึงรากเลย ก่อนเลือกวิธีแก้ เราจึงดูภาพรวมของค่าเดาตั้งต้นทั้งหมดก่อน แล้วดูเครื่องมือสามอย่างที่ช่วยลดความเสี่ยง ได้แก่ นิวตันแบบหน่วงก้าว การเร่งแบบไอทเคน และวิธีผสมที่เก็บช่วงล้อมไว้เสมอ เรื่องเกณฑ์การหยุดก็สำคัญพอกัน เพราะตัวเลขที่ดูเล็กอาจหลอกเราได้

แอ่งดึงดูด: จุดเริ่มไหนไปถึงรากไหน

ถ้ามองย้อนกลับ เราถามได้ว่าค่าเดาตั้งต้นชุดไหนบ้างที่ทำให้วิธีเดียวกันลงที่รากเดียวกัน คำตอบนี้มีชื่อว่า

นิยาม 6.3.1: แอ่งดึงดูด

ให้ α เป็นราก และให้วิธีทำซ้ำ $x_{n+1} = \Phi(x_n)$ **แอ่งดึงดูด** (basin of attraction) ของ α คือเซตของค่าเดาตั้งต้น x_0 ทั้งหมดที่ทำให้ $x_n \to \alpha$ (แนวคิดจาก Iserles 2009, §7.3, p. 131)

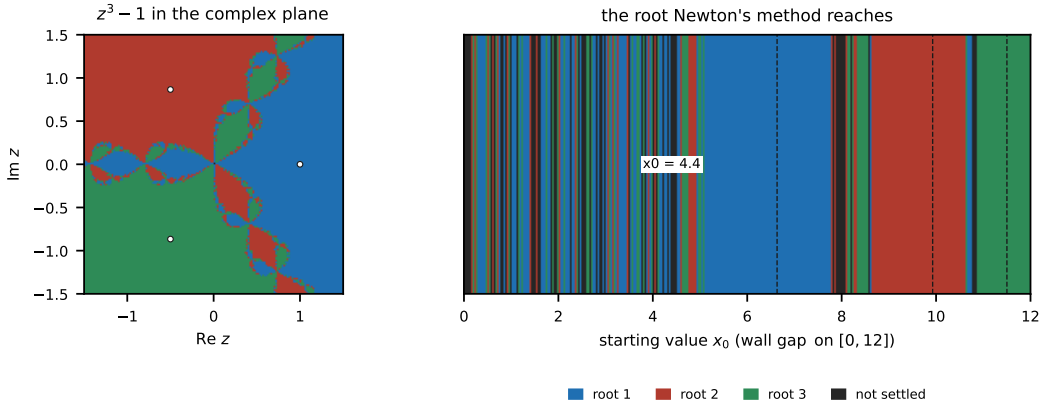
อ่านแบบเข้าใจง่าย. แอ่งดึงดูดเหมือนพื้นที่รับของรากแต่ละราก จุดเริ่มที่อยู่ในพื้นที่ของ α จะไปถึง α จุดเริ่มที่อยู่ในพื้นที่ของรากอื่นจะไปถึงรากนั้น และจุดที่ไม่อยู่ในพื้นที่ใดเลยอาจวนไปมาหรือหลุดออกไป

รูปต่อไปนี้แสดงแอ่งของวิธีนิวตันสองแบบ ทางขวาคือผนังของเราบน $[0, 12]$ ซึ่งมีสามราก จุดเริ่มทุกจุด x_0 ถูกระบายสีตามรากที่นิวตันไปถึง ทางซ้ายเป็นอีกตัวอย่างหนึ่ง คือพหุนาม $z^3 - 1$ บนระนาบจำนวนเชิงซ้อน ซึ่งมีรูปแบบซับซ้อนกว่า และใช้เพียงให้เห็นภาพ ไม่ใช่เนื้อหาหลักของบท

ในแถบขวาจะเห็นว่าสีสลับกันไม่เรียบร้อยโดยเฉพาะช่วง x_0 ระหว่าง 0 ถึง 5 จุดสองจุดที่อยู่ใกล้กันมากอาจไปคนละราก ตารางต่อไปนี้ให้ตัวอย่างบนผนังของเรา

ค่า $x_0 = 4$ กับ $x_0 = 4.4$ ห่างกันเพียง 0.4 แต่ไปคนละราก ส่วน $x_0 = 3$ กับ $x_0 = 4$ ไปรากเดียวกัน แต่ $x_0 = 3$ ต้องใช้ 28 รอบ ส่วน $x_0 = 4$ ใช้เพียง 6 รอบ จุดเริ่มที่ดูดีไม่ได้รับประกันว่าจะดีที่สุด

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 6.8: แอ่งดึงดูดของวิธีนิวตัน ทางซ้าย: จุดเริ่มเชิงซ้อนของ $z^3 - 1$ สีแสดงรากสามราก ทางขวา: จุดเริ่ม x_0 บน $[0, 12]$ ของผนัง สีแสดงรากที่วิธีไปถึง เส้นประคือรากทั้งสาม สีดำคือจุดที่ไม่ลู่เข้า (ทางซ้ายภายใน 60 รอบ ทางขวาภายใน 50 รอบ) (สร้างจาก `numanal/ch06.py`)

ตารางที่ 6.9: จุดเริ่มบนผนังและรากที่นิวตันไปถึง รอบคือจำนวนรอบจนค่า $|\,\text{gap}\,|$ ต่ำกว่า 10^{-10} (ตัวเลขจากโค้ดของเรา)

x_0	x_1	ราก	รอบ	จุดชนแรก
2	-1.555304	11.500616	6	ไม่
3	1.501498	6.632181	28	ใช่
4	-0.352360	6.632181	6	ใช่
4.4	-27.895646	9.927600	12	ไม่
6	6.644054	6.632181	3	ใช่
9	10.001495	9.927600	4	ไม่

นิวตันแบบหน่วง: ก้าวเล็กลงเมื่อก้าวเต็มทำให้แย่ลง

ถ้าก้าวเต็มของนิวตันทำให้ค่า $|f|$ โตขึ้น เราก็ลดก้าวลงครึ่งหนึ่งแล้วลองใหม่ ทำซ้ำจนค่าลดลงจริง วิธีนี้เรียกว่า **นิวตันแบบหน่วง** (damped Newton's method) ชื่อนี้เราตั้งเองโดยเทียบกับการสั่นแบบหน่วงในฟิสิกส์ คือการลดขนาดการเคลื่อนที่ลงทุกรอบ

$$x_{n+1} = x_n + t_n\, d_n, \qquad d_n = -\frac{f(x_n)}{f'(x_n)}, \tag{6.4}$$

โดย t_n เป็น 1 หรือ $\frac{1}{2}, \frac{1}{4}, \ldots$ ตัวแรกที่ทำให้ $|f(x_{n+1})| < |f(x_n)|$

ก่อนใช้ เราต้องรู้ว่าการลดก้าวจะหยุดได้เสมอหรือไม่ ทฤษฎีบทต่อไปนี้ตอบว่า ถ้าอนุพันธ์ไม่เป็นศูนย์ ก้าวสั้นพอจะลด $|f|$ ได้เสมอ

ทฤษฎีบท 6.3.1: ทิศลงของนิวตัน

ให้ f หาอนุพันธ์ได้ที่ x โดย $f(x) \neq 0$ และ $f'(x) \neq 0$ ให้ $d = -f(x)/f'(x)$ จะมี $\bar{t} > 0$ ที่ทำให้ $|f(x+td)| < |f(x)|$ สำหรับทุก $0 < t < \bar{t}$

พิสูจน์. **แนวคิด.** ดูฟังก์ชัน $\varphi(t) = f(x + td)^2$ ที่ $t = 0$ ว่าความชันเป็นลบ แล้วใช้นิยามของอนุพันธ์

1. **ความชันที่ศูนย์.** $\varphi'(0) = 2f(x)f'(x)\,d = 2f(x)f'(x)\big(-\frac{f(x)}{f'(x)}\big) = -2f(x)^2 < 0$ ค่านี้ใช้ได้เพราะ $f'(x) \neq 0$

ฉบับร่าง Preprint, 2026-10-23

ตารางที่ 6.10: นิวตันแบบหน่วงจาก $x_0 = 4.4$ บน $\mathrm{gap}(x) = 0$ (t คือตัวคูณที่พาไปถึง x_n) (ตัวเลขจากโค้ดของเรา)

n	x_n	t	$\lvert\mathrm{gap}(x_n)\rvert$
0	4.4	--	1.851602
1	0.363044	$\frac{1}{8}$	1.554117
2	1.674735	1	0.586713
3	1.925469	$\frac{1}{16}$	0.580873
4	1.738878	$\frac{1}{32}$	0.579373
5	1.848331	$\frac{1}{64}$	0.576183

2. **ลดลงเมื่อ t เล็ก.** จากนิยามของอนุพันธ์ $\varphi(t) - \varphi(0) = t\big(\varphi'(0) + o(1)\big)$ เมื่อ $t \to 0$ ดังนั้นมี $\bar{t} > 0$ ที่ทำให้ $\varphi(t) < \varphi(0)$ สำหรับ $0 < t < \bar{t}$
3. **สรุป.** $\varphi(t) < \varphi(0)$ หมายความว่า $f(x + td)^2 < f(x)^2$ จึง $|f(x + td)| < |f(x)|$

สรุป: ทิศของนิวตันเป็นทิศลงของ f^2 เสมอ เมื่อ $f' \neq 0$ ก้าวที่สั้นพอจึงลด $|f|$ ได้เสมอ □

อ่านแบบเข้าใจง่าย

ทฤษฎีบทนี้บอกว่า นิวตันมีทิศที่พาเราลงสู่ค่าต่ำของ $|f|$ ตราบที่ f' ไม่เป็นศูนย์ แต่ไม่ได้บอกว่าก้าวควรยาวเท่าไร การลดครึ่งเรื่อย ๆ จะพบก้าวที่ใช้ได้ภายในจำนวนครั้งจำกัดเสมอ ผลคือวิธีลดค่า $|f|$ ได้ทุกรอบ แต่ไม่รับประกันว่าจะไปถึงรากแรกที่ต้องการ

ในโค้ดของเรา การลดครึ่งทำได้ไม่เกิน 40 ครั้ง คือลองตัวคูณ $t = 1, \frac{1}{2}, \ldots, 2^{-39}$ (ตัวเล็กสุดราว 1.8×10^{-12}) ถ้าไม่มีตัวคูณใดในนี้ที่ทำให้ $|f|$ ลดลง จุดจะไม่ขยับ ($t = 0$) และการทำซ้ำหยุด ทฤษฎีบทบอกเพียงว่ามี $\bar{t} > 0$ จึงไม่ได้รับประกันว่าจะพบภายใน 40 ครั้ง เช่น $f(x) = \arctan x$ ที่ $x_0 = 10^{13}$ ต้องใช้ $t = 2^{-43}$ ซึ่งเกินขอบเขตนี้ และถ้าเพิ่มจำนวนครั้งเป็น 60 ก้าวนั้นจึงถูกรับ ค่า $t = 0$ ยังเกิดเมื่อ $f'(x) = 0$ หรือ $f(x) = 0$ ด้วย

ตารางต่อไปนี้ทำนิวตันแบบหน่วงจาก $x_0 = 4.4$ บนผนังของเรา

ค่า $|\,\mathrm{gap}\,|$ ลดลงทุกรอบจริง แต่ความคืบหน้าช้าลงมากเมื่อ x อยู่ใกล้ 1.82 ที่ $x = \arccos(-\frac{1}{4}) \approx 1.823477$ อนุพันธ์ gap' เป็นศูนย์ และค่า gap ที่จุดนี้คือ 0.575885 ซึ่งไม่ใช่ศูนย์ จึงไม่ใช่ราก การทำซ้ำติดอยู่แถวนี้ราวหกรอบ จนในที่สุดหลุดไปที่ $x \approx 10.88$ ด้วยตัวคูณ $\frac{1}{512}$ แล้วลงที่ 11.500616 ภายใน 14 รอบ ซึ่งเป็นรากที่สาม ไม่ใช่รากแรก

ข้อควรระวัง: ค่า $|f|$ ที่เล็กลงไม่ได้หมายความว่าใกล้ราก

นิวตันแบบหน่วงลด $|f|$ ทุกรอบ แต่ถ้าไปถึงจุดที่ $f' = 0$ และ $f \neq 0$ วิธีจะติดอยู่ที่นั่น จุดนั้นเป็นจุดต่ำสุดเฉพาะที่ของ $|f|$ ที่ไม่ใช่ราก ดังนั้นการหน่วงช่วยไม่ให้ก้าวกระโดดหลุด แต่ไม่ได้บอกว่าได้รากไหนเลย

ลองเอง

ใช้ $f(x) = x^2 - 2$ และ $x_0 = 0.1$ ลองคำนวณก้าวแรกของนิวตันแบบหน่วงด้วยมือ ก้าวเต็มทำให้ $|f|$ โตขึ้น เราต้องลองตัวคูณ t เป็นกี่ค่า (คำตอบอยู่ในแบบฝึกหัด 6.3.2)

ร่างเพื่อตรวจ · Preprint · 2026-10-23

ตารางที่ 6.11: วิธีนิวตันบนผนังจาก $x_0 = 6$: ค่าตกค้าง ขนาดก้าว และค่าคลาดเคลื่อนสัมบูรณ์จริง (ตัวเลขจากโค้ดของเรา)

n	x_n	$\lvert\operatorname{gap}(x_n)\rvert$	$\lvert x_{n+1} - x_n\rvert$	$\lvert x_n - \alpha\rvert$
0	6	0.779415	0.644054	0.632181
1	6.644054	0.014101	0.011894	0.011873
2	6.632161	2.47×10^{-5}	2.08×10^{-5}	2.08×10^{-5}
3	6.632181	7.4×10^{-11}	6.2×10^{-11}	6.2×10^{-11}

เกณฑ์การหยุด: ค่าตกค้าง ขั้น และค่าคลาดเคลื่อน

เมื่อทำซ้ำไปเรื่อย ๆ เราต้องตัดสินใจว่าเมื่อไรพอ มีสามตัวเลขที่ใช้ได้ ตัวแรกคือ **ค่าตกค้าง** (residual) $|f(x_n)|$ ซึ่งบอกว่าสมการพลาดไปเท่าไรในรอบนี้ ตัวที่สองคือขนาดก้าว $|x_{n+1} - x_n|$ และตัวที่สามคือค่าคลาดเคลื่อนสัมบูรณ์จริง $|x_n - \alpha|$ ตัวที่สามเราไม่รู้ในทางปฏิบัติ ส่วนสองตัวแรกคำนวณได้ทุกรอบ ตารางนี้เป็นวิธีนิวตันบนผนังจาก $x_0 = 6$

สำหรับรากเดี่ยวอย่างผนังของเรา สามตัวเลขนี้ให้ข้อสรุปเดียวกัน แต่มีสองกรณีที่ตัวเลขใดตัวเลขหนึ่งหลอกเราได้

ข้อควรระวัง: ตัวเลขเล็กที่หลอกเรา

- **ค่าตกค้างหลอก.** ดู $f(x) = (x-1)^3$ ซึ่งมีรากซ้ำที่ 1 จาก $x_0 = 1.1$ วิธีนิวตันลดค่าคลาดเคลื่อนเป็น $\frac{2}{3}$ ของเดิมทุกรอบ (แบบฝึกหัด 6.3.3) ค่าตกค้างเป็น e_n^3 จึงลดเร็วกว่าค่าคลาดเคลื่อนมาก ที่ $n = 11$ ค่าตกค้างยังเป็น 1.5×10^{-9} แต่ที่ $n = 12$ ค่าตกค้างต่ำกว่า 10^{-9} ทั้งที่ค่าคลาดเคลื่อนยังเป็น 7.7×10^{-4}
- **ขั้นหลอก.** ดูการทำซ้ำช้า $g(x) = 0.999x + 0.001$ จาก $x_0 = 0$ ซึ่งมีจุดตรึง 1 ขั้นของมันเท่ากับ 0.001 คูณค่าคลาดเคลื่อน เมื่อขั้นน้อยกว่า 10^{-6} เป็นครั้งแรกที่ $n = 6905$ แต่ค่าคลาดเคลื่อนยังเป็น 9.99×10^{-4} หรือ 1000 เท่าของขั้น

ดังนั้นเราใช้เกณฑ์หลายตัวร่วมกัน และตั้งจำนวนรอบสูงสุดไว้ด้วย ถ้าเกินจำนวนนั้นแล้วยังไม่ลู่เข้า ควรถือว่าจุดเริ่มนี้ไม่ดี แล้วเปลี่ยนวิธีหรือจุดเริ่ม ข้อคิดนี้มาจากบริบทของสมการเชิงอนุพันธ์ที่ Iserles ใช้ ซึ่งการหยุดและเปลี่ยนขนาดขั้นเป็นทางเลือกที่ดีกว่าการทำซ้ำต่อไปเรื่อย ๆ Iserles 2009, §7.3, pp. 131--132

การเร่งแบบไอทเคน: ทายจุดหมายจากสามรอบ

ในหัวข้อ 6.1 ค่าคลาดเคลื่อนของ $g_{1/2}$ ลดลงโดยประมาณด้วยตัวคูณคงที่ $\rho \approx 0.405$ ถ้าลำดับลู่เข้าเชิงเส้นแบบนี้ เราทายจุดหมายได้จากสามค่าติดกัน **การเร่งแบบไอทเคน** (Aitken extrapolation) คือการเดาจุดหมาย $\hat{x}$ จาก x_0, x_1, x_2 ด้วยสูตร

$$\hat{x} = x_0 - \frac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0}. \tag{6.5}$$

ทฤษฎีบท 6.3.2: การเร่งแบบไอทเคนบนลำดับเชิงเส้นที่แน่นอน

ให้ $x_n = \alpha + A\rho^n$ สำหรับทุก $n \geq 0$ โดย $A \neq 0$ และ $\rho \neq 1$ แล้วสูตร (6.5) ให้ $\hat{x} = \alpha$ เสมอ

พิสูจน์. **แนวคิด.** ผลต่างของลำดับเชิงเส้นมีตัวประกอบร่วม $A(\rho - 1)$ ซึ่งตัดกันได้พอดี

1. **ผลต่างลำดับแรก.** $x_1 - x_0 = A(\rho - 1)$
2. **ผลต่างลำดับที่สอง.** $x_2 - 2x_1 + x_0 = A(\rho^2 - 2\rho + 1) = A(\rho - 1)^2$ ซึ่งไม่เป็นศูนย์เพราะ $A \neq 0$ และ $\rho \neq 1$

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 6.12: การเร่งแบบไอทเคนบนลำดับ $g_{1/2}$ จาก $x_0 = 6$ โดย $\hat{x}_n$ ใช้ x_n, x_{n+1}, x_{n+2} (ตัวเลขจากโค้ดของเรา)

n	x_n	$\lvert x_n - \alpha\rvert$	$\hat{x}_n$	$\lvert \hat{x}_n - \alpha\rvert$
0	6	0.632181	6.628657	0.003524
1	6.389708	0.242474	6.630008	0.002173
2	6.537834	0.094348	6.631742	0.000439
3	6.594652	0.037530	6.632105	0.0000769
4	6.617093	0.015089	6.632169	0.0000129

3. **แทนค่า.** $\hat{x} = x_0 - \dfrac{A^2(\rho-1)^2}{A(\rho-1)^2} = x_0 - A = \alpha$

สรุป: ถ้าค่าคลาดเคลื่อนลดลงด้วยตัวคูณคงที่ สูตรเร่งจะให้จุดหมายทันที □

ลำดับจริงไม่ได้ตรงตามรูปแบบนี้พอดี แต่เมื่อ n โตขึ้น อัตราส่วนค่าคลาดเคลื่อนเริ่มคงที่ เราจึงได้ค่าประมาณที่ดีขึ้น Atkinson 1989, §2.6, pp. 83–86 ตารางต่อไปนี้ใช้ลำดับ $g_{1/2}$ จาก $x_0 = 6$ ของหัวข้อ 6.1

จากสามค่าแรก $\hat{x}_0$ ห่างจากราก 0.0035 ขณะที่ x_2 ห่างถึง 0.094 คือไกลกว่าเกือบ 27 เท่า

ข้อควรระวัง: ไอทเคนใช้กับลำดับเชิงเส้นเท่านั้น

สูตร (6.5) สมมติว่าค่าคลาดเคลื่อนลดลงด้วยตัวคูณคงที่ ถ้าใช้กับนิวตันซึ่งลู่เข้าแบบกำลังสอง ตัวคูณนั้นไม่คงที่ ค่าที่ได้จึงไม่น่าเชื่อถือ ส่วนตัวหารที่เป็นศูนย์ก็เกิดได้เมื่อค่าสามค่าเท่ากันถึงระดับการปัดเศษ เราจึงให้ค่ากลับเป็น x_2 ในกรณีนี้

วิธีผสม: เก็บช่วงล้อมไว้ แล้วให้นิวตันเดินเมื่อปลอดภัย

ในหัวข้อ 6.1 เราเก็บช่วงล้อมไว้เสมอ จึงปลอดภัยแต่ช้า ในหัวข้อ 6.2 นิวตันเร็ว แต่อาจกระโดดออกจากช่วงที่เราควบคุมได้ **วิธีผสม** (hybrid method) คือวิธีที่รวมสองอย่างเข้าด้วยกันโดยใช้กฎที่ตรวจได้ง่าย

1. เริ่มจากช่วงล้อม $[a, b]$ และค่าเดาตั้งต้น x ในช่วงนั้น (ถ้าไม่มีค่าเดาตั้งต้น ให้ใช้จุดกลาง)
2. คำนวณจุดนิวตัน $c = x - f(x)/f'(x)$
3. ถ้า c อยู่ในช่วงเปิด (a, b) และรอบก่อนหน้า *ไม่ใช่*นิวตันที่ทำให้ช่วงแคบลงไม่ถึงครึ่ง ให้รับ c เป็นก้าวนิวตัน มิฉะนั้นให้ใช้จุดกลาง $c = \frac{a+b}{2}$ เป็นก้าวแบ่งครึ่ง
4. คำนวณ $f(c)$ ถ้า $f(c) = 0$ ให้หยุดที่ c เพราะได้รากพอดี มิฉะนั้นแทนปลายของช่วงที่มีเครื่องหมายเดียวกับ $f(c)$ ด้วย c ช่วงใหม่ยังล้อมรากเสมอ
5. ให้ $x = c$ แล้วทำซ้ำจนความกว้าง $b - a$ ไม่เกินค่าที่ต้องการ

ทฤษฎีบท 6.3.3: ความกว้างของวิธีผสม

ให้ f หาอนุพันธ์ได้ต่อเนื่อง $f(a)f(b) < 0$ และให้ w_k เป็นความกว้างของช่วงก่อนรอบที่ k ของวิธีผสม จะได้

1. ถ้า $f(c) \neq 0$ ในทุกรอบ $1, \ldots, k+1$ แล้ว ช่วงที่เก็บหลังแต่ละรอบเหล่านั้นยังมีเครื่องหมายต่างกันที่ปลาย และ $w_{k+2} \le \frac{1}{2} w_k$
2. ถ้ารอบใดได้ $f(c) = 0$ วิธีหยุดที่ c ซึ่งเป็นรากพอดี ช่วงที่เหลือคือ $[c, c]$ มีความกว้างศูนย์ และปลายทั้งสองมี

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 6.13: วิธีผสมบนผนังจาก $x_0 = 4.4$ ในช่วง $[4, 8]$ (ตัวเลขจากโค้ดของเรา)

n	ก้าว	x_n	$\lvert x_n - \alpha\rvert$	ความกว้างช่วง
1	แบ่งครึ่ง	6.000000	0.632181	2.000000
2	นิวตัน	6.644054	0.011873	0.644054
3	นิวตัน	6.632161	2.08×10^{-5}	0.011894
4	นิวตัน	6.632181	6×10^{-11}	0.011873
5	แบ่งครึ่ง	6.638118	0.005936	0.005936

$f = 0$ จึงไม่มีเครื่องหมายต่างกัน นี่คือข้อยกเว้นของข้อ 1

3. ถ้าปลายข้างหนึ่งคือ x และ α เป็นรากในช่วงนั้น จะได้ $\lvert x - \alpha\rvert \le b - a$

พิสูจน์. **แนวคิด.** ช่วงไม่เคยขยาย และการแบ่งครึ่งลดความกว้างลงครึ่งหนึ่งพอดี กฎข้อที่ 3 ทำให้ไม่มีคู่รอบใดที่ไม่ลดครึ่ง ถ้าเจอ $f(c) = 0$ วิธีหยุดทันที

1. **ช่วงยังล้อม.** ถ้า $f(c) \neq 0$ ทุกก้าวเลือกช่วงย่อยที่ปลายมีเครื่องหมายต่างกัน ดังนั้นช่วงใหม่อยู่ในช่วงเดิมและยังล้อมราก
2. **ความกว้างไม่เพิ่ม.** ช่วงใหม่อยู่ในช่วงเดิม จึง $w_{k+1} \le w_k$ และก้าวแบ่งครึ่งที่ $f(c) \neq 0$ ทำให้ $w_{k+1} = \frac{1}{2} w_k$ พอดี ถ้า $f(c) = 0$ ช่วงยุบเป็นจุดเดียวและวิธีหยุด (ข้อ 2 ของทฤษฎีบท)
3. **สองรอบ.** ถ้ารอบ k เป็นการแบ่งครึ่งแล้ว $w_{k+1} = \frac{1}{2} w_k \le \frac{1}{2} w_k$ และ $w_{k+2} \le w_{k+1} \le \frac{1}{2} w_k$ ถ้ารอบ k เป็นนิวตันที่ทำให้ $w_{k+1} \le \frac{1}{2} w_k$ แล้ว $w_{k+2} \le w_{k+1} \le \frac{1}{2} w_k$ ถ้ารอบ k เป็นนิวตันที่ไม่ถึงครึ่ง กฎข้อที่ 3 บังคับให้รอบ $k + 1$ เป็นการแบ่งครึ่ง จึง $w_{k+2} = \frac{1}{2} w_{k+1} \le \frac{1}{2} w_k$
4. **ระยะถึงราก.** ทั้ง x และ α อยู่ในช่วง $[a, b]$ จึง $\lvert x - \alpha\rvert \le b - a$

สรุป: ความกว้างลดลงอย่างน้อยครึ่งหนึ่งทุกสองรอบ และความกว้างของช่วงเป็นขอบเขตค่าคลาดเคลื่อนสัมบูรณ์ที่แน่นอน ยกเว้นเมื่อพบรากพอดีแล้วหยุดที่จุดนั้น □

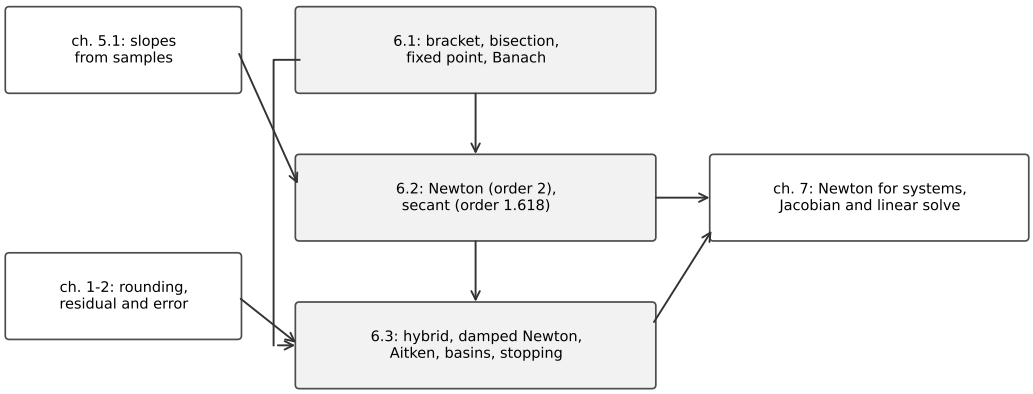
อ่านแบบเข้าใจง่าย

ข้อ 1 บอกว่า วิธีผสมไม่แย่กว่าการแบ่งครึ่งเกินสองรอบ ข้อ 2 บอกว่า ถ้าเจอรากพอดีก็หยุดได้ทันที ส่วนข้อ 3 บอกว่าความกว้างของช่วงเป็นขอบเขตค่าคลาดเคลื่อนสัมบูรณ์ที่เราได้มาโดยไม่ต้องรู้รากจริง ราคาที่ต้องจ่ายคือ นิวตันอาจเข้าใกล้ราก แต่ช่วงยังกว้างอยู่ จนกว่าจะแคบลงก็ยังหยุดไม่ได้

ตารางต่อไปนี้แสดงวิธีผสมจาก $x_0 = 4.4$ บนช่วง $[4, 8]$ ตัวเลขทั้งหมดมาจากโค้ดของเรา

ในรอบที่ 4 ค่า x_4 ห่างจากราก 6×10^{-11} แต่ช่วงยังกว้าง 0.0119 เพราะการเลื่อนข้างเดียวของนิวตันไม่ทำให้ช่วงแคบลงครึ่งหนึ่ง กฎข้อ 3 จึงบังคับให้รอบที่ 5 แบ่งครึ่ง เมื่อกำหนดให้หยุดเมื่อช่วงกว้างไม่เกิน 10^{-10} วิธีผสมจากจุดนี้ต้องใช้ 40 รอบ ส่วนการแบ่งครึ่งบนช่วงเดิมใช้ 35 รอบ วิธีผสมไม่ได้เร็วกว่าเสมอไป แต่ไม่มีการกระโดดหลุดช่วง และได้ขอบเขตค่าคลาดเคลื่อนสัมบูรณ์ที่รับประกัน

สัมมนา · Preprint · 2026-10-23



รูปที่ 6.9: แผนผังบทที่ 6 ทั้งบท: อินพุตจากบทที่ 5 และบทที่ 1--2 วิธีในหัวข้อ 6.1 ถึง 6.3 และสิ่งที่บทที่ 7 ใช้ต่อ (สร้างจาก `numanal/ch06.py`)

ลองเอง

เริ่มจาก $x_0 = 4.4$ บนช่วง $[4, 8]$ จงบอกว่ารอบแรกของวิธีผสมเป็นนิวตันหรือแบ่งครึ่ง และเพราะอะไร (คำตอบอยู่ในแบบฝึกหัด 6.3.5)

สรุปใจความ

- ค่าเดาตั้งต้นกำหนดว่าวิธีจะลงที่รากไหน ภาพของแอ่งดึงดูดแสดงว่าค่าที่ใกล้กันอาจไปคนละราก: จาก $x_0 = 4.4$ เราได้จุดชนที่สอง ไม่ใช่จุดชนแรก
- นิวตันแบบหน่วงลด $|f|$ ได้ทุกรอบและไม่กระโดดหลุด แต่ $|f|$ ที่เล็กไม่ได้แปลว่าใกล้ราก เพราะอาจติดจุดที่ $f' = 0$
- เกณฑ์การหยุดควรดูค่าตกค้าง ขนาดก้าว และจำนวนรอบพร้อมกัน ค่าที่เล็กอย่างเดียวหลอกได้ทั้งในรากซ้ำและการทำซ้ำช้า
- การเร่งแบบไอทเคนทายจุดหมายจากสามค่าติดกันได้แม่นสำหรับลำดับเชิงเส้น แต่ไม่ใช้กับนิวตัน
- วิธีผสมเก็บช่วงล้อมไว้ ได้ขอบเขตค่าคลาดเคลื่อนสัมบูรณ์ที่แน่นอน ความกว้างลดลงอย่างน้อยครึ่งหนึ่งทุกสองรอบ แต่อาจใช้รอบมากกว่าการแบ่งครึ่งบนช่วงเดิม

อ่านต่อ. ตำรา Atkinson 1989, §2.5--2.6, pp. 76--86 อธิบายการทำซ้ำทั่วไปและการเร่งแบบไอทเคนด้วยสูตรที่ละเอียดกว่า และ Iserles 2009, §7.3, pp. 130--132 อธิบายว่าค่าเดาตั้งต้นกับเกณฑ์การหยุดมีผลต่อการลู่เข้าอย่างไร ในบริบทของการทำซ้ำในวิธีขั้นเวลา ส่วนนิวตันแบบหน่วงที่เราเขียนขึ้นเองไม่ได้อยู่ในหนังสือเหล่านี้

รูปที่ 6.9 แสดงแผนผังของบทที่ 6 ฉบับสมบูรณ์ ว่าแต่ละวิธีต้องใช้อะไร และส่งต่อไปยังบทที่ 7 อย่างไร

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 6.3.1 (⋆) ให้ $f(x) = x^3 - x$ มีรากที่ $-1, 0$ และ 1 วิธีนิวตันคือ $N(x) = x - \dfrac{f(x)}{f'(x)}$

1. จงแสดงว่า $N(x) = \dfrac{2x^3}{3x^2 - 1}$
2. จงคำนวณ $N(0.5)$ และ $N(0.1)$ แล้วบอกว่าแต่ละค่าเดาตั้งต้นเข้าใกล้ราก ± 1 หรือ 0 ในหนึ่งรอบ หรือต้องมากกว่าหนึ่งรอบ
3. จงหาจุด $x > 0$ ที่ $f'(x) = 0$ และอธิบายว่าทำไมจุดนี้ไม่ควรใช้เป็นค่าเดาตั้งต้น

แบบฝึกหัด 6.3.2 (⋆⋆) ให้ $f(x) = x^2 - 2$ และ $x_0 = 0.1$ วิธีนิวตันแบบหน่วงทำก้าวแรกดังนี้ $d = -f(x_0)/f'(x_0)$ แล้วลองตัวคูณ $t = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$ ตัวแรกที่ทำให้ $|f|$ เล็กลง

1. จงหา d และค่า $|f|$ ที่จุด $x_0 + td$ สำหรับ $t = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ บอกว่า t ใดถูกรับ
2. จงหา x_1 แล้วคำนวณ x_2 โดยนิวตันเต็มรูป (ตัวคูณ $t = 1$) และตรวจว่า x_2 ใกล้ $\sqrt{2}$
3. จงอธิบายว่าทำไมก้าวเต็มจาก x_0 จึงไม่ถูกรับ ทั้งที่ x_0 อยู่ใกล้ 0 มาก

แบบฝึกหัด 6.3.3 (⋆⋆) ให้ $f(x) = (x - 1)^3$ ซึ่งมีรากซ้ำที่ 1 และเริ่มจาก $x_0 = 1.1$ ด้วยวิธีนิวตัน

1. จงแสดงว่า $N(x) - 1 = \frac{2}{3}(x - 1)$ จึงได้ $e_n = \frac{1}{10}\left(\frac{2}{3}\right)^n$ โดย $e_n = x_n - 1$
2. จงคำนวณค่าตกค้าง $|f(x_n)|$ สำหรับ $n = 11$ และ $n = 12$ แล้วบอกว่าเกณฑ์ค่าตกค้างต่ำกว่า 10^{-9} หยุดที่รอบใด
3. จงหาค่าคลาดเคลื่อนจริง e_n ที่รอบนั้น และอธิบายว่าเหตุใดเกณฑ์ค่าตกค้างจึงหลอก

แบบฝึกหัด 6.3.4 (⋆⋆) ให้ลำดับ $x_n = 2 + \frac{1}{2}(0.2)^n$ และให้ $\hat{x}_0$ เป็นค่าเร่งของไอทเคนจาก x_0, x_1, x_2

1. จงคำนวณ x_0, x_1, x_2 แล้วหา $\hat{x}_0$ ด้วยสูตร (6.5) ตรวจว่าได้ 2
2. จงแสดงว่าถ้าเติมค่าคงที่ B ลงในทุก x_n แล้ว $\hat{x}$ เลื่อนไปด้วย B เท่านั้น
3. จงอธิบายว่าเมื่อ $\rho = 1$ สูตรไอทเคนใช้ไม่ได้ และเพราะเหตุใดตามทฤษฎีบท

แบบฝึกหัด 6.3.5 (⋆⋆) ใช้ผนัง gap บนช่วง $[4, 8]$ ซึ่งมี $\mathrm{gap}(4) > 0 > \mathrm{gap}(8)$ (ใช้เครื่องคิดเลขได้)

1. จาก $x_0 = 4.4$ จงดูว่าก้าวนิวตันแรกอยู่ในช่วงหรือไม่ และบอกว่ารอบแรกต้องเป็นการแบ่งครึ่งหรือไม่ เพราะอะไร
2. จาก $x_0 = 6.5$ จงคำนวณก้าวนิวตัน $c = x_0 - \mathrm{gap}(x_0)/\mathrm{gap}'(x_0)$ (กำหนด $\mathrm{gap}(6.5) \approx 0.159880$ และ $\mathrm{gap}'(6.5) \approx -1.226588$) และบอกว่า c อยู่ในช่วงหรือไม่
3. จงบอกว่าถ้ารับก้าวนิวตันจากข้อ 2 ช่วงใหม่คืออะไร และความกว้างเป็นกี่เท่าของความกว้างเดิม 4 (ตรวจว่าแคบลงอย่างน้อยครึ่งหนึ่งหรือไม่)

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 6.3.6 (★★) *สถานการณ์สมมติ:* ยืมเงิน $400{,}000$ บาทมาผ่อนเดือนละ $9{,}000$ บาทเป็นเวลา 60 เดือน อัตราดอกเบี้ยรายเดือน r ทำให้

$$F(r) = 9000 \cdot \frac{1-(1+r)^{-60}}{r} - 400000 = 0.$$

ตัวเลขนี้เป็นตัวอย่างสมมติเพื่อการคำนวณ ไม่ได้อ้างอิงสัญญาของใคร

1. จงแสดงว่า $F(0.001) > 0$ และ $F(0.1) < 0$ จึงมีรากในช่วง $[0.001, 0.1]$
2. จงหาจำนวนรอบของการแบ่งครึ่งที่รับประกันว่า $|c_n - r| \le 10^{-6}$
3. จงแสดงว่า $F'(r) = 9000 \cdot \dfrac{60(1+r)^{-61} r - \left(1-(1+r)^{-60}\right)}{r^2}$ และทำนิวตันจาก $r_0 = 0.01$ สามรอบ

แบบฝึกหัด 6.3.7 (★★★) *โครงงานเล็กของบทที่ 6.* ยิงกระสุนจากจุด $(0, 2)$ ด้วยเส้นตรง $y = 2 + mx$ ไปชนผนัง $y = \sin x$ โดยให้ความชัน m เปลี่ยนไป ฟังก์ชันจุดชนคือ $g_m(x) = 2 + mx - \sin x$

> **ตัวช่วยคิด**
>
> ก่อนรันนิวตัน ตรวจให้แน่ใจว่าช่วง $[4, 8]$ มีรากเดียวของ g_m โดยดูการเปลี่ยนเครื่องหมายบนตารางจุดที่ถี่พอ และในวิธีผสมให้ใช้ความกว้างของช่วงเป็นเกณฑ์การหยุด ไม่ใช่ค่าตกค้าง

ทำกับความชัน $m \in \{-0.2,\ -0.25,\ -0.3,\ -0.4\}$ (-0.25 คือกระสุนของเรา)

1. จงตรวจว่าแต่ละ g_m มีรากเดียวบน $[4, 8]$ และหาราก
2. จงทำนิวตันจาก $x_0 = 0$ และจาก $x_0 = 4.4$ นาน 30 รอบ บอกว่าแต่ละกรณีถึงรากใดในช่วง $[4, 8]$ หรือบินหลุดหรือค้าง
3. จงทำวิธีผสมบน $[4, 8]$ จาก $x_0 = 4.4$ จนความกว้างช่วงไม่เกิน 10^{-10} นับจำนวนรอบ และเทียบกับการแบ่งครึ่งบนช่วงเดียวกัน
4. สรุปเป็นย่อหน้า: ความชันใดทำให้นิวตันจาก 0 ไปถึงรากแรก และผลนี้บอกอะไรเกี่ยวกับค่าเดาตั้งต้น และวิธีผสมคุ้มค่าหรือไม่ในกรณีเหล่านี้

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 6.3.1.

1. $f'(x) = 3x^2 - 1$ จึง $N(x) = x - \dfrac{x^3 - x}{3x^2 - 1} = \dfrac{x(3x^2-1) - x^3 + x}{3x^2 - 1} = \dfrac{2x^3}{3x^2 - 1}$
2. $N(0.5) = \dfrac{2 \cdot 0.125}{3 \cdot 0.25 - 1} = \dfrac{0.25}{-0.25} = -1$ ดังนั้นค่าเดาตั้งต้น 0.5 ไปถึงรากเต็ม -1 ในหนึ่งรอบ ส่วน $N(0.1) = \dfrac{0.002}{0.03 - 1} \approx -0.0020619$ แล้ว $N(-0.0020619) \approx 1.754 \times 10^{-8}$ และรอบที่สามเล็กกว่า 10^{-23} จึงใกล้ราก 0 ภายในสามรอบ
3. $f'(x) = 0$ เมื่อ $3x^2 = 1$ คือ $x = \frac{1}{\sqrt{3}} \approx 0.57735$ ที่จุดนี้ตัวหารของ N เป็นศูนย์ เส้นสัมผัสขนานแกนนอนจึงไม่ตัดแกนที่จุดจำกัด ค่าเดาตั้งต้นที่อยู่ใกล้จุดนี้จะทำให้ก้าวยาวมากหรือไม่ได้คำตอบเลย

ตรวจด้วยการคำนวณ: test_ex_6_3_1_newton_map_for_x_cubed_minus_x_from_one_half_and_one_tenth ในไฟล์ทดสอบ.

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 6.3.2.

1. $f(0.1) = -1.99$ และ $f'(0.1) = 0.2$ จึง $d = -\dfrac{f(0.1)}{f'(0.1)} = 9.95$ แล้วตรวจ $|f|$ ที่จุด $0.1 + td$:
   - $t = 1$: $x = 10.05$, $f = 99.0025$ มากกว่า 1.99 ไม่รับ
   - $t = \frac{1}{2}$: $x = 5.075$, $f = 23.7556$ ไม่รับ
   - $t = \frac{1}{4}$: $x = 2.5875$, $f = 4.6952$ ไม่รับ
   - $t = \frac{1}{8}$: $x = 1.34375$, $f = -0.19434$ มี $|f| \approx 0.19434 < 1.99$ จึงรับ
2. $x_1 = 1.34375$ และ $f'(x_1) = 2.6875$ ก้าวนิวตันเต็มได้ $x_2 = x_1 - \dfrac{f(x_1)}{f'(x_1)} = 1.34375 + \dfrac{0.19434}{2.6875} \approx 1.416061$ ค่า $|f(x_2)| \approx 0.00523$ น้อยกว่า 0.19434 จึงรับ $t = 1$ และ x_2 ใกล้ $\sqrt{2} \approx 1.414214$ อยู่ห่างประมาณ 0.0018
3. ความชันของเส้นสัมผัสที่ 0.1 มีค่า 0.2 เท่านั้น เส้นจึงแทบขนานแกนนอน ก้าวนิวตันเต็มจึงยาวถึง 9.95 ทำให้ไปไกลจากรากมาก ตัวคูณ t ที่เล็กพอทำให้ก้าวสั้นและพาเข้าใกล้รากได้

ตรวจด้วยการคำนวณ: test_damped_newton_on_x_squared_minus_two_from_one_tenth ในไฟล์ทดสอบ.

เฉลย 6.3.3.

1. $N(x) = x - \dfrac{(x-1)^3}{3(x-1)^2} = x - \dfrac{x-1}{3}$ เมื่อ $x \neq 1$ ดังนั้น $N(x) - 1 = (x-1) - \dfrac{x-1}{3} = \dfrac{2}{3}(x-1)$ ซึ่งหมายความว่า $e_{n+1} = \frac{2}{3}e_n$ และ $e_n = \frac{1}{10}\left(\frac{2}{3}\right)^n$ เพราะ $e_0 = 0.1$
2. ค่าตกค้างคือ $|f(x_n)| = e_n^3$ เมื่อ $n = 11$: $e_{11} = \frac{1}{10}(\frac{2}{3})^{11} \approx 1.15610 \times 10^{-3}$ จึง $|f| \approx 1.5452 \times 10^{-9}$ ซึ่งยังมากกว่า 10^{-9} เมื่อ $n = 12$: $e_{12} \approx 7.7073 \times 10^{-4}$ และ $|f| \approx 4.578 \times 10^{-10}$ ซึ่งต่ำกว่า 10^{-9} เกณฑ์จึงหยุดที่ $n = 12$
3. ที่ $n = 12$ ค่าคลาดเคลื่อนจริงคือ $e_{12} \approx 7.7 \times 10^{-4}$ ซึ่งยังใหญ่กว่า 10^{-6} มาก เกณฑ์ค่าตกค้างหลอกเพราะ $|f| = e^3$ เป็นกำลังสามของค่าคลาดเคลื่อน ค่า 10^{-9} ของค่าตกค้างจึงตรงกับค่าคลาดเคลื่อนราว 10^{-3} เท่านั้น

ตรวจด้วยการคำนวณ: test_ex_6_3_3_triple_root_residual_at_n_eleven_and_twelve ในไฟล์ทดสอบ.

เฉลย 6.3.4.

1. $x_0 = 2 + \frac{1}{2} = 2.5$, $x_1 = 2 + \frac{1}{2}(0.2) = 2.1$, $x_2 = 2 + \frac{1}{2}(0.04) = 2.02$ จึง $x_1 - x_0 = -0.4$ และ $x_2 - 2x_1 + x_0 = 2.02 - 4.2 + 2.5 = 0.32$ ดังนั้น $\hat{x}_0 = 2.5 - \dfrac{(-0.4)^2}{0.32} = 2.5 - 0.5 = 2$
2. ถ้า $x'_n = x_n + B$ แล้ว $x'_1 - x'_0 = x_1 - x_0$ และ $x'_2 - 2x'_1 + x'_0 = x_2 - 2x_1 + x_0$ เพราะค่าคงที่หักล้างกัน จึง $\hat{x}' = x'_0 - \dfrac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0} = \hat{x} + B$ ดังนั้นจุดหมายเลื่อนไปด้วย B เท่านั้น
3. เมื่อ $\rho = 1$ ลำดับเป็น $x_n = \alpha + A$ คงที่ ผลต่างทั้งสองเป็นศูนย์ สูตรจึงเป็น $0/0$ ทฤษฎีบทต้องการ $\rho \neq 1$ เพราะที่นั่นค่าคลาดเคลื่อนไม่ลดลงเลย ค่าคงที่ของลำดับไม่บอกว่าจุดหมายอยู่ที่ α หรือที่ $\alpha + A$ จึงไม่มีข้อมูลพอจะเร่ง

ตรวจด้วยการคำนวณ: test_ex_6_3_4_aitken_recovers_the_limit_of_a_two_term_sequence และการเลื่อนค่าคงที่ในไฟล์ทดสอบ.

ฉบับร่าง Preprint 2026-10-23

เฉลย 6.3.5.

1. $\mathrm{gap}(4.4) \approx 1.851602$ และ $\mathrm{gap}'(4.4) \approx 0.057333$ จึงก้าวนิวตันคือ $4.4 - \dfrac{1.851602}{0.057333} \approx -27.9$ ซึ่งอยู่นอกช่วง $(4, 8)$ วิธีผสมจึงไม่รับก้าวนิวตัน รอบแรกต้องเป็นการแบ่งครึ่ง ได้จุดกลาง 6

2. $c = 6.5 - \dfrac{0.159880}{-1.226588} \approx 6.630345$ และ $c \in (4, 8)$ จึงรับเป็นก้าวนิวตัน (ค่าแม่นคือ $\mathrm{gap}(c) \approx 0.00218 > 0$ จึง $c < \alpha$)

3. $\mathrm{gap}(c) > 0$ จึงเลื่อนปลายซ้ายไปที่ c ช่วงใหม่คือ $[6.630345,\ 8]$ มีความกว้าง $8 - 6.630345 = 1.369655$ ซึ่งเป็น 0.342 เท่าของความกว้างเดิม 4 น้อยกว่า 2 จึงแคบลงอย่างน้อยครึ่งหนึ่งแล้ว ดังนั้นกฎข้อ 3 ไม่บังคับให้รอบถัดไปเป็นการแบ่งครึ่ง

ตรวจด้วยการคำนวณ: test_ex_6_3_5_hybrid_from_six_keeps_the_sign_change_and_the_bound ในไฟล์ทดสอบ (ค่าของข้อ 1 และข้อ 2 ตรวจด้วยการคำนวณในโค้ดด้วย).

เฉลย 6.3.6.

1. $F(0.001) = 9000 \cdot \dfrac{1 - 1.001^{-60}}{0.001} - 400000 \approx 123865 > 0$ และ $F(0.1) = 9000 \cdot \dfrac{1 - 1.1^{-60}}{0.1} - 400000 \approx -310296 < 0$ ดังนั้น $[0.001, 0.1]$ เป็นช่วงล้อม และ F มีรากหนึ่งรากในช่วงนี้ (ตรวจด้วยการคำนวณในโค้ดว่ารากมีเพียงหนึ่งราก)

2. ความยาวช่วงคือ 0.099 ต้องการ $\dfrac{0.099}{2^{n+1}} \le 10^{-6}$ จึง $2^{n+1} \ge 99000$ และ $2^{16} = 65536 < 99000 \le 2^{17}$ ดังนั้น $n + 1 = 17$ คือ $n = 16$ รอบ

3. กระจายจาก $F(r) = 9000h(r) - 400000$ เมื่อ $h(r) = \dfrac{1 - u^{-60}}{r}$ และ $u = 1 + r$ ได้ $h'(r) = \dfrac{60u^{-61}r - (1 - u^{-60})}{r^2}$ จึง $F'(r) = 9000\,h'(r)$ ตามที่ต้องการ ที่ $r_0 = 0.01$ ได้ $F(r_0) \approx 4595.35$ และ $F'(r_0) \approx -1.10296 \times 10^7$ จึง $r_1 = 0.01 + \dfrac{4595.35}{1.10296 \times 10^7} \approx 0.0104166$ รอบถัดไป $r_2 \approx 0.0104200$ และ $r_3 \approx 0.0104200$ (เท่ากับ r_2 ถึงหลักที่แปด) ดังนั้นอัตราดอกเบี้ยรายเดือนคือราว 1.042%

ตรวจด้วยการคำนวณ: ชุด test_loan ในไฟล์ทดสอบ (สามการทดสอบ).

เฉลย 6.3.7.

1. บนตารางจุดละเอียดของ $[4, 8]$ แต่ละ g_m เปลี่ยนเครื่องหมายเพียงครั้งเดียว ราก (จากโค้ดของเรา) คือ 6.941415 ($m = -0.2$), 6.632181 ($m = -0.25$), 6.371770 ($m = -0.3$) และ 5.910438 ($m = -0.4$)

2. ผลของนิวตันจาก $x_0 = 0$ และจาก $x_0 = 4.4$ (30 รอบ) เป็นดังนี้

m	นิวตันจาก 0	นิวตันจาก 4.4
-0.20	บินหลุด ($x \approx 5982$)	ค้าง ($x \approx -3.52$)
-0.25	ค้าง ($x \approx 10.85$)	9.927600 (จุดชนที่สอง)
-0.30	6.371770 (ถึงราก)	ค้าง ($x \approx 1.96$)
-0.40	บินหลุด ($x \approx 1.3 \times 10^5$)	5.910438 (ถึงราก)

3. วิธีผสมจาก $x_0 = 4.4$ บน $[4, 8]$ ต้องใช้ 50, 40, 36 และ 58 รอบตามลำดับ ส่วนการแบ่งครึ่งบนช่วงเดียวกันต้องใช้ 35 รอบ วิธีผสมจึงใช้รอบมากกว่า แต่ไม่หลุดช่วงและได้ขอบเขตที่แน่นอนทุกรอบ

ฉบับร่าง · Preprint · 2026-10-23

4. ความชันที่นิวตันจาก 0 ถึงรากแรกได้คือ $m = -0.3$ เท่านั้น จากค่าเดาตั้งต้น 4.4 นิวตันถึงรากแรกได้เฉพาะ $m = -0.4$ ไม่มีค่าเดาตั้งต้นเดียวที่ดีสำหรับทุกความชัน ในสี่กรณีนี้วิธีผสมได้รากเดียวกับที่การแบ่งครึ่งจะได้ แต่แลกด้วยรอบที่มากขึ้นเมื่อหยุดด้วยความกว้างช่วง จึงคุ้มเมื่อต้องการความแน่นอน ไม่คุ้มเมื่อต้องการความเร็วเพียงอย่างเดียว

ตรวจด้วยการคำนวณ: ชุด test_mini_project และ test_hybrid_steps ในไฟล์ทดสอบ.

สรุปบทที่ 6 ในหน้าเดียว

ตารางต่อไปนี้เทียบวิธีทั้งหมดของบทนี้ ตัวอักษร α คือราก

วิธี	ต้องการอะไร	อัตราหรืออันดับ	ค่าใช้จ่ายต่อรอบ
แบ่งครึ่ง (6.1)	ช่วงล้อม	เชิงเส้น $\frac{1}{2}$	ค่า f หนึ่งค่า
$x_{n+1} = g(x_n)$ (6.1)	การหดตัว $L < 1$	เชิงเส้น L	ค่า g หนึ่งค่า
นิวตัน (6.2)	f' และค่าเดาตั้งต้นใกล้ราก	กำลังสอง	f และ f'
เซแคนต์ (6.2)	ค่าตั้งต้นสองค่า	≈ 1.618	ค่า f หนึ่งค่า
นิวตันแบบหน่วง (6.3)	เหมือนนิวตัน	ขึ้นกับจุดเริ่ม	f และ f' กับการทดสอบ t
ไอทเคน (6.3)	ลำดับเชิงเส้นที่ ρ คงที่	ขึ้นกับลำดับเดิม	สามค่าติดกัน
วิธีผสม (6.3)	ช่วงล้อมและ f'	กว้างลดครึ่งทุกสองรอบ	เหมือนนิวตันกับแบ่งครึ่ง

สูตรที่ควรจำ

- แบ่งครึ่ง: $|c_n - \alpha| \le \dfrac{b-a}{2^{n+1}}$
- บานัค: $|x_n - \alpha| \le \dfrac{L^n}{1-L}|x_1 - x_0|$ และ $|x_n - \alpha| \le \dfrac{L}{1-L}|x_n - x_{n-1}|$
- นิวตัน: $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$ และ $|x_{n+1} - \alpha| \le C|x_n - \alpha|^2$ เมื่อเริ่มใกล้พอ
- ไอทเคน: $\hat{x} = x_0 - \dfrac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0}$
- วิธีผสม: $|x - \alpha| \le b - a$ เมื่อ α เป็นรากในช่วง $[a, b]$

อ่านต่อ. เมื่อตัวไม่รู้มีหลายตัว วิธีนิวตันในแต่ละรอบต้องแก้ระบบเชิงเส้นหนึ่งระบบ บทที่ 7 สมการเชิงเส้นหลายตัว จะบอกวิธีแก้ระบบนั้น และกลับมาที่วิธีนิวตันสำหรับระบบในหัวข้อ 7.4 โครงงานเล็กของบทนี้คือแบบฝึกหัด 6.3.7

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 7

สมการเชิงเส้นหลายตัว

สถานการณ์สมมติ: โซ่ที่แขวนแล้วนิ่ง

ลองนึกถึงตัวละครที่ถือโซ่ยาว หรือผ้าคลุมที่ห้อยจากจุดสองจุด เมื่อปล่อยแล้วมันแกว่งอยู่สักพักแล้วก็หยุดในรูปที่ไม่ขยับอีก ลูกปัดแต่ละลูกในโซ่หยุดตรงที่แรงดึงจากเพื่อนบ้านสองข้างกับน้ำหนักของมันหักล้างกันพอดี

ถ้าเราจำลองโซ่ด้วยลูกปัดห้าลูก เราต้องหาตำแหน่งของลูกปัดห้าลูก และตำแหน่งของแต่ละลูกขึ้นกับเพื่อนบ้าน จึงได้สมการห้าสมการที่ผูกกันทั้งหมด เราแก้ลูกแรกก่อนไม่ได้ เพราะมันรู้จักลูกที่สองด้วย

เอกสารของ Thomas Jakobsen ระบุว่าอัลกอริทึมของเขาถูกพัฒนาขึ้นสำหรับเกม Hitman: Codename 47 และเขาเลือกปรับตำแหน่งทีละข้อจำกัดซ้ำ ๆ หลายรอบ แทนการแก้ระบบทั้งชุดในคราวเดียว Jakobsen 2001, บทคัดย่อ และ §4 ส่วนเอกสารของ Baraff และ Witkin ปี 1998 บอกว่าการจำลองผ้าที่ใช้ก้าวเวลาใหญ่ ต้องแก้ระบบเชิงเส้นขนาดใหญ่ที่เป็น**เมทริกซ์มากเลขศูนย์** (sparse matrix) ทุกก้าว Baraff and Witkin 1998, บทคัดย่อ

ทั้งสองวิธีมุ่งไปที่คำตอบเดียวกัน แต่ต้นทุนต่างกัน คำถามของบทนี้จึงเป็นว่า เมื่อตัวไม่ทราบหลายตัวผูกกันเป็นระบบ เราแก้พร้อมกันได้อย่างไร และการทำซ้ำทีละน้อยเข้าใกล้คำตอบได้เร็วแค่ไหน

(แบบจำลองโซ่เป็นสถานการณ์สมมติที่เราสร้างเอง ตัวเลขทั้งหมดในบทนี้มาจากโค้ดของเรา และไม่ได้บอกว่าเกมหรือเอนจินใดทำงานอย่างไรในรายละเอียด)

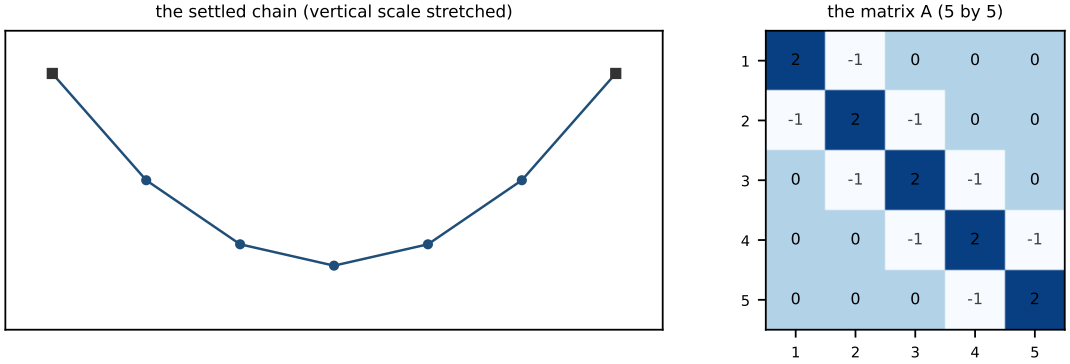
ภาพรวมบท

เราเริ่มจากโซ่ลูกปัดที่แขวนนิ่ง แล้วจะเจออีกด้านหนึ่งคือภาพเบลอที่ทำให้ได้ง่ายแต่แก้คืนยาก ตามลำดับนี้

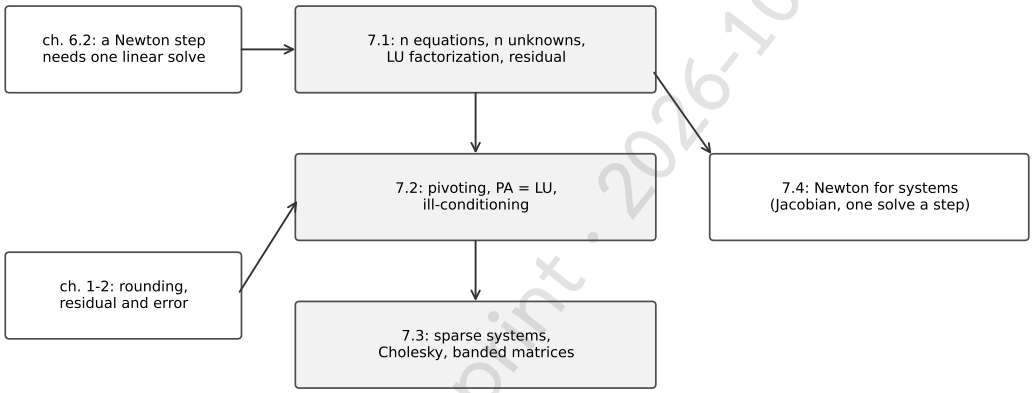
- หัวข้อ 7.1: ตัวไม่ทราบหลายตัวพร้อมกัน เราเขียนโซ่เป็นระบบ $A\mathbf{x} = \mathbf{b}$ แล้วแก้ด้วยการกำจัดแบบเกาส์ที่เก็บผลไว้เป็นการแยกตัวประกอบ LU และเทียบกับการทำซ้ำทีละลูก
- หัวข้อ 7.2: ทำไมต้องสลับแถว และทำไมภาพเบลอแก้คืนยาก การทำให้ภาพเบลอทำได้ง่าย แต่การกลับด้านขยายสัญญาณรบกวนเล็กน้อยให้ใหญ่มาก
- หัวข้อ 7.3: เมื่อเมทริกซ์ส่วนใหญ่เป็นศูนย์ ผ้าที่มีจุดหลายหมื่นจุดทำให้ระบบใหญ่มาก แต่แต่ละแถวมีสมาชิกที่ไม่เป็นศูนย์เพียงไม่กี่ตัว
- หัวข้อ 7.4: กลับไปสมการไม่เชิงเส้น แขนสองท่อนที่ต้องไปถึงจุดที่ผู้เล่นคลิก ต้องแก้ระบบเชิงเส้นในทุกรอบของนิวตัน

ภาพแรกแสดงโซ่ที่นิ่งแล้ว และรูปแบบของเมทริกซ์ที่เกิดจากสมการของแต่ละลูก

ฉบับร่าง · preprint · 2026-10-28



รูปที่ 7.1: ซ้าย: โซ่ห้าลูกที่นิ่งแล้ว ระยะตกลงของแต่ละลูกคำนวณจากสูตรในหัวข้อ 7.1 ขวา: เมทริกซ์ A ของระบบ ค่าที่ไม่เป็นศูนย์อยู่บนแนวทแยงกับแนวที่ติดกันเท่านั้น (สร้างจาก `numanal/ch07.py`)



รูปที่ 7.2: ลำดับของหัวข้อในบทที่ 7 และสิ่งที่บทนี้รับมาจากบทที่ 6 และบทที่ 1--2 (สร้างจาก `numanal/ch07.py`)

ภาพรวมของลำดับหัวข้อในบทนี้แสดงไว้ในรูปต่อไปนี้

7.1 ตัวไม่ทราบหลายตัวพร้อมกัน

สถานการณ์สมมติ: ลูกปัดแต่ละลูกมองเพื่อนบ้าน

ลูกปัดลูกที่สองถูกดึงขึ้นจากลูกแรก และถูกดึงลงจากลูกที่สาม เมื่อแรงสองข้างกับน้ำหนักหักล้างกันพอดี มันก็หยุด แต่ลูกที่สามก็กำลังรอลูกข้างเคียงของมันอยู่เหมือนกัน คำถามคือ จะหาตำแหน่งที่ทุกลูกพอใจพร้อมกันได้อย่างไร

ภาพรวมก่อนอ่าน

ความคิดของหัวข้อนี้คือ ถ้าเขียนสมการของลูกปัดทุกลูกไว้ด้วยกัน จะได้ระบบเชิงเส้นที่มีจำนวนสมการเท่ากับจำนวนตัวไม่ทราบ แล้วแก้ทั้งชุดในคราวเดียวได้ด้วยการกำจัดแบบเกาส์ ผลที่ได้คือเมทริกซ์สามเหลี่ยมสองตัวที่ใช้ซ้ำได้เมื่อแรงดึงเปลี่ยน อีกทางหนึ่งคือปรับทีละลูกไปเรื่อย ๆ ซึ่งทำงานน้อยในแต่ละรอบ แต่ต้องทำหลายรอบเมื่อโซ่ยาวขึ้น

ฉบับร่าง · Preprint · 2026-10-23

จากลูกปัดสู่สมการ

เราสร้างแบบจำลองขึ้นเองให้ง่ายพอจะคำนวณได้ทั้งหมด ลูกปัดแต่ละลูกผูกกับเพื่อนบ้านด้วยสปริงที่มีค่าคงที่ $K = 40$ นิวตันต่อเมตร น้ำหนักของลูกละ $W = 2$ นิวตัน ตำแหน่งในแนวนอนของลูกปัดคงที่ เราดูเฉพาะการตกลงในแนวดิ่ง ให้ u_i เป็นระยะที่ลูกที่ i ตกลงจากจุดแขวน (หน่วยเมตร ทิศลงเป็นบวก) และหมุดที่ปลายทั้งสองไม่ขยับ จึงมี $u_0 = u_6 = 0$

สปริงซ้ายดึงลูกที่ i ลงด้วยแรง $K(u_{i-1} - u_i)$ และสปริงขวาดึงลงด้วย $K(u_{i+1} - u_i)$ เมื่อรวมกับน้ำหนัก W ที่ดึงลงด้วย สมดุลคือผลรวมของแรงเป็นศูนย์

$$K\,(2u_i - u_{i-1} - u_{i+1}) = W, \qquad i = 1, 2, \ldots, 5. \tag{7.1}$$

หารด้วย K แล้วใช้ $c = W/K = 0.05$ เมตร จะได้ระบบ $A\mathbf{u} = \mathbf{b}$

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}, \qquad \mathbf{b} = c\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}. \tag{7.2}$$

ระบบนี้มีห้าสมการและห้าตัวไม่ทราบค่า แต่ละแถวมีค่าไม่เป็นศูนย์อยู่เพียงสามช่องที่อยู่ติดกัน เมทริกซ์แบบนี้เรียกว่า **เมทริกซ์สามแนวเฉียง** (tridiagonal matrix) ซึ่งหมายถึงเมทริกซ์ที่มีค่าไม่เป็นศูนย์เฉพาะบนแนวทแยงหลักกับแนวที่อยู่ติดข้างบนและข้างล่าง

นิยาม 7.1.1: ระบบสมการเชิงเส้น

ให้ A เป็นเมทริกซ์จัตุรัสขนาด $n \times n$ และ $\mathbf{b} \in \mathbb{R}^n$ เราเรียก $A\mathbf{x} = \mathbf{b}$ ว่า **ระบบสมการเชิงเส้น** (linear system) ที่มี n สมการและ n ตัวไม่ทราบค่า ถ้า A เป็น **เมทริกซ์ไม่เอกฐาน** (nonsingular matrix) ซึ่งหมายความว่า $\det A \neq 0$ ระบบจะมีคำตอบเดียว $\mathbf{x} = A^{-1}\mathbf{b}$ Iserles 2009, ภาคผนวก A.1.4.1, p. 434

อ่านแบบเข้าใจง่าย. ระบบที่มีคำตอบเดียวคือระบบที่สมการแต่ละอันให้ข้อมูลใหม่ ไม่มีสมการไหนซ้ำหรือขัดกัน ตัวอย่างที่ไม่ดีคือ $x + y = 1$ กับ $x + y = 2$ ซึ่งไม่มีคำตอบเลยเพราะขัดกัน และ $x + y = 1$ กับ $2x + 2y = 2$ ซึ่งมีคำตอบมากมายเพราะสมการที่สองเป็นสมการเดิมคูณสอง ทั้งสองกรณีเมทริกซ์เป็นเอกฐานจึง $\det A = 0$

วิธีที่หนึ่ง: กำจัดทีละตัวแล้วแทนค่ากลับ

ใช้สมการแรกกำจัด u_1 ออกจากสมการถัดไป โดยคูณสมการแรกด้วยตัวคูณที่เหมาะสมแล้วลบ จากนั้นทำซ้ำกับสมการที่เหลือจนได้ระบบสามเหลี่ยมบนที่แก้จากล่างขึ้นบนได้ทีละตัว วิธีนี้เรียกว่า **การกำจัดแบบเกาส์** (Gaussian elimination) Iserles 2009, ภาคผนวก A.1.4.3, pp. 434--435 และตำรา Atkinson 1989, §8.1, pp. 508--511 มีการไล่ขั้นตอนอย่างละเอียด

ตัวหารที่ใช้กำจัดในแต่ละขั้นเรียกว่า **ตัวหลัก** (pivot) ตัวหลักแรกของโซ่ของเราคือ 2 ตัวคูณของแถวที่สองคือ $-\frac{1}{2}$ เมื่อลบแถวแรกคูณตัวคูณนี้ออก แถวที่สองจะกลายเป็น $(0,\ \frac{3}{2},\ -1,\ 0,\ 0 \,|\, \frac{3}{2}c)$ ตัวหลักของขั้นนี้คือ $\frac{3}{2}$ ซึ่งให้ตัวคูณถัดไปเป็น $-\frac{2}{3}$ ทำต่อไปจนสิ้นสุด ได้ตารางต่อไปนี้ ซึ่งสรุปตัวหลักและตัวคูณของโซ่ห้าลูก

ตัวหลักเป็น $\frac{k+1}{k}$ และตัวคูณเป็น $-\frac{k}{k+1}$ ซึ่งเราจะพิสูจน์ในแบบฝึกหัด 7.1.4 ตัวหลักทุกตัวเป็นบวก ไม่มีตัวไหนเป็นศูนย์ การกำจัดจึงทำได้ตามลำดับโดยไม่ต้องสลับแถว

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 7.1: ตัวหลักและตัวคูณของการกำจัดแบบเกาส์สำหรับโซ่ห้าลูก (ค่าเป็นเศษส่วนที่ตรงกับผลจากโค้ดของเรา)

k	ตัวหลัก u_{kk}	ตัวคูณ $\ell_{k+1,k}$
1	2	$-\frac{1}{2}$
2	$\frac{3}{2}$	$-\frac{2}{3}$
3	$\frac{4}{3}$	$-\frac{3}{4}$
4	$\frac{5}{4}$	$-\frac{4}{5}$
5	$\frac{6}{5}$	(ไม่มี)

นิยาม 7.1.2: การแยกตัวประกอบ LU

ให้ A เป็นเมทริกซ์จัตุรัส ถ้าเราหาเมทริกซ์สามเหลี่ยมล่าง L ที่มีหนึ่งทุกช่องบนแนวทแยง และเมทริกซ์สามเหลี่ยมบน U ซึ่ง $A = LU$ เราเรียกการเขียนนี้ว่า **การแยกตัวประกอบ LU** (LU factorization) ของ A สมาชิกบนแนวทแยงของ U คือตัวหลักของการกำจัด

อ่านแบบเข้าใจง่าย. ลองนึกว่าการกำจัดเป็นการจดบันทึกการทำงาน ทุกตัวคูณที่ใช้ถูกจดไว้ใต้แนวทแยงของ L ส่วนที่ทำให้ระบบกลายเป็นสามเหลี่ยมบนคือ U ดังนั้นการกำจัดหนึ่งครั้งให้คู่ L, U ที่ใช้ซ้ำได้ เมื่อ $\mathbf{b}$ เปลี่ยน เราไม่ต้องกำจัดใหม่ เพียงแก้ $L\mathbf{y} = \mathbf{b}$ แล้วแก้ $U\mathbf{x} = \mathbf{y}$ ซึ่งแต่ละขั้นเป็นการแทนค่าทีละตัวในระบบสามเหลี่ยม Iserles 2009, ภาคผนวก A.1.4.5, pp. 435--436

ลองเอง

ทำซ้ำขั้นตอนเดียวกันกับโซ่สามลูก ซึ่งมีเมทริกซ์ A_3 ขนาด 3×3 และ $\mathbf{b} = c(1,1,1)^T$ ที่ $c = 0.05$ หาตัวหลักทั้งสามตัว ตัวคูณสองตัว และแก้ $L\mathbf{y} = \mathbf{b}$ แล้ว $U\mathbf{x} = \mathbf{y}$ ผลที่ควรได้คือตัวหลัก $2,\ \frac{3}{2},\ \frac{4}{3}$ ตัวคูณ $-\frac{1}{2},\ -\frac{2}{3}$ และ $\mathbf{x} = (0.075,\ 0.1,\ 0.075)$ ตรวจว่าตรงกับ $u_i = c\,i(4-i)/2$ หรือไม่

ทฤษฎีบท 7.1.1: การมีอยู่ของการแยกตัวประกอบ LU

ให้ A เป็นเมทริกซ์จริงขนาด $n \times n$ ซึ่งเมทริกซ์ย่อยมุมบนซ้ายขนาด $k \times k$ เป็นเมทริกซ์ไม่เอกฐานสำหรับทุก $k = 1, \ldots, n$ แล้วมี L สามเหลี่ยมล่างที่แนวทแยงเป็น 1 และ U สามเหลี่ยมบน ที่ $A = LU$ และตัวหลัก $u_{kk} = \det A_k / \det A_{k-1}$ ไม่เป็นศูนย์ทุกตัว โดยที่ A_k คือเมทริกซ์ย่อยมุมบนซ้ายขนาด $k \times k$ และ $\det A_0 = 1$

พิสูจน์. **แนวคิด.** เราทำการอุปนัยเชิงคณิตศาสตร์บน n: ส่วนบนซ้ายขนาด $n-1$ แยกได้ตามสมมติฐาน แล้วที่เหลือต่อได้ด้วยการเลือกแถวและคอลัมน์สุดท้าย และตัวหลักตัวสุดท้ายเป็นอัตราส่วนของดีเทอร์มิแนนต์

1. **กรณี** $n = 1$. $A = (a_{11})$ โดยที่ $a_{11} \neq 0$ ตั้ง $L = (1)$ และ $U = (a_{11})$ ก็เสร็จ
2. **แบ่งบล็อก.** เขียน $A = \begin{pmatrix} A_{11} & \mathbf{a} \\ \mathbf{r}^T & \alpha \end{pmatrix}$ เมื่อ A_{11} คือ A_{n-1} ขนาด $(n-1) \times (n-1)$ ซึ่งเป็นเมทริกซ์ไม่เอกฐานตามสมมติฐาน
3. **ใช้สมมติฐานอุปนัย.** จาก $A_{11} = L_{11}U_{11}$ โดยที่ L_{11} มีแนวทแยงเป็น 1 และ U_{11} มีตัวหลักไม่เป็นศูนย์ ทั้งสองกลับเมทริกซ์ได้ เพราะ $\det U_{11} = \det A_{11} \neq 0$

สงวนลิขสิทธิ์ Preprint, 2026-10-23

ตารางที่ 7.2: จำนวนการดำเนินการของการแยก LU แบบเมทริกซ์เต็ม (ตัวเลขจากสูตรและจากโค้ดของเรา ตรงกัน)

n	คูณและหาร	บวกและลบ	รวม	$\frac{2}{3}n^3$
5	40	30	70	83.3
10	330	285	615	666.7
20	2 660	2 470	5 130	5 333.3
100	333 300	328 350	661 650	666 666.7

4. **สร้าง L และ U.** ตั้ง $\mathbf{y} = L_{11}^{-1}\mathbf{a}$ และ $s = \alpha - \mathbf{r}^T U_{11}^{-1}\mathbf{y}$ แล้วตั้ง

$$L = \begin{pmatrix} L_{11} & 0 \\ \mathbf{r}^T U_{11}^{-1} & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} U_{11} & \mathbf{y} \\ 0 & s \end{pmatrix}.$$

   ตรวจได้ตรง ๆ ว่า LU มีบล็อกบนซ้าย $L_{11}U_{11} = A_{11}$ บล็อกบนขวา $L_{11}\mathbf{y} = \mathbf{a}$ บล็อกล่างซ้าย $\mathbf{r}^T$ และมุมล่างขวา $\mathbf{r}^T U_{11}^{-1}\mathbf{y} + s = \alpha$ จึง $LU = A$

5. **ตัวหลักสุดท้ายไม่เป็นศูนย์.** $\det A = \det L \det U = \det U_{11} \cdot s$ และ $\det A_{11} = \det U_{11}$ จึงได้ $s = \det A / \det A_{11}$ ซึ่งไม่เป็นศูนย์ เพราะ $\det A \neq 0$ และ $\det A_{11} \neq 0$ ดังนั้น $u_{nn} = s$ และ $u_{kk} = \det A_k / \det A_{k-1}$ ไม่เป็นศูนย์ทุกตัว

สรุป: บทพิสูจน์และการเขียนเงื่อนไขในรูปดีเทอร์มิแนนต์นี้เราเขียนเองจากการกำจัดแบบเกาส์ ตัวหลักแต่ละตัวเป็นอัตราส่วนของดีเทอร์มิแนนต์ที่ไม่เป็นศูนย์ □

อ่านแบบเข้าใจง่าย

ทฤษฎีบทนี้บอกว่า ถ้าเมทริกซ์ย่อยมุมบนซ้ายทุกขนาดไม่ยุบ การกำจัดแบบเกาส์จะเดินได้โดยไม่ต้องสลับแถว และได้ $A = LU$ เสมอ สำหรับโซ่ของเรา $\det A_k = k + 1$ จึงได้ $u_{kk} = (k+1)/k$ ตรงกับตารางที่ 7.1 พอดี

ข้อควรระวัง: ไม่เอกฐานไม่ได้แปลว่าแยก LU ได้เสมอ

เมทริกซ์ $B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ ไม่เอกฐาน เพราะ $\det B = -1$ แต่ตัวหลักตัวแรกคือ 0 การกำจัดแบบเกาส์โดยไม่สลับแถวหยุดทันที เราจึงแยก $B = LU$ แบบนี้ไม่ได้ ทางแก้คือสลับแถวก่อน ซึ่งจะเป็นเรื่องของหัวข้อ 7.2 (ตัวอย่างนี้เป็นของเราเอง)

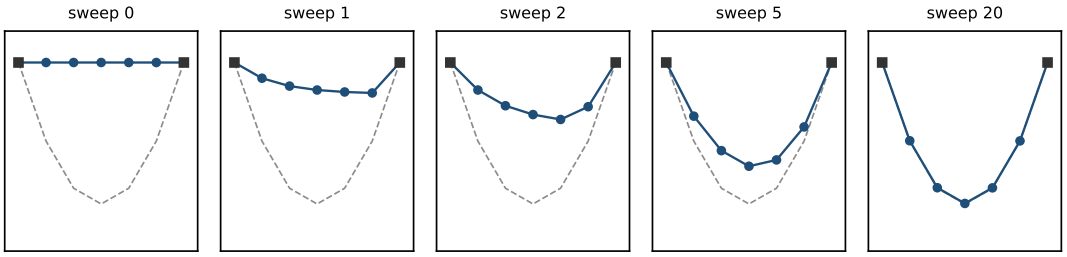
จ่ายกี่การดำเนินการ

เราจะเปรียบเทียบวิธีต่าง ๆ จึงต้องนับการดำเนินการ เรานับจำนวนการคูณและการหาร กับจำนวนการบวกและการลบแยกกัน **จำนวนการดำเนินการ** (operation count) ที่ขั้น k ของการกำจัด เราหาตัวคูณ $n - k$ ตัว (การหาร) แล้วปรับสมาชิก $(n-k)^2$ ตัว (การคูณและการลบ) ผลรวมทุกขั้นให้ตารางต่อไปนี้ ตามสูตรในตำรา Atkinson 1989, §8.1, Table 8.1, p. 513

ข้อควรระวัง: ตัวเลขนี้นับทุกช่องของเมทริกซ์เต็ม

ตัวเลข 70 ของโซ่ห้าลูกนับการคูณกับศูนย์ด้วย ทั้งที่ A มีศูนย์เกือบทุกช่อง โซ่ที่ยาวขึ้นจึงควรใช้วิธีที่ไม่ต้องนับศูนย์ เรื่องนี้จะกลับมาในหัวข้อ 7.3

ฉบร่าง Preprint · 2026-10-23



รูปที่ 7.3: โซ่ห้าลูกหลังการทำซ้ำ 0, 1, 2, 5 และ 20 รอบ เส้นประคือรูปที่นิ่งแล้ว (สร้างจาก `numanal/ch07.py`)

ตารางที่ 7.3: ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดของโซ่ห้าลูกหลัง s รอบของการทำซ้ำแบบเกาส์–ไซเดล เริ่มจากศูนย์ (ตัวเลขจากโค้ดของเรา)

s	ค่าคลาดเคลื่อนสัมบูรณ์สูงสุด
0	2.2500×10^{-1}
1	1.8125×10^{-1}
2	1.4219×10^{-1}
3	1.0664×10^{-1}
5	5.9985×10^{-2}
10	1.4235×10^{-2}
20	8.0161×10^{-4}
30	4.5142×10^{-5}

ร่าง · preprint 2026-10-23

แนวคิดของการแยกนี้ คือการกำจัดหนึ่งครั้งได้ L และ U ที่ใช้แก้ระบบด้วยการแทนค่ากลับ สำหรับโซ่ของเรา การแทนค่ากลับให้คำตอบ $(0.125,\ 0.2,\ 0.225,\ 0.2,\ 0.125)$ เมตร ซึ่งตรงกับสูตร $u_i = c\,i(6-i)/2$ ผิดจากค่าที่แม่นจริงไม่เกิน 3×10^{-17} ซึ่งอยู่ในระดับการปัดเศษของเลขทศนิยมแบบคู่ (double) ในคอมพิวเตอร์

วิธีที่สอง: ทำซ้ำทีละลูก

ลองนึกภาพว่าลูกปัดแต่ละลูกปรับตำแหน่งตามเพื่อนบ้านที่เพิ่งขยับล่าสุด ถ้าเขียนสมการของลูกที่ i ใหม่ให้เป็น $u_i = \tfrac{1}{2}(u_{i-1} + u_{i+1} + c)$ เราจะเริ่มจากค่าศูนย์ทุกลูก แล้วปรับลูกที่ 1 ก่อน จากนั้นลูกที่ 2 โดยใช้ค่า u_1 ที่เพิ่งได้มา แล้วไปต่อจนลูกที่ 5 หนึ่งรอบของการปรับนี้เรียกว่า **การทำซ้ำแบบเกาส์–ไซเดล** (Gauss--Seidel iteration) และทำซ้ำหลายรอบจนค่าหยุดเปลี่ยน

เอกสารของ Jakobsen เรียกแนวคิดคล้ายนี้ว่า **การผ่อน** (relaxation) และอธิบายว่ามันเป็นการทำซ้ำแบบ Jacobi หรือ Gauss--Seidel ขึ้นกับวิธีลงมือทำ Jakobsen 2001, §4 ในที่นี้เราใช้รูปแบบเกาส์–ไซเดลกับระบบของเราเอง ซึ่งไม่ใช่โค้ดของเอกสารนั้น

ลองเอง

เริ่มจาก $\mathbf{u} = (0,0,0,0,0)$ และ $c = 0.05$ จงคำนวณรอบแรกด้วยมือ: $u_1 = \tfrac{1}{2}(0+0+c) = 0.025$ แล้ว $u_2 = \tfrac{1}{2}(0.025+0+c)$ ไปเรื่อย ๆ ผลที่ควรได้คือ $(0.025,\ 0.0375,\ 0.04375,\ 0.046875,\ 0.0484375)$ เทียบกับคำตอบที่ถูกต้องว่าลูกแรกยังห่างไปมากแค่ไหน

ค่าคลาดเคลื่อนสัมบูรณ์ (absolute error) คือระยะห่างระหว่างค่าที่คำนวณได้กับค่าจริง ตาราง 7.3 แสดงค่าคลาดเคลื่อนสัมบูรณ์สูงสุดของโซ่ห้าลูกหลังจำนวนรอบต่าง ๆ เริ่มจากศูนย์ วัดด้วย $\max_i |u_i^{(s)} - u_i|$ (ตัวเลขจากโค้ดของเรา)

ค่าคลาดเคลื่อนสัมบูรณ์ลดลงทุกรอบ และในช่วงหลังลดลงประมาณ 0.75 เท่าต่อรอบ (วัดจากโค้ดของเรา) รูป 7.3 แสดงการ

เปลี่ยนแปลงเดียวกันเป็นภาพ จำนวนรอบที่ต้องใช้เพื่อให้ค่าคลาดเคลื่อนสัมบูรณ์ของทุกลูกไม่เกิน 10^{-3} คือ 20 รอบ ถ้าต้องการไม่เกิน 10^{-4} ต้องใช้ 28 รอบ และไม่เกิน 10^{-6} ต้องใช้ 44 รอบ

เมื่อโซ่ยาวขึ้น จำนวนรอบเพิ่มเร็วมาก ที่ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-3} โซ่สิบลูกต้องใช้ 81 รอบ และโซ่ยี่สิบลูกต้องใช้ 355 รอบ ตัวเลขเหล่านี้มาจากโค้ดของเรา ไม่ใช่สูตรปิด ในขณะที่การแยก LU ของโซ่สิบลูกใช้ 615 การดำเนินการ และยี่สิบลูกใช้ $5\,130$ การดำเนินการ (ตาราง 7.2) เราจะเห็นในแบบฝึกหัด 7.1.6 ว่าสองวิธีนี้เทียบกันอย่างไร

ข้อควรระวัง: การเปลี่ยนแปลงเล็กไม่ได้แปลว่าค่าคลาดเคลื่อนสัมบูรณ์เล็ก

ที่รอบที่ 20 การเปลี่ยนแปลงสูงสุดระหว่างรอบที่ 19 กับ 20 คือ 2.7×10^{-4} แต่ค่าคลาดเคลื่อนสัมบูรณ์จริงยังอยู่ที่ 8.0×10^{-4} เมื่อการลู่เข้าช้า การหยุดเมื่อค่าเปลี่ยนน้อยจึงหยุดเร็วเกินไปได้ เราจะดูเกณฑ์การหยุดให้ละเอียดขึ้นในบทที่ 6 และหัวข้อ 7.2

นอร์มและค่าตกค้าง

เพื่อบอกว่าคำตอบเวกเตอร์ใกล้แค่ไหน เราต้องวัดขนาดของเวกเตอร์ก่อน

นิยาม 7.1.3: นอร์มของเวกเตอร์

นอร์มของเวกเตอร์ (vector norm) คือฟังก์ชัน $\|\cdot\|$ จาก $\mathbb{R}^n$ เป็นจำนวนจริงที่ (N1) $\|\mathbf{x}\| \geq 0$ และ $\|\mathbf{x}\| = 0$ ก็ต่อเมื่อ $\mathbf{x} = \mathbf{0}$ (N2) $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ และ (N3) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ Atkinson 1989, §7.3, p. 481

ตัวอย่างที่เราใช้บ่อยมีสามแบบ **นอร์มหนึ่ง** (1-norm) คือผลรวมของค่าสัมบูรณ์ขององค์ประกอบ $\|\mathbf{x}\|_1 = \sum_i |x_i|$ **นอร์มสอง** (2-norm) คือความยาวของเวกเตอร์ตามทฤษฎีบทพีทาโกรัส $\|\mathbf{x}\|_2 = \left(\sum_i x_i^2\right)^{1/2}$ และ **นอร์มอนันต์** (infinity norm) คือค่าสัมบูรณ์ที่มากที่สุดขององค์ประกอบ $\|\mathbf{x}\|_\infty = \max_i |x_i|$ Iserles 2009, ภาคผนวก A.1.3.3, p. 433 สำหรับโซ่ของเรา $\mathbf{u} = (0.125,\ 0.2,\ 0.225,\ 0.2,\ 0.125)$ ได้ $\|\mathbf{u}\|_1 = 0.875$ เมตร $\|\mathbf{u}\|_2 \approx 0.40234$ เมตร และ $\|\mathbf{u}\|_\infty = 0.225$ เมตร ตัวเลขเหล่านี้มาจากโค้ดของเรา

นิยาม 7.1.4: นอร์มเมทริกซ์ที่ถูกเหนี่ยวนำ

เมื่อมีนอร์มของเวกเตอร์ $\|\cdot\|$ อยู่แล้ว เราวัดว่าเมทริกซ์ A ขยายเวกเตอร์ได้มากที่สุดเท่าไร ด้วยอัตราส่วน $\|A\mathbf{x}\|/\|\mathbf{x}\|$ เมื่อ $\mathbf{x} \neq \mathbf{0}$ แล้วใช้ค่าที่มากที่สุดนั้นเป็น **นอร์มเมทริกซ์ที่ถูกเหนี่ยวนำ** (induced matrix norm) ของ A

$$\|A\| = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}.$$

ดังนั้น $\|A\mathbf{x}\| \leq \|A\|\,\|\mathbf{x}\|$ สำหรับทุก $\mathbf{x}$ ทันทีจากนิยาม Iserles 2009, ภาคผนวก A.1.3.4, pp. 433–434

นอร์มเมทริกซ์ที่ถูกเหนี่ยวนำจาก $\|\cdot\|_1$ เท่ากับผลรวมค่าสัมบูรณ์ของคอลัมน์ที่มากที่สุด และจาก $\|\cdot\|_\infty$ เท่ากับผลรวมค่าสัมบูรณ์ของแถวที่มากที่สุด Atkinson 1989, §7.3, Table 7.1, p. 485 สำหรับเมทริกซ์ A ของโซ่ห้าลูก คอลัมน์และแถวที่มากที่สุดรวม $1 + 2 + 1 = 4$ ดังนั้น $\|A\|_1 = \|A\|_\infty = 4$

นิยาม 7.1.5: ค่าตกค้าง

ถ้า $\hat{\mathbf{x}}$ เป็นคำตอบที่เราหาได้โดยประมาณ **ค่าตกค้าง** (residual) ของ $\hat{\mathbf{x}}$ คือ $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$ และค่าคลาดเคลื่อนสัมบูรณ์คือ $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$ โดยที่ $\mathbf{x}$ คือคำตอบจริง

ฉบับร่าง Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย. ค่าตกค้างบอกว่าคำตอบที่เดานั้นพลาดจากสมการจริงเท่าไร ส่วนค่าคลาดเคลื่อนสัมบูรณ์บอกว่าพลาดจากคำตอบจริงเท่าไร ทั้งสองเชื่อมกันด้วยสมการ $A\mathbf{e} = \mathbf{r}$ เพราะ $A\mathbf{e} = A\mathbf{x} - A\hat{\mathbf{x}} = \mathbf{b} - A\hat{\mathbf{x}} = \mathbf{r}$ จึงได้ $\mathbf{e} = A^{-1}\mathbf{r}$ ถ้าทราบ A^{-1} เราจะแปลงค่าตกค้างเป็นค่าคลาดเคลื่อนสัมบูรณ์ได้

ข้อควรระวัง: ค่าตกค้างเล็กไม่รับประกันค่าคลาดเคลื่อนสัมบูรณ์เล็ก

จาก $\mathbf{e} = A^{-1}\mathbf{r}$ ถ้า A^{-1} มีขนาดใหญ่ ค่าตกค้างที่เล็กมากยังให้ค่าคลาดเคลื่อนสัมบูรณ์ที่ใหญ่ได้ ผลนี้ขึ้นกับความไวของระบบต่อการเปลี่ยนแปลงเล็กน้อย ซึ่งเป็นเรื่องของหัวข้อ 7.2 โดยตรง

สรุปใจความ

- โซ่ที่ผูกกันมีตัวไม่ทราบเท่ากับจำนวนสมการ จึงเขียนเป็นระบบ $A\mathbf{x} = \mathbf{b}$ และแก้ทั้งชุดได้
- การกำจัดแบบเกาส์เก็บผลไว้เป็น LU ซึ่งใช้กับ $\mathbf{b}$ ที่ต่างกันได้โดยไม่ต้องกำจัดใหม่ และต้องทำประมาณ $\frac{2}{3}n^3$ การดำเนินการ
- การทำซ้ำแบบเกาส์–ไซเดลทำงานน้อยในแต่ละรอบ แต่จำนวนรอบเพิ่มเร็วเมื่อโซ่ยาวขึ้น (20, 81 และ 355 รอบ สำหรับโซ่ 5, 10 และ 20 ลูก เมื่อต้องการค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-3})
- นอร์มกับค่าตกค้างวัดขนาดของคำตอบและความพลาด แต่ค่าตกค้างเล็กไม่รับประกันค่าคลาดเคลื่อนสัมบูรณ์เล็กเสมอไป

กลับไปที่โซ่ที่แขวนนิ่ง การแก้พร้อมกันให้คำตอบตรงทุกลูกด้วยการดำเนินการ 70 ครั้ง ส่วนการทำซ้ำทีละลูกต้องใช้ 20 รอบเพื่อให้ผิดไม่เกิน 10^{-3} คำถามที่เหลือคือ เมื่อเมทริกซ์ใหญ่ขึ้นมากจนเก็บทุกช่องไม่ไหว จะทำอย่างไร

อ่านต่อ. ตำรา Atkinson 1989, §7.3, pp. 481--485 อธิบายนอร์มของเวกเตอร์และเมทริกซ์อย่างละเอียด และ Atkinson 1989, §8.1, pp. 508--515 กับ Atkinson 1989, §8.3, pp. 522--524 อธิบายการกำจัดแบบเกาส์และรูปแบบที่คำนวณตัวประกอบ LU ทีละสมาชิก (Doolittle และ Crout) ส่วนเอกสารของ Jakobsen Jakobsen 2001, §2 และ §4 และเอกสารของ Baraff และ Witkin Baraff and Witkin 1998 บอกว่าการจำลองผ้าในงานของเขาใช้การทำซ้ำและระบบขนาดใหญ่อย่างไรในระดับแนวคิด

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 7.1.1 (⋆) โซ่สี่ลูกมีน้ำหนักลูกละ $W = 3$ นิวตัน และสปริงมีค่าคงที่ $K = 40$ นิวตันต่อเมตร แบบจำลองเหมือนในหัวข้อ 7.1

1. จงเขียนเมทริกซ์ A ขนาด 4×4 และเวกเตอร์ $\mathbf{b}$ ของระบบ $A\mathbf{u} = \mathbf{b}$ โดยใช้ $c = W/K$
2. จงตรวจว่า $\mathbf{u} = (0.15,\ 0.225,\ 0.225,\ 0.15)$ ทำให้สมการทั้งสี่เป็นจริง

แบบฝึกหัด 7.1.2 (⋆) ใช้โซ่สี่ลูกในแบบฝึกหัด 7.1.1

1. จงหาตัวหลัก $u_{11}, u_{22}, u_{33}, u_{44}$ และตัวคูณ $\ell_{21}, \ell_{32}, \ell_{43}$ ด้วยมือ
2. จงแก้ $L\mathbf{y} = \mathbf{b}$ แล้วแก้ $U\mathbf{x} = \mathbf{y}$ และตรวจว่าคำตอบตรงกับแบบฝึกหัด 7.1.1

แบบฝึกหัด 7.1.3 (⋆⋆) ให้ $n \geq 2$ และให้ $u_i = \dfrac{c\,i\,(n+1-i)}{2}$ สำหรับ $i = 0, 1, \ldots, n+1$ (จึงได้ $u_0 = u_{n+1} = 0$)

1. จงแสดงว่า $u_{i-1} - 2u_i + u_{i+1} = -c$ สำหรับ $i = 1, \dots, n$ (ใช้ว่า $q(i) = i(n+1-i)$ เป็นพหุนามกำลังสองในตัวแปร i)
2. จงสรุปว่า $\mathbf{u} = (u_1, \dots, u_n)$ เป็นคำตอบของระบบ $A\mathbf{u} = c\mathbf{1}$ สำหรับทุก $n \geq 2$ เมื่อ A เป็นเมทริกซ์ $\operatorname{tridiag}(-1, 2, -1)$

แบบฝึกหัด 7.1.4 (★★) ให้ T_n เป็นเมทริกซ์ $n \times n$ ที่มี 2 บนแนวทแยงและ -1 บนแนวที่อยู่ติดกัน

1. จงแสดงว่า $u_{11} = 2$ และถ้า $u_{kk} = \frac{k+1}{k}$ แล้วตัวคูณ $\ell_{k+1,k} = -\frac{k}{k+1}$
2. จากข้อ 1 จงแสดงว่า $u_{k+1,k+1} = 2 - \ell_{k+1,k} \cdot (-1) = \frac{k+2}{k+1}$
3. จงใช้การอุปนัยเชิงคณิตศาสตร์สรุปว่าตัวหลัก $u_{kk} = \frac{k+1}{k}$ สำหรับทุก $k = 1, \dots, n$ ของ T_n

แบบฝึกหัด 7.1.5 (★★) ให้ $B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

1. จงหา $\det B$ และคำตอบของ $B\mathbf{x} = (3,\ 5)^T$
2. จงอธิบายว่าทำไมการกำจัดแบบเกาส์โดยไม่สลับแถวหยุดทันทีกับ B
3. ให้ $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ จงแสดงว่า $PB = I$ แล้วอธิบายว่า $PB = LU$ ได้โดย L และ U เป็นอะไร และคำตอบของ $B\mathbf{x} = \mathbf{b}$ หาได้จาก $P\mathbf{b}$ อย่างไร

แบบฝึกหัด 7.1.6 (★★★) ใช้โค้ดของบทนี้ หรือทำด้วยมือบางส่วน

1. จงยืนยันจากตาราง 7.3 ว่าโซ่ห้าลูกต้องใช้ 20 รอบของการทำซ้ำแบบเกาส์–ไซเดลจากศูนย์ จึงมีค่าคลาดเคลื่อนสัมบูรณ์สูงสุดไม่เกิน 10^{-3} และรอบที่ 19 ยังไม่พอ
2. ถ้าแต่ละรอบเก็บเมทริกซ์เต็มและทำการคูณประมาณ n^2 ครั้ง จงเทียบจำนวนการคูณรวมของการทำซ้ำกับจำนวนการคูณและหารของการแยก LU ที่ตาราง 7.2 สำหรับ $n = 5, 10, 20$ แล้วบอกว่าที่ค่า n ใดการทำซ้ำใช้น้อยกว่า
3. ถ้าเปลี่ยน c จาก 0.05 เป็น 0.5 จงทำนายจำนวนรอบที่ต้องใช้เพื่อให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-3} โดยพิจารณาว่าค่าคลาดเคลื่อนสัมบูรณ์ขึ้นกับ c อย่างไร แล้วตรวจด้วยโค้ด
4. จงอธิบายด้วยถ้อยคำของคุณเองว่า ข้อมูลในหัวข้อนี้มีอะไรบ้างที่ช่วยตัดสินว่าวิธีใดเหมาะกับการคำนวณทุกเฟรมของเกมและอะไรยังขาดอยู่

แบบฝึกหัด 7.1.7 (★★★) โครงงานเล็ก: นับและเทียบการแยก LU

1. เขียนโค้ดที่นับการคูณหรือหาร และการบวกหรือลบของการแยก LU สำหรับเมทริกซ์ $M = R + nI$ ขนาด $n \times n$ เมื่อ R เป็นเมทริกซ์สุ่มที่มีค่าจากการแจกแจงปกติ (ตั้งเมล็ดสุ่มไว้) สำหรับ $n = 4, 8, 16$ แล้วเทียบกับสูตรของตำราในหัวข้อ 7.1
2. ใช้สูตร $\frac{n(n^2-1)}{3} + \frac{n(n-1)(2n-1)}{6}$ คำนวณจำนวนการดำเนินการรวมสำหรับ $n = 50$ และ $n = 100$ แล้วหาอัตราส่วนเมื่อ n เพิ่มเป็นสองเท่า
3. ถ้า $n = 1000$ และคอมพิวเตอร์ทำได้ 10^9 การดำเนินการต่อวินาที (สมมติ) จงประมาณเวลาที่ใช้ในการแยก LU
4. (ลองดู) เมทริกซ์ของโซ่มีสมาชิกที่ไม่เป็นศูนย์เพียง $3n - 2$ ตัว ถ้าเรานับเฉพาะการดำเนินการที่แตะสมาชิกเหล่านี้ ตัวเลขจะต่างจากตาราง 7.2 อย่างไร (ไม่ต้องทำให้ลึก เพียงให้เห็นภาพ)

ฉบับร่าง · preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 7.1.1.

1. $c = W/K = 3/40 = 0.075$ เมตร เมทริกซ์คือ

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}, \qquad \mathbf{b} = 0.075 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

2. แทน $\mathbf{u} = (0.15,\ 0.225,\ 0.225,\ 0.15)$: แถวแรก $2(0.15) - 0.225 = 0.075$; แถวที่สอง $-0.15 + 2(0.225) - 0.225 = 0.075$; แถวที่สาม $-0.225 + 2(0.225) - 0.15 = 0.075$; แถวที่สี่ $-0.225 + 2(0.15) = 0.075$ ทั้งสี่สมการเป็นจริง

เฉลย 7.1.2.

1. ตัวหลักและตัวคูณ คำนวณทีละขั้น

   - $u_{11} = 2$
   - $\ell_{21} = \dfrac{-1}{2} = -\dfrac{1}{2}$ และ $u_{22} = 2 - \ell_{21} \cdot (-1) = 2 - \dfrac{1}{2} = \dfrac{3}{2}$
   - $\ell_{32} = \dfrac{-1}{3/2} = -\dfrac{2}{3}$ และ $u_{33} = 2 - \ell_{32} \cdot (-1) = 2 - \dfrac{2}{3} = \dfrac{4}{3}$
   - $\ell_{43} = \dfrac{-1}{4/3} = -\dfrac{3}{4}$ และ $u_{44} = 2 - \ell_{43} \cdot (-1) = 2 - \dfrac{3}{4} = \dfrac{5}{4}$

2. แก้ $L\mathbf{y} = \mathbf{b}$ จากบนลงล่าง: $y_1 = 0.075$; $y_2 = 0.075 - (-\frac{1}{2})(0.075) = 0.1125$; $y_3 = 0.075 - (-\frac{2}{3})(0.1125) = 0.15$; $y_4 = 0.075 - (-\frac{3}{4})(0.15) = 0.1875$ แล้วแก้ $U\mathbf{x} = \mathbf{y}$ จากล่างขึ้นบน: $x_4 = \dfrac{0.1875}{5/4} = 0.15$; $x_3 = \dfrac{0.15 - (-1)(0.15)}{4/3} = \dfrac{0.3}{4/3} = 0.225$; $x_2 = \dfrac{0.1125 - (-1)(0.225)}{3/2} = \dfrac{0.3375}{3/2} = 0.225$; $x_1 = \dfrac{0.075 - (-1)(0.225)}{2} = 0.15$ ได้ $\mathbf{x} = (0.15,\ 0.225,\ 0.225,\ 0.15)$ ตรงกับแบบฝึกหัด 7.1.1

เฉลย 7.1.3.

1. ให้ $q(i) = i(n+1-i) = (n+1)i - i^2$ เราได้ $u_i = \dfrac{c}{2} q(i)$ การคำนวณตรงไปตรงมาให้

$$q(i-1) - 2q(i) + q(i+1) = -2,$$

   เพราะส่วน $(n+1)i$ เป็นเชิงเส้นจึงหักล้างกันหมด และ $-i^2$ ให้ผลต่างที่สองเป็น -2 เมื่อแทนแล้ว $u_{i-1} - 2u_i + u_{i+1} = \dfrac{c}{2} \cdot (-2) = -c$ ตามที่ต้องการ

2. จากข้อ 1 ได้ $2u_i - u_{i-1} - u_{i+1} = c$ สำหรับ $i = 1, \ldots, n$ โดยมี $u_0 = u_{n+1} = 0$ ซึ่งเป็นระบบ $A\mathbf{u} = c\mathbf{1}$ ของเรา เพราะแถวแรกของ A ใช้ $u_0 = 0$ และแถวสุดท้ายใช้ $u_{n+1} = 0$

ตรวจด้วยโค้ด: ทดสอบสำหรับ $n = 2, \ldots, 30$ ใน `test_ch07.py`

ฉบับร่าง Preprint 2026-10-23

เฉลย 7.1.4.

1. $u_{11} = a_{11} = 2 = \frac{1+1}{1}$ ตรงตามสูตร ถ้า $u_{kk} = \frac{k+1}{k}$ แล้วตัวคูณ $\ell_{k+1,k} = \dfrac{a_{k+1,k}}{u_{kk}} = \dfrac{-1}{(k+1)/k} = -\dfrac{k}{k+1}$
2. ในแถว $k+1$ ของ T_n มีสมาชิกไม่เป็นศูนย์เฉพาะคอลัมน์ k และ $k+1$ และในแถว k หลังการกำจัด ยังมีเพียง $u_{k,k+1} = -1$ ที่คอลัมน์ $k+1$ ดังนั้น
$$u_{k+1,k+1} = a_{k+1,k+1} - \ell_{k+1,k}\, u_{k,k+1} = 2 - \Big(-\frac{k}{k+1}\Big)(-1) = 2 - \frac{k}{k+1} = \frac{k+2}{k+1}$$
3. ถ้าสูตรเป็นจริงที่ k ข้อ 2 ให้ $u_{k+1,k+1} = \frac{(k+1)+1}{k+1}$ ซึ่งเป็นสูตรที่ต้องการที่ขั้น $k+1$ เริ่มจากฐาน $k=1$ จึงได้ $u_{kk} = \frac{k+1}{k}$ สำหรับทุก $k = 1, \ldots, n$

ตรวจด้วยโค้ด: ทดสอบสำหรับ $n = 2, \ldots, 11$ ใน `test_ch07.py`

เฉลย 7.1.5.

1. $\det B = 0 \cdot 0 - 1 \cdot 1 = -1 \neq 0$ ดังนั้น B ไม่เอกฐาน ระบบ $B\mathbf{x} = (3,5)^T$ เขียนได้เป็น $x_2 = 3$ จากแถวแรก และ $x_1 = 5$ จากแถวที่สอง จึงได้ $\mathbf{x} = (5,3)$
2. ขั้นแรกของการกำจัดคือหารด้วยตัวหลัก $a_{11} = 0$ ซึ่งหารด้วยศูนย์ไม่ได้ วิธีกำจัดโดยไม่สลับแถวจึงหยุดทันที ทั้งที่ B ไม่เอกฐาน
3. $PB = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$ ดังนั้น $PB = LU$ โดย $L = I$ และ $U = I$ (ทั้งคู่สามเหลี่ยมและมีแนวทแยงที่ถูกต้อง) เมื่อ $B\mathbf{x} = \mathbf{b}$ จะได้ $PB\mathbf{x} = P\mathbf{b}$ จึงได้ $\mathbf{x} = P\mathbf{b} = (b_2,\ b_1)$ เช่น $\mathbf{b} = (3,5)$ ได้ $\mathbf{x} = (5,3)$ ตรงกับข้อ 1

เฉลย 7.1.6.

1. จากตาราง 7.3 ที่รอบ 10 ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดคือ 1.4235×10^{-2} ส่วนรอบ 20 คือ 8.0161×10^{-4} ซึ่งไม่เกิน 10^{-3} ต้องดูรอบระหว่างนี้ด้วย ที่รอบ 19 ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดคือ 1.0688×10^{-3} ซึ่งยังเกิน 10^{-3} จึงสรุปว่า 20 รอบเป็นจำนวนน้อยที่สุด
2. หนึ่งรอบของการทำซ้ำเต็มเมทริกซ์ทำการคูณ $n(n-1)$ ครั้ง (แถวละ $n-1$ ครั้ง) จำนวนการคูณรวมคือจำนวนรอบคูณ $n(n-1)$ และเราได้:
   - $n = 5$: $20 \times 20 = 400$ ครั้ง เทียบกับการแยก LU 40 การคูณหรือหาร
   - $n = 10$: $81 \times 90 = 7\,290$ ครั้ง เทียบกับ 330 การคูณหรือหาร
   - $n = 20$: $355 \times 380 = 134\,900$ ครั้ง เทียบกับ $2\,660$ การคูณหรือหาร

   ในทุกค่า n ที่ทดสอบ การแยก LU ใช้การดำเนินการน้อยกว่า การทำซ้ำเต็มเมทริกซ์ไม่มีจุดที่ได้เปรียบในกรณีนี้ เมื่อโซ่ยาวขึ้น จำนวนรอบเพิ่มประมาณสี่เท่าทุกการเพิ่ม n เป็นสองเท่า จึงแปลว่าการทำซ้ำเต็มเมทริกซ์ใช้งานประมาณ n^4 ส่วน LU ใช้ประมาณ n^3
3. ค่าคลาดเคลื่อนสัมบูรณ์ของการทำซ้ำจากศูนย์เป็นเชิงเส้นในเวกเตอร์ $\mathbf{b}$ ซึ่งคือ $c\mathbf{1}$ ดังนั้นเมื่อ c คูณสิบ ค่าคลาดเคลื่อนสัมบูรณ์ทุกรอบก็คูณสิบด้วย เพื่อให้ค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-3} ที่ $c = 0.5$ เราต้องให้ค่าคลาดเคลื่อนสัมบูรณ์ที่ $c = 0.05$ ไม่เกิน 10^{-4} ซึ่งต้องใช้ 28 รอบ (ตามที่ได้ที่ 10^{-4} ในหัวข้อ 7.1) โค้ดยืนยันว่าได้ 28 รอบจริง

ฉบับร่าง · Preprint · 2026-10-23

4. ข้อมูลในหัวข้อนี้มีสามอย่างที่เกี่ยวข้อง คือจำนวนการคำนวณต่อรอบ จำนวนรอบที่ต้องใช้เพื่อความแม่นที่กำหนด และความเร็วที่ได้จากการหยุดก่อน ข้อมูลที่ยังขาดคือจำนวนรอบจริงที่เกมยอมได้ในหนึ่งเฟรม ค่าเริ่มต้นที่ดีจากเฟรมก่อน และระบบจริงที่มีลูกปัดนับพันลูก ซึ่งเราจะกลับมาพิจารณาในหัวข้อ 7.3

เฉลย 7.1.7.

1. ใช้ `numanal/ch07.py`ซึ่งนับการคูณหรือหารและการบวกหรือลบของ `lu_factor` เมื่อตั้งเมล็ดสุ่มที่ 2 ได้ผลว่าสำหรับ $n = 4, 8, 16$ จำนวนที่นับตรงกับสูตรทุกค่า ซึ่งการทดสอบยืนยันใน `test_ch07.py`แล้ว

2. ผลรวม $n(n^2-1)/3 + n(n-1)(2n-1)/6$ เป็น $661\,650$ สำหรับ $n = 100$ และ $82\,075$ สำหรับ $n = 50$ อัตราส่วนเมื่อเพิ่มเป็นสองเท่าคือ $661\,650/82\,075 \approx 8.06$ ซึ่งใกล้ 8 ตามที่ค่าลำดับนำ $\frac{2}{3}n^3$ บอก

3. ที่ $n = 1000$ การคูณหารคือ $n(n^2-1)/3 = 333\,333\,000$ และการบวกหรือลบคือ $n(n-1)(2n-1)/6 = 332\,833\,500$ รวมประมาณ 6.66×10^8 การดำเนินการ ถ้าเครื่องทำได้ 10^9 ต่อวินาที (สมมติ) ใช้เวลาประมาณ 0.67 วินาที

4. โซ่มีสมาชิกที่ไม่เป็นศูนย์เพียง $3n - 2 = 13$ ตัวสำหรับ $n = 5$ (ห้าตัวบนแนวทแยงและแปดตัวนอกแนวทแยง) ถ้านับเฉพาะการดำเนินการที่แตะสมาชิกเหล่านี้ จำนวนจะน้อยกว่าตาราง 7.2 มาก โดยไม่ต้องมีตัวเลข n^3 นำหน้า ซึ่งจะเป็นเรื่องของหัวข้อ 7.3

7.2 ทำไมต้องสลับแถว และทำไมภาพเบลอแก้คืนยาก

สถานการณ์สมมติ: อีกด้านหนึ่งของปัญหาเดียวกัน

ลองดูแถวของ **พิกเซล** (pixel) สามสิบสองจุดที่มีแถบสว่างอยู่ตรงกลาง การทำให้แถวนี้**เบลอ** (blur) ทำได้ง่ายมาก เพียงให้แต่ละจุดเป็นค่าเฉลี่ยถ่วงน้ำหนักของตัวมันเองกับเพื่อนบ้านสองข้าง ทำซ้ำหนึ่งรอบ แถบจะนุ่มขึ้นนิดหนึ่ง ทำซ้ำสามรอบ แถบจะนุ่มขึ้นอีก

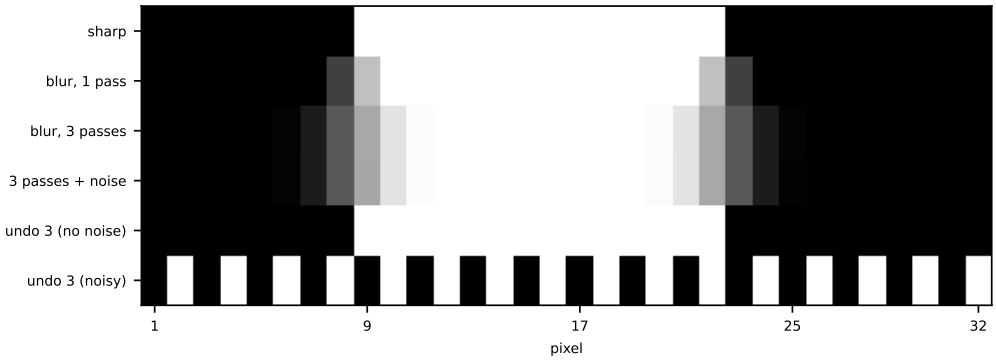
คำถามกลับทิศคือ ถ้ามีแถวที่เบลอแล้ว และอยากได้แถวที่คมเดิมคืนมา เราต้องแก้ระบบสมการ $B^k\mathbf{x} = \mathbf{y}$ ซึ่งมีรูปแบบเดียวกับโซ่ของเรา ในทางคณิตศาสตร์คำตอบตรงเป๊ะ แต่ถ้าแถวที่เบลอมี**สัญญาณรบกวน** (noise) ไม่เกินหนึ่งระดับความเข้มของภาพแปดบิต คำตอบที่คืนมาอาจเพี้ยนจนแทบจำแถบเดิมไม่ได้ ทำไมความเพี้ยนเล็กจึงขยายเป็นความเพี้ยนใหญ่ และเรามีวิธีไหนบ้างที่ช่วยได้

นี่คือหน้าตาอีกด้านของปัญหาเดียวกัน ในโซ่ระบบให้คำตอบที่ดี ที่นี่ระบบมีรูปแบบคล้ายกันแต่ไวต่อข้อมูลมาก (แถบการเบลอ และสัญญาณรบกวนเป็นสถานการณ์ที่เราสร้างขึ้นเอง ไม่ได้บอกว่าแอปหรือกล้องใดทำงานอย่างไร)

ภาพรวมก่อนอ่าน

การเบลอคือการคูณด้วยเมทริกซ์ที่ง่าย แต่การคืนภาพคือการแก้ระบบ และระบบนี้อาจไวมากต่อข้อมูลที่รบกวน ความไวนี้วัดด้วย**คอนดิชันนัมเบอร์** (condition number) ซึ่งเป็นแนวคิดเดียวกับที่บทที่ 2 ใช้วัดปัญหา ถ้าความผิดมาจากการคำนวณ เราช่วยได้ด้วยการเลือกตัวหลักให้ดีและการปรับปรุงคำตอบซ้ำ แต่ถ้าความผิดมาจากข้อมูลที่รบกวนมาแล้ว วิธีคำนวณดีแค่ไหนก็ช่วยไม่ได้

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 7.4: แถวจุดภาพสามสิบสองจุด จากบนลงล่าง: แถวคมเดิม เบลอหนึ่งรอบ เบลอสามรอบ เบลอสามรอบแล้วมีสัญญาณรบกวน คืนจากการเบลอสามรอบโดยไม่มีสัญญาณรบกวน และคืนจากแถวที่มีสัญญาณรบกวน ค่าที่อยู่นอกช่วง 0--1 ถูกตัดเพื่อการแสดงผลเท่านั้น (สร้างจาก `numanal/ch07_blur.py`)

เบลอคือการคูณด้วยเมทริกซ์

ลองให้แต่ละจุดเป็นค่าเฉลี่ยถ่วงน้ำหนักของตัวมันเองกับเพื่อนบ้านสองข้าง โดยใช้น้ำหนัก $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$ จุดนอกแถวนับเป็นศูนย์ ถ้าเรียงค่าของ n จุดเป็นเวกเตอร์ $\mathbf{x}$ การเบลอหนึ่งรอบคือ $\mathbf{y} = B\mathbf{x}$ เมื่อ B เป็นเมทริกซ์ที่มี $\frac{1}{2}$ บนแนวทแยงและ $\frac{1}{4}$ บนแนวที่ติดกัน ตัวอย่างสำหรับ $n = 5$ คือ

$$B = \frac{1}{4}\begin{bmatrix} 2 & 1 & & & \\ 1 & 2 & 1 & & \\ & 1 & 2 & 1 & \\ & & 1 & 2 & 1 \\ & & & 1 & 2 \end{bmatrix}.$$

จุดสว่างจุดเดียวจึงกลายเป็นสามจุดที่สว่างลดหลั่น ถ้า $\mathbf{x} = (0, 0, 4, 0, 0)$ ได้ $B\mathbf{x} = (0, 1, 2, 1, 0)$

ลองเอง

ทำซ้ำอีกหนึ่งรอบด้วยมือ คือคำนวณ $B(0, 1, 2, 1, 0)^T$ ผลที่ควรได้คือ $(\frac{1}{4},\ 1,\ \frac{3}{2},\ 1,\ \frac{1}{4})$ สังเกตว่ายอดลดลงแต่ความกว้างของแถบเพิ่มขึ้น

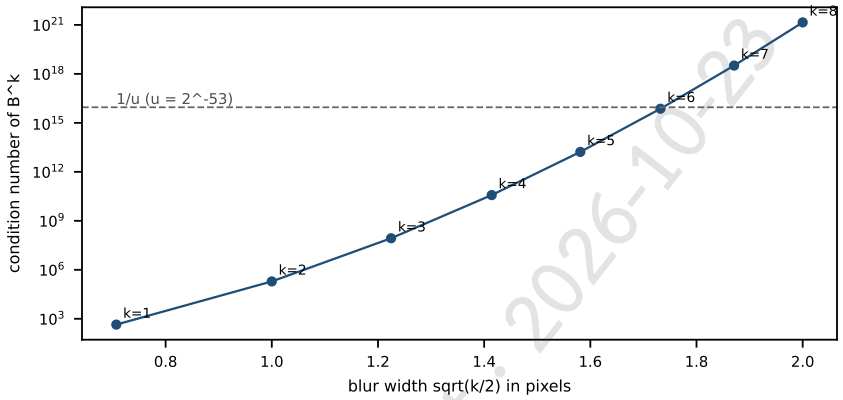
การเบลอแต่ละรอบทำให้จุดหนึ่งกระจายออกไปด้วยความแปรปรวน $\frac{1}{4}\cdot 1^2 + \frac{1}{4}\cdot 1^2 = \frac{1}{2}$ พิกเซล^2 ความแปรปรวนรวมกันเมื่อเบลอหลายรอบ ดังนั้นหลัง k รอบ ความกว้างการเบลอคือ $\sigma = \sqrt{k/2}$ พิกเซล หนึ่งรอบได้ประมาณ 0.707 และสามรอบได้ประมาณ 1.225 พิกเซล

ค่าลักษณะเฉพาะ (eigenvalue) คือตัวเลข λ ที่ทำให้ $B\mathbf{v} = \lambda\mathbf{v}$ มีคำตอบ $\mathbf{v} \neq \mathbf{0}$ มันบอกว่าเมทริกซ์ขยายหรือหดเวกเตอร์ในทิศนั้นกี่เท่า ค่าลักษณะเฉพาะของ B คือ $\lambda_j = \frac{1}{2}\big(1 + \cos\frac{j\pi}{n+1}\big)$ สำหรับ $j = 1, \ldots, n$ ทั้งหมดเป็นบวก และค่าที่เล็กที่สุดสอดคล้องกับลวดลายที่สลับเครื่องหมายไปมาทุกจุด ซึ่งถูกการเบลอกดให้เล็กที่สุด เมื่อ $n = 32$ ค่านี้เท่ากับ 0.002264 เมื่อเบลอ k รอบ ค่าลักษณะเฉพาะของ B^k คือกำลัง k ของค่าเหล่านี้

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 7.4: การคืนแถวจุดภาพที่เบลอ k รอบ ($n = 32$) โดยแก้ $B^k\mathbf{x} = \mathbf{y}$ ด้วยเลือกตัวหลักตามหลัก เมื่อข้อมูลมีและไม่มีสัญญาณรบกวน ค่าที่มากกว่า 1 หมายความว่าคำตอบผิดมากกว่าตัวคำตอบเอง (ตัวเลขจากโค้ดของเรา)

k	σ	$\kappa_2(B^k)$	$\|\mathbf{e}\|/\|\mathbf{y}\|$	ไม่มีสัญญาณรบกวน	มีสัญญาณรบกวน	ขอบเขต $\kappa\|\mathbf{e}\|/\|\mathbf{y}\|$
1	0.707	441	3.33×10^{-3}	3.58×10^{-15}	9.54×10^{-2}	1.47
2	1.00	1.94×10^{5}	3.38×10^{-3}	2.72×10^{-13}	26.8	656
3	1.22	8.56×10^{7}	3.41×10^{-3}	1.24×10^{-9}	1.14×10^{4}	2.92×10^{5}
4	1.41	3.77×10^{10}	3.44×10^{-3}	4.89×10^{-7}	5.01×10^{6}	1.3×10^{8}
5	1.58	1.66×10^{13}	3.47×10^{-3}	2.12×10^{-4}	2.21×10^{9}	5.76×10^{10}
6	1.73	7.32×10^{15}	3.49×10^{-3}	4.32×10^{-2}	9.57×10^{11}	2.56×10^{13}



รูปที่ 7.5: คอนดิชันนัมเบอร์ $\kappa_2(B^k)$ เทียบกับความกว้างการเบลอ $\sqrt{k/2}$ แกนตั้งเป็นลอการิทึม เส้นประคือระดับ $1/u$ เมื่อ $u = 2^{-53} \approx 1.1 \times 10^{-16}$ (สร้างจาก `numanal/ch07_blur.py`)

ฉบับร่าง · Preprint · 2026-10-23

การคืนภาพคือการแก้ระบบ

ถ้ารู้ว่าเบลอกี่รอบ และได้แถว $\mathbf{y}$ มาแล้ว การคืนภาพคือการแก้ $B^k\mathbf{x} = \mathbf{y}$ ระบบนี้ไม่มีอะไรแปลกจากโซ่ในหัวข้อ 7.1 เพราะเป็นเมทริกซ์สมมาตรที่มีแนวทแยงเป็นบวก และแยก LU ได้ ในทางคณิตศาสตร์ $\mathbf{x}$ คือแถวคมเดิม เรื่องยากอยู่ที่เครื่องคำนวณ และอยู่ที่ข้อมูลที่ไม่ตรงกับแถวจริง

ตาราง 7.4 แสดงผลสำหรับแถวสามสิบสองจุดข้างต้น สัญญาณรบกวนสุ่มอยู่ในช่วง $\pm 1/255$ ซึ่งเท่ากับหนึ่งระดับความเข้มของภาพแปดบิต ค่าคลาดเคลื่อนสัมพัทธ์คือ $\|\hat{\mathbf{x}} - \mathbf{x}\|_2 / \|\mathbf{x}\|_2$

แถวแรกบอกเรื่องสำคัญ ข้อมูลผิดราว 0.33% เทียบกับขนาดของ $\mathbf{y}$ แต่คำตอบผิด 9.5% ซึ่งขยายราว 29 เท่า เมื่อเป็นสองรอบ คำตอบผิดมากกว่าตัวมันเอง 26.8 เท่า และเมื่อเป็นสามรอบค่าที่คืนมามีค่าสูงถึงราว $10{,}500$ ทั้งที่ค่าของภาพจริงอยู่ระหว่าง 0 กับ 1 แถวสุดท้ายแสดงอีกเรื่องหนึ่ง แม้ไม่มีสัญญาณรบกวนเลย เมื่อเบลอหกรอบค่าคลาดเคลื่อนสัมพัทธ์ก็ยังถึงราว 4% เพราะการปัดเศษเล็กน้อยถูกขยายด้วย κ

รูป 7.5 แสดงว่า κ เพิ่มเป็นกำลังของจำนวนรอบ จุดที่เส้นประตัดคือจุดที่การปัดเศษอย่างเดียวทำลายคำตอบได้ ถ้าต้องการรู้ว่าความไวนี้วัดอย่างไร เรานิยามคอนดิชันนัมเบอร์ของเมทริกซ์ดังนี้

นิยาม 7.2.1: คอนดิชันนัมเบอร์ของเมทริกซ์

ให้ A เป็นเมทริกซ์ไม่เอกฐานขนาด $n \times n$ และให้ $\|\cdot\|$ เป็นนอร์มเมทริกซ์ที่ถูกเหนี่ยวนำจากนอร์มของเวกเตอร์ เรียก $\kappa(A) = \|A\|\,\|A^{-1}\|$ ว่า **คอนดิชันนัมเบอร์** (condition number) ของ A เมื่อไม่ระบุนอร์ม ให้หมายถึงนอร์มสอง

$\kappa_2(A)$ Ern and Guermond 2004, Definition 9.1, p. 384 Atkinson 1989, §8.4, p. 530

สำหรับเมทริกซ์สมมาตร A ค่าคอนดิชันนัมเบอร์ $\kappa_2(A)$ เท่ากับอัตราส่วนของค่าลักษณะเฉพาะที่ใหญ่สุดต่อค่าที่เล็กสุดในค่าสัมบูรณ์ Ern and Guermond 2004, Proposition 9.2(ii), p. 384 ซึ่งเป็นสูตรที่เราใช้กับ B และ B^k ในหัวข้อนี้

อ่านแบบเข้าใจง่าย

- $\kappa(A)$ เป็นตัวขยาย ถ้าข้อมูลเปลี่ยนสัมพัทธ์เล็กน้อย คำตอบอาจเปลี่ยนสัมพัทธ์มากขึ้นได้ราว $\kappa(A)$ เท่า (ทฤษฎีบทถัดไปบอกขอบเขตที่แน่นอน)
- ระบบที่มี $\kappa(A)$ ใหญ่เรียกว่า**ระบบที่มีสภาวะเลว** (ill-conditioned system) ซึ่งเป็นแนวคิดเดียวกับ**ปัญหาสภาวะเลว**ในหัวข้อ 2.3 คือข้อมูลเปลี่ยนนิดเดียวแล้วคำตอบเปลี่ยนมาก
- $\kappa(A) \geq 1$ เสมอ (แบบฝึกหัด 7.2.3) และ $\kappa(I) = 1$

ลองเอง

จากค่าลักษณะเฉพาะ $\lambda_j = \frac{1}{2}(1 + \cos\frac{j\pi}{4})$ ของ B_3 (สามจุดภาพ) จงหา $\kappa_2(B_3)$ ด้วยมือ ผลที่ควรได้คือประมาณ 5.83 ตรวจว่าตรงกับ $3 + 2\sqrt{2}$ หรือไม่

ทฤษฎีบท 7.2.1: ขอบเขตของการเปลี่ยนแปลงคำตอบ

ให้ A ไม่เอกฐาน $\mathbf{b} \neq \mathbf{0}$ และ $\mathbf{x} = A^{-1}\mathbf{b}$ ให้ $\|\cdot\|$ เป็นนอร์มเมทริกซ์ที่ถูกเหนี่ยวนำ แล้ว

(ก) ถ้า $\tilde{\mathbf{b}} = \mathbf{b} + \delta\mathbf{b}$ และ $\tilde{\mathbf{x}} = A^{-1}\tilde{\mathbf{b}}$ จะได้ $\dfrac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(A)\dfrac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$

(ข) ถ้า $\hat{\mathbf{x}}$ เป็นคำตอบโดยประมาณและ $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$ จะได้ $\dfrac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(A)\dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$

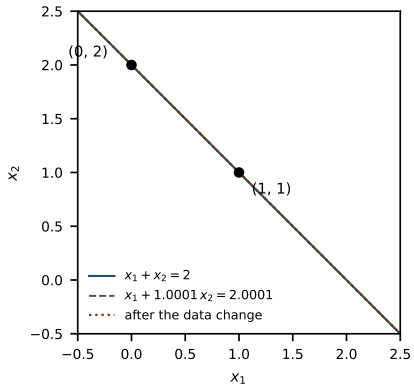
พิสูจน์. **แนวคิด.** ค่าคลาดเคลื่อนสัมบูรณ์ของคำตอบคือ A^{-1} คูณค่าคลาดเคลื่อนสัมบูรณ์ของข้อมูล และเราเทียบขนาดกับ $\mathbf{b} = A\mathbf{x}$ ในทั้งสองส่วน

1. **(ก).** ให้ $\delta\mathbf{x} = \tilde{\mathbf{x}} - \mathbf{x} = A^{-1}\delta\mathbf{b}$ จากนิยามของนอร์มเหนี่ยวนำได้ $\|\delta\mathbf{x}\| \leq \|A^{-1}\|\,\|\delta\mathbf{b}\|$
2. **เทียบกับขนาดของ $\mathbf{x}$.** จาก $\mathbf{b} = A\mathbf{x}$ ได้ $\|\mathbf{b}\| \leq \|A\|\,\|\mathbf{x}\|$ เพราะ $\mathbf{x} \neq \mathbf{0}$ จึงได้ $\dfrac{1}{\|\mathbf{x}\|} \leq \dfrac{\|A\|}{\|\mathbf{b}\|}$
3. **รวมกัน.** คูณอสมการในขั้น 1 ด้วยอสมการในขั้น 2 จะได้ $\dfrac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\|\,\|A^{-1}\|\dfrac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$ ซึ่งคือ (ก)
4. **(ข).** ให้ $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$ จาก $A\mathbf{e} = \mathbf{b} - A\hat{\mathbf{x}} = \mathbf{r}$ ได้ $\mathbf{e} = A^{-1}\mathbf{r}$ แล้วทำขั้น 1–3 เดิมโดยแทน $\delta\mathbf{b}$ ด้วย $\mathbf{r}$ และ $\tilde{\mathbf{x}}$ ด้วย $\hat{\mathbf{x}}$

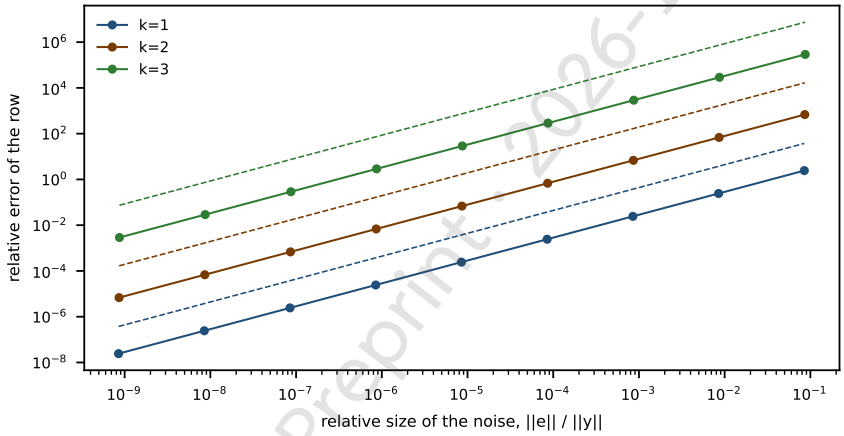
สรุป: ค่าคลาดเคลื่อนสัมพัทธ์ของคำตอบถูกขยายด้วย $\kappa(A)$ เมื่อเทียบกับค่าคลาดเคลื่อนสัมพัทธ์ของข้อมูลหรือค่าตกค้าง □

ตำรา Atkinson 1989, §8.4, p. 530 บอกว่า สำหรับ A ที่กำหนด มี $\mathbf{b}$ และ $\mathbf{r}$ ที่ทำให้อสมการทั้งสองเป็นสมการได้ ขอบเขตจึงแน่นตามความหมายนั้น

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 7.6: เส้นสองเส้นที่ชันเกือบเท่ากัน เส้นทึบคือ $x_1 + x_2 = 2$ เส้นประคือ $x_1 + 1.0001\,x_2 = 2.0001$ และเส้นจุดคือเส้นหลังการเปลี่ยนข้อมูลเป็น 2.0002 ทั้งสามเกือบซ้อนกันในภาพ จุด $(1, 1)$ และ $(0, 2)$ อยู่บนเส้นเดียวกันโดยประมาณ การเปลี่ยนข้อมูลเล็กน้อยจึงเลื่อนจุดตัดไปตามเส้นนั้นได้ไกล (สร้างจาก `numanal/ch07_blur.py`)



รูปที่ 7.7: ค่าคลาดเคลื่อนสัมพัทธ์ของแถวที่คืนมา เทียบกับขนาดสัมพัทธ์ของสัญญาณรบกวน บนแกนลอการิทึม จุดเป็นค่าที่คำนวณได้ เส้นประของสีเดียวกันคือขอบเขต $\kappa_2(B^k)\,\|\mathbf{e}\|/\|\mathbf{y}\|$ ของทฤษฎีบท (ก) สีแยกตามจำนวนรอบ $k = 1, 2, 3$ (สร้างจาก `numanal/ch07_blur.py`)

ฉบับร่าง · Preprint · 2026-10-23

ตัวอย่างเส้นเกือบขนาน. ลองดูระบบสองสมการ $x_1 + x_2 = 2$ กับ $x_1 + 1.0001\,x_2 = 2.0001$ เป็นเส้นตรงสองเส้นที่ชันเกือบเท่ากัน คำตอบคือ $(1, 1)$ ซึ่งเป็นจุดที่เส้นทั้งสองตัดกัน ถ้าข้อมูลด้านขวาขยับเป็น $(2,\ 2.0002)$ ซึ่งเปลี่ยนไปเพียง 5×10^{-5} สัมพัทธ์ คำตอบจะกลายเป็น $(0, 2)$ ซึ่งเปลี่ยนไปถึง 100% ค่า $\kappa_\infty(A)$ ของระบบนี้ประมาณ $40\,004$ ทั้งที่ A มีสมาชิกขนาดราว 1 ขอบเขตของทฤษฎีบท (ก) ให้ $\kappa_\infty(A) \cdot 5 \times 10^{-5} \approx 2.0$ ซึ่งไม่ขัดกับผลจริง เพราะขอบเขตเป็นค่าสูงสุดที่เกิดได้ ไม่ได้บอกว่าจะเกิดเท่านั้นเสมอไป

รูป 7.7 แสดงว่าค่าคลาดเคลื่อนสัมบูรณ์เพิ่มเป็นสัดส่วนกับสัญญาณรบกวน (เส้นตรงในกราฟลอการิทึม) และทุกจุดอยู่ใต้เส้นประ เมื่อ $k = 1$ อัตราส่วนของค่าคลาดเคลื่อนสัมบูรณ์ต่อสัญญาณรบกวนจริงเท่ากับราว 29 ซึ่งต่ำกว่าค่า $\kappa_2(B) \approx 441$ มาก ขอบเขตจึงมองโลกในแง่ร้ายเกินจริงได้ แต่ไม่เคยต่ำกว่าความจริง

ตารางที่ 7.5: ค่าตกค้างสัมพัทธ์เล็กถึงระดับเครื่อง แต่คำตอบห่างจากแถวคมมาก เมื่อข้อมูล $\tilde{\mathbf{y}}$ มีสัญญาณรบกวน ($n = 32$; $\tilde{\mathbf{x}} = A^{-1}\tilde{\mathbf{y}}$; ตัวเลขจากโค้ดของเรา)

k	$\|\mathbf{r}\|/\|\tilde{\mathbf{y}}\|$	$\|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|/\|\tilde{\mathbf{x}}\|$	ขอบเขต $\kappa\|\mathbf{r}\|/\|\tilde{\mathbf{y}}\|$	$\|\hat{\mathbf{x}} - \mathbf{x}\|/\|\mathbf{x}\|$
1	4.32×10^{-17}	4.59×10^{-15}	1.90×10^{-14}	9.54×10^{-2}
2	1.89×10^{-15}	1.59×10^{-13}	3.66×10^{-10}	26.8
3	7.21×10^{-13}	5.90×10^{-11}	6.17×10^{-5}	1.14×10^{4}

ข้อควรระวัง: กฎคร่าว ๆ ของการเสียหลัก

ตำรา Atkinson 1989, §8.4, p. 538 ให้กฎคร่าว ๆ ว่า ถ้า $\kappa(A) \approx 10^{m}$ คำตอบจะเสียหลักราว m หลักเทียบกับจำนวนหลักของเลขที่ใช้คำนวณ เมื่อเบลอหกรอบ $\kappa \approx 7 \times 10^{15}$ ซึ่งเกินจำนวนหลักของเลขจุดลอยตัว 64 บิตแทบทั้งหมด จึงเข้าใจได้ว่าทำไมคำตอบในตารางพังแม้ไม่มีสัญญาณรบกวน

ค่าตกค้างเล็กแต่คำตอบผิด

ในหัวข้อ 7.1 เราเขียนว่า $\mathbf{e} = A^{-1}\mathbf{r}$ จึงรู้ว่าค่าตกค้างกับค่าคลาดเคลื่อนสัมบูรณ์เชื่อมกันด้วย A^{-1} ตารางต่อไปนี้แสดงตัวเลขจริงของการคืนแถวที่มีสัญญาณรบกวนจากหัวข้อนี้ โดยแก้ $B^{k}\hat{\mathbf{x}} = \tilde{\mathbf{y}}$ ด้วย LU ในเลข 64 บิต

ค่าตกค้างของแถวที่สามเล็กเพียง 7×10^{-13} ทฤษฎีบท (ข) ให้ว่าคำตอบห่างจาก $\tilde{\mathbf{x}}$ ไม่เกินค่านี้คูณ κ ซึ่งจริงตามตาราง แต่คำตอบห่างจากแถวคมเดิมถึง 1.14×10^{4} เท่า ไม่ขัดกันเลย เพราะค่าตกค้างวัดว่าคำตอบแก้ระบบที่ข้อมูลกำหนดได้ดีแค่ไหน ส่วนสัญญาณรบกวนที่ติดอยู่ในข้อมูลไม่ปรากฏในค่าตกค้างเลย

ข้อควรระวัง: อย่าตรวจคำตอบด้วยค่าตกค้างอย่างเดียว

ค่าตกค้างที่เล็กเป็นหลักฐานว่าการแก้ทำได้ดี แต่ไม่เป็นหลักฐานว่าคำตอบตรงกับความจริง ถ้าข้อมูลรบกวนมาแล้ว ค่าตกค้างอาจเล็กจนเชื่อได้ ขณะที่คำตอบเพี้ยนมาก เพื่อตรวจว่าคำตอบแม่นจริงเราต้องรู้ κ ของระบบด้วย

สลับแถวเมื่อตัวหลักเล็กหรือเป็นศูนย์

ในหัวข้อ 7.1 เราพบว่าการกำจัดแบบเกาส์หยุดเมื่อตัวหลักเป็นศูนย์ แม้เมทริกซ์จะไม่เอกฐาน เช่น $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ ทางแก้คือสลับแถว แต่ตัวหลักที่เล็กมากแต่ไม่เป็นศูนย์ก็อันตรายไม่น้อย ลองดูระบบ

$$\begin{bmatrix} 10^{-5} & 1 \\ 1 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

คำตอบจริงคือ $x_1 = \dfrac{1}{1 - 10^{-5}} \approx 1.00001$ และ $x_2 \approx 0.99999$ ลองแก้ด้วยเลขทศนิยมสี่หลักนัยสำคัญ คือทุกการคำนวณถูกปัดเหลือสี่หลัก

- **ไม่สลับแถว.** ตัวคูณ $\ell_{21} = 10^{5}$ แถวที่สองได้ $u_{22} = 1 - 10^{5} = -99999 \approx -1.000 \times 10^{5}$ และ $y_2 = 2 - 10^{5} \approx -1.000 \times 10^{5}$ เลข 1 หายไปในผลลัพธ์ จึงได้ $x_2 = 1.000$ และ $x_1 = \dfrac{1 - 1.000}{10^{-5}} = 0$ คำตอบ x_1 ผิด 100%
- **สลับแถวก่อน.** สลับเป็น $\begin{bmatrix} 1 & 1 \\ 10^{-5} & 1 \end{bmatrix}$ กับ $(2, 1)$ ตัวคูณ $\ell_{21} = 10^{-5}$ แถวที่สองได้ $u_{22} = 1 - 10^{-5} \approx 1.000$ และ $y_2 = 1 - 2 \times 10^{-5} \approx 1.000$ จึงได้ $x_2 = 1.000$ และ $x_1 = 2 - 1.000 = 1.000$ ผลถูกต้องถึงสี่หลัก

ฉบับร่าง · Preprint 2026-10-23

ทำไมต่างกัน การหารด้วยตัวหลักเล็กทำให้ตัวคูณใหญ่ และตัวคูณใหญ่ทำให้เลขที่เล็กกว่าหายไปในผลรวม การสลับแถวจะเลือกตัวหลักที่ใหญ่ที่สุดในคอลัมน์ ทำให้ตัวคูณทุกตัวมีค่าสัมบูรณ์ไม่เกิน 1 ด้วยเลขจุดลอยตัว 64 บิต ระบบเดียวกันที่ 10^{-17} ก็เกิดเรื่องแบบเดียวกันได้ ซึ่งโค้ดของเราแสดงให้เห็นว่าคำตอบเป็น $(0, 1)$ เมื่อไม่สลับ และ $(1, 1)$ เมื่อสลับ

นิยาม 7.2.2: เลือกตัวหลักตามหลัก

ในขั้นที่ k ของการกำจัดแบบเกาส์ ให้เลือกแถว $i \geq k$ ที่ทำให้ $|a_{ik}|$ มากที่สุด (ถ้ามีหลายแถวให้เลือกแถวที่อยู่บนสุด) แล้วสลับแถว k กับแถว i ก่อนกำจัด เรียกวิธีนี้ว่า **เลือกตัวหลักตามหลัก** (partial pivoting) ตำรา Iserles เรียกวิธีเดียวกันว่าการเลือกตัวหลักตามคอลัมน์ Iserles 2009, ภาคผนวก A.1.4.4, p. 435 และตำรา Atkinson ให้นิยามเดียวกันว่า partial pivoting Atkinson 1989, §8.2, p. 515

อ่านแบบเข้าใจง่าย. พูดง่าย ๆ คือ ก่อนหารด้วยตัวหลัก ให้ยกตัวที่ใหญ่ที่สุดในคอลัมน์ขึ้นมาเป็นตัวหลัก ผลคือตัวคูณทั้งหมดไม่เกิน 1 ในค่าสัมบูรณ์ วิธีนี้ทำให้ตัวปัดเศษไม่ขยายตามตัวคูณใหญ่

ทฤษฎีบท 7.2.2: การแยกตัวประกอบ $PA = LU$

ถ้า A เป็นเมทริกซ์จัตุรัสไม่เอกฐาน จะมี**เมทริกซ์การเรียงสับเปลี่ยน** (permutation matrix) P ซึ่ง $PA = LU$ โดย L เป็นสามเหลี่ยมล่างที่มีหนึ่งทุกช่องบนแนวทแยง และ U เป็นสามเหลี่ยมบน ถ้าใช้เลือกตัวหลักตามหลักทุกขั้น ทุกตัวคูณใน L มีค่าสัมบูรณ์ไม่เกิน 1 Atkinson 1989, §8.2, สมการ (8.2.7), p. 517

เราไม่พิสูจน์ทฤษฎีบทนี้ที่นี่ แนวคิดคือ เมื่อ A ไม่เอกฐาน คอลัมน์ที่ยังเหลือในแต่ละขั้นต้องมีสมาชิกที่ไม่เป็นศูนย์ จึงสลับแถวนั้นขึ้นมาเป็นตัวหลักได้เสมอ (อ้างจากตำราที่ระบุข้างบน) ตัวอย่างนี้ต้องใช้ $PA = LU$ คือ $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ และ $L = U = I$ ตามแบบฝึกหัด 7.1.5

เกร็ดเพิ่มเติม

ทำไมไม่เลือกตัวหลักจากทั้งเมทริกซ์ ตำรา Atkinson อธิบายว่าการเลือกตัวหลักแบบเต็มทำให้ค่าคลาดเคลื่อนจากการปัดเศษแพร่ช้าอย่างพิสูจน์ได้ แต่ต้องเปรียบเทียบมากกว่ามากในแต่ละขั้น ทฤษฎีของเลือกตัวหลักตามหลักอ่อนกว่าเล็กน้อย แต่ในปัญหาปฏิบัติส่วนใหญ่พฤติกรรมเกือบเหมือนกัน จึงนิยมใช้เลือกตัวหลักตามหลัก Atkinson 1989, §8.2, p. 516

ในระบบเบลอของเรา เมื่อ $k = 1$ โค้ดของเราไม่สลับแถวเลยสักขั้น แต่เมื่อ $k = 2$ การสลับแถวเกิดขึ้นจริง ดูขั้นที่สองของ B^2 เมื่อ $n = 5$ โดยคูณทั้งเมทริกซ์ด้วย 16 ให้เป็นจำนวนเต็ม

$$16B^2 = \begin{bmatrix} 5 & 4 & 1 & 0 & 0 \\ 4 & 6 & 4 & 1 & 0 \\ 1 & 4 & 6 & 4 & 1 \\ 0 & 1 & 4 & 6 & 4 \\ 0 & 0 & 1 & 4 & 5 \end{bmatrix}$$

ขั้นแรกใช้ตัวหลัก $\frac{5}{16}$ ซึ่งเป็นตัวที่ใหญ่ที่สุดในคอลัมน์แรก หลังกำจัดแล้ว สมาชิกที่เหลือในคอลัมน์ที่สองคือ แถวที่สองได้ $\frac{3}{8} - \frac{(1/4)(1/4)}{5/16} = \frac{7}{40} = 0.175$ และแถวที่สามได้ $\frac{1}{4} - \frac{(1/16)(1/4)}{5/16} = \frac{1}{5} = 0.200$ ตัวหลักของขั้นที่สองจึงต้องเป็นแถวที่สาม ซึ่งต้องสลับกับแถวที่สอง โค้ดของเรายืนยันการสลับนี้ และเมื่อ $n = 32$ การเลือกตัวหลักสลับแถวใน 30 จาก 31 ขั้น ดังนั้นเมื่อเบลอตั้งแต่สองรอบขึ้นไป การสลับแถวจึงเป็นส่วนของการแก้คืนจริง ๆ

ฉบับร่าง · Preprint · 2026-10-23

ข้อควรระวัง: สลับแถวแก้วิธี ไม่ได้แก้ปัญหา

เลือกตัวหลักตามหลักแก้ค่าคลาดเคลื่อนที่มาจากตัวหลักเล็ก ซึ่งเป็นเรื่องของวิธี แต่ไม่ได้ลดคอนดิชันนัมเบอร์ของระบบ ถ้า $\kappa(A)$ ใหญ่ การสลับแถวไม่ช่วยให้ข้อมูลที่รบกวนมาแล้วกลับมาแม่นขึ้น ดูแบบฝึกหัด 7.2.6 ที่ระบบมี κ เพียงราว 4 แต่วิธีที่ไม่สลับแถวให้คำตอบผิด

การปรับปรุงคำตอบซ้ำ

ถ้าเราได้คำตอบ $\hat{\mathbf{x}}$ จากการแยก LU ที่คำนวณด้วยเลขความแม่นต่ำ การปรับปรุงคำตอบซ้ำ (**การปรับปรุงคำตอบซ้ำ** (iterative refinement)) คือการใช้ค่าตกค้างของคำตอบนั้นเพื่อปรับแก้ซ้ำ ตำรา Atkinson เรียกวิธีนี้ว่าการแก้ค่าตกค้าง (residual correction) Atkinson 1989, §8.5, pp. 540--541 มีขั้นตอนดังนี้

1. แยก A เป็น LU ครั้งเดียวด้วยเลขความแม่นต่ำ
2. คำนวณค่าตกค้าง $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$ ด้วยเลขความแม่นสูงกว่า
3. แก้ $A\mathbf{d} = \mathbf{r}$ ด้วย LU เดิม จะได้ค่าประมาณค่าคลาดเคลื่อนสัมบูรณ์ $\mathbf{d}$
4. ตั้ง $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} + \mathbf{d}$ แล้วกลับไปข้อ 2

การแยก LU ทำครั้งเดียว ส่วนแต่ละรอบของการปรับปรุงใช้การคูณเมทริกซ์กับเวกเตอร์และการแทนค่าสองครั้ง ซึ่งตำรา Atkinson 1989, §8.5, p. 541 ประมาณว่าใช้ราว n^2 ต่อขั้น ดังนั้นเมื่อ $n = 32$ การแยก LU ใช้ราว $\frac{2}{3}n^3 \approx 21\,800$ การดำเนินการ ส่วนหนึ่งรอบของการปรับปรุงใช้ราว $2n^2 = 2\,048$ การดำเนินการ ปรับซ้ำได้ราวสิบรอบในราคาเท่ากับการแยกหนึ่งครั้ง และตำราเตือนว่าค่าตกค้างต้องคำนวณให้แม่น เพราะ $A\hat{\mathbf{x}}$ กับ $\mathbf{b}$ เกือบเท่ากัน จนการคำนวณด้วยเลขความแม่นเดียวกับที่แยกจะสูญเสียหลักไปมาก

ทฤษฎีบท 7.2.3: การลู่เข้าของการปรับปรุงซ้ำ

ให้ A ไม่เอกฐาน ให้ C เป็นเมทริกซ์ และให้ $\mathbf{x}_{m+1} = \mathbf{x}_m + C(\mathbf{b} - A\mathbf{x}_m)$ เริ่มจาก $\mathbf{x}_0$ ถ้า $\|I - CA\| < 1$ สำหรับนอร์มเหนี่ยวนำบางนอร์ม แล้ว $\|\mathbf{x} - \mathbf{x}_m\| \le \|I - CA\|^m \|\mathbf{x} - \mathbf{x}_0\|$ เมื่อ $\mathbf{x} = A^{-1}\mathbf{b}$ ดังนั้น $\mathbf{x}_m \to \mathbf{x}$

พิสูจน์. **แนวคิด.** ค่าคลาดเคลื่อนสัมบูรณ์ของรอบถัดไปเท่ากับ $I - CA$ คูณค่าคลาดเคลื่อนสัมบูรณ์ของรอบปัจจุบัน

1. เพราะ $\mathbf{b} = A\mathbf{x}$ จึงได้ $\mathbf{b} - A\mathbf{x}_m = A(\mathbf{x} - \mathbf{x}_m)$ แทนลงในนิยามของ $\mathbf{x}_{m+1}$ จะได้ $\mathbf{x} - \mathbf{x}_{m+1} = (\mathbf{x} - \mathbf{x}_m) - CA(\mathbf{x} - \mathbf{x}_m) = (I - CA)(\mathbf{x} - \mathbf{x}_m)$
2. จากอสมการของนอร์มเหนี่ยวนำ $\|\mathbf{x} - \mathbf{x}_{m+1}\| \le \|I - CA\|\,\|\mathbf{x} - \mathbf{x}_m\|$
3. ทำซ้ำโดยอุปนัยเชิงคณิตศาสตร์บน m จะได้ $\|\mathbf{x} - \mathbf{x}_m\| \le \|I - CA\|^m \|\mathbf{x} - \mathbf{x}_0\|$ (ฐาน $m = 0$ เห็นได้ทันที)
4. ถ้า $\|I - CA\| < 1$ ด้านขวาไปที่ศูนย์เมื่อ $m \to \infty$ จึงได้ $\mathbf{x}_m \to \mathbf{x}$

สรุป: การปรับปรุงซ้ำลู่เข้าเมื่อ C ทำให้ $I - CA$ เล็กกว่าหนึ่ง คือ C ต้องประมาณ A^{-1} ได้ดี □

ตัวแก้ด้วย LU ที่คำนวณด้วยเลขความแม่นต่ำไม่ใช่เมทริกซ์ C ที่แน่นอนเสมอไป ทฤษฎีบทให้เงื่อนไขพอ และตำรา Atkinson 1989, §8.5, pp. 541--542 พูดถึงเงื่อนไขที่ทำให้ $\|I - CA\|$ เล็กพอ เราดูผลจริงในตารางต่อไปนี้

ฉบับร่าง preprint 2026-10-23

ตารางที่ 7.6: การปรับปรุงซ้ำของ B (หนึ่งรอบ, $n = 32$) โดยแยก LU ในเลขความแม่นเดี่ยว (32 บิต) และคำนวณค่าตกค้างในเลขจุดลอยตัว 64 บิต (ตัวเลขจากโค้ดของเรา)

รอบ m	ไม่มีสัญญาณรบกวน (เทียบกับ $\mathbf{x}$)	มีสัญญาณรบกวน (เทียบกับ $A^{-1}\tilde{\mathbf{b}}$)
0	4.08×10^{-6}	1.54×10^{-6}
1	1.97×10^{-12}	4.57×10^{-13}
2	1.07×10^{-18}	1.75×10^{-15}

ตารางที่ 7.7: LU ในเลขความแม่นเดี่ยวสำหรับ B^k ($n = 32$, ไม่มีสัญญาณรบกวน) เทียบ $\kappa \cdot 2^{-24}$ กับค่าคลาดเคลื่อนสัมพัทธ์ก่อนและหลังการปรับปรุงสองรอบ (ตัวเลขจากโค้ดของเรา)

k	$\kappa_2(B^k)$	$\kappa \cdot 2^{-24}$	ก่อนปรับปรุง	หลังปรับปรุงสองรอบ
1	441	2.6×10^{-5}	4.08×10^{-6}	1.07×10^{-18}
2	1.94×10^{5}	1.2×10^{-2}	3.67×10^{-4}	3.54×10^{-11}
3	8.56×10^{7}	5.1	8.56×10^{-3}	5.79×10^{-6}

ตารางแสดงว่า การปรับปรุงซ้ำแก้ค่าคลาดเคลื่อนของการคำนวณได้ดีมาก เพียงสองรอบก็ถึงระดับของเลข 64 บิตเมื่อข้อมูลไม่มีสัญญาณรบกวน แต่เมื่อข้อมูลมีสัญญาณรบกวน คอลัมน์ขวาแสดงว่าทุกรอบคำตอบลู่เข้าสู่คำตอบของระบบที่ข้อมูลผิดไปแล้วอย่างแม่นยำ ขณะที่คำตอบนั้นยังห่างจากแถวคมเดิมเท่าเดิม ราว 9.5%

ตารางต่อไปนี้ใช้ LU ในเลขความแม่นเดี่ยวกับแถวที่ไม่มีสัญญาณรบกวน เพื่อดูว่าค่าคลาดเคลื่อนขึ้นกับ κ อย่างไร

ค่าคลาดเคลื่อนก่อนปรับปรุงอยู่ใต้ $\kappa \cdot 2^{-24}$ ในทุกแถว สอดคล้องกับการที่การปัดเศษถูกขยายด้วย κ ที่ $k = 3$ ค่า $\kappa \cdot 2^{-24}$ มีค่าประมาณ 5 ซึ่งเป็นเพียงตัวบ่งชี้ว่าทฤษฎีบทข้างบนอาจไม่รับประกันการลู่เข้าแล้ว ในการรันของเรา ค่าคลาดเคลื่อนสัมพัทธ์ยังลดลงทุกรอบ จาก 8.6×10^{-3} เป็น 7.5×10^{-4} แล้วเป็น 5.8×10^{-6} แต่การลดลงไม่สม่ำเสมอ และเราไม่มีทฤษฎีบทรับประกันมาก่อน จึงไม่ควรวางใจการลู่เข้าเพียงเพราะเห็นมันลดลงในไม่กี่รอบ

ข้อควรระวัง: การปรับปรุงซ้ำแก้การคำนวณ ไม่ได้แก้ข้อมูล

นี่คือเรื่องปัญหาไม่ดีกับวิธีไม่ดีในหัวข้อ 2.3 ความผิดที่มาจากการปัดเศษเป็นความผิดของวิธี การปรับปรุงซ้ำช่วยได้ แต่สัญญาณรบกวนในข้อมูลเป็นความผิดของปัญหา มันติดไปกับคำตอบที่ได้ไม่ว่าจะปรับซ้ำกี่รอบ

เกร็ดเพิ่มเติม

คอนดิชันนัมเบอร์ยังวัดระยะห่างสัมพัทธ์จาก A ไปยังเมทริกซ์เอกฐานที่ใกล้ที่สุด ตามทฤษฎีบทของ Gastinel ที่ตำราอ้าง $1/\kappa(A) = \min\{\|A - B\|/\|A\|\}$ เมื่อ B ทั่วไปเป็นเมทริกซ์เอกฐาน Atkinson 1989, §8.4, Theorem 8.3, p. 533 ดังนั้นเมทริกซ์ที่มี κ ใหญ่คือเมทริกซ์ที่อยู่ใกล้เอกฐานมาก

สรุปใจความ

- การเบลอคือการคูณด้วย B^k ซึ่งทำได้ง่าย แต่การคืนภาพคือการแก้ระบบ ซึ่งไวต่อข้อมูลตาม $\kappa_2(B^k)$ ที่เพิ่มเป็นกำลัง k
- ค่าคลาดเคลื่อนสัมพัทธ์ของคำตอบถูกขยายด้วย $\kappa(A)$ เทียบกับค่าคลาดเคลื่อนสัมพัทธ์ของข้อมูลหรือค่าตกค้าง
- ตัวหลักเล็กทำให้ผลพัง เลือกตัวหลักตามหลักให้ $PA = LU$ ที่ตัวคูณไม่เกิน 1 และในระบบเบลอตั้งแต่สองรอบ

ฉบับร่าง · Preprint · 2026-10-23

ขึ้นไปต้องสลับแถวจริง

- การปรับปรุงซ้ำแก้ค่าคลาดเคลื่อนจากการคำนวณได้ แต่แก้ข้อมูลที่รบกวนแล้วไม่ได้
- ค่าตกค้างเล็กเป็นหลักฐานของการแก้ที่ดี แต่ไม่รับประกันว่าคำตอบตรงกับแถวคมเดิม เพราะไม่เห็นสัญญาณรบกวนในข้อมูล

อ่านต่อ. ตำรา Atkinson 1989, §7.4, Thm 7.12, p. 493 ให้ขอบเขตของการผกผันเมื่อเมทริกซ์เปลี่ยนนิดเดียว Atkinson 1989, §8.4--8.5, pp. 529--544 ให้การวิเคราะห์ค่าคลาดเคลื่อนของการแก้ระบบอย่างละเอียด ส่วน Ern and Guermond 2004, Definition 9.1 และ Proposition 9.5, pp. 384--385 นิยามคอนดิชันนัมเบอร์ในบริบทของไฟไนต์เอลิเมนต์ และ Iserles 2009, ภาคผนวก A.1.4.4, p. 435 อธิบายการเลือกตัวหลักตามคอลัมน์เป็นวิธีมาตรฐาน

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 7.2.1 (⋆) ให้แถวจุดภาพสามจุด และเมทริกซ์การเบลอหนึ่งรอบ $B_3 = \frac{1}{4}\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$

1. จงคำนวณ $B_3(1,2,1)^T$ แล้วเทียบค่าที่สูงที่สุดของแถวก่อนและหลังการเบลอ
2. ค่าลักษณะเฉพาะของ B_3 คือ $\lambda_j = \frac{1}{2}\big(1+\cos\frac{j\pi}{4}\big)$ สำหรับ $j=1,2,3$ จงหาค่าทั้งสามเป็นตัวเลขสี่ตำแหน่ง
3. จงหา $\kappa_2(B_3)$ ด้วยมือ และแสดงว่าได้ $3+2\sqrt{2}$

แบบฝึกหัด 7.2.2 (⋆) แก้ระบบ $\begin{bmatrix} 0.0005 & 2 \\ 1 & 3 \end{bmatrix}\mathbf{x} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$ โดยใช้เลขทศนิยมสี่หลักนัยสำคัญ ทุกการคำนวณปัดเป็นสี่หลัก (ปัดคู่เมื่อครึ่ง)

1. แก้โดยกำจัดแบบเกาส์โดยไม่สลับแถว ทำทุกขั้นด้วยมือ
2. แก้อีกครั้งโดยสลับแถวก่อนขั้นแรก ทำทุกขั้นด้วยมือ
3. คำตอบจริงคือ $x_1 \approx -1.00075$ และ $x_2 \approx 2.00025$ จงหาค่าคลาดเคลื่อนสัมพัทธ์ของ x_1 ในข้อ 1 และข้อ 2 แล้วบอกว่าวิธีใดดีกว่า

แบบฝึกหัด 7.2.3 (⋆⋆) ให้ A เป็นเมทริกซ์ไม่เอกฐาน และให้ $\|\cdot\|$ เป็นนอร์มเมทริกซ์ที่ถูกเหนี่ยวนำจากนอร์มของเวกเตอร์

1. จงแสดงว่า $\|I\| = 1$
2. จงแสดงว่า $\kappa(A) = \|A\|\,\|A^{-1}\| \geq 1$ โดยใช้ $I = AA^{-1}$ และอสมการ $\|PQ\| \leq \|P\|\,\|Q\|$
3. จงหา $\kappa_1(A)$ ของ $A = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$ ด้วยมือ (ใช้นอร์มหนึ่ง คือผลรวมค่าสัมบูรณ์ของคอลัมน์ที่มากที่สุด)

แบบฝึกหัด 7.2.4 (⋆⋆) กลับไปที่การเบลอ ให้ B เป็นเมทริกซ์การเบลอหนึ่งรอบขนาด $n \times n$ และ B^k คือการเบลอ k รอบ

1. ถ้า B สมมาตรและมีค่าลักษณะเฉพาะบวก $\lambda_1, \ldots, \lambda_n$ จงแสดงว่า $\kappa_2(B^k) = \kappa_2(B)^k$ (ใช้ว่า B^k มีค่าลักษณะเฉพาะ λ_j^k และเวกเตอร์ลักษณะเฉพาะเดิม)

ฉบับร่าง · Preprint · 2025-10-23

2. ใช้สูตร $\kappa_2(B) = \dfrac{1+\cos\big(\pi/(n+1)\big)}{1-\cos\big(\pi/(n+1)\big)}$ จงหาค่า k ที่น้อยที่สุดที่ทำให้ $\kappa_2(B^k)\,2^{-53} \geq 1$ สำหรับ $n = 16, 32$ และ 64

3. จากค่า k ในข้อ 2 จงหาความกว้างการเบลอ $\sqrt{k/2}$ ของแต่ละ n และอธิบายว่าเหตุใดระบบเดียวกันจึงไวกว่าเมื่อการเบลอกว้างขึ้น

แบบฝึกหัด 7.2.5 (⋆⋆) ให้ A และ C เป็นเมทริกซ์ไม่เอกฐาน และให้ $\mathbf{x}_{m+1} = \mathbf{x}_m + C(\mathbf{b} - A\mathbf{x}_m)$ สมมติว่า $\mathbf{x}_m \to \mathbf{x}^*$

1. จงแสดงว่า $\mathbf{b} = A\mathbf{x}^*$ และสรุปว่า $\mathbf{x}^* = A^{-1}\mathbf{b}$

2. ถ้าข้อมูลผิดไปเป็น $\tilde{\mathbf{b}} = \mathbf{b} + \mathbf{e}$ และเราทำการปรับปรุงซ้ำด้วย $\tilde{\mathbf{b}}$ จงบอกลิมิต และคำนวณค่าคลาดเคลื่อนสัมบูรณ์ของลิมิตเทียบกับ $A^{-1}\mathbf{b}$

3. ใช้ข้อ 2 อธิบายว่าเหตุใดการปรับปรุงซ้ำในตาราง 7.6 จึงไม่ช่วยให้คำตอบกลับไปตรงกับแถวคมเดิมเมื่อข้อมูลมีสัญญาณรบกวน

แบบฝึกหัด 7.2.6 (⋆⋆⋆) ให้ $A_m = \begin{bmatrix} 10^{-m} & 1 \\ 1 & 1 \end{bmatrix}$ และ $\mathbf{b} = (1,2)^T$ สำหรับ $m = 1, 2, \ldots$

1. จงแสดงว่า $\kappa_1(A_m) = \dfrac{4}{1-10^{-m}}$ โดยใช้ว่า $A_m^{-1} = \dfrac{1}{10^{-m}-1}\begin{bmatrix} 1 & -1 \\ -1 & 10^{-m} \end{bmatrix}$ แล้วสรุปว่าปัญหาของระบบนี้ไวต่อข้อมูลหรือไม่

2. ใช้เลขทศนิยมสี่หลักนัยสำคัญแก้โดยไม่สลับแถว จงหา m ที่น้อยที่สุดซึ่งค่าคลาดเคลื่อนสัมพัทธ์ของ x_1 มากกว่า 50%

3. ทำซ้ำข้อ 2 ด้วยเลขจุดลอยตัว 64 บิตสำหรับ $m = 1, \ldots, 16$ จงหา m แรกที่ค่าคลาดเคลื่อนสัมพัทธ์ของ x_1 เกิน 50%

4. จากข้อ 1 ถึง 3 จงอธิบายว่าความผิดนี้มาจากปัญหาหรือจากวิธี และเทียบกับเรื่องปัญหาไม่ดีหรือวิธีไม่ดีในหัวข้อ 2.3

แบบฝึกหัด 7.2.7 (⋆⋆⋆) โครงงานเล็ก: เบลอกับขนาดของแถว ให้ $n \in \{16, 32, 64\}$ ให้แถวคมเป็นแถบสว่าง $x_i = 1$ เมื่อ $n/4 \leq i < 11n/16$ (นับจาก $i = 0$) และอื่น ๆ เป็นศูนย์ ให้สัญญาณรบกวน $\mathbf{e}$ สุ่มในช่วง $[-1/255, 1/255]$ โดยตั้งเมล็ดสุ่มเป็น 2 เมื่อ $n = 32$ แถวนี้ตรงกับแถวในหัวข้อนี้

1. จงคำนวณ $\kappa_2(B)$ จากสูตรสำหรับแต่ละ n แล้วเทียบกับ $\big(2(n+1)/\pi\big)^2$

2. ใช้โค้ดของบทนี้หาค่าคลาดเคลื่อนสัมพัทธ์ของการคืนแถวที่เบลอหนึ่งรอบ เมื่อข้อมูลมีสัญญาณรบกวน $\mathbf{e}$ จงบอกว่า n ใดที่ผิดไม่เกิน 10%

3. จงหาจำนวนรอบ k ที่น้อยที่สุดที่การปัดเศษอย่างเดียวทำให้ $\kappa_2(B^k)\,2^{-53} \geq 1$ สำหรับแต่ละ n แล้วเทียบกับผลในข้อ 2

4. (ลองดู) ถ้าอยากให้การเบลอหนึ่งรอบคืนได้ดีเมื่อ $n = 64$ จงบอกทางเลือกสองทางที่เกี่ยวกับข้อมูลหรือวิธีคำนวณ โดยไม่ต้องทำให้ลึก

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

ฉบับร่าง Preprint · 2026-10-23

เฉลย 7.2.1.

1. $B_3(1,2,1)^T$: แถวแรกได้ $\frac{1}{4}(2\cdot1+1\cdot2)=1$ แถวกลางได้ $\frac{1}{4}(1+2\cdot2+1)=\frac{3}{2}$ แถวสุดท้ายได้ $\frac{1}{4}(2+2\cdot1)=1$ จึงได้ $(1,\frac{3}{2},1)$ ค่าสูงสุดลดจาก 2 เป็น $\frac{3}{2}$ เพราะยอดถูกเฉลี่ยกับเพื่อนบ้านที่ต่ำกว่า

2. $\lambda_1=\frac{1}{2}(1+\cos\frac{\pi}{4})\approx 0.8536$, $\lambda_2=\frac{1}{2}(1+0)=0.5000$ และ $\lambda_3=\frac{1}{2}(1+\cos\frac{3\pi}{4})\approx 0.1464$

3. เมทริกซ์สมมาตรบวกแน่นอนจึงได้ $\kappa_2(B_3)=\dfrac{\lambda_{\max}}{\lambda_{\min}}=\dfrac{\frac{1}{2}(1+\frac{\sqrt{2}}{2})}{\frac{1}{2}(1-\frac{\sqrt{2}}{2})}=\dfrac{2+\sqrt{2}}{2-\sqrt{2}}=\dfrac{(2+\sqrt{2})^2}{2}=3+2\sqrt{2}\approx 5.83$

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_blur.py` ว่า $\kappa_2(B_3)=3+2\sqrt{2}$ และค่าลักษณะเฉพาะตรงกับสูตร

เฉลย 7.2.2.

1. โดยไม่สลับแถว: $\ell_{21}=1/0.0005=2000=2.000\times10^3$ แถวที่สองได้ $u_{22}=3-2000\cdot2=3-4000=-3997$ และ $y_2=5-2000\cdot4=-7995$ (ทั้งสองค่าตรงกับสี่หลักพอดี) จึงได้ $x_2=\dfrac{-7995}{-3997}\approx 2.0002\to 2.000$ แล้ว $x_1=\dfrac{4-2\cdot2.000}{0.0005}=\dfrac{0}{0.0005}=0$ ได้ $\mathbf{x}=(0,\,2.000)$

2. สลับแถวก่อน ได้ระบบ $\begin{bmatrix}1 & 3\\ 0.0005 & 2\end{bmatrix}$ กับ $(5,4)$ แล้ว $\ell_{21}=0.0005$ $u_{22}=2-0.0005\cdot3=1.9985\to1.998$ (ปัดคู่) $y_2=4-0.0005\cdot5=3.9975\to3.998$ จึงได้ $x_2=\dfrac{3.998}{1.998}=2.001$ และ $x_1=5-3\cdot2.001=-1.003$ ได้ $\mathbf{x}=(-1.003,\,2.001)$

3. ข้อ 1: $|0-(-1.00075)|/1.00075=100\%$ ข้อ 2: $|-1.003+1.00075|/1.00075\approx0.22\%$ วิธีที่สลับแถวดีกว่ามาก เพราะหลีกเลี่ยงการหารด้วยตัวหลักเล็ก 0.0005 ซึ่งทำให้ตัวคูณ 2000 ไปกลบตัวเลขที่มีนัยสำคัญของแถวแรก

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_blur.py` โดยตรวจทุกขั้นของทั้งสองทาง

เฉลย 7.2.3.

1. $\|I\|=\max_{\mathbf{x}\neq\mathbf{0}}\|I\mathbf{x}\|/\|\mathbf{x}\|=\max_{\mathbf{x}\neq\mathbf{0}}\|\mathbf{x}\|/\|\mathbf{x}\|=1$

2. จาก $I=AA^{-1}$ และอสมการของนอร์มเมทริกซ์ได้ $1=\|I\|=\|AA^{-1}\|\le\|A\|\,\|A^{-1}\|=\kappa(A)$ ดังนั้น $\kappa(A)\ge1$ สำหรับทุกนอร์มเหนี่ยวนำ

3. $\|A\|_1=\max(2+1,\,1+3)=4$ และ $A^{-1}=\frac{1}{5}\begin{bmatrix}3 & -1\\ -1 & 2\end{bmatrix}$ จึงได้ $\|A^{-1}\|_1=\max(\frac{3+1}{5},\frac{1+2}{5})=\frac{4}{5}$ ดังนั้น $\kappa_1(A)=4\cdot\frac{4}{5}=3.2$

ตรวจด้วยโค้ด: ทดสอบ $\kappa_1\ge1$ บนเมทริกซ์สุ่มด้วยเมล็ดสุ่มคงที่ และค่า 3.2 ในข้อ 3 โดยเทียบกับ `numpy`

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 7.2.4.

1. เนื่องจาก B สมมาตร $B = Q\Lambda Q^T$ โดย Q เป็นเมทริกซ์ตั้งฉาก ดังนั้น $B^k = Q\Lambda^k Q^T$ มีค่าลักษณะเฉพาะ λ_j^k และเวกเตอร์ลักษณะเฉพาะเดิม เมื่อ B^k สมมาตรบวกแน่นอน $\|B^k\|_2 = \max_j \lambda_j^k$ และ $\|(B^k)^{-1}\|_2 = 1/\min_j \lambda_j^k$ จึงได้ $\kappa_2(B^k) = (\max_j \lambda_j / \min_j \lambda_j)^k = \kappa_2(B)^k$

2. จากสูตร $\kappa_2(B) \approx 116.5$ ($n=16$) ≈ 440.7 ($n=32$) และ ≈ 1712 ($n=64$) และ $2^{53} \approx 9.007 \times 10^{15}$ จึงต้องการ $k \geq \dfrac{\log(9.007 \times 10^{15})}{\log \kappa_2(B)}$ ซึ่งได้ $k \geq 7.7$ สำหรับ $n=16$ ได้ $k \geq 6.04$ สำหรับ $n=32$ และได้ $k \geq 4.93$ สำหรับ $n=64$ ดังนั้นค่า k ที่น้อยที่สุดคือ 8, 7 และ 5 ตามลำดับ

3. ความกว้างการเบลอ $\sqrt{k/2}$ คือ 2.00, 1.87 และ 1.58 พิกเซลตามลำดับ การเบลอกว้างขึ้นหมายถึงค่าลักษณะเฉพาะที่เล็กที่สุดของ B^k เล็กลงในอัตราเลขชี้กำลัง จึงทำให้ค่าปัดเศษเดียวกันถูกขยายมากขึ้น ส่วนการเพิ่มจำนวนจุดภาพก็ทำให้ $\kappa_2(B)$ โตตามกำลังสองของ n เช่นกัน

ตรวจด้วยโค้ด: ทดสอบความเท่ากัน $\kappa_2(B^k) = \kappa_2(B)^k$ ที่ $n = 16, 32, 64$ และค่า k ที่ได้จากการคำนวณตรงกับค่าข้างบน

เฉลย 7.2.5.

1. สมมติ $\mathbf{x}_m \to \mathbf{x}^*$ ทั้งสองข้างของ $\mathbf{x}_{m+1} = \mathbf{x}_m + C(\mathbf{b} - A\mathbf{x}_m)$ ได้ $\mathbf{x}^* = \mathbf{x}^* + C(\mathbf{b} - A\mathbf{x}^*)$ จึง $C(\mathbf{b} - A\mathbf{x}^*) = \mathbf{0}$ เมื่อ C ไม่เอกฐาน จะได้ $\mathbf{b} = A\mathbf{x}^*$ และ $\mathbf{x}^* = A^{-1}\mathbf{b}$

2. ข้อ 1 ใช้ได้กับข้อมูล $\tilde{\mathbf{b}}$ ด้วย ดังนั้นลิมิตคือ $A^{-1}\tilde{\mathbf{b}} = A^{-1}\mathbf{b} + A^{-1}\mathbf{e}$ ค่าคลาดเคลื่อนสัมบูรณ์ของลิมิตเทียบกับ $A^{-1}\mathbf{b}$ คือ $A^{-1}\mathbf{e}$ ซึ่งเป็นค่าคลาดเคลื่อนสัมบูรณ์ที่การแก้โดยตรงได้จากข้อมูลผิดนั้นเอง

3. ตาราง 7.6 ให้ค่าที่ลู่เข้าไปที่ $A^{-1}\tilde{\mathbf{b}}$ (ผิดจาก $A^{-1}\tilde{\mathbf{b}}$ เพียงประมาณ 10^{-15}) ซึ่งยังห่างจากแถวคมเดิม 9.5% เพราะการปรับปรุงซ้ำเปลี่ยนคำตอบให้ตรงกับระบบที่ผิดไปแล้ว ไม่ได้ย้อนกลับไปหาข้อมูลเดิมที่ไม่มีสัญญาณรบกวน

เฉลย 7.2.6.

1. $\det A_m = 10^{-m} - 1$ และ $A_m^{-1} = \frac{1}{10^{-m}-1}\begin{bmatrix} 1 & -1 \\ -1 & 10^{-m} \end{bmatrix}$ นอร์มหนึ่งของ A_m คือ $\max(1+10^{-m}, 2) = 2$ และผลรวมค่าสัมบูรณ์ของคอลัมน์ของ A_m^{-1} คือ $\frac{2}{1-10^{-m}}$ กับ $\frac{1+10^{-m}}{1-10^{-m}}$ ค่าแรกมากกว่า จึง $\|A_m^{-1}\|_1 = \frac{2}{1-10^{-m}}$ และ $\kappa_1(A_m) = \frac{4}{1-10^{-m}}$ ซึ่งอยู่ราว 4 ทุก m ปัญหาของระบบนี้จึงไม่ไวต่อข้อมูลเลย

2. กับ $m = 1, 2, 3, 4$ คำตอบ x_1 ผิดไม่เกิน 0.1% สัมพัทธ์ แต่กับ $m = 5$ ทำดังนี้: $\ell_{21} = 10^5$, $u_{22} = 1 - 10^5 \approx -1.000 \times 10^5$, $y_2 = 2 - 10^5 \approx -1.000 \times 10^5$ จึงได้ $x_2 = 1.000$ และ $x_1 = \dfrac{1 - 1.000}{10^{-5}} = 0$ ค่าคลาดเคลื่อนสัมพัทธ์ 100% ดังนั้น $m = 5$

3. ในเลขจุดลอยตัว 64 บิต ค่าคลาดเคลื่อนสัมพัทธ์ของ x_1 สำหรับ $m = 1, \ldots, 15$ ทั้งหมดต่ำกว่า 10^{-3} แต่ที่ $m = 16$ ได้ $x_1 \approx 2.22$ ค่าคลาดเคลื่อนสัมพัทธ์ 122% เพราะที่ 10^{16} ช่องว่างระหว่างเลขที่เก็บได้เท่ากับ 2 ทำให้ $1 - 10^{16}$ เก็บเป็นตัวเลขได้ไม่ตรง ตัวหาร $1 - x_2$ จึงเหลือแต่ผลของการปัดเศษ ดังนั้น $m = 16$

4. จากข้อ 1 ปัญหาดีมาก ($\kappa_1 \approx 4$) ความผิดที่เกิดจึงมาจากวิธี คือการหารด้วยตัวหลักเล็ก 10^{-m} ซึ่งตรงกับเรื่อง ``ปัญหาไม่ดีหรือวิธีไม่ดี'' ในหัวข้อ 2.3 วิธีแก้คือสลับแถวให้ตัวหลักเป็นแถวที่ใหญ่กว่า โดยที่ κ ของระบบไม่เปลี่ยน

ตรวจด้วยโค้ด: ทดสอบค่า $\kappa_1(A_m)$ เทียบกับสูตร และค่า $m = 5$ และ $m = 16$ ทั้งเลขทศนิยมสี่หลักและเลขจุดลอยตัว 64 บิต

สงวนลิขสิทธิ์ · Preprint · 2026-10-23

เฉลย 7.2.7.

1. จากสูตร $\kappa_2(B) = \cot^2 \frac{\pi}{2(n+1)}$ ได้ 116.5 ($n = 16$), 440.7 ($n = 32$) และ 1712 ($n = 64$) และ $(2(n+1)/\pi)^2 \approx 117.1, 441.4$ และ 1712 ตามลำดับ ค่าประมาณนี้ดีภายในไม่เกินหนึ่งเปอร์เซ็นต์

2. ใช้แถวคมที่ระบุและสัญญาณรบกวนเมล็ด 2 ค่าคลาดเคลื่อนสัมพัทธ์ของการคืนหนึ่งรอบคือ 3.85% ($n = 16$), 9.54% ($n = 32$) และ 21.7% ($n = 64$) ดังนั้น $n = 16$ และ $n = 32$ ผิดไม่เกิน 10% ส่วน $n = 64$ ไม่ผ่าน ผลนี้เปลี่ยนตามเมล็ดสุ่ม จึงต้องบอกเมล็ดด้วยทุกครั้ง

3. ค่า k ที่การปัดเศษอย่างเดียวทำให้ $\kappa_2(B^k)2^{-53} \geq 1$ คือ 8, 7 และ 5 ตามข้อ 2 ของแบบฝึกหัด 7.2.4 กรณี $n = 64$ สัญญาณรบกวนทำให้ผิดตั้งแต่ $k = 1$ ทั้งที่การปัดเศษจะทำลายคำตอบเมื่อ k ถึง 5 เท่านั้น ในตัวอย่างนี้สัญญาณรบกวนจึงเป็นตัวกำหนดก่อนการปัดเศษ

4. (ทางเลือก) หนึ่ง ลดหรือกรองสัญญาณรบกวนในข้อมูลก่อนแก้ระบบ (เปลี่ยนข้อมูล) สอง ใช้วิธีที่เปลี่ยนปัญหาให้คำตอบเรียบขึ้น ซึ่งเป็นหัวข้อนอกบทนี้ (เปลี่ยนวิธี) ทั้งสองทางแลกกับการบอกว่าคำตอบที่ได้เป็นคำตอบของปัญหาอื่น

ตรวจด้วยโค้ด: ทดสอบค่าทั้งหมดของข้อ 1 ถึง 3 ด้วยแถวคมที่ระบุและเมล็ดสุ่ม 2

7.3 เมื่อเมทริกซ์ส่วนใหญ่เป็นศูนย์

สถานการณ์สมมติ: ผ้าที่มีหลายหมื่นจุด

ลองนึกถึงผ้าที่ตัวละครสวมอยู่ เราจำลองผ้าเป็นตารางของจุดเล็ก ๆ ที่ผูกกับเพื่อนบ้านทั้งสี่ทิศ เมื่อผ้าขยับ จุดแต่ละจุดต้องหาตำแหน่งใหม่ที่เข้ากับเพื่อนบ้าน ซึ่งทำให้เกิดระบบสมการเชิงเส้นอีกระบบในทุกก้าวเวลา
ถ้าผ้ามี 100×100 จุด ก็มีตัวไม่ทราบ $10{,}000$ ตัว แต่สมการของจุดหนึ่งเกี่ยวข้องกับตัวเองและเพื่อนบ้านไม่เกินห้าตัว ส่วนใหญ่ของเมทริกซ์จึงเป็นศูนย์ ทั้งที่มันมีขนาด $10{,}000 \times 10{,}000$ เอกสารของ Baraff และ Witkin ระบุว่าการจำลองผ้าด้วยก้าวเวลาใหญ่สร้างระบบเชิงเส้นขนาดใหญ่ที่ไม่เป็นแถบและมีศูนย์อยู่มาก ในทุกก้าวเวลา Baraff and Witkin 1998, บทคัดย่อ
คำถามของบทนี้จึงเป็นว่า ถ้าเมทริกซ์ใหญ่มากและเกือบทั้งหมดเป็นศูนย์ เราจะเก็บมันอย่างไรโดยไม่เสียหน่วยความจำไปกับศูนย์ และแก้มันได้อย่างไรโดยไม่สร้างค่าไม่เป็นศูนย์ใหม่มากเกินไป
(ผ้าและแบบจำลองในหัวข้อนี้เป็นสถานการณ์สมมติที่เราสร้างเอง ตัวเลขทั้งหมดมาจากโค้ดของเรา และไม่ได้บอกว่าเกมหรือเอนจินใดสร้างระบบนี้ในรายละเอียด)

ภาพรวมก่อนอ่าน

มีสามความคิดที่ตามมาในหัวข้อนี้

- เก็บเฉพาะตำแหน่งที่ไม่เป็นศูนย์ และคำนวณเฉพาะตำแหน่งเหล่านั้น
- การกำจัดแบบเกาส์สร้างค่าไม่เป็นศูนย์ที่เดิมเป็นศูนย์ ซึ่งเรียกว่า **ฟิลอิน** (fill-in) เป็นศัตรูหลักของวิธีนี้
- ลำดับที่เรานับจุดมีผลมากต่อจำนวนฟิลอิน และเราเห็นโครงสร้างนี้ได้จาก **กราฟของเมทริกซ์** (graph of a matrix)

ร่างพิมพ์ · Preprint · 2026-10-23

ตารางที่ 7.8: จำนวนไบต์ของ A ของผ้า $n \times n$ จุด เก็บเต็มและเก็บแบบ CSR (ตัวเลขจากโค้ดของเรา)

n	N	ค่าไม่เป็นศูนย์	เก็บเต็ม (ไบต์)	เก็บแบบ CSR (ไบต์)	อัตราส่วน
10	100	460	80 000	5 924	13.5
30	900	4 380	6 480 000	56 164	115.4
100	10 000	49 600	800 000 000	635 204	1 259.4

ผ้าเป็นตารางของจุด และเมทริกซ์ของมัน

ให้ผ้ามี $n \times n$ จุดเรียงเป็นตาราง เรานับจุดที่แถว i คอลัมน์ j เป็นหมายเลข $(i-1)n + j$ โดยเรียงตามแถว (นับจาก 1) ให้ x_p เป็นตำแหน่งของจุด p เราสมมติว่าขอบของผ้ายึดกับกรอบที่ไม่ขยับ สมการของจุดหนึ่งจึงมีรูป

$$4x_p - \sum_{q \in N(p)} x_q = b_p,$$

เมื่อ $N(p)$ คือเพื่อนบ้านของจุด p ที่อยู่ในตาราง (ไม่เกินสี่จุด) และ b_p คือแรงที่จุด p ได้รับ ดังนั้นสมการแต่ละอันเกี่ยวกับตัวเองและเพื่อนบ้านไม่เกินห้าตัว เมทริกซ์ A ขนาด $N \times N$ โดย $N = n^2$ จึงมี 4 บนแนวทแยง และ -1 ที่ตำแหน่งของทุกคู่เพื่อนบ้าน ทำให้แต่ละแถวมีค่าไม่เป็นศูนย์ไม่เกินห้าช่อง

เมื่อ $n = 100$ จำนวนค่าไม่เป็นศูนย์ของ A คือ $5n^2 - 4n = 49{,}600$ จากทั้งหมด 10^8 ช่อง คิดเป็นราว 0.05% และเฉลี่ยแถวละ 4.96 ค่า (ตัวเลขนี้มาจากโค้ดของเรา)

นิยาม 7.3.1: กราฟของเมทริกซ์

ให้ A เป็นเมทริกซ์สมมาตรขนาด $N \times N$ **กราฟ** (graph) ของ A มี**จุดยอด** (vertex) $1, 2, \ldots, N$ และมี**เส้นเชื่อม** (edge) ระหว่างจุดยอด $i \neq j$ ก็ต่อเมื่อ $a_{ij} \neq 0$ Iserles 2009, §11.2, p. 238

อ่านแบบเข้าใจง่าย. ให้ตัวไม่ทราบแต่ละตัวเป็นจุดบนแผนที่ ถ้าสองตัวผูกกันในสมการ (คือ $a_{ij} \neq 0$) ก็ลากเส้นระหว่างจุดทั้งสอง กราฟของผ้าจึงเป็นตารางของจุดเดิมนั่นเอง เส้นชัดกว่าตัวเลขศูนย์ในเมทริกซ์มาก

ลองเอง

นับเส้นเชื่อมของผ้าขนาด 3×3 จุด ผลที่ควรได้คือ 12 เส้น จากนั้นนับช่องศูนย์ของเมทริกซ์ 9×9 ที่สอดคล้องกัน ผลที่ควรได้คือ 48 ช่อง ตรวจว่าแต่ละเส้นเชื่อมทำให้มีค่าไม่เป็นศูนย์สองช่องใน A ที่สมมาตรกัน

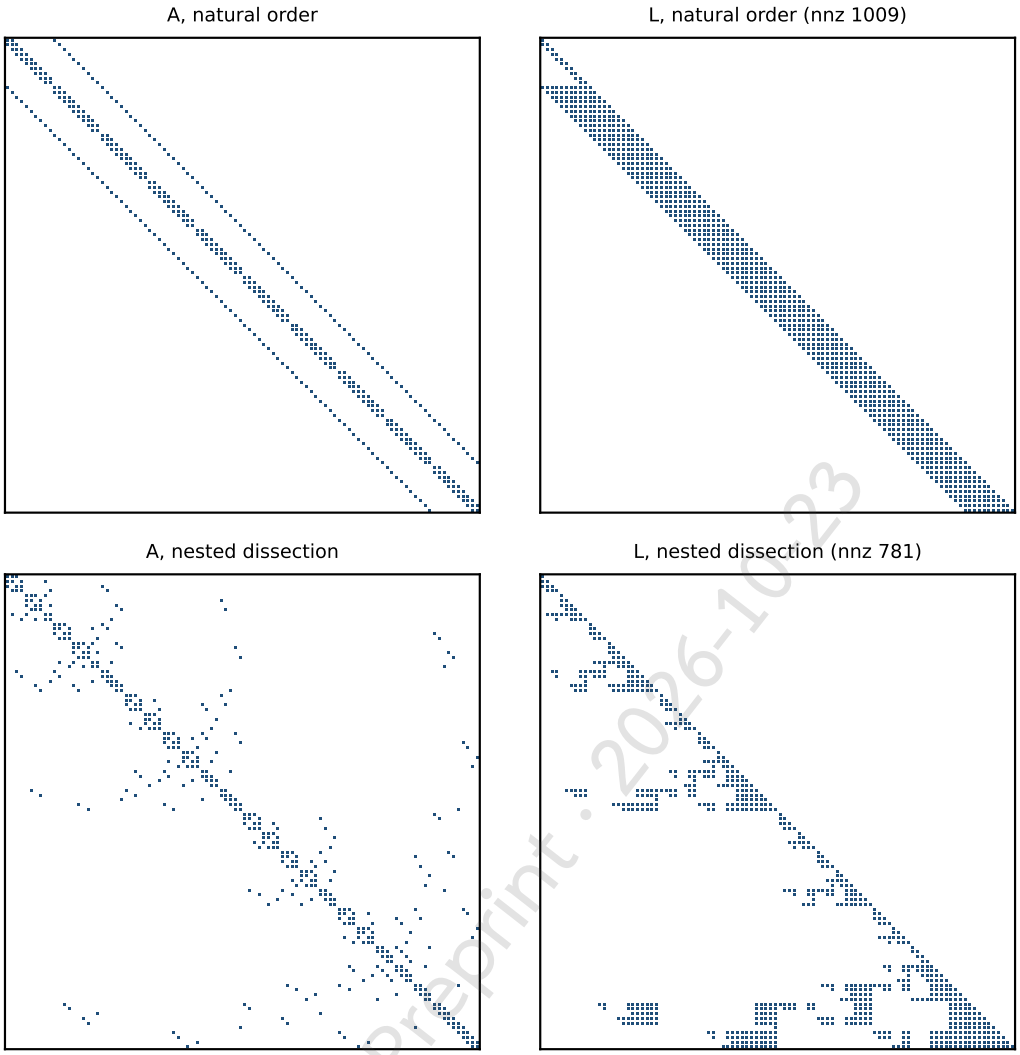
รูป 7.8 แสดงรูปแบบของ A และ L ของผ้าสิบคูณสิบจุด

เก็บเฉพาะที่ไม่เป็นศูนย์

ถ้าเก็บ A เป็นตารางเต็ม ต้องใช้ N^2 ช่อง สำหรับผ้า 100×100 จุดคือ 10^8 ช่อง ซึ่งเป็นราว 800 MB ถ้าแต่ละค่าเป็นเลขจุดลอยตัว 8 ไบต์ วิธีที่ใช้กันบ่อยคือ**การเก็บแบบแถวบีบอัด** (compressed sparse row, CSR) ซึ่งเก็บสามอาร์เรย์ คือค่าที่ไม่เป็นศูนย์เรียงตามแถว อาร์เรย์ของดัชนีคอลัมน์ของแต่ละค่า และอาร์เรย์ที่บอกว่าแต่ละแถวเริ่มที่ตำแหน่งใดในสองอาร์เรย์แรก Ern and Guermond 2004, §8.3, p. 373 การคูณ $A\mathbf{x}$ จึงวนเฉพาะค่าที่เก็บไว้ ใช้การคูณเท่ากับจำนวนค่าที่ไม่เป็นศูนย์ Ern and Guermond 2004, §8.3, p. 375

ตาราง 7.8 เทียบจำนวนไบต์ของการเก็บสองแบบ ค่าใช้ 8 ไบต์ และดัชนีใช้ 4 ไบต์

ฉบับร่าง Preprint · 2026-10-23



ฉบับร่าง · Preprint · 2026-10-23

รูปที่ 7.8: รูปแบบศูนย์และไม่ศูนย์ของผ้าสิบคูณสิบจุด ($N = 100$) ซ้าย: เมทริกซ์ A ขวา: ตัวประกอบ L บนแถวบนเรียงตามแถว แถวล่างเรียงด้วยการตัดแบ่งซ้อนที่จะกล่าวถึงในหัวข้อนี้ ตำแหน่งที่ L มีแต่ A ไม่มีคือฟิลอิน (สร้างจาก `numanal/ch07_sparse.py`)

แถบ: ศูนย์นอกแถบรอด

เราจะเห็นว่าค่าไม่เป็นศูนย์ของผ้าที่เรียงตามแถวอยู่ใกล้เส้นทแยงมุมเป็นพิเศษ วัดความใกล้นี้ด้วยความกว้างแถบ

นิยาม 7.3.2: เมทริกซ์แถบและความกว้างแถบ

ให้ A เป็นเมทริกซ์จัตุรัสขนาด $N \times N$ ถ้ามีจำนวนเต็ม $s \geq 0$ ที่ $a_{ij} = 0$ ทุกคู่ (i, j) ที่ $|i - j| > s$ เราเรียก A ว่า **เมทริกซ์แถบ** (banded matrix) และเรียก s ที่น้อยที่สุดว่า **ความกว้างแถบ** (bandwidth) ของ A Iserles 2009, §11.1, p. 233

อ่านแบบเข้าใจง่าย. ความกว้างแถบคือระยะที่ไกลที่สุดจากเส้นทแยงมุมที่ยังมีค่าไม่เป็นศูนย์ ถ้าความกว้างแถบเล็ก ตัวไม่ทราบแต่ละตัวเกี่ยวข้องกับตัวไม่ทราบที่อยู่ใกล้เท่านั้น ตัวอย่างของเราคือโซ่ของหัวข้อ 7.1 ซึ่งมีความกว้างแถบ 1 และ B^k ของการเบลอในหัวข้อ 7.2 ซึ่งมีความกว้างแถบ k เพราะแต่ละรอบของการเบลอขยายแถบออกไปหนึ่งแถว (ตรวจด้วยโค้ดของเรา)

สำหรับผ้าที่เรียงตามแถว จุดที่อยู่เหนือและใต้จุดหนึ่งห่างกัน n หมายเลข ความกว้างแถบจึงเป็น n ไม่ว่าตารางจะมีขนาด

เท่าใด (ตัวอย่าง $n = 10$ ได้ 10 และ $n = 100$ ได้ 100) ตำรา Iserles 2009, §11.1, pp. 235--236 นับว่าการแยก LU ของเมทริกซ์แถบความกว้าง s ทำได้ด้วยราว $O(s^2N)$ การดำเนินการ แทนที่จะเป็น $O(N^3)$ ของเมทริกซ์เต็ม

ข้อควรระวัง: ศูนย์นอกแถบรอด แต่ศูนย์ในแถบไม่รอด

การแยกแบบแถบรับประกันเพียงว่าศูนย์นอกแถบจะยังเป็นศูนย์ ส่วนศูนย์ในแถบอาจกลายเป็นไม่ศูนย์ ซึ่งเป็นฟิลอินที่เราจะนับในหัวข้อย่อยถัดไป Iserles 2009, §11.1, p. 237

ฟิลอิน: ศูนย์ที่กลายเป็นไม่ศูนย์

นิยาม 7.3.3: ฟิลอิน

ให้ A เป็นเมทริกซ์สมมาตรที่เรียงลำดับแล้ว และ $A = LL^T$ เป็นการแยก **ฟิลอิน** (fill-in) คือจำนวนตำแหน่ง (i, j) ที่ $i > j$ ซึ่ง $a_{ij} = 0$ แต่ $\ell_{ij} \neq 0$ Iserles 2009, §11.2, p. 239

อ่านแบบเข้าใจง่าย. ฟิลอินคือศูนย์ที่เดิมอยู่ในเมทริกซ์ แต่หลังการกำจัดกลายเป็นตัวเลขที่ไม่เป็นศูนย์ ตัวประกอบต้องเก็บมันและต้องคำนวณเพื่อมันด้วย จำนวนฟิลอินจึงเป็นต้นทุนที่แท้จริง

สำหรับผ้า 10×10 จุดเรียงตามแถว L มี $1{,}009$ ค่าไม่เป็นศูนย์ ขณะที่ส่วนล่างของ A รวมแนวทแยงมี 280 ค่า นั่นคือ L มากกว่าราว 3.6 เท่า สำหรับการเรียงตามแถว เราตรวจด้วยการนับแบบสัญลักษณ์ว่าฟิลอินอยู่ในซองหุ้มของ A คือช่วงตั้งแต่ค่าไม่เป็นศูนย์แรกของแต่ละแถวถึงแนวทแยง และทุกตำแหน่งในซองนั้นกลายเป็นไม่ศูนย์ ทำให้จำนวนค่าใน L เท่ากับ

$$1 + 2(n-1) + (N-n)(n+1).$$

สูตรนี้เราตรวจด้วยการนับแบบสัญลักษณ์สำหรับ $n \leq 30$ ทุกกรณีตรงกัน และใช้ต่อสำหรับ $n = 100$ (การพิสูจน์ทั่วไปว่าฟิลอินไม่หลุดจากซองหุ้มเราไม่ได้ทำในที่นี้)

สมมาตรและบวกแน่นอน: การแยกโชเลสกี

เมทริกซ์ของผ้าเป็นสมมาตร และเป็น**บวกแน่นอน** (positive definite) คือ $\mathbf{x}^T A\mathbf{x} > 0$ สำหรับทุก $\mathbf{x} \neq \mathbf{0}$ เมื่อ A สมมาตร ตัวประกอบ U ของ LU มีรูปเดียวกับ L^T คูณด้วยแนวทแยงบางอย่าง เราจึงไม่ต้องเก็บหรือคำนวณอีกครึ่งหนึ่ง ผลคือการแยกแบบ**โชเลสกี** (Cholesky factorization)

นิยาม 7.3.4: การแยกตัวประกอบโชเลสกี

ถ้าเมทริกซ์สมมาตรบวกแน่นอน A เขียนได้ในรูป $A = LL^T$ โดย L เป็นสามเหลี่ยมล่าง และสมาชิกบนแนวทแยงของ L เป็นบวกทั้งหมด เราเรียกการเขียนนี้ว่า การแยกตัวประกอบโชเลสกี

อ่านแบบเข้าใจง่าย. ต่างจาก LU ที่มีสองตัว โชเลสกีมีตัวเดียวคือ L เพราะ U ถูกแทนด้วย L^T จึงเก็บครึ่งเดียวและคำนวณประมาณครึ่งหนึ่งของ LU Iserles 2009, ภาคผนวก A.1.4.6, p. 436

ลองเอง

สำหรับโซ่สามลูก $A_3 = \mathrm{tridiag}(-1, 2, -1)$ จงหา L_{11} และ L_{21} จาก $LL^T = A_3$ ผลที่ควรได้คือ $L_{11} = \sqrt{2}$ และ $L_{21} = -1/\sqrt{2}$ ต่อไปตรวจว่า $L_{22} = \sqrt{3/2}$, $L_{32} = -\sqrt{2/3}$ และ $L_{33} = \sqrt{4/3}$

ร่างปรับปรุง Preprint 2026-10-23

ทฤษฎีบท 7.3.1: การมีอยู่และความเป็นหนึ่งเดียวของการแยกโชเลสกี

ถ้า A เป็นเมทริกซ์สมมาตรบวกแน่นอนขนาด $N \times N$ แล้วมีการแยก $A = LL^T$ ที่ L เป็นสามเหลี่ยมล่างซึ่งสมาชิกบนแนวทแยงเป็นบวก และการแยกนี้มีเพียงแบบเดียว

พิสูจน์. **แนวคิด.** ใช้ทฤษฎีบท LU ของหัวข้อ 7.1 ตัวหลักเป็นบวกเพราะเมทริกซ์ย่อยของเมทริกซ์บวกแน่นอนก็บวกแน่นอน แล้วใช้ความสมมาตรบังคับให้ตัวประกอบสองตัวเป็นคู่ของกันและกัน

1. **เมทริกซ์ย่อยบวกแน่นอน.** ให้ A_k เป็นเมทริกซ์ย่อยมุมบนซ้ายขนาด $k \times k$ สำหรับ $\mathbf{y} \in \mathbb{R}^k$ ที่ไม่เป็นศูนย์ ให้ $\mathbf{x} = (\mathbf{y}, \mathbf{0})$ จะได้ $\mathbf{y}^T A_k \mathbf{y} = \mathbf{x}^T A \mathbf{x} > 0$ เมทริกซ์บวกแน่นอนมีค่าลักษณะเฉพาะเป็นบวกทั้งหมด จึง $\det A_k > 0$ สำหรับทุก k

2. **ใช้ทฤษฎีบท LU.** ทฤษฎีบทในหัวข้อ 7.1 ให้ $A = \tilde{L}U$ เมื่อ $\tilde{L}$ มีหนึ่งบนแนวทแยง และ $u_{kk} = \det A_k / \det A_{k-1} > 0$ ตั้ง $D = \operatorname{diag}(u_{11}, \ldots, u_{NN})$ และ $M^T = D^{-1}U$ จะได้ M เป็นสามเหลี่ยมบนที่มีหนึ่งบนแนวทแยง และ $A = \tilde{L}DM^T$

3. **ความสมมาตรบังคับให้ $M = \tilde{L}$.** จาก $A^T = A$ ได้ $MD\tilde{L}^T = \tilde{L}DM^T$ จึง $M^{-1}\tilde{L}D = D\tilde{L}^T(M^T)^{-1}$ ด้านซ้ายเป็นสามเหลี่ยมล่างและด้านขวาเป็นสามเหลี่ยมบน ทั้งสองข้างจึงเป็นเมทริกซ์แนวทแยง และแนวทแยงของมันคือ D ดังนั้น $M^{-1}\tilde{L}D = D$ นั่นคือ $M^{-1}\tilde{L} = I$ และ $\tilde{L} = M$

4. **สร้าง L.** ได้ $A = \tilde{L}D\tilde{L}^T$ ตั้ง $L = \tilde{L}D^{1/2}$ จะได้ $LL^T = A$ และแนวทแยงของ L เป็นบวก

5. **ความเป็นหนึ่งเดียว.** ถ้า $A = LL^T = L'L'^T$ โดยทั้งสองมีแนวทแยงเป็นบวก ตั้ง $Q = L'^{-1}L$ จะได้ $QQ^T = I$ จึง Q เป็นเมทริกซ์ตั้งฉากที่เป็นสามเหลี่ยมล่าง ดังนั้น $Q^{-1} = Q^T$ เป็นทั้งสามเหลี่ยมล่างและบน จึงเป็นแนวทแยง สมาชิกบนแนวทแยงของ Q คือ $\ell_{kk}/\ell'_{kk} > 0$ และมีกำลังสองเป็น 1 จึงเป็น 1 ทั้งหมด ดังนั้น $Q = I$ และ $L = L'$

สรุป: การกำจัดแบบเกาส์ให้ตัวหลักเป็นบวกสำหรับเมทริกซ์บวกแน่นอน และความสมมาตรทำให้ตัวประกอบสองตัวกลายเป็นคู่ที่เป็นกระจกเงาของกันและกัน □

สำหรับโซ่ของหัวข้อ 7.1 ตัวหลักของการกำจัดคือ $u_{kk} = (k+1)/k$ และตัวคูณคือ $-k/(k+1)$ จากการสร้าง $L = \tilde{L}D^{1/2}$ ในขั้นที่ 4 ของการพิสูจน์จึงได้ตัวประกอบโชเลสกีในรูปปิด

$$L_{kk} = \sqrt{\frac{k+1}{k}}, \qquad L_{k+1,k} = -\sqrt{\frac{k}{k+1}},$$

เมื่อ $A = \operatorname{tridiag}(-1, 2, -1)$ มีขนาด $N \times N$ (แบบฝึกหัด 7.3.3 ตรวจกรณี $N = 4$)

ต้นไม้ไม่มีฟิลอิน

โซ่ของเรามีกราฟที่เป็นเส้นตรง ซึ่งเป็น**ต้นไม้** (tree) ตัวอย่างนี้ให้เบาะแสว่าโครงสร้างของกราฟอาจกำจัดฟิลอินได้ทั้งหมด

นิยาม 7.3.5: ต้นไม้มีรากและการเรียงแบบขึ้นสู่ราก

ต้นไม้ (tree) คือกราฟที่เชื่อมกันทั้งหมดและไม่มีวงปิด เลือกจุดหนึ่งเป็นราก แล้วทุกจุดที่ไม่ใช่รากมี*แม่* หนึ่งจุด คือเพื่อนบ้านที่อยู่ใกล้รากกว่า ถ้าการเรียงหมายเลขทำให้ทุกจุดมีหมายเลขน้อยกว่าแม่ของมันเสมอ เรียกว่าการเรียงแบบ*ขึ้นสู่ราก*
Iserles 2009, §11.2, p. 238

ฉบับร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย. ลองนึกถึงต้นไม้ครอบครัว รากคือบรรพบุรุษ แต่ละจุดมีแม่คนเดียว ที่เหลือที่ติดกับมันเป็นลูกหลาน การเรียงแบบขึ้นสู่รากคือการตั้งหมายเลขให้ลูกมีเลขน้อยกว่าแม่เสมอ ราก (บรรพบุรุษ) จึงได้เลขมากที่สุด

ทฤษฎีบท 7.3.2: ต้นไม้ไม่มีฟิลอิน

ให้ A เป็นเมทริกซ์สมมาตรบวกแน่นอนที่กราฟเป็นต้นไม้ และการเรียงเป็นแบบขึ้นสู่ราก แล้วการแยกโชเลสกีไม่มีฟิลอินเลย นั่นคือ ถ้า $i > j$ และ $a_{ij} = 0$ แล้ว $\ell_{ij} = 0$ Iserles 2009, ทฤษฎีบท 11.1, p. 238

พิสูจน์. **แนวคิด.** ตัดจุดที่มีหมายเลขน้อยที่สุดออก จุดนั้นเป็นใบของต้นไม้ การกำจัดใบเปลี่ยนแนวทแยงของแม่เท่านั้น ไม่เพิ่มเส้นเชื่อมใหม่ จึงทำซ้ำได้โดยอุปนัย

1. **จุดที่หมายเลขน้อยที่สุดเป็นใบ.** ถ้าจุด 1 มีลูกสักคน ลูกนั้นต้องมีหมายเลขน้อยกว่า 1 ซึ่งเป็นไปไม่ได้ ดังนั้นจุด 1 มีเพื่อนบ้านเพียงคนเดียว คือแม่ γ ของมัน
2. **การกำจัดใบเปลี่ยนเฉพาะแนวทแยงของแม่.** คอลัมน์ที่ 1 ของ A มีค่านอกแนวทแยงเพียงตัวเดียวที่ตำแหน่ง γ ดังนั้นเมทริกซ์ที่เหลือหลังการกำจัด (คอมพลีเมนต์ชูร์) ต่างจาก A_{22} เฉพาะที่ตำแหน่ง (γ, γ) จึงไม่มีตำแหน่งใหม่ที่จะกลายเป็นไม่ศูนย์
3. **ทำซ้ำ.** คอมพลีเมนต์ชูร์ยังบวกแน่นอน และกราฟของมันคือต้นไม้เดิมตัดจุด 1 ออก ซึ่งยังเรียงแบบขึ้นสู่รากอยู่ ใช้สมมติฐานเดิมกับขนาด $N - 1$ ซ้ำไปเรื่อย ๆ จนครบ

สรุป: ทุกขั้นของการกำจัดคือการตัดใบหนึ่งใบ และการตัดใบไม่เพิ่มเส้นเชื่อม จึงไม่มีฟิลอิน □

อ่านแบบเข้าใจง่าย

ผลนี้บอกว่า ถ้าการผูกกันของตัวไม่ทราบเป็นต้นไม้และเรียงแบบขึ้นสู่ราก ตัวประกอบ L มีค่าไม่เป็นศูนย์เท่ากับส่วนล่างของ A เท่านั้น ไม่มีตัวใหม่เพิ่มขึ้น โซ่ของหัวข้อ 7.1 เป็นแบบนี้ ตัวคูณทั้งหมดอยู่ที่ใต้แนวทแยงหนึ่งแถวเท่านั้น

เราสุ่มต้นไม้สี่สิบจุดห้าต้น (เมล็ดสุ่ม 0 ถึง 4) ตรวจด้วยโค้ดแล้วไม่มีฟิลอินในทุกต้น ส่วนวงปิดทำให้มีฟิลอินได้ ดังนี้

ข้อควรระวัง: ต้นไม้ไม่มีฟิลอิน แต่วงปิดมี

ถ้าผูกจุดหกจุดเป็นวงแหวน โดยเรียงตามลำดับรอบวง การกำจัดจะสร้างฟิลอินสามตำแหน่ง (ตัวอย่างของเรา) วงปิดทำให้ต้องเติมตำแหน่งเพิ่ม ดังนั้นต้องตรวจโครงสร้างก่อนพึ่งพาผลของต้นไม้

จัดลำดับให้ศูนย์รอด

ถ้าสลับหมายเลขของตัวไม่ทราบพร้อมกันทั้งแถวและคอลัมน์ เมทริกซ์ใหม่ยังสมมาตรและบวกแน่นอนเหมือนเดิม การสลับแบบนี้เป็นการ**เรียงสับเปลี่ยน** (permutation) ที่ไม่เปลี่ยนคำตอบของระบบ เพียงแต่เปลี่ยนจำนวนฟิลอิน Iserles 2009, §11.2, p. 239

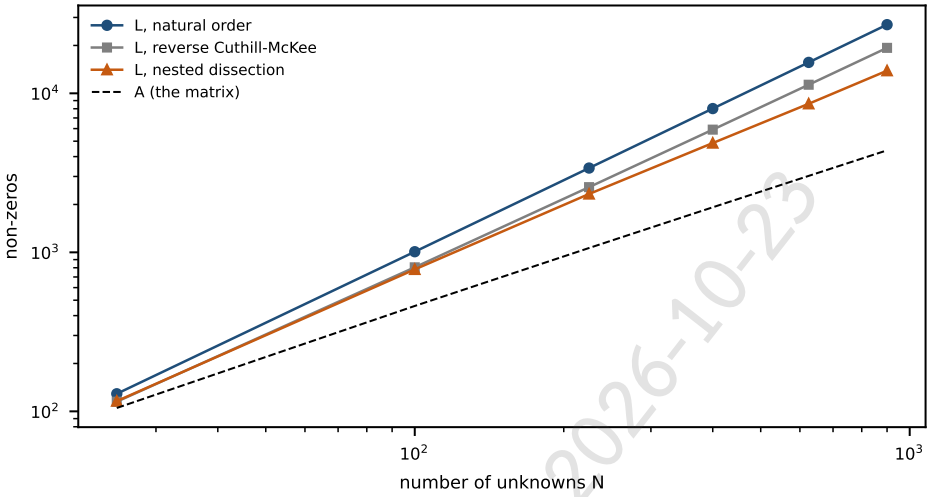
เราจะเปรียบเทียบลำดับสามแบบ ทุกแบบบนผ้าที่ $n \times n$ จุด

- **ตามแถว** คือลำดับธรรมชาติ
- **ขั้นตอนวิธีคัธฮิลล์–แมกคีแบบกลับลำดับ** (reverse Cuthill--McKee) คือวิธีจัดลำดับตัวไม่ทราบใหม่ให้ค่าที่ไม่เป็นศูนย์รวมกันใกล้แนวทแยง วิธีคือค้นหากว้างก่อนจากจุดที่มีเพื่อนบ้านน้อยที่สุด เรียงเพื่อนบ้านตามจำนวนเพื่อนบ้านจากน้อยไปมาก แล้วกลับลำดับทั้งหมด ตำรา Ern and Guermond 2004, §9.2, p. 399 ระบุว่าการกลับลำดับให้ผลดีกว่าสำหรับการกำจัดแบบเกาส์แบบเบาบาง

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 7.9: จำนวนค่าไม่เป็นศูนย์ของ L และจำนวนการคูณและหารของการแยกโชเลสกี สำหรับลำดับสามแบบ (ตัวเลขจากโค้ดของเรา)

n	N	ของ A	L ตามแถว	L RCM	L ตัดแบ่ง	การคูณ ตามแถว	การคูณ RCM
10	100	460	1 009	805	781	5 748	3 840
20	400	1 920	8 019	5 910	4 867	86 298	50 730
30	900	4 380	27 029	19 315	13 874	426 648	238 670



รูปที่ 7.9: จำนวนค่าไม่เป็นศูนย์ของ L เทียบกับจำนวนตัวไม่ทราบ $N = n^2$ บนแกนลอการิทึม สามเส้นคือสามลำดับ เส้นประคือจำนวนค่าไม่เป็นศูนย์ของ A (สร้างจาก `numanal/ch07_sparse.py`)

- **การตัดแบ่งซ้อน.** ตัดผ้าด้วยเส้นกลางของด้านที่ยาวกว่า นับสองฝั่งก่อน แล้วนับเส้นตัดทีหลัง ทำซ้ำในแต่ละครึ่ง ลำดับนี้เป็นของเราที่สร้างขึ้นเอง ไม่ได้อ้างหน้าตำรา

ตาราง 7.9 และรูป 7.9 แสดงจำนวนค่าไม่เป็นศูนย์ของ L และจำนวนการคูณ ของผ้าสามขนาด

ค่าในตารางแสดงว่า การเรียงแบบตัดแบ่งซ้อนให้ฟิลอินน้อยที่สุดในสามแบบ ความกว้างแถบของการเรียงตามแถวและแบบ RCM ยังเป็น n เท่ากัน จึงแสดงว่าความกว้างแถบไม่ได้บอกจำนวนฟิลอินทั้งหมด การตัดแบ่งซ้อนมีความกว้างแถบใหญ่ที่สุดในสามแบบ (326 สำหรับ $n = 20$) แต่ฟิลอินน้อยที่สุด ตัวอย่างหนึ่งคือการเรียงแบบ red--black ที่ตำรา Iserles 2009, §11.2, p. 238 บอกว่าไม่ดีเมื่อต้องการความกว้างแถบเล็ก ในผ้า 10×10 จุดการเรียงแบบ red--black มีความกว้างแถบ 55 แต่ให้ 713 ค่าไม่เป็นศูนย์ใน L น้อยกว่าการเรียงตามแถว ($1{,}009$) และเป็นการสังเกตจากโค้ดของเรา

ข้อควรระวัง: ลำดับเป็นเรื่องของวิธี ไม่ใช่ของปัญหา

ลำดับไม่เปลี่ยนคำตอบในทางคณิตศาสตร์ แต่เปลี่ยนหน่วยความจำและจำนวนการคำนวณ ความไวของระบบต่อข้อมูลไม่เปลี่ยนตามลำดับเลย ความไวนั้นเป็นเรื่องของหัวข้อ 7.2

ตำรา Grossmann, Roos, and Stynes 2007, Remark 8.1, p. 504 กล่าวถึงวิธีแก้แบบแนวหน้า (frontal solution method) ซึ่งประกอบเมทริกซ์และแยก LU ไปพร้อมกันทีละส่วน จึงไม่ต้องเก็บเมทริกซ์ทั้งหมดในหน่วยความจำ ตำรา Iserles ระบุว่าวิธีทำซ้ำและตัวแก้แบบเร็วอยู่ในบทที่ 12--15 Iserles 2009, p. 243

ร่างพิมพ์ · preprint · 2026-10-23

กลับไปที่ผ้า 100×100 จุด

สรุปตัวเลขของผ้า 100×100 จุด ทุกตัวมาจากสูตรที่ตรวจแล้วในหัวข้อนี้ เมทริกซ์ของผ้าที่มี $10{,}000$ ตัวไม่ทราบ มีค่าไม่เป็นศูนย์ $49{,}600$ ค่า และถ้าเก็บเต็มต้องใช้ราว 800 MB ส่วนการเก็บแบบ CSR ใช้ราว 0.64 MB

การแยกโชเลสกีแบบเต็มต้องการการคูณและหารราว 1.67×10^{11} ครั้ง (ราว $N^3/6$) เมื่อเรียงตามแถวแบบแถบ จำนวนนี้ลดเหลือ 5.08×10^{7} ครั้ง และ L มี 1.0×10^{6} ค่าไม่เป็นศูนย์ ซึ่งยังมากกว่าค่าไม่เป็นศูนย์ของ A ราว 20 เท่า ตัวเลขการเรียงตามแถวมาจากสูตรที่ตรวจแล้ว ไม่ได้มาจากการแยกเมทริกซ์ขนาดนี้จริงในหนังสือ

สรุปใจความ

- ผ้าที่ประกอบจากจุดที่ผูกกับเพื่อนบ้านทำให้เมทริกซ์มีศูนย์ส่วนใหญ่ เก็บแบบ CSR ได้โดยใช้หน่วยความจำเพียงเศษเสี้ยวของการเก็บเต็ม
- ความกว้างแถบกำหนดค่าใช้จ่ายของการแยกแบบแถบ และเมทริกซ์สมมาตรบวกแน่นอนแยกได้ด้วยโชเลสกีซึ่งครึ่งหนึ่งของ LU
- ฟิลอินคือศูนย์ที่กลายเป็นไม่ศูนย์ ถ้ากราฟเป็นต้นไม้และเรียงแบบขึ้นสู่ราก ไม่มีฟิลอินเลย แต่วงปิดสร้างฟิลอิน
- ลำดับการเรียงจุดเปลี่ยนจำนวนฟิลอินและการคำนวณ แต่ไม่เปลี่ยนคำตอบ การตัดแบ่งซ้อนให้ฟิลอินน้อยที่สุดในสามแบบที่ทดสอบ
- กลับไปที่ผ้า: การเรียงตามแถวใช้ราว 5×10^{7} การคูณ แทนที่จะเป็น 1.7×10^{11} ของเมทริกซ์เต็ม

อ่านต่อ. ตำรา Iserles 2009, §11.1--11.2, pp. 233--243 อธิบายเมทริกซ์แถบ กราฟ และการแยกโชเลสกีของเมทริกซ์บางอย่างที่ไม่มีฟิลอิน Iserles 2009, ภาคผนวก A.1.4.6, p. 436 ให้สูตรของโชเลสกี Ern and Guermond 2004, §8.3, pp. 372--375 และ Ern and Guermond 2004, §9.2, pp. 393--400 อธิบายการเก็บแบบ CSR และการเรียงลำดับเพื่อลดฟิลอิน ส่วน Grossmann, Roos, and Stynes 2007, Remarks 8.1--8.3, pp. 504--509 อธิบายวิธีแก้แบบแนวหน้าและตัวแก้แบบเร็ว

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 7.3.1 (⋆) ผ้าสามคูณสามจุด ($N = 9$) มีเมทริกซ์ A ตามแบบของหัวข้อนี้ คือ 4 บนแนวทแยง และ -1 ที่ทุกคู่เพื่อนบ้าน

1. จงนับค่าไม่เป็นศูนย์ของ A และยืนยันว่าได้ $5n^2 - 4n = 33$
2. ถ้าค่าแต่ละตัวใช้ 8 ไบต์ และดัชนีคอลัมน์ใช้ 4 ไบต์ จงหาจำนวนไบต์ของการเก็บเต็ม และของการเก็บแบบ CSR ที่มีอาร์เรย์ชี้แถว $N + 1$ ตัว
3. เมื่อ N ใหญ่ขึ้น อัตราส่วนระหว่างการเก็บเต็มกับการเก็บแบบ CSR เปลี่ยนไปอย่างไร ให้ใช้ N^2 เทียบกับ $5N$ เป็นแนวคิด

แบบฝึกหัด 7.3.2 (⋆) ผ้าขนาด 3×5 จุด ($N = 15$) มีเพื่อนบ้านในแนวตั้งและแนวนอน จุดที่แถว i คอลัมน์ j อยู่ในตำแหน่ง $(i-1) \cdot 5 + j$ เมื่อเรียงตามแถว

1. เรียงหมายเลขตามแถว (แถวละห้าจุด) จงหาความกว้างแถบ
2. เรียงหมายเลขตามคอลัมน์ (คอลัมน์ละสามจุด) จงหาความกว้างแถบ

ฉบับร่าง · Preprint · 2026-10-23

3. จงนับจำนวนค่าไม่เป็นศูนย์ของ L (รวมแนวทแยงและนับเฉพาะส่วนล่าง) สำหรับทั้งสองการเรียง ผลนี้บอกอะไรเกี่ยวกับการเลือกเรียงตามด้านที่สั้นกว่า (ข้อนี้ใช้โค้ดของบท)

แบบฝึกหัด 7.3.3 (⋆⋆) ให้ $A_4 = \operatorname{tridiag}(-1, 2, -1)$ ขนาด 4×4 (โซ่สี่ลูก)

1. จงหาการแยกโชเลสกี $A_4 = LL^T$ ด้วยมือ โดยหา $L_{11}, \ldots, L_{44}$ และตัวนอกแนวทแยงทั้งสามตัว
2. จงแสดงว่า $LL^T = A_4$ โดยคูณเมทริกซ์ตรง ๆ
3. จงนับจำนวนค่าไม่เป็นศูนย์ของ L และอธิบายว่าเหตุใดในการแยกนี้ไม่มีฟิลอิน

แบบฝึกหัด 7.3.4 (⋆⋆) ให้ A เป็นเมทริกซ์สมมาตรบวกแน่นอนขนาด $N \times N$ และให้ A_k เป็นเมทริกซ์ย่อยมุมบนซ้ายขนาด $k \times k$ ให้ $A = LL^T$ เป็นการแยกโชเลสกี

1. จงแสดงว่า $\det A_k > 0$ สำหรับทุก k
2. จงแสดงว่า $\ell_{kk}^2 = \det A_k / \det A_{k-1}$ (ใช้ว่า L_k คือเมทริกซ์ย่อยมุมบนซ้ายของ L แล้ว $A_k = L_k L_k^T$)
3. ตรวจด้วยมือสำหรับ $A = \begin{bmatrix} 4 & 2 \\ 2 & 5 \end{bmatrix}$ ว่า $\ell_{11}^2 = 4$ และ $\ell_{22}^2 = 16/4$

แบบฝึกหัด 7.3.5 (⋆⋆) กราฟดาวคือกราฟที่มีจุดศูนย์กลางหนึ่งจุดผูกกับใบ k จุด โดยใบไม่ผูกกันเอง ให้ A เป็นเมทริกซ์สมมาตรบวกแน่นอนของกราฟนี้ โดยแนวทแยงของทุกจุดเท่ากับจำนวนเพื่อนบ้านบวกหนึ่ง และสมาชิกนอกแนวทแยงของเส้นเชื่อมเป็น -1

1. จงเรียงให้ใบมาก่อนและจุดศูนย์กลางท้ายสุด (เป็นการเรียงแบบขึ้นสู่ราก) จงหาจำนวนค่าไม่เป็นศูนย์ของ L และแสดงว่าไม่มีฟิลอิน
2. จงเรียงให้จุดศูนย์กลางมาก่อน จงหาจำนวนค่าไม่เป็นศูนย์ของ L และบอกว่าฟิลอินเกิดที่ตำแหน่งใด
3. จงหาสูตรของจำนวนค่าไม่เป็นศูนย์ของ L ในข้อ 2 เป็นฟังก์ชันของ k และตรวจว่าเมื่อ $k = 4$ ได้ 15

แบบฝึกหัด 7.3.6 (⋆ ⋆ ⋆) โครงงานเล็ก: เรียงผ้า 20×20 จุดด้วยลำดับ red--black คือเรียงจุดที่ $i + j$ เป็นเลขคู่ก่อน แล้วตามด้วยจุดที่ $i + j$ เป็นเลขคี่ ให้เทียบกับลำดับตามแถวและลำดับแบบตัดแบ่งซ้อนของตาราง 7.9

1. จงหาความกว้างแถบ และจำนวนค่าไม่เป็นศูนย์ของ L สำหรับการเรียงแบบ red--black (ใช้โค้ดของบท)
2. จากตาราง 7.9 และข้อ 1 จงเปรียบเทียบสามลำดับนี้ทั้งความกว้างแถบและฟิลอิน ลำดับใดให้ฟิลอินน้อยที่สุด
3. อธิบายว่าเหตุใดความกว้างแถบกับจำนวนฟิลอินจึงให้ลำดับที่ต่างกัน และลองทำซ้ำข้อ 1 ถึง 2 สำหรับ $n = 30$ (แถวและแบบตัดแบ่งซ้อนจากตาราง)

แบบฝึกหัด 7.3.7 (⋆⋆) เครื่องหนึ่งมีหน่วยความจำ 1 กิบิไบต์ (2^{30} ไบต์) และเก็บแต่ละค่าเป็นเลขจุดลอยตัว 8 ไบต์ สำหรับผ้า $n \times n$ จุด

1. จงหาค่า n ที่มากที่สุดซึ่งเก็บเมทริกซ์ A แบบเต็มได้ในหน่วยความจำนี้ (ตรวจด้วยการคำนวณ ไม่ใช่การประมาณ)
2. สำหรับค่า n นั้น จงหาจำนวนไบต์ของการเก็บแบบ CSR และเทียบกับการเก็บเต็ม
3. การแยกโชเลสกีแบบเต็มของ N ตัวไม่ทราบต้องการการคูณและหารราว $N^3/6$ ครั้ง จงหาจำนวนนี้โดยประมาณสำหรับค่า n นั้น และอธิบายว่าเหตุใดความจำจึงไม่ใช่ข้อจำกัดเดียว

ฉบับร่าง Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 7.3.1.

1. แนวทแยงมี 9 ค่า และมีเส้นเชื่อม 12 เส้น (ผ้าสามคูณสามมีคู่เพื่อนบ้านแนวนอน 6 คู่และแนวตั้ง 6 คู่) แต่ละเส้นให้ค่า -1 สองช่องที่สมมาตรกัน จึงได้ $9 + 24 = 33 = 5 \cdot 9 - 4 \cdot 3$
2. การเก็บเต็ม: $9 \cdot 9 \cdot 8 = 648$ ไบต์ การเก็บแบบ CSR: ค่า $33 \cdot (8 + 4)$ ไบต์ บวกอาร์เรย์ชี้แถว $10 \cdot 4 = 40$ ไบต์ รวม 436 ไบต์
3. การเก็บเต็มโตเป็น $8N^2$ ไบต์ ส่วน CSR โตเป็นราว $12 \cdot 5N = 60N$ ไบต์ อัตราส่วนจึงโตราว $2N/15$ เมื่อ N ใหญ่ สำหรับผ้า 100×100 จุดได้ราว 1,300 (ตรงกับ 1,259.4 ในตารางของหัวข้อนี้)

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_sparse.py` ว่า 33 ค่า และ 648 กับ 436 ไบต์

เฉลย 7.3.2.

1. จุด (i, j) กับ $(i + 1, j)$ ห่างกัน 5 หมายเลข ไม่มีคู่ที่ห่างกันมากกว่านี้ ความกว้างแถบคือ 5
2. จุด (i, j) กับ $(i, j + 1)$ ห่างกัน 3 หมายเลข ความกว้างแถบคือ 3 เมื่อเรียงตามคอลัมน์สั้น
3. ตรวจด้วยการนับแบบสัญลักษณ์ของโค้ด: การเรียงตามแถวได้ 69 ค่า และการเรียงตามคอลัมน์ได้ 53 ค่า ดังนั้นเรียงตามด้านสั้นให้ฟิลอินน้อยกว่า และทั้งสองแบบมีความกว้างแถบต่างกันด้วย

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_sparse.py` ทั้งความกว้างแถบ 5 และ 3 และจำนวน 69 กับ 53

เฉลย 7.3.3.

1. จาก $L_{11}^2 = 2$ ได้ $L_{11} = \sqrt{2}$ และ $L_{11}L_{21} = -1$ จึง $L_{21} = -1/\sqrt{2}$ ต่อไป $L_{21}^2 + L_{22}^2 = 2$ ให้ $L_{22} = \sqrt{3/2}$ แล้ว $L_{32} = -1/L_{22} = -\sqrt{2/3}$ และ $L_{33} = \sqrt{2 - 2/3} = \sqrt{4/3}$ สุดท้าย $L_{43} = -1/L_{33} = -\sqrt{3/4}$ และ $L_{44} = \sqrt{2 - 3/4} = \sqrt{5/4}$
2. แถวที่ 4 ของ LL^T คือ $(0,\ 0,\ L_{43}L_{33},\ L_{43}^2 + L_{44}^2) = (0,\ 0,\ -1,\ 2)$ ซึ่งเป็นแถวที่ 4 ของ A_4 แถวอื่นตรวจแบบเดียวกัน
3. L มี 4 ค่าบนแนวทแยงและ 3 ค่าใต้แนวทแยง รวม 7 ค่า โครงสร้างของโซ่เป็นต้นไม้ และการเรียงตามลำดับ $1, \ldots, 4$ เป็นแบบขึ้นสู่ราก (ราก 4) ตามทฤษฎีบทต้นไม้ไม่มีฟิลอิน

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_sparse.py` ว่า $L_{44} = \sqrt{5/4}$, $L_{43} = -\sqrt{3/4}$ และ $LL^T = A_4$ และว่ามี 7 ค่า

เฉลย 7.3.4.

1. ให้ $\mathbf{y} \in \mathbb{R}^k$ ไม่เป็นศูนย์ และ $\mathbf{x} = (\mathbf{y}, \mathbf{0})$ จะได้ $\mathbf{y}^T A_k \mathbf{y} = \mathbf{x}^T A \mathbf{x} > 0$ จึง A_k บวกแน่นอน ค่าลักษณะเฉพาะทั้งหมดของ A_k เป็นบวก ดังนั้น $\det A_k$ ซึ่งเป็นผลคูณของค่าลักษณะเฉพาะเป็นบวก
2. เพราะ L เป็นสามเหลี่ยมล่าง บล็อกบนซ้ายขนาด $k \times k$ ของ LL^T เท่ากับ $L_k L_k^T$ โดย L_k คือบล็อกบนซ้ายของ L จึง $\det A_k = (\det L_k)^2 = \prod_{i \le k} \ell_{ii}^2$ หารด้วย $\det A_{k-1}$ จะได้ $\ell_{kk}^2 = \det A_k / \det A_{k-1}$
3. $\det A_1 = 4$ และ $\det A_2 = 4 \cdot 5 - 2 \cdot 2 = 16$ จึง $\ell_{11}^2 = 4/1 = 4$ และ $\ell_{22}^2 = 16/4 = 4$ ตรงกับการแยกจริง $\ell_{11} = 2$, $\ell_{21} = 1$, $\ell_{22} = 2$

ตรวจด้วยโค้ด: ทดสอบ $\ell_{kk}^2 = \det A_k / \det A_{k-1}$ บนเมทริกซ์สุ่มขนาด 5×5 ด้วยเมล็ดสุ่มคงที่ และค่า 4 กับ 4 ของข้อ 3

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 7.3.5.

1. ใบทุกตัวมีเพื่อนบ้านเพียงศูนย์กลาง ซึ่งอยู่ท้ายสุด ดังนั้นคอลัมน์ของใบแต่ละตัวมีค่าใต้แนวทแยงหนึ่งตัวที่ตำแหน่งศูนย์กลาง และไม่มีใบใดสร้างเส้นเชื่อมใหม่ คอลัมน์ของศูนย์กลางไม่มีค่าใต้แนวทแยง L มี $(k+1)+k = 2k+1$ ค่า และเท่ากับส่วนล่างของ A จึงไม่มีฟิลอิน สำหรับ $k=4$ ได้ 9

2. ศูนย์กลางถูกกำจัดก่อน มันผูกกับใบทั้ง k ตัว การกำจัดจึงทำให้ใบทุกคู่ผูกกันเอง ได้ฟิลอิน $\binom{k}{2}$ ตำแหน่ง ที่ตำแหน่ง $(\text{ใบ}_i, \text{ใบ}_j)$ สำหรับ $i > j$ ทั้งหมด จำนวนค่าไม่เป็นศูนย์ของ L เท่ากับ $(k+1)+k+\binom{k}{2}$ ซึ่งสำหรับ $k=4$ ได้ $5+4+6=15$

3. จาก $(k+1)+k+k(k-1)/2 = (k+1)(k+2)/2$ ซึ่งเป็นจำนวนสมาชิกของสามเหลี่ยมล่างทั้งหมดของเมทริกซ์ขนาด $(k+1)\times(k+1)$ นั่นคือ L กลายเป็นเมทริกซ์สามเหลี่ยมล่างเต็มหลังจากกำจัดศูนย์กลางก่อน สำหรับ $k=4$ ได้ $5\cdot 6/2 = 15$

ตรวจด้วยโค้ด: ทดสอบจำนวน $(k+1)+k+\binom{k}{2}$ และ $2k+1$ สำหรับ $k = 3, 4, 5, 6$

เฉลย 7.3.6.

1. ผลจากโค้ด: การเรียงแบบ red--black ของผ้า 20×20 จุดมีความกว้างแถบ 210 และ L มี $4\,878$ ค่าไม่เป็นศูนย์ การแยกใช้การคูณและหาร $44\,382$ ครั้ง

2. ตาราง 7.9 กับข้อ 1: สำหรับ $n=20$ การเรียงตามแถวมีความกว้างแถบ 20 และ $8\,019$ ค่า การตัดแบ่งซ้อนมี $4\,867$ ค่า และ red--black มี $4\,878$ ค่า ดังนั้นการตัดแบ่งซ้อนและ red--black ให้ฟิลอินน้อยกว่าการเรียงตามแถวมาก ขณะที่ red--black มีความกว้างแถบ 210 ซึ่งกว้างกว่าสองลำดับแรกแต่แคบกว่าการตัดแบ่งซ้อน (326) สำหรับ $n=10$ red--black น้อยที่สุดที่ 713 ค่า ลำดับที่ดีที่สุดจึงขึ้นกับขนาดของผ้า

3. ความกว้างแถบบอกระยะที่ไกลที่สุดของการผูก แต่ฟิลอินนับว่าการกำจัดแต่ละจุดสร้างการผูกใหม่กี่คู่ ในข้อ 1 และ 2 red--black แยกผ้าเป็นสองชั้นที่แต่ละชั้นไม่ผูกกันเอง จึงทำให้จุดสร้างการผูกใหม่น้อย แม้ระยะไกลที่สุดจะกว้าง สำหรับ $n=30$ red--black ได้ $15\,493$ ค่า การตัดแบ่งซ้อนได้ $13\,874$ ค่า ซึ่งน้อยที่สุดในสี่ลำดับ (ตามแถว $27\,029$ และ RCM $19\,315$)

ตรวจด้วยโค้ด: ทดสอบความกว้างแถบ 210, $\mathrm{nnz}(L) = 4\,878$ และการคูณ $44\,382$ สำหรับ $n=20$ และ $\mathrm{nnz}(L) = 15\,493$ สำหรับ $n=30$ (red--black)

เฉลย 7.3.7.

1. ค่า n ที่เก็บเต็มได้ต้องมี $8N^2 \le 2^{30}$ โดย $N = n^2$ สำหรับ $n=107$ ได้ $N = 11\,449$ และ $8N^2 = 1\,048\,636\,808$ ไบต์ ซึ่งไม่เกิน $2^{30} = 1\,073\,741\,824$ ส่วน $n=108$ ได้ $8\cdot 11\,664^2 = 1\,088\,391\,168$ ซึ่งเกิน ดังนั้น $n=107$

2. จำนวนค่าไม่เป็นศูนย์คือ $5\cdot 11\,449 - 4\cdot 107 = 56\,817$ จึงได้ CSR $= 56\,817\cdot 12 + 11\,450\cdot 4 = 727\,604$ ไบต์ เล็กกว่าการเก็บเต็มราว $1\,440$ เท่า

3. การแยกโชเลสกีเต็มต้องการการคูณและหารราว $N^3/6 \approx 2.5\times 10^{11}$ ครั้ง (นับแม่นยำได้ $250\,187\,257\,476$) จำนวนนี้มากมหาศาล ดังนั้นแม้หน่วยความจำจะพอ จำนวนการคำนวณก็ยังเป็นข้อจำกัด

ตรวจด้วยโค้ด: ทดสอบ $n=107$, $n=108$, $727\,604$ ไบต์ และ $250\,187\,257\,476$ การคูณ

ฉบับร่าง preprint · 2026-10-23

7.4 กลับไปสมการไม่เชิงเส้น: หลายตัวแปร

สถานการณ์สมมติ: แขนสองท่อนที่ต้องไปถึงจุดที่คลิก

ลองนึกถึงตัวละครที่มีแขนสองท่อน ท่อนบนติดกับไหล่ ท่อนล่างติดกับข้อศอก แล้วผู้เล่นคลิกจุดหนึ่งบนจอ ตัวเกมต้องหาว่าไหล่ควรหมุนไปเท่าไร และข้อศอกควรงอเท่าไร เพื่อให้มือไปถึงจุดนั้นพอดี
มุมสองตัวนี้คือตัวไม่ทราบสองตัว และการที่มือไปถึงจุดที่คลิกให้สมการสองสมการ สมการเหล่านี้ไม่เป็นเชิงเส้น เพราะมีไซน์และโคไซน์ของมุมอยู่ ผู้เล่นไม่อยากรอ ตัวเกมจึงควรหาคำตอบในไม่กี่รอบ และเราอยากรู้ว่าแต่ละรอบต้องทำอะไร
(แขนนี้เป็นแบบจำลองสมมติที่เราสร้างเอง ความยาวท่อนบน 1 และท่อนล่าง 0.8 เป็นค่าสมมติ ตัวเลขทั้งหมดในหัวข้อนี้มาจากโค้ดของเรา และไม่ได้บอกว่าเกมหรือเอนจินใดคำนวณท่าแขนด้วยวิธีนี้)

ภาพรวมก่อนอ่าน

ในหัวข้อ 6.2 วิธีนิวตันหารากของ $f(x) = 0$ ด้วยเส้นสัมผัส เมื่อตัวไม่ทราบมีหลายตัว เส้นสัมผัสกลายเป็นเมทริกซ์ที่เรียกว่า**เมทริกซ์จาโคเบียน** (Jacobian matrix) ทุกรอบของวิธีนิวตันจึงแก้ระบบเชิงเส้นหนึ่งระบบ ด้วยเครื่องมือจากหัวข้อ 7.1 และ 7.2 ภาพแรกแสดงแขนที่ค่อย ๆ ไปถึงจุดที่คลิกในสามรอบ

มือของแขนเป็นฟังก์ชันของมุม

ให้ท่อนแรกยาว L_1 และท่อนที่สองยาว L_2 วัดมุม θ_1 ที่ไหล่ และมุม θ_2 ที่ข้อศอก โดยวัดข้อศอกจากทิศของท่อนแรก ท่อนแรกชี้ไปทางมุม $\varphi_1 = \theta_1$ และท่อนที่สองชี้ไปทางมุม $\varphi_2 = \theta_1 + \theta_2$ ตำแหน่งของมือจึงเป็น

$$x(\theta) = L_1 \cos \varphi_1 + L_2 \cos \varphi_2, \qquad y(\theta) = L_1 \sin \varphi_1 + L_2 \sin \varphi_2.$$

ถ้าผู้เล่นคลิกจุด $T = (T_x, T_y)$ เราต้องหา θ ที่ทำให้ค่าตกค้าง $F(\theta) = \big(x(\theta) - T_x,\ y(\theta) - T_y\big)$ เป็นศูนย์ทั้งสององค์ประกอบ นี่คือระบบสมการไม่เชิงเส้นสองสมการ สองตัวไม่ทราบ θ_1 และ θ_2

ที่ท่าเริ่มต้น $\theta^{(0)} = (0.2,\ 1.0)$ กับเป้า $T = (1.2,\ 0.9)$ เราได้ $x = \cos 0.2 + 0.8 \cos 1.2 = 1.269953$ และ $y = \sin 0.2 + 0.8 \sin 1.2 = 0.944301$ จึง $F(\theta^{(0)}) = (0.069953,\ 0.044301)$ และ $\|F(\theta^{(0)})\|_2 = 0.0828$ มือยังไม่ถึงจุดที่คลิก

อ่านแบบเข้าใจง่าย

ค่าตกค้างหนึ่งค่ามีสององค์ประกอบ องค์ประกอบแรกบอกว่ามือต้องขยับในแนวนอนเท่าไร องค์ประกอบที่สองบอกแนวตั้ง ปัญหาคือหา θ ที่ทำให้ทั้งสองค่าเป็นศูนย์พร้อมกัน ซึ่งเหมือนหาจุดตัดของเส้นโค้งสองเส้นบนระนาบ

เส้นสัมผัสกลายเป็นเมทริกซ์จาโคเบียน

นิยาม 7.4.1: เมทริกซ์จาโคเบียน

ให้ $F = (F_1, \ldots, F_m)$ เป็นฟังก์ชันของตัวไม่ทราบ $\theta = (\theta_1, \ldots, \theta_n)$ เมทริกซ์จาโคเบียนของ F คือเมทริกซ์ขนาด $m \times n$ ที่สมาชิกแถว i คอลัมน์ j เป็น $\partial F_i / \partial \theta_j$ Atkinson 1989, §2.10, p. 105

อ่านแบบเข้าใจง่าย. คอลัมน์ที่ j บอกว่าขยับมุม θ_j นิดเดียว ทั้งสองสมการเปลี่ยนไปเท่าไร เมทริกซ์นี้คือเส้นสัมผัสของระนาบ ถ้าเราขยับ θ ไปเป็น $\theta + d$ โดย d เล็ก ค่าตกค้างจะเปลี่ยนประมาณ $J\,d$

ฉบับร่าง · preprint · 2026-10-23

สำหรับแขนของเรา มุม θ_1 อยู่ทั้งใน φ_1 และ φ_2 จึงหมุนไหล่แล้วข้อศอกเหยียดตรงหรือพับทับ $\theta_2 = 0$ หรือ π และ J ไม่มีอินเวอร์ส

$$J(\theta) = \begin{bmatrix} -L_1 \sin\varphi_1 - L_2 \sin\varphi_2 & -L_2 \sin\varphi_2 \\ L_1 \cos\varphi_1 + L_2 \cos\varphi_2 & L_2 \cos\varphi_2 \end{bmatrix}.$$

ที่ท่าเริ่มต้น $J(\theta^{(0)}) \approx \begin{bmatrix} -0.9443 & -0.7456 \\ 1.2700 & 0.2899 \end{bmatrix}$ และดีเทอร์มิแนนต์ของแขนสองท่อนมีค่า $\det J = L_1 L_2 \sin\theta_2$ (ดูแบบฝึกหัด 7.4.4) ดังนั้นเมื่อข้อศอกเหยียดตรงหรือพับทับ $\theta_2 = 0$ หรือ π และ J ไม่มีอินเวอร์ส

นิวตันสำหรับระบบ: แก้ระบบเชิงเส้นทุกรอบ

แนวคิด. ถ้า θ อยู่ใกล้คำตอบ เราประมาณ $F(\theta + d) \approx F(\theta) + J(\theta)\, d$ แล้วหา d ที่ทำให้ค่าประมาณนี้เป็นศูนย์ ซึ่งคือระบบเชิงเส้น $J(\theta)\, d = -F(\theta)$

นิยาม 7.4.2: วิธีนิวตันสำหรับระบบ

ให้ $\theta^{(0)}$ เป็นค่าเดาตั้งต้น **วิธีนิวตัน** (Newton's method) สำหรับระบบคือ สำหรับ $k = 0, 1, \ldots$ แก้ระบบเชิงเส้น $J(\theta^{(k)})\, d^{(k)} = -F(\theta^{(k)})$ แล้วตั้ง $\theta^{(k+1)} = \theta^{(k)} + d^{(k)}$ Atkinson 1989, §2.11, (2.11.4)--(2.11.5), p. 109

อ่านแบบเข้าใจง่าย

ในแต่ละรอบเราไม่กลับ J เป็นอินเวอร์สทั้งเมทริกซ์ เราแก้ระบบสองสมการสองตัวไม่ทราบสำหรับ d โดยแยก J ด้วย LU พร้อมการเลือกตัวหลักของหัวข้อ 7.2 แล้วแทนค่าไปข้างหน้าและย้อนกลับ ตำรา Atkinson ระบุว่าการแก้ระบบใช้การดำเนินการราวหนึ่งในสามของการหาอินเวอร์ส และการแยก LU ของระบบ m ตัวไม่ทราบใช้ราว $\frac{2}{3}m^3$ การดำเนินการตามการนับของหัวข้อ 7.1 ส่วนการสร้าง J ต้องหาอนุพันธ์ย่อยราว m^2 ตัว

ทำรอบแรกด้วยมือ ระบบคือ

$$-0.94430\, d_1 - 0.74563\, d_2 = -0.069953,$$
$$1.26995\, d_1 + 0.28989\, d_2 = -0.044301.$$

องค์ประกอบแรกของคอลัมน์แรกคือ -0.94430 ซึ่งมีค่าสัมบูรณ์น้อยกว่า 1.26995 ตามการเลือกตัวหลักตามหลัก เราจึงสลับสองแถว ตัวหลักเป็น 1.26995 และตัวคูณคือ $m = -0.94430/1.26995 \approx -0.74354$ แถวที่สองกลายเป็น $-0.53009\, d_2 = -0.10289$ จึง $d_2 \approx 0.19410$ แล้วแทนกลับได้ $d_1 \approx -0.07920$ ดังนั้น $\theta^{(1)} = (0.2 - 0.0792,\ 1.0 + 0.1941) \approx (0.1208,\ 1.1941)$ และ $\|F(\theta^{(1)})\|_2 \approx 0.0073$ ค่าตกค้างลดลงราว 11 เท่าในรอบเดียว ตัวเลขในโค้ดละเอียดกว่าที่พิมพ์ในที่นี้

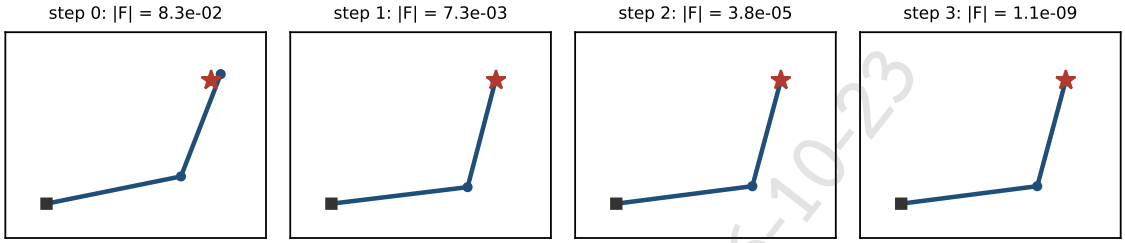
รอบต่อไปทำแบบเดียวกันด้วยโค้ด ตาราง 7.10 แสดงสี่รอบแรก และรูป 7.10 แสดงแขนในแต่ละรอบ

ตารางแสดงว่า ตั้งแต่รอบที่สอง จำนวนหลักที่ถูกต้องของคำตอบราวเพิ่มเป็นสองเท่าทุกรอบ ค่าในคอลัมน์สุดท้ายอยู่ที่ราว 0.33 ถึง 0.35 ในสองรอบท้าย นี่คือค่าคลาดเคลื่อนสัมบูรณ์แบบกำลังสองที่ตำรา Atkinson 1989, §2.11, (2.11.7), p. 110 และ Iserles 2009, §7.2, p. 127 กล่าวถึงว่าเมื่อเริ่มใกล้พอ $\|\theta^{(k+1)} - \theta^*\| \le C\|\theta^{(k)} - \theta^*\|^2$ โดยมีค่าคงที่ $C > 0$ ตัวเลข 0.33 เป็นค่าที่วัดได้ของแขนนี้ ไม่ใช่ค่าที่ทฤษฎีกำหนด

ร่างไม่เป็นทางการ · Preprint · 2026-10-23

ตารางที่ 7.10: วิธีนิวตันบนแขนสองท่อนจากท่าเริ่มต้น $(0.2,\ 1.0)$ ค่า $\|\theta^{(k)}-\theta^*\|_2$ วัดจากคำตอบแม่นที่สูตรปิดให้ (ตัวเลขจากโค้ดของเรา)

k	$\theta_1^{(k)}$	$\theta_2^{(k)}$	$\|F(\theta^{(k)})\|_2$	$\|\theta^{(k)}-\theta^*\|_2$	$\|\theta^{(k+1)}-\theta^*\|_2/\|\theta^{(k)}-\theta^*\|_2^2$
0	0.200000	1.000000	8.280×10^{-2}	1.936×10^{-1}	0.430
1	0.120808	1.194109	7.332×10^{-3}	1.611×10^{-2}	0.327
2	0.127874	1.179726	3.847×10^{-5}	8.483×10^{-5}	0.348
3	0.127907	1.179648	1.133×10^{-9}	2.505×10^{-9}	--
4	0.127907	1.179648	2.5×10^{-16}	$\approx10^{-16}$	--



รูปที่ 7.10: แขนสองท่อนในรอบที่ 0 ถึง 3 ของวิธีนิวตัน ดาวคือจุดที่คลิก และ $|F|$ คือขนาดของค่าตกค้างในแต่ละรอบ (สร้างจาก `numanal/ch07_arm.py`)

ทฤษฎีบท 7.4.1: นิวตันลู่เข้าแบบกำลังสองเมื่อเริ่มใกล้

ให้ F มีอนุพันธ์ต่อเนื่องสองครั้งใกล้ θ^* โดย $F(\theta^*)=0$ และ $J(\theta^*)$ ไม่เอกฐาน แล้วมี $C>0$ และย่าน U ของ θ^* ซึ่งถ้า $\theta^{(k)}\in U$ แล้ว $\theta^{(k+1)}\in U$ และ $\|\theta^{(k+1)}-\theta^*\|\le C\|\theta^{(k)}-\theta^*\|^2$ Atkinson 1989, §2.11, (2.11.7), p. 110

พิสูจน์. **แนวคิด.** ขยาย F รอบ θ^* ด้วยเทย์เลอร์ ส่วนที่เหลือเป็นกำลังสองของระยะจาก θ^* เมื่อแก้ระบบเชิงเส้นด้วย J ที่ใกล้ $J(\theta^*)$ ค่าคลาดเคลื่อนสัมบูรณ์ที่เหลือหลังหนึ่งรอบจึงเป็นกำลังสองด้วย และย่านเล็กพอจะถูกรักษาไว้ในรอบต่อไป

1. เนื่องจาก $J(\theta^*)$ ไม่เอกฐาน และ J ต่อเนื่อง มีค่าคงที่ c และ $r_0>0$ ที่ $\|J(\theta)^{-1}\|\le c$ เมื่อ $\|\theta-\theta^*\|\le r_0$
2. เพราะ F หาอนุพันธ์ได้ต่อเนื่องสองครั้ง และ $F(\theta^*)=0$ มีค่าคงที่ C_1 และ C_2 (ลด r_0 ได้ถ้าจำเป็น) ที่ $\|F(\theta)-J(\theta^*)(\theta-\theta^*)\|\le C_1\|\theta-\theta^*\|^2$ และ $\|J(\theta)-J(\theta^*)\|\le C_2\|\theta-\theta^*\|$ เมื่อ $\|\theta-\theta^*\|\le r_0$ ให้ $\theta^+=\theta-J(\theta)^{-1}F(\theta)$ จะได้ $\theta^+-\theta^*=J(\theta)^{-1}\big[(J(\theta)-J(\theta^*))(\theta-\theta^*)-\big(F(\theta)-J(\theta^*)(\theta-\theta^*)\big)\big]$ จึง $\|\theta^+-\theta^*\|\le C\|\theta-\theta^*\|^2$ โดย $C=c(C_1+C_2)$
3. เลือกรัศมี $r\le r_0$ ที่ $Cr\le\tfrac12$ และให้ $U=\{\theta:\|\theta-\theta^*\|\le r\}$ ถ้า $\theta\in U$ แล้ว $\|\theta^+-\theta^*\|\le C\|\theta-\theta^*\|^2\le Cr\|\theta-\theta^*\|\le\tfrac12\|\theta-\theta^*\|\le r$ ดังนั้น $\theta^+\in U$ จึงใช้ข้อ 1 และข้อ 2 ซ้ำได้ในรอบถัดไป ค่าคลาดเคลื่อนสัมบูรณ์ของทุกรอบจึงไม่เกินกำลังสองของค่าคลาดเคลื่อนสัมบูรณ์ก่อนหน้า และไม่เกินครึ่งหนึ่งของค่าคลาดเคลื่อนสัมบูรณ์ก่อนหน้า ทำให้ลำดับลู่เข้าสู่ θ^*

สรุป: เมื่อ J ไม่เอกฐานที่คำตอบ และเริ่มในย่าน U ค่าคลาดเคลื่อนสัมบูรณ์หนึ่งรอบเป็นกำลังสองของค่าคลาดเคลื่อนสัมบูรณ์ก่อนหน้า และรอบต่อ ๆ ไปไม่หลุดออกจากย่าน U □

ฉบับร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย. ข้อนี้บอกว่า เมื่อเริ่มใกล้พอ จำนวนหลักที่ถูกต้องเพิ่มขึ้นราวเป็นสองเท่าทุกรอบ แต่ทฤษฎีบทไม่ได้บอกว่า "ใกล้พอ" คือเท่าไร ตัวอย่างข้างหน้าจะแสดงว่าถ้าเริ่มไกลเกินไปอาจไม่ลู่เข้าเลย

การทำซ้ำจุดตรึงในปริภูมิ $\mathbb{R}^n$

เขียนนิวตันเป็นการทำซ้ำ $\theta^{(k+1)} = G(\theta^{(k)})$ โดย $G(\theta) = \theta - J(\theta)^{-1}F(\theta)$ คำตอบ θ^* ของ $F = 0$ ทำให้ $G(\theta^*) = \theta^*$ จึงเป็น **จุดตรึง** (fixed point) ของ G การทำซ้ำแบบนี้มีทฤษฎีที่เป็นอิสระจากนิวตัน

ทฤษฎีบท 7.4.2: ทฤษฎีบทจุดตรึงในปริภูมิ $\mathbb{R}^n$ (การขยายผลโดยผู้เรียบเรียง)

ให้ $D \subset \mathbb{R}^n$ เป็นเซตปิด มีขอบเขต และนูน ให้ $G : D \to D$ และมี $0 \le \lambda < 1$ ที่ $\|G(\mathbf{u}) - G(\mathbf{v})\|_\infty \le \lambda\|\mathbf{u} - \mathbf{v}\|_\infty$ ทุก $\mathbf{u}, \mathbf{v} \in D$ แล้ว G มีจุดตรึงเดียว $\mathbf{x}^* \in D$ และสำหรับทุก $\mathbf{x}_0 \in D$ การทำซ้ำ $\mathbf{x}_{k+1} = G(\mathbf{x}_k)$ ลู่เข้าหา $\mathbf{x}^*$ โดย $\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_\infty \le \lambda\|\mathbf{x}_k - \mathbf{x}^*\|_\infty$

ที่มาและการขยายผล. ตำรา Atkinson 1989, §2.10, Thm 2.9, p. 105 ให้ผลจุดตรึงในระนาบ โดยสมมติว่าองค์ประกอบของ g หาอนุพันธ์ได้ต่อเนื่องและ $\|G\|_\infty < 1$ บน D ส่วน Iserles 2009, Thm 7.1, p. 125 ให้ผลแบบบานาคในปริภูมิ $\mathbb{R}^d$ บนลูกบอลรอบจุดเริ่ม ทฤษฎีบทข้างบนใช้ปริภูมิ $\mathbb{R}^n$ และเงื่อนไขลิพชิทซ์บนเซต D โดยไม่สมมติการหาอนุพันธ์ จึงเป็น**การขยายผลโดยผู้เรียบเรียง** รูปนี้ไม่ตรงกับทั้งสองเล่มโดยตรง ส่วนที่ทั้งสองเล่มให้ไว้คือจุดตรึงเดียวและการลู่เข้าภายใต้การหดตัว การพิสูจน์ด้านล่างเป็นของผู้เรียบเรียงเอง

พิสูจน์. **แนวคิด.** การทำซ้ำติดกันหดตัวด้วยตัวคูณ λ ผลรวมของการเปลี่ยนแปลงจึงเป็นอนุกรมเรขาคณิต

1. ความต่างของการทำซ้ำติดกันคือ $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \lambda^k\|\mathbf{x}_1 - \mathbf{x}_0\|$ โดยอุปนัย
2. การทำซ้ำเป็นลำดับโคชี เพราะผลรวมของความต่างตามอสมการสามเหลี่ยมมีขอบเขตด้วยอนุกรมเรขาคณิตที่มีอัตราส่วน $\lambda < 1$ เซต D ปิดและมีขอบเขต จึงลู่เข้าไปยังจุด $\mathbf{x}^* \in D$
3. จาก G ต่อเนื่องเรามี $G(\mathbf{x}^*) = \mathbf{x}^*$ และถ้ามีจุดตรึงสองจุด $\|\mathbf{u} - \mathbf{v}\| \le \lambda\|\mathbf{u} - \mathbf{v}\|$ บังคับให้ $\mathbf{u} = \mathbf{v}$

สรุป: การหดตัวทำให้ได้จุดตรึงเดียว และการทำซ้ำลู่เข้าด้วยอัตรา λ □

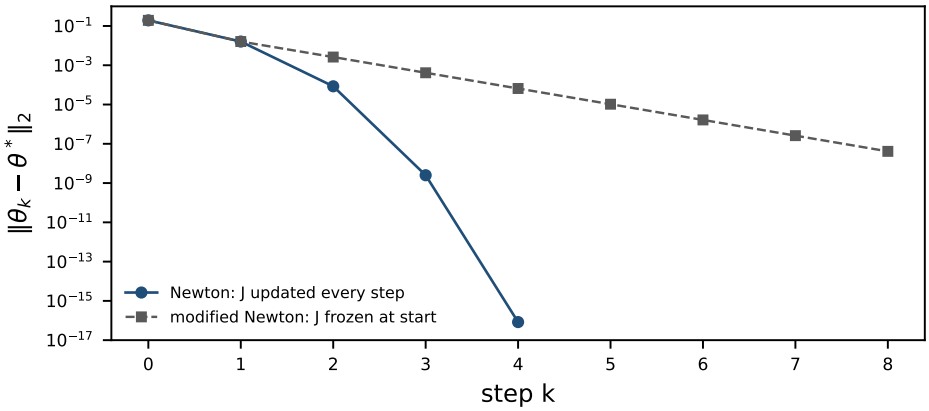
อ่านแบบเข้าใจง่าย. ค่า λ คืออัตราที่ค่าคลาดเคลื่อนสัมบูรณ์ลดลงต่อรอบ ถ้า $\lambda = 0.5$ ค่าคลาดเคลื่อนสัมบูรณ์ราวครึ่งหนึ่งในแต่ละรอบ นี่คือการลู่เข้าแบบเชิงเส้น (linear convergence) ซึ่งต่างจากแบบกำลังสองที่ค่าคลาดเคลื่อนสัมบูรณ์ยกกำลังสองทุกรอบ ฉบับนี้ใช้เงื่อนไขลิพชิทซ์แทนอนุพันธ์ เพื่อให้ตรงกับการใช้ในหัวข้อนี้

วิธีนิวตัน-ราฟสันแบบดัดแปลง: จุดตรึงที่ลู่เข้าเชิงเส้น

ถ้าเราแยก J เพียงครั้งเดียวที่ท่าเริ่มต้น แล้วใช้ $J_0 = J(\theta^{(0)})$ ตลอดทุกรอบ เรียกว่า**วิธีนิวตัน-ราฟสันแบบดัดแปลง** (modified Newton--Raphson method) ตัวทำซ้ำคือ $G_0(\theta) = \theta - J_0^{-1}F(\theta)$ ซึ่งมี θ^* เป็นจุดตรึง ค่าอนุพันธ์ที่คำตอบคือ $I - J_0^{-1}J(\theta^*)$ ซึ่งมี $\|\cdot\|_\infty \approx 0.1798$ และ**รัศมีสเปกตรัม** (spectral radius) $\rho \approx 0.1584$ รูป 7.11 แสดงว่าค่าคลาดเคลื่อนสัมบูรณ์ของวิธีนี้ลดลงราว 0.158 เท่าทุกรอบ ซึ่งตรงกับ ρ ส่วนนิวตันเต็มมี $G'(\theta^*) = 0$ ตำรา Atkinson 1989, §2.11, p. 110 ระบุว่าอนุพันธ์ของ G ในรูปนี้เป็นศูนย์ที่คำตอบ จึงได้การลู่เข้าแบบกำลังสอง

วิธีกึ่งนิวตัน. ระหว่างสองขั้วนี้มี**วิธีกึ่งนิวตัน** (quasi-Newton method) ที่ไม่คำนวณ J ใหม่ทุกรอบ แต่ปรับเมทริกซ์ที่ใช้แทนจากค่าที่ผ่านมา ตำรา Atkinson 1989, §2.11, p. 110 จัดให้วิธีนี้เป็นกลุ่มเดียวกับ Newton-like และนิวตันแบบดัดแปลง เราไม่ลงรายละเอียดของการปรับปรุงในหนังสือเล่มนี้ วิธีนิวตัน-ราฟสันแบบดัดแปลงที่เราศึกษาเป็นกรณีที่ง่ายที่สุดของกลุ่มนี้

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 7.11: ค่าคลาดเคลื่อนสัมบูรณ์ $\|\theta^{(k)} - \theta^*\|_2$ บนแกนลอการิทึม: นิวตันที่อัปเดต J ทุกรอบ (วงกลม) กับวิธีนิวตัน-ราฟสันแบบดัดแปลง (สี่เหลี่ยม) ที่ท่าเริ่มเดียวกัน (สร้างจาก `numanal/ch07_arm.py`)

ข้อควรระวัง: นิวตันอาจพังในสามกรณี

- **ท่าเหยียดตรงหรือพับ.** เมื่อ $\theta_2 = 0$ ดีเทอร์มิแนนต์ของ J เป็นศูนย์ การแยก LU หยุดด้วยข้อความ `zero last pivot` และเมื่อ θ_2 ใกล้ศูนย์แต่ไม่เป็นศูนย์ $\kappa_\infty(J)$ ก็โตมาก ที่ $\theta = (0.3,\ 0.01)$ เราได้ $\kappa_\infty(J) \approx 699$
- **เป้าอยู่นอกระยะ.** ถ้า $\|T\|_2 > L_1 + L_2 = 1.8$ ไม่มีคำตอบเลย และนิวตันกระโดดไปไกล จากเป้า $(2.5,\ 0)$ ค่าตกค้างในสี่รอบแรกคือ $0.965,\ 2.948,\ 4.241,\ 3.022$ หลัง 12 รอบยังเป็น 2.10 และมุมไปถึงหลายร้อยเรเดียน
- **จุดเริ่มกำหนดคำตอบ.** เมื่อเป้าอยู่ในระยะแขนมีคำตอบสองท่า (หรือท่าเดียวที่ขอบระยะ) ดูหัวข้อถัดไป จุดเริ่มต่างกันอาจลงที่ท่าต่างกัน

สองท่าที่ไปถึงจุดเดียวกัน

สูตรปิดของแขนสองท่อนทำให้เรารู้คำตอบแม่นได้โดยไม่ต้องทำนิวตัน ถ้า $\|T\|_2^2 = L_1^2 + L_2^2 + 2L_1L_2\cos\theta_2$ จะได้

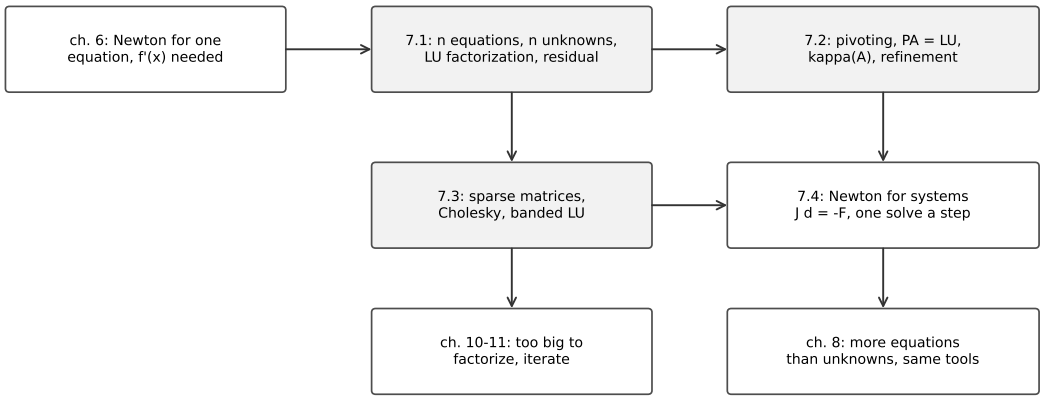
$$\cos\theta_2 = \frac{\|T\|_2^2 - L_1^2 - L_2^2}{2L_1L_2},$$

และ θ_2 มีสองค่าที่เป็นเครื่องหมายตรงข้ามกัน คือข้อศอกงอไปทางหนึ่งหรืออีกทางหนึ่ง เมื่อเป้า $T = (1.2,\ 0.9)$ เราได้ $\cos\theta_2 = 0.38125$ จึง $\theta_2 = \pm 1.179648$ และ θ_1 ตามสูตร $\theta_1 = \operatorname{atan2}(T_y, T_x) - \operatorname{atan2}(L_2\sin\theta_2,\ L_1 + L_2\cos\theta_2)$ ได้สองคำตอบ

$$\theta^a = (0.127907,\ 1.179648), \qquad \theta^b = (1.159095,\ -1.179648).$$

ทั้งสองค่าให้มือที่ $(1.200000,\ 0.900000)$ จากจุดเริ่ม $(0.2,\ 1.0)$ นิวตันลงที่ θ^a ส่วนจาก $(0.8,\ -0.3)$ ก้าวแรกพาไปไกลมากจน $\theta_2 \approx -2.41$ แล้วลงที่ θ^b ในตัวอย่างนี้ จุดเริ่มเป็นตัวกำหนดคำตอบ

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 7.12: แผนผังบทที่ 7 ทั้งบท: ลำดับของหัวข้อ 7.1 ถึง 7.4 สิ่งที่แต่ละหัวข้อให้หัวข้อถัดไป และสิ่งที่บทที่ 8 และบทที่ 10--11 ใช้ต่อ เส้นจากบทที่ 6 เข้าสู่หัวข้อ 7.1 คือวิธีนิวตันที่ขยายเป็นระบบ (สร้างจาก `numanal/ch07_arm.py`)

สรุปใจความ

- ตัวไม่ทราบหลายตัวทำให้เส้นสัมผัสกลายเป็นเมทริกซ์จาโคเบียน J และนิวตันแต่ละรอบแก้ระบบเชิงเส้น $J\,d = -F$ ด้วยเครื่องมือจากหัวข้อ 7.1 และ 7.2
- ถ้า J ไม่เอกฐานที่คำตอบและเริ่มใกล้พอ นิวตันลู่เข้าแบบกำลังสอง ในแขนนี้ค่าตกค้างลดจาก 0.083 เป็นราว 10^{-9} ในสามรอบ
- การทำซ้ำ $\mathbf{x} = G(\mathbf{x})$ ลู่เข้าแบบเชิงเส้นเมื่อ G หดตัว และวิธีนิวตัน-ราฟสันแบบดัดแปลงเป็นตัวอย่างที่อัตราเท่ากับรัศมีสเปกตรัม 0.158
- ข้อจำกัดของนิวตันคือ J เอกฐานเมื่อข้อศอกเหยียดตรง เป้าที่อยู่นอกระยะ และจุดเริ่มที่ต่างกันให้คำตอบต่างกัน
- $\kappa_\infty(J)$ จากหัวข้อ 7.2 บอกว่าการปัดเศษในแต่ละรอบจะถูกขยายได้เท่าไร

อ่านต่อ. ตำรา Iserles 2009, §7.1--7.2, pp. 123--130 อธิบายการทำซ้ำจุดตรึงในปริภูมิ $\mathbb{R}^d$ และนิวตัน-ราฟสันของระบบรวมถึงรูปแบบดัดแปลง Atkinson 1989, §2.10--2.11, pp. 103--111 อธิบายระบบไม่เชิงเส้น ทฤษฎีจุดตรึงในปริภูมิ และนิวตันสำหรับระบบอย่างละเอียดกว่าเรา ระบบในหัวข้อนี้มีสองตัวไม่ทราบ ระบบที่ใหญ่กว่านี้มากอาจมีตัวไม่ทราบหลายพันตัว และการแยก J ทุกรอบไม่คุ้ม ซึ่งเป็นเรื่องของบทที่ 10--11

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 7.4.1 (⋆) แขนสองท่อนมีความยาว $L_1 = 2$ และ $L_2 = 1$ ที่มุม $\theta_1 = 0$ และ $\theta_2 = \pi/2$

1. จงหาตำแหน่งของข้อศอกและของมือ (x, y)
2. จงหาเมทริกซ์จาโคเบียน J ด้วยมือ แล้วคำนวณ $\det J$ และเทียบกับ $L_1 L_2 \sin\theta_2$

แบบฝึกหัด 7.4.2 (⋆) ใช้แขนของแบบฝึกหัด 7.4.1 และเป้าหมาย $T = (2,\ 1.5)$ เริ่มจากท่า $\theta^{(0)} = (0,\ \pi/2)$

1. จงหา $F(\theta^{(0)})$ และ $\|F(\theta^{(0)})\|_2$

ฉบับร่าง · Preprint · 2026-10-23

2. จงแก้ $J\,d = -F$ ด้วยการกำจัดแบบเกาส์บนระบบสองสมการ แล้วหา $\theta^{(1)} = \theta^{(0)} + d$
3. จงคำนวณ $\|F(\theta^{(1)})\|_2$ และตอบว่าค่าตกค้างลดลงกี่เท่า

แบบฝึกหัด 7.4.3 (★★) ให้ $G(x, y) = \left(\frac{1}{2}\cos y,\ \frac{1}{4}\sin x\right)$ บนระนาบ $\mathbb{R}^2$ และให้ $\mathbf{x}_{k+1} = G(\mathbf{x}_k)$ เริ่มที่ $\mathbf{x}_0 = (0, 0)$ ใช้นอร์มอนันต์ $\|\cdot\|_\infty$

1. จงแสดงว่า $\|G(\mathbf{u}) - G(\mathbf{v})\|_\infty \le \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_\infty$ สำหรับทุก $\mathbf{u}, \mathbf{v}$ (ใช้ $|\cos a - \cos b| \le |a - b|$ และ $|\sin a - \sin b| \le |a - b|$)
2. จงหา $\mathbf{x}_1$, $\mathbf{x}_2$ และ $\mathbf{x}_3$ ด้วยมือ (ทศนิยมห้าตำแหน่ง)
3. จงหาจำนวนรอบ k ที่น้อยที่สุดซึ่งขอบเขตล่วงหน้า $\|\mathbf{x}_k - \mathbf{x}^*\|_\infty \le \dfrac{\lambda^k}{1-\lambda}\|\mathbf{x}_1 - \mathbf{x}_0\|_\infty$ รับประกันว่าค่าคลาดเคลื่อนสัมบูรณ์ไม่เกิน 10^{-8}
4. ที่จุดตรึง $\mathbf{x}^*$ อนุพันธ์ของ G เป็นเมทริกซ์ที่มีศูนย์บนแนวทแยง จงแสดงว่ารัศมีสเปกตรัมของอนุพันธ์นี้คือ $\sqrt{|G'_{12}G'_{21}|}$ แล้วหาค่าโดยประมาณ และตรวจว่าน้อยกว่า $\frac{1}{2}$

แบบฝึกหัด 7.4.4 (★★) แขนของเนื้อเรื่องมี $L_1 = 1$ และ $L_2 = 0.8$

1. จงใช้สูตรอนุพันธ์ย่อยของหัวข้อ 7.4 และเอกลักษณ์ $\sin(\phi_2 - \phi_1) = \sin\phi_2\cos\phi_1 - \cos\phi_2\sin\phi_1$ แสดงว่า $\det J = L_1 L_2 \sin\theta_2$
2. จงหาค่า θ_2 ทั้งหมดใน $(-\pi, \pi]$ ที่ทำให้ J ไม่มีการผกผัน
3. ที่ $\theta = (0.3,\ 0.01)$ จงหา $\det J$ และ $\kappa_\infty(J) = \|J\|_\infty\|J^{-1}\|_\infty$ (ดูหัวข้อ 7.2) แล้วอธิบายว่าการแก้ระบบเชิงเส้นในรอบนิวตันจะไวต่อข้อมูลอย่างไรที่ท่านี้
4. จงอธิบายว่าเหตุใดขั้นตอนของนิวตันที่ $\theta = (0.3,\ 0)$ ทำไม่ได้ และการแยก LU จะรายงานอะไร

แบบฝึกหัด 7.4.5 (★★) แขนของเนื้อเรื่อง ($L_1 = 1$, $L_2 = 0.8$) และเป้าหมาย $T = (1.2,\ 0.9)$

1. จงแสดงว่า $\cos\theta_2 = 0.38125$ จาก $\|T\|_2^2 = L_1^2 + L_2^2 + 2L_1L_2\cos\theta_2$
2. จงตรวจด้วยการแทนค่า (ทศนิยมหกตำแหน่ง) ว่าทั้งสองค่าต่อไปนี้เป็นคำตอบของ $F = 0$

$$\theta^a = (0.127907,\ 1.179648), \qquad \theta^b = (1.159095,\ -1.179648).$$

3. จงทำนิวตันจากจุดเริ่ม $(0.2,\ 1.0)$ และจาก $(0.8,\ -0.3)$ ด้วยโค้ดของบทนี้ แล้วบอกว่าแต่ละจุดเริ่มลงที่คำตอบใด และอธิบายด้วยเครื่องหมายของ θ_2 เริ่มต้น
4. จงตรวจด้วยเงื่อนไขระยะเอื้อมว่า $T = (2.5,\ 0)$ ไม่มีคำตอบ โดยไม่ต้องทำนิวตัน

แบบฝึกหัด 7.4.6 (★ ★ ★) โครงงานเล็ก: ผู้เล่นลากเมาส์เป็นเส้นตรงจาก $(1.2,\ 0.9)$ ไปยัง $(1.7,\ 0.4)$ โดยเราแบ่งเป็นหกจุด $T_k = (1.2 + 0.1k,\ 0.9 - 0.1k)$ เมื่อ $k = 0, \dots, 5$ ใช้แขนของเนื้อเรื่อง เริ่มจากท่า $(0.2,\ 1.0)$ แล้วทำนิวตันต่อเนื่อง โดยให้คำตอบของ T_k เป็นจุดเริ่มของ T_{k+1}

1. จงทำตารางที่มีคอลัมน์ k, จำนวนรอบนิวตันจนค่าตกค้างต่ำกว่า 10^{-12} และ $\kappa_\infty(J)$ ที่คำตอบของแต่ละจุด
2. จุดใดมี $\kappa_\infty(J)$ มากที่สุด และจำนวนรอบเพิ่มขึ้นที่จุดใด จงอธิบายโดยดูจากรูปร่างของแขนและ $\det J$

ร่างต้น · Preprint · 2026-10-23

3. ลากต่อไปยัง $(1.8,\ 0.3)$ จงตรวจด้วยเงื่อนไขระยะเอื้อมว่าเป้านี้อยู่นอกระยะ แล้วทำนิวตันด้วยโค้ดจากคำตอบของ T_5 สามสิบรอบ ตรวจว่าค่าตกค้างไม่ต่ำกว่า $\|T\|_2 - (L_1 + L_2)$ และอธิบายว่าเหตุใดขอบล่างนี้ต้องเป็นจริง

4. จงอธิบายว่าทำไม $\kappa_\infty(J)$ ที่ใหญ่ทำให้ผลของการปัดเศษในการแก้ระบบเชิงเส้นแต่ละรอบใหญ่ขึ้น และการเลือกตัวหลักช่วยอะไรในกรณีแขนสองท่อนนี้ (ดูหัวข้อ 7.2)

ตัวช่วยคิด

ใช้ `follow_path` ของโค้ดบทนี้ซึ่งทำนิวตันต่อเนื่องและคืนจำนวนรอบ ค่าตกค้าง และ κ_∞ ทุกจุด ก่อนเช็กระยะเอื้อมให้คำนวณ $\|T\|_2$ เทียบกับ $L_1 + L_2 = 1.8$ ในข้อ 2 ให้ดูว่า $\det J = L_1 L_2 \sin\theta_2$ ลดลงเมื่อ θ_2 เข้าใกล้ศูนย์อย่างไร

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 7.4.1.

1. มุมสัมบูรณ์ของท่อนแรกคือ $\varphi_1 = \theta_1 = 0$ และของท่อนที่สองคือ $\varphi_2 = \theta_1 + \theta_2 = \pi/2$ ข้อศอกอยู่ที่ $(L_1\cos\varphi_1,\ L_1\sin\varphi_1) = (2, 0)$ และมืออยู่ที่ $(2\cos 0 + \cos(\pi/2),\ 2\sin 0 + \sin(\pi/2)) = (2,\ 1)$

2. ใช้สูตรอนุพันธ์ย่อยของหัวข้อ 7.4: $\partial x/\partial\theta_1 = -L_1\sin\varphi_1 - L_2\sin\varphi_2 = -1$, $\partial x/\partial\theta_2 = -L_2\sin\varphi_2 = -1$, $\partial y/\partial\theta_1 = L_1\cos\varphi_1 + L_2\cos\varphi_2 = 2$ และ $\partial y/\partial\theta_2 = L_2\cos\varphi_2 = 0$ จึง $J = \begin{bmatrix} -1 & -1 \\ 2 & 0 \end{bmatrix}$ และ $\det J = (-1)(0) - (-1)(2) = 2$ ตรงกับ $L_1 L_2 \sin\theta_2 = 2 \cdot 1 \cdot 1 = 2$

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_arm.py` ว่าตำแหน่งมือ $(2, 1)$ เมทริกซ์ J และ $\det J = 2$

เฉลย 7.4.2.

1. $F(\theta^{(0)}) = (2, 1) - (2,\ 1.5) = (0,\ -0.5)$ และ $\|F(\theta^{(0)})\|_2 = 0.5$

2. ระบบ $J\,d = -F = (0,\ 0.5)$ คือ $-d_1 - d_2 = 0$ และ $2d_1 = 0.5$ ใช้การกำจัดแบบเกาส์โดยไม่สลับแถวในข้อนี้ ตัวหลักแรกคือ -1 และตัวคูณของแถวที่สองคือ $m = 2/(-1) = -2$ ได้แถวใหม่ $(0,\ -2 \mid 0.5)$ จึง $d_2 = -0.25$ และ $d_1 = 0.25$ ดังนั้น $\theta^{(1)} = (0.25,\ \pi/2 - 0.25) \approx (0.25,\ 1.32080)$

3. แทนค่า $x = 2\cos(0.25) + \cos(\pi/2) = 1.937825$ และ $y = 2\sin(0.25) + \sin(\pi/2) = 1.494808$ จึง $F(\theta^{(1)}) = (-0.062175,\ -0.005192)$ และ $\|F(\theta^{(1)})\|_2 \approx 0.06239$ ค่าตกค้างลดลงจาก 0.5 เป็น 0.0624 ราว 8 เท่าในรอบเดียว

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_arm.py` ว่า $d = (0.25, -0.25)$ และ $\|F(\theta^{(1)})\|_2 \approx 0.06239$

ร่างพิมพ์ Preprint · 2026-10-23

เฉลย 7.4.3.

1. สำหรับทุก a, b เรามี $|\cos a - \cos b| \le |a-b|$ และ $|\sin a - \sin b| \le |a-b|$ เพราะอนุพันธ์ของ cos และ sin มีค่าสัมบูรณ์ไม่เกิน 1 ดังนั้น

$$|G_1(\mathbf{u}) - G_1(\mathbf{v})| = \tfrac{1}{2}|\cos u_2 - \cos v_2| \le \tfrac{1}{2}\|\mathbf{u}-\mathbf{v}\|_\infty,$$
$$|G_2(\mathbf{u}) - G_2(\mathbf{v})| = \tfrac{1}{4}|\sin u_1 - \sin v_1| \le \tfrac{1}{4}\|\mathbf{u}-\mathbf{v}\|_\infty.$$

   นอร์มอนันต์คือค่าสัมบูรณ์มากที่สุดของสององค์ประกอบ จึง $\|G(\mathbf{u}) - G(\mathbf{v})\|_\infty \le \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_\infty$ ดังนั้น $\lambda = \frac{1}{2}$

2. ใช้ G ทีละรอบ ได้

$$\mathbf{x}_1 = G(0,0) = (0.5,\ 0),$$
$$\mathbf{x}_2 = G(\mathbf{x}_1) = (\tfrac{1}{2}\cos 0,\ \tfrac{1}{4}\sin 0.5) = (0.5,\ 0.119856),$$
$$\mathbf{x}_3 = G(\mathbf{x}_2) = (\tfrac{1}{2}\cos 0.119856,\ \tfrac{1}{4}\sin 0.5) = (0.496413,\ 0.119856).$$

3. ขอบเขตล่วงหน้าคือ $\|\mathbf{x}_k - \mathbf{x}^*\|_\infty \le \dfrac{\lambda^k}{1-\lambda}\|\mathbf{x}_1 - \mathbf{x}_0\|_\infty = \dfrac{(1/2)^k}{1/2}\cdot\tfrac{1}{2} = (\tfrac{1}{2})^k$ ต้องการให้ $(\tfrac{1}{2})^k \le 10^{-8}$ นั่นคือ $k \ge \log_2 10^8 \approx 26.58$ จึงได้ $k = 27$ (เพราะ $(\tfrac{1}{2})^{27} \approx 7.45\times10^{-9}$ แต่ $(\tfrac{1}{2})^{26} \approx 1.49\times10^{-8}$)

4. จุดตรึง $\mathbf{x}^* \approx (0.496459,\ 0.119079)$ อนุพันธ์ของ G คือ $\begin{bmatrix} 0 & -\frac{1}{2}\sin x_2 \\ \frac{1}{4}\cos x_1 & 0 \end{bmatrix}$ ซึ่งมีค่าลักษณะเฉพาะ μ ที่ $\mu^2 = G'_{12}G'_{21}$ ดังนั้น $\rho = \sqrt{|G'_{12}G'_{21}|}$ เมื่อแทนค่า $G'_{12} \approx -0.059399$ และ $G'_{21} \approx 0.219819$ ได้ $\rho \approx 0.1143$ ซึ่งน้อยกว่า $\frac{1}{2}$ ขอบเขตในข้อ 3 จึงมองโลกในแง่ร้ายเกินจริง ค่าคลาดเคลื่อนสัมบูรณ์จริงหลังรอบที่ 27 น้อยกว่าที่ขอบเขตให้ไว้มาก

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_arm.py` ว่า $\lambda \le \frac{1}{2}$ จากการสุ่มคู่จุด ว่า $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ ตรงกับข้อ 2 ว่า $\rho \approx 0.1142$ และว่าค่าคลาดเคลื่อนสัมบูรณ์ที่ $k = 27$ น้อยกว่า 10^{-8}

เฉลย 7.4.4.

1. เขียน $s_i = \sin\varphi_i$ และ $c_i = \cos\varphi_i$ เมตริกซ์คือ $J = \begin{bmatrix} -L_1 s_1 - L_2 s_2 & -L_2 s_2 \\ L_1 c_1 + L_2 c_2 & L_2 c_2 \end{bmatrix}$ จึง

$$\begin{aligned}\det J &= (-L_1 s_1 - L_2 s_2)L_2 c_2 + L_2 s_2 (L_1 c_1 + L_2 c_2) \\ &= L_1 L_2 (s_2 c_1 - s_1 c_2) = L_1 L_2 \sin(\varphi_2 - \varphi_1) = L_1 L_2 \sin\theta_2.\end{aligned}$$

   ที่สุดท้ายใช้ $\sin(A-B) = \sin A\cos B - \cos A \sin B$ กับ $A = \varphi_2$ และ $B = \varphi_1$

2. $\sin\theta_2 = 0$ เมื่อ $\theta_2 = k\pi$ ในช่วง $(-\pi, \pi]$ ได้สองค่าคือ $\theta_2 = 0$ (ข้อศอกเหยียดตรง) และ $\theta_2 = \pi$ (ท่อนล่างพับทับท่อนบน)

3. ที่ $\theta = (0.3,\ 0.01)$ และ $L = (1,\ 0.8)$ จะได้ $\det J = 0.8\sin 0.01 \approx 0.0080$ และ

$$J \approx \begin{bmatrix} -0.5396 & -0.2440 \\ 1.7172 & 0.7619 \end{bmatrix},$$
$$J^{-1} = \frac{1}{0.0080}\begin{bmatrix} 0.7619 & 0.2440 \\ -1.7172 & -0.5396 \end{bmatrix} \approx \begin{bmatrix} 95.23 & 30.51 \\ -214.65 & -67.45 \end{bmatrix}.$$

ฉบับร่าง · Preprint · 2026-10-23

ดังนั้น $\|J\|_\infty = \max(0.7836,\ 2.4791) = 2.4791$ และ $\|J^{-1}\|_\infty = \max(125.74,\ 282.10) = 282.10$ จึง $\kappa_\infty(J) \approx 2.4791 \times 282.10 \approx 699.3$ ตามทฤษฎีบท κ ในหัวข้อ 7.2 ค่าคลาดเคลื่อนสัมพัทธ์ของ d อาจโตได้ถึงราว 700 เท่าของค่าคลาดเคลื่อนสัมพัทธ์ของ F ที่ท่านี้ จึงการปัดเศษเล็กน้อยในแต่ละรอบอาจทำให้ก้าวของนิวตันผิดมาก

4. ที่ $\theta_2 = 0$ ค่า $\det J = 0$ จึงไม่มี J^{-1} เมื่อแยกด้วยการเลือกตัวหลักตามหลัก ตัวหลักสุดท้ายจะเป็นศูนย์ และโค้ดจะหยุดด้วยข้อผิดพลาด `zero last pivot` นิวตันจึงคำนวณก้าวถัดไปไม่ได้ ที่ท่านี้เราต้องเปลี่ยนจุดเริ่มหรือยอมรับว่าคำตอบอาจอยู่คนละท่า

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_arm.py` ว่า $\det J = 0.0080$ ที่ $\theta_2 = 0.01$ ว่า $\kappa_\infty(J) \approx 699.35$ และว่า `newton_step` ที่ $\theta_2 = 0$ ยกข้อผิดพลาด `ValueError`

เฉลย 7.4.5.

1. $\|T\|_2^2 = 1.2^2 + 0.9^2 = 2.25$ และ $L_1^2 + L_2^2 + 2L_1L_2\cos\theta_2 = 1 + 0.64 + 1.6\cos\theta_2$ จึง $\cos\theta_2 = (2.25 - 1.64)/1.6 = 0.61/1.6 = 0.38125$

2. สำหรับ $\theta^a = (0.127907,\ 1.179648)$ เรามี $\cos\theta_1 = 0.991831$, $\sin\theta_1 = 0.127559$, $\varphi_2 = 1.307555$ จึง $\cos\varphi_2 = 0.260212$, $\sin\varphi_2 = 0.965552$ และ $x = 0.991831 + 0.8 \cdot 0.260212 = 1.200000$, $y = 0.127559 + 0.8 \cdot 0.965552 = 0.900000$ สำหรับ $\theta^b = (1.159095,\ -1.179648)$ เรามี $\varphi_2 = -0.020553$ จึง $x = 0.400169 + 0.8 \cdot 0.999789 = 1.200000$ และ $y = 0.916441 + 0.8 \cdot (-0.020552) = 0.900000$ ทั้งสองค่าตรงกับ T ทศนิยมหกตำแหน่ง

3. จาก $(0.2,\ 1.0)$ นิวตันลงที่ θ^a ซึ่งมีมุมข้อศอก $\theta_2 > 0$ จาก $(0.8,\ -0.3)$ ก้าวแรกพาไปถึง $\theta_2 \approx -2.41$ ซึ่งไกลจากคำตอบมาก แล้วกลับเข้ามาลงที่ θ^b ความแตกต่างมาจากจุดเริ่มเท่านั้น ท่าที่เริ่มต่างกันอยู่ในแอ่งดึงดูดของคำตอบคนละจุด

4. $\|T\|_2 = 2.5 > L_1 + L_2 = 1.8$ มือของแขนอยู่ห่างจากไหล่ได้ไม่เกิน 1.8 จึงไม่มีมุมใดที่ทำให้ $F = 0$ และ `two_bone_solutions` คืนรายการว่าง

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_arm.py` ว่าค่าตกค้างของทั้งสองรากน้อยกว่า 10^{-12} ว่านิวตันจากสองจุดเริ่มลงที่รากตามที่ระบุ และว่า $T = (2.5, 0)$ ไม่มีคำตอบ

เฉลย 7.4.6.

1. ผลจากโค้ดของบท (ค่าตกค้างต่ำกว่า 10^{-12} ทุกจุด):

k	T_k	รอบนิวตัน	$\kappa_\infty(J)$
0	$(1.2,\ 0.9)$	4	4.749
1	$(1.3,\ 0.8)$	4	4.660
2	$(1.4,\ 0.7)$	4	5.479
3	$(1.5,\ 0.6)$	4	6.603
4	$(1.6,\ 0.5)$	4	8.468
5	$(1.7,\ 0.4)$	5	13.330

2. $\kappa_\infty(J)$ มากที่สุดที่ $T_5 = (1.7,\ 0.4)$ และจำนวนรอบเพิ่มขึ้นจาก 4 เป็น 5 ที่จุดนี้ เมื่อเป้าหมายเคลื่อนออกไปจนเข้าใกล้ขอบระยะเอื้อม $|T_5|_2 = 1.746$ มุม θ_2 ลดจาก 1.180 เป็น 0.492 และ $\det J = 0.8\sin\theta_2$ ลดจาก 0.740 เป็น 0.378 ทำให้ κ โตขึ้น ส่วนที่ว่ารอบเพิ่มขึ้นเพราะเหตุใดนั้นเป็นการสังเกตจากตารางนี้ ไม่ใช่การพิสูจน์

ฉบับร่าง Preprint · 2026-10-23

3. $\|T\|_2 = 1.8248$ สำหรับ $(1.8,\ 0.3)$ และ 1.9105 สำหรับ $(1.9,\ 0.2)$ ทั้งสองมากกว่า 1.8 จึงอยู่นอกระยะ จากคำตอบของ T_5 เราทำนิวตันสามสิบรอบสำหรับ $(1.8,\ 0.3)$ ได้ค่าตกค้างที่รอบสุดท้ายประมาณ 0.0536 และต่ำสุดประมาณ 0.0250 ขอบล่างคือ $\|T\|_2 - (L_1 + L_2) = 1.8248 - 1.8 = 0.0248$ เป็นจริงเพราะ $\|\text{มือ} - T\|_2 \geq \|T\|_2 - \|\text{มือ}\|_2 \geq \|T\|_2 - (L_1 + L_2)$ เพราะมืออยู่ห่างจากไหล่ไม่เกิน $L_1 + L_2$ ค่าตกค้างจึงไม่มีทางเป็นศูนย์

4. ตามทฤษฎีบท κ ในหัวข้อ 7.2 ค่าคลาดเคลื่อนสัมพัทธ์ของ d โตได้ราว κ เท่าของค่าคลาดเคลื่อนสัมพัทธ์ของ F เมื่อ κ ใหญ่ เล็กน้อยในการปัดเศษของ F ก็ขยายเป็นก้าวที่ผิดมาก สำหรับแขนสองท่อนที่ 2×2 การเลือกตัวหลักตามหลักทำให้แถวที่มีค่าสัมบูรณ์มากอยู่บนตัวหลัก ตัวคูณจึงไม่เกิน 1 ซึ่งป้องกันการขยายที่ไม่จำเป็นในการกำจัด แต่ไม่ได้ลด κ ของ J เอง เพราะ κ เป็นคุณสมบัติของเมทริกซ์

ตรวจด้วยโค้ด: ทดสอบใน `test_ch07_arm.py` ว่าตาราง k, รอบนิวตัน และ κ ตรงกับข้อ 1 ว่า κ สูงสุดที่ T_5 ว่า $(1.8, 0.3)$ และ $(1.9, 0.2)$ อยู่นอกระยะ และว่าค่าตกค้างต่ำสุดไม่ต่ำกว่า 0.0248

สรุปบทที่ 7 ในหน้าเดียว

ตารางต่อไปนี้เทียบวิธีทั้งหมดของบทนี้ ตัวอักษร n คือขนาดของระบบ m คือจำนวนสมการของแขน s คือความกว้างแถบ และ d คือจำนวนตัวไม่ทราบ ค่าใช้จ่ายเป็นการคูณและหารโดยประมาณ

วิธี	ต้องการอะไร	ค่าใช้จ่าย	หมายเหตุ
การกำจัดแบบเกาส์ (7.1)	ไม่เอกฐาน	แยก $\frac{2}{3}n^3$ แก้ซ้ำ $\approx n^2$	ใช้ตัวประกอบเดิมได้หลายชุด
การทำซ้ำแบบเกาส์--ไซเดล (7.1)	ลู่เข้า	$\approx n^2$ ต่อรอบ	ลู่เข้าเชิงเส้น
การปรับปรุงคำตอบซ้ำ (7.2)	ตัวประกอบเดิม	$\approx n^2$ ต่อรอบ	แก้ค่าคลาดเคลื่อนจากการปัดเศษ
โชเลสกี (7.3)	สมมาตรบวกแน่นอน	$\frac{1}{3}n^3$	ครึ่งหนึ่งของ LU
แถบ (7.3)	ความกว้างแถบ s	$\approx s^2 d$	ศูนย์นอกแถบรอด
CSR และการคูณเมทริกซ์เบาบาง (7.3)	เก็บเฉพาะค่าไม่เป็นศูนย์	ราวจำนวนค่าไม่เป็นศูนย์	ประหยัดความจำ
นิวตันสำหรับระบบ (7.4)	J ไม่เอกฐานที่คำตอบ	แยก LU ทุกรอบ $\frac{2}{3}m^3$	ลู่เข้าแบบกำลังสอง
นิวตัน-ราฟสันแบบดัดแปลง (7.4)	J_0 ใกล้พอ	แยกครั้งเดียว แล้ว $\approx m^2$	ลู่เข้าเชิงเส้น ด้วย ρ

สูตรที่ควรจำ

- การกำจัดแบบเกาส์: การคูณและหารราว $n(n^2-1)/3$ และการบวกและลบราว $n(n-1)(2n-1)/6$ (หัวข้อ 7.1)
- ค่าคอนดิชันนัมเบอร์: $\kappa(A) = \|A\|\,\|A^{-1}\|$ และ $\|\delta x\|/\|x\| \leq \kappa\,\|\delta b\|/\|b\|$ (หัวข้อ 7.2)
- โชเลสกี: $A = LL^T$ โดย $\ell_{ii} > 0$ (หัวข้อ 7.3)
- นิวตันสำหรับระบบ: $J(\theta^{(k)})\,d^{(k)} = -F(\theta^{(k)})$ และ $\theta^{(k+1)} = \theta^{(k)} + d^{(k)}$ (หัวข้อ 7.4)
- การทำซ้ำจุดตรึง: $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \lambda \|\mathbf{x}_k - \mathbf{x}^*\|$ เมื่อ $\lambda < 1$ (หัวข้อ 7.4)
- แขนสองท่อน: $\det J = L_1 L_2 \sin\theta_2$ (หัวข้อ 7.4)

อ่านต่อ. บทที่ 8 ข้อมูลมากกว่าตัวไม่รู้ เริ่มจากระบบที่สมการมากกว่าตัวไม่รู้ เช่น ตำแหน่งจากดาวเทียมหลายดวงซึ่งไม่มีคำตอบที่ตรงทุกสมการ เครื่องมือของบทนี้ยังใช้ได้ แต่ต้องตั้งปัญหาใหม่ก่อน โครงงานเล็กของบทนี้คือแบบฝึกหัด 7.4.6 ซึ่งใช้โค้ดของหัวข้อ 7.4 และแนวคิดของ κ จากหัวข้อ 7.2

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 8

ข้อมูลมากกว่าตัวไม่รู้

สถานการณ์สมมติ: จุดสีน้ำเงินบนแผนที่

ลองนึกถึงจุดสีน้ำเงินบนแผนที่ในโทรศัพท์ ที่บอกว่าเราอยู่ตรงไหน จุดนั้นขยับตามเราเมื่อเดิน และดูเหมือนรู้ตำแหน่งได้ทันที แต่เบื้องหลังมีตัวไม่ทราบอยู่สี่ตัว คือพิกัดสามตัวของตำแหน่ง กับค่าเวลาที่คลาดเคลื่อน (เขียนเป็นระยะทางได้ โดยคูณด้วยความเร็วแสง)

ถ้าเราวัดระยะถึงจุดอ้างอิงได้สี่จุด เราได้สี่สมการ และแก้ตัวไม่ทราบทั้งสี่ได้ตรง ๆ แต่ถ้าวัดได้แปดจุด เราได้แปดสมการสำหรับตัวไม่ทราบสี่ตัว และโดยทั่วไปไม่มีตำแหน่งใดที่ทำให้ทุกสมการเป็นจริงพอดี เพราะการวัดมีความคลาดเคลื่อนเล็กน้อยทุกครั้ง คำถามของบทนี้จึงเป็นว่า เมื่อสมการมากกว่าตัวไม่ทราบ เราจะเลือกคำตอบอย่างไร และคำนวณอย่างไรให้เชื่อถือได้

(แบบจำลองนี้เป็นของเราเอง ไม่ได้บอกว่าโทรศัพท์หรือระบบนำทางใดทำงานอย่างไรในรายละเอียด ตัวเลขทั้งหมดในบทนี้มาจากโค้ดของเรา)

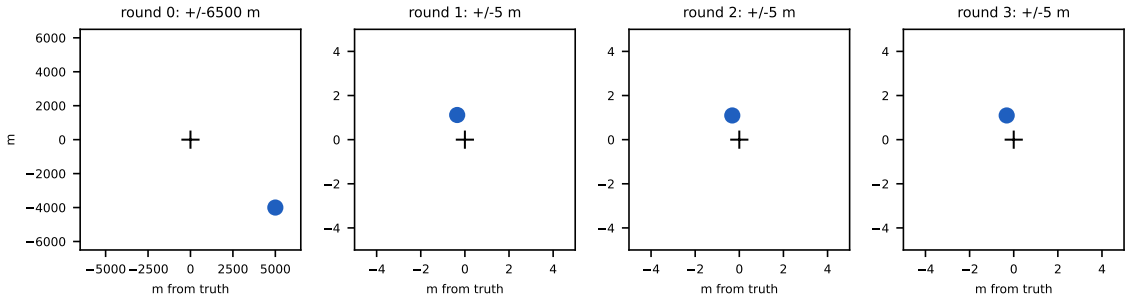
ภาพรวมบท

บทที่ 7 จบด้วยคำถามว่า เมื่อเราแก้ระบบไม่ได้ตรงตัว จะทำอย่างไร บทนี้เริ่มจากจุดสีน้ำเงินบนแผนที่ ที่มีสมการมากกว่าตัวไม่ทราบ แล้วจะเจออีกด้านหนึ่งคือภาพที่เก็บไว้ได้ด้วยรูปแบบเล็ก ๆ ของข้อมูลเดิม ตามลำดับนี้

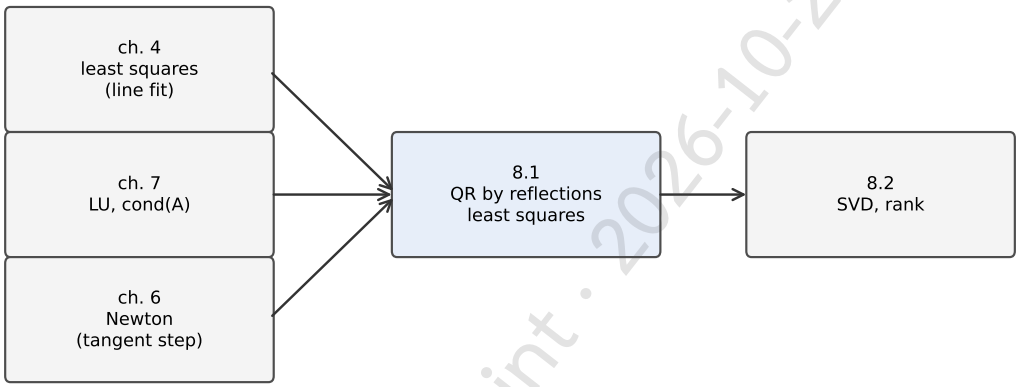
- หัวข้อ 8.1: คำตอบที่พอดีที่สุดของระบบที่มีสมการเกิน เราเลือกคำตอบที่ทำให้ผลบวกกำลังสองของค่าตกค้างน้อยที่สุด แล้วแก้ด้วยการแยก QR ด้วยการสะท้อนของเฮาส์โฮลเดอร์ ซึ่งเสถียรกว่าสมการปรกติ
- หัวข้อ 8.2: ข้อมูลที่ซ้ำซ้อนกัน ภาพขาวดำที่มีแถวและคอลัมน์มากหลายแถวอาจบรรจุได้ด้วยรูปแบบไม่กี่รูปแบบที่ใหญ่ที่สุด เราจะดูว่าทำอย่างไรด้วยการแยกค่าเอกฐาน

ภาพแรกแสดงจุดสีน้ำเงินที่ขยับเข้าหาตำแหน่งจริงในแบบจำลองสองมิติ ภาพที่สองแสดงลำดับของหัวข้อในบทนี้ และสิ่งที่บทนี้รับมาจากบทที่ 4, 6 และ 7

สัมมนา. Preprint. 2026-10-23



รูปที่ 8.1: จุดสีน้ำเงินในแบบจำลองสองมิติ การทำซ้ำรอบที่ 0 ถึง 3 เครื่องหมาย + คือตำแหน่งจริง ดาวเทียมสมมติแปดดวงอยู่นอกกรอบ (ไม่ได้วาด) กรอบแต่ละกรอบมีมาตราของตัวเอง สร้างจาก `numanal/ch08.py`



ฉบับร่าง · Preprint · 2026-10-23

รูปที่ 8.2: ลำดับของหัวข้อในบทที่ 8 และสิ่งที่บทนี้รับมาจากบทที่ 4, 6 และ 7 (สร้างจาก `numanal/ch08.py`)

8.1 คำตอบที่พอดีที่สุดของระบบที่มีสมการเกิน

สถานการณ์สมมติ: แว่นตาเสมือนบนใบหน้า

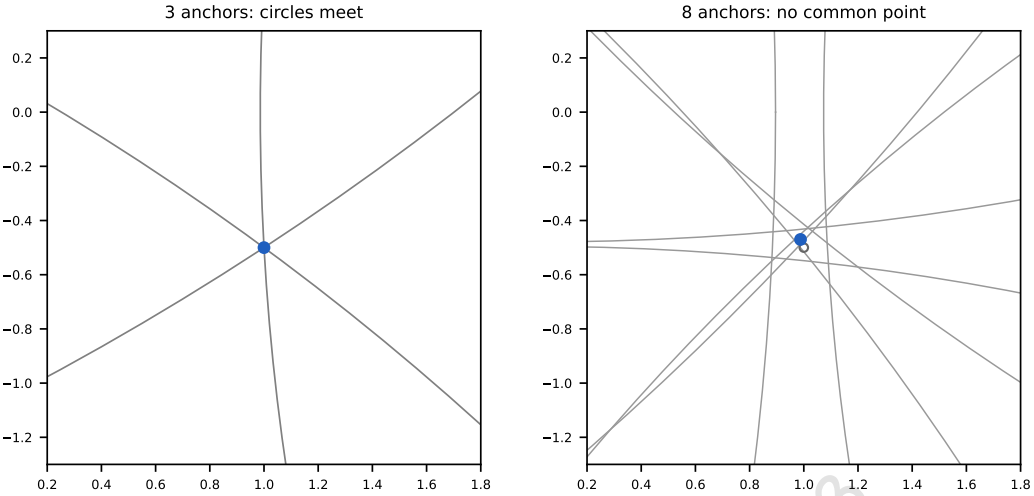
ลองนึกถึงแว่นตาเสมือนที่วาดซ้อนบนใบหน้าในภาพจากกล้อง เมื่อหน้าหมุนหรือเลื่อน แว่นต้องหมุนและเลื่อนตามให้พอดี ระบบจึงต้องรู้ว่าใบหน้าอยู่ในท่าใด จากจุดสำคัญบนภาพ เช่น ตาซ้าย ตาขวา และปาก ซึ่งมีมากกว่าตัวไม่ทราบที่ต้องหา

เอกสารของไลบรารี OpenCV บอกว่าวิธีแบบวนซ้ำของฟังก์ชัน `solvePnP` ปรับท่าทางให้ดีขึ้นด้วยการทำให้น้อยที่สุดแบบเลเวนเบิร์ก–มาร์ควอร์ต OpenCV contributors 2026 (เราพูดถึงเอกสารของไลบรารีนี้เท่านั้น ไม่ได้บอกว่าแอปหรือเกมใดใช้มัน) ในรุ่นเล็กของเรา จุดสมมติหกจุดถูกหมุนและเลื่อนในระนาบ แล้วเราวัดตำแหน่งที่มีสัญญาณรบกวนเล็กน้อย

คำถามคือ เมื่อจุดที่วัดได้มากกว่าตัวไม่ทราบ เราจะเลือกท่าทางที่เข้ากับทุกจุดมากที่สุดได้อย่างไร

ภาพรวมก่อนอ่าน

ความคิดของหัวข้อนี้คือ เมื่อสมการมีมากกว่าตัวไม่ทราบ เราไม่หาจุดที่ทำให้ทุกสมการเป็นจริง (มักไม่มี) แต่หาจุดที่ผลบวกกำลังสองของค่าตกค้างน้อยที่สุด ถ้าวาดวงกลมรอบจุดอ้างอิงแต่ละจุดด้วยระยะที่วัดได้ วงกลมเหล่านั้นไม่ตัดกันที่



รูปที่ 8.3: ซ้าย: จุดอ้างอิงสามจุดและระยะที่วัดได้ไม่มีความคลาดเคลื่อน วงกลมทั้งสามตัดกันที่ตำแหน่งจริงพอดี ขวา: จุดอ้างอิงแปดจุดและระยะที่มีความคลาดเคลื่อนเล็กน้อย วงกลมไม่ตัดกันที่จุดเดียว (สร้างจาก `numanal/ch08.py`)

จุดเดียว และเราเลือกจุดที่ทำให้ช่องว่างระหว่างวงกลมกับจุดนั้นรวมกันน้อยที่สุด เราจะเรียกปัญหานี้ว่าปัญหากำลังสองน้อยสุด และเครื่องมือที่แก้มันได้อย่างเสถียรคือการแยก QR

สมการที่ขัดกันเล็กน้อย

สมมติว่าเรามีระยะ b_i จากจุด s_i ถึงจุดที่เราหา x ถ้าระยะนั้นถูกต้องทุกจุด สมการ $\|s_i - x\| = b_i$ ทุกสมการจะเป็นจริงพร้อมกัน แต่ระยะจริงมีความคลาดเคลื่อน ทำให้ไม่มี x ที่ทำได้ เมื่อเรามีสมการมากกว่าตัวไม่ทราบ ระบบเชิงเส้น $A\mathbf{x} = \mathbf{b}$ จึงมักไม่มีคำตอบเลย ซึ่งเราเรียกว่า **ระบบสมการเกิน** (overdetermined system) เมื่อ A มี m แถว (สมการ) และ n คอลัมน์ (ตัวไม่ทราบ) กับ $m > n$

เราจึงเปลี่ยนคำถาม จากการหาคำตอบที่ทำให้ทุกสมการเป็นจริง เป็นการหาคำตอบที่ทำให้ผลบวกกำลังสองของค่าตกค้างน้อยที่สุด สมการ i มี **ค่าตกค้าง** (residual) $r_i = b_i - (A\mathbf{x})_i$ และเราวัดค่าตกค้างทั้งหมดด้วยผลบวกกำลังสอง

นิยาม 8.1.1: ปัญหากำลังสองน้อยสุด

ให้ A เป็นเมทริกซ์จริงขนาด $m \times n$ เมื่อ $m \geq n$ และ $\mathbf{b} \in \mathbb{R}^m$ เราเรียกปัญหา

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - A\mathbf{x}\|_2$$

ว่า **ปัญหากำลังสองน้อยสุด** (least squares problem) คำตอบ $\mathbf{x}^*$ ที่ทำให้ค่านี้น้อยที่สุดเรียกว่า **คำตอบกำลังสองน้อยสุด** (least squares solution) ถ้า A มีลำดับชั้นเต็มในแนวคอลัมน์ (คอลัมน์เป็นอิสระเชิงเส้นทั้ง n คอลัมน์) คำตอบนั้นมีเพียงคำตอบเดียว

อ่านแบบเข้าใจง่าย. ลองนึกภาพว่าทุกสมการเป็นสายยางที่ดึงตัวไม่ทราบไปทางของมัน แล้วเราหาจุดที่แรงดึงรวมกำลังสองน้อยที่สุด ถ้าสมการทั้งหมดเห็นด้วยกัน จุดนั้นทำให้ทุกสมการเป็นจริงพอดี ถ้าไม่เห็นด้วย จุดที่เราได้จะเป็นจุดกลาง ๆ ที่มีค่าตกค้างกระจายกันไป เงื่อนไขอันอาจแย่ที่สุดสำหรับปัญหานี้คือคอลัมน์ของ A ซ้ำกันเกือบหมด เพราะจุดกลางจะไม่มีทางเดียวที่ชัดเจน

ฉบับร่าง · Preprint · 2026-10-23

สมการปรกติ: วิธีแรกที่คิดได้

ถ้าเลือก $\mathbf{x}$ ที่ดีที่สุดแล้ว การเปลี่ยน $\mathbf{x}$ เล็กน้อยไม่ควรทำให้ผลบวกกำลังสองลดลงได้อีก เงื่อนไขนี้คือค่าตกค้างต้องตั้งฉากกับคอลัมน์ของ A และเขียนเป็นสมการได้ตรงไปตรงมา เราเรียกสมการนี้ว่า **สมการปรกติ** (normal equations)

ทฤษฎีบท 8.1.1: สมการปรกติ

ให้ $A \in \mathbb{R}^{m\times n}$ และ $\mathbf{b} \in \mathbb{R}^m$ เวกเตอร์ $\mathbf{x}^*$ เป็นคำตอบกำลังสองน้อยสุดของ $A\mathbf{x} = \mathbf{b}$ ก็ต่อเมื่อ

$$A^T A\,\mathbf{x}^* = A^T\mathbf{b}.$$

ถ้า A มีลำดับชั้นเต็มในแนวคอลัมน์ A^TA ไม่เอกฐาน และคำตอบมีเพียงคำตอบเดียวคือ $\mathbf{x}^* = (A^TA)^{-1}A^T\mathbf{b}$

พิสูจน์. **แนวคิด.** เราเขียนความต่างของผลบวกกำลังสองของค่าตกค้างสองค่าให้เห็นว่าอยู่ในรูปของ $A(\mathbf{y}-\mathbf{x})$ แล้วใช้ว่าความต่างนี้ไม่ติดลบเสมอเมื่อเงื่อนไขเป็นจริง

1. **กระจายผลต่าง.** ให้ $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ และ $\mathbf{e} = \mathbf{y} - \mathbf{x}$ ดังนั้น $\mathbf{b} - A\mathbf{y} = \mathbf{r} - A\mathbf{e}$ แล้ว
$$\|\mathbf{b} - A\mathbf{y}\|_2^2 = \|\mathbf{r}\|_2^2 - 2\,\mathbf{e}^TA^T\mathbf{r} + \|A\mathbf{e}\|_2^2.$$
2. **ถ้าสมการปรกติเป็นจริง.** สมมติว่า $A^TA\mathbf{x} = A^T\mathbf{b}$ จะได้ $A^T\mathbf{r} = 0$ ทำให้ $\|\mathbf{b} - A\mathbf{y}\|_2^2 - \|\mathbf{r}\|_2^2 = \|A\mathbf{e}\|_2^2 \geq 0$ สำหรับทุก $\mathbf{y}$ ดังนั้น $\mathbf{x}$ เป็นคำตอบกำลังสองน้อยสุด
3. **ถ้า $\mathbf{x}$ เป็นคำตอบกำลังสองน้อยสุด.** ถ้า $A^T\mathbf{r} \neq 0$ เลือก $\mathbf{e} = tA^T\mathbf{r}$ โดยที่ $t > 0$ เล็กพอ จะได้ $-2t\|A^T\mathbf{r}\|_2^2 + t^2\|AA^T\mathbf{r}\|_2^2 < 0$ ซึ่งขัดกับความเป็นคำตอบ จึงต้องมี $A^T\mathbf{r} = 0$ นั่นคือสมการปรกติ
4. **ความไม่เอกฐาน.** ถ้า A มีลำดับชั้นเต็ม $A\mathbf{z} = 0$ ได้เฉพาะเมื่อ $\mathbf{z} = 0$ ดังนั้น $\mathbf{z}^TA^TA\mathbf{z} = \|A\mathbf{z}\|_2^2 > 0$ สำหรับ $\mathbf{z} \neq 0$ จึง A^TA ไม่เอกฐาน

สรุป: สมการปรกติเป็นเงื่อนไขที่ตรงกับความหมายของการทำให้ผลบวกกำลังสองของค่าตกค้างน้อยที่สุด เราเขียนบทพิสูจน์นี้เอง □

อ่านแบบเข้าใจง่าย. ทฤษฎีบทบอกว่า เราไม่ต้องค้นหาคำตอบจากกราฟทั้งหมด เพียงแก้ระบบสมการ $A^TA\mathbf{x} = A^T\mathbf{b}$ ซึ่งมี n สมการและ n ตัวไม่ทราบ แม้ระบบเดิมมี m สมการ ระบบใหม่นี้เป็นระบบเล็กและสมมาตร จึงแก้ด้วยการแยกโชเลสกีได้

ลองเอง

ใช้จุดสามจุด $(0,1)$, $(1,2)$, $(2,2)$ ให้ $A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$ และ $\mathbf{b} = (1,2,2)^T$ จงคำนวณ A^TA และ $A^T\mathbf{b}$ แล้วแก้ระบบ 2×2 ที่ได้ ผลที่ควรได้คือ $A^TA = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}$, $A^T\mathbf{b} = (5,6)^T$ และคำตอบ $x_1 = \frac{7}{6}$, $x_2 = \frac{1}{2}$ จากนั้นตรวจว่าค่าตกค้างคือ $-\frac{1}{6},\ \frac{1}{3},\ -\frac{1}{6}$ และรวมกันเป็นศูนย์

ฉบับร่าง · Preprint · 2026-10-23

$$A = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}, \qquad \mathbf{b} = A\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \varepsilon = 10^{-9}.$$

รูปที่ 8.4: ระบบที่คำตอบที่แท้จริงคือ $(1, 1)$ เมื่อ ε เล็กมาก คอลัมน์ทั้งสองของ A เกือบเหมือนกัน

ข้อควรระวัง: สมการปรกติทำให้ความไวต่อความคลาดเคลื่อนเป็นกำลังสอง

เมื่อ A มีลำดับชั้นเต็ม $\operatorname{cond}(A^T A) = \operatorname{cond}(A)^2$ ตำรา Atkinson 1989, §9.7, pp. 638--639 แสดงว่าความไวของ $\mathbf{x}^*$ ต่อความคลาดเคลื่อนในข้อมูลขึ้นกับ $\operatorname{cond}(A)^2$ ไม่ใช่ $\operatorname{cond}(A)$ เท่านั้น ถ้า $\operatorname{cond}(A)$ เป็น 10^4 สมการปรกติก็ต้องรับผิดชอบต่อ 10^8 ในการแก้ จึงอาจเสียหลักไปมาก ตัวอย่างต่อไปแสดงว่าอาจแย่จนระบบปรกติเอกฐานในเลขคอมพิวเตอร์

สำหรับ $\varepsilon = 10^{-9}$ เมทริกซ์ $A^T A$ เป็น $\begin{pmatrix} 1+\varepsilon^2 & 1 \\ 1 & 1+\varepsilon^2 \end{pmatrix}$ ซึ่งในเลขทศนิยมของคอมพิวเตอร์ $\varepsilon^2 = 10^{-18}$ ถูกปัดทิ้งเมื่อบวกกับ 1 ทำให้กลายเป็น $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ ที่เอกฐาน การแยกโชเลสกีจึงหยุดเพราะตัวหลักไม่เป็นบวก ในทางกลับกัน การแยก QR (ที่เราจะเรียนต่อไป) ให้ $(1, 1)$ โดยความคลาดเคลื่อนไม่เกิน 10^{-14} ขณะที่ $\operatorname{cond}(A) \approx \sqrt{2} \times 10^9$ (ตัวเลขจากโค้ดของเรา)

การสะท้อนที่ทำให้เวกเตอร์เป็นแกน

ก่อนจะแยก QR เราต้องมีเครื่องมือที่รักษาความยาวของเวกเตอร์ เมทริกซ์ที่เราต้องการคือ **เมทริกซ์เชิงตั้งฉาก** (orthogonal matrix)

นิยาม 8.1.2: เมทริกซ์เชิงตั้งฉาก

ให้ Q เป็นเมทริกซ์จริงขนาด $n \times n$ เราเรียก Q ว่า**เมทริกซ์เชิงตั้งฉาก** (orthogonal matrix) ถ้าผลคูณของทรานสโพสกับตัวมันเองได้เอกลักษณ์ คือ $Q^T Q = I$

อ่านแบบเข้าใจง่าย. คอลัมน์ของ Q ตั้งฉากกันเป็นคู่ และแต่ละคอลัมน์ยาวหนึ่งหน่วย การคูณด้วย Q จึงหมุนหรือสะท้อนเวกเตอร์โดยไม่เปลี่ยนความยาว เพราะ $\|Q\mathbf{x}\|_2^2 = \mathbf{x}^T Q^T Q \mathbf{x} = \|\mathbf{x}\|_2^2$ ข้อเท็จจริงนี้เราจะใช้ในการพิสูจน์ทฤษฎีบทของการแก้ด้วย QR

ตัวอย่างที่ง่ายที่สุดของเมทริกซ์แบบนี้คือ **การสะท้อนแบบเฮาส์โฮลเดอร์** (Householder reflection) ตำรา Atkinson 1989, §9.3, pp. 609--610 ใช้เมทริกซ์รูปนี้เพื่อแปลงเวกเตอร์ให้เป็นรูปที่ง่ายขึ้น เราเขียนเองตามแนวทางนั้นดังนี้

ภาพรวมก่อนอ่าน

ถ้าเราต้องการให้เวกเตอร์ $\mathbf{x}$ ชี้ไปทางแกนแรกโดยไม่เปลี่ยนความยาว เราสะท้อนมันผ่านระนาบที่ตั้งฉากกับผลต่างระหว่าง $\mathbf{x}$ กับเป้าหมาย ระนาบนั้นมีเวกเตอร์ปกติ $\mathbf{v}$ และการสะท้อนผ่านมันคือ $I - 2\mathbf{v}\mathbf{v}^T/(\mathbf{v}^T\mathbf{v})$ ทิศของการสะท้อนต้องเลือกให้ดี ไม่เช่นนั้นการลบจะหักล้างกันเอง

ฉบับร่าง · Preprint · 2026-10-23

ทฤษฎีบท 8.1.2: การสะท้อนแบบเฮาส์โฮลเดอร์

ให้ $\mathbf{x} \in \mathbb{R}^n$ ไม่เป็นศูนย์ กำหนด

$$\alpha = -\operatorname{sign}(x_1)\,\|\mathbf{x}\|_2, \qquad \mathbf{v} = \mathbf{x} - \alpha\mathbf{e}_1, \qquad H = I - \frac{2\,\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}},$$

โดยที่ $\operatorname{sign}(0) = 1$ และ $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ แล้ว H เป็นเมทริกซ์สมมาตรและเชิงตั้งฉาก และ $H\mathbf{x} = \alpha\mathbf{e}_1$

พิสูจน์. **แนวคิด.** H สมมาตรเพราะเป็น I ลบเมทริกซ์สมมาตร ตั้งฉากเพราะการคำนวณ $H^TH = I$ ตรงไปตรงมา และ $H\mathbf{x}$ ได้เพราะ $\mathbf{x}$ กับ $\alpha\mathbf{e}_1$ มีความยาวเท่ากัน

1. $\mathbf{v} \neq 0$. ถ้า $\mathbf{v} = 0$ จะได้ $\mathbf{x} = \alpha\mathbf{e}_1$ จึง $|\alpha| = \|\mathbf{x}\|_2$ และ $x_1 = \alpha$ แต่ $\alpha = -\operatorname{sign}(x_1)\|\mathbf{x}\|_2 = -\operatorname{sign}(\alpha)\,|\alpha| = -\alpha$ ทำให้ $\alpha = 0$ จึง $\mathbf{x} = 0$ ขัดกับสมมติฐาน
2. **สมมาตร.** $H^T = I - \frac{2}{\mathbf{v}^T\mathbf{v}}(\mathbf{v}\mathbf{v}^T)^T = H$
3. **ตั้งฉาก.** ให้ $c = \frac{2}{\mathbf{v}^T\mathbf{v}}$ ได้ $H^TH = (I - c\mathbf{v}\mathbf{v}^T)^2 = I - 2c\mathbf{v}\mathbf{v}^T + c^2\mathbf{v}(\mathbf{v}^T\mathbf{v})\mathbf{v}^T = I - 2c\mathbf{v}\mathbf{v}^T + c\cdot 2\mathbf{v}\mathbf{v}^T = I$ เพราะ $c^2\mathbf{v}^T\mathbf{v} = c \cdot 2$
4. **ผลของ H กับ $\mathbf{x}$.** จาก $\mathbf{v}^T\mathbf{v} = \|\mathbf{x}\|^2 - 2\alpha x_1 + \alpha^2$ และ $\alpha^2 = \|\mathbf{x}\|^2$ ได้ $\mathbf{v}^T\mathbf{v} = 2(\|\mathbf{x}\|^2 - \alpha x_1) = 2\,\mathbf{v}^T\mathbf{x}$ ดังนั้น $H\mathbf{x} = \mathbf{x} - 2\mathbf{v}\frac{\mathbf{v}^T\mathbf{x}}{\mathbf{v}^T\mathbf{v}} = \mathbf{x} - \mathbf{v} = \alpha\mathbf{e}_1$
5. **เครื่องหมายของ α.** α มีเครื่องหมายตรงข้ามกับ x_1 จึงไม่มีการหักล้างใน $v_1 = x_1 - \alpha$ เพราะ x_1 กับ $-\alpha$ มีเครื่องหมายเดียวกัน

สรุป: ผลคือ H เป็นการสะท้อนที่รักษาความยาว และส่ง $\mathbf{x}$ ไปที่แกนแรก บทพิสูจน์นี้เราเขียนเองตามข้อเท็จจริงจากตำรา □

ฉบับร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

ตัวอย่าง $\mathbf{x} = (3, 4)$ มีความยาว 5 ถ้าเลือก $\alpha = -5$ จะได้ $\mathbf{v} = (8, 4)$ และ

$$H = \begin{pmatrix} -0.6 & -0.8 \\ -0.8 & 0.6 \end{pmatrix}, \qquad H\begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} -5 \\ 0 \end{pmatrix}.$$

ทุกค่าตรวจได้ด้วยมือ และตรงกับโค้ดของเรา (`test_ch08.py`)

ข้อควรระวัง: เลือกเครื่องหมายผิดแล้วเสียความแม่นยำ

ถ้าเลือก $\alpha = +\|\mathbf{x}\|$ แทน จะได้ $v_1 = x_1 - \|\mathbf{x}\|$ ซึ่งเป็นการลบเลขสองจำนวนที่ใกล้กัน (เรื่องการลบแบบหักล้างจากบทที่ 1) สำหรับ $\mathbf{x} = (1,\ 10^{-6})$ ค่า v_1 ที่ได้จากการลบโดยตรงมีค่าคลาดเคลื่อนสัมพัทธ์ประมาณ 8.9×10^{-5} ขณะที่ใช้กฎของทฤษฎีบท ค่าคลาดเคลื่อนเกือบเป็นศูนย์ (2×10^{-17}) ตัวเลขทั้งหมดนี้มาจากโค้ดของเรา ค่าที่แม่นจริงคำนวณด้วยทศนิยม 50 หลัก

เกร็ดเพิ่มเติม

การหมุนกีฟเวนส์ (Givens rotation) เป็นเมทริกซ์เชิงตั้งฉากอีกแบบหนึ่งที่ทำงานกับเพียงสองแถวในแต่ละครั้ง ถ้า

$c = a/r$, $s = b/r$ และ $r = \sqrt{a^2 + b^2}$ เมทริกซ์ $\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$ ส่ง (a, b) ไปที่ $(r, 0)$ ตัวอย่าง $(a, b) = (3, 4)$ ได้ $c = 0.6$, $s = 0.8$, $r = 5$ ซึ่งเราใช้เมื่อต้องการกำจัดสมาชิกเพียงตัวเดียวโดยไม่แตะส่วนอื่นมาก

จาก QR ด้วยการสะท้อนสู่คำตอบ

แนวคิด. ถ้าเราสะท้อนคอลัมน์แรกของ A ให้เป็นทิศของแกนแรก สมาชิกที่อยู่ใต้ตัวหลักจะเป็นศูนย์หมด แล้วทำซ้ำกับคอลัมน์ที่สอง ผลสุดท้ายคือ A ที่ถูกคูณด้วยเมทริกซ์เชิงตั้งฉากหลายตัวจนเป็นรูปสามเหลี่ยมบน

นิยาม 8.1.3: การแยกตัวประกอบ QR

ให้ $A \in \mathbb{R}^{m \times n}$ การเขียน $A = QR$ เมื่อ Q เป็นเมทริกซ์เชิงตั้งฉากขนาด $m \times m$ และ R เป็นเมทริกซ์สามเหลี่ยมบนขนาด $m \times n$ เรียกว่า **การแยกตัวประกอบ QR** (QR factorization) ของ A

ทฤษฎีบท 8.1.3: การแก้ปัญหากำลังสองน้อยสุดด้วย QR

ให้ $A \in \mathbb{R}^{m \times n}$ มีลำดับชั้นเต็มในแนวคอลัมน์ และ $m \geq n$ ให้ $A = QR$ โดยที่ $R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$ และ R_1 เป็นเมทริกซ์สามเหลี่ยมบน $n \times n$ ให้ $\mathbf{c} = Q^T\mathbf{b} = (\mathbf{c}_1, \mathbf{c}_2)$ โดยที่ $\mathbf{c}_1 \in \mathbb{R}^n$ แล้ว R_1 ไม่เอกฐาน และคำตอบกำลังสองน้อยสุดเพียงคำตอบเดียวคือ $\mathbf{x}^* = R_1^{-1}\mathbf{c}_1$ โดยค่าตกค้างต่ำสุดคือ $\|\mathbf{c}_2\|_2$

พิสูจน์. **แนวคิด.** เพราะ Q ไม่เปลี่ยนความยาว เราแทน $\mathbf{b} - A\mathbf{x}$ ด้วย $Q^T(\mathbf{b} - A\mathbf{x})$ แล้วแยกเป็นส่วนที่แก้ได้กับส่วนที่แก้ไม่ได้

1. **R_1 ไม่เอกฐาน.** $\operatorname{rank} A = \operatorname{rank} R$ เพราะ Q เป็นเมทริกซ์ไม่เอกฐาน และ $\operatorname{rank} R = \operatorname{rank} R_1$ เพราะแถวล่างเป็นศูนย์ เนื่องจาก $\operatorname{rank} A = n$ จึง R_1 ไม่เอกฐาน

2. **แปลงค่าตกค้าง.** เมทริกซ์ Q จัตุรัสและ $Q^TQ = I$ จึง $QQ^T = I$ และ Q^T ก็เชิงตั้งฉากด้วย ดังนั้น $\|Q^T\mathbf{y}\|_2^2 = \mathbf{y}^TQQ^T\mathbf{y} = \|\mathbf{y}\|_2^2$ ดังนั้น
$$\|\mathbf{b} - A\mathbf{x}\|_2 = \|Q^T(\mathbf{b} - A\mathbf{x})\|_2 = \|\mathbf{c} - R\mathbf{x}\|_2.$$

3. **แยกเป็นสองส่วน.** เขียน $\mathbf{c} - R\mathbf{x} = (\mathbf{c}_1 - R_1\mathbf{x},\ \mathbf{c}_2)$ จึง $\|\mathbf{c} - R\mathbf{x}\|_2^2 = \|\mathbf{c}_1 - R_1\mathbf{x}\|_2^2 + \|\mathbf{c}_2\|_2^2$ ส่วนที่สองไม่ขึ้นกับ $\mathbf{x}$

4. **เลือก $\mathbf{x}$.** ส่วนแรกเป็นศูนย์เมื่อ $R_1\mathbf{x} = \mathbf{c}_1$ ซึ่งมีคำตอบเดียวเพราะ R_1 ไม่เอกฐาน ดังนั้น $\mathbf{x}^* = R_1^{-1}\mathbf{c}_1$ และค่าตกค้างต่ำสุดคือ $\|\mathbf{c}_2\|_2$

สรุป: การแยก QR แปลงปัญหาที่ดูยุ่งให้เป็นระบบสามเหลี่ยม ซึ่งแก้จากล่างขึ้นบนได้ บทพิสูจน์นี้เราเขียนเอง □

อ่านแบบเข้าใจง่าย. ตัวเลขที่เราต้องคำนวณมีเพียงสองอย่าง คือ $Q^T\mathbf{b}$ กับการแทนค่ากลับในระบบสามเหลี่ยม ไม่มีการสร้าง A^TA จึงไม่มีการยกกำลังสองของคอนดิชันนัมเบอร์ ตรงนี้คือเหตุผลที่ QR มักปลอดภัยกว่าสมการปรกติ

ฉบับร่าง Preprint · 2026-10-23

ตารางที่ 8.1: ความไม่ตั้งฉากของ Q จากวิธีต่าง ๆ (ตัวเลขจากโค้ดของเรา ใช้เป็นเปรียบเทียบเท่านั้น)

วิธี	$\lVert Q^{T}Q - I \rVert_{F}$
แกรม–ชมิดท์แบบคลาสสิก	6.0×10^{-7}
แกรม–ชมิดท์แบบปรับแก้	5.7×10^{-12}

ตารางที่ 8.2: จุดสีน้ำเงินในแบบจำลองสามมิติของเรา (แปดดาวเทียม ตำแหน่งจริงและค่าเวลาจริง 30 เมตร) ตามรอบของเกาส์–นิวตัน (ตัวเลขจากโค้ดของเรา)

รอบ	ค่าคลาดเคลื่อนสัมบูรณ์ของตำแหน่ง (ม.)	ค่าเวลาที่ประมาณ (ม.)	ค่าตกค้าง (ม.)
0	1.526×10^{6}	0	2.173×10^{6}
1	3.304×10^{4}	$51\,263.5$	1.352×10^{5}
2	10.30	44.83	61.68
3	7.646	23.04	4.266
4	7.646	23.04	4.266

สองวิธีของแกรม–ชมิดท์ในเลขคอมพิวเตอร์

วิธีที่อาจคิดได้ก่อนคือ **กระบวนการแกรม–ชมิดท์** (Gram--Schmidt process) ซึ่งหาเวกเตอร์ตั้งฉากทีละคอลัมน์ การทำตามสูตรตรง ๆ เรียกว่าแบบคลาสสิก และการทำโดยใช้ผลล่าสุดของแต่ละขั้นเรียกว่าแบบปรับแก้ ทั้งสองให้ผลทางคณิตศาสตร์เหมือนกัน แต่ในเลขคอมพิวเตอร์ไม่เหมือนกัน ตารางต่อไปนี้วัด $\lVert Q^{T}Q - I \rVert_{F}$ ของเมทริกซ์ Vandermonde ขนาด 12×8 ที่มีคอนดิชันนัมเบอร์ประมาณ 1.2×10^{5}

แบบคลาสสิกเสียความตั้งฉากไปราว 6×10^{-7} ขณะที่แบบปรับแก้เหลือราว 5.7×10^{-12} ดังนั้นเมื่อเราต้องสร้าง Q จริง ๆ ควรใช้การสะท้อนของเฮาส์โฮลเดอร์ ซึ่งในโค้ดของเราให้ความไม่ตั้งฉากไม่เกิน 10^{-13} หรือใช้แบบปรับแก้

จุดสีน้ำเงินอีกครั้ง: การปรับแบบวนซ้ำ

กลับไปที่แบบจำลองของเรา ระยะที่วัดได้ y_i ไม่เป็นเชิงเส้นในตำแหน่ง $\mathbf{p}$ เพราะ $f_i(\mathbf{p}, b) = \lVert \mathbf{s}_i - \mathbf{p} \rVert_2 + b$ เราจึงใช้วิธี **เกาส์–นิวตัน** (Gauss--Newton) ซึ่งแต่ละรอบแก้ปัญหากำลังสองน้อยสุดเชิงเส้น $J\boldsymbol{\delta} \approx \mathbf{y} - f(\mathbf{x})$ ด้วยการแยก QR แล้วปรับ $\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\delta}$ เมทริกซ์ J คือ **เมทริกซ์จาโคเบียน** (Jacobian matrix) ของ f โดยแถวที่ i คือ $-(\mathbf{s}_i - \mathbf{p})^{T} / \lVert \mathbf{s}_i - \mathbf{p} \rVert_2$ และคอลัมน์ของ b เป็นหนึ่งทั้งหมด

ในแบบจำลองสามมิติของเรา ดาวเทียมแปดดวงอยู่บนทรงกลมรัศมี 2×10^{7} เมตรที่ระยะสมมติ ระยะที่วัดมีสัญญาณรบกวนราว 2 เมตร เริ่มจากจุดกำเนิดที่ห่างจากตำแหน่งจริงราว 1.5×10^{6} เมตร หลังสองรอบ ตำแหน่งอยู่ห่างจากตำแหน่งจริงราว 10.3 เมตร หลังสามรอบอยู่ห่างราว 7.6 เมตร และไม่เปลี่ยนอีกเมื่อทำต่อ ค่าเวลาที่ประมาณได้คือ 23.0 เมตร เทียบกับค่าจริง 30 เมตร (ตัวเลขจากโค้ดของเรา) ค่าตกค้างสุดท้ายคือ 4.3 เมตร ไม่มีความหมายว่าเราพบตำแหน่งจริงแล้ว เพราะความคลาดเคลื่อนในตำแหน่งและเวลาผูกกันในการคำนวณ ตาราง 8.2 แสดงลำดับของการทำซ้ำ

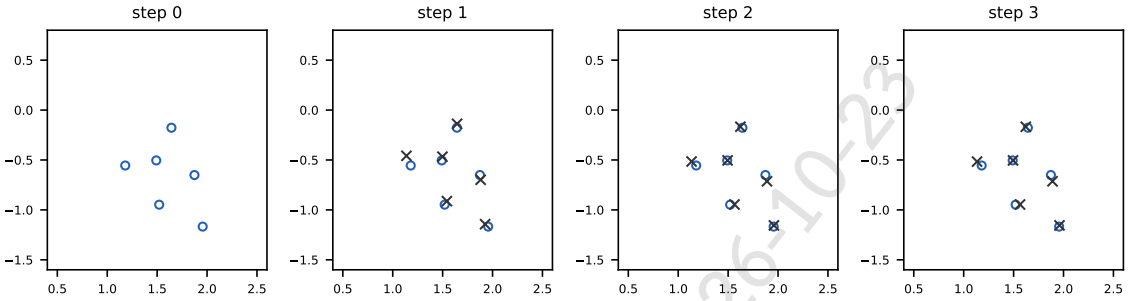
เกร็ดเพิ่มเติม

ตัวขยายความคลาดเคลื่อนจากเรขาคณิต (dilution of precision) เป็นชื่อที่เราใช้เรียก $\sqrt{\operatorname{tr}\left((J^{T}J)^{-1}\right)}$ ที่ตำแหน่งจริง มันบอกว่าเรขาคณิตของดาวเทียมขยายความคลาดเคลื่อนของระยะเท่าไร ดาวเทียมแปดดวงที่กระจายดี ได้ค่าราว 3.6 ส่วนดาวเทียมแปดดวงที่กระจุกอยู่ในย่านท้องฟ้าเล็ก ๆ ได้ค่าราว 2696 ที่ความคลาดเคลื่อน 2 เมตรเท่ากัน ความคลาดเคลื่อนของตำแหน่งที่คาดได้จึงเป็นประมาณ 7 เมตรกับประมาณ 5.4 กิโลเมตร (ตัวเลขจากโค้ดของเรา ชื่อนี้เป็นศัพท์ที่เราเลือกใช้ในบทนี้)

ร่างต้น Preprint 2026-10-23

ตารางที่ 8.3: การปรับท่าทางสองมิติ: ค่าประมาณตามรอบ และค่าจริงคือ $\theta = 0.6$, $t_x = 1.5$, $t_y = -0.5$ (ตัวเลขจากโค้ดของเรา)

รอบ	θ	t_x	t_y	ค่าตกค้าง
0	0	0	0	4.1558
1	0.564128	1.497926	−0.467069	0.140745
2	0.619280	1.494495	−0.503956	0.102540
3	0.620993	1.494383	−0.504409	0.102521
4	0.621045	1.494375	−0.504415	0.102521



รูปที่ 8.5: จุดสมมติหกจุดที่สังเกตได้ (วงกลมว่าง) กับจุดที่แบบจำลองทำนาย (กากบาท) ในรอบที่ 0 ถึง 3 ของการปรับท่าทาง ค่าตกค้างลดลงเร็วในรอบแรกแล้วคงที่ (สร้างจาก `numanal/ch08.py`)

ข้อควรระวัง: ค่าตกค้างเล็กไม่ได้แปลว่าคำตอบถูก

ค่าตกค้างบอกว่าเราเข้าใกล้ระยะที่วัดได้เพียงใด ไม่ได้บอกว่าตำแหน่งนั้นใกล้ตำแหน่งจริง ถ้าเรขาคณิตไม่ดี เช่นดาวเทียมกระจุกกัน คำตอบอาจไปไกลจากจริงมาก ทั้งที่ค่าตกค้างเล็ก เมื่อคอนดิชันนัมเบอร์ของ J ใหญ่ ต้องดูความไวของคำตอบด้วย ไม่ใช่ดูค่าตกค้างอย่างเดียว

ท่าทางของใบหน้าสองมิติ

ในแบบจำลองใบหน้า เราหมุนจุดสมมติหกจุด $\mathbf{P}_k$ ด้วยมุม θ แล้วเลื่อนด้วย $\mathbf{t} = (t_x, t_y)$ ได้ตำแหน่ง $R(\theta)\mathbf{P}_k + \mathbf{t}$ ซึ่งไม่เป็นเชิงเส้นใน θ จึงใช้เกาส์–นิวตันเหมือนเดิม เมทริกซ์จาโคเบียนมี 3 คอลัมน์ คือ $R'(\theta)\mathbf{P}_k$ และคอลัมน์ของการเลื่อนสองแบบ ค่าจริงคือ $\theta = 0.6$ $t_x = 1.5$ $t_y = -0.5$ และสัญญาณรบกวนราว 0.02 (ตัวเลขจากโค้ดของเรา) เริ่มจาก $\theta = 0$ และ $\mathbf{t} = 0$ ค่าตกค้างเริ่มต้นคือ 4.156 หลังหนึ่งรอบลดเหลือ 0.1407 และหลังสี่รอบ $\theta = 0.6210$ ค่าตกค้างคือ 0.1025 ซึ่งเล็กกว่าค่าตกค้างที่จุดจริง 0.1050 เพราะคำตอบกำลังสองน้อยสุดปรับตามสัญญาณรบกวนในข้อมูลด้วย จึงไม่จำเป็นต้องตรงกับค่าจริงพอดี ตาราง 8.3 แสดงรอบต่าง ๆ

สรุปใจความ

- เมื่อสมการมากกว่าตัวไม่ทราบ เราหาคำตอบที่ทำให้ผลบวกกำลังสองของค่าตกค้างน้อยที่สุด เรียกว่าปัญหากำลังสองน้อยสุด
- สมการปรกติ $A^T A\mathbf{x} = A^T\mathbf{b}$ ให้คำตอบ แต่ทำให้คอนดิชันนัมเบอร์เป็นกำลังสอง จึงไม่เสถียรเมื่อคอลัมน์เกือบซ้ำกัน

ฉบับร่าง · Preprint · 2026-10-23

- การสะท้อนของเฮาส์โฮลเดอร์รักษาความยาว และเลือกเครื่องหมายเพื่อหลีกเลี่ยงการหักล้าง ผลคือการแยก QR ที่ได้ Q ตั้งฉากเกือบแม่นตรง
- แก้ปัญหากำลังสองน้อยสุดด้วย QR ได้ $\mathbf{x}^* = R_1^{-1}\mathbf{c}_1$ ไม่ต้องสร้าง A^TA
- ค่าตกค้างเล็กไม่รับประกันว่าคำตอบถูก เรขาคณิตของจุดอ้างอิงกำหนดความไวต่อความคลาดเคลื่อน

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 8.1.1 (⋆) ให้ $\mathbf{x} = (5,\ 12)^T$

1. จงหา α และ $\mathbf{v}$ ตามทฤษฎีบทการสะท้อนของเฮาส์โฮลเดอร์
2. จงเขียนเมทริกซ์ $H = I - 2\mathbf{v}\mathbf{v}^T/(\mathbf{v}^T\mathbf{v})$
3. จงตรวจว่า $H\mathbf{x} = (-13,\ 0)^T$

แบบฝึกหัด 8.1.2 (⋆) ใช้การหมุนกีฟเวนส์ $\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$ ซึ่ง $c = a/r,\ s = b/r,\ r = \sqrt{a^2+b^2}$

1. จงหา c, s, r สำหรับ $(a, b) = (8, 6)$ และตรวจว่าการหมุนส่ง $(8, 6)$ ไปที่ $(10, 0)$
2. จงหา c, s, r สำหรับ $(a, b) = (-3, 4)$ และตรวจว่าการหมุนส่ง $(-3, 4)$ ไปที่ $(5, 0)$

แบบฝึกหัด 8.1.3 (⋆⋆) ให้ $\mathbf{v} \neq \mathbf{0}$ และ $H = I - \dfrac{2\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}$

1. จงแสดงว่า H เป็นเมทริกซ์สมมาตร
2. จงแสดงว่า $H^TH = I$ โดยใช้ $c = 2/(\mathbf{v}^T\mathbf{v})$ และ $c^2\,\mathbf{v}^T\mathbf{v} = 2c$
3. จงตรวจด้วยมือสำหรับ $\mathbf{v} = (1, 2, -2)^T$ ว่า $H = \begin{pmatrix} 7/9 & -4/9 & 4/9 \\ -4/9 & 1/9 & 8/9 \\ 4/9 & 8/9 & 1/9 \end{pmatrix}$ และแถวของ H ยาวหนึ่งหน่วยทุกแถว

แบบฝึกหัด 8.1.4 (⋆⋆) ให้ $\mathbf{x} = (1,\ \varepsilon)^T$ โดยที่ $\varepsilon > 0$ เล็ก และให้ $\|\mathbf{x}\|_2 = \sqrt{1+\varepsilon^2}$

1. จงแสดงว่า $1 + \sqrt{1+\varepsilon^2}$ เป็นค่าประมาณ 2 ไม่มีการหักล้าง และ $1 - \sqrt{1+\varepsilon^2} = -\dfrac{\varepsilon^2}{1+\sqrt{1+\varepsilon^2}}$
2. จงใช้ข้อ 1 อธิบายว่าเมื่อ $\varepsilon = 10^{-6}$ การคำนวณ $1 - \sqrt{1+\varepsilon^2}$ ตรงไปตรงมาในเลขทศนิยม จะเสียความแม่นยำไปมากเพียงใด ทำนายลำดับของค่าคลาดเคลื่อนสัมพัทธ์ก่อน แล้วเทียบกับโค้ดของเรา

แบบฝึกหัด 8.1.5 (⋆⋆) ให้ $A = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$ และ $\varepsilon > 0$

1. จงแสดงว่า $A^TA = \begin{pmatrix} 1+\varepsilon^2 & 1 \\ 1 & 1+\varepsilon^2 \end{pmatrix}$ และค่าลักษณะเฉพาะคือ $2+\varepsilon^2$ กับ ε^2 โดยมีเวกเตอร์ลักษณะเฉพาะ $(1, 1)^T$ และ $(1, -1)^T$

ฉบับร่าง · Preprint · 2026-10-23

2. จงหา $\operatorname{cond}(A^T A) = (2+\varepsilon^2)/\varepsilon^2$ และเทียบกับ $\operatorname{cond}(A)^2$ เมื่อ $\operatorname{cond}(A) = \sqrt{2}/\varepsilon$

3. ให้ $\varepsilon = 10^{-4}$ และ 10^{-5} จงตรวจด้วยโค้ดว่าการแยก QR ให้คำตอบ $(1, 1)$ ได้ใกล้มาก (ค่าความคลาดเคลื่อนไม่เกิน 10^{-12})

แบบฝึกหัด 8.1.6 (★★) ข้อมูลห้าจุด (t, b) คือ $(0, 1)$, $(1, 2.9)$, $(2, 5.2)$, $(3, 6.8)$, $(4, 9.1)$ ต้องการหาเส้น $b \approx x_0 + x_1 t$ ที่ทำให้ผลบวกกำลังสองของค่าตกค้างน้อยที่สุด

1. จงสร้าง A และ $\mathbf{b}$ แล้วเขียนสมการปรกติ $A^T A\mathbf{x} = A^T\mathbf{b}$

2. จงแก้สมการปรกติด้วยมือ ได้ $x_0 = 0.98$ และ $x_1 = 2.01$

3. จงหาค่าตกค้างทั้งห้า รวมกันเป็นศูนย์ และ $\sum t_i r_i = 0$ แล้วหา $\sum r_i^2$ และ $\|\mathbf{r}\|_2$

แบบฝึกหัด 8.1.7 (★ ★ ★) ใช้ตัวขยายความคลาดเคลื่อนจากเรขาคณิต (เกร็ดเพิ่มเติมของหัวข้อ 8.1) ที่ตำแหน่งจริง ให้ค่า $d_{\text{good}} \approx 3.62$ สำหรับดาวเทียมกระจายดี และ $d_{\text{cluster}} \approx 2696$ สำหรับดาวเทียมกระจุกกัน สมมติว่าความคลาดเคลื่อนของระยะคือ $\sigma = 2$ เมตร

1. จงอธิบายว่าเหตุใดขนาดของความคลาดเคลื่อนของตำแหน่งและเวลาจึงประมาณได้ด้วย σd (ใช้การประมาณเชิงเส้นและความแปรปรวนของ $\mathbf{y}$ เป็น $\sigma^2 I$ และตรวจว่าความแปรปรวนของ $\mathbf{x}$ เป็น $\sigma^2 (J^T J)^{-1}$)

2. จงคำนวณ σd ของทั้งสองกรณี และบอกว่าสองกรณีต่างกันกี่เท่า

3. จงอธิบายด้วยถ้อยคำของคุณเองว่าทำไมดาวเทียมที่กระจุกกันทำให้ $J^T J$ เกือบเอกฐาน และทำไมค่าตกค้างอาจยังเล็กในกรณีนั้น

ตัวช่วยคิด

ข้อ 1: เขียน $\mathbf{x} - \mathbf{x}_{\text{true}} \approx (J^T J)^{-1} J^T \boldsymbol{\varepsilon}$ เมื่อ $\boldsymbol{\varepsilon}$ คือความคลาดเคลื่อนของระยะ แล้วใช้ $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I$ ข้อ 3: ดูที่ค่าลักษณะเฉพาะที่เล็กที่สุดของ $J^T J$

แบบฝึกหัด 8.1.8 (★ ★ ★) ใช้แบบจำลองท่าทางสองมิติในหัวข้อ 8.1 ซึ่ง $f_k(\theta, \mathbf{t}) = R(\theta)\mathbf{P}_k + \mathbf{t}$ และเมทริกซ์จาโคเบียนมีสามคอลัมน์

1. ให้มีจุดอ้างอิงเพียงจุดเดียวที่ $\mathbf{P}_1 = (0, 0)$ จงแสดงว่าคอลัมน์ของ θ ของ J เป็นศูนย์ และลำดับชั้นของ J เท่ากับ 2

2. ให้จุดอ้างอิงสองจุดคือ $(-0.3,\ 0.2)$ และ $(0.3,\ 0.2)$ และข้อมูลไม่มีสัญญาณรบกวน (ค่าที่วัดคือ $f(\theta, \mathbf{t})$ ที่ $\theta = 0.6$, $\mathbf{t} = (1.5, -0.5)$) จงแสดงว่าลำดับชั้นของ J เป็น 3 และตรวจว่าการทำซ้ำเกาส์–นิวตันจากจุดเริ่ม $(0, 0, 0)$ ลู่เข้าหาค่าจริงได้

ตัวช่วยคิด

ข้อ 1: คำนวณ $R'(\theta)\mathbf{P}_1$ โดยตรงเมื่อ $\mathbf{P}_1 = 0$ ข้อ 2: ดูว่าสองจุดนี้ทำให้คอลัมน์ทั้งสามเป็นอิสระเชิงเส้นได้หรือไม่

ฉบับร่าง · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 8.1.1.

1. $\|\mathbf{x}\|_2 = \sqrt{25+144} = 13$ และ $x_1 = 5 > 0$ จึง $\alpha = -13$ แล้ว $\mathbf{v} = \mathbf{x} - \alpha\mathbf{e}_1 = (5+13,\ 12)^T = (18,\ 12)^T$

2. $\mathbf{v}^T\mathbf{v} = 324 + 144 = 468$ แล้ว
$$H_{11} = 1 - \frac{2\cdot 324}{468} = -\frac{5}{13}, \quad H_{12} = -\frac{2\cdot 216}{468} = -\frac{12}{13}, \quad H_{22} = 1 - \frac{2\cdot 144}{468} = \frac{5}{13}.$$
ดังนั้น $H = \begin{pmatrix} -5/13 & -12/13 \\ -12/13 & 5/13 \end{pmatrix}$

3. $H\mathbf{x}$: แถวแรก $\frac{-25-144}{13} = -13$ แถวที่สอง $\frac{-60+60}{13} = 0$ จึง $H\mathbf{x} = (-13,\ 0)^T$ ตรงกับที่ต้องการ

ตรวจด้วยโค้ด: ใน `test_ch08.py`(`test_householder_exercise_five_twelve`)

เฉลย 8.1.2.

1. $r = \sqrt{64+36} = 10, c = 0.8, s = 0.6$ การหมุนให้ $(c\cdot 8 + s\cdot 6,\ -s\cdot 8 + c\cdot 6) = (6.4+3.6,\ -4.8+4.8) = (10,\ 0)$

2. $r = \sqrt{9+16} = 5$, $c = -0.6$, $s = 0.8$ การหมุนให้ $(c\cdot(-3) + s\cdot 4,\ -s\cdot(-3) + c\cdot 4) = (1.8+3.2,\ 2.4-2.4) = (5,\ 0)$

ตรวจด้วยโค้ด: ใน `test_ch08.py`(`test_givens_examples`)

เฉลย 8.1.3.

1. $(\mathbf{v}\mathbf{v}^T)^T = \mathbf{v}\mathbf{v}^T$ จึง $H^T = I - \frac{2}{\mathbf{v}^T\mathbf{v}}\mathbf{v}\mathbf{v}^T = H$

2. ให้ $c = \frac{2}{\mathbf{v}^T\mathbf{v}}$ เราได้ $H^TH = H^2 = (I - c\mathbf{v}\mathbf{v}^T)^2 = I - 2c\mathbf{v}\mathbf{v}^T + c^2\mathbf{v}(\mathbf{v}^T\mathbf{v})\mathbf{v}^T$ และ $c^2\mathbf{v}^T\mathbf{v} = 2c$ ทำให้พจน์ที่สามเท่ากับ $2c\mathbf{v}\mathbf{v}^T$ หักล้างพจน์ที่สอง จึง $H^TH = I$

3. สำหรับ $\mathbf{v} = (1, 2, -2)^T$ ได้ $\mathbf{v}^T\mathbf{v} = 9$ และ $H = I - \frac{2}{9}\mathbf{v}\mathbf{v}^T$ คูณออกได้ตามเมทริกซ์ในโจทย์ แถวแรกยาว $\frac{49+16+16}{81} = 1$ แถวที่สองยาว $\frac{16+1+64}{81} = 1$ แถวที่สามยาว $\frac{16+64+1}{81} = 1$ และแถวคู่ที่ตั้งฉากกันได้ผลคูณ 0 ทั้งหมด

ตรวจด้วยโค้ด: ใน `test_ch08.py`(`test_reflector_is_symmetric_and_orthogonal`) ด้วยเวกเตอร์สุ่ม

เฉลย 8.1.4.

1. เมื่อ $\varepsilon > 0$ ทั้ง 1 และ $\sqrt{1+\varepsilon^2}$ เป็นบวก จึง $1 + \sqrt{1+\varepsilon^2} \approx 2$ ไม่มีการลบเลขใกล้กันเลย ส่วน $1 - \sqrt{1+\varepsilon^2} = \frac{(1-\sqrt{1+\varepsilon^2})(1+\sqrt{1+\varepsilon^2})}{1+\sqrt{1+\varepsilon^2}} = \frac{1-(1+\varepsilon^2)}{1+\sqrt{1+\varepsilon^2}} = -\frac{\varepsilon^2}{1+\sqrt{1+\varepsilon^2}}$ ซึ่งมีค่าประมาณ $-\varepsilon^2/2$

2. เมื่อ $\varepsilon = 10^{-6}$ ค่าที่แม่นคือ $\approx -5\times 10^{-13}$ แต่ถ้าคำนวณ $1 - \sqrt{1+\varepsilon^2}$ โดยตรง ผลของ $\sqrt{1+10^{-12}}$ ถูกปัดด้วยค่าคลาดเคลื่อนสัมบูรณ์ราว 10^{-16} ซึ่งเป็นระดับของ u การลบเลข 1 ออกเหลือค่าเล็ก 5×10^{-13} จึงมีค่าคลาดเคลื่อนสัมพัทธ์ราว $10^{-16}/(5\times 10^{-13}) \approx 2\times 10^{-4}$ ลำดับ 10^{-4} นี้ตรงกับโค้ดของเราที่ได้ 8.9×10^{-5} ส่วนสูตรที่ข้อ 1 ให้ค่าคลาดเคลื่อนสัมบูรณ์ราว 10^{-17} ตามที่บันทึกในข้อควรระวัง

ตรวจด้วยโค้ด: ใน `test_ch08.py`(`test_sign_rule_cancellation_numbers`)

ฉบับร่าง Preprint 2026-10-23

เฉลย 8.1.5.

1. คอลัมน์ของ A คือ $\mathbf{a}_1 = (1, \varepsilon, 0)^T$ และ $\mathbf{a}_2 = (1, 0, \varepsilon)^T$ จึง $\mathbf{a}_1^T\mathbf{a}_1 = \mathbf{a}_2^T\mathbf{a}_2 = 1 + \varepsilon^2$ และ $\mathbf{a}_1^T\mathbf{a}_2 = 1$ ดังนั้น A^TA ตามที่โจทย์ให้ เมื่อคูณ $(1,1)^T$ ได้ $(2+\varepsilon^2)(1,1)^T$ และคูณ $(1,-1)^T$ ได้ $\varepsilon^2(1,-1)^T$

2. ค่าลักษณะเฉพาะมากสุดและน้อยสุดคือ $2+\varepsilon^2$ และ ε^2 จึง $\operatorname{cond}(A^TA) = \dfrac{2+\varepsilon^2}{\varepsilon^2}$ ส่วน A มีค่าเอกฐาน $\sqrt{2+\varepsilon^2}$ กับ ε จึง $\operatorname{cond}(A) = \sqrt{2+\varepsilon^2}/\varepsilon$ ซึ่งเมื่อ ε เล็ก ประมาณ $\sqrt{2}/\varepsilon$ และ $\operatorname{cond}(A)^2 \approx 2/\varepsilon^2$ ตรงกับ $\operatorname{cond}(A^TA)$ ที่ลำดับเดียวกัน

3. ที่ $\varepsilon = 10^{-4}$ ได้ $\operatorname{cond}(A^TA) \approx 2\times10^{8}$ และที่ 10^{-5} ได้ $\approx 2\times10^{10}$ โค้ดให้ $(1,1)$ ทั้งสองกรณี โดย ความคลาด เคลื่อน ไม่ เกิน 10^{-12} ซึ่ง ทดสอบ ใน `test_ch08.py`(`test_exercise_815_normal_equations_eigenvalues`)

เฉลย 8.1.6.

1. $A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$ และ $\mathbf{b} = (1,\ 2.9,\ 5.2,\ 6.8,\ 9.1)^T$ จากนั้น $A^TA = \begin{pmatrix} 5 & 10 \\ 10 & 30 \end{pmatrix}$ และ $A^T\mathbf{b} = (25,\ 70.1)^T$ เพราะ $\sum b_i = 25$ และ $\sum t_i b_i = 0 + 2.9 + 10.4 + 20.4 + 36.4 = 70.1$

2. จาก $5x_0 + 10x_1 = 25$ ได้ $x_0 = 5 - 2x_1$ แทนในสมการที่สอง $10x_0 + 30x_1 = 70.1$ ได้ $50 - 20x_1 + 30x_1 = 70.1$ จึง $x_1 = 2.01$ และ $x_0 = 0.98$

3. เส้นคือ $0.98 + 2.01t$ ค่าตกค้างคือ $r = (0.02,\ -0.09,\ 0.20,\ -0.21,\ 0.08)$ รวมกันได้ 0 และ $\sum t_i r_i = 0 - 0.09 + 0.40 - 0.63 + 0.32 = 0$ ตามที่ทฤษฎีบทสมการปรกติบอก $\sum r_i^2 = 0.0004 + 0.0081 + 0.04 + 0.0441 + 0.0064 = 0.099$ และ $\|\mathbf{r}\|_2 = \sqrt{0.099} \approx 0.3146$

ตรวจด้วยโค้ด: ใน `test_ch08.py`(`test_exercise_816_five_points_line_fit`)

เฉลย 8.1.7.

1. ถ้าค่าคลาดเคลื่อนของระยะ $\boldsymbol{\varepsilon}$ มีความแปรปรวน $\sigma^2 I$ และเราประมาณเชิงเส้นรอบตำแหน่งจริง จะได้ $\mathbf{x} - \mathbf{x}_{\text{true}} \approx (J^TJ)^{-1}J^T\boldsymbol{\varepsilon}$ ความแปรปรวนของ $\mathbf{x}$ จึงเป็น $\sigma^2(J^TJ)^{-1}$ ผลรวมของความแปรปรวนของแต่ละตัวไม่ทราบคือ $\sigma^2 \operatorname{tr}\big((J^TJ)^{-1}\big)$ และขนาดของความคลาดเคลื่อนรวมจึงประมาณได้ด้วย $\sigma\sqrt{\operatorname{tr}\big((J^TJ)^{-1}\big)} = \sigma d$ ตามที่กำหนดในเกร็ดเพิ่มเติม (นี่เป็นการประมาณเชิงเส้น ไม่ใช่สูตรที่แน่นอน)

2. กรณีดี $\sigma d \approx 2 \times 3.62 = 7.2$ เมตร กรณีกระจุก $\sigma d \approx 2 \times 2696 = 5392$ เมตร ต่างกันราว 745 เท่า $(2696/3.62 \approx 745)$

3. ดาวเทียมที่กระจุกกันทำให้ทิศของแถวใน J เกือบเหมือนกัน คอลัมน์ของตำแหน่งจึงเกือบพึ่งพากัน และคอลัมน์ของค่าเวลาก็พึ่งพาได้เกือบเต็มที่ ค่าลักษณะเฉพาะที่เล็กที่สุดของ J^TJ จึงเล็กมาก ค่าตกค้างอาจยังเล็กได้เพราะสมการที่เกือบเหมือนกันทุกสมการเห็นด้วยกันได้ดี แต่ตำแหน่งที่เลื่อนไปตามทิศที่ไม่ค่อยมีข้อมูลจะไม่ทำให้ค่าตกค้างเปลี่ยนมากนัก

ตรวจด้วยโค้ด: ค่า 3.6188 และ 2696.2 ทดสอบใน `test_ch08.py`(`test_dilution_good_and_clustered_geometry`) และ `test_exercise_817_clustered_geometry_rank_still_full` ตรวจว่า J ยังมีลำดับชั้น 4 แต่คอนดิชันนัมเบอร์ใหญ่

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 8.1.8.

1. $\frac{d}{d\theta}R(\theta)\mathbf{P}_1 = R'(\theta)\mathbf{0} = \mathbf{0}$ จึงคอลัมน์แรกของ J เป็นศูนย์ ส่วนคอลัมน์ของการเลื่อนเป็น $(1,0)^T$ และ $(0,1)^T$ สำหรับจุดเดียว J มีลำดับชั้น 2 เพราะมีเพียงสองคอลัมน์ที่เป็นอิสระ จุดเดียวที่อยู่ที่จุดหมุนจึงบอกไม่ได้ว่าหมุนไปเท่าไร

2. เมื่อจุดทั้งสองไม่ตรงกับจุดหมุน คอลัมน์ของ θ คือ $R'(\theta)\mathbf{P}_k$ ซึ่งไม่เป็นศูนย์ และไม่ขนานกับคอลัมน์การเลื่อน จึงได้ลำดับชั้น 3 ข้อมูลไม่มีสัญญาณรบกวนมีค่าตรงกับ $f(\theta, \mathbf{t})$ ดังนั้นการทำซ้ำเกาส์–นิวตันจากจุดเริ่ม $(0,0,0)$ ลู่เข้าหาค่าจริง $(0.6,\ 1.5,\ -0.5)$ ได้จนค่าคลาดเคลื่อนไม่เกิน 10^{-9} ตามที่ทดสอบ

ตรวจ ด้วย โค้ด: `test_pose_single_landmark_at_origin_has_no_angle_information` ใน `test_ch08.py`

ตรวจด้วยโค้ด: `test_pose_two_landmarks_fit_is_exact_without_noise` ใน `test_ch08.py`

อ่านต่อ. ตำรา Atkinson 1989, §9.3, pp. 609--618 ให้การสะท้อนและการแยก QR ที่ใช้ในการคำนวณค่าลักษณะเฉพาะ และ Atkinson 1989, §9.7, pp. 633--645 อธิบายปัญหากำลังสองน้อยสุดโดยใช้การแยกค่าเอกฐาน (SVD) ซึ่งเป็นเครื่องมือของหัวข้อ 8.2 และให้ทฤษฎีบท 9.7 ที่ใช้กับเมทริกซ์ที่ลำดับชั้นไม่เต็มด้วย เมื่อ A มีลำดับชั้นไม่เต็ม คำตอบไม่มีเพียงคำตอบเดียว หัวข้อ 8.2 จะกลับมาที่ปัญหานี้

8.2 ข้อมูลที่ซ้ำซ้อนกัน

สถานการณ์สมมติ: อีกด้านหนึ่งของปัญหาเดียวกัน

ลองนึกถึงภาพขาวดำบนจอ ภาพขนาด 128×128 จุดคือตารางค่าความสว่าง $16\,384$ ค่า ถ้าเก็บทุกค่า ภาพก็กลับมาครบ แต่ภาพที่เราเห็นบนจอมักไม่ได้มีข้อมูลอิสระมากขนาดนั้น พื้นที่ที่สว่างขึ้นอย่างค่อยเป็นค่อยไป และขอบวัตถุที่เรียบทำให้ตัวเลขหลายตัวบอกเรื่องเดียวกันซ้ำ ๆ

ลองถามว่า ถ้าเราเก็บไว้เพียงรูปแบบใหญ่สุดไม่กี่รูป ภาพที่กลับมายังจำได้หรือไม่ และต้องเก็บตัวเลขเท่าไร ภาพในบทนี้เป็นภาพสมมติที่เราสร้างเอง มีพื้นหลังไล่ระดับ วงกลม แถบเฉียง จุดสว่างนุ่ม ๆ และสัญญาณรบกวนเล็กน้อย ไม่ใช่รูปถ่ายของใคร และไม่ได้บอกว่าโปรแกรมใดบีบอัดภาพอย่างไร

หัวข้อ 8.1 มีสมการมากกว่าตัวไม่ทราบ และเราเลือกคำตอบที่ใกล้ข้อมูลที่สุด หัวข้อนี้เป็นอีกด้านของเรื่องเดียวกัน คือ ข้อมูลมีมากกว่าสิ่งที่อยู่เบื้องหลัง เราจึงเก็บเฉพาะส่วนที่สำคัญให้ใกล้ข้อมูลเดิมที่สุด

ภาพรวมก่อนอ่าน

ทุกเมทริกซ์แยกได้เป็นผลบวกของชั้น เรียงจากชั้นที่แรงที่สุดไปชั้นที่อ่อนที่สุด แต่ละชั้นเป็นผลคูณของเวกเตอร์หนึ่งคู่ ค่าเอกฐานบอกความแรงของแต่ละชั้น เมื่อเราตัดชั้นอ่อน ๆ ทิ้ง ภาพที่เหลือยังใกล้ภาพเดิม และเครื่องมือเดียวกันนี้ยังช่วยหาคำตอบกำลังสองน้อยสุดเมื่อระบบไม่มีลำดับชั้นเต็มได้ด้วย

ฉบับร่าง. Preprint · 2026-10-23

ชั้นที่ง่ายที่สุด: ลำดับชั้นหนึ่ง

ลองดูภาพที่ง่ายที่สุด ถ้าความสว่างที่แถว i และคอลัมน์ j เท่ากับ $a_i b_j$ โดย a_i ขึ้นกับแถว และ b_j ขึ้นกับคอลัมน์ ภาพทั้งภาพคือเมทริกซ์ $\mathbf{a}\mathbf{b}^T$ ภาพแบบนี้เก็บได้ด้วยตัวเลข $m+n$ ตัว แทนที่จะเป็น mn ตัว เพราะทั้งภาพถูกกำหนดด้วยความสว่างแถวหนึ่งกับความสว่างคอลัมน์หนึ่ง

นิยาม 8.2.1: ลำดับชั้นของเมทริกซ์

จำนวนคอลัมน์ที่อิสระเชิงเส้นต่อกันได้มากที่สุดของเมทริกซ์ A เรียกว่า **ลำดับชั้นของเมทริกซ์** (rank) และเขียนเป็น $\operatorname{rank} A$ (จำนวนแถวที่อิสระได้มากที่สุดเท่ากันเสมอ) เมทริกซ์ที่มีลำดับชั้นหนึ่งเรียกว่า **เมทริกซ์ลำดับชั้นหนึ่ง** (rank-one matrix)

อ่านแบบเข้าใจง่าย. เมทริกซ์ A มีลำดับชั้นหนึ่ง ก็ต่อเมื่อ $A = \mathbf{a}\mathbf{b}^T$ สำหรับเวกเตอร์ $\mathbf{a}$ และ $\mathbf{b}$ ที่ไม่เป็นศูนย์ คอลัมน์ทุกคอลัมน์จึงเป็นพหุคูณของเวกเตอร์เดียวกัน ภาพที่มีขอบคมหรือสัญญาณรบกวนมักมีลำดับชั้นเต็ม เพราะต้องใช้หลายรูปแบบซ้อนกัน

ลองเอง

จงตรวจว่า $\begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix}$ มีลำดับชั้นหนึ่ง โดยเขียนเป็น $\mathbf{a}\mathbf{b}^T$ ด้วย $\mathbf{a} = (2,1)^T$ และ $\mathbf{b} = (1,2)^T$ ผลที่ควรได้คือ $\mathbf{a}\mathbf{b}^T$ เท่ากับเมทริกซ์เดิม และคอลัมน์ที่สองเป็นสองเท่าของคอลัมน์แรก

การแยกค่าเอกฐาน: หมุน ยืด แล้วหมุน

สำหรับเมทริกซ์จริงขนาด 2×2 ภาพของวงกลมหนึ่งหน่วยภายใต้เมทริกซ์นั้นเป็นวงรี (บางครั้งยุบเป็นส่วนของเส้นตรง) แกนยาวและแกนสั้นของวงรีตั้งฉากกัน และความยาวของครึ่งแกนคือ **ค่าเอกฐาน** (singular value) ของเมทริกซ์ การแยกค่าเอกฐานบอกว่า ทุกเมทริกซ์ทำงานได้เหมือนกัน คือหมุน แล้วยืดตามแกนฉากสองสามแกน แล้วหมุนอีกครั้ง

นิยาม 8.2.2: การแยกค่าเอกฐาน

ให้ $A \in \mathbb{R}^{m \times n}$ และ $p = \min(m, n)$ การเขียน

$$A = U \Sigma V^T$$

โดยที่ $U \in \mathbb{R}^{m \times m}$ และ $V \in \mathbb{R}^{n \times n}$ เป็นเมทริกซ์เชิงตั้งฉาก และ Σ เป็นเมทริกซ์ขนาด $m \times n$ ที่สมาชิกนอกแนวทแยงเป็นศูนย์ และสมาชิกในแนวทแยงเป็น $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ เรียกว่า **การแยกค่าเอกฐาน** (singular value decomposition) ของ A โดยที่ σ_i คือค่าเอกฐานของ A และคอลัมน์ $\mathbf{u}_i$ ของ U กับ $\mathbf{v}_i$ ของ V เป็นคู่ที่ตรงกับค่าเอกฐานนั้น

ทฤษฎีบท 8.2.1: การมีอยู่ของการแยกค่าเอกฐาน

ทุกเมทริกซ์จริง A มีการแยกค่าเอกฐานเสมอ และจำนวนค่าเอกฐานที่ไม่เป็นศูนย์เท่ากับ $\operatorname{rank} A$

พิสูจน์. **แนวคิด.** ตำรา Atkinson 1989, Thm 7.5, pp. 478--479 ใช้ค่าลักษณะเฉพาะของ $A^T A$ เราทำตามแนวเดียวกันและเขียนขั้นตอนเอง ขั้นที่ 1 ยอมรับจากทฤษฎีบทสเปกตรัมของเมทริกซ์สมมาตร ซึ่งเป็นความรู้จากพีชคณิตเชิงเส้น ส่วนที่ต้องระวังคือคอลัมน์ของ U ที่ค่าเอกฐานเป็นศูนย์

ฉบับร่าง · Preprint · 2026-10-23

1. **ค่าลักษณะเฉพาะไม่ติดลบ และจำนวนที่เป็นบวก.** A^TA เป็นเมทริกซ์สมมาตร และ $\mathbf{x}^TA^TA\mathbf{x} = \|A\mathbf{x}\|_2^2 \geq 0$ ดังนั้นมีเมทริกซ์เชิงตั้งฉาก V ที่ $A^TA = V\,\mathrm{diag}(\lambda_1,\ldots,\lambda_n)V^T$ โดย $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ ให้ r เป็นจำนวน λ_i ที่เป็นบวก จึง $\lambda_i > 0$ ก็ต่อเมื่อ $i \leq r$ เพราะ $A^TA\mathbf{x} = \mathbf{0}$ ก็ต่อเมื่อ $A\mathbf{x} = \mathbf{0}$ จึงได้ $r = \mathrm{rank}\,A^TA = \mathrm{rank}\,A \leq p$

2. **ได้ σ และ $\mathbf{u}_1,\ldots,\mathbf{u}_r$.** ให้ $\sigma_i = \sqrt{\lambda_i}$ สำหรับ $i \leq p$ และ $\sigma_i = 0$ สำหรับ $i > p$ จึงได้ $\sigma_i > 0$ ก็ต่อเมื่อ $i \leq r$ เพราะ $\|A\mathbf{v}_i\|_2^2 = \lambda_i$ จะได้ $A\mathbf{v}_i = \mathbf{0}$ เมื่อ $i > r$ สำหรับ $i \leq r$ ให้ $\mathbf{u}_i = A\mathbf{v}_i/\sigma_i$ จะได้ $(A\mathbf{v}_i)^T(A\mathbf{v}_j) = \lambda_j\delta_{ij}$ จึง $\mathbf{u}_1,\ldots,\mathbf{u}_r$ ยาวหนึ่งหน่วยและตั้งฉากกันเป็นคู่

3. **เลือก $\mathbf{u}_{r+1},\ldots,\mathbf{u}_m$ ให้ U เชิงตั้งฉาก.** ถ้า $r < m$ เลือก $\mathbf{u}_{r+1}$ เป็นเวกเตอร์หน่วยที่ตั้งฉากกับ $\mathbf{u}_1,\ldots,\mathbf{u}_r$ แล้วเลือกตัวถัดไปให้ตั้งฉากกับทุกตัวที่เลือกมาแล้ว ทำได้เสมอ เพราะเวกเตอร์ $k < m$ ตัวสร้างปริภูมิมิติ k ส่วนที่ตั้งฉากกับปริภูมินั้นมีมิติ $m - k \geq 1$ ผลคือ U เป็นเมทริกซ์เชิงตั้งฉาก การเลือกนี้ไม่ได้มีวิธีเดียว

4. **ตรวจว่า $AV = U\Sigma$.** ดูทีละคอลัมน์: เมื่อ $j \leq r$ ทั้งสองฝั่งเป็น $\sigma_j\mathbf{u}_j$ เมื่อ $j > r$ ทั้งสองฝั่งเป็นศูนย์ เพราะ $A\mathbf{v}_j = \mathbf{0}$ และ $\sigma_j = 0$ (คอลัมน์ $j > m$ ของ $U\Sigma$ ก็เป็นศูนย์ และ $j > m \geq r$) ดังนั้น $A = U\Sigma V^T$ เพราะ V เชิงตั้งฉาก

5. **ลำดับชั้น.** เพราะ U และ V ไม่เอกฐาน $\mathrm{rank}\,A = \mathrm{rank}\,\Sigma = r$ ตามที่ขั้นที่ 1 และ 2 แสดงไว้

สรุป: การแยกค่าเอกฐานมาจากค่าลักษณะเฉพาะของ A^TA เราเขียนตามแนวของตำรา ยกเว้นขั้นที่ 1 ที่เรายอมรับ และการเลือกคอลัมน์ของ U ที่เหลือซึ่งเราเขียนเอง □

ลองเอง: ตัวอย่าง 2×2

ให้ $A = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}$ จงคำนวณ A^TA แล้วหาค่าลักษณะเฉพาะ ผลที่ควรได้คือ $A^TA = \begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix}$ มีค่าลักษณะเฉพาะ 45 และ 5 จึง $\sigma_1 = 3\sqrt{5}$ และ $\sigma_2 = \sqrt{5}$ และ $\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1,1)^T$, $\mathbf{u}_1 = \frac{1}{\sqrt{10}}(1,3)^T$, $\mathbf{v}_2 = \frac{1}{\sqrt{2}}(1,-1)^T$, $\mathbf{u}_2 = \frac{1}{\sqrt{10}}(3,-1)^T$ จงตรวจว่า $A = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T$

อ่านแบบเข้าใจง่าย. ค่าเอกฐานทั้งสองคือ 6.708 และ 2.236 โดยประมาณ ชั้นที่แรงกว่าคือ $\sigma_1\mathbf{u}_1\mathbf{v}_1^T$ เครื่องหมายของคู่ $\mathbf{u}_i$ กับ $\mathbf{v}_i$ เปลี่ยนพร้อมกันได้โดยผลคูณไม่เปลี่ยน ตัวเลขเหล่านี้เราคำนวณด้วยมือและตรงกับโค้ดของเรา (`test_ch08_svd.py`)

เก็บชั้นแรกไม่กี่ชั้น

ถ้าเราต้องการเมทริกซ์ที่เก็บได้เพียง k ชั้น ควรเก็บชั้นไหน ให้ $A_k = \sum_{i=1}^{k}\sigma_i\mathbf{u}_i\mathbf{v}_i^T$ ซึ่งเก็บชั้นแรก k ชั้น ทฤษฎีบทต่อไปบอกว่า ไม่มีเมทริกซ์ที่มีลำดับชั้นไม่เกิน k ใดที่ใกล้ A มากกว่า A_k เมื่อวัดความต่างด้วย **นอร์มสเปกตรัม** (spectral norm) $\|B\|_2 = \max_{\mathbf{x}\neq\mathbf{0}}\|B\mathbf{x}\|_2/\|\mathbf{x}\|_2$

ทฤษฎีบท 8.2.2: การประมาณลำดับชั้นต่ำที่ดีที่สุด

ให้ A มีค่าเอกฐาน $\sigma_1 \geq \cdots \geq \sigma_p$ โดยตั้ง $\sigma_i = 0$ เมื่อ $i > p$ และให้ $0 \leq k < p$ แล้ว $\mathrm{rank}\,A_k \leq k$ และ $\|A - A_k\|_2 = \sigma_{k+1}$ และถ้า $B \in \mathbb{R}^{m\times n}$ มี $\mathrm{rank}\,B \leq k$ แล้ว $\|A - B\|_2 \geq \sigma_{k+1}$

พิสูจน์. **แนวคิด.** ส่วนแรกคำนวณตรงจากการแยก ส่วนที่สองใช้ว่าเคอร์เนลของ B มีมิติใหญ่พอ จนตัดกับปริภูมิที่สร้างจาก $\mathbf{v}_1,\ldots,\mathbf{v}_{k+1}$ ได้

ฉบับร่าง · Preprint · 2026-10-20

1. **ส่วนแรก.** $A - A_k = \sum_{i>k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ และ $\mathbf{u}_i$ ตั้งฉากกัน จึง $\|(A - A_k)\mathbf{x}\|_2^2 = \sum_{i>k} \sigma_i^2 (\mathbf{v}_i^T \mathbf{x})^2 \leq \sigma_{k+1}^2 \sum_i (\mathbf{v}_i^T \mathbf{x})^2 \leq \sigma_{k+1}^2 \|\mathbf{x}\|_2^2$ โดยใช้ $\sigma_i \leq \sigma_{k+1}$ เมื่อ $i > k$ และ V เชิงตั้งฉาก ที่ $\mathbf{x} = \mathbf{v}_{k+1}$ ทุกความไม่เท่ากันเป็นสมการ จึงได้ $\|A - A_k\|_2 = \sigma_{k+1}$ ส่วน $\operatorname{rank} A_k \leq k$ เพราะ A_k เป็นผลบวกของ k เมทริกซ์ลำดับชั้นหนึ่ง

2. **ส่วนที่สอง.** สมมติ $\operatorname{rank} B \leq k$ เคอร์เนลของ B มีมิติอย่างน้อย $n - k$ และปริภูมิ $W = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$ มีมิติ $k+1$ เพราะ $(n-k) + (k+1) = n + 1 > n$ จึงมีเวกเตอร์ $\mathbf{z} \neq \mathbf{0}$ ใน $\ker B \cap W$ ปรับให้ $\|\mathbf{z}\|_2 = 1$ ได้ $\mathbf{z} = \sum_{i \leq k+1} c_i \mathbf{v}_i$ โดย $\sum_i c_i^2 = 1$

3. **ประเมินที่ $\mathbf{z}$.** เพราะ $B\mathbf{z} = \mathbf{0}$ จึง $(A-B)\mathbf{z} = A\mathbf{z} = \sum_{i \leq k+1} c_i \sigma_i \mathbf{u}_i$ และ $\|A\mathbf{z}\|_2^2 = \sum_{i \leq k+1} \sigma_i^2 c_i^2 \geq \sigma_{k+1}^2$ ดังนั้น $\|A - B\|_2 \geq \|(A - B)\mathbf{z}\|_2 \geq \sigma_{k+1}$

สรุป: A_k ดีที่สุดในนอร์มสเปกตรัม และบทพิสูจน์นี้เราเขียนเองทั้งหมด □

อ่านแบบเข้าใจง่าย

ทฤษฎีบทบอกว่า ถ้าจำกัดให้มีได้เพียง k ชั้น การเก็บ k ชั้นแรกทำได้ดีที่สุดเมื่อวัดด้วยนอร์มสเปกตรัม ค่าคลาดเคลื่อนสัมบูรณ์ที่เหลือคือค่าเอกฐานชั้นถัดไป ถ้าชั้นถัดไปเล็กมาก การตัดทิ้งก็ไม่เสียมาก สำหรับนอร์มโฟรเบเนียส ค่าคลาดเคลื่อนสัมบูรณ์ของ A_k คือ $\sqrt{\sum_{i>k} \sigma_i^2}$ สูตรนี้เห็นได้จากความตั้งฉาก เพราะการคูณด้วยเมทริกซ์เชิงตั้งฉากไม่เปลี่ยนนอร์มโฟรเบเนียส

ตัวผกผันเทียม: กำลังสองน้อยสุดเมื่อลำดับชั้นไม่เต็ม

ในหัวข้อ 8.1 เราแก้กำลังสองน้อยสุดด้วย QR เมื่อ A มีลำดับชั้นเต็มในแนวคอลัมน์ ถ้าลำดับชั้นไม่เต็ม คำตอบมีได้หลายคำตอบ ตำรา Atkinson 1989, §9.7, pp. 634--636 ให้คำตอบทั่วไปด้วย SVD ผู้อ่านที่ต้องการคำตอบเดียวมักเลือกคำตอบที่มีนอร์มเล็กที่สุด เราจะนิยามและพิสูจน์เอง

นิยาม 8.2.3: ตัวผกผันเทียม

ให้ $A = U\Sigma V^T$ เป็นการแยกค่าเอกฐาน ให้ Σ^+ เป็นเมทริกซ์ขนาด $n \times m$ ที่สมาชิกแนวทแยงเป็น $1/\sigma_i$ เมื่อ $\sigma_i > 0$ และเป็นศูนย์ในกรณีอื่น เรียกว่า **ตัวผกผันเทียม** (pseudo-inverse) ของ A คือ $A^+ = V\Sigma^+ U^T$

ทฤษฎีบท 8.2.3: คำตอบกำลังสองน้อยสุดที่มีนอร์มเล็กที่สุด

ให้ $A \in \mathbb{R}^{m \times n}$ และ $\mathbf{b} \in \mathbb{R}^m$ แล้ว $\mathbf{x}^+ = A^+\mathbf{b}$ เป็นคำตอบกำลังสองน้อยสุดของ $\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|_2$ และในบรรดาคำตอบกำลังสองน้อยสุดทั้งหมด $\mathbf{x}^+$ เป็นเพียงคำตอบเดียวที่มีนอร์ม $\|\mathbf{x}\|_2$ เล็กที่สุด

พิสูจน์. **แนวคิด.** การคูณด้วย U^T และ V^T ไม่เปลี่ยนนอร์ม ปัญหาเดิมจึงกลายเป็นปัญหาของเมทริกซ์แนวทแยงที่แก้ได้ทีละตัว

1. **แปลงตัวแปร.** ให้ $\mathbf{c} = U^T\mathbf{b}$ และ $\mathbf{y} = V^T\mathbf{x}$ เพราะ U^T เชิงตั้งฉาก $\|\mathbf{b} - A\mathbf{x}\|_2 = \|\mathbf{c} - \Sigma\mathbf{y}\|_2$ และ $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$

2. **แยกเป็นส่วน.** ให้ $r = \operatorname{rank} A$ จะได้ $(\Sigma\mathbf{y})_i = \sigma_i y_i$ เมื่อ $i \leq r$ และเป็นศูนย์เมื่อ $i > r$ จึง $\|\mathbf{c} - \Sigma\mathbf{y}\|_2^2 = \sum_{i \leq r} (c_i - \sigma_i y_i)^2 + \sum_{i>r} c_i^2$

3. **ค่าต่ำสุด.** ผลบวกหลังไม่ขึ้นกับ $\mathbf{y}$ ผลบวกแรกเป็นศูนย์เมื่อ $y_i = c_i/\sigma_i$ สำหรับ $i \leq r$ และส่วนประกอบ $y_{r+1}, \ldots, y_n$ เลือกได้อย่างอิสระ ค่าตกค้างต่ำสุดจึงเป็น $\sum_{i>r} c_i^2$

ฉบับร่าง Preprint · 2026-10-23

4. **นอร์มเล็กที่สุด.** สำหรับ $\mathbf{y}$ ที่ทำให้ค่าตกค้างต่ำสุด $\|\mathbf{y}\|_2^2 = \sum_{i\le r}(c_i/\sigma_i)^2 + \sum_{i>r} y_i^2$ เล็กที่สุดเมื่อ $y_i = 0$ สำหรับ $i > r$ ซึ่งให้ $\mathbf{y} = \Sigma^+\mathbf{c}$ และ $\mathbf{x} = V\mathbf{y} = V\Sigma^+U^T\mathbf{b} = A^+\mathbf{b}$ โดยคำตอบนี้มีเพียงคำตอบเดียว

สรุป: $A^+\mathbf{b}$ เป็นคำตอบที่สั้นที่สุดในกลุ่มคำตอบกำลังสองน้อยสุด เมื่อลำดับชั้นเต็มคำตอบมีเพียงคำตอบเดียว และตรงกับ QR ของหัวข้อ 8.1 บทพิสูจน์นี้เราเขียนเอง □

อ่านแบบเข้าใจง่าย

ให้ $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ และ $\mathbf{b} = (2, 0)^T$ ระบบนี้ไม่มีคำตอบตรง เพราะแถวทั้งสองเหมือนกัน ค่าตกค้างเล็กที่สุดเกิดเมื่อ $x_1 + x_2 = 1$ ซึ่งมีคำตอบ $(t, 1-t)$ ทุก t และค่าตกค้างคือ $\sqrt{2}$ จากการแยกค่าเอกฐานได้ $A^+ = \frac{1}{4}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ และ $\mathbf{x}^+ = (\frac{1}{2}, \frac{1}{2})^T$ ซึ่งมีนอร์ม $1/\sqrt{2}$ น้อยกว่า $(1, 0)$ และ $(0, 1)$ ที่ให้ค่าตกค้างเท่ากัน

ค่าเอกฐานเล็กมากแต่ไม่เป็นศูนย์

ข้อควรระวัง: ตัดลำดับชั้นก่อนดูค่าเอกฐานเล็ก

ที่ $\varepsilon = 10^{-9}$ เมทริกซ์ $A = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$ จากหัวข้อ 8.1 มีค่าเอกฐาน $\sigma_1 = \sqrt{2+\varepsilon^2} \approx 1.41421$ และ $\sigma_2 = \varepsilon = 10^{-9}$ พอดี (เห็นได้จาก $A^TA = \begin{pmatrix} 1+\varepsilon^2 & 1 \\ 1 & 1+\varepsilon^2 \end{pmatrix}$ ซึ่งมีค่าลักษณะเฉพาะ $2+\varepsilon^2$ และ ε^2) สมการปรกติทำให้ $\varepsilon^2 = 10^{-18}$ หายไปในเลขคอมพิวเตอร์ แต่ SVD เห็น σ_2 โดยตรง
ถ้า $\mathbf{b} = (2, 0, 0)^T$ คำตอบเต็มลำดับชั้นคือ $\approx (1, 1)$ แต่ถ้า $\mathbf{b} = (2, 0, 10^{-9})^T$ ซึ่งต่างกันเพียง 10^{-9} คำตอบเต็มลำดับชั้นเปลี่ยนเป็น $(0.5,\ 1.5)$ ค่าตกค้างของทั้งสองคำตอบมีขนาดราว 10^{-9} จึงบอกไม่ได้ว่าคำตอบไหนถูก ถ้าเราตัดค่าเอกฐานที่เล็กกว่า $10^{-8}\sigma_1$ ทิ้ง เราจะได้ $(1, 1)$ ทั้งสองกรณี แต่ความเสถียรนี้แลกกับการทิ้ง σ_2 ที่ไม่เป็นศูนย์ ต้องรู้ว่ากำลังทำแบบนั้น

เกร็ดเพิ่มเติม

สำหรับเมทริกซ์สมมาตรกึ่งบวกแน่นอน $A = V\Lambda V^T$ โดย $\lambda_i \ge 0$ เราได้ $A^TA = A^2 = V\Lambda^2V^T$ จึง $\sigma_i = \lambda_i$ ค่าเอกฐานของเมทริกซ์ประเภทนี้จึงเป็นค่าลักษณะเฉพาะเอง จุดเชื่อมนี้จะใช้ต่อในบทที่ 9

ภาพสมมติของเรา: เก็บชั้นแรกไม่กี่ชั้น

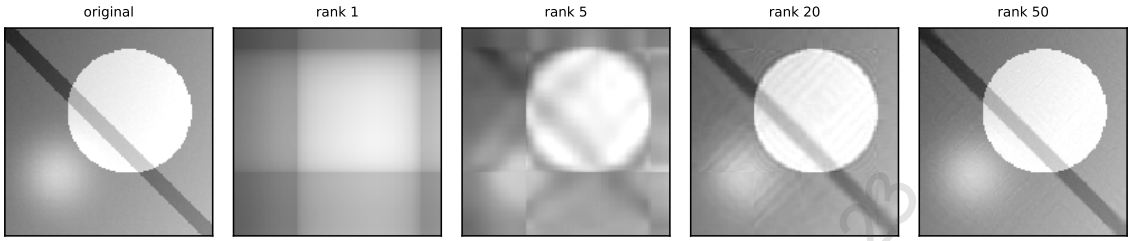
กลับไปที่ภาพของเรา ภาพขนาด 128×128 มีค่าเอกฐาน $\sigma_1 \approx 87.70$ และ $\sigma_2 \approx 11.95$ ค่าที่ตามมาลดลงเร็วในตอนแรกและช้าลงเมื่อ k โต ภาพนี้มีลำดับชั้นเต็ม 128 ค่าเอกฐานทุกตัวไม่เป็นศูนย์ แม้ค่าท้ายสุดจะเล็กมากราว 10^{-3} สิ่งที่ทำให้ภาพกลับมาจำได้คือการเก็บเพียงชั้นแรกไม่กี่ชั้น ตาราง 8.4 แสดงจำนวนตัวเลขที่เก็บ และค่าคลาดเคลื่อนสัมพัทธ์ในนอร์มโฟรเบเนียส $\sqrt{\sum_{i>k}\sigma_i^2}/\sqrt{\sum_i \sigma_i^2}$

ที่ $k = 1$ ภาพเป็นบล็อกไล่ระดับ ที่ $k = 5$ วงกลมชัดขึ้น ที่ $k = 20$ วงกลมและแถบเฉียงคมพอจำได้ แต่ยังมีลายเล็ก ๆ ที่ไม่มีในภาพเดิม และที่ $k = 50$ ภาพใกล้ต้นฉบับมาก (รูป 8.6) ค่าคลาดเคลื่อนสัมพัทธ์ลดลงเร็วในตอนแรกแล้วช้าลง (รูป 8.7)

ร่างต้น · Preprint · 2026-10-23

ตารางที่ 8.4: ภาพของเราที่เก็บ k ชั้นแรก ($m = n = 128$ และ $mn = 16\,384$ ตัวเลขจากโค้ดของเรา)

k	ตัวเลขที่เก็บ $k(m+n+1)$	เทียบกับ mn (%)	ค่าคลาดเคลื่อนสัมพัทธ์
1	257	1.6	0.178
5	1 285	7.8	0.080
20	5 140	31.4	0.037
50	12 850	78.4	0.017



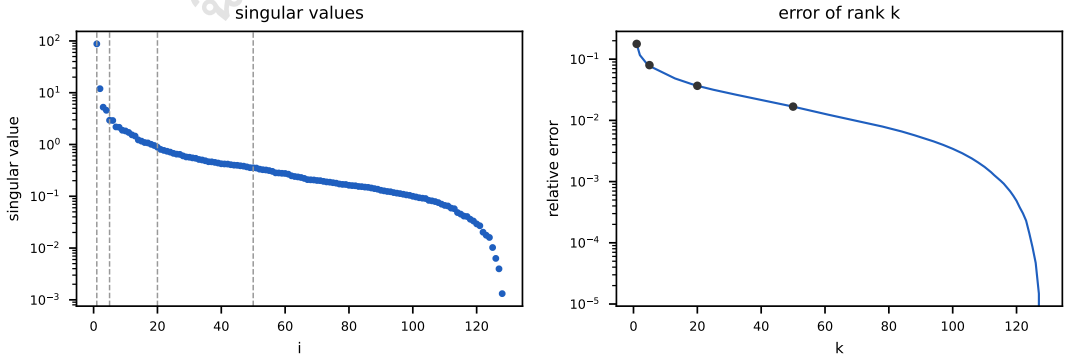
รูปที่ 8.6: ภาพสมมติ 128×128 และการประมาณด้วย $k = 1, 5, 20, 50$ ชั้นแรก (สร้างจาก `numanal/ch08_svd.py`)

สรุปใจความ

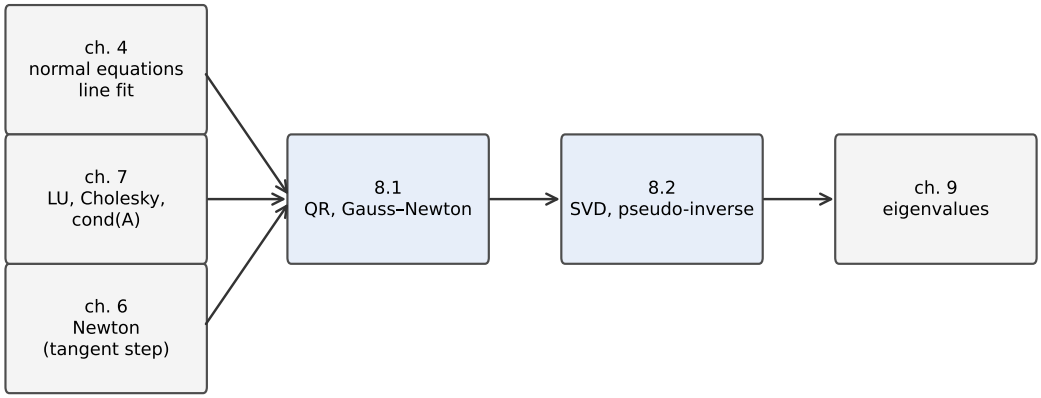
- ทุกเมทริกซ์แยกเป็นผลบวกของชั้นลำดับชั้นหนึ่ง เรียงตามค่าเอกฐานจากมากไปน้อย จำนวนค่าเอกฐานที่ไม่เป็นศูนย์คือลำดับชั้นของเมทริกซ์
- เก็บ k ชั้นแรกได้เมทริกซ์ที่ดีที่สุดในนอร์มสเปกตรัม ค่าคลาดเคลื่อนสัมบูรณ์คือ σ_{k+1} ภาพของเราเก็บ 5 ชั้นได้ราว 8 เปอร์เซ็นต์ของตัวเลข และได้ค่าคลาดเคลื่อนสัมพัทธ์ราว 0.08
- ตัวผกผันเทียม $A^{+}\mathbf{b}$ ให้คำตอบกำลังสองน้อยสุดที่มีนอร์มเล็กที่สุด ใช้ได้แม้ลำดับชั้นไม่เต็ม
- ค่าเอกฐานเล็กมากแต่ไม่เป็นศูนย์ต้องตัดสินใจก่อนว่าจะตัดหรือไม่ มิฉะนั้นคำตอบเปลี่ยนมาก

อ่านต่อ. ตำรา Atkinson 1989, Thm 7.5, pp. 478--479 ให้การแยกค่าเอกฐานพร้อมบทพิสูจน์ผ่านค่าลักษณะเฉพาะของ $A^{T}A$ และ Atkinson 1989, §9.7, pp. 634--636 ใช้ SVD กับปัญหากำลังสองน้อยสุดและนิยาม A^{+} ส่วนนอร์มโฟรเบเนียส

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 8.7: ซ้าย: ค่าเอกฐานของภาพบนสเกลลอการิทึม เส้นประคือ $k = 1, 5, 20, 50$ ขวา: ค่าคลาดเคลื่อนสัมพัทธ์ของ A_k จุดคือค่าที่ใช้ในตาราง (สร้างจาก `numanal/ch08_svd.py`)



รูปที่ 8.8: แผนผังบทที่ 8 ทั้งบท: ลำดับของหัวข้อ 8.1 และ 8.2 สิ่งที่แต่ละหัวข้อให้หัวข้อถัดไป และสิ่งที่บทที่ 4, 6 และ 7 ให้บทนี้ เส้นจากหัวข้อ 8.2 ไปบทที่ 9 คือค่าเอกฐานกับค่าลักษณะเฉพาะ (สร้างจาก `numanal/ch08_svd.py`)

ของการประมาณลำดับชั้นต่ำ เราไม่ได้พิสูจน์ในเล่มนี้

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 8.2.1 (⋆) ให้ $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{pmatrix}$

1. จงเขียน $A = \mathbf{a}\mathbf{b}^T$ โดยมี $\mathbf{a} = (1, 2, 3)^T$ และหา $\mathbf{b}$ แล้วบอกลำดับชั้นของ A
2. จงคำนวณ $A^T A$ และค่าลักษณะเฉพาะ แล้วแสดงว่า $\sigma_1 = \sqrt{70}$ และ $\sigma_2 = 0$

แบบฝึกหัด 8.2.2 (⋆) ให้ $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$

1. จงแสดงว่า $\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1, 1)^T$ และ $\mathbf{v}_2 = \frac{1}{\sqrt{2}}(1, -1)^T$ เป็นเวกเตอร์ลักษณะเฉพาะของ A และหาค่าลักษณะเฉพาะที่สอดคล้องกัน
2. จงหาค่าเอกฐาน σ_1, σ_2 และ $\mathbf{u}_i = A\mathbf{v}_i/\sigma_i$ แล้วตรวจว่า $A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T$

แบบฝึกหัด 8.2.3 (⋆⋆) ให้ $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ และ $\mathbf{b} = (4, 0)^T$

1. จงหา A^+ จากการแยกค่าเอกฐาน แล้วคำนวณ $\mathbf{x}^+ = A^+ \mathbf{b}$
2. จงแสดงว่า $\mathbf{x} = (2, 0)^T$ และ $\mathbf{x} = (0, 2)^T$ ให้ค่าตกค้างเท่ากับค่าที่ $\mathbf{x}^+$ ให้ แต่มีนอร์มมากกว่า $\|\mathbf{x}^+\|_2$
3. จงอธิบายว่าทำไมสมการปรกติ $A^T A\mathbf{x} = A^T \mathbf{b}$ ในข้อนี้ไม่ให้คำตอบเพียงคำตอบเดียว

แบบฝึกหัด 8.2.4 (⋆⋆) ให้ $A = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$ เมื่อ $\varepsilon > 0$

ฉบับร่าง · Preprint · 2026-10-23

1. จงแสดงว่า $\sigma_1 = \sqrt{2+\varepsilon^2}$ และ $\sigma_2 = \varepsilon$
2. สำหรับ $\mathbf{b} = (2,\, 0,\, \delta)^T$ จงแสดงว่าคำตอบกำลังสองน้อยสุดเพียงคำตอบเดียวคือ $x_1 = \dfrac{2-\delta/\varepsilon}{2+\varepsilon^2}$ และ $x_2 = x_1 + \delta/\varepsilon$
3. ที่ $\varepsilon = 10^{-3}$ จงหาคำตอบเมื่อ $\delta = 0$ และเมื่อ $\delta = 10^{-3}$ แล้วอธิบายว่าทำไมการเปลี่ยน δ เพียง 10^{-3} จึงเปลี่ยน $\mathbf{x}$ ราว 0.5 (ดู σ_2 และข้อควรระวังในหัวข้อนี้)

แบบฝึกหัด 8.2.5 (★★) ให้ $A = U\Sigma V^T$ เป็นการแยกค่าเอกฐาน

1. จงแสดงว่า $\|A\mathbf{x}\|_2 \le \sigma_1 \|\mathbf{x}\|_2$ สำหรับทุก $\mathbf{x}$ และ $\|A\|_2 = \sigma_1$
2. จงแสดงว่า ถ้า A เป็นเมทริกซ์จัตุรัสไม่เอกฐาน แล้ว $A^{+} = A^{-1}$

แบบฝึกหัด 8.2.6 (★ ★ ★, โครงงานเล็ก) ใช้ภาพสมมติขนาด 128×128 ของหัวข้อ 8.2 (สัญญาณรบกวน 0.02 เมล็ดสุ่ม 8) และโค้ดใน `test_ch08_svd.py`เป็นตัวช่วย

1. จงหา k ที่เล็กที่สุดที่ค่าคลาดเคลื่อนสัมพัทธ์นอร์มโฟรเบเนียสไม่เกิน 5% แล้วนับตัวเลขที่เก็บ $k(m+n+1)$ เทียบกับ $16\,384$
2. ทำซ้ำข้อ 1 สำหรับภาพเดียวกันแต่ตั้งสัญญาณรบกวนเป็นศูนย์ จงนับลำดับชั้นเชิงตัวเลขโดยเกณฑ์ $10^{-12}\sigma_1$ และ $10^{-3}\sigma_1$ แล้วอธิบายว่าทำไมจำนวนทั้งสองต่างกัน
3. ถ้าต้องการค่าคลาดเคลื่อนไม่เกิน 2% สำหรับภาพที่มีสัญญาณรบกวน ต้องเก็บกี่ชั้น และตัวเลขที่เก็บคิดเป็นกี่เปอร์เซ็นต์ของ $16\,384$

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 8.2.1.

1. แถวที่ i ของ $\mathbf{a}\mathbf{b}^T$ คือ $a_i(1,2)$ ได้ $(1,2)$, $(2,4)$, $(3,6)$ ซึ่งตรงกับ A ทุกแถว จึงลำดับชั้นของ A เท่ากับ 1 เพราะเป็นเมทริกซ์ลำดับชั้นหนึ่ง
2. $A^TA = \begin{pmatrix} 14 & 28 \\ 28 & 56 \end{pmatrix}$ ผลรวมของค่าลักษณะเฉพาะ (ร่องรอย) คือ 70 และดีเทอร์มิแนนต์คือ $14\cdot 56 - 28^2 = 0$ จึงค่าลักษณะเฉพาะคือ 70 และ 0 ดังนั้น $\sigma_1 = \sqrt{70}$ และ $\sigma_2 = 0$ ตรวจได้ด้วย $\|\mathbf{a}\|_2\|\mathbf{b}\|_2 = \sqrt{14}\sqrt{5} = \sqrt{70}$

ตรวจด้วยโค้ด: ใน `test_ch08_svd.py` (`test_exercise_821_rank_one_layer_and_sigma`)

เฉลย 8.2.2.

1. $A\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1{+}2,\, 2{+}1)^T = \frac{1}{\sqrt{2}}(3,3)^T = 3\mathbf{v}_1$ และ $A\mathbf{v}_2 = \frac{1}{\sqrt{2}}(1{-}2,\, 2{-}1)^T = \frac{1}{\sqrt{2}}(-1,1)^T = -\mathbf{v}_2$ ค่าลักษณะเฉพาะคือ 3 และ -1
2. เพราะ A สมมาตร $A^TA = A^2$ มีค่าลักษณะเฉพาะ 9 และ 1 จึง $\sigma_1 = 3$ และ $\sigma_2 = 1$ ได้ $\mathbf{u}_1 = A\mathbf{v}_1/3 = \mathbf{v}_1$ และ $\mathbf{u}_2 = A\mathbf{v}_2/1 = -\mathbf{v}_2 = \frac{1}{\sqrt{2}}(-1,1)^T$ จากนั้น

$$3\mathbf{u}_1\mathbf{v}_1^T + \mathbf{u}_2\mathbf{v}_2^T = \frac{3}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix} + \frac{1}{2}\begin{pmatrix}-1&1\\1&-1\end{pmatrix} = \begin{pmatrix}1&2\\2&1\end{pmatrix} = A.$$

ตรวจด้วยโค้ด: ใน `test_ch08_svd.py` (`test_exercise_822_symmetric_matrix_svd_by_hand`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 8.2.3.

1. $A = 2 \cdot \frac{1}{\sqrt{2}}(1,1)^T \cdot \frac{1}{\sqrt{2}}(1,1)$ จึง $\sigma_1 = 2$, $\mathbf{u}_1 = \mathbf{v}_1 = \frac{1}{\sqrt{2}}(1,1)^T$ และ $\sigma_2 = 0$ ดังนั้น

$$A^+ = \frac{1}{\sigma_1}\mathbf{v}_1\mathbf{u}_1^T = \frac{1}{2}\cdot\frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} = \frac{1}{4}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}, \qquad \mathbf{x}^+ = A^+\mathbf{b} = \frac{1}{4}(4,4)^T = (1,1)^T.$$

2. $A(2,0)^T = (2,2)^T$ และ $A(0,2)^T = (2,2)^T$ ทั้งสองให้ค่าตกค้าง $\mathbf{b} - A\mathbf{x} = (2,-2)^T$ เท่ากับของ $\mathbf{x}^+$ (ซึ่งได้ $\mathbf{b} - A\mathbf{x}^+ = (4,0)^T - (2,2)^T = (2,-2)^T$) แต่ $\|(2,0)^T\|_2 = 2 > \sqrt{2} = \|\mathbf{x}^+\|_2$ จึง $\mathbf{x}^+$ มีนอร์มเล็กกว่า

3. $A^TA = \begin{pmatrix}2 & 2\\ 2 & 2\end{pmatrix}$ มีดีเทอร์มิแนนต์ 0 สมการ $A^TA\mathbf{x} = A^T\mathbf{b} = (4,4)^T$ จึงเป็นสมการ $x_1 + x_2 = 2$ ซึ่งมีคำตอบมากมาย ไม่มี $(A^TA)^{-1}$ ให้ใช้ จึงต้องเลือกคำตอบด้วยเกณฑ์อื่น เช่นนอร์มเล็กที่สุด

ตรวจ ด้วย โค้ด: ใน `test_ch08_svd.py` (`test_exercise_823_pseudo_inverse_and_the_other_minimisers`)

เฉลย 8.2.4.

1. $A^TA = \begin{pmatrix}1+\varepsilon^2 & 1\\ 1 & 1+\varepsilon^2\end{pmatrix} = \begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} + \varepsilon^2 I$ ค่าลักษณะเฉพาะของ $\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$ คือ 2 (ที่ $(1,1)$) และ 0 (ที่ $(1,-1)$) เติม $\varepsilon^2 I$ แล้วได้ $2+\varepsilon^2$ และ ε^2 จึง $\sigma_1 = \sqrt{2+\varepsilon^2}$ และ $\sigma_2 = \varepsilon$

2. ระบบปรกติคือ $(1+\varepsilon^2)x_1 + x_2 = 2$ และ $x_1 + (1+\varepsilon^2)x_2 = 2 + \varepsilon\delta$ ลบสมการหลังจากสมการแรกได้ $\varepsilon^2(x_1 - x_2) = -\varepsilon\delta$ จึง $x_2 = x_1 + \delta/\varepsilon$ แทนลงในสมการแรก $(2+\varepsilon^2)x_1 + \delta/\varepsilon = 2$ ได้ $x_1 = (2 - \delta/\varepsilon)/(2+\varepsilon^2)$ เพราะ A มีลำดับชั้นเต็ม คำตอบนี้มีเพียงคำตอบเดียว

3. เมื่อ $\delta = 0$ ได้ $x_1 = x_2 = 2/(2+10^{-6}) \approx 0.9999995$ เมื่อ $\delta = 10^{-3}$ ได้ $\delta/\varepsilon = 1$ จึง $x_1 = 1/(2+10^{-6}) \approx 0.4999998$ และ $x_2 \approx 1.4999998$ ค่าเปลี่ยนไปราว 0.5 ทั้งที่ข้อมูลเปลี่ยนเพียง 10^{-3} เพราะการเปลี่ยน δ ถูกขยายด้วย $1/\sigma_2 = 10^3$ ตามที่ข้อควรระวังบอก

ตรวจด้วยโค้ด: ใน `test_ch08_svd.py` (`test_exercise_824_closed_form_for_epsilon_three` และ `test_exercise_824_change_is_about_one_half`)

เฉลย 8.2.5.

1. ให้ $\mathbf{y} = V^T\mathbf{x}$ เพราะ U เชิงตั้งฉาก $\|A\mathbf{x}\|_2^2 = \|\Sigma\mathbf{y}\|_2^2 = \sum_i \sigma_i^2 y_i^2 \le \sigma_1^2 \sum_i y_i^2 = \sigma_1^2\|\mathbf{x}\|_2^2$ โดยใช้ V เชิงตั้งฉากในขั้นสุดท้าย ที่ $\mathbf{x} = \mathbf{v}_1$ ความไม่เท่ากันเป็นสมการ จึง $\|A\|_2 = \sigma_1$

2. ถ้า A จัตุรัสไม่เอกฐาน ทุก $\sigma_i > 0$ จึง $\Sigma^+ = \Sigma^{-1}$ ดังนั้น $A^+ = V\Sigma^{-1}U^T$ และ $A^{-1} = (U\Sigma V^T)^{-1} = V\Sigma^{-1}U^T$ เพราะ U และ V เชิงตั้งฉาก จึงได้ $A^+ = A^{-1}$

ตรวจ ด้วย โค้ด: ใน `test_ch08_svd.py` (`test_exercise_825_spectral_norm_and_inverse_of_square_matrix`)

ฉบับร่าง Preprint 2026-10-23

เฉลย 8.2.6.

1. จากโค้ด ค่าคลาดเคลื่อนสัมพัทธ์ที่ $k = 12$ มากกว่า 0.05 และที่ $k = 13$ ได้ 0.0484 จึง $k = 13$ ตัวเลขที่เก็บคือ $13 \cdot 257 = 3\,341$ คิดเป็นราว 20.4% ของ $16\,384$

2. ภาพไม่มีสัญญาณรบกวนยังมีค่าเอกฐานครบ 128 ตัวที่เกณฑ์ $10^{-12}\sigma_1$ (ค่าเล็กสุดราว 1.1×10^{-3}) จึงลำดับชั้นเชิงตัวเลขเท่ากับ 128 แต่ที่เกณฑ์ $10^{-3}\sigma_1$ เหลือ 103 ค่าที่นับต่างกันเพราะเกณฑ์ต่างกัน ไม่ใช่เพราะภาพมีสัญญาณรบกวน ค่าเอกฐานชั้นท้ายที่ไม่เป็นศูนย์จึงต้องตัดสินใจว่าจะนับหรือไม่ นอกจากนี้ค่าคลาดเคลื่อนไม่เกิน 5% ยังต้องใช้ $k = 13$ เท่าเดิม

3. สำหรับภาพที่มีสัญญาณรบกวน $k = 44$ เพราะ $k = 43$ ยังให้ค่าคลาดเคลื่อนเกิน 2% ตัวเลขที่เก็บคือ $44 \cdot 257 = 11\,308$ คิดเป็นราว 69.0% ของ $16\,384$ ซึ่งแสดงว่าการลดค่าคลาดเคลื่อนลงอีกนิดต้องเก็บชั้นเพิ่มมาก

ตรวจด้วยโค้ด: ใน `test_ch08_svd.py` (`test_exercise_826_picture_project_numbers`)

สรุปบทที่ 8 ในหน้าเดียว

ตารางต่อไปนี้เทียบวิธีทั้งหมดของบทนี้ ตัวอักษร m และ n คือขนาดของเมทริกซ์ k คือจำนวนชั้นที่เก็บ ค่าใช้จ่ายเป็นลำดับโดยประมาณ

วิธี	ต้องการอะไร	ค่าใช้จ่าย	หมายเหตุ
สมการปรกติกับโชเลสกี (8.1)	ลำดับชั้นเต็ม	$O(mn^2 + n^3)$	คอนดิชันนัมเบอร์ยกกำลังสอง
QR ด้วยเฮาส์โฮลเดอร์ (8.1)	ลำดับชั้นเต็ม	$O(mn^2)$	ไม่สร้าง A^TA
เกาส์–นิวตันด้วย QR (8.1)	ลำดับชั้นเต็มของ J	$O(mn^2)$ ต่อรอบ	ขึ้นกับจุดเริ่ม
การแยกค่าเอกฐาน (8.2)	ไม่ต้องการอะไร	$O(mn^2)$ ต่อรอบ	ได้ค่าเอกฐานทั้งหมด
ตัวผกผันเทียม (8.2)	ตัดสินใจเรื่องค่าเอกฐานเล็กแล้ว	ตาม SVD	คำตอบนอร์มเล็กที่สุด
การประมาณลำดับชั้นต่ำ (8.2)	ชั้น k	เก็บ $k(m+n+1)$ ตัว	ค่าคลาดเคลื่อนสัมบูรณ์ σ_{k+1}

สูตรที่ควรจำ

- การแยกค่าเอกฐาน: $A = U\Sigma V^T = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ (หัวข้อ 8.2)
- นอร์มสเปกตรัม: $\|A\|_2 = \sigma_1$ และ $\|A - A_k\|_2 = \sigma_{k+1}$ (หัวข้อ 8.2)
- ค่าคลาดเคลื่อนสัมบูรณ์ในนอร์มโฟรเบเนียสของ A_k: $\sqrt{\sum_{i>k} \sigma_i^2}$ (หัวข้อ 8.2)
- ตัวผกผันเทียม: $\mathbf{x}^+ = A^+\mathbf{b} = V\Sigma^+U^T\mathbf{b}$ เป็นคำตอบกำลังสองน้อยสุดที่มีนอร์มเล็กที่สุด (หัวข้อ 8.2)
- คอนดิชันนัมเบอร์: $\operatorname{cond}_2 A = \sigma_1/\sigma_r$ เมื่อ $r = \operatorname{rank} A$ (หัวข้อ 7.2 และ 8.2)
- ตัวเลขที่เก็บของชั้น k: $k(m+n+1)$ (หัวข้อ 8.2)

อ่านต่อ. บทที่ 9 ทิศทางพิเศษของเมทริกซ์ ต่อจากนี้ ค่าเอกฐานของเมทริกซ์สมมาตรกึ่งบวกแน่นอนคือค่าลักษณะเฉพาะ ซึ่งเป็นมุมมองเดียวกับการแยกค่าเอกฐาน โครงงานเล็กของบทนี้คือแบบฝึกหัด 8.2.6 ซึ่งใช้ภาพสมมติของหัวข้อ 8.2 และค่าเอกฐานจากโค้ดของเรา

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 9

ทิศทางพิเศษของเมทริกซ์

สถานการณ์สมมติ: เจลลี่ที่สั่นตามรูปแบบของมันเอง

ลองนึกถึงวัตถุนิ่ม ๆ บนจอ เช่นเจลลี่ก้อนหนึ่ง หรือเชือกที่ประกอบจากลูกบอลเล็ก ๆ ต่อกันด้วยสปริง เมื่อเราเคาะมันเบา ๆ ตรงกลาง มันจะโยกไปมา และการโยกนั้นมักดูเป็นแบบแผน บางจุดขยับมากที่สุด บางจุดแทบไม่ขยับเลย แบบแผนแต่ละแบบค่อย ๆ ขยายและหดตามเวลา แต่รูปร่างของมันไม่เปลี่ยน แบบแผนนี้เรียกว่า **โหมด** (mode) คือการสั่นแบบหนึ่งที่รูปร่างคงที่

ถ้าเจลลี่มีลูกบอลเจ็ดลูกระหว่างกำแพงสองข้าง การโยกทั้งหมดดูซับซ้อน แต่ทุกการโยกเป็นผลรวมของโหมดเหล่านี้ และแต่ละโหมดสั่นด้วยความเร็วของตัวเอง เมื่อเคาะตรงกลาง บางโหมดก็ไม่ถูกกระตุ้นเลย คำถามของบทนี้จึงมีสองข้อ คือ โหมดเหล่านั้นคืออะไร และเราจะหาได้อย่างไรโดยไม่ต้องจำลองการสั่นทั้งหมด

(แบบจำลองนี้เป็นของเราเอง ไม่ได้บอกว่าวัตถุหรือเอนจินใดในเกมทำงานอย่างไร ตัวเลขทั้งหมดในบทนี้มาจากโค้ดของเรา)

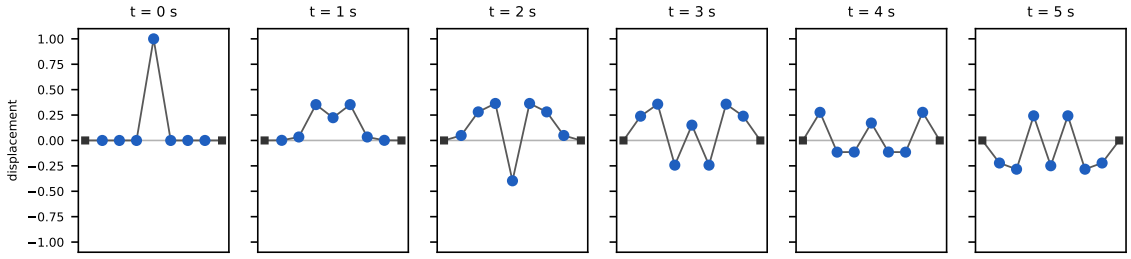
ภาพรวมบท

บทที่ 8 จบด้วยการแยกค่าเอกฐาน และบอกว่าสำหรับเมทริกซ์สมมาตรกึ่งบวกแน่นอน ค่าเอกฐานคือค่าลักษณะเฉพาะ บทนี้จึงเริ่มจากเจลลี่ที่สั่นตามรูปแบบของตัวเองแล้วจะเจออีกด้านหนึ่งคือ การจัดอันดับหน้าเว็บที่เป็นทิศทางเด่นของเมทริกซ์ที่ใหญ่มาก ตามลำดับนี้

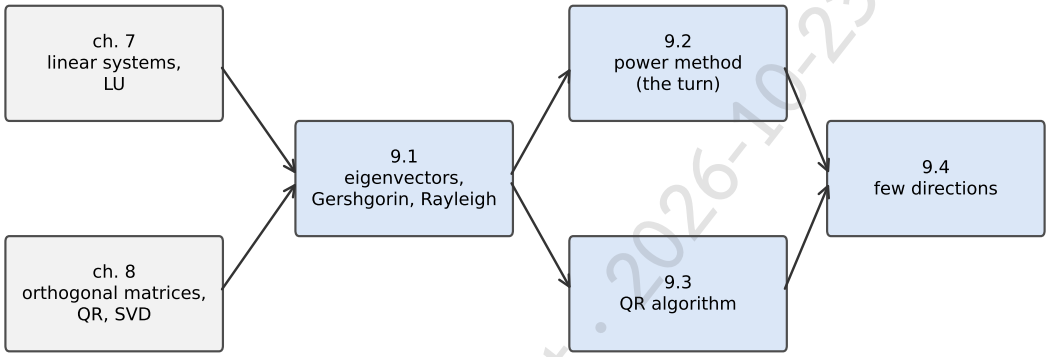
- หัวข้อ 9.1: ทิศทางที่ไม่เปลี่ยน เราดูค่าลักษณะเฉพาะและเวกเตอร์ลักษณะเฉพาะจากภาพของโหมด แล้วให้เครื่องมือตรวจตำแหน่งและความไวของค่าเหล่านั้น
- หัวข้อ 9.2: ยกกำลังซ้ำ ๆ วิธีกำลังหาทิศทางที่ใหญ่ที่สุด ซึ่งเป็นอีกด้านของปัญหาเดียวกัน
- หัวข้อ 9.3: ขั้นตอนวิธี QR หาทุกโหมดพร้อมกัน
- หัวข้อ 9.4: ใช้เพียงไม่กี่ทิศ เมื่อเมทริกซ์ใหญ่เกินกว่าจะหาทุกโหมดได้

ภาพแรกแสดงเจลลี่เจ็ดลูกที่ถูกเคาะตรงกลางในหลายเวลา ภาพที่สองแสดงลำดับของหัวข้อในบทนี้ และสิ่งที่บทนี้รับมาจากบทที่ 7 และ 8

ฉบับร่าง Preprint 2026-10-28



รูปที่ 9.1: เจลลี่เจ็ดลูกระหว่างกำแพงสองข้าง ถูกเคาะที่ลูกที่สี่ (ตรงกลาง) หนึ่งหน่วย ภาพแต่ละกรอบคือการโยกที่เวลาต่างกันในหน่วยของเรา รูปร่างการโยกสมมาตรเพราะการเคาะอยู่ตรงกลาง (สร้างจาก `numanal/ch09.py`)



รูปที่ 9.2: ลำดับของหัวข้อในบทที่ 9 และสิ่งที่บทนี้รับมาจากบทที่ 7 และ 8 (สร้างจาก `numanal/ch09.py`)

ฉบับร่าง · Preprint · 2026-10-23

9.1 ทิศทางที่ไม่เปลี่ยน

สถานการณ์สมมติ: แรงดึงกลับของสปริง

ให้ $\mathbf{x}$ เป็นเวกเตอร์ของการขยับของลูกบอลทุกลูกจากตำแหน่งพัก แรงดึงกลับของสปริงที่ดึงทุกลูกกลับที่เดิม เขียนได้เป็น $-A\mathbf{x}$ โดยที่ A เป็นเมทริกซ์ของการเชื่อมต่อ (แบบจำลองของเรา) ถ้าการขยับแบบหนึ่งทำให้แรงดึงกลับชี้ไปทางเดิมกับการขยับ การโยกแบบนั้นก็จะรักษารูปร่างไว้ได้ คำถามคือมีการขยับแบบนี้กี่แบบ และแต่ละแบบเร็วแค่ไหน

ทิศทางที่ A ไม่เปลี่ยน

ภาพรวมก่อนอ่าน

เมทริกซ์ส่วนใหญ่หมุนและยืดเวกเตอร์ไปทิศใหม่ แต่มีบางทิศที่เมทริกซ์ทำได้เพียงยืดหรือหดเวกเตอร์ไปตามทิศนั้น ทิศนี้คือ**ทิศทางพิเศษ** ของเมทริกซ์ และตัวเลขที่บอกว่ายืดเท่าไรคือค่าพิเศษของมัน ในเจลลี่ ทิศพิเศษคือรูปร่างของโหมด และค่าพิเศษคือกำลังสองของความถี่ของโหมดนั้น

นิยาม 9.1.1: ค่าลักษณะเฉพาะและเวกเตอร์ลักษณะเฉพาะ

ให้ A เป็นเมทริกซ์จัตุรัสขนาด $n \times n$ จำนวนเชิงซ้อน λ เรียกว่า **ค่าลักษณะเฉพาะ** (eigenvalue) ของ A ถ้ามีเวกเตอร์ $\mathbf{v} \neq \mathbf{0}$ ที่ทำให้ $A\mathbf{v} = \lambda\mathbf{v}$ เวกเตอร์ $\mathbf{v}$ นั้นเรียกว่า **เวกเตอร์ลักษณะเฉพาะ** (eigenvector) ที่สอดคล้องกับ λ เซตของค่าลักษณะเฉพาะทั้งหมดเรียกว่า **สเปกตรัม** (spectrum) และเขียนเป็น $\sigma(A)$

อ่านแบบเข้าใจง่าย. คำจำกัดความนี้บอกว่า เวกเตอร์ $\mathbf{v}$ เป็นทิศพิเศษ ก็ต่อเมื่อ $A\mathbf{v}$ ชี้ไปทางเดียวกับ $\mathbf{v}$ หรือทางตรงข้าม ตัวเลข λ บอกว่ายืดหรือหดเท่าไร และ λ ติดลบได้ ซึ่งหมายความว่าทิศนั้นกลับด้าน ตัวอย่างเช่น ถ้า $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ และ $\mathbf{v} = (1,1)^T$ จะได้ $A\mathbf{v} = (3,3)^T = 3\mathbf{v}$ จึง $\lambda = 3$ ส่วน $\mathbf{e}_1 = (1,0)^T$ ไม่ใช่ทิศพิเศษ เพราะ $A\mathbf{e}_1 = (2,1)^T$ ไม่ขนานกับ $\mathbf{e}_1$

ถ้าเขียน $A\mathbf{v} = \lambda\mathbf{v}$ ใหม่เป็น $(A - \lambda I)\mathbf{v} = \mathbf{0}$ และ $\mathbf{v} \neq \mathbf{0}$ เราจะได้ว่า $\det(A - \lambda I) = 0$ ซึ่งเป็นสมการพหุนามในตัวแปร λ สำหรับเมทริกซ์ขนาด $n \times n$ จึงมีค่าลักษณะเฉพาะ n ตัว นับจำนวนซ้ำตามรากของพหุนามนี้ Iserles 2009, §A.1.5, pp. 437--438 **ภาวะรากซ้ำเชิงพีชคณิต** (algebraic multiplicity) คือจำนวนครั้งที่ค่าหนึ่งเป็นรากซ้ำ ของพหุนามนี้

ลองเอง

หาค่าลักษณะเฉพาะของ $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ โดยแก้ $(2-\lambda)^2 - 1 = 0$ จะได้ $\lambda = 1$ และ $\lambda = 3$ จากนั้นหาเวกเตอร์ของ $\lambda = 1$ คือ $(1,-1)^T$ ตรวจว่า $A(1,-1)^T = (1,-1)^T$ และลองหาเวกเตอร์ของ $\lambda = 3$ เองด้วย

โซ่ของลูกบอลคือเมทริกซ์แบบไหน

สำหรับลูกบอล n ลูกที่เรียงกันระหว่างกำแพงสองข้าง โดยทุกลูกมีมวล 1 และทุกสปริงมีค่าคงที่ 1 สมการการเคลื่อนที่ของการขยับ $\mathbf{x}(t)$ คือ $\mathbf{x}'' = -A\mathbf{x}$ โดยที่

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n \times n},$$

ซึ่งมี 2 บนแนวทแยงมุม และ -1 ทันทีข้างแนวทแยงมุม เมทริกซ์แบบนี้เรียกว่า **เมทริกซ์สามแนวเฉียง** (tridiagonal matrix) (ตัวที่ไม่เป็นศูนย์อยู่ที่สามแนวเท่านั้น)

ถ้า $A\mathbf{v} = \lambda\mathbf{v}$ ลองแทนค่า $\mathbf{x}(t) = \mathbf{v}\cos(\omega t)$ ลงในสมการ จะได้ $\mathbf{x}'' = -\omega^2\mathbf{v}\cos(\omega t) = -\lambda\mathbf{v}\cos(\omega t) = -A\mathbf{x}$ เมื่อ $\omega^2 = \lambda$ ดังนั้น ทุกเวกเตอร์ลักษณะเฉพาะคือโหมดหนึ่ง และความถี่เชิงมุมของโหมดนั้นคือ $\omega = \sqrt{\lambda}$ ค่าลักษณะเฉพาะจึงเป็นกำลังสองของความถี่

ค่าลักษณะเฉพาะของโซ่คือ $\lambda_j = 2 - 2\cos\left(\frac{j\pi}{n+1}\right)$ สำหรับ $j = 1, \ldots, n$ และเวกเตอร์ลักษณะเฉพาะที่สอดคล้องกันมีสมาชิกลำดับที่ i เป็น $\sin\left(\frac{ij\pi}{n+1}\right)$ ที่มาของสูตรนี้จะกลับมาในบทที่ 10 ส่วนในบทนี้เราตรวจด้วยโค้ดของเรา โดยเทียบกับการคำนวณค่าลักษณะเฉพาะด้วยวิธีอื่น ตารางที่ 9.1 แสดงค่าสำหรับ $n = 7$ ที่ใช้ในภาพ

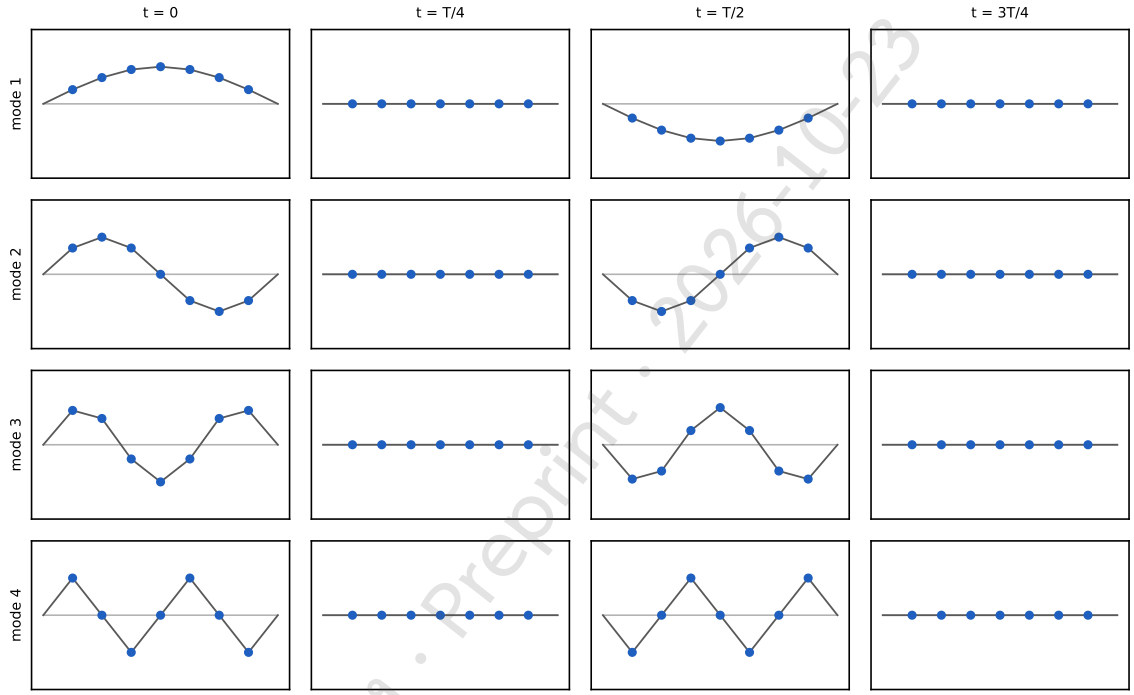
ข้อควรระวัง: ไม่ใช่ทุกเมทริกซ์มีทิศทางจริง

เมทริกซ์ $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ หมุนทุกเวกเตอร์ไปทางอื่นเสมอ (ทำมุม $90°$) ไม่มีเวกเตอร์จริงที่คูณแล้วชี้ทางเดิม ค่าลักษณะ

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 9.1: โซ่เจ็ดลูก ($n = 7$): ค่าลักษณะเฉพาะ λ_j และความถี่ $\omega_j = \sqrt{\lambda_j}$ (ตัวเลขจากโค้ดของเรา ปัดเป็นสี่ตำแหน่ง)

โหมด j	λ_j	ω_j
1	0.1522	0.3902
2	0.5858	0.7654
3	1.2346	1.1111
4	2.0000	1.4142
5	2.7654	1.6629
6	3.4142	1.8478
7	3.8478	1.9616



รูปที่ 9.3: สี่โหมดแรกของโซ่เจ็ดลูก แต่ละแถวคือโหมดหนึ่ง คอลัมน์คือเฟสของการสั่นที่ $t = 0,\ T/4,\ T/2,\ 3T/4$ เมื่อ T คือคาบของโหมดนั้น รูปร่างของแต่ละแถวไม่เปลี่ยน มีเพียงขนาดและเครื่องหมายที่สลับไปมา (สร้างจาก `numanal/ch09.py`)

> เฉพาะของมันคือ $\pm i$ ซึ่งเป็นจำนวนเชิงซ้อน ดังนั้นอย่าคาดหวังว่าทุกเมทริกซ์จะมีทิศพิเศษเป็นจำนวนจริงเสมอไป

ร่าง · Preprint · 2026-10-23

วงกลมของเกอร์ชโกริน: ค่าลักษณะเฉพาะอยู่ตรงไหน

ภาพรวมก่อนอ่าน

ถ้ายังไม่รู้ค่าลักษณะเฉพาะ เราอาจรู้ได้ว่ามันอยู่ตรงไหนจากสมาชิกของเมทริกซ์เอง ทุกแถวให้วงกลมหนึ่งวง มีศูนย์กลางอยู่ที่สมาชิกบนแนวทแยงมุม และมีรัศมีเท่าผลรวมขนาดของสมาชิกอื่นในแถวนั้น ค่าลักษณะเฉพาะทุกตัวอยู่ในวงกลมอย่างน้อยหนึ่งวง โดยที่เราไม่ต้องคำนวณค่าลักษณะเฉพาะก่อน

นิยาม 9.1.2: จานเกอร์ชโกริน

ให้ $A = [a_{ij}]$ เป็นเมทริกซ์จัตุรัสขนาด $n \times n$ ให้ $r_i = \sum_{j \neq i} |a_{ij}|$ และให้ **จานเกอร์ชโกริน** (Gerschgorin disc) ที่แถว i เป็นเซต $D_i = \{z \in \mathbb{C} : \ |z - a_{ii}| \leq r_i\}$

ทฤษฎีบท 9.1.1: ทฤษฎีบทจานเกอร์ชโกริน

ให้ A เป็นเมทริกซ์ขนาด $n \times n$

1. ค่าลักษณะเฉพาะทุกตัวของ A อยู่ในอย่างน้อยหนึ่งจาน D_i
2. ถ้า m จานรวมกันเป็นเซต S ที่เชื่อมกันเป็นชิ้นเดียวและแยกขาดจากจานที่เหลือ $n - m$ จาน แล้ว S มีค่าลักษณะเฉพาะ m ตัวพอดี นับตามภาวะรากซ้ำ

พิสูจน์. **แนวคิด.** ส่วน (i) พิสูจน์โดยเลือกองค์ประกอบที่ใหญ่ที่สุดของเวกเตอร์ลักษณะเฉพาะ แล้วใช้สมการแถวนั้น ส่วน (ii) พิสูจน์แบบแนวคิดเท่านั้น (ดูด้านล่าง)

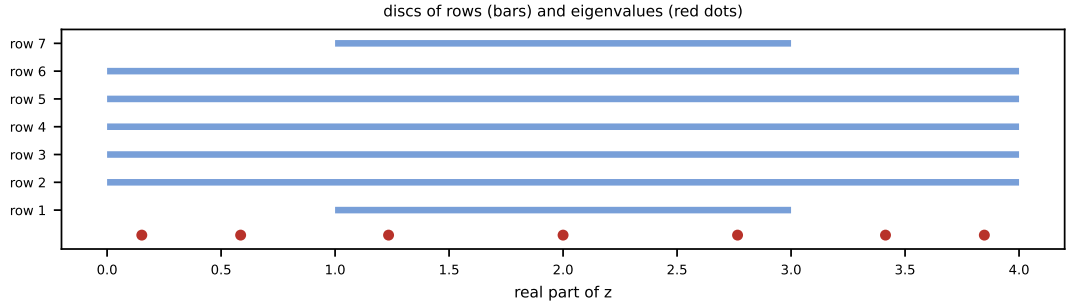
1. **เลือกองค์ประกอบ.** ให้ λ เป็นค่าลักษณะเฉพาะ และ $\mathbf{x} \neq \mathbf{0}$ เป็นเวกเตอร์ลักษณะเฉพาะ เลือก k ที่ทำให้ $|x_k| = \max_i |x_i| > 0$
2. **เขียนสมการแถว k.** จาก $A\mathbf{x} = \lambda \mathbf{x}$ แถวที่ k ได้ $(\lambda - a_{kk})x_k = \sum_{j \neq k} a_{kj} x_j$
3. **ประมาณค่าสัมบูรณ์.** เพราะ $|x_j| \leq |x_k|$ สำหรับทุก j จึงได้ $|\lambda - a_{kk}|\,|x_k| \leq \sum_{j \neq k} |a_{kj}|\,|x_j| \leq r_k |x_k|$
4. **หารด้วย $|x_k|$.** เพราะ $|x_k| > 0$ จึงได้ $|\lambda - a_{kk}| \leq r_k$ นั่นคือ $\lambda \in D_k$

สรุป: ส่วน (i) พิสูจน์ครบแล้ว ส่วน (ii) เราอ้างอิงแนวคิดของ Atkinson 1989, §9.1, Thm 9.1, pp. 588--589 คือ เปลี่ยนสมาชิกนอกแนวทแยงมุมด้วยตัวคูณ ε จาก 0 ถึง 1 แล้วใช้ความต่อเนื่องของค่าลักษณะเฉพาะ เพื่อนับว่าเซตที่เชื่อมกันมีกี่ค่า $\square$

อ่านแบบเข้าใจง่าย. ทฤษฎีบทบอกว่า ค่าลักษณะเฉพาะทุกตัวซ่อนอยู่ในจานใดจานหนึ่ง ถ้าจานไม่ทับกันเลย แต่ละจานจะมีค่าลักษณะเฉพาะหนึ่งตัวพอดี ถ้ามีจานทับกันเป็นกลุ่ม กลุ่มนั้นก็นับค่าลักษณะเฉพาะได้เท่าจำนวนจาน (ส่วน (ii)) ในทางกลับกัน จานเหล่านี้ให้ขอบเขต ไม่ใช่ค่าที่แน่นอน

สำหรับโซ่เจ็ดลูก แถวปลายมีจานรัศมี 1 คือ $[1, 3]$ ส่วนแถวกลางมีรัศมี 2 คือ $[0, 4]$ จานทั้งเจ็ดทับกันเป็นกลุ่มเดียว จึงมีค่าลักษณะเฉพาะครบเจ็ดตัวอยู่ในช่วง $[0, 4]$ ทุกค่า จากจานจึงได้ว่า $\lambda \leq 4$ และ $\omega \leq 2$ สำหรับทุกโหมด ค่าจริงที่มากที่สุดคือ 3.8478 (ตรงกับตารางที่ 9.1 และ `test_ch09.py`ตรวจแล้ว) ซึ่งให้ $\omega = 1.9616$

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 9.4: จานเกอร์ชโกรินของโซ่เจ็ดลูกบนเส้นจำนวน แถบสีฟ้าคือจานหนึ่งแถว (แถวปลายสั้นกว่า) จุดสีแดงคือค่าลักษณะเฉพาะทั้งเจ็ดตัว ซึ่งอยู่ในช่วง $[0, 4]$ ทั้งหมด (สร้างจาก `numanal/ch09.py`)

ผลหารเรย์ลี: เดาทิศแล้วได้ค่า

สมมติว่าเลือกทิศการโยกมาหนึ่งทิศ โดยที่ยังไม่รู้ค่าลักษณะเฉพาะ เราจะได้ตัวเลขที่ใกล้ค่าลักษณะเฉพาะบ้างหรือไม่ สำหรับเมทริกซ์สมมาตรมีคำตอบที่ตรงไปตรงมา

นิยาม 9.1.3: ผลหารเรย์ลี

ให้ A เป็นเมทริกซ์จริงสมมาตร และ $\mathbf{x} \neq \mathbf{0}$ เป็นเวกเตอร์จริง **ผลหารเรย์ลี** (Rayleigh quotient) ของ $\mathbf{x}$ คือ

$$R(\mathbf{x}) = \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}}.$$

อ่านแบบเข้าใจง่าย. ตัวเศษคือ $\mathbf{x}^T A\mathbf{x}$ ซึ่งบอกว่าเมทริกซ์ทำงานกับทิศ $\mathbf{x}$ มากน้อยเพียงใด ตัวส่วนคือกำลังสองของความยาว $\mathbf{x}$ ดังนั้น $R(\mathbf{x})$ เป็นค่าที่ไม่ขึ้นกับการขยายหรือหด $\mathbf{x}$ ถ้า $\mathbf{x}$ เป็นเวกเตอร์ลักษณะเฉพาะ $R(\mathbf{x})$ เท่ากับค่าลักษณะเฉพาะนั้นพอดี สำหรับโซ่เจ็ดลูก ถ้า $\mathbf{x} = (1, 1, \ldots, 1)^T$ จะได้ $R(\mathbf{x}) = 2/7 \approx 0.2857$ ซึ่งอยู่ในช่วง $[0.1522,\ 3.8478]$ ของค่าลักษณะเฉพาะ

ทฤษฎีบท 9.1.2: ผลหารเรย์ลีให้ค่าตกค้างต่ำสุด

ให้ A เป็นเมทริกซ์จริงสมมาตร และ $\mathbf{x} \neq \mathbf{0}$ สำหรับทุกจำนวนจริง μ จะได้ $\|A\mathbf{x} - \mu\mathbf{x}\|_2 \geq \|A\mathbf{x} - R(\mathbf{x})\,\mathbf{x}\|_2$ และเท่ากันเมื่อ $\mu = R(\mathbf{x})$

พิสูจน์. **แนวคิด.** เราเขียน $\|A\mathbf{x} - \mu\mathbf{x}\|_2^2$ เป็นพหุนามกำลังสองในตัวแปร μ แล้วเติมให้เป็นกำลังสองสมบูรณ์

1. **กระจายกำลังสอง.** $\|A\mathbf{x} - \mu\mathbf{x}\|_2^2 = \|A\mathbf{x}\|_2^2 - 2\mu\,\mathbf{x}^T A\mathbf{x} + \mu^2\|\mathbf{x}\|_2^2$
2. **เติมกำลังสองสมบูรณ์.** ให้ $R = R(\mathbf{x})$ เราได้ $\mathbf{x}^T A\mathbf{x} = R\|\mathbf{x}\|_2^2$ จึง
$$\|A\mathbf{x} - \mu\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2(\mu - R)^2 + \big(\|A\mathbf{x}\|_2^2 - R^2\|\mathbf{x}\|_2^2\big).$$
3. **สังเกตพจน์แรก.** พจน์ $\|\mathbf{x}\|_2^2(\mu - R)^2$ ไม่เป็นลบ และเป็นศูนย์เมื่อ $\mu = R$ พจน์ที่สองไม่ขึ้นกับ μ ดังนั้นค่าน้อยที่สุดเกิดที่ $\mu = R$ พอดี

สรุป: การกระจายกำลังสองทำให้เห็นทั้งว่า $R(\mathbf{x})$ เป็นค่าที่ดีที่สุดสำหรับเวกเตอร์ $\mathbf{x}$ และว่าค่าตกค้างที่เหลือไม่ขึ้นกับการเลือก μ □

ฉบับร่าง · Preprint · 2026-10-23

ลองเอง

ให้ $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ จงคำนวณ $R(\mathbf{x})$ สำหรับ $\mathbf{x} = (1,0)^T$, $(1,1)^T$ และ $(1,-1)^T$ ควรได้ 2, 3 และ 1 ตามลำดับ ค่าสองตัวสุดท้ายคือค่าลักษณะเฉพาะของ A เพราะ $(1,1)^T$ และ $(1,-1)^T$ เป็นเวกเตอร์ลักษณะเฉพาะ ส่วน $(1,0)^T$ ไม่ใช่ จึงได้ค่าที่อยู่ระหว่างกลาง

ค่าตกค้างกับขอบเขตของค่าลักษณะเฉพาะ

ถ้าเรามีเวกเตอร์ $\tilde{\mathbf{x}}$ ที่เป็นค่าประมาณของเวกเตอร์ลักษณะเฉพาะ เราเลือกค่า $\tilde{\lambda}$ ของตัวคูณ แล้วคำนวณ **ค่าตกค้าง** (residual) $\boldsymbol{\eta} = A\tilde{\mathbf{x}} - \tilde{\lambda}\tilde{\mathbf{x}}$ ซึ่งเป็นผลต่างที่เหลืออยู่ (เราเคยเจอค่าตกค้างในบทที่ 8) ถ้า $\boldsymbol{\eta}$ เล็ก เราควรได้ค่าลักษณะเฉพาะใกล้ $\tilde{\lambda}$ หรือไม่ คำตอบคือใช่ สำหรับเมทริกซ์สมมาตร

ทฤษฎีบท 9.1.3: ขอบเขตของค่าตกค้าง

ให้ A เป็นเมทริกซ์จริงสมมาตรขนาด $n \times n$ ที่มีค่าลักษณะเฉพาะ $\lambda_1, \ldots, \lambda_n$ ให้ $\mathbf{x} \neq \mathbf{0}$ และ $\tilde{\lambda} \in \mathbb{R}$ กำหนด $\boldsymbol{\eta} = A\mathbf{x} - \tilde{\lambda}\mathbf{x}$ แล้ว

$$\min_{1 \le i \le n} |\tilde{\lambda} - \lambda_i| \le \frac{\|\boldsymbol{\eta}\|_2}{\|\mathbf{x}\|_2}.$$

พิสูจน์. **แนวคิด.** เมทริกซ์สมมาตรมีการแยกตัวประกอบ $A = QDQ^T$ โดย Q เป็นเมทริกซ์เชิงตั้งฉาก และ D เป็นเมทริกซ์ทแยงมุมของค่าลักษณะเฉพาะ Vorst 2003, §2.2, p. 17 ใช้การแยกนี้เปลี่ยนค่าตกค้างเป็นการคูณด้วย $D - \tilde{\lambda}I$ ซึ่งนอร์มของมันคือค่าที่เล็กที่สุดของ $|\lambda_i - \tilde{\lambda}|$

1. **เปลี่ยนฐาน.** ให้ $\mathbf{y} = Q^T\mathbf{x}$ เพราะ $Q^TQ = I$ จึง $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2$ และ $\|Q^T\boldsymbol{\eta}\|_2 = \|\boldsymbol{\eta}\|_2$
2. **เขียนค่าตกค้างใหม่.** $Q^T\boldsymbol{\eta} = Q^TAQ\,Q^T\mathbf{x} - \tilde{\lambda}Q^T\mathbf{x} = (D - \tilde{\lambda}I)\mathbf{y}$
3. **ประมาณนอร์ม.** $\|(D - \tilde{\lambda}I)\mathbf{y}\|_2^2 = \sum_i (\lambda_i - \tilde{\lambda})^2 y_i^2 \;\ge\; \min_i(\lambda_i - \tilde{\lambda})^2 \sum_i y_i^2 = \min_i(\lambda_i - \tilde{\lambda})^2 \, \|\mathbf{x}\|_2^2$
4. **สรุป.** จาก $\|\boldsymbol{\eta}\|_2 = \|Q^T\boldsymbol{\eta}\|_2$ จึง $\|\boldsymbol{\eta}\|_2 \ge \min_i |\lambda_i - \tilde{\lambda}| \, \|\mathbf{x}\|_2$ ซึ่งเป็นข้อความที่ต้องการ

สรุป: ค่าตกค้างเล็กรับประกันว่ามีค่าลักษณะเฉพาะอยู่ใกล้ $\tilde{\lambda}$ และถ้าเลือก $\tilde{\lambda} = R(\mathbf{x})$ ขอบเขตนี้ดีที่สุดเพราะค่าตกค้างมีนอร์มเล็กที่สุดที่ตัวคูณนั้น (ทฤษฎีบท 9.1.2) □

อ่านแบบเข้าใจง่าย. ทฤษฎีบทนี้ให้เครื่องมือตรวจคำตอบโดยไม่ต้องรู้ค่าลักษณะเฉพาะ เราคำนวณค่าตกค้างแล้วหารด้วยความยาวของ $\mathbf{x}$ ถ้าได้ตัวเลขเล็ก เราก็รู้ว่ามีค่าลักษณะเฉพาะอยู่ใกล้ $\tilde{\lambda}$ ระยะที่ได้จริงอาจน้อยกว่าขอบเขตนี้มาก

สมมติว่า $\mathbf{x} = \mathbf{u}_1 + 0.1\,\mathbf{u}_2$ โดยที่ $\mathbf{u}_1, \mathbf{u}_2$ เป็นโหมดที่ยาวหนึ่งหน่วยของโซ่เจ็ดลูก ผลหารเรย์ลีของ $\mathbf{x}$ คือ $\tilde{\lambda} = 0.15653$ ค่าลักษณะเฉพาะที่ใกล้ที่สุดคือ 0.15224 ระยะห่างจริงคือ 0.00429 ส่วนขอบเขต $\|\boldsymbol{\eta}\|_2/\|\mathbf{x}\|_2$ คือ 0.04293 ขอบเขตนี้ถูกต้องแต่หลวมไปราวสิบเท่า (ตัวเลขจากโค้ดของเรา)

ข้อควรระวัง: ค่าตกค้างเล็กไม่ได้บอกว่าเวกเตอร์ถูก

ทฤษฎีบทบอกเพียงว่ามีค่าลักษณะเฉพาะอยู่ใกล้ $\tilde{\lambda}$ ไม่ได้บอกว่าเป็นค่าที่เราต้องการ และไม่ได้บอกว่า $\tilde{\mathbf{x}}$ ใกล้เวกเตอร์ลักษณะเฉพาะนั้น ถ้าค่าลักษณะเฉพาะสองตัวอยู่ใกล้กันมาก ค่าตกค้างเล็กอาจตรงกับตัวใดก็ได้ จึงต้องดูให้แน่ใจด้วยวิธีอื่นด้วย

ฉบับร่าง · Preprint · 2026-10-23

ความไว: เมื่อเมทริกซ์เปลี่ยนเล็กน้อย

ในทางปฏิบัติ สมาชิกของเมทริกซ์มาจากการวัดหรือการปัดเศษ จึงมีความคลาดเคลื่อนเล็กน้อยเสมอ คำถามคือ ค่าลักษณะเฉพาะเปลี่ยนไปมากแค่ไหน

ทฤษฎีบท 9.1.4: ความไวของค่าลักษณะเฉพาะ

1. (แบบทั่วไป) ให้ $A = PDP^{-1}$ โดย D เป็นเมทริกซ์ทแยงมุม ถ้า μ เป็นค่าลักษณะเฉพาะของ $A + E$ แล้ว
$$\min_i |\mu - \lambda_i| \le \|P\|_2 \, \|P^{-1}\|_2 \, \|E\|_2,$$
โดย λ_i คือสมาชิกทแยงมุมของ D
2. (แบบสมมาตร) ถ้า A เป็นเมทริกซ์จริงสมมาตร แล้วทุกค่าลักษณะเฉพาะ μ ของ $A+E$ สอดคล้อง $\min_i |\mu - \lambda_i| \le \|E\|_2$ สำหรับเมทริกซ์ E ใดก็ได้

พิสูจน์. **แนวคิด.** ส่วน (ii) เป็นกรณีพิเศษของ (i) เมื่อเลือก P เป็นเมทริกซ์เชิงตั้งฉาก (นอร์ม 2 ของ P และ P^{-1} เท่ากับ 1) สำหรับส่วน (i) ให้ μ เป็นค่าลักษณะเฉพาะของ $A + E$ โดยที่ไม่ตรงกับ λ_i ใด ๆ (กรณีอื่นเป็นจริงทันที)

1. **ย้ายไปฐาน P.** ให้ $(A + E)\mathbf{x} = \mu\mathbf{x}$ โดย $\mathbf{x} \neq \mathbf{0}$ จะได้ $P(\mu I - D)P^{-1}\mathbf{x} = E\mathbf{x}$ แล้วให้ $\mathbf{z} = P^{-1}\mathbf{x}$ จะได้ $(\mu I - D)\mathbf{z} = P^{-1}EP\,\mathbf{z}$
2. **ประมาณนอร์ม.** เพราะ $\mu I - D$ เป็นเมทริกซ์ทแยงมุม นอร์ม 2 ของ $(\mu I - D)^{-1}$ คือ $1/\min_i |\mu - \lambda_i|$ จึง $\|\mathbf{z}\|_2 \le \|P^{-1}EP\|_2 \, \|\mathbf{z}\|_2 / \min_i |\mu - \lambda_i|$
3. **สรุป.** หารด้วย $\|\mathbf{z}\|_2 > 0$ แล้วได้ $\min_i |\mu - \lambda_i| \le \|P^{-1}EP\|_2 \le \|P\|_2 \|P^{-1}\|_2 \|E\|_2$

สรุป: การเปลี่ยนเมทริกซ์เล็กน้อยทำให้ค่าลักษณะเฉพาะเปลี่ยนได้ไม่เกินนอร์มของการเปลี่ยนนั้น คูณด้วยตัวเลข $\|P\|\|P^{-1}\|$ ซึ่งเป็นตัวบอกความไวของเมทริกซ์นั้น Atkinson 1989, §9.1, Thm 9.2 และ (9.1.11), pp. 592--593 □

อ่านแบบเข้าใจง่าย. สำหรับเมทริกซ์สมมาตร ไม่มีตัวคูณขยายเลย ค่าลักษณะเฉพาะเปลี่ยนไม่เกินขนาดของการเปลี่ยนสมาชิก เจลลี่ของเราจึงทนต่อการวัดที่คลาดเคลื่อนเล็กน้อย ตัวอย่างนี้ใช้การเปลี่ยน E ที่มีนอร์ม 2 เท่ากับ 0.01 แบบสุ่มด้วยเมล็ดสุ่มห้า ค่า ค่าลักษณะเฉพาะที่เปลี่ยนมากที่สุดเปลี่ยนไปเพียง 0.0033, 0.0078, 0.0049, 0.0036 และ 0.0072 ตามลำดับ ซึ่งทุกค่าน้อยกว่า 0.01 (ตัวเลขจากโค้ดของเรา)

ข้อควรระวัง: เมทริกซ์ที่เวกเตอร์ลักษณะเฉพาะไม่ครบ

เมทริกซ์ $A_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ มีค่าลักษณะเฉพาะ 0 สองซ้ำ แต่มีเวกเตอร์ลักษณะเฉพาะเพียงทิศเดียว $(1, 0)^T$ จึงเรียกว่า **เมทริกซ์ที่เวกเตอร์ลักษณะเฉพาะไม่ครบ** (defective matrix) และไม่มี P ที่ทำให้ $A_0 = PDP^{-1}$ ทฤษฎีบทข้างต้นจึงใช้ไม่ได้ ถ้าเพิ่มสมาชิกที่ตำแหน่ง $(2, 1)$ ให้เป็น ε จะได้ค่าลักษณะเฉพาะ $\pm\sqrt{\varepsilon}$ เมื่อ $\varepsilon = 10^{-6}$ ค่าลักษณะเฉพาะเคลื่อนไป 10^{-3} ซึ่งมากกว่าการเปลี่ยน ε เองถึงพันเท่า (ตัวเลขจากโค้ดของเรา) ตำรา Atkinson 1989, §9.1, p. 594 แสดงปรากฏการณ์คล้ายกันสำหรับเมทริกซ์ไม่สมมาตร ตัวอย่างของเราเป็นรูปที่เรียบง่ายที่สุดของเรื่องนี้

ฉบับร่าง · Preprint · 2026-10-23

สรุปใจความ

- ทิศพิเศษของเมทริกซ์คือเวกเตอร์ $\mathbf{v}$ ที่ $A\mathbf{v} = \lambda\mathbf{v}$ ค่า λ คือค่าลักษณะเฉพาะ ในเจลลี่ $\mathbf{v}$ คือรูปร่างของโหมด และ $\omega = \sqrt{\lambda}$ คือความถี่ของโหมดนั้น (หัวข้อ 9.1)
- จานเกอร์ชโกรินบอกตำแหน่งของค่าลักษณะเฉพาะโดยไม่ต้องคำนวณก่อน สำหรับโซ่เจ็ดลูก ค่าลักษณะเฉพาะทั้งหมดอยู่ใน $[0, 4]$ (หัวข้อ 9.1)
- ผลหารเรย์ลี $R(\mathbf{x})$ ให้ค่าตกค้างต่ำสุดสำหรับเวกเตอร์ที่เลือก และขอบเขต $\min_i |\tilde{\lambda} - \lambda_i| \le \|\boldsymbol{\eta}\|_2 / \|\mathbf{x}\|_2$ ตรวจคำตอบได้โดยไม่ต้องรู้ค่าลักษณะเฉพาะ (หัวข้อ 9.1)
- สำหรับเมทริกซ์สมมาตร การเปลี่ยนเล็กน้อยเปลี่ยนค่าลักษณะเฉพาะไม่เกินนอร์มของการเปลี่ยนนั้น แต่สำหรับเมทริกซ์ที่เวกเตอร์ลักษณะเฉพาะไม่ครบ ค่าลักษณะเฉพาะอาจไวมาก (หัวข้อ 9.1)
- กลับไปที่เจลลี่: การเคาะตรงกลางกระตุ้นเฉพาะโหมดที่สมมาตร คือโหมด $1, 3, 5, 7$ ของโซ่เจ็ดลูก โหมดคู่ไม่ถูกกระตุ้นเลย จึงเห็นการโยกเป็นแบบแผนไม่กี่แบบ

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 9.1.1 (⋆) ให้ $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$

1. จงหาค่าลักษณะเฉพาะทั้งสองตัวจาก $\det(A - \lambda I) = 0$
2. จงหาเวกเตอร์ลักษณะเฉพาะที่ทิศ $(1, 1)^T$ และ $(1, -1)^T$ แล้วตรวจว่าแต่ละตัวสอดคล้องกับนิยาม $A\mathbf{v} = \lambda\mathbf{v}$

แบบฝึกหัด 9.1.2 (⋆) ให้ $A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ (โซ่สามลูก)

1. จงแสดงว่า $\lambda = 2 - \sqrt{2},\ 2,\ 2 + \sqrt{2}$ เป็นค่าลักษณะเฉพาะ โดยใช้เวกเตอร์ $(1, \sqrt{2}, 1)^T$, $(1, 0, -1)^T$ และ $(1, -\sqrt{2}, 1)^T$ ตามลำดับ
2. จงตรวจว่าผลรวมของค่าลักษณะเฉพาะเท่ากับผลรวมของสมาชิกบนแนวทแยงมุม

แบบฝึกหัด 9.1.3 (⋆⋆) ลองพิสูจน์ว่า ถ้า T เป็นเมทริกซ์สามเหลี่ยมบน (สมาชิกใต้แนวทแยงมุมเป็นศูนย์ทั้งหมด) แล้วค่าลักษณะเฉพาะของ T คือสมาชิกบนแนวทแยงมุม แล้วใช้ผลนี้หาค่าลักษณะเฉพาะของ

$$T = \begin{pmatrix} 5 & 7 & -2 \\ 0 & -1 & 4 \\ 0 & 0 & 3 \end{pmatrix}.$$

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 9.1.4 (★★) ให้ $G = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -3 \end{pmatrix}$

1. จงหาจานเกอร์ชโกรินทั้งสามจาน และช่วงบนเส้นจำนวนที่แต่ละจานครอบคลุม
2. จงแสดงว่าค่าลักษณะเฉพาะของ G คือ $2-\sqrt{5}$, -3 และ $2+\sqrt{5}$ และแต่ละตัวอยู่ในจานที่ทฤษฎีบทบอก
3. จานทั้งสามแยกขาดจากกัน จงอธิบายว่าทำไมแต่ละจานจึงมีค่าลักษณะเฉพาะหนึ่งตัวพอดี

แบบฝึกหัด 9.1.5 (★★) ให้ $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ และ $\mathbf{x} = (1, t)^T$ โดย t เป็นจำนวนจริง

1. จงแสดงว่า $R(\mathbf{x}) = \dfrac{2 + 2t + 2t^2}{1 + t^2}$
2. จงแสดงว่า $R(\mathbf{x}) - 1 = \dfrac{(1+t)^2}{1+t^2} \geq 0$ และ $3 - R(\mathbf{x}) = \dfrac{(1-t)^2}{1+t^2} \geq 0$ แล้วสรุปว่า $1 \leq R(\mathbf{x}) \leq 3$ จงบอกว่าค่าสุดขีดเกิดที่ t ใด และเวกเตอร์นั้นเป็นเวกเตอร์ลักษณะเฉพาะหรือไม่

แบบฝึกหัด 9.1.6 (★★, ตรวจด้วยโค้ดได้) ให้ A เหมือนข้อ 9.1.5 และ $\tilde{\mathbf{x}} = (1,\ 0.9)^T$

1. จงหา $\tilde{\lambda} = R(\tilde{\mathbf{x}})$ และค่าตกค้าง $\boldsymbol{\eta} = A\tilde{\mathbf{x}} - \tilde{\lambda}\tilde{\mathbf{x}}$
2. จงคำนวณ $\|\boldsymbol{\eta}\|_2 / \|\tilde{\mathbf{x}}\|_2$ และ $\min_i |\tilde{\lambda} - \lambda_i|$ โดย λ_i คือค่าลักษณะเฉพาะของ A แล้วตรวจว่าทฤษฎีบทขอบเขตของค่าตกค้างเป็นจริง
3. ขอบเขตหลวมไปกี่เท่าของระยะห่างจริง

แบบฝึกหัด 9.1.7 (★★★, สิ่งที่ผิดพลาดได้) ให้ $A_\varepsilon = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix}$ โดย $\varepsilon > 0$ เมื่อ $\varepsilon = 0$ เมทริกซ์นี้คือ A_0 ในหัวข้อ 9.1 ซึ่งเป็นเมทริกซ์ที่เวกเตอร์ลักษณะเฉพาะไม่ครบ และ $A_\varepsilon - A_0$ คือการเพิ่มสมาชิกที่ตำแหน่ง $(2,1)$

1. จงแสดงว่าค่าลักษณะเฉพาะของ A_ε คือ $\pm\sqrt{\varepsilon}$
2. ถ้า $\varepsilon = 10^{-8}$ จงหาค่าลักษณะเฉพาะ และเทียบกับการเปลี่ยนสมาชิก ε
3. ถ้า $\varepsilon = 10^{-16}$ (ขนาดใกล้เคียงกับการปัดเศษในเลขทศนิยมคู่) จงหาค่าลักษณะเฉพาะ แล้วอธิบายว่าเรื่องนี้บอกอะไรเกี่ยวกับการคำนวณค่าลักษณะเฉพาะของเมทริกซ์ที่เวกเตอร์ไม่ครบ

สัปดาห์ · preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 9.1.1.

1. $\det(A-\lambda I) = (3-\lambda)^2 - 1 = (2-\lambda)(4-\lambda)$ จึงได้ $\lambda = 2$ และ $\lambda = 4$
2. เมื่อ $\lambda = 4$ เมทริกซ์ $A - 4I = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ มีเวกเตอร์ในแนวว่าง $(1,1)^T$ ตรวจ $A(1,1)^T = (4,4)^T = 4(1,1)^T$ ส่วนเมื่อ $\lambda = 2$ เมทริกซ์ $A-2I = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ มีเวกเตอร์ $(1,-1)^T$ ตรวจ $A(1,-1)^T = (2,-2)^T = 2(1,-1)^T$

ตรวจด้วยโค้ด: ใน `test_ch09.py`(`test_exercise_911_two_by_two_eigen`)

เฉลย 9.1.2.

1. ผลคูณ $A(1,\sqrt{2},1)^T$ ได้ $(2-\sqrt{2},\ 2\sqrt{2}-2,\ 2-\sqrt{2})^T$ ซึ่งเท่ากับ $(2-\sqrt{2})(1,\sqrt{2},1)^T$ เพราะ $(2-\sqrt{2})\sqrt{2} = 2\sqrt{2}-2$ ส่วน $A(1,0,-1)^T = (2,0,-2)^T = 2(1,0,-1)^T$ และ $A(1,-\sqrt{2},1)^T = (2+\sqrt{2},\ -2-2\sqrt{2},\ 2+\sqrt{2})^T = (2+\sqrt{2})(1,-\sqrt{2},1)^T$
2. ผลรวมของค่าลักษณะเฉพาะคือ $(2-\sqrt{2})+2+(2+\sqrt{2}) = 6$ และผลรวมของสมาชิกบนแนวทแยงคือ $2+2+2 = 6$

ตรวจด้วยโค้ด: ใน `test_ch09.py`(`test_exercise_912_chain_three`)

เฉลย 9.1.3. เราเขียน $T - \lambda I$ ซึ่งเป็นเมทริกซ์สามเหลี่ยมบนที่มีสมาชิกบนแนวทแยงมุมเป็น $t_{ii} - \lambda$ ดีเทอร์มิแนนต์ของเมทริกซ์สามเหลี่ยมเท่ากับผลคูณของสมาชิกบนแนวทแยงมุม จึง $\det(T-\lambda I) = \prod_i (t_{ii}-\lambda)$ ค่านี้เป็นศูนย์ก็ต่อเมื่อ $\lambda = t_{ii}$ สำหรับบาง i ดังนั้นค่าลักษณะเฉพาะของ T คือสมาชิกบนแนวทแยงมุมทั้งหมด

สำหรับเมทริกซ์ในโจทย์ สมาชิกบนแนวทแยงมุมคือ $5, -1$ และ 3 ค่าลักษณะเฉพาะจึงเป็น $-1, 3$ และ 5

ตรวจด้วยโค้ด: ใน `test_ch09.py`(`test_exercise_913_triangular_example`)

เฉลย 9.1.4.

1. แถวที่ 1: จุดศูนย์กลาง 4 รัศมี $|1| = 1$ จึงได้จาน $[3,5]$ แถวที่ 2: จุดศูนย์กลาง 0 รัศมี $|1| + |0| = 1$ จึงได้จาน $[-1,1]$ แถวที่ 3: จุดศูนย์กลาง -3 รัศมี 0 จึงได้จาน $\{-3\}$
2. ส่วนแรกของ G เป็นบล็อกต่อไปนี้ ส่วนสมาชิก -3 แยกต่างหาก
$$\begin{pmatrix} 4 & 1 \\ 1 & 0 \end{pmatrix}.$$
ค่าลักษณะเฉพาะของบล็อกหาได้จาก $\det\begin{pmatrix} 4-\lambda & 1 \\ 1 & -\lambda \end{pmatrix} = \lambda^2 - 4\lambda - 1 = 0$ จึงได้ $\lambda = 2 \pm \sqrt{5}$ รวมกับ -3 ได้ค่าลักษณะเฉพาะ $2-\sqrt{5} \approx -0.236$, -3 และ $2+\sqrt{5} \approx 4.236$ ตามลำดับ ค่า 4.236 อยู่ใน $[3,5]$ ค่า -0.236 อยู่ใน $[-1,1]$ และ -3 อยู่ใน $\{-3\}$
3. ช่วง $[3,5]$, $[-1,1]$ และ $\{-3\}$ ไม่ทับกันเลย จึงเป็นสามเซตที่เชื่อมกันเป็นชิ้นเดียวและแยกขาดจากกัน ทฤษฎีบทส่วน (ii) จึงบอกว่าแต่ละจานมีค่าลักษณะเฉพาะหนึ่งตัวพอดี ซึ่งตรงกับข้อ 2

ตรวจด้วยโค้ด: ใน `test_ch09.py`(`test_exercise_914_block_eigen` และ `test_gershgorin_block_components_each_one_eigenvalue`)

ฉบับร่าง · preprint · 2026-10-23

เฉลย 9.1.5.

1. $\mathbf{x}^T A\mathbf{x} = (1,t)\begin{pmatrix}2+t\\1+2t\end{pmatrix} = (2+t)+t(1+2t) = 2+2t+2t^2$ และ $\mathbf{x}^T\mathbf{x} = 1+t^2$ จึงได้ผลลัพธ์

2. คำนวณได้

$$R(\mathbf{x}) - 1 = \frac{2+2t+2t^2-(1+t^2)}{1+t^2} = \frac{1+2t+t^2}{1+t^2} = \frac{(1+t)^2}{1+t^2} \geq 0,$$

$$3 - R(\mathbf{x}) = \frac{3(1+t^2)-(2+2t+2t^2)}{1+t^2} = \frac{1-2t+t^2}{1+t^2} = \frac{(1-t)^2}{1+t^2} \geq 0.$$

ดังนั้น $1 \leq R(\mathbf{x}) \leq 3$ ค่าต่ำสุด 1 เกิดที่ $t=-1$ คือ $\mathbf{x} = (1,-1)^T$ และค่าสูงสุด 3 เกิดที่ $t=1$ คือ $\mathbf{x} = (1,1)^T$ ทั้งสองเป็นเวกเตอร์ลักษณะเฉพาะ และค่าสุดขีดของผลหารเรย์ลีคือค่าลักษณะเฉพาะ

ตรวจด้วยโค้ด: ใน `test_ch09.py`(`test_exercise_915_rayleigh_range_and_extremes`) ด้วยการไล่ค่า t บนช่วง $[-5,5]$

เฉลย 9.1.6.

1. $\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}} = 2+2(0.9)+2(0.81) = 5.42$ และ $\tilde{\mathbf{x}}^T\tilde{\mathbf{x}} = 1.81$ จึง $\tilde{\lambda} = 5.42/1.81 \approx 2.99448$ จาก $A\tilde{\mathbf{x}} = (2.9,\ 2.8)^T$ และ $\tilde{\lambda}\tilde{\mathbf{x}} \approx (2.99448,\ 2.69503)^T$ ได้ $\boldsymbol{\eta} \approx (-0.09448,\ 0.10497)^T$

2. $\|\boldsymbol{\eta}\|_2 \approx \sqrt{0.09448^2 + 0.10497^2} \approx 0.14123$ และ $\|\tilde{\mathbf{x}}\|_2 = \sqrt{1.81} \approx 1.34536$ จึงได้ $\|\boldsymbol{\eta}\|_2/\|\tilde{\mathbf{x}}\|_2 \approx 0.10497$ ค่าลักษณะเฉพาะของ A คือ 1 และ 3 ดังนั้น $\min_i|\tilde{\lambda}-\lambda_i| = |2.99448-3| \approx 0.00552$ และ $0.00552 \leq 0.10497$ ทฤษฎีบทเป็นจริง

3. ขอบเขตหลวมราว $0.10497/0.00552 \approx 19$ เท่า ทฤษฎีบทใช้ได้กับทุกเวกเตอร์ $\tilde{\mathbf{x}}$ โดยไม่ต้องรู้ค่าลักษณะเฉพาะ ความหลวมนี้จึงเป็นราคาของการไม่รู้ค่าลักษณะเฉพาะนั้นเอง

ตรวจด้วยโค้ด: ใน `test_ch09.py`(`test_residual_bound_exercise_one_point_nine`)

เฉลย 9.1.7.

1. $\det(A_\varepsilon - \lambda I) = \det\begin{pmatrix}-\lambda & 1\\ \varepsilon & -\lambda\end{pmatrix} = \lambda^2 - \varepsilon$ จึงได้ $\lambda = \pm\sqrt{\varepsilon}$

2. เมื่อ $\varepsilon = 10^{-8}$ ได้ $\lambda = \pm 10^{-4}$ การเคลื่อนของค่าลักษณะเฉพาะจาก 0 คือ 10^{-4} ซึ่งมากกว่าการเปลี่ยนสมาชิก $\varepsilon = 10^{-8}$ ถึง 10^4 เท่า

3. เมื่อ $\varepsilon = 10^{-16}$ ได้ $\lambda = \pm 10^{-8}$ การเปลี่ยนสมาชิกเพียงขนาดการปัดเศษ 10^{-16} ทำให้ค่าลักษณะเฉพาะเคลื่อนถึง 10^{-8} ซึ่งเป็นรากที่สองของการเปลี่ยนนั้น นี่คือเหตุผลที่ค่าลักษณะเฉพาะของเมทริกซ์ที่เวกเตอร์ลักษณะเฉพาะไม่ครบคำนวณได้แม่นน้อยกว่าที่คาด ความแม่นหายไปราวครึ่งหนึ่งของหลักทศนิยม และทฤษฎีบทความไวในหัวข้อนี้ใช้ไม่ได้ เพราะไม่มี P ที่ทำให้ $A_0 = PDP^{-1}$

ตรวจด้วยโค้ด: ใน `test_ch09.py`(`test_exercise_917_jordan_eps_1e8` และ `test_exercise_917_jordan_eps_1e16`)

อ่านต่อ. ตำรา Atkinson 1989, §9.1, pp. 588--596 ให้การหาตำแหน่งของค่าลักษณะเฉพาะด้วยวงกลมของเกอร์ชโกริน ผลหารเรย์ลี และขอบเขตของค่าตกค้างสำหรับเมทริกซ์สมมาตร ที่เราอ้างอิงในหัวข้อนี้ และ Iserles 2009, §A.1.5, pp. 437--438 ให้นิยามค่าลักษณะเฉพาะที่ใช้ในบทนี้ หัวข้อ 9.2 จะถามต่อว่า ถ้าเราสนใจเพียงทิศที่ใหญ่ที่สุด เราไม่จำเป็นต้องหาค่าทั้งหมด ทิศนั้นคือทิศทางเด่นที่เราจะเห็นในอันดับของหน้าเว็บ

ฉบับร่าง · Preprint · 2026-10-23

9.2 ยกกำลังซ้ำ ๆ

สถานการณ์สมมติ: อีกด้านหนึ่งของปัญหาเดียวกัน: เรียงหน้าเว็บตามการคลิกลิงก์

เมื่อกี้เราเห็นเจลลี่ที่โยกตามโหมดของมัน และโหมดที่สั่นเร็วที่สุดคือทิศที่ใหญ่ที่สุดของเมทริกซ์ ตอนนี้ลองดูอีกด้านหนึ่งของเรื่องเดียวกัน คราวนี้ทิศที่เด่นคือ *คะแนนของหน้าเว็บ*

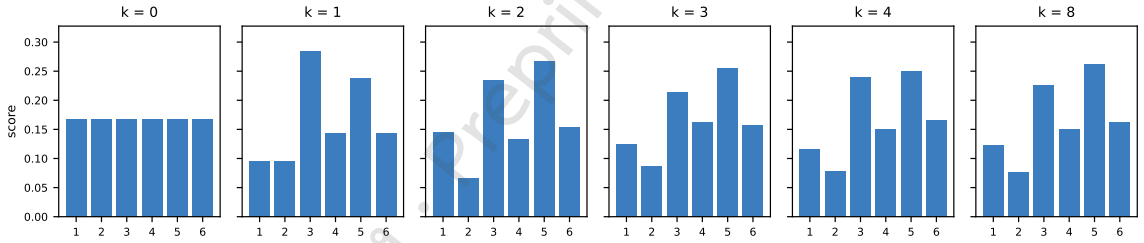
สมมติว่ามีหน้าเว็บหกหน้า และบางหน้ามีลิงก์ไปยังหน้าอื่น ลองนึกถึงผู้อ่านคนหนึ่งที่เริ่มจากหน้าใดก็ได้ แล้วทุกครั้งเขาคลิกลิงก์บนหน้าปัจจุบันแบบสุ่ม หน้าไหนที่เขากลับไปเยี่ยมบ่อยที่สุด หน้านั้นก็น่าจะสำคัญที่สุด ถ้ารู้สัดส่วนนี้ เราก็เรียงหน้าได้ทันที

เว็บจริงมีหน้านับล้าน เราเขียนเมทริกซ์ทั้งก้อนลงกระดาษไม่ได้ และแยกค่าลักษณะเฉพาะทุกตัวก็เกินกำลัง สิ่งที่ทำได้คือคูณเวกเตอร์ด้วยเมทริกซ์ทีละรอบ คำถามคือ ถ้าทำซ้ำไปเรื่อย ๆ สัดส่วนนั้นจะนิ่งไปที่ค่าใด

(หน้าเว็บทั้งหกและแบบจำลองการคลิกเป็นของเราเอง ไม่ได้บอกว่าเครื่องมือค้นหาใดทำงานอย่างไร ตัวเลขทั้งหมดในหัวข้อนี้มาจากโค้ดของเรา)

ภาพรวมก่อนอ่าน

นี่คืออีกด้านของเรื่องเดียวกัน ในเจลลี่เราอยากรู้ทิศเด่นที่สุด ในหน้าเว็บเราอยากรู้ทิศเด่นของเมทริกซ์การคลิกทั้งก้อน วิธีที่ง่ายที่สุดคือคูณเวกเตอร์ด้วยเมทริกซ์ แล้วปรับขนาดให้มีความยาวหนึ่ง ทำซ้ำไปเรื่อย ๆ เวกเตอร์จะค่อย ๆ หันเข้าหาทิศเด่น วิธีนี้เรียกว่า **วิธีกำลัง** (power method) บทนี้จะถามว่าวิธีนี้ลู่เข้าเมื่อไร และเมื่อไม่ลู่เข้าเราควรทำอย่างไร



รูปที่ 9.5: คะแนนของหน้าทั้งหกหลังคูณ k รอบ ($k = 0, 1, 2, 3, 4, 8$) กรอบซ้ายสุดคือจุดเริ่มต้นที่ทุกหน้าเท่ากัน ทุกกรอบใช้แกนแนวตั้งเดียวกัน ที่ $k = 8$ ลำดับเริ่มชัด (สร้างจาก `numanal/ch09_power.py`)

ฉบับร่าง · preprint · 2026-10-23

คูณซ้ำ ๆ แล้วทิศจะหันเข้าหาทิศเด่น

ก่อนคำนวณ เราต้องตกลงว่าผู้อ่านคลิกอย่างไร ผู้อ่านที่อยู่บนหน้า j จะคลิกลิงก์หนึ่งในลิงก์ของหน้านั้นแบบเท่ากันทุกลิงก์ด้วยความน่าจะเป็น $d = 0.85$ และด้วยความน่าจะเป็น $1 - d = 0.15$ เขาจะกระโดดไปหน้าใดก็ได้แบบเท่ากันทุกหน้า ค่า d นี้เรียกว่า **ตัวประกอบการหน่วง** (damping factor) ซึ่งบอกว่าเราเชื่อการคลิกลิงก์มากแค่ไหน (เหตุผลที่ต้องมีการกระโดดจะกลับมาอธิบายในหัวข้อย่อยที่สาม) ทุกหน้าในตัวอย่างนี้มีลิงก์ออกอย่างน้อยหนึ่งลิงก์

ให้ P เป็นเมทริกซ์การลิงก์ โดย $P_{ij} = 1/L_j$ ถ้าหน้า j มีลิงก์ไปยังหน้า i และ $P_{ij} = 0$ ถ้าไม่มี เมื่อ L_j คือจำนวนลิงก์ออกของหน้า j เมทริกซ์การคลิกทั้งหมดคือ

$$G = d\,P + \frac{1-d}{6}\,\mathbf{1}\mathbf{1}^{T},$$

เมื่อ $\mathbf{1} = (1, 1, 1, 1, 1, 1)^T$ ถ้า $\mathbf{x}$ คือคะแนนของหน้าทั้งหกที่รวมเป็นหนึ่ง คะแนนรอบถัดไปคือ $G\mathbf{x}$

ตารางที่ 9.2: คะแนนของหน้าทั้งหกที่รอบ k ต่าง ๆ (ตัวเลขจากโค้ดของเรา ปัดเป็นสี่ตำแหน่ง)

k	หน้า 1	หน้า 2	หน้า 3	หน้า 4	หน้า 5	หน้า 6
0	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
1	0.0958	0.0958	0.2847	0.1431	0.2375	0.1431
2	0.1460	0.0657	0.2345	0.1330	0.2676	0.1531
4	0.1156	0.0780	0.2404	0.1502	0.2494	0.1664
8	0.1231	0.0764	0.2267	0.1498	0.2622	0.1618
30	0.1219	0.0768	0.2280	0.1505	0.2602	0.1627

เริ่มจาก $\mathbf{x}_0 = \frac{1}{6}\mathbf{1}$ ซึ่งทุกหน้าเท่ากัน เมื่อคูณหนึ่งรอบ คะแนนของหน้า i คือ

$$(G\mathbf{x}_0)_i = \frac{d\, s_i}{6} + \frac{1-d}{6}, \qquad s_i = \sum_j P_{ij}.$$

หน้า 1 มีลิงก์เข้าจากหน้า 3 เท่านั้น และหน้า 3 มีลิงก์ออกสองลิงก์ จึง $s_1 = \frac{1}{2}$ และได้คะแนน $0.85 \cdot \frac{1}{12} + 0.025 = 0.0958$ ส่วนหน้า 3 มีลิงก์เข้าจากหน้า 2, 4 และ 5 ได้ $s_3 = 1 + \frac{1}{2} + \frac{1}{3} = \frac{11}{6}$ จึงได้คะแนน $0.85 \cdot \frac{11}{36} + 0.025 = 0.2847$ ทั้งสองค่าตรงกับกรอบที่ $k = 1$ ในรูป

ตารางที่ 9.2 แสดงคะแนนที่รอบต่าง ๆ จากรอบที่ 2 เป็นต้นไป หน้า 5 มีคะแนนสูงสุด และที่รอบที่ 30 ลำดับของหน้าที่เหลือคือ 3, 6, 4, 1, 2 ความคลาดเคลื่อนรวมจากค่านิ่งเหลือไม่ถึง 10^{-6} (ตัวเลขจากโค้ดของเรา) คำถามคือ ทำไมมันนิ่งไปที่ค่านี้ และนิ่งเร็วแค่ไหน

ในโลกจริง: แนวคิดนี้มาจากเครื่องมือค้นหาต้นแบบ (Brin และ Page, 1998)

ตัวอย่างหกหน้าข้างต้นเป็นของเรา แต่แนวคิดในการเรียงหน้าด้วยการคลิกลิงก์มาจากบทความปี 1998 ที่อธิบายเครื่องมือค้นหาต้นแบบ Google Brin and Page 1998, §2.1.1--2.1.2 ผู้เขียนบอกไว้ดังนี้

- ลองนึกถึงผู้ท่องเว็บแบบสุ่ม (random surfer) ที่เริ่มจากหน้าใดก็ได้ แล้วคลิกลิงก์ต่อไปเรื่อย ๆ จนบางครั้งเบื่อและขอหน้าสุ่มใหม่ ความน่าจะเป็นที่ผู้ท่องเว็บจะไปถึงหน้าใดคือ **คะแนนความสำคัญของหน้า** (PageRank) ของหน้านั้น
- โดยทั่วไปผู้เขียนตั้งตัวประกอบการหน่วง d ไว้ที่ 0.85 ซึ่งตรงกับค่าที่เราเลือกในตัวอย่างนี้
- คะแนนนี้สอดคล้องกับทิศเด่น (principal eigenvector) ของเมทริกซ์การลิงก์ของเว็บที่ทำให้เป็นมาตรฐานแล้ว

เอกสารนั้นไม่มีคะแนนของหน้าทั้งหกนี้ หน้าเหล่านี้และตารางข้างบนเป็นของเราเอง ใช้เพื่อดูแนวคิดเท่านั้น

ฉบับร่าง. Preprint · 2026-10-23

วิธีกำลังและอัตราการลู่เข้า

1. เลือกเวกเตอร์เริ่มต้น $\mathbf{x}_0 \neq \mathbf{0}$ ที่มีส่วนประกอบบนทิศเด่นไม่เป็นศูนย์ (สุ่มก็ได้)
2. คำนวณ $\mathbf{y} = A\mathbf{x}_{m-1}$
3. ปรับขนาด $\mathbf{x}_m = \mathbf{y}/\|\mathbf{y}\|_2$ แล้วประมาณค่าลักษณะเฉพาะด้วย $\mathbf{x}_m^T A \mathbf{x}_m$ (ถ้า A สมมาตร ค่านี้คือผลหารเรย์ลี)

แต่ละรอบต้องคูณเมทริกซ์กับเวกเตอร์เพียงหนึ่งครั้ง ไม่ต้องแยกตัวประกอบเมทริกซ์ จึงเหมาะกับเมทริกซ์ใหญ่ที่สมาชิกส่วนใหญ่เป็นศูนย์ วิธีนี้เป็นวิธีคลาสสิกที่หนังสือ Atkinson 1989, §9.2, pp. 602--606 อธิบายไว้ โดยใช้ตัวปรับขนาดแบบค่าสูงสุดแทนนอร์ม 2

ทฤษฎีบท 9.2.1: การลู่เข้าของวิธีกำลัง

ให้ A เป็นเมทริกซ์จัตุรัสขนาด $n \times n$ ที่มีเวกเตอร์ลักษณะเฉพาะ $\mathbf{v}_1, \ldots, \mathbf{v}_n$ เป็นฐานของ $\mathbb{C}^n$ ให้ $\lambda_1, \ldots, \lambda_n$ เป็นค่าลักษณะเฉพาะที่สอดคล้องกัน โดย λ_1 เป็นจำนวนจริง และ $|\lambda_1| > |\lambda_i|$ ทุก $i \geq 2$ ให้ $q = \max_{i \geq 2} |\lambda_i/\lambda_1|$ ให้ $\mathbf{x}_0 = \sum_i c_i \mathbf{v}_i$ โดย $c_1 \neq 0$ และให้ $\mathbf{x}_m = A\mathbf{x}_{m-1}/\|A\mathbf{x}_{m-1}\|_2$ แล้ว

$$\left\| \mathbf{x}_m - \operatorname{sgn}(\lambda_1)^m \frac{c_1 \mathbf{v}_1}{\|c_1 \mathbf{v}_1\|_2} \right\|_2 \;\leq\; \frac{2K}{|c_1|\,\|\mathbf{v}_1\|_2}\, q^m, \qquad K = \sum_{i \geq 2} |c_i|\,\|\mathbf{v}_i\|_2 .$$

อ่านแบบเข้าใจง่าย. ทุกรอบ ส่วนประกอบของ $\mathbf{x}_0$ บนแต่ละโหมดถูกคูณด้วยค่าลักษณะเฉพาะยกกำลัง m โหมดที่ค่าขนาดใหญ่ที่สุดจึงครองทิศไปเรื่อย ๆ อัตรา $q = |\lambda_2/\lambda_1|$ บอกว่าโหมดอื่นจางหายเร็วแค่ไหน ถ้า $q = \frac{1}{2}$ ความคลาดเคลื่อนลดครึ่งหนึ่งทุกรอบ แต่ถ้า q ใกล้ 1 ต้องใช้รอบมาก สำหรับโซ่เจ็ดลูกของหัวข้อ 9.1 เราได้ $q = 3.4142/3.8478 \approx 0.8873$ หลังสิบรอบค่าประมาณของ λ จึงยังต่างจาก 3.8478 อยู่ราว 0.016 (ตัวเลขจากโค้ดของเรา)

พิสูจน์. **แนวคิด.** เขียนเวกเตอร์เริ่มต้นเป็นผลรวมของโหมด แล้วดูว่าการคูณซ้ำทำให้โหมดหนึ่งครองทิศอย่างไร

1. **กระจายเวกเตอร์เริ่มต้น.** เมื่อ $c_1 \neq 0$ การคูณ m รอบให้ $A^m \mathbf{x}_0 = \sum_i c_i \lambda_i^m \mathbf{v}_i = \lambda_1^m (\mathbf{a} + \mathbf{r}_m)$ โดย $\mathbf{a} = c_1 \mathbf{v}_1$ และ $\mathbf{r}_m = \sum_{i \geq 2} c_i (\lambda_i/\lambda_1)^m \mathbf{v}_i$
2. **ประมาณส่วนที่เหลือ.** เพราะ $|\lambda_i/\lambda_1| \leq q$ ทุก $i \geq 2$ จึง $\|\mathbf{r}_m\|_2 \leq K q^m$
3. **ทิศของ $\mathbf{x}_m$.** แต่ละรอบปรับขนาดด้วยจำนวนบวกเท่านั้น ดังนั้น $\mathbf{x}_m$ เป็นจำนวนบวกคูณ $A^m \mathbf{x}_0$ นั่นคือ $\mathbf{x}_m = \operatorname{sgn}(\lambda_1)^m (\mathbf{a} + \mathbf{r}_m)/\|\mathbf{a} + \mathbf{r}_m\|_2$ (เครื่องหมายมาจากการดึง λ_1^m ออกมา)
4. **เทียบกับ $\mathbf{a}$.** สำหรับเวกเตอร์ $\mathbf{a}, \mathbf{r}$ ที่ $\mathbf{a} + \mathbf{r} \neq \mathbf{0}$ จะได้ $\left\| \frac{\mathbf{a}+\mathbf{r}}{\|\mathbf{a}+\mathbf{r}\|_2} - \frac{\mathbf{a}}{\|\mathbf{a}\|_2} \right\|_2 \leq \frac{2\|\mathbf{r}\|_2}{\|\mathbf{a}\|_2}$ เพราะผลต่างนี้เท่ากับผลรวมของ $(\mathbf{a}+\mathbf{r})\big(\frac{1}{\|\mathbf{a}+\mathbf{r}\|} - \frac{1}{\|\mathbf{a}\|}\big)$ กับ $\frac{\mathbf{r}}{\|\mathbf{a}\|}$ และพจน์แรกมีนอร์ม $\frac{|\,\|\mathbf{a}\| - \|\mathbf{a}+\mathbf{r}\|\,|}{\|\mathbf{a}\|} \leq \frac{\|\mathbf{r}\|}{\|\mathbf{a}\|}$
5. **สรุป.** รวมข้อ 2–4 และใช้ $\|\mathbf{a}\|_2 = |c_1|\,\|\mathbf{v}_1\|_2$ จะได้ขอบเขตที่ต้องการ

สรุป: ทิศของ $\mathbf{x}_m$ เข้าหา $\pm \mathbf{v}_1$ ด้วยอัตรา q^m เครื่องหมายสลับทุกรอบเมื่อ $\lambda_1 < 0$ และนิ่งเมื่อ $\lambda_1 > 0$ บทพิสูจน์นี้เราเขียนเอง โดยใช้ข้อสมมติเดียวกับ Atkinson 1989, §9.2, pp. 602--606 □

สำหรับเว็บหกหน้า เมทริกซ์ G มีค่าลักษณะเฉพาะ 1 เพียงตัวเดียวที่ขนาดเท่ากับ 1 และค่าลักษณะเฉพาะตัวที่สองมีขนาด 0.612 (ตรวจด้วยโค้ดของเรา) ค่าลักษณะเฉพาะทั้งหกตัวของ G แยกกันคนละค่า ทำให้มีเวกเตอร์ลักษณะเฉพาะครบฐาน ตัวอย่างนี้จึงเข้าเงื่อนไขของทฤษฎีบท อัตรา 0.612 จึงบอกว่าคะแนนนิ่งเร็วกว่าโซ่ในหัวข้อ 9.1 ที่มีอัตรา 0.887

ฉบับร่าง Preprint · 2026-10-23

ข้อควรระวัง: เมื่อค่าลักษณะเฉพาะขนาดเท่ากัน

ทฤษฎีบทต้องการ $|\lambda_1| > |\lambda_2|$ ถ้าสองค่าขนาดเท่ากัน คะแนนจะไม่นิ่ง ตัวอย่างที่ง่ายที่สุดคือหน้าสองหน้าที่ลิงก์ไปหากันเท่านั้น ดังหัวข้อย่อยถัดไป เมทริกซ์การลิงก์ P_2 มีค่าลักษณะเฉพาะ ± 1 และถ้าเริ่มจาก $(1,0)^T$ คะแนนจะสลับระหว่าง $(1,0)^T$ กับ $(0,1)^T$ ไปเรื่อย ๆ โดยไม่นิ่ง

ทำไมต้องมีการกระโดด: หน้าสองหน้าที่วนไปมา

ลองดูหน้าสองหน้า โดยหน้า 1 ลิงก์ไปหน้า 2 และหน้า 2 ลิงก์กลับไปหน้า 1 เท่านั้น ถ้าไม่มีการกระโดด ผู้อ่านจะวนไปมาตลอด เมทริกซ์การลิงก์คือ $P_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ และมีค่าลักษณะเฉพาะ 1 กับ -1 ที่ขนาดเท่ากัน ดังที่ระบุไว้ในข้อควรระวัง

เติมการกระโดดด้วย d เราได้ $G_2 = dP_2 + \frac{1-d}{2}\mathbf{1}\mathbf{1}^T$ ซึ่งสมาชิกทุกตัวเป็นบวกเมื่อ $0 < d < 1$ ตรวจได้ว่า $G_2\mathbf{1} = \mathbf{1}$ และ $G_2(1,-1)^T = -d\,(1,-1)^T$ จึงได้ค่าลักษณะเฉพาะ 1 กับ $-d$ เท่านั้น เมื่อ $d = 0.85$ และเริ่มจาก $(1,0)^T = \frac{1}{2}\mathbf{1} + \frac{1}{2}(1,-1)^T$ คะแนนของหน้า 1 หลัง k รอบคือ $\frac{1}{2} + \frac{1}{2}(-0.85)^k$ ความคลาดเคลื่อนจึงเท่ากับ $\frac{1}{2}(0.85)^k$ ซึ่งลดลงทุกรอบ แม้คะแนนจะแกว่งรอบ $\frac{1}{2}$ อยู่บ้าง

อ่านแบบเข้าใจง่าย

การกระโดดทำให้ทุกหน้าเชื่อมกับทุกหน้าอย่างน้อยเล็กน้อย ผู้อ่านที่ติดวงจรจึงหลุดออกมาได้ และค่าลักษณะเฉพาะที่ขนาดเท่ากันก็ถูกทำให้แยกจากกัน ค่า d เป็นค่าที่เราเลือกเพื่อให้เห็นภาพ ไม่ได้บอกว่าเว็บหรือเครื่องมือใดใช้ค่าใด

ตัวผกผัน: หาทิศที่อยู่ใกล้ค่าที่กำหนด

วิธีกำลังหาทิศที่ใหญ่ที่สุด แต่บางครั้งเราต้องการค่าลักษณะเฉพาะที่อยู่ใกล้ค่า σ ที่เรากำหนดไว้ ถ้าใช้เมทริกซ์ $(A-\sigma I)^{-1}$ แทน A ค่าลักษณะเฉพาะ λ_i จะกลายเป็น $1/(\lambda_i - \sigma)$ และค่าที่อยู่ใกล้ σ จะมีขนาดใหญ่ที่สุด จึงกลายเป็นทิศเด่น วิธีนี้เรียกว่า **การวนซ้ำแบบผกผัน** (inverse iteration) ในทางปฏิบัติเราไม่คำนวณ $(A-\sigma I)^{-1}$ แต่แก้สมการ $(A-\sigma I)\mathbf{w} = \mathbf{x}_{m-1}$ ทุกรอบด้วยการแยกตัวประกอบ Atkinson 1989, §9.6, pp. 628--629 ใช้วิธีนี้เป็นเครื่องมือหลักในการหาเวกเตอร์ลักษณะเฉพาะ

ทฤษฎีบท 9.2.2: การลู่เข้าของการวนซ้ำแบบผกผัน

ให้ A มีเวกเตอร์ลักษณะเฉพาะ $\mathbf{v}_1, \dots, \mathbf{v}_n$ เป็นฐาน โดยค่าลักษณะเฉพาะ λ_i เป็นจำนวนจริงทั้งหมด ให้ σ ไม่ตรงกับ λ_i ใด และให้ λ_k เป็นค่าเดียวที่อยู่ใกล้ σ มากที่สุด คือ $|\lambda_k - \sigma| < |\lambda_i - \sigma|$ ทุก $i \neq k$ ถ้า $\mathbf{x}_0$ มีส่วนประกอบบน $\mathbf{v}_k$ ไม่เป็นศูนย์ และ $\mathbf{x}_m = (A-\sigma I)^{-1}\mathbf{x}_{m-1}/\|\cdot\|_2$ แล้ว $\mathbf{x}_m$ เข้าหา $\pm\mathbf{v}_k$ ด้วยอัตรา

$$q = \frac{|\lambda_k - \sigma|}{\min_{i \neq k} |\lambda_i - \sigma|}.$$

พิสูจน์. **แนวคิด.** การวนซ้ำแบบผกผันเป็นวิธีกำลังของเมทริกซ์ $B = (A-\sigma I)^{-1}$ จึงใช้ทฤษฎีบทก่อนหน้าได้ทันที

1. **เปลี่ยนเมทริกซ์.** $B\mathbf{v}_i = \mu_i \mathbf{v}_i$ โดย $\mu_i = 1/(\lambda_i - \sigma)$ จริง ดังนั้น B มีเวกเตอร์ลักษณะเฉพาะชุดเดียวกับ A และ $\mathbf{x}_m$ ตรงกับวิธีกำลังของ B
2. **ตรวจเงื่อนไข.** เพราะ $|\mu_k| = 1/|\lambda_k - \sigma|$ มีขนาดมากกว่า $|\mu_i|$ ทุก $i \neq k$ ทฤษฎีบทการลู่เข้าของวิธีกำลังใช้ได้กับ B และส่วนประกอบของ $\mathbf{x}_0$ บน $\mathbf{v}_k$ ไม่เป็นศูนย์

ฉบับร่าง · preprint · 2026-10-23

3. **อัตรา.** $\max_{i\neq k}|\mu_i|/|\mu_k| = |\lambda_k-\sigma|/\min_{i\neq k}|\lambda_i-\sigma|$ ซึ่งเป็น q ที่ต้องการ

สรุป: ผลนี้ได้จากการนำทฤษฎีบทวิธีกำลังมาใช้กับ B ทั้งหมด □

สำหรับโซ่เจ็ดลูก เลือก $\sigma = 0.6$ และเริ่มจาก $\mathbf{x}_0 = \mathbf{e}_1$ (มวลลูกแรกเท่านั้นมีค่าหนึ่ง) ค่าลักษณะเฉพาะที่ใกล้ 0.6 ที่สุดคือ $\lambda_2 = 0.585786$ ระยะทาง 0.0142 ส่วนตัวที่ใกล้รองลงมาคือ $\lambda_1 = 0.152241$ ระยะทาง 0.4478 ดังนั้นอัตราตามทฤษฎีบทคือ $0.0142/0.4478 \approx 0.0317$ ค่าประมาณจากผลหารเรย์ลีคือ 2.000000, 0.586755, 0.585787 และ 0.585786 ตามลำดับ (ตัวเลขจากโค้ดของเรา) หลังสามรอบค่าที่ได้ตรงกับ λ_2 ถึงหกตำแหน่ง

ข้อควรระวัง: เวกเตอร์เริ่มต้นสมมาตร

ถ้าเริ่มจาก $(1,1,\ldots,1)^T$ แทน $\mathbf{e}_1$ เวกเตอร์เริ่มต้นนี้สมมาตรรอบจุดกึ่งกลางของโซ่ และไม่มีส่วนประกอบบนโหมด 2 เลย (ในเลขคณิตที่แม่นยำ) ผลคือการวนซ้ำไปลู่เข้าที่ $\lambda_1 = 0.152241$ ซึ่งเป็นค่าลักษณะเฉพาะที่ใกล้ 0.6 รองลงมาในกลุ่มโหมดคี่ ในโค้ดของเรา ค่าประมาณไปอยู่ราว 0.1524 ที่รอบที่ 10 แล้วกระโดดไป 0.5858 ภายในสามรอบต่อมา เพราะการปัดเศษทำให้มีส่วนประกอบบนโหมด 2 เล็กมาก (ราว 10^{-16}) ซึ่งถูกขยายประมาณ 31 เท่าทุกรอบ คำตอบที่ถูกต้องจึงต้องตรวจด้วยวิธีอื่นเสมอ

การเลื่อนค่า: ทำให้ค่าบนสุดแยกจากค่าถัดไป

ถ้าค่าลักษณะเฉพาะที่ใหญ่ที่สุดกับค่าถัดไปอยู่ใกล้กัน วิธีกำลังจะช้า เราจึงเลื่อนทุกค่าลงด้วยจำนวน b เท่ากัน แล้วทำวิธีกำลังกับ $A - bI$ วิธีนี้เรียกว่า **การเลื่อนค่า** (shift) ค่าลักษณะเฉพาะของ $A-bI$ คือ $\lambda_i - b$ การเลือก b จึงเปลี่ยนอัตรา q ได้ ถ้าค่าลักษณะเฉพาะเป็นจำนวนจริงทั้งหมด Atkinson 1989, §9.2, pp. 606--607 ให้สูตรของค่า b ที่ทำให้อัตรานี้เล็กที่สุดคือ $b^* = \tfrac{1}{2}(\lambda_2 + \lambda_n)$ เมื่อ λ_2 คือค่าที่ใหญ่ที่สุดรองลงมา และ λ_n คือค่าที่เล็กที่สุด เราอ้างสูตรนี้โดยไม่พิสูจน์ที่นี่

สำหรับโซ่เจ็ดลูก $b^* = \tfrac{1}{2}(3.414214 + 0.152241) \approx 1.783227$ เมื่อเลื่อนด้วย b^* ค่า $\lambda_7 - b^* \approx 2.0645$ มากที่สุด ส่วน $|\lambda_6 - b^*|$ และ $|\lambda_1 - b^*|$ เท่ากันที่ 1.6310 อัตราจึงเป็น $1.6310/2.0645 \approx 0.790$ แทนที่จะเป็น 0.887 (ตัวเลขจากโค้ดของเรา) ซึ่งดีขึ้นแต่ไม่มาก

ข้อควรระวัง: การเลื่อนที่ทำให้ขนาดเท่ากัน

ถ้าเลือก $b = 2$ ซึ่งเป็นจุดกึ่งกลางของค่าลักษณะเฉพาะสุดขั้วของโซ่ ค่า $\lambda_7 - 2 = 1.8478$ กับ $\lambda_1 - 2 = -1.8478$ มีขนาดเท่ากัน วิธีกำลังจะไม่ลู่เข้า ค่าประมาณจากผลหารเรย์ลีจะคงที่ที่ -1.8420 ซึ่งไม่ใช่ค่าลักษณะเฉพาะของ $A-2I$ (ค่าลักษณะเฉพาะคือ ± 1.8478) ค่าคงที่ที่ไม่ตรงกับค่าลักษณะเฉพาะจึงไม่ใช่หลักฐานว่าลู่เข้าแล้ว

การวนซ้ำผลหารเรย์ลี: เปลี่ยนค่าเดาทุกรอบ

สำหรับเมทริกซ์สมมาตร ผลหารเรย์ลี $R(\mathbf{x}_m)$ เป็นค่าประมาณที่ดีกว่าที่ได้จากวิธีกำลังธรรมดา Atkinson 1989, §9.2, pp. 608--609 บอกว่าเมื่อใช้วิธีกำลัง ผลหารเรย์ลีลู่เข้าด้วยอัตรา $(\lambda_2/\lambda_1)^2$ ซึ่งดีกว่าอัตรา λ_2/λ_1 ของตัวเวกเตอร์เอง *การวนซ้ำผลหารเรย์ลี* (Rayleigh quotient iteration) ใช้ประโยชน์จากเรื่องนี้ โดยเปลี่ยนค่าเดาทุกรอบ: $\sigma_m = R(\mathbf{x}_m)$ แล้วแก้ $(A-\sigma_m I)\mathbf{w} = \mathbf{x}_m$ และปรับขนาดเหมือนเดิม หนังสือเดียวกันนี้ระบุว่าเป็นรูปแบบที่ลู่เข้าเร็วสำหรับเมทริกซ์สมมาตร แต่เราไม่ได้พิสูจน์อัตราของมันในหัวข้อนี้

สำหรับโซ่เจ็ดลูก เริ่มจากเวกเตอร์ที่เป็นโหมดที่ 2 บวกโหมดที่ 7 ประมาณ 0.2 เท่า แล้วทำวิธีนี้ สามรอบแรกให้ค่า σ คือ 0.711247, 0.585995, 0.585786 และความคลาดเคลื่อนของค่าที่ได้คือ 0.1255, 2.09×10^{-4} และ 8.6×10^{-13} ตามลำดับ (ตัวเลขจากโค้ดของเรา) ลำดับนี้ลดแบบกำลังสามในตัวอย่างนี้ แต่เป็นการสังเกต ยังไม่ใช่ข้อพิสูจน์

ฉบับร่าง · Preprint · 2026-10-23

สรุปใจความ

- วิธีกำลังคูณเวกเตอร์กับเมทริกซ์แล้วปรับขนาดซ้ำ ทิศจะหันเข้าหาทิศเด่น อัตราการลู่เข้าคือ $|\lambda_2/\lambda_1|$ (หัวข้อย่อยแรก)
- คะแนนของหน้าเว็บคือทิศเด่นของเมทริกซ์การคลิก หน้า 5 อยู่บนสุดในตัวอย่างของเรา (อีกด้านของเจลลี่)
- การกระโดดแบบ $1-d$ ทำให้วงจรของหน้าไม่ค้างอยู่ที่เดิม ค่าลักษณะเฉพาะที่ขนาดเท่ากันทำให้วิธีกำลังไม่ลู่เข้า
- การวนซ้ำแบบผกผันหาทิศที่อยู่ใกล้ค่าที่เดา และการเลื่อนค่าเปลี่ยนอัตราการลู่เข้า แต่ต้องเลือกค่าเลื่อนไม่ให้ขนาดเท่ากัน
- การวนซ้ำผลหารเรย์ลีเปลี่ยนค่าเดาทุกรอบ และลู่เข้าเร็วขึ้นสำหรับเมทริกซ์สมมาตร

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 9.2.1 (⋆) ให้ $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ และ $\mathbf{x}_0 = (1,0)^T$

1. จงคำนวณ $A\mathbf{x}_0$, $A^2\mathbf{x}_0$ และ $A^3\mathbf{x}_0$ โดยไม่ต้องหาค่าลักษณะเฉพาะ
2. จงปรับ $A^3\mathbf{x}_0$ ให้มีความยาวหนึ่ง แล้วเทียบกับ $\frac{1}{\sqrt{2}}(1,1)^T$ ซึ่งเป็นทิศของค่าลักษณะเฉพาะที่ใหญ่ที่สุด

แบบฝึกหัด 9.2.2 (⋆) ใช้เวกเตอร์ $\mathbf{x}_2 = (5,4)^T$ จากแบบฝึกหัด 9.2.1 จงคำนวณผลหารเรย์ลี $R(\mathbf{x}_2)$ แล้วเทียบกับค่าลักษณะเฉพาะที่ใหญ่ที่สุดของ A (ไม่จำเป็นต้องปรับขนาด $\mathbf{x}_2$ ก่อน)

แบบฝึกหัด 9.2.3 (⋆⋆, พิสูจน์) ให้ $P_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ และ $0 < d < 1$ ให้ $G_2 = dP_2 + \frac{1-d}{2}\mathbf{1}\mathbf{1}^T$ เมื่อ $\mathbf{1} = (1,1)^T$

1. จงแสดงว่า $G_2\mathbf{1} = \mathbf{1}$ และ $G_2(1,-1)^T = -d\,(1,-1)^T$
2. เริ่มจาก $\mathbf{x}_0 = (1,0)^T$ จงแสดงว่าสมาชิกตัวแรกของ $G_2^k\mathbf{x}_0$ เท่ากับ $\frac{1}{2} + \frac{1}{2}(-d)^k$ แล้วหาความคลาดเคลื่อนจาก $\frac{1}{2}$ เมื่อ $d = 0.85$ และ $k = 5$

แบบฝึกหัด 9.2.4 (⋆⋆, พิสูจน์) ให้ P เป็นเมทริกซ์ขนาด $n \times n$ ที่สมาชิกทุกตัวไม่ติดลบ และผลรวมของแต่ละคอลัมน์เป็นหนึ่ง ให้ $0 < d < 1$ และ $G = dP + \frac{1-d}{n}\mathbf{1}\mathbf{1}^T$

1. จงแสดงว่าผลรวมของแต่ละคอลัมน์ของ G เป็นหนึ่ง
2. จงแสดงว่าทุกสมาชิกของ G มีค่าอย่างน้อย $\frac{1-d}{n}$ จึงเป็นบวก
3. จงแสดงว่า $\mathbf{1}^T G = \mathbf{1}^T$ แล้วสรุปว่า 1 เป็นค่าลักษณะเฉพาะของ G

ฉบับร่าง Preprint: 2026-10-23

แบบฝึกหัด 9.2.5 (⋆⋆, คำนวณและตีความ) ใช้หน้าเว็บหกหน้าของหัวข้อนี้ (ลิงก์เดิม) แต่เปลี่ยนตัวประกอบการหน่วงเป็น $d = 0.5$ และ $d = 0.99$

1. จงเรียงลำดับหน้าทั้งหกจากคะแนนสูงไปต่ำที่ $d = 0.5$ และที่ $d = 0.99$ แล้วเทียบกับ $d = 0.85$
2. จงหาจำนวนรอบ k ที่น้อยที่สุด ซึ่งผลรวมของ $|\mathbf{x}_k - \mathbf{r}|$ ไม่เกิน 10^{-6} เมื่อ $\mathbf{r}$ คือคะแนนนิ่ง และ $\mathbf{x}_0$ เป็นคะแนนเท่ากันทุกหน้า
3. เมื่อ d เพิ่มขึ้น จำนวนรอบเปลี่ยนไปอย่างไร และลำดับของหน้าสองหน้าที่ใกล้กันมากเปลี่ยนหรือไม่ ตอบด้วยตัวเลขที่ได้

แบบฝึกหัด 9.2.6 (⋆ ⋆ ⋆, สิ่งที่ผิดพลาดได้) ให้ A เป็นเมทริกซ์โซ่เจ็ดลูกของหัวข้อ 9.1 ($\sigma = 0.6$) และเริ่มจาก $\mathbf{x}_0 = \frac{1}{\sqrt{7}}(1, 1, \ldots, 1)^T$

1. จงแสดงว่า $\mathbf{v}_2 = (\sin\frac{\pi}{4}, \sin\frac{2\pi}{4}, \ldots, \sin\frac{7\pi}{4})^T$ ตั้งฉากกับ $(1, 1, \ldots, 1)^T$ ดังนั้นในเลขคณิตที่แม่นยำ $\mathbf{x}_0$ ไม่มีส่วนประกอบบน $\mathbf{v}_2$
2. ในเลขคณิตที่แม่นยำ การวนซ้ำแบบผกผันจะลู่เข้าค่าลักษณะเฉพาะใด และเพราะเหตุใด
3. ในการคำนวณจริง ค่าประมาณจะไปอยู่ที่ค่าลักษณะเฉพาะที่ผิดก่อน แล้วกระโดดไปค่าที่ถูกต้องภายหลัง จงอธิบายด้วยอัตราขยาย $|\lambda_1 - \sigma|/|\lambda_2 - \sigma|$ และขนาดของการปัดเศษ 10^{-16} ว่าการกระโดดเกิดราวรอบที่เท่าไร

ตัวช่วยคิด

สำหรับข้อ 1 จับคู่พจน์ i กับ $8 - i$ และสังเกต $\sin(2\pi - x) = -\sin x$ สำหรับข้อ 3 ลองหาจำนวนรอบ m ที่ 31.5^m เท่ากับ 10^{16}

แบบฝึกหัด 9.2.7 (⋆⋆, พิสูจน์) ให้ A มีเวกเตอร์ลักษณะเฉพาะ $\mathbf{v}_i$ ที่สอดคล้องกับ λ_i และให้ σ ไม่ตรงกับ λ_i ใด

1. จงแสดงว่า $(A - \sigma I)^{-1}\mathbf{v}_i = \frac{1}{\lambda_i - \sigma}\mathbf{v}_i$ ดังนั้น $(A - \sigma I)^{-1}$ มีเวกเตอร์ลักษณะเฉพาะชุดเดียวกับ A
2. จงแสดงว่าการวนซ้ำแบบผกผันหนึ่งรอบเท่ากับวิธีกำลังของ $B = (A - \sigma I)^{-1}$ ที่ปรับขนาดแบบเดียวกัน
3. ในทางปฏิบัติ เราไม่ต้องคำนวณ B ทั้งก้อน จงอธิบายว่าทำไมการแก้สมการ $(A - \sigma I)\mathbf{w} = \mathbf{x}$ ก็เพียงพอ

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 9.2.1.

1. $A\mathbf{x}_0 = (2, 1)^T$ จากนั้น $A^2\mathbf{x}_0 = A(2, 1)^T = (5, 4)^T$ และ $A^3\mathbf{x}_0 = A(5, 4)^T = (14, 13)^T$
2. ความยาวของ $(14, 13)^T$ คือ $\sqrt{365} \approx 19.105$ จึงได้เวกเตอร์หนึ่งหน่วย $(0.732793,\ 0.680451)^T$ ส่วน $\frac{1}{\sqrt{2}}(1, 1)^T \approx (0.707107,\ 0.707107)^T$ ระยะห่างระหว่างสองเวกเตอร์นี้ลดจาก 0.1106 ที่ $\mathbf{x}_2$ เป็น 0.0370 ที่ $\mathbf{x}_3$ คืออัตราราว $\frac{1}{3}$ ซึ่งเท่ากับ $\lambda_2/\lambda_1 = 1/3$ ของค่าลักษณะเฉพาะ 3 และ 1 ของ A

ตรวจ ด้วย โค้ด: ใน `test_ch09_power.py`(`test_exercise_921_normalised_iterates_approach_one_one_slowly`; ทดสอบเวกเตอร์ด้วย `test_exercise_921_hand_matrix_power_two_by_two`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 9.2.2. $\mathbf{x}_2^T A\mathbf{x}_2 = (5,4)\cdot(14,13) = 70+52 = 122$ และ $\mathbf{x}_2^T\mathbf{x}_2 = 25+16 = 41$ จึง

$$R(\mathbf{x}_2) = \frac{122}{41} \approx 2.9756.$$

ค่านี้ต่างจาก 3 อยู่เพียง 0.0244 ทั้งที่ทิศของ $\mathbf{x}_2/\|\mathbf{x}_2\|_2$ ยังห่างจาก $\frac{1}{\sqrt{2}}(1,1)^T$ ราว 0.11 นี่คือเหตุผลที่ใช้ผลหารเรย์ลีเป็นค่าประมาณค่าลักษณะเฉพาะได้ดีกว่าตัวเวกเตอร์เอง เพราะ A สมมาตร

ตรวจด้วยโค้ด: ใน `test_ch09_power.py`(`test_exercise_922_rayleigh_quotient_of_x2`)

เฉลย 9.2.3.

1. สมาชิกของ G_2 คือ $\frac{1-d}{2}$ บนแนวทแยงมุม และ $d+\frac{1-d}{2} = \frac{1+d}{2}$ นอกแนวทแยงมุม ดังนั้น $G_2\mathbf{1} = \mathbf{1}$ (แถวทั้งสองรวมเป็น $\frac{1-d}{2}+\frac{1+d}{2}=1$) และ

$$G_2(1,-1)^T = \left(\tfrac{1-d}{2} - \tfrac{1+d}{2}\right)(1,-1)^T = -d\,(1,-1)^T.$$

2. เขียน $(1,0)^T = \frac{1}{2}(1,1)^T + \frac{1}{2}(1,-1)^T$ แล้วใช้ข้อ 1 ซ้ำ k ครั้ง ได้ $G_2^k\mathbf{x}_0 = \frac{1}{2}(1,1)^T + \frac{1}{2}(-d)^k(1,-1)^T$ สมาชิกตัวแรกคือ $\frac{1}{2}+\frac{1}{2}(-d)^k$ ความคลาดเคลื่อนจาก $\frac{1}{2}$ คือ $\frac{1}{2}d^k$ เมื่อ $d=0.85$ และ $k=5$ ได้

$$\tfrac{1}{2}(0.85)^5 = \tfrac{1}{2}\times 0.443705 \approx 0.22185.$$

   เครื่องหมายแกว่งรอบ $\frac{1}{2}$ แต่ขนาดของความคลาดเคลื่อนลดลงทุกรอบ

ตรวจด้วยโค้ด: ใน `test_ch09_power.py`(`test_exercise_923_proof_error_formula_for_damped_cycle_at_k_5`) และ `test_cycle_damped_error_is_0_85_power_k`

เฉลย 9.2.4.

1. ผลรวมของคอลัมน์ j ของ G คือ $d\sum_i P_{ij} + \frac{1-d}{n}\cdot n = d\cdot 1 + (1-d) = 1$
2. เพราะ $P_{ij}\ge 0$ จึง $G_{ij}\ge\frac{1-d}{n}$ และ $\frac{1-d}{n}>0$ เพราะ $d<1$ ดังนั้นทุกสมาชิกของ G เป็นบวก
3. จากข้อ 1 ผลรวมแต่ละคอลัมนเป็นหนึ่ง นั่นคือ $\mathbf{1}^TG = \mathbf{1}^T$ หรือ $G^T\mathbf{1}=\mathbf{1}$ จึงได้ว่า 1 เป็นค่าลักษณะเฉพาะของ G^T และ G กับ G^T มีค่าลักษณะเฉพาะชุดเดียวกัน ดังนั้น 1 เป็นค่าลักษณะเฉพาะของ G

ตรวจ ด้วย โค้ด: ใน `test_ch09_power.py`(`test_exercise_924_damping_positive_and_column_sums`) ตรวจผลรวมคอลัมน์และขอบเขต $\frac{1-d}{6}$ กับเว็บหกหน้าที่ $d=0.5$

เฉลย 9.2.5.

1. ที่ $d=0.5$ คะแนนนิ่งประมาณ $(0.1375,\ 0.1177,\ 0.2167,\ 0.1538,\ 0.2164,\ 0.1579)$ ลำดับคือ หน้า $3,5,6,4,1,2$ ที่ $d=0.99$ คะแนนประมาณ $(0.1139,\ 0.0580,\ 0.2267,\ 0.1509,\ 0.2814,\ 0.1692)$ ลำดับคือ $5,3,6,4,1,2$ ซึ่งเหมือนกับที่ $d=0.85$ (ตัวเลขจากโค้ดของเรา) การสลับลำดับที่ $d=0.5$ เกิดขึ้นเพราะหน้า 3 และหน้า 5 ต่างกันเพียง 0.0003 จึงเป็นคู่ที่ใกล้กันมาก
2. จำนวนรอบที่ต้องใช้คือ 13 เมื่อ $d=0.5$, 27 เมื่อ $d=0.85$ และ 38 เมื่อ $d=0.99$ (ตัวเลขจากโค้ดของเรา)
3. เมื่อ d เพิ่มขึ้น ค่า $|\lambda_2|$ ของ G เพิ่มจาก 0.360 เป็น 0.612 และ 0.713 จำนวนรอบจึงมากขึ้น หมายความว่า d มากขึ้นทำให้การกระโดดในแต่ละรอบน้อยลง และคะแนนลู่เข้าช้าลง ส่วนลำดับหน้าที่ห่างกันมากไม่เปลี่ยน มีเพียงคู่ที่ใกล้กันที่สลับได้

ตรวจด้วยโค้ด: ใน `test_ch09_power.py`(`test_exercise_925_damping_variants_of_the_toy_web`)

ฉบับร่าง · preprint · 2026-10-23

เฉลย 9.2.6.

1. สมาชิกของ $\mathbf{v}_2$ คือ $\sin\frac{i\pi}{4}$ เมื่อ $i = 1, \ldots, 7$ พจน์ i กับ $8-i$ มีค่าตรงข้ามกัน เพราะ $\sin(2\pi - x) = -\sin x$ และพจน์กลาง $i = 4$ เป็น $\sin\pi = 0$ ผลรวมของสมาชิกจึงเป็นศูนย์ ดังนั้น $\mathbf{v}_2 \perp \mathbf{1}$ ในเลขคณิตที่แม่นยำ $\mathbf{x}_0$ ไม่มีส่วนประกอบบน $\mathbf{v}_2$

2. ส่วนประกอบทุกรอบบน $\mathbf{v}_2$ ยังคงเป็นศูนย์ ดังนั้นการวนซ้ำลู่เข้าค่าลักษณะเฉพาะที่ใกล้ 0.6 ที่สุดในกลุ่มโหมดคี่ คือ $\lambda_1 = 0.152241$ เพราะระยะทาง $|\lambda_1 - 0.6| = 0.4478$ น้อยกว่า $|\lambda_3 - 0.6| = 0.6346$ (ค่า $\lambda_2 = 0.585786$ ไม่ถูกเลือก)

3. ในการคำนวณจริงมีส่วนประกอบบน $\mathbf{v}_2$ ราว 10^{-16} จากการปัดเศษ อัตราขยายต่อรอบของส่วนนี้เทียบกับส่วนบน $\mathbf{v}_1$ คือ $\frac{|\lambda_1-\sigma|}{|\lambda_2-\sigma|} = \frac{0.4478}{0.0142} \approx 31.5$ เมื่อผ่าน m รอบ ส่วนประกอบนี้โตราว $10^{-16}\cdot 31.5^m$ ซึ่งถึงระดับ 1 เมื่อ $m \approx \frac{16\ln 10}{\ln 31.5} \approx 10.7$ จึงเกิดการกระโดดราวรอบที่ 11 ถึง 12 ซึ่งตรงกับที่โค้ดของเราเห็น (ค่าประมาณ 0.1524 ที่รอบ 10 แล้วไป 0.5858 ที่รอบ 13)

ตรวจด้วยโค้ด: ใน `test_ch09_power.py`(`test_exercise_926_symmetric_start_has_no_mode_two_component_in_exact_arithmetic`) และ `test_inverse_iteration_symmetric_start_finds_the_wrong_mode_first`

เฉลย 9.2.7.

1. จาก $(A-\sigma I)\mathbf{v}_i = (\lambda_i-\sigma)\mathbf{v}_i$ คูณทั้งสองข้างด้วย $(A-\sigma I)^{-1}$ แล้วหารด้วย $\lambda_i-\sigma \neq 0$ จะได้ $(A-\sigma I)^{-1}\mathbf{v}_i = \frac{1}{\lambda_i-\sigma}\mathbf{v}_i$ ดังนั้นเวกเตอร์ลักษณะเฉพาะของ $(A-\sigma I)^{-1}$ คือชุดเดียวกับ A

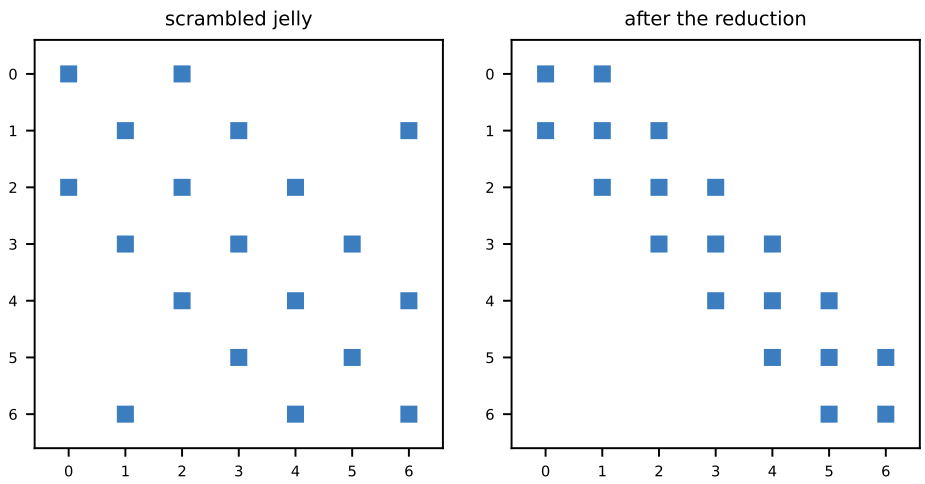
2. ถ้า $\mathbf{x}_{m-1}$ คือเวกเตอร์ที่ปรับขนาดแล้ว การวนซ้ำแบบผกผันหาได้ $\mathbf{w} = (A-\sigma I)^{-1}\mathbf{x}_{m-1}$ และปรับขนาดเป็น $\mathbf{x}_m = \mathbf{w}/\|\mathbf{w}\|_2$ ซึ่งคือวิธีกำลังของ B โดยเริ่มจาก $\mathbf{x}_0$ ที่ปรับขนาดแล้วทุกรอบ

3. การแก้สมการ $(A-\sigma I)\mathbf{w} = \mathbf{x}$ ให้เวกเตอร์เดียวกับ $(A-\sigma I)^{-1}\mathbf{x}$ โดยตรง ไม่ต้องหาเมทริกซ์ผกผันทั้งก้อน และการแยกตัวประกอบ LU ของ $A-\sigma I$ ทำได้ครั้งเดียวแล้วใช้ซ้ำทุกรอบ

ตรวจ ด้วย โค้ด: ใน `test_ch09_power.py`(`test_exercise_927_inverse_iteration_is_power_method_on_inverse`)

อ่านต่อ. หนังสือ Atkinson 1989, §9.2, pp. 602--609 ให้วิธีกำลัง การเลื่อนค่า และผลหารเรย์ลีในรายละเอียดมากกว่านี้ และ Atkinson 1989, §9.6, pp. 628--633 ให้การวนซ้ำแบบผกผันสำหรับการหาเวกเตอร์ลักษณะเฉพาะ หัวข้อ 9.3 จะถามต่อว่า ถ้าเราต้องการค่าลักษณะเฉพาะทุกตัวพร้อมกัน ไม่ใช่เพียงตัวเดียว ควรทำอย่างไร

ฉบับร่าง Preprint · 2026-10-23



รูปที่ 9.6: ซ้าย: เมทริกซ์ของเจลลี่ที่ติดหมายเลขใหม่ จุดแต่ละจุดคือสมาชิกที่ไม่เป็นศูนย์ ไม่มีจุดใดอยู่ข้างแนวทแยงมุม ขวา: หลังการลดรูปด้วยการสะท้อน ได้เมทริกซ์สามแนวเฉียงที่มีค่าลักษณะเฉพาะชุดเดียวกัน (สร้างจาก `numanal/ch09_qr.py`)

9.3 ขั้นตอนวิธี QR: หาทุกโหมดพร้อมกัน

สถานการณ์สมมติ: เลขลูกบอลสลับกัน แต่โหมดไม่เปลี่ยน

ในหัวข้อ 9.2 วิธีกำลังหาโหมดที่แรงที่สุดได้ทีละหนึ่งทิศ คราวนี้เราอยากรู้โหมดทั้งเจ็ดของเจลลี่พร้อมกัน ทั้งโหมดที่สั่นเร็วและโหมดที่สั่นช้า

สมมติว่าเราสร้างแบบจำลองโดยติดหมายเลขลูกบอลตามลำดับที่สะดวกของโปรแกรม ไม่ใช่ตามลำดับจากกำแพงซ้ายไปขวา ลูกบอลทุกลูกยังต่อกับเพื่อนบ้านเหมือนเดิม แต่เมทริกซ์ที่เขียนลงไปไม่มีสมาชิกใดอยู่ข้างแนวทแยงมุมเลย ดูไม่ออกว่าใครต่อกับใคร คำถามคือ เราจะหาค่าลักษณะเฉพาะทั้งเจ็ดได้อย่างไร โดยไม่ต้องรู้ลำดับที่แท้จริงก่อน

(แบบจำลองและการติดหมายเลขนี้เป็นของเราเอง ไม่ได้บอกว่าโปรแกรมใดทำอย่างไร ตัวเลขทั้งหมดในหัวข้อนี้มาจากโค้ดของเรา)

ภาพรวมก่อนอ่าน

เมทริกซ์ที่คล้ายกันมีค่าลักษณะเฉพาะชุดเดียวกัน ดังนั้นเราเปลี่ยนรูปเมทริกซ์ให้ง่ายขึ้นได้โดยค่าลักษณะเฉพาะไม่เปลี่ยน ขั้นตอนมีสามช่วง ช่วงแรกสะท้อนเวกเตอร์ซ้ำ ๆ จนได้เมทริกซ์สามแนวเฉียง ช่วงที่สองแยกเมทริกซ์เป็นเมทริกซ์เชิงตั้งฉากคูณสามเหลี่ยมบน แล้วคูณกลับลำดับ ช่วงที่สามทำซ้ำ ค่าบนแนวทแยงมุมจะค่อย ๆ ลู่เข้าค่าลักษณะเฉพาะ เมื่อสมาชิกมุมล่างนิ่งแล้วเราตัดแถวนั้นทิ้งและทำต่อกับเมทริกซ์ที่เล็กลง

ฉบับร่าง · Preprint · 2026-10-23

สะท้อนเวกเตอร์ให้ไปที่แกนเดียว

เครื่องมือแรกคือการสะท้อนแบบเฮาส์โฮลเดอร์ (Householder reflection) ที่เราเจอในบทที่ 8 ลองนึกถึงกระจกที่ตั้งฉากกับเวกเตอร์ $\mathbf{w}$ เวกเตอร์ที่ขนานกับ $\mathbf{w}$ จะกลับด้าน ส่วนเวกเตอร์ที่ตั้งฉากกับ $\mathbf{w}$ จะไม่ขยับเลย

นิยาม 9.3.1: การสะท้อนแบบเฮาส์โฮลเดอร์

ให้ $\mathbf{w} \in \mathbb{R}^n$ มีความยาว $\|\mathbf{w}\|_2 = 1$ เมทริกซ์ $U = I - 2\mathbf{w}\mathbf{w}^T$ เรียกว่าการสะท้อนแบบเฮาส์โฮลเดอร์

อ่านแบบเข้าใจง่าย. ถ้า $\mathbf{v} \perp \mathbf{w}$ แล้ว $\mathbf{w}^T\mathbf{v} = 0$ จึง $U\mathbf{v} = \mathbf{v}$ ส่วน $U\mathbf{w} = -\mathbf{w}$ เพราะ $\mathbf{w}^T\mathbf{w} = 1$ เมทริกซ์ U สมมาตร และ $U^2 = I$ จึงเป็นเมทริกซ์เชิงตั้งฉาก (แบบฝึกหัด 9.3.3 ให้ตรวจเอง)

ทฤษฎีบท 9.3.1: การสะท้อนที่ส่งเวกเตอร์ไปแกนแรก

ให้ $\mathbf{b} \neq \mathbf{0}$ ให้ $\alpha = -\operatorname{sgn}(b_1)\,\|\mathbf{b}\|_2$ โดยกำหนด $\operatorname{sgn}(0) = 1$ ให้ $\mathbf{u} = \mathbf{b} - \alpha\mathbf{e}_1$ และ $\mathbf{w} = \mathbf{u}/\|\mathbf{u}\|_2$ แล้ว $U = I - 2\mathbf{w}\mathbf{w}^T$ ทำให้ $U\mathbf{b} = \alpha\mathbf{e}_1$

พิสูจน์. **แนวคิด.** ใช้ว่า $\|\mathbf{u}\|_2^2 = 2\,\mathbf{u}^T\mathbf{b}$ ซึ่งเป็นผลจาก $\alpha^2 = \|\mathbf{b}\|_2^2$

1. **ตัวหารไม่เป็นศูนย์.** เครื่องหมายของ α ตรงข้ามกับ b_1 (หรือ $\alpha < 0$ เมื่อ $b_1 = 0$) จึง $|u_1| = |b_1| + \|\mathbf{b}\|_2 > 0$ ดังนั้น $\mathbf{u} \neq \mathbf{0}$
2. **สองปริมาณ.** $\mathbf{u}^T\mathbf{b} = \|\mathbf{b}\|_2^2 - \alpha b_1$ และ $\|\mathbf{u}\|_2^2 = \|\mathbf{b}\|_2^2 - 2\alpha b_1 + \alpha^2 = 2\,\mathbf{u}^T\mathbf{b}$
3. **คำนวณ.** $U\mathbf{b} = \mathbf{b} - 2\mathbf{w}(\mathbf{w}^T\mathbf{b}) = \mathbf{b} - \dfrac{2\,\mathbf{u}^T\mathbf{b}}{\|\mathbf{u}\|_2^2}\,\mathbf{u} = \mathbf{b} - \mathbf{u} = \alpha\mathbf{e}_1$

สรุป: บทพิสูจน์นี้เราเขียนเอง ตามแนวของ Atkinson 1989, §9.3, pp. 610--612 เครื่องหมายของ α ที่ตรงข้ามกับ b_1 ทำให้ไม่มีการหักล้างใน u_1 (ดูข้อควรระวังด้านล่าง) □

อ่านแบบเข้าใจง่าย. ความยาวของ $U\mathbf{b}$ ต้องเท่ากับความยาวของ $\mathbf{b}$ เสมอ ดังนั้นเราเลือก α ให้มีขนาด $\|\mathbf{b}\|_2$ ส่วนเครื่องหมายเลือกให้ตรงข้ามกับ b_1 เพื่อให้ $\mathbf{u}$ ไม่เล็กเกินไป ขั้นตอนนี้ทำให้สมาชิกใต้ตำแหน่งแรกของ $\mathbf{b}$ เป็นศูนย์ได้ในครั้งเดียว

ลองเอง

ให้ $\mathbf{b} = (-1, 2, 2)^T$ มีความยาว 3 และ $b_1 < 0$ จึงได้ $\alpha = +3$ และ $\mathbf{u} = (-4, 2, 2)^T$ ได้ $\mathbf{w} = (-2, 1, 1)^T/\sqrt{6}$ และ

$$U = \frac{1}{3}\begin{pmatrix} -1 & 2 & 2 \\ 2 & 2 & -1 \\ 2 & -1 & 2 \end{pmatrix}, \qquad U\mathbf{b} = (3, 0, 0)^T.$$

ลองคูณเองให้ครบ แล้วตรวจว่า U สมมาตร และ $U^2 = I$

ข้อควรระวัง: เลือกเครื่องหมายผิดแล้วศูนย์กลายเป็นค่าเล็กที่ผิด

ให้ $\mathbf{b} = (1, 10^{-8}, 0)^T$ ในทางคณิตศาสตร์ทั้งสองเครื่องหมายของ α ให้ผล $U\mathbf{b} = \alpha\mathbf{e}_1$ แต่ในเลขทศนิยมลอยตัว 64 บิต $\|\mathbf{b}\|_2$ ปัดเป็น 1 เพราะ $1 + 10^{-16}$ ไม่ต่างจาก 1 ถ้าเลือก $\alpha = +1$ จะได้ $u_1 = 1 - 1 = 0$ การสะท้อน

ฉบับร่าง · Preprint · 2026-10-23

กลายเป็น $\mathbf{w} = \mathbf{e}_2$ และได้ $U\mathbf{b} = (1, -10^{-8}, 0)^T$ ค่าที่ควรเป็นศูนย์กลายเป็น -10^{-8} ค่าคลาดเคลื่อนสัมบูรณ์จึงเป็น 10^{-8} ถ้าเลือก $\alpha = -1$ จะได้ $U\mathbf{b} = (-1, 0, 0)^T$ โดยไม่มีการหักล้าง (ตัวเลขจากโค้ดของเรา)

แยก $A = QR$: เมทริกซ์เชิงตั้งฉากคูณสามเหลี่ยมบน

การแยกตัวประกอบ QR ที่เราเห็นในบทที่ 8 เขียนเมทริกซ์เป็น $A = QR$ โดย Q เป็นเมทริกซ์เชิงตั้งฉาก และ R เป็นเมทริกซ์สามเหลี่ยมบน คือสมาชิกใต้แนวทแยงมุมเป็นศูนย์ทั้งหมด การแยกนี้ทำได้เสมอสำหรับเมทริกซ์จริงจัตุรัส

ทฤษฎีบท 9.3.2: การแยกตัวประกอบ QR

ให้ A เป็นเมทริกซ์จริงขนาด $n \times n$ จะมีเมทริกซ์เชิงตั้งฉาก Q และเมทริกซ์สามเหลี่ยมบน R ที่ $A = QR$

พิสูจน์. **แนวคิด.** ใช้การสะท้อนทีละคอลัมน์ ครั้งที่ k ทำให้สมาชิกใต้แนวทแยงมุมของคอลัมน์ k เป็นศูนย์ โดยไม่แตะแถวก่อนหน้า

1. **ครั้งที่ k.** ให้ $B_{k-1} = P_{k-1} \cdots P_1 A$ ใช้ทฤษฎีบทการสะท้อนกับเวกเตอร์ใต้ตำแหน่ง (k, k) ของคอลัมน์ k ของ B_{k-1} ได้การสะท้อน P_k ที่ไม่แตะตำแหน่งสมาชิกของแถว $1, \ldots, k-1$
2. **ศูนย์เดิมคงอยู่.** เพราะ P_k ไม่แตะแถว $1, \ldots, k-1$ สมาชิกใต้แนวทแยงมุมของคอลัมน์ $1, \ldots, k-1$ ที่เป็นศูนย์แล้วจึงยังเป็นศูนย์
3. **สรุป.** หลัง $n-1$ ครั้งได้ $P_{n-1} \cdots P_1 A = R$ เป็นสามเหลี่ยมบน ให้ $Q = P_1 P_2 \cdots P_{n-1}$ เชิงตั้งฉาก เพราะ $P_k^T = P_k = P_k^{-1}$ จึงได้ $A = QR$

สรุป: เราเขียนแนวคิดนี้เอง รายละเอียดการเลือกเครื่องหมายและการนับการคำนวณอยู่ใน Atkinson 1989, §9.3, pp. 612--614 □

ลองเอง

ให้ $A = \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix}$ คอลัมน์แรก $(3, 4)^T$ มีความยาว 5 จึงได้ $\alpha = -5$ และ $Q = \begin{pmatrix} -3/5 & -4/5 \\ -4/5 & 3/5 \end{pmatrix}$ ได้ $R = \begin{pmatrix} -5 & -11/5 \\ 0 & 2/5 \end{pmatrix}$ ตรวจว่า $QR = A$ และ $|\det A| = 2 = |R_{11} R_{22}|$

การหมุนกีฟเวนส์ที่เห็นในบทที่ 8 คือเมทริกซ์ $\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$ เมื่อ $c^2 + s^2 = 1$ มันหมุนระนาบของคู่แกนหนึ่งคู่เพื่อทำให้สมาชิกหนึ่งตัวเป็นศูนย์ เช่น เมื่อ $c = \frac{3}{5}$ และ $s = \frac{4}{5}$ การหมุนเปลี่ยน $(3, 4)^T$ เป็น $(5, 0)^T$ สำหรับเมทริกซ์สามแนวเฉียงหรือเฮสเซนเบิร์ก การหมุนแบบนี้ประหยัดกว่าการสะท้อนในขั้นตอน QR Atkinson 1989, §9.3, p. 619, §9.5, p. 625

ฉบับร่าง · Preprint · 2026-10-23

ลดรูปก่อนทำซ้ำ: เมทริกซ์เฮสเซนเบิร์กและสามแนวเฉียง

ถ้าเราแยก QR กับเมทริกซ์เต็มทุกรอบ แต่ละรอบจะแพงมาก ตำราจึงลดรูปเมทริกซ์ให้ง่ายขึ้นก่อนด้วย **การแปลงภาวะคล้าย** (similarity transformation) ที่ใช้เมทริกซ์เชิงตั้งฉาก Q คือเปลี่ยน A เป็น Q^TAQ ค่าลักษณะเฉพาะไม่เปลี่ยน เพราะถ้า $A\mathbf{v}=\lambda\mathbf{v}$ แล้ว $(Q^TAQ)(Q^T\mathbf{v}) = Q^TA\mathbf{v} = \lambda Q^T\mathbf{v}$ (ใช้ $QQ^T=I$)

นิยาม 9.3.2: เมทริกซ์เฮสเซนเบิร์ก

เมทริกซ์จัตุรัส $H=[h_{ij}]$ เรียกว่า **เมทริกซ์เฮสเซนเบิร์ก** (Hessenberg matrix) ถ้า $h_{ij}=0$ ทุกคู่ (i,j) ที่ $i>j+1$

อ่านแบบเข้าใจง่าย. เมทริกซ์เฮสเซนเบิร์กเป็นสามเหลี่ยมบนเกือบทั้งหมด ยกเว้นแนวที่อยู่ติดใต้แนวทแยงมุมหนึ่งแนว ถ้ามันสมมาตรด้วย ก็คือเมทริกซ์สามแนวเฉียง

ทฤษฎีบท 9.3.3: การลดรูปด้วยการสะท้อน

ให้ A เป็นเมทริกซ์จริงขนาด $n\times n$ จะมีเมทริกซ์เชิงตั้งฉาก Q ที่ทำให้ $H=Q^TAQ$ เป็นเมทริกซ์เฮสเซนเบิร์ก ถ้า A สมมาตร H เป็นเมทริกซ์สามแนวเฉียงที่สมมาตร

พิสูจน์. **แนวคิด.** ทำเหมือนการแยก QR แต่สะท้อนทั้งจากซ้ายและจากขวา การสะท้อนครั้งที่ k ใช้เฉพาะดัชนี $k+1,\dots,n$

1. **ครั้งที่ k.** ให้ U_k เป็นการสะท้อนบนดัชนี $k+1,\dots,n$ ที่ทำให้สมาชิกใต้ $(k+1,k)$ ของคอลัมน์ k เป็นศูนย์ และให้ P_k เป็นเมทริกซ์ที่บล็อกล่างขวาเป็น U_k ส่วนที่เหลือเป็นเอกลักษณ์ แล้ว $H_k = P_kH_{k-1}P_k$
2. **ศูนย์เดิมคงอยู่.** การคูณจากซ้ายไม่แตะแถว $1,\dots,k$ และการคูณจากขวาไม่แตะคอลัมน์ $1,\dots,k$ จึงคงศูนย์ของคอลัมน์ $1,\dots,k-1$ ไว้
3. **สมมาตร.** ถ้า A สมมาตร H_k สมมาตรด้วย เพราะ $(P_kH_{k-1}P_k)^T = P_kH_{k-1}^TP_k$ ศูนย์ใต้แนวทแยงมุมของคอลัมน์ k จึงสะท้อนเป็นศูนย์เหนือแนวทแยงมุมของแถว k ด้วย

สรุป: หลัง $n-2$ ครั้งได้เมทริกซ์เฮสเซนเบิร์ก และถ้าสมมาตรได้สามแนวเฉียง ดู Atkinson 1989, §9.3, pp. 615--617 สำหรับกรณีสมมาตร และ Atkinson 1989, §9.5, p. 624 สำหรับนิยามเฮสเซนเบิร์ก □

ตัวอย่าง. ให้ $A=\begin{pmatrix}1&3&4\\3&2&0\\4&0&5\end{pmatrix}$ คอลัมน์แรกใต้แนวทแยงมุมคือ $(3,4)^T$ จึงใช้การสะท้อนเพียงครั้งเดียวกับ $\alpha=-5$ และได้

$$T=\begin{pmatrix}1&-5&0\\-5&\frac{98}{25}&-\frac{36}{25}\\0&-\frac{36}{25}&\frac{77}{25}\end{pmatrix}=\begin{pmatrix}1&-5&0\\-5&3.92&-1.44\\0&-1.44&3.08\end{pmatrix}.$$

ตรวจได้ว่า $\operatorname{tr}T = 8 = \operatorname{tr}A$ และผลรวมกำลังสองของสมาชิกทั้งหมดเป็น 80 ทั้งสองกรณี เพราะการแปลงภาวะคล้ายด้วยเมทริกซ์เชิงตั้งฉากไม่เปลี่ยนค่าเหล่านี้

ขั้นตอนนี้ทำเพียงครั้งเดียว แล้วทุกรอบของ QR ที่ตามมาทำบนเมทริกซ์เฮสเซนเบิร์กหรือสามแนวเฉียง ซึ่งถูกกว่ามาก Atkinson เรียกการลดรูปนี้ว่าการดำเนินการที่เสถียรมาก Atkinson 1989, §9.3, Thm 9.4, pp. 617--618

สำหรับเจลลี่ที่ติดหมายเลขใหม่ ผลการลดรูปคือ $T=DCD$ โดย D เป็นเมทริกซ์ทแยงมุมของ ±1 และ C คือโซ่ของหัวข้อ 9.1 นั่นคือ T เป็นโซ่เดิมที่เครื่องหมายของสมาชิกนอกแนวทแยงมุมบางตัวต่างออกไป ซึ่งมาจากการเลือกทิศการขยับของลูกบอลแต่ละลูก ค่าลักษณะเฉพาะจึงยังเป็น $2-2\cos\!\left(\frac{j\pi}{8}\right)$ (ตัวเลขจากโค้ดของเรา)

ฉบับร่าง · preprint · 2026-10-23

ตารางที่ 9.3: โซ่เจ็ดลูก: ค่าบนแนวทแยงมุมซ้ายสุด $(A_{m+1})_{11}$ หลัง m รอบ และค่าคลาดเคลื่อนสัมบูรณ์เทียบกับ $\lambda_1 = 3.8478$ (ค่าที่แน่นอนของสองแถวแรกเป็นเศษส่วน ปัดเป็นสี่ตำแหน่ง ตัวเลขจากโค้ดของเรา)

m	$(A_{m+1})_{11}$	ค่าคลาดเคลื่อนสัมบูรณ์
1	2.8000 ($\frac{14}{5}$)	1.0478
4	3.4545 ($\frac{38}{11}$)	0.3932
8	3.6842	0.1636
14	3.8007	0.0470

ขั้นตอนวิธี QR: แยกแล้วคูณกลับลำดับ ทำซ้ำ

ขั้นตอนวิธี QR (QR method) เริ่มจาก $A_1 = A$ แล้วทำซ้ำ

$$A_m = Q_m R_m, \qquad A_{m+1} = R_m Q_m, \qquad m = 1, 2, \ldots \tag{9.1}$$

ในแต่ละรอบเราแยกเมทริกซ์ปัจจุบันเป็นเมทริกซ์เชิงตั้งฉากคูณสามเหลี่ยมบน แล้วคูณกลับลำดับ ผลที่ได้ยังคล้ายกับเมทริกซ์เดิม

ทฤษฎีบท 9.3.4: ผลของขั้นตอนวิธี QR

1. ทุก A_m คล้ายกับ A ด้วยเมทริกซ์เชิงตั้งฉาก จึงมีค่าลักษณะเฉพาะชุดเดียวกับ A ถ้า A สมมาตร A_m ก็สมมาตรด้วย
2. ถ้า $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$ แล้ว A_m ลู่เข้าสู่เมทริกซ์สามเหลี่ยมบนที่มีค่าลักษณะเฉพาะอยู่บนแนวทแยงมุม

พิสูจน์. **แนวคิด.** ส่วน (i) พิสูจน์จากสมการ $A_{m+1} = Q_m^T A_m Q_m$ ส่วน (ii) พิสูจน์แนวคิดเฉพาะค่าบนซ้ายสุด

1. **ส่วน (i).** จาก $R_m = Q_m^T A_m$ จึง $A_{m+1} = R_m Q_m = Q_m^T A_m Q_m$ ถ้า A_m สมมาตร $A_{m+1}^T = Q_m^T A_m^T Q_m = A_{m+1}$ ส่วนที่ว่า A_m เป็นเฮสเซนเบิร์กต่อไปได้ด้วยดู Atkinson 1989, §9.5, p. 625
2. **สูตรของ A^m.** ให้ $\hat{Q}_m = Q_1 \cdots Q_m$ และ $\hat{R}_m = R_m \cdots R_1$ จาก (i) จะได้ $A_{m+1} = \hat{Q}_m^T A \hat{Q}_m$ จึง $A\hat{Q}_m = \hat{Q}_m A_{m+1}$ และ
$$A^{m+1} = A\hat{Q}_m\hat{R}_m = \hat{Q}_m A_{m+1}\hat{R}_m = \hat{Q}_m Q_{m+1} R_{m+1}\hat{R}_m = \hat{Q}_{m+1}\hat{R}_{m+1}.$$
3. **คอลัมน์แรก.** เพราะ $\hat{R}_m$ เป็นสามเหลี่ยมบน $A^m\mathbf{e}_1 = \hat{Q}_m\hat{R}_m\mathbf{e}_1 = r_{11}\hat{\mathbf{q}}_1$ โดย $\hat{\mathbf{q}}_1$ คือคอลัมน์แรกของ $\hat{Q}_m$ ดังนั้น $\hat{\mathbf{q}}_1$ มีทิศเดียวกับ $A^m\mathbf{e}_1$ ซึ่งคือวิธีกำลังในหัวข้อ 9.2 ที่เริ่มจาก $\mathbf{e}_1$
4. **ค่าบนซ้ายสุด.** $(A_{m+1})_{11} = \hat{\mathbf{q}}_1^T A \hat{\mathbf{q}}_1$ เมื่อ A สมมาตร นี่คือผลหารเรย์ลีของ $A^m\mathbf{e}_1$ ถ้า $\mathbf{e}_1$ มีส่วนประกอบบนเวกเตอร์ลักษณะเฉพาะของ λ_1 ไม่เป็นศูนย์ ค่านี้เข้าหา λ_1 ตามแนวคิดของทฤษฎีบท 9.2.1

สรุป: ค่าบนแนวทแยงมุมซ้ายสุดคือผลหารเรย์ลีของวิธีกำลังที่เริ่มจาก $\mathbf{e}_1$ ส่วนการลู่เข้าของค่าที่เหลือเราอ้างจาก Atkinson 1989, §9.5, Thm 9.6, p. 625 □

ตัวอย่างโซ่. ลองดูโซ่เจ็ดลูก C โดยไม่สลับหมายเลข คอลัมน์แรกของ C คือ $C\mathbf{e}_1 = (2, -1, 0, \ldots, 0)^T$ ดังนั้น $(A_2)_{11} = R(C\mathbf{e}_1) = \frac{14}{5}$ เพราะ $(C\mathbf{e}_1)^T C(C\mathbf{e}_1) = 14$ และ $(C\mathbf{e}_1)^T(C\mathbf{e}_1) = 5$ ตารางที่ 9.3 ดูต่อไปตามรอบ

ค่าคลาดเคลื่อนสัมบูรณ์ลดลงทุกรอบ อัตราต่อรอบขึ้นไปถึงราว 0.817 ก่อนจะค่อย ๆ เข้าหา $(\lambda_2/\lambda_1)^2 = 0.7873$ เมื่อ m โตมาก อัตราที่ช้านี้มาจาก $\lambda_2/\lambda_1 = 0.887$ ซึ่งใกล้หนึ่งมาก เราจึงต้องการการเลื่อนค่าในหัวข้อย่อยถัดไป

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 9.4: โซ่เจ็ดลูก: ขนาดของสมาชิกล่างสุดนอกแนวทแยงมุม (ตำแหน่ง $(n, n-1)$) หลัง k รอบ เมื่อไม่เลื่อนค่าและเมื่อเลื่อนแบบวิลกินสัน (ตัวเลขจากโค้ดของเรา)

k	ไม่เลื่อนค่า	เลื่อนแบบวิลกินสัน
0	1	1
1	0.5451	0.4000
2	0.0752	0.1633
3	0.0151	0.005834
4	0.0037	1.39×10^{-7}

เลื่อนค่า: ทำให้ค่าล่างสุดนิ่งเร็วขึ้น

ในโซ่ สมาชิกล่างสุดนอกแนวทแยงมุม ตำแหน่ง $(n, n-1)$ ลดลงช้า อัตราที่ควรเป็นคือ $|\lambda_n/\lambda_{n-1}| = 0.2599$ ต่อรอบ การเลื่อนค่าที่เราเห็นในหัวข้อ 9.2 ช่วยได้ (shift) เลือกค่า σ_m แล้วทำ

$$A_m - \sigma_m I = Q_m R_m, \qquad A_{m+1} = R_m Q_m + \sigma_m I. \tag{9.2}$$

การเลื่อนไม่เปลี่ยนค่าลักษณะเฉพาะ เพราะ $R_m Q_m + \sigma_m I = Q_m^T (Q_m R_m + \sigma_m I) Q_m = Q_m^T A_m Q_m$

นิยาม 9.3.3: การเลื่อนแบบวิลกินสัน

ให้ A_m เป็นเมทริกซ์สมมาตรสามแนวเฉียง และให้บล็อกล่างขวาขนาด 2×2 เป็น $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$ โดย c คือสมาชิกล่างขวาสุด **การเลื่อนแบบวิลกินสัน** (Wilkinson shift) คือค่าลักษณะเฉพาะของบล็อกนี้ที่อยู่ใกล้ c มากที่สุด

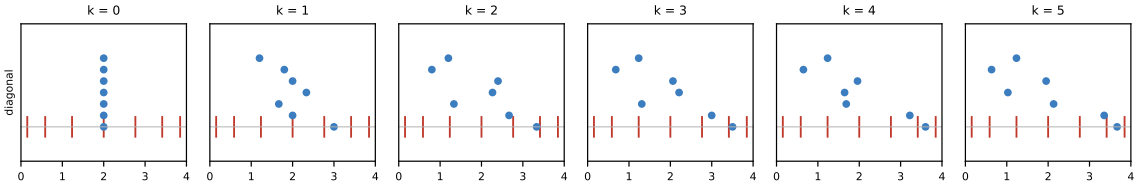
อ่านแบบเข้าใจง่าย. บล็อก 2×2 มีค่าลักษณะเฉพาะสองค่า เราเลือกค่าที่อยู่ใกล้ c เพราะมุมล่างขวาควรเข้าไปหาค่านั้น ถ้าสองค่าห่างจาก c เท่ากัน เราเลือกค่าที่มากกว่า (เป็นข้อตกลงของเรา ทฤษฎีบทไม่ได้กำหนด) ตำราแนะนำค่าเลื่อนแบบนี้ Atkinson 1989, §9.5, pp. 626--627

ในรอบแรกและรอบที่สอง การเลื่อนแบบวิลกินสันยังไม่ดีกว่า (0.1633 มากกว่า 0.0752) แต่ตั้งแต่รอบที่สามเป็นต้นไปมันนิ่งเร็วกว่ามาก

ข้อควรระวัง: ลำดับบนแนวทแยงมุมไม่ใช่ลำดับตามขนาด

เมื่อเลื่อนค่า ค่าที่ลู่เข้าที่มุมล่างขวาก่อนไม่จำเป็นต้องเป็นค่าที่ใหญ่หรือเล็กที่สุด ในเจลลี่ที่สลับหมายเลขของเรา ค่าแรกที่ได้คือ 1.234633 ซึ่งเป็นโหมดที่สามของโซ่ ทั้งที่มีค่าที่เล็กกว่า (0.152241) อยู่ด้วย ลำดับจึงขึ้นกับการเลื่อนและเมทริกซ์เริ่มต้น อย่าสรุปเอาว่าค่าที่ได้ก่อนเป็นค่าที่เล็กที่สุด

ฉบับร่าง · preprint · 2026-10-23



รูปที่ 9.7: ค่าบนแนวทแยงมุมของ T หลัง k รอบ ($k = 0, \ldots, 5$) ด้วยการเลื่อนแบบวิลกินสัน จุดสีฟ้าคือสมาชิกบนแนวทแยงมุม ขีดสีแดงคือค่าลักษณะเฉพาะจริงทั้งเจ็ดตัว ความสูงของจุดไม่มีความหมาย ใช้เพียงแยกสมาชิกออกจากกัน (สร้างจาก `numanal/ch09_qr.py`)

การตัดทิ้ง: เมื่อโหมดหนึ่งแยกออกมาแล้ว

เมื่อสมาชิกล่างสุดนอกแนวทแยงมุม β_{n-1} ของเมทริกซ์สามแนวเฉียงเป็นศูนย์ หรือเล็กมากจนปัดเป็นศูนย์ สมาชิกมุมล่างขวาคือค่าลักษณะเฉพาะหนึ่งค่า เราตัดแถวและคอลัมน์สุดท้ายทิ้ง แล้วทำต่อกับเมทริกซ์ที่เล็กลงหนึ่งขนาด การทำแบบนี้เรียกว่า **การตัดทิ้ง** (deflation) คำนี้ใกล้กับความหมายที่ตำราใช้สำหรับวิธีกำลังในหัวข้อ 9.2 Atkinson 1989, §9.2, p. 609

ทฤษฎีบท 9.3.5: การแยกบล็อก

ให้ T เป็นเมทริกซ์สมมาตรสามแนวเฉียงขนาด $n \times n$ ถ้า $\beta_k = 0$ แล้ว $T = \operatorname{diag}(T_1, T_2)$ โดย T_1 ขนาด $k \times k$ และค่าลักษณะเฉพาะของ T คือค่าลักษณะเฉพาะของ T_1 รวมกับของ T_2 นับตามภาวะรากซ้ำ

พิสูจน์.

1. **บล็อก.** เพราะ $\beta_k = 0$ สมาชิก $(k, k+1)$ และ $(k+1, k)$ เป็นศูนย์ และสมาชิกอื่นที่ข้ามรอยต่อเป็นศูนย์อยู่แล้วเพราะสามแนวเฉียง จึงแยกได้เป็นบล็อกทแยงมุม

2. **ดีเทอร์มินันต์.** $\det(T - \lambda I) = \det(T_1 - \lambda I)\det(T_2 - \lambda I)$ ดังนั้นศูนย์ของพหุนามนี้คือศูนย์ของ T_1 รวมกับของ T_2

สรุป: การแยกบล็อกเป็นเหตุผลของการตัดทิ้ง ตำราอธิบายเรื่องนี้ไว้ใน Atkinson 1989, §9.4, p. 621 □

อ่านแบบเข้าใจง่าย. ถ้าสปริงระหว่างลูกที่ k กับลูกที่ $k+1$ ขาด สมาชิก β_k ของเจลลี่กลายเป็นศูนย์ และสมาชิกแนวทแยงมุมของลูกทั้งสองที่ปลายรอยขาดลดลงหนึ่ง เจลลี่จึงแยกเป็นสองก้อน ค่าลักษณะเฉพาะของเจลลี่ทั้งหมดคือค่าของก้อนหนึ่งรวมกับของอีกก้อนหนึ่ง (แบบฝึกหัด 9.3.5 ให้ลองกับลูกบอลสี่ลูก)

กลับไปที่เจลลี่ที่ติดหมายเลขใหม่ หลังการลดรูปเราได้ T แล้วทำขั้นตอนวิธี QR แบบเลื่อนค่าแบบวิลกินสัน พร้อมการตัดทิ้ง ผลคือค่าลักษณะเฉพาะทั้งเจ็ดตัวตรงกับสูตรปิด $2 - 2\cos\left(\frac{j\pi}{8}\right)$ ของตารางที่ 9.1 ภายใน 3×10^{-15} ใช้ทั้งหมด 15 รอบ ค่าแรกที่ตัดทิ้งได้เกิดหลังห้ารอบ เมื่อสมาชิกล่างสุดเหลือ 1.2×10^{-21} ค่านั้นคือ 1.234633 ซึ่งตรงกับโหมดที่สามของโซ่ (ตัวเลขจากโค้ดของเรา) รูป 9.7 แสดงค่าบนแนวทแยงมุมทีละรอบ

สรุปใจความ

- กลับไปที่เจลลี่: ลูกบอลที่ติดหมายเลขสลับกันยังให้โหมดเดิมทั้งเจ็ด (ค่าลักษณะเฉพาะเท่ากับของโซ่)
- การสะท้อนแบบเฮาส์โฮลเดอร์ทำให้ส่วนใต้ตำแหน่งแรกของเวกเตอร์เป็นศูนย์ได้ในครั้งเดียว ต้องเลือกเครื่องหมายของ α ให้ตรงข้ามกับ b_1 (หัวข้อย่อยแรก)
- การแยก $A = QR$ ใช้การสะท้อนทีละคอลัมน์ และการลดรูปด้วยการสะท้อนทำให้เมทริกซ์สมมาตรเป็นสามแนวเฉียงโดยไม่เปลี่ยนค่าลักษณะเฉพาะ

ร่างต้น · Preprint · 2026-10-23

- ขั้นตอนวิธี QR $A_{m+1} = R_m Q_m$ คล้ายกับ A เสมอ ค่าบนซ้ายสุดของแนวทแยงมุมคือผลหารเรย์ลีของวิธีกำลังที่เริ่มจาก $\mathbf{e}_1$
- การเลื่อนแบบวิลกินสันทำให้มุมล่างขวานิ่งเร็วขึ้น แต่ลำดับบนแนวทแยงมุมไม่ใช่ลำดับตามขนาด
- เมื่อสมาชิกล่างสุดเป็นศูนย์ เราตัดทิ้งหนึ่งแถวหนึ่งคอลัมน์ และสปริงที่ขาดทำให้เจลลี่แยกเป็นสองก้อน

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 9.3.1 (⋆) ให้ $\mathbf{b} = (0, 3, 4)^T$

1. จงหา α ตามกฎของทฤษฎีบทการสะท้อน (ให้ $\operatorname{sgn}(0) = 1$) แล้วหาเวกเตอร์ $\mathbf{w}$
2. จงคำนวณ $U\mathbf{b}$ เมื่อ $U = I - 2\mathbf{w}\mathbf{w}^T$ แล้วตรวจว่าได้ $(-5, 0, 0)^T$

แบบฝึกหัด 9.3.2 (⋆) ให้ $A = \begin{pmatrix} 12 & 3 \\ 5 & 4 \end{pmatrix}$ จงหาเมทริกซ์เชิงตั้งฉาก Q และสามเหลี่ยมบน R ที่ $A = QR$ โดยใช้การสะท้อนเพียงหนึ่งครั้ง แล้วตรวจว่า $QR = A$ และ $|\det A| = |R_{11}R_{22}|$

แบบฝึกหัด 9.3.3 (⋆⋆, พิสูจน์) ให้ $\mathbf{w}$ เป็นเวกเตอร์ที่ $\|\mathbf{w}\|_2 = 1$ และ $U = I - 2\mathbf{w}\mathbf{w}^T$

1. จงแสดงว่า $U^T = U$ และ $U^2 = I$ ดังนั้น U เป็นเมทริกซ์เชิงตั้งฉาก
2. จงแสดงว่า $U\mathbf{w} = -\mathbf{w}$ และ $U\mathbf{v} = \mathbf{v}$ สำหรับทุก $\mathbf{v}$ ที่ $\mathbf{v} \perp \mathbf{w}$
3. จงใช้ข้อ 2 แสดงว่า $\det U = -1$

ตัวช่วยคิด

ในข้อ 3 ลองใช้ฐานของ $\mathbb{R}^n$ ที่ประกอบด้วย $\mathbf{w}$ และเวกเตอร์ที่ตั้งฉากกับ $\mathbf{w}$ แล้วดูค่าลักษณะเฉพาะของ U ในฐานนั้น

แบบฝึกหัด 9.3.4 (⋆⋆, คำนวณและตีความ) ให้ C เป็นเมทริกซ์โซ่เจ็ดลูก $\operatorname{tridiag}(-1, 2, -1)$ ของหัวข้อ 9.1 ทำขั้นตอนวิธี QR สามรอบแบบไม่เลื่อนค่า และสามรอบแบบเลื่อนแบบวิลกินสัน แล้วให้ β_k เป็นขนาดของสมาชิกล่างสุดนอกแนวทแยงมุมหลัง k รอบ

1. จงเปรียบ β_2 และ β_3 ของสองวิธี วิธีใดดีกว่าในแต่ละจังหวะ
2. บล็อกมุมล่างขวาขนาด 2×2 ของ C คือ $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ จงหาค่าลักษณะเฉพาะทั้งสองของบล็อกนี้ แล้วอธิบายว่าทำไมการเลื่อนแบบวิลกินสันจึงเลือกค่าเลื่อน $\sigma = 3$ ได้ตามข้อตกลงของเรา

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 9.3.5 (★★, กลับไปที่เจลลี่) ลูกบอลสี่ลูกเรียงระหว่างกำแพงสองข้าง มวลทุกลูกเท่ากับ 1 สปริงทุกเส้นมีค่าคงที่ 1 แต่สปริงระหว่างลูกที่ 2 กับลูกที่ 3 ขาด

1. จงเขียนเมทริกซ์ 4×4 ของการขยับ โดยปรับสมาชิกแนวทแยงมุมของลูกที่ 2 และ 3 ให้เหมาะกับสปริงที่หายไป
2. จงแสดงว่าเมทริกซ์แยกเป็นสองบล็อก และหาค่าลักษณะเฉพาะของแต่ละบล็อก
3. จงอธิบายว่าทำไมค่าลักษณะเฉพาะซ้ำกันสองค่า โดยดูจากรูปร่างของเจลลี่ที่แยกเป็นสองก้อน

แบบฝึกหัด 9.3.6 (★★★, สิ่งที่ผิดพลาดได้) ให้ $\mathbf{b} = (1, 10^{-8}, 0)^T$

1. ในเลขคณิตที่แม่นยำ จงแสดงว่าทั้ง $\alpha = \|\mathbf{b}\|_2$ และ $\alpha = -\|\mathbf{b}\|_2$ ให้ $U\mathbf{b} = \alpha \mathbf{e}_1$ เมื่อ $U = I - 2\mathbf{w}\mathbf{w}^T$ และ $\mathbf{w} = (\mathbf{b} - \alpha\mathbf{e}_1)/\|\mathbf{b} - \alpha\mathbf{e}_1\|_2$
2. ในเลขทศนิยม 64 บิต $\|\mathbf{b}\|_2$ ปัดเป็น 1 จงคำนวณ $U\mathbf{b}$ สำหรับ $\alpha = +1$ และ $\alpha = -1$ แล้วหาค่าคลาดเคลื่อนสัมบูรณ์ของสมาชิกที่สองในแต่ละกรณี

> **ตัวช่วยคิด**
>
> สำหรับข้อ 2 ลองดูว่า $1 - \alpha$ เป็นเท่าไรในเลขทศนิยมเมื่อ $\alpha = 1$ เทียบกับเมื่อ $\alpha = -1$ และเทียบกับการคำนวณในเลขคณิตที่แม่นยำ

แบบฝึกหัด 9.3.7 (★★★, พิสูจน์) ให้ A เป็นเมทริกซ์สมมาตร ให้ $A_1 = A$ และ $A_m = Q_m R_m$, $A_{m+1} = R_m Q_m$ ตามขั้นตอนวิธี QR แบบไม่เลื่อนค่า ให้ $\hat{Q}_m = Q_1 \cdots Q_m$ และ $\hat{R}_m = R_m \cdots R_1$

1. จงแสดงว่า $A^m = \hat{Q}_m \hat{R}_m$ สำหรับทุก $m \geq 1$
2. จงแสดงว่า $(A_{m+1})_{11} = R(A^m \mathbf{e}_1)$ เมื่อ R คือผลหารเรย์ลี
3. เมื่อ $A = \mathrm{tridiag}(-1, 2, -1)$ ขนาด 7×7 จงหา $(A_2)_{11}$ และตรวจว่าได้ $\frac{14}{5}$

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 9.3.1.

1. เพราะ $b_1 = 0$ จึง $\mathrm{sgn}(b_1) = +1$ และ $\alpha = -\|\mathbf{b}\|_2 = -5$ จากนั้น $\mathbf{u} = \mathbf{b} - \alpha\mathbf{e}_1 = (5, 3, 4)^T$ มี $\|\mathbf{u}\|_2^2 = 50$ จึง $\mathbf{w} = (5, 3, 4)^T/\sqrt{50}$
2. $\mathbf{w}^T\mathbf{b} = \frac{25}{\sqrt{50}}$ ดังนั้น $2(\mathbf{w}^T\mathbf{b})\mathbf{w} = 2 \cdot \frac{25}{\sqrt{50}} \cdot \frac{(5,3,4)^T}{\sqrt{50}} = (5, 3, 4)^T$ จึง $U\mathbf{b} = \mathbf{b} - (5, 3, 4)^T = (-5, 0, 0)^T$

ตรวจด้วยโค้ด: ใน `test_ch09_qr.py`(`test_ex931_b_0_3_4`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 9.3.2. คอลัมน์แรก $(12,5)^T$ มีความยาว 13 และ $12>0$ จึง $\alpha=-13$ และ $\mathbf{u}=(25,5)^T$ มี $\|\mathbf{u}\|_2^2=650$ จึง $\mathbf{w}=(5,1)^T/\sqrt{26}$ แล้ว

$$Q=I-2\mathbf{w}\mathbf{w}^T=I-\frac{1}{13}\begin{pmatrix}25&5\\5&1\end{pmatrix}=\begin{pmatrix}-\frac{12}{13}&-\frac{5}{13}\\-\frac{5}{13}&\frac{12}{13}\end{pmatrix}.$$

ได้ $R=Q^TA=QA$ โดยแถวแรกของ QA คือ $\left(\frac{-144-25}{13},\frac{-36-20}{13}\right)=\left(-13,-\frac{56}{13}\right)$ และแถวที่สองคือ $\left(\frac{-60+60}{13},\frac{-15+48}{13}\right)=\left(0,\frac{33}{13}\right)$ จึง

$$R=\begin{pmatrix}-13&-\frac{56}{13}\\0&\frac{33}{13}\end{pmatrix}.$$

ตรวจ QR: สมาชิกแรกคือ $\frac{-12}{13}\cdot(-13)=12$ และสมาชิก $(1,2)$ คือ $\frac{-12}{13}\cdot\frac{-56}{13}+\frac{-5}{13}\cdot\frac{33}{13}=\frac{672-165}{169}=3$ ทั้งสองตรงกับ A ส่วนแถวที่สองได้ 4 เช่นกัน และ $\det A=48-15=33$ ขณะที่ $|R_{11}R_{22}|=13\cdot\frac{33}{13}=33$

ตรวจด้วยโค้ด: ใน `test_ch09_qr.py`(`test_ex932_qr_12_3_5_4`)

เฉลย 9.3.3.

1. $U^T=(I-2\mathbf{w}\mathbf{w}^T)^T=I-2\mathbf{w}\mathbf{w}^T=U$ และ

   $$U^2=I-4\mathbf{w}\mathbf{w}^T+4\mathbf{w}(\mathbf{w}^T\mathbf{w})\mathbf{w}^T=I-4\mathbf{w}\mathbf{w}^T+4\mathbf{w}\mathbf{w}^T=I.$$

   ดังนั้น $U^TU=U^2=I$ จึงเป็นเมทริกซ์เชิงตั้งฉาก

2. $U\mathbf{w}=\mathbf{w}-2\mathbf{w}(\mathbf{w}^T\mathbf{w})=-\mathbf{w}$ และถ้า $\mathbf{v}\perp\mathbf{w}$ แล้ว $\mathbf{w}^T\mathbf{v}=0$ จึง $U\mathbf{v}=\mathbf{v}-2\mathbf{w}(\mathbf{w}^T\mathbf{v})=\mathbf{v}$

3. เลือกฐานมุมฉากของ $\mathbb{R}^n$ เป็น $\mathbf{w},\mathbf{v}_2,\ldots,\mathbf{v}_n$ โดยที่ $\mathbf{v}_i\perp\mathbf{w}$ ตามข้อ 2 U คงที่บน $\mathbf{v}_i$ ทุกตัวและกลับเครื่องหมายบน $\mathbf{w}$ ดังนั้นในฐานนี้ $U=\operatorname{diag}(-1,1,\ldots,1)$ จึง $\det U=-1$

ตรวจ ด้วย โค้ด: ใน `test_ch09_qr.py`(`test_ex933_reflector_w_flips`) และ `test_reflector_orthogonal_involution`

เฉลย 9.3.4.

1. ตัวเลขจากโค้ดของเรา (ขนาดของสมาชิกล่างสุดนอกแนวทแยงมุม β_k):

k	ไม่เลื่อนค่า	เลื่อนแบบวิลกินสัน
2	0.0752	0.1633
3	0.0151	0.00583

   หลังสองรอบการไม่เลื่อนค่าดีกว่า ($0.0752<0.1633$) แต่หลังสามรอบการเลื่อนแบบวิลกินสันดีกว่า ($0.00583<0.0151$) การเลื่อนไม่ได้ช่วยในรอบแรก ๆ เสมอไป

2. บล็อก $\begin{pmatrix}2&-1\\-1&2\end{pmatrix}$ มีค่าลักษณะเฉพาะจาก $(2-\lambda)^2-1=0$ คือ $\lambda=1$ และ $\lambda=3$ ทั้งสองห่างจาก $c=2$ เป็นระยะ 1 เท่ากัน จึงเป็นกรณีเสมอ ข้อตกลงของเราเลือกค่าที่มากกว่า คือ $\sigma=3$ ทฤษฎีบทไม่ได้กำหนดกรณีเสมอนี้ จึงเป็นข้อตกลง

ตรวจด้วยโค้ด: ใน `test_ch09_qr.py`(`test_ex934_plain_vs_wilkinson`) และ `test_wilkinson_tie_takes_larger`

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 9.3.5.

1. สมาชิกแนวทแยงมุมของลูกที่ 2 และ 3 ลดลงหนึ่งจาก 2 เป็น 1 สมาชิก $(2,3)$ เป็นศูนย์ และสปริงอื่นยังเป็น -1:

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}.$$

2. เมทริกซ์ที่มีสมาชิก $(2,3)$ และ $(3,2)$ เป็นศูนย์แยกเป็นบล็อกทแยงมุมได้เสมอ คือ $A = \operatorname{diag}(T_1, T_2)$ โดย $T_1 = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ และ $T_2 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$ ค่าลักษณะเฉพาะของ T_1 มาจาก $(2-\lambda)(1-\lambda)-1 = \lambda^2-3\lambda+1 = 0$ จึงได้ $\lambda = \frac{3\pm\sqrt{5}}{2} \approx 0.381966$ และ 2.618034 ส่วน T_2 มีพหุนามเดียวกัน $\lambda^2 - 3\lambda + 1$ ดังนั้นค่าลักษณะเฉพาะของ A คือ 0.381966 สองครั้ง และ 2.618034 สองครั้ง

3. T_2 คือ T_1 เมื่อสลับดัชนีลำดับ (ลูกที่ 1 กับ 4 และลูกที่ 2 กับ 3) ดังนั้นสองก้อนเป็นภาพกระจกของกันและกัน เจลลี่ทั้งสองก้อนจึงมีโหมดและความถี่เท่ากัน เพราะ $\omega = \sqrt{\lambda}$ ค่าลักษณะเฉพาะจึงซ้ำกัน

ตรวจด้วยโค้ด: ใน `test_ch09_qr.py`(`test_ex935_cut_spring_blocks`) และ `test_ex935_cut_seven_balls` (สปริงที่ขาดที่ลูก 4 ของโซ่เจ็ดลูก ค่าลักษณะเฉพาะเป็นการรวมของสองบล็อก)

เฉลย 9.3.6.

1. ให้ $\alpha = \pm\|\mathbf{b}\|_2$ จะได้ $\alpha^2 = \|\mathbf{b}\|_2^2$ และ $\mathbf{u} = \mathbf{b} - \alpha\mathbf{e}_1 \neq \mathbf{0}$ เพราะ $u_2 = 10^{-8}$ จาก $\|\mathbf{u}\|_2^2 = \|\mathbf{b}\|_2^2 - 2\alpha b_1 + \alpha^2 = 2(\|\mathbf{b}\|_2^2 - \alpha b_1) = 2\,\mathbf{u}^T\mathbf{b}$ จึง $U\mathbf{b} = \mathbf{b} - \frac{2\mathbf{u}^T\mathbf{b}}{\|\mathbf{u}\|_2^2}\mathbf{u} = \mathbf{b} - \mathbf{u} = \alpha\mathbf{e}_1$ ทั้งสองเครื่องหมาย ขั้นตอนนี้ไม่ต้องการว่า $\alpha b_1 \leq 0$ เพียงว่า $\mathbf{u} \neq \mathbf{0}$ ซึ่งเป็นจริงที่นี่

2. ในเลขทศนิยม $\|\mathbf{b}\|_2 = 1$ จึง $1 - \alpha$ เป็น 0 เมื่อ $\alpha = +1$ และเป็น 2 เมื่อ $\alpha = -1$

   - $\alpha = +1$: $\mathbf{u} = (0, 10^{-8}, 0)^T$ จึง $\mathbf{w} = \mathbf{e}_2$ และ $U = \operatorname{diag}(1,-1,1)$ ได้ $U\mathbf{b} = (1, -10^{-8}, 0)^T$ ค่าคลาดเคลื่อนสัมบูรณ์ของสมาชิกที่สองคือ 10^{-8}
   - $\alpha = -1$: $\mathbf{u} = (2, 10^{-8}, 0)^T$ ไม่มีการหักล้าง ได้ $U\mathbf{b}$ ที่สมาชิกที่สองเป็นศูนย์ในระดับการปัดเศษ (ตรวจด้วยโค้ด)

   เครื่องหมายผิดทำให้ $1 - \alpha$ เสียข้อมูลของ 10^{-16} ที่อยู่ในเลขนั้น

ตรวจด้วยโค้ด: ใน `test_ch09_qr.py`(`test_ex936_wrong_sign_window`) และ `test_sign_caution_error_1e_minus_8`

ฉบับร่าง Preprint 2026-10-23

เฉลย 9.3.7.

1. ใช้การเหนี่ยวนำ เมื่อ $m = 1$ ได้ $A = Q_1 R_1 = \hat{Q}_1 \hat{R}_1$ สมมติว่า $A^m = \hat{Q}_m \hat{R}_m$ และ $A_{m+1} = \hat{Q}_m^T A \hat{Q}_m$ (จากการสังเกตว่า $A_{m+1} = Q_m^T A_m Q_m$ ซ้ำ) จึง $A\hat{Q}_m = \hat{Q}_m A_{m+1}$ และ

$$A^{m+1} = A\hat{Q}_m \hat{R}_m = \hat{Q}_m A_{m+1} \hat{R}_m = \hat{Q}_m Q_{m+1} R_{m+1} \hat{R}_m = \hat{Q}_{m+1} \hat{R}_{m+1}.$$

2. ให้ $\mathbf{q} = \hat{Q}_m \mathbf{e}_1$ เพราะ $\hat{R}_m$ เป็นสามเหลี่ยมบน $A^m \mathbf{e}_1 = \hat{Q}_m \hat{R}_m \mathbf{e}_1 = r_{11}\mathbf{q}$ จึง $\mathbf{q} = \pm A^m \mathbf{e}_1 / \|A^m \mathbf{e}_1\|_2$ จาก $A_{m+1} = \hat{Q}_m^T A \hat{Q}_m$ จึง $(A_{m+1})_{11} = \mathbf{q}^T A \mathbf{q}$ ซึ่งเป็นผลหารเรย์ลีของ $\mathbf{q}$ และผลหารเรย์ลีไม่เปลี่ยนเมื่อคูณเวกเตอร์ด้วยจำนวนไม่เป็นศูนย์ ดังนั้น $(A_{m+1})_{11} = R(A^m \mathbf{e}_1)$

3. จาก $C\mathbf{e}_1 = (2, -1, 0, \ldots, 0)^T$ จึงได้ $\|C\mathbf{e}_1\|_2^2 = 5$ และ

$$C(C\mathbf{e}_1) = (5, -4, 1, 0, \ldots, 0)^T.$$

ดังนั้น $(C\mathbf{e}_1)^T C(C\mathbf{e}_1) = 2 \cdot 5 + (-1)(-4) = 14$ และ $(A_2)_{11} = \frac{14}{5}$ ตรงกับตัวอย่างของหัวข้อนี้

ตรวจ ด้วย โค้ด: ใน `test_ch09_qr.py`(`test_ex937_identity_chain_power`) และ `test_top_left_exact_fractions`

อ่านต่อ. ตำรา Atkinson 1989, §9.3--9.5, pp. 609--628 ให้รายละเอียดการเลือกเครื่องหมาย การลู่เข้า และการเลือกค่าเลื่อนที่ละเอียดกว่านี้ สำหรับค่าลักษณะเฉพาะบางตัวในช่วงที่กำหนด ตำราใช้ลำดับสตูร์มกับการแบ่งครึ่งแทนขั้นตอนวิธี QR Atkinson 1989, §9.4, pp. 619--623 หัวข้อ 9.4 จะถามต่อว่า ถ้าเมทริกซ์ใหญ่จนแม้ขั้นตอนนี้ทำไม่ไหว ควรทำอย่างไร

ร่างปฐม Preprint 2026-10-23

9.4 ใช้เพียงไม่กี่ทิศ

สถานการณ์สมมติ: หน้าเว็บหนึ่งล้านหน้า ทำได้เพียงไม่กี่ทิศ

ในหัวข้อ 9.2 เราจัดอันดับหน้าเว็บหกหน้าด้วยวิธีกำลัง ลองนึกถึงเว็บที่มีหนึ่งล้านหน้า เมทริกซ์การคลิกของมันจึงมีหนึ่งล้านแถวและหนึ่งล้านคอลัมน์ ถ้าเก็บเป็นตารางเต็มต้องมีตัวเลข 10^{12} ตัว หรือประมาณ 8 TB แต่สมมติว่าแต่ละหน้ามีลิงก์ออกประมาณสิบลิงก์ เมทริกซ์จึงเบาบาง เก็บได้ประมาณ 10^7 ตัว และการคูณกับเวกเตอร์หนึ่งครั้งทำได้เร็ว สิ่งที่ทำไม่ไหวคือการแยกตัวประกอบหรือการหาค่าลักษณะเฉพาะทั้งหมด คำถามของหัวข้อนี้คือ ถ้าทำได้เพียงไม่กี่รอบของการคูณ เราจะได้ทิศทางอะไรบ้าง

(ตัวเลขทั้งหมดเป็นการคำนวณของเรา เว็บและการลิงก์เป็นแบบจำลองของเราเอง ไม่ได้บอกว่าเครื่องมือค้นหาใดทำงานอย่างไร)

ภาพรวมก่อนอ่าน

วิธีกำลังในหัวข้อ 9.2 คูณ A ซ้ำกับเวกเตอร์เดียว ทิศที่ได้เอียงเข้าหาทิศเด่นทิศเดียว และรอบก่อนหน้าไม่ถูกเก็บไว้ หัวข้อนี้เก็บทุกเวกเตอร์ที่คูณได้ โดยทำให้ตั้งฉากกันทีละตัว ได้ปริภูมิที่มี k ทิศ แล้วเรามองเมทริกซ์ A ผ่านปริภูมินั้น คือหาค่าลักษณะเฉพาะของเมทริกซ์เล็ก $k \times k$ ค่าเหล่านี้คือ **ค่าริตซ์** ซึ่งเรานิยามในหัวข้อนี้ และค่าริตซ์บนสุดกับล่างสุดจะเข้าหาปลายทั้งสองของสเปกตรัมก่อนค่าอื่น

ปริภูมิไครลอฟ: ทิศจากการคูณซ้ำ

เริ่มจากเวกเตอร์ $\mathbf{r}_0$ แล้วคูณด้วย A ซ้ำ ๆ เก็บผลทุกรอบไว้ ปริภูมิที่สร้างจากผลเหล่านี้มีชื่อเฉพาะ

นิยาม 9.4.1: ปริภูมิย่อยไครลอฟ

ให้ A เป็นเมทริกซ์จัตุรัสขนาด $n \times n$ และ $\mathbf{r}_0 \neq \mathbf{0}$ **ปริภูมิย่อยไครลอฟ** (Krylov subspace) ขนาด k คือ

$$\mathcal{K}_k(A, \mathbf{r}_0) = \operatorname{span}\{\mathbf{r}_0,\ A\mathbf{r}_0,\ \ldots,\ A^{k-1}\mathbf{r}_0\}.$$

อ่านแบบเข้าใจง่าย. ทุกเวกเตอร์ในปริภูมินี้เขียนได้เป็น $p(A)\mathbf{r}_0$ เมื่อ p เป็นพหุนามดีกรีไม่เกิน $k-1$ ดังนั้นการคูณ A ซ้ำ $k-1$ ครั้งให้ k ทิศ เช่น โซ่สองลูก $C_2 = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ กับ $\mathbf{e}_1$: $C_2\mathbf{e}_1 = (2,-1)^T$ จึงได้ $\mathcal{K}_2(C_2, \mathbf{e}_1) = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ เพราะ $(2,-1)^T - 2\mathbf{e}_1 = -\mathbf{e}_2$ ทิศจากการคูณซ้ำจะเอียงเข้าหากันเรื่อย ๆ (ดูหัวข้อ 9.2) จึงต้องทำให้ตั้งฉากกันก่อนใช้

การวนซ้ำอาร์โนลดี: ทำให้ทิศตั้งฉากกันทีละทิศ

ขั้นตอนต่อไปนี้สร้างฐานตั้งฉากของปริภูมิไครลอฟทีละรอบ โดยไม่ต้องคำนวณปริภูมิทั้งหมดก่อน

นิยาม 9.4.2: การวนซ้ำอาร์โนลดี

ให้ $\mathbf{v}_1 = \mathbf{r}_0 / \|\mathbf{r}_0\|_2$ แล้วสำหรับ $j = 1, 2, \ldots$ ทำดังนี้

1. คำนวณ $\mathbf{t} = A\mathbf{v}_j$
2. สำหรับ $i = 1, \ldots, j$ ให้ $h_{ij} = \mathbf{v}_i^T \mathbf{t}$ แล้วแทน $\mathbf{t} \leftarrow \mathbf{t} - h_{ij}\mathbf{v}_i$
3. ให้ $h_{j+1,j} = \|\mathbf{t}\|_2$ ถ้า $h_{j+1,j} = 0$ ให้หยุด มิฉะนั้นให้ $\mathbf{v}_{j+1} = \mathbf{t} / h_{j+1,j}$

เรียกขั้นตอนนี้ว่า **การวนซ้ำอาร์โนลดี** (Arnoldi iteration)

อ่านแบบเข้าใจง่าย. ในแต่ละรอบ เราคูณ A กับทิศล่าสุด แล้วตัดส่วนที่ซ้ำกับทิศเก่าออก สิ่งที่เหลือคือทิศใหม่ที่ตั้งฉากกับทุกทิศก่อนหน้า ในโค้ดของเรา การตัดนี้ทำซ้ำอีกหนึ่งรอบ เพื่อให้ตั้งฉากจริงในเลขทศนิยม

ทฤษฎีบท 9.4.1: ความสัมพันธ์อาร์โนลดี

ให้ A เป็นเมทริกซ์จัตุรัสจริง $\mathbf{r}_0 \neq \mathbf{0}$ และสมมติว่า $h_{j+1,j} > 0$ สำหรับ $j < k$ ให้ $V_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ และ $H_k = (h_{ij})_{i,j=1}^{k}$ จะได้ว่า

1. $\mathbf{v}_1, \ldots, \mathbf{v}_k$ เป็นเวกเตอร์หนึ่งหน่วยที่ตั้งฉากกันเป็นคู่ และ $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} = \mathcal{K}_k(A, \mathbf{r}_0)$
2. H_k เป็นเมทริกซ์เฮสเซนเบิร์กบน และ $H_k = V_k^T A V_k$
3. $AV_k = V_k H_k + h_{k+1,k}\,\mathbf{v}_{k+1}\mathbf{e}_k^T$ เมื่อ $h_{k+1,k} > 0$ ส่วนถ้า $h_{k+1,k} = 0$ จะได้ $AV_k = V_k H_k$ โดยไม่มีพจน์เพิ่ม และไม่ต้องนิยาม $\mathbf{v}_{k+1}$

พิสูจน์. 1. **อุปนัยบน j.** สมมติว่า $\mathbf{v}_1, \ldots, \mathbf{v}_j$ เป็นหนึ่งหน่วยที่ตั้งฉากกันเป็นคู่ และ span เท่ากับ $\mathcal{K}_j$ กรณี $j = 1$ เป็นจริงเพราะ $\mathcal{K}_1 = \operatorname{span}\{\mathbf{r}_0\}$ ในเลขคณิตที่แม่นยำ $h_{ij} = \mathbf{v}_i^T A\mathbf{v}_j$ เมื่อ $i \le j$ เพราะ $\mathbf{v}_i$ ตั้งฉากกันอยู่แล้ว ดังนั้น

ร่างปริญญา Preprint · 2026-10-23

$\mathbf{v}_m^T\mathbf{t} = \mathbf{v}_m^T A\mathbf{v}_j - \sum_{i\le j} h_{ij}\mathbf{v}_m^T\mathbf{v}_i = h_{mj} - h_{mj} = 0$ สำหรับ $m \le j$ จึง $\mathbf{t}$ ตั้งฉากกับ $\mathbf{v}_1, \ldots, \mathbf{v}_j$

2. **ทิศถัดไป.** เพราะ $h_{j+1,j} > 0$ จึง $\mathbf{t} \neq \mathbf{0}$ และ $\mathbf{v}_{j+1} = \mathbf{t}/h_{j+1,j}$ เป็นหนึ่งหน่วยที่ตั้งฉากกับทุกตัวก่อนหน้า นอกจากนี้ $\mathbf{t} = h_{j+1,j}\mathbf{v}_{j+1}$ จึง
$$A\mathbf{v}_j = \sum_{i\le j+1} h_{ij}\mathbf{v}_i.$$

3. **ปริภูมิ.** ทุกเวกเตอร์ของ $\mathcal{K}_j$ คูณ A แล้วอยู่ใน $\mathcal{K}_{j+1}$ ดังนั้น $A\mathbf{v}_j \in \mathcal{K}_{j+1}$ และ $\mathbf{v}_{j+1} \in \mathcal{K}_{j+1}$ จึง $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{j+1}\} \subseteq \mathcal{K}_{j+1}$ เวกเตอร์ $j+1$ ตัวที่ตั้งฉากกันเป็นอิสระเชิงเส้น ส่วน $\mathcal{K}_{j+1}$ ถูกสร้างจาก $j+1$ เวกเตอร์ จึงสองปริภูมินี้เท่ากัน ข้อ (i) จึงเป็นจริง

4. **ข้อ (ii).** จากข้อ 2 สมาชิก h_{ij} เป็นศูนย์เมื่อ $i > j+1$ จึง H_k เป็นเฮสเซนเบิร์ก ส่วน $(V_k^T A V_k)_{ij} = \mathbf{v}_i^T A\mathbf{v}_j$ ซึ่งเท่ากับ h_{ij} เมื่อ $i \le j+1$ และเป็นศูนย์เมื่อ $i > j+1$ เพราะ $A\mathbf{v}_j \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{j+1}\}$

5. **ข้อ (iii).** เขียนข้อ 2 ของแต่ละคอลัมน์ $j = 1, \ldots, k$ ได้ $A\mathbf{v}_j = \sum_{i\le j} h_{ij}\mathbf{v}_i + h_{j+1,j}\mathbf{v}_{j+1}$ ซึ่งเป็นคอลัมน์ที่ j ของสมการในข้อ (iii) ผลรวมเหล่านี้รวมกันเป็นเมทริกซ์ และพจน์ $h_{k+1,k}\mathbf{v}_{k+1}$ ปรากฏเฉพาะในคอลัมน์สุดท้าย ถ้า $h_{k+1,k} = 0$ แล้ว $\mathbf{t} = \mathbf{0}$ จึง $A\mathbf{v}_k = \sum_{i\le k} h_{ik}\mathbf{v}_i$ โดยตรง และไม่มีพจน์นี้

สรุป: ทุกทิศได้จากการคูณ A ซ้ำ และการตั้งฉากทำให้ได้ฐานของปริภูมิไครลอฟ เมทริกซ์ที่เหลือจากการฉาย A ลงบนฐานนี้เป็นเฮสเซนเบิร์ก (ทฤษฎีบทนี้เป็นการเรียบเรียงของเราเอง ตามแนวทางของ Vorst 2003, §3.3, pp. 27--29) □

อ่านแบบเข้าใจง่าย. ทฤษฎีบทนี้บอกว่า ถ้าเราคูณ A กับทิศเดิมแล้วตั้งฉาก เราได้ฐานตั้งฉากของปริภูมิไครลอฟ และการฉาย A ลงบนฐานนั้นได้เมทริกซ์ H_k ที่มีรูปร่างง่ายอยู่แล้ว ส่วนที่เหลือ $h_{k+1,k}\mathbf{v}_{k+1}$ คือส่วนที่หลุดออกนอกปริภูมิ ซึ่งจะกลับมาในบทที่ 11 (GMRES)

ข้อควรระวัง: เมื่อ $h_{j+1,j} = 0$

ถ้า $h_{j+1,j} = 0$ แล้ว $\mathbf{t} = \mathbf{0}$ แสดงว่า $A\mathbf{v}_j$ อยู่ในปริภูมิที่มีอยู่แล้ว ปริภูมิจึงไม่ขยายอีก และการวนซ้ำหยุดโดยไม่มีทิศถัดไป ปริภูมิที่ได้มีสมบัติว่า A คูณเวกเตอร์ใดในปริภูมินั้นแล้วยังอยู่ในปริภูมินั้น เรียกว่า **ปริภูมิยืนยง** (invariant subspace) ของ A ดังนั้นเราพบปริภูมิยืนยงแล้ว และค่าริตซ์ที่ได้จากปริภูมินี้เป็นค่าลักษณะเฉพาะของ A จริง ตัวอย่างเช่น $C_3 = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ กับ $\mathbf{r}_0 = (1, 0, -1)^T$ ได้ $C_3\mathbf{r}_0 = 2\mathbf{r}_0$ ทั้งที่ C_3 มีค่าลักษณะเฉพาะอีกสองค่า การวนซ้ำหยุดหลังหนึ่งรอบและให้ค่าริตซ์เพียง 2 (แบบฝึกหัด 9.4.5) จุดเริ่มต้นจึงควรมีส่วนประกอบในทุกทิศที่ต้องการ

เมทริกซ์สมมาตร: เหลือสามพจน์

ถ้า A สมมาตร เมทริกซ์ H_k จะง่ายกว่ามาก กรณีนี้เรียกว่า **วิธีลานซอส** (Lanczos method) ตามชื่อผู้เสนอ และการตั้งฉากในแต่ละรอบเหลือเพียงกับทิศก่อนหน้าสองทิศ ค่าใช้จ่ายต่อรอบจึงคงที่ คือการคูณ A หนึ่งครั้ง ผลคูณภายในสองครั้ง และการปรับเวกเตอร์สองครั้ง Vorst 2003, §3.3, p. 30

ฉบับร่าง · Preprint · 2026-10-23

ทฤษฎีบท 9.4.2: การเวียนเกิดสามพจน์ของลานซอส

ถ้า A สมมาตรและเงื่อนไขของทฤษฎีบท 9.4.1 เป็นจริง แล้ว H_k เป็นเมทริกซ์สามแนวเฉียง มีแนวทแยงมุม $\alpha_j = h_{jj}$ และข้างแนวทแยงมุม $\beta_j = h_{j+1,j}$ (สำหรับ $j < k$) และ

$$A\mathbf{v}_j = \beta_{j-1}\mathbf{v}_{j-1} + \alpha_j\mathbf{v}_j + \beta_j\mathbf{v}_{j+1}, \qquad \beta_0 = 0.$$

สมการนี้ใช้ได้สำหรับ $j < k$ และสำหรับ $j = k$ เมื่อ $\beta_k = h_{k+1,k} > 0$ ถ้า $\beta_k = 0$ การวนซ้ำหยุดที่รอบ k ไม่มี $\mathbf{v}_{k+1}$ และสมการสำหรับ $j = k$ เหลือเพียง $A\mathbf{v}_k = \beta_{k-1}\mathbf{v}_{k-1} + \alpha_k\mathbf{v}_k$

พิสูจน์. 1. **สมมาตร.** จากทฤษฎีบท 9.4.1 ข้อ (ii) $h_{ij} = \mathbf{v}_i^T A\mathbf{v}_j = \mathbf{v}_j^T A^T\mathbf{v}_i = \mathbf{v}_j^T A\mathbf{v}_i = h_{ji}$ จึง H_k สมมาตร

2. **สามแนวเฉียง.** H_k เป็นเฮสเซนเบิร์ก จึง $h_{ij} = 0$ เมื่อ $i > j+1$ เมื่อสมมาตรแล้ว $h_{ij} = h_{ji} = 0$ ด้วยเมื่อ $j > i+1$ ดังนั้นมีสมาชิกไม่เป็นศูนย์เฉพาะแนวทแยงมุมและสองแนวข้างแนวทแยงมุม

3. **สูตร.** คอลัมน์ที่ j ของ $AV_k = V_kH_k + \cdots$ มีสมาชิกไม่เป็นศูนย์เฉพาะ $h_{j-1,j} = \beta_{j-1}$, $h_{jj} = \alpha_j$ และ $h_{j+1,j} = \beta_j$ จึงได้สูตร

สรุป: การตั้งฉากลดเหลือสองทิศก่อนหน้า ซึ่งเป็นเหตุผลที่ลานซอสทำได้ถูกกว่าอาร์โนลดีทั่วไป □

ลองเอง

ให้ $A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 4 \end{pmatrix}$ และ $\mathbf{r}_0 = \mathbf{e}_1$ จงหา α_1, β_1 และ $\mathbf{v}_2$ แล้วตรวจว่า $A\mathbf{v}_1 = \alpha_1\mathbf{v}_1 + \beta_1\mathbf{v}_2$ (ผลที่ควรได้คือ $\alpha_1 = 2$, $\beta_1 = \sqrt{2}$ และ $\mathbf{v}_2 = (0,1,1)^T/\sqrt{2}$)

อ่านแบบเข้าใจง่าย. ทำต่ออีกสองรอบกับเมทริกซ์ในกล่องลองเอง จะได้ $\alpha_2 = \alpha_3 = \frac{7}{2}$ และ $\beta_2 = \frac{1}{2}$ เมทริกซ์ T_3 ที่ได้มีค่าลักษณะเฉพาะเท่ากับของ A เพราะเราใช้ทิศครบ 3 ทิศ (แบบฝึกหัด 9.4.1 ทำทุกขั้น และตรวจ trace กับ det)

ค่าริตซ์: ค่าลักษณะเฉพาะของเมทริกซ์เล็ก

นิยาม 9.4.3: ค่าริตซ์และเวกเตอร์ริตซ์

ให้ $H_k = V_k^T AV_k$ ค่าลักษณะเฉพาะ θ ของ H_k เรียกว่า **ค่าริตซ์** (Ritz value) ของ A เทียบกับ V_k ถ้า $H_k\mathbf{s} = \theta\mathbf{s}$ แล้ว $\mathbf{x} = V_k\mathbf{s}$ เรียกว่า **เวกเตอร์ริตซ์** (Ritz vector) Vorst 2003, §2.2, p. 19

อ่านแบบเข้าใจง่าย. H_k คือภาพของ A บนปริภูมิที่เราเลือก ค่าริตซ์จึงเป็นค่าประมาณของค่าลักษณะเฉพาะของ A โดยใช้เพียง k ทิศ เมทริกซ์ H_k มีขนาดเล็ก เราจึงหาค่าลักษณะเฉพาะของมันด้วยขั้นตอนวิธี QR ของหัวข้อ 9.3 ได้ทันที

สำเนาร่าง · Preprint · 2025-10-23

ทฤษฎีบท 9.4.3: ค่าริตซ์อยู่ในช่วงของค่าลักษณะเฉพาะ

ให้ A เป็นเมทริกซ์สมมาตรจริงขนาด $n \times n$ มีค่าลักษณะเฉพาะ $\lambda_1 \le \cdots \le \lambda_n$ ให้ $\mathbf{r}_0 \neq \mathbf{0}$ และให้ $\theta_1^{(k)} \le \cdots \le \theta_k^{(k)}$ เป็นค่าริตซ์หลัง k รอบ โดยการวนซ้ำไม่หยุดก่อนรอบที่ k จะได้

1. $\lambda_1 \le \theta_1^{(k)}$ และ $\theta_k^{(k)} \le \lambda_n$
2. ถ้าทำรอบที่ $k+1$ ได้ด้วย แล้ว $\theta_1^{(k+1)} \le \theta_1^{(k)}$ และ $\theta_k^{(k)} \le \theta_{k+1}^{(k+1)}$
3. ถ้าทำได้ n รอบโดยไม่หยุด แล้ว $\theta_j^{(n)} = \lambda_j$ สำหรับทุก j

พิสูจน์. 1. **ผลหารเรย์ลีของ A.** ให้ $\mathbf{u}_1, \ldots, \mathbf{u}_n$ เป็นเวกเตอร์ลักษณะเฉพาะหนึ่งหน่วยที่ตั้งฉากกันเป็นคู่ เขียน $\mathbf{x} = \sum_j c_j \mathbf{u}_j$ จะได้ $\|\mathbf{x}\|_2^2 = \sum_j c_j^2$ และ $\mathbf{x}^T A\mathbf{x} = \sum_j \lambda_j c_j^2$ จึง $\lambda_1 \le R(\mathbf{x}) \le \lambda_n$ ทุก $\mathbf{x} \neq \mathbf{0}$ โดยมีค่าเท่ากับ λ_1 และ λ_n ที่ $\mathbf{u}_1$ และ $\mathbf{u}_n$ (ดูแบบฝึกหัด 9.4.3)

2. **ค่าริตซ์เป็นผลหารเรย์ลี.** H_k สมมาตร จึงแยกเป็น $H_k = \sum_j \theta_j^{(k)} \mathbf{s}_j \mathbf{s}_j^T$ โดย $\mathbf{s}_j$ ตั้งฉากกันเป็นคู่ ดังนั้น $\mathbf{y}^T H_k \mathbf{y} / \mathbf{y}^T \mathbf{y}$ มีค่าสูงสุด $\theta_k^{(k)}$ และต่ำสุด $\theta_1^{(k)}$ ตั้งแต่ $\mathbf{y} = \mathbf{s}_k$ และ $\mathbf{s}_1$ เมื่อใส่ $\mathbf{x} = V_k \mathbf{y}$ จะได้ $\|\mathbf{x}\| = \|\mathbf{y}\|$ และ $\mathbf{x}^T A\mathbf{x} = \mathbf{y}^T H_k \mathbf{y}$ เมื่อ $\mathbf{y}$ วิ่งผ่านทุกเวกเตอร์ $\mathbf{x}$ วิ่งผ่านทุกเวกเตอร์ของ $\mathcal{K}_k$ (ทฤษฎีบท 9.4.1 ข้อ (i)) จึงได้

$$\theta_k^{(k)} = \max_{\mathbf{x} \in \mathcal{K}_k,\, \mathbf{x} \neq \mathbf{0}} R(\mathbf{x}), \qquad \theta_1^{(k)} = \min_{\mathbf{x} \in \mathcal{K}_k,\, \mathbf{x} \neq \mathbf{0}} R(\mathbf{x}).$$

3. **ทฤษฎีบทข้อ 1 และ 2.** ข้อ 1 ตามจากข้อ 1 และ 2 ของการพิสูจน์ เพราะ $\mathcal{K}_k \subseteq \mathbb{R}^n$ ส่วนข้อ 2 ตามจาก $\mathcal{K}_k \subseteq \mathcal{K}_{k+1}$: เซตที่ใช้หาค่าสูงสุดของรอบ k อยู่ในเซตของรอบ $k+1$ จึงสูงสุดไม่เกิน และเซตที่ใช้หาค่าต่ำสุดของรอบ $k+1$ ใหญ่กว่าจึงต่ำสุดไม่เกิน

4. **ข้อ 3.** ถ้าทำได้ n รอบ V_n เป็นเมทริกซ์จัตุรัสที่คอลัมน์ตั้งฉากกัน จึงเป็นเมทริกซ์เชิงตั้งฉาก และ $H_n = V_n^T A V_n = V_n^{-1} A V_n$ คล้ายกับ A ค่าลักษณะเฉพาะของเมทริกซ์ที่คล้ายกันเท่ากัน

สรุป: ค่าริตซ์ทุกค่าอยู่ในช่วง $[\lambda_1, \lambda_n]$ ค่าบนสุดเพิ่มขึ้นและค่าล่างสุดลดลงเมื่อเพิ่มรอบ และเมื่อใช้ n ทิศเต็ม ค่าริตซ์กลับมาเป็นค่าลักษณะเฉพาะจริงทุกตัว □

โซ่จากลูกแรก: ค่าริตซ์คือค่าลักษณะเฉพาะของโซ่ที่สั้นลง

ลองเริ่มจากลูกบอลลูกแรกของโซ่ เมทริกซ์ C_n เป็นสามแนวเฉียง และ $\mathbf{e}_1$ มีส่วนประกอบในทุกโหมด ค่าริตซ์จะคำนวณได้เป็นรูปปิดทั้งหมด

ทฤษฎีบท 9.4.4: ค่าริตซ์ของโซ่จากลูกแรก

ให้ $C_n = \operatorname{tridiag}(-1, 2, -1)$ ของโซ่ n ลูกในหัวข้อ 9.1 และ $\mathbf{r}_0 = \mathbf{e}_1$ ถ้า $1 \le k \le n$ การวนซ้ำอาร์โนลดีไม่หยุดก่อนรอบที่ k และค่าริตซ์หลัง k รอบคือ $2 - 2\cos\frac{j\pi}{k+1}$ สำหรับ $j = 1, \ldots, k$ โดยเฉพาะ

$$\theta_{\max}^{(k)} = 2 + 2\cos\frac{\pi}{k+1}, \qquad \lambda_{\max}(C_n) - \theta_{\max}^{(k)} = 2\cos\frac{\pi}{n+1} - 2\cos\frac{\pi}{k+1}.$$

พิสูจน์. 1. **เวกเตอร์ของอาร์โนลดี.** จาก $C_n \mathbf{e}_j = -\mathbf{e}_{j-1} + 2\mathbf{e}_j - \mathbf{e}_{j+1}$ (โดย $\mathbf{e}_0 = \mathbf{e}_{n+1} = \mathbf{0}$) เราแสดงโดยอุปนัยว่า $\mathbf{v}_j = (-1)^{j-1}\mathbf{e}_j$ และ $\alpha_j = 2$, $\beta_j = 1$ สำหรับ $j < k$ รอบแรกได้ $\mathbf{v}_1 = \mathbf{e}_1$, $\alpha_1 = 2$

ฉบับร่าง Preprint · 2026-05-23

ตารางที่ 9.5: โซ่เจ็ดลูก ($n = 7$) จากลูกแรก: ค่าริตซ์ล่างสุดและบนสุดหลัง k รอบ ค่าของ $k = 7$ คือค่าลักษณะเฉพาะจริงของโซ่ (ตัวเลขจากโค้ดของเรา)

k	$\theta_{\min}^{(k)}$	$\theta_{\max}^{(k)}$
1	2.000000	2.000000
2	1.000000	3.000000
3	0.585786	3.414214
4	0.381966	3.618034
5	0.267949	3.732051
6	0.198062	3.801938
7	0.152241	3.847759

และ $\mathbf{t} = C_n\mathbf{e}_1 - 2\mathbf{e}_1 = -\mathbf{e}_2$ จึง $\beta_1 = 1$ และ $\mathbf{v}_2 = -\mathbf{e}_2$ สมมติว่าถึงรอบ j แล้ว ทฤษฎีบท 9.4.2 ให้ $\mathbf{t} = C_n\mathbf{v}_j - \beta_{j-1}\mathbf{v}_{j-1} - \alpha_j\mathbf{v}_j$ และ $\alpha_j = \mathbf{v}_j^T C_n \mathbf{v}_j = 2$ ใส่ $\mathbf{v}_{j-1} = (-1)^{j-2}\mathbf{e}_{j-1}$ และ $\beta_{j-1} = 1$

$$\mathbf{t} = (-1)^{j-1}\big(-\mathbf{e}_{j-1} + 2\mathbf{e}_j - \mathbf{e}_{j+1}\big) + (-1)^{j-1}\mathbf{e}_{j-1} - 2(-1)^{j-1}\mathbf{e}_j = (-1)^j\mathbf{e}_{j+1}.$$

จึง $\beta_j = 1$ และ $\mathbf{v}_{j+1} = (-1)^j\mathbf{e}_{j+1}$ ซึ่งเป็นรูปแบบเดียวกันสำหรับ $j+1$

2. **ค่าริตซ์.** จากข้อ 1 H_k มีแนวทแยงมุมเป็น 2 และข้างแนวทแยงมุมเป็น 1 นั่นคือ $H_k = DC_kD$ โดย $D = \operatorname{diag}(1, -1, 1, \dots)$ และ C_k คือโซ่ k ลูก ค่าลักษณะเฉพาะของ H_k จึงเท่ากับของ C_k คือ $2 - 2\cos\frac{j\pi}{k+1}$ (สูตรของหัวข้อ 9.1 ซึ่งเราจะพิสูจน์ในหัวข้อ 10.1)

3. **ค่าสุดขั้ว.** สำหรับ $j = k$ ได้ $2 - 2\cos\frac{k\pi}{k+1} = 2 + 2\cos\frac{\pi}{k+1}$ และ $\lambda_{\max}(C_n) = 2 + 2\cos\frac{\pi}{n+1}$ เมื่อลบกันได้ผลต่างตามที่ต้องการ

สรุป: ค่าริตซ์ของโซ่จากลูกแรกคือค่าลักษณะเฉพาะของโซ่ที่สั้นลง ทุกรอบไม่ต้องแยกเมทริกซ์ใหญ่ทั้งก้อน □

อ่านแบบเข้าใจง่าย. ค่าริตซ์ของรอบที่ k คือค่าลักษณะเฉพาะของโซ่ k ลูก ดังนั้นเมื่อ $k = n$ เราได้ค่าลักษณะเฉพาะของโซ่ n ลูกพอดี ตารางต่อไปนี้ให้ค่าริตซ์บนสุดและล่างสุดของโซ่เจ็ดลูก

ค่าบนสุดขยับขึ้นจาก 2 ไป 3.847759 และค่าล่างสุดลดลงจาก 2 ไป 0.152241 ตามทฤษฎีบท ค่าเหล่านี้ไม่กระโดดออกนอกช่วง $[\lambda_1, \lambda_n]$ ของโซ่เจ็ดลูก

โซ่ใหญ่ขึ้นทำให้เห็นความเร็วชัดขึ้น เมื่อ $n = 100$ มีค่าลักษณะเฉพาะบนสุด $\lambda_{\max} = 3.999033$ เราได้

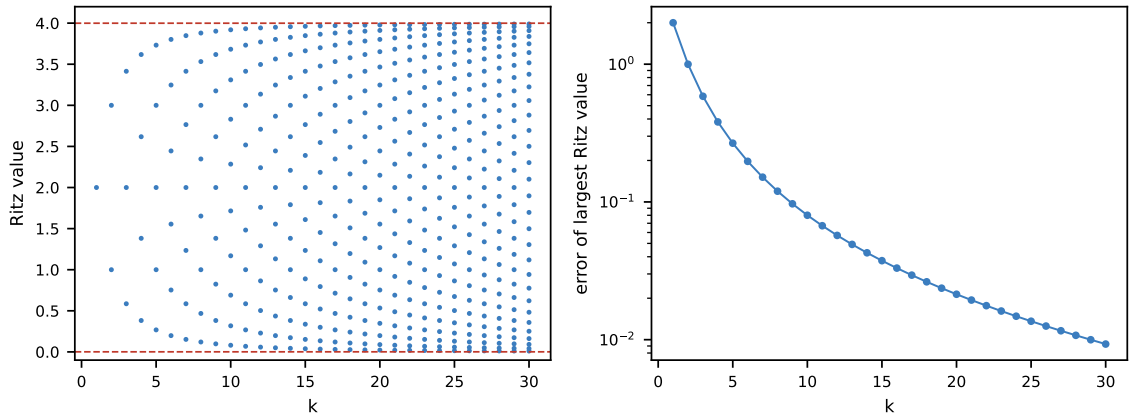
- หลัง 10 รอบ ค่าริตซ์บนสุดคือ 3.918986 ค่าคลาดเคลื่อนสัมบูรณ์ 0.080047
- หลัง 20 รอบ ค่าริตซ์บนสุดคือ 3.977662 ค่าคลาดเคลื่อนสัมบูรณ์ 0.021371
- หลัง 30 รอบ ค่าริตซ์บนสุดคือ 3.989739 ค่าคลาดเคลื่อนสัมบูรณ์ 0.009294

ค่าคลาดเคลื่อนนี้ลดลงตามสูตรในทฤษฎีบท 9.4.4 ซึ่งสำหรับ k ที่น้อยกว่า n มาก ประมาณ $\pi^2/(k+1)^2$ (การประมาณจากสูตร ไม่ใช่ทฤษฎีบท) การลดลงจึงช้า ตามกำลังสองของ k รูป 9.8 แสดงค่าริตซ์ทั้งหมดทีละรอบ ในรูปนี้ค่าที่อยู่ปลายเข้าหาขอบของสเปกตรัมก่อนค่าตรงกลาง

ข้อควรระวัง: ค่าริตซ์ไม่ใช่ค่าลักษณะเฉพาะจนกว่าจะครบทิศ

ค่าริตซ์ที่ปลายสุดเข้าใกล้ค่าลักษณะเฉพาะจริงเมื่อรอบเพิ่มขึ้น แต่ในรอบแรก ๆ ยังไม่เท่ากับค่านั้น ค่าริตซ์ที่อยู่ตรงกลางอาจไม่ใกล้ค่าลักษณะเฉพาะใดเลย และถ้าหยุดก่อนครบ เราไม่รู้ว่ามีค่าอื่นอยู่ตรงกลางหรือไม่ นอกจากนี้ในเลขทศนิยม

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 9.8: ซ้าย: ค่าริตซ์ของโซ่ 100 ลูกจากลูกแรก หลัง $k = 1, \ldots, 30$ รอบ แต่ละคอลัมน์คือค่าริตซ์ k ค่า เส้นประคือ $\lambda_{\min}$ และ $\lambda_{\max}$ ของโซ่ ขวา: ค่าคลาดเคลื่อนสัมบูรณ์ของค่าริตซ์บนสุดในสเกลลอการิทึม (สร้างจาก `numanal/ch09_krylov.py`)

การตั้งฉากอาจเสียไปได้ การวิเคราะห์ของเพจ (Paige) ใน Vorst 2003, §8.3.1, p. 123 อธิบายเรื่องนี้ ซึ่งเราไม่พิสูจน์ในเล่มนี้ โค้ดของเราทำการตั้งฉากซ้ำทั้งหมดเพื่อหลีกเลี่ยงปัญหานี้

อ่านแบบเข้าใจง่าย. ตำรา Vorst 2003, §5.3, p. 51 ระบุว่า ตามทฤษฎีของลานซอส ค่าลักษณะเฉพาะที่อยู่ปลายมักถูกประมาณได้ดีตั้งแต่รอบต้น ๆ และความเร็วขึ้นกับระยะห่างจากค่าอื่น เรายังไม่ได้พิสูจน์ข้อความนี้ในเล่มนี้ สิ่งที่พิสูจน์แล้วคือการอยู่ในช่วงและการเพิ่มขึ้นของค่าริตซ์ปลายสุดในทฤษฎีบท 9.4.3 กับสูตรปิดของโซ่ในทฤษฎีบท 9.4.4

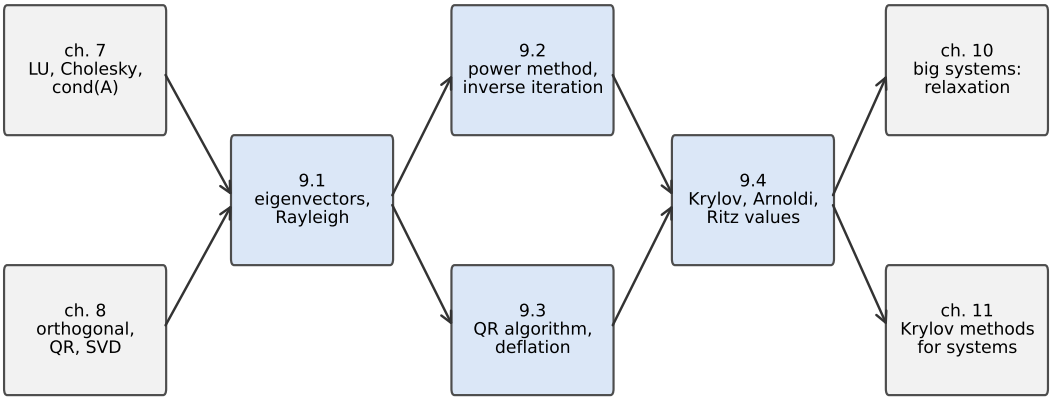
เกร็ดเพิ่มเติม

ถ้า A ไม่สมมาตร การวนซ้ำอาร์โนลดียังให้ H_k เป็นเฮสเซนเบิร์ก แต่ H_k ไม่จำเป็นต้องสมมาตร ค่าลักษณะเฉพาะของ H_k อาจเป็นจำนวนเชิงซ้อน และเราไม่มีทฤษฎีบทแบบข้อ 1 ถึง 3 ของหัวข้อนี้ บทที่ 11 จะใช้แนวทางนี้กับระบบสมการที่ไม่สมมาตร

สรุปใจความ

- ปริภูมิไครลอฟ $\mathcal{K}_k(A, \mathbf{r}_0)$ สร้างจาก $\mathbf{r}_0$ กับการคูณ A ซ้ำ $k-1$ ครั้ง การวนซ้ำอาร์โนลดีให้ฐานตั้งฉากของมัน พร้อม $H_k = V_k^T A V_k$ ที่เป็นเฮสเซนเบิร์ก
- ถ้า A สมมาตร H_k เป็นสามแนวเฉียง (ลานซอส) และค่าใช้จ่ายต่อรอบคงที่
- ค่าริตซ์คือค่าลักษณะเฉพาะของ H_k ทุกค่าอยู่ในช่วง $[\lambda_1, \lambda_n]$ ค่าบนสุดขยับขึ้นและค่าล่างสุดขยับลงเมื่อเพิ่มรอบ และเมื่อครบ n ทิศจะได้ค่าลักษณะเฉพาะจริง
- โซ่จากลูกแรก: ค่าริตซ์หลัง k รอบคือค่าลักษณะเฉพาะของโซ่ k ลูก ค่าบนสุดคือ $2 + 2\cos\frac{\pi}{k+1}$
- การเข้าหาปลายสุดของสเปกตรัมช้าเมื่อ $\lambda_{\max}$ อยู่ชิดกับค่าอื่น หลังจากทิศจำนวนน้อย ค่าริตซ์ยังไม่ใช่ค่าลักษณะเฉพาะ
- จุดเริ่มต้นที่เป็นเวกเตอร์ลักษณะเฉพาะให้ค่าเดียว จึงควรมีส่วนประกอบในทุกทิศที่ต้องการ

ร่าง: Preprint, 2026-10-23



รูปที่ 9.9: แผนผังบทที่ 9 ทั้งบท: หัวข้อ 9.1 ถึง 9.4 และสิ่งที่บทที่ 7 กับ 8 ให้บทนี้ เส้นจากหัวข้อ 9.4 ไปบทที่ 10 และ 11 คือสิ่งที่บทนี้ให้ต่อ (สร้างจาก `numanal/ch09_krylov.py`)

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 9.4.1 (⋆, ทำด้วยมือ) ให้ $A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 4 \end{pmatrix}$ และ $\mathbf{r}_0 = \mathbf{e}_1$

1. จงทำการวนซ้ำลานซอสสามรอบ แล้วหา $\alpha_1, \alpha_2, \alpha_3$ และ β_1, β_2 พร้อมเวกเตอร์ $\mathbf{v}_2, \mathbf{v}_3$
2. จงเขียนเมทริกซ์ T_3 และตรวจว่า $\operatorname{tr} T_3 = \operatorname{tr} A$ และ $\det T_3 = \det A$

แบบฝึกหัด 9.4.2 (⋆⋆, พิสูจน์) ให้ A เป็นเมทริกซ์จัตุรัสจริง และให้ V_k, H_k เป็นผลของการวนซ้ำอาร์โนลดีตามทฤษฎีบท 9.4.1 ข้อ (ii)

1. จงแสดงว่า ถ้า A สมมาตร แล้ว H_k สมมาตร
2. จงสรุปจากข้อ 1 และความเป็นเฮสเซนเบิร์กว่า H_k เป็นเมทริกซ์สามแนวเฉียง
3. จงตรวจว่า ถ้า $A = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$ และ $\mathbf{r}_0 = \mathbf{e}_2$ แล้ว $H_2 = \begin{pmatrix} 3 & 0 \\ 2 & 1 \end{pmatrix}$ ซึ่งไม่สมมาตร

แบบฝึกหัด 9.4.3 (⋆⋆, พิสูจน์) ให้ A เป็นเมทริกซ์สมมาตรจริงขนาด $n \times n$ มีค่าลักษณะเฉพาะ $\lambda_1 \le \cdots \le \lambda_n$ และเวกเตอร์ลักษณะเฉพาะหนึ่งหน่วย $\mathbf{u}_1, \ldots, \mathbf{u}_n$ ที่ตั้งฉากกันเป็นคู่

1. จงแสดงว่า $\mathbf{x}^T A \mathbf{x} \in [\lambda_1 \|\mathbf{x}\|_2^2,\ \lambda_n \|\mathbf{x}\|_2^2]$ สำหรับทุก $\mathbf{x} \in \mathbb{R}^n$
2. จงใช้ข้อ 1 แสดงว่า $\lambda_n = \max_{\mathbf{x} \neq \mathbf{0}} R(\mathbf{x})$ และ $\lambda_1 = \min_{\mathbf{x} \neq \mathbf{0}} R(\mathbf{x})$
3. จงใช้ข้อ 2 แสดงว่า ถ้า V_k มีคอลัมน์ตั้งฉากกันเป็นคู่และหนึ่งหน่วย ค่าริตซ์บนสุดของ k รอบไม่เกิน λ_n

ฉบับร่าง · preprint · 2026-10-23

แบบฝึกหัด 9.4.4 (★★, โซ่ลูกบอล) ให้ $C_7 = \operatorname{tridiag}(-1, 2, -1)$ ขนาด 7×7 ของโซ่เจ็ดลูก และ $\mathbf{r}_0 = \mathbf{e}_1$

1. จงแสดงว่าหลังสี่รอบ $\mathbf{v}_j = (-1)^{j-1}\mathbf{e}_j$ สำหรับ $j = 1, \dots, 4$ และ H_4 เป็นสามแนวเฉียงที่มีสมาชิกแนวทแยงมุมทุกตัวเท่ากับ 2 และสมาชิกข้างแนวทแยงมุมเท่ากับ 1
2. จงหาค่าริตซ์บนสุดและล่างสุดของรอบที่สี่ในรูปแบบปิด
3. จงตรวจว่าค่าริตซ์บนสุดคือ $\frac{5+\sqrt{5}}{2} \approx 3.618034$

แบบฝึกหัด 9.4.5 (★★★, จุดเริ่มที่ผิด) ให้ $C_3 = \operatorname{tridiag}(-1, 2, -1)$ และ $\mathbf{r}_0 = (1, 0, -1)^T$

1. จงแสดงว่า $\mathbf{r}_0$ เป็นเวกเตอร์ลักษณะเฉพาะของ C_3 และหาค่าลักษณะเฉพาะที่สอดคล้องกัน
2. จงแสดงว่าการวนซ้ำอาร์โนลดีหยุดหลังหนึ่งรอบ และค่าริตซ์มีเพียงค่าเดียว
3. เมื่อเริ่มจาก $\mathbf{e}_1$ แทน จงหาค่าริตซ์ทั้งสามหลังสามรอบ และอธิบายว่าทำไมจุดเริ่มต้นควรมีส่วนประกอบในทุกทิศที่ต้องการ

แบบฝึกหัด 9.4.6 (★★★, โครงงานเล็กของบทนี้) ให้ C_{100} เป็นโซ่หนึ่งร้อยลูก และ $\mathbf{r}_0 = \mathbf{e}_1$ ใช้ทฤษฎีบทของโซ่ 9.4.4 เพื่อตอบ

1. จงหาจำนวนรอบ k น้อยที่สุดที่ทำให้ $\lambda_{\max}(C_{100}) - \theta_{\max}^{(k)} \le 0.01$ และตรวจว่ารอบที่ $k-1$ ยังไม่พอ
2. สมมติว่าหน้าเว็บของสถานการณ์สมมติมีหนึ่งล้านหน้า แต่ละหน้ามีลิงก์ออกประมาณสิบลิงก์ เก็บเมทริกซ์แบบตารางเต็มกับแบบเบาบาง (เก็บเฉพาะค่าที่ไม่เป็นศูนย์) จะต้องเก็บตัวเลขต่างกันกี่ตัว และการทำ k รอบของข้อ 1 ด้วยเมทริกซ์เบาบางใช้การคูณประมาณกี่ครั้ง

ตัวช่วยคิด

ในข้อ 2 การคูณเมทริกซ์เบาบางกับเวกเตอร์หนึ่งครั้งใช้หนึ่งการคูณต่อสมาชิกที่ไม่เป็นศูนย์ ส่วนการตั้งฉากเพิ่มค่าใช้จ่ายอีกเล็กน้อย ซึ่งเราไม่นับในข้อนี้

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 9.4.1.

1. **รอบแรก.** $\mathbf{v}_1 = \mathbf{e}_1$ และ $A\mathbf{e}_1 = (2, 1, 1)^T$ จึง $\alpha_1 = 2$ และ $\mathbf{t} = A\mathbf{v}_1 - \alpha_1 \mathbf{v}_1 = (0, 1, 1)^T$ มีความยาว $\sqrt{2}$ ดังนั้น $\beta_1 = \sqrt{2}$ และ $\mathbf{v}_2 = (0, 1, 1)^T/\sqrt{2}$

   รอบที่สอง. $A\mathbf{v}_2 = (2, 3, 4)^T/\sqrt{2}$ และ $\alpha_2 = \mathbf{v}_2^T A\mathbf{v}_2 = \frac{1}{2}(0,1,1)\cdot(2,3,4) = \frac{7}{2}$ จากนั้น

   $$\mathbf{t} = A\mathbf{v}_2 - \beta_1\mathbf{v}_1 - \alpha_2\mathbf{v}_2 = \frac{1}{\sqrt{2}}\big[(2,3,4) - (0,\tfrac{7}{2},\tfrac{7}{2})\big] - (\sqrt{2},0,0) = \Big(0, -\frac{1}{2\sqrt{2}}, \frac{1}{2\sqrt{2}}\Big)^T.$$

   ความยาวของ $\mathbf{t}$ คือ $\frac{1}{2}$ จึง $\beta_2 = \frac{1}{2}$ และ $\mathbf{v}_3 = (0, -1, 1)^T/\sqrt{2}$

   รอบที่สาม. $A\mathbf{v}_3 = (0, -3, 4)^T/\sqrt{2}$ และ $\alpha_3 = \frac{1}{2}(0,-1,1)\cdot(0,-3,4) = \frac{7}{2}$

   ตรวจด้วยโค้ด: `test_ch09_krylov.py` (`test_ex941_lanczos`)

ฉบับร่าง · Preprint · 2026-10-23

2. $T_3 = \begin{pmatrix} 2 & \sqrt{2} & 0 \\ \sqrt{2} & \frac{7}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{7}{2} \end{pmatrix}$ มี $\operatorname{tr} T_3 = 2 + \frac{7}{2} + \frac{7}{2} = 9 = 2 + 3 + 4 = \operatorname{tr} A$ และ

$$\det T_3 = 2\Big(\frac{49}{4} - \frac{1}{4}\Big) - 2 \cdot \frac{7}{2} = 24 - 7 = 17.$$

$\det A = 2(12-0) - 1(4-0) + 1(0-3) = 17$ ตรงกัน เพราะ V_3 เป็นเมทริกซ์เชิงตั้งฉากจัตุรัส จึง $T_3 = V_3^T A V_3$ คล้ายกับ A (ดูทฤษฎีบทค่าริตซ์ ข้อ (iii))

เฉลย 9.4.2.

1. ถ้า $A^T = A$ แล้ว $(V_k^T A V_k)^T = V_k^T A^T V_k = V_k^T A V_k$ จึง H_k สมมาตร

2. จากทฤษฎีบท 9.4.1 ข้อ (ii) H_k เป็นเฮสเซนเบิร์กบน คือ $h_{ij} = 0$ เมื่อ $i > j+1$ เมื่อ H_k สมมาตร ได้ $h_{ij} = h_{ji} = 0$ เมื่อ $j > i + 1$ ด้วย ดังนั้นมีสมาชิกไม่เป็นศูนย์เฉพาะแนวทแยงมุมและสองแนวข้างแนวทแยงมุม จึงเป็นสามแนวเฉียง

3. $A = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$ และ $\mathbf{r}_0 = \mathbf{e}_2$: $\mathbf{v}_1 = \mathbf{e}_2$, $A\mathbf{v}_1 = (2,3)^T$ จึง $h_{11} = 3$ และ $\mathbf{t} = (2,3)^T - 3\mathbf{e}_2 = (2,0)^T$ ดังนั้น $h_{21} = 2$ และ $\mathbf{v}_2 = \mathbf{e}_1$ ต่อไป $A\mathbf{v}_2 = (1,0)^T$ จึง $h_{12} = \mathbf{v}_1^T A\mathbf{v}_2 = 0$ และ $h_{22} = \mathbf{v}_2^T A\mathbf{v}_2 = 1$ ได้ $H_2 = \begin{pmatrix} 3 & 0 \\ 2 & 1 \end{pmatrix}$ ซึ่งไม่สมมาตร เพราะ A ไม่สมมาตร

   ตรวจด้วยโค้ด: `test_ch09_krylov.py` (`test_ex942_band`)

เฉลย 9.4.3.

1. เขียน $\mathbf{x} = \sum_j c_j \mathbf{u}_j$ โดย $\mathbf{u}_j$ ตั้งฉากกันเป็นคู่และหนึ่งหน่วย จะได้ $\|\mathbf{x}\|_2^2 = \sum_j c_j^2$ และ $\mathbf{x}^T A\mathbf{x} = \sum_j \lambda_j c_j^2$ เพราะ $A\mathbf{u}_j = \lambda_j \mathbf{u}_j$ เนื่องจาก $\lambda_1 \le \lambda_j \le \lambda_n$ จึงได้ $\lambda_1 \|\mathbf{x}\|_2^2 \le \mathbf{x}^T A\mathbf{x} \le \lambda_n \|\mathbf{x}\|_2^2$

2. ถ้า $\mathbf{x} = \mathbf{u}_n$ แล้ว $R(\mathbf{u}_n) = \lambda_n$ และข้อ 1 บอกว่า $R(\mathbf{x}) \le \lambda_n$ ทุก $\mathbf{x}$ จึง $\lambda_n = \max R$ ในทำนองเดียวกัน $\lambda_1 = \min R$ โดยใช้ $\mathbf{u}_1$

3. ให้ $H_k = V_k^T A V_k$ ซึ่งสมมาตรตามแบบฝึกหัด 9.4.2 ข้อ 1 และแยกได้เป็น $\sum_j \theta_j s_j s_j^T$ โดย s_j ตั้งฉากกันเป็นคู่ ดังนั้น ค่าริตซ์บนสุด $\theta_k^{(k)} = \max_{\mathbf{y} \neq \mathbf{0}} \mathbf{y}^T H_k \mathbf{y} / \mathbf{y}^T \mathbf{y}$ เมื่อ $\mathbf{x} = V_k \mathbf{y}$ จะได้ $\|\mathbf{x}\| = \|\mathbf{y}\|$ และ $\mathbf{y}^T H_k \mathbf{y} = \mathbf{x}^T A\mathbf{x}$ จึง $\mathbf{y}^T H_k \mathbf{y} / \mathbf{y}^T \mathbf{y} = R(\mathbf{x}) \le \lambda_n$ จากข้อ 2

   ตรวจด้วยโค้ด: `test_ch09_krylov.py` (`test_ex943_rayleigh`) และ (`test_ex943_monotone`)

เฉลย 9.4.4.

1. ใช้การอุปนัยเหมือนการพิสูจน์ทฤษฎีบท 9.4.4 รอบแรก $\mathbf{v}_1 = \mathbf{e}_1$, $\alpha_1 = 2$, $\mathbf{t} = -\mathbf{e}_2$, $\beta_1 = 1$ และ $\mathbf{v}_2 = -\mathbf{e}_2$ รอบที่สอง $C_7 \mathbf{v}_2 = \mathbf{e}_1 - 2\mathbf{e}_2 + \mathbf{e}_3$ จึง $\alpha_2 = 2$ และ $\mathbf{t} = \mathbf{e}_3$ ได้ $\beta_2 = 1$ และ $\mathbf{v}_3 = \mathbf{e}_3$ รอบที่สามและสี่ทำซ้ำแบบเดียวกันจนได้ $\mathbf{v}_4 = -\mathbf{e}_4$ และ $\alpha_j = 2$, $\beta_j = 1$ สำหรับ $j \le 3$ ดังนั้น H_4 มีแนวทแยงมุมเป็น 2 และข้างแนวทแยงมุมเป็น 1

2. $H_4 = D C_4 D$ โดย $D = \operatorname{diag}(1,-1,1,-1)$ จึงมีค่าลักษณะเฉพาะชุดเดียวกับ C_4 คือ $2 - 2\cos\frac{j\pi}{5}$ ค่าบนสุดคือ $2 - 2\cos\frac{4\pi}{5} = 2 + 2\cos\frac{\pi}{5}$ และค่าล่างสุดคือ $2 - 2\cos\frac{\pi}{5}$

3. $\cos\frac{\pi}{5} = \frac{1+\sqrt{5}}{4}$ จึง $\theta_{\max}^{(4)} = 2 + \frac{1+\sqrt{5}}{2} = \frac{5+\sqrt{5}}{2} \approx 3.618034$ และ $\theta_{\min}^{(4)} = \frac{3-\sqrt{5}}{2} \approx 0.381966$

   ตรวจด้วยโค้ด: `test_ch09_krylov.py` (`test_ex944_chain4`)

สัมมนา · preprint · 2026-10-29

เฉลย 9.4.5.

1. คูณแถวต่อแถว: แถวแรก $2 \cdot 1 + (-1) \cdot 0 = 2$ แถวที่สอง $(-1) \cdot 1 + 2 \cdot 0 + (-1)(-1) = 0$ แถวที่สาม $(-1) \cdot 0 + 2 \cdot (-1) = -2$ ได้ $C_3\mathbf{r}_0 = (2, 0, -2)^T = 2\mathbf{r}_0$ ดังนั้น $\mathbf{r}_0$ เป็นเวกเตอร์ลักษณะเฉพาะที่สอดคล้องกับค่า 2

2. $\mathbf{v}_1 = \mathbf{r}_0/\sqrt{2}$ และ $\alpha_1 = 2$ จึง $\mathbf{t} = C_3\mathbf{v}_1 - \alpha_1\mathbf{v}_1 = \mathbf{0}$ ดังนั้น $h_{21} = 0$ และการวนซ้ำหยุดหลังหนึ่งรอบ $H_1 = (2)$ ค่าริตซ์จึงมีเพียง 2

3. เริ่มจาก $\mathbf{e}_1$ ได้ค่าริตซ์หลังสามรอบเท่ากับค่าลักษณะเฉพาะของ C_3 คือ $2 - \sqrt{2}$, 2 และ $2 + \sqrt{2}$ (ทฤษฎีบท 9.4.4 กับ $k = 3$) เหตุผลคือ $\mathbf{r}_0$ ต้องมีส่วนประกอบบนทุกเวกเตอร์ลักษณะเฉพาะที่เราอยากได้ ปริภูมิไครลอฟจึงจะเห็นทุกค่า ข้อ 2 แสดงว่าถ้า $\mathbf{r}_0$ เป็นเวกเตอร์ลักษณะเฉพาะ ปริภูมิจะไม่ขยายอีกและค่าอื่นหายไป ในทางปฏิบัติจุดเริ่มแบบสุ่มมักมีส่วนประกอบในทุกทิศ

   ตรวจด้วยโค้ด: `test_ch09_krylov.py` (`test_ex945_breakdown`)

เฉลย 9.4.6.

1. จากทฤษฎีบท 9.4.4 $\lambda_{\max}(C_{100}) - \theta_{\max}^{(k)} = 2\cos\frac{\pi}{101} - 2\cos\frac{\pi}{k+1}$ ค่านี้ลดลงเมื่อ k เพิ่ม

   - $k = 28$: $2\cos\frac{\pi}{29} \approx 1.988277$ จึงได้ $\lambda_{\max} - \theta_{\max}^{(28)} \approx 0.010756 > 0.01$ ยังไม่พอ
   - $k = 29$: $2\cos\frac{\pi}{30} \approx 1.989044$ จึงได้ $\approx 0.009989 \le 0.01$

   ดังนั้น $k = 29$ รอบ (ค่าคลาดเคลื่อนสัมบูรณ์ประมาณ 0.00999)

2. เมทริกซ์ตารางเต็มมี $(10^6)^2 = 10^{12}$ ตัวเลข ซึ่งคือ 8×10^{12} ไบต์หรือประมาณ 8 TB เมทริกซ์เบาบางเก็บค่าที่ไม่เป็นศูนย์ $10^6 \times 10 = 10^7$ ตัว หรือประมาณ 80 MB (เฉพาะค่า ไม่นับดัชนี) การคูณเมทริกซ์เบาบางกับเวกเตอร์หนึ่งครั้งใช้ประมาณ 10^7 การคูณ จึง 29 รอบใช้ประมาณ $29 \times 10^7 = 2.9 \times 10^8$ การคูณ (ส่วนการตั้งฉากไม่นับในข้อนี้)

   ตรวจด้วยโค้ด: `test_ch09_krylov.py` (`test_ex946_steps29`) และ (`test_ex946_web`)

สรุปบทที่ 9 ในหน้าเดียว

ตารางต่อไปนี้เทียบวิธีหลักของบทนี้ n คือขนาดของเมทริกซ์ k คือจำนวนรอบที่ทำ ค่าใช้จ่ายเป็นลำดับโดยประมาณ

วิธี	ต้องการอะไร	ค่าใช้จ่าย	หมายเหตุ
วิธีกำลัง (9.2)	คูณ A กับเวกเตอร์ได้	ต่อรอบ 1 การคูณ A	ทิศเด่นหนึ่งทิศ อัตรา $\lvert\lambda_2/\lambda_1\rvert$
การวนซ้ำแบบผกผัน (9.2)	แยกตัวประกอบ $A - \sigma I$	ต่อรอบ 1 การแก้ระบบ	ได้ทิศที่ค่าใกล้ σ
การลดรูปด้วยการสะท้อน (9.3)	เมทริกซ์เต็ม	$O(n^3)$	ได้เฮสเซนเบิร์กหรือสามแนวเฉียง
ขั้นตอนวิธี QR แบบเลื่อนค่า (9.3)	สามแนวเฉียง	ต่อรอบ $O(n)$	ทุกค่าลักษณะเฉพาะ (ตัดทิ้ง)
อาร์โนลดี / ลานซอส (9.4)	คูณ A กับเวกเตอร์ได้	k การคูณ A; ตั้งฉาก $O(nk^2)$	ค่าริตซ์ปลายสุดมาก่อน

ฉบับร่าง · Preprint · 2026-10-23

สูตรที่ควรจำ

- ค่าลักษณะเฉพาะของโซ่ n ลูก: $\lambda_j = 2 - 2\cos\frac{j\pi}{n+1}$ (หัวข้อ 9.1 พิสูจน์ในบทที่ 10)
- จานเกอร์ชโกริน: $D_i = \{z : |z - a_{ii}| \le \sum_{j\neq i} |a_{ij}|\}$ ทุกค่าลักษณะเฉพาะอยู่ในอย่างน้อยหนึ่งจาน (หัวข้อ 9.1)
- ค่าตกค้าง: $\min_i |\tilde{\lambda} - \lambda_i| \le \|A\mathbf{x} - \tilde{\lambda}\mathbf{x}\|_2 / \|\mathbf{x}\|_2$ เมื่อ A สมมาตร (หัวข้อ 9.1)
- ผลหารเรย์ลี $R(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}/\mathbf{x}^T\mathbf{x}$ อยู่ในช่วง $[\lambda_1, \lambda_n]$ เมื่อ A สมมาตร (หัวข้อ 9.1 และ 9.4)
- วิธีกำลัง: อัตราการลู่เข้า $|\lambda_2/\lambda_1|$ (หัวข้อ 9.2)
- การวนซ้ำแบบผกผัน: อัตรา $|\lambda_k - \sigma| / \min_{i\neq k} |\lambda_i - \sigma|$ (หัวข้อ 9.2)
- การสะท้อน: $U = I - 2\mathbf{w}\mathbf{w}^T$ และ $\alpha = -\operatorname{sgn}(b_1)\|\mathbf{b}\|_2$ (หัวข้อ 9.3)
- เมทริกซ์เฮสเซนเบิร์ก: $h_{ij} = 0$ เมื่อ $i > j + 1$ (หัวข้อ 9.3)
- ขั้นตอนวิธี QR: $A_{m+1} = R_m Q_m$ (หัวข้อ 9.3)
- ความสัมพันธ์อาร์โนลดี: $AV_k = V_k H_k + h_{k+1,k}\mathbf{v}_{k+1}\mathbf{e}_k^T$ และ $H_k = V_k^T A V_k$ (หัวข้อ 9.4)
- ค่าริตซ์: $\theta_{\max}^{(k)} = \max_{\mathbf{x}\in\mathcal{K}_k} R(\mathbf{x}) \le \lambda_n$ (หัวข้อ 9.4)
- โซ่จากลูกแรก: $\theta_{\max}^{(k)} = 2 + 2\cos\frac{\pi}{k+1}$ (หัวข้อ 9.4)

อ่านต่อ. บทที่ 10 *เมื่อเมทริกซ์ใหญ่เกินกว่าจะแยกตัวประกอบ* เริ่มจากสมการปัวซงของควัน และพิสูจน์สูตร $2 - 2\cos\frac{j\pi}{n+1}$ ที่เราใช้ในบทนี้โดยไม่พิสูจน์ (หัวข้อ 10.1) บทที่ 10 ใช้วิธีวนซ้ำแบบเฉลี่ยกับเพื่อนบ้านแทนการแยกตัวประกอบ บทที่ 11 นำปริภูมิไครลอฟของหัวข้อ 9.4 มาแก้ระบบสมการโดยตรง ด้วยการไล่ค่าซ้ำแบบคอนจูเกตกราเดียนต์ และ GMRES สำหรับระบบที่ไม่สมมาตร

ฉบับร่าง · Preprint · 2026-10-23

ภาค IV

ระบบขนาดใหญ่

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 10

เมื่อเมทริกซ์ใหญ่เกินกว่าจะแยกตัวประกอบ

สถานการณ์: ควันในเกมที่ต้องแก้สมการทุกเฟรม

คุณคงเคยเห็นควันพวยพุ่งจากปล่องไฟ หรือควันที่ปลิวตามตัวละครที่วิ่งผ่านในเกม ควันเหล่านี้ไม่ได้ถูกวาดไว้ล่วงหน้า โปรแกรมต้องคำนวณว่าควันจะไปอยู่ที่ไหนในทุกเฟรม ในปี ค.ศ. 2003 โจส สแตม (Jos Stam) เสนอวิธีคำนวณการไหลของของไหลสำหรับงานภาพและเกม และบอกว่าวิธีของเขา **ไม่ระเบิด** ไม่ว่าค่าพารามิเตอร์จะใหญ่เพียงใด Stam 2003 ในโค้ดตัวอย่างของบทความ ขั้นที่ทำให้ควันกระจายตัว (ขั้นการแพร่) ต้องแก้ระบบที่ค่าใหม่ของแต่ละช่องขึ้นกับค่าเดิมของช่องนั้นและค่าใหม่ของเพื่อนบ้านสี่ช่อง ไม่ใช่การหาค่าเฉลี่ยของเพื่อนบ้านตรง ๆ โค้ดวนซ้ำการคำนวณนี้ 20 รอบ ขั้นที่ปรับความเร็วให้มวลไม่หายไปต้องแก้สมการปัวซง ด้วยโค้ดวนซ้ำแบบเดียวกันอีก 20 รอบ

ตัวเลขต่อไปนี้เป็นของเราเอง ไม่ได้มาจากบทความ ลองสมมติกริดขนาด 128×128 ช่อง ซึ่งมี $16{,}384$ ค่าที่ไม่รู้ในแต่ละครั้งที่แก้ ถ้าเก็บระบบนี้เป็นตารางเต็มขนาด $16{,}384 \times 16{,}384$ ต้องใช้ $16{,}384^2 \times 8 = 2^{31}$ ไบต์ คือราว 2 กิกะไบต์ ทั้งที่แต่ละแถวมีค่าไม่เป็นศูนย์ไม่เกินห้าตัว

บนจอ ควันนี้ต้องเปลี่ยนรูปทุกเฟรม สมมติว่าเกมทำงานที่ 60 เฟรมต่อวินาที ทุกเฟรมจึงต้องแก้ระบบขนาดนี้ซ้ำหลายครั้งภายในเวลาราว 16.7 มิลลิวินาทีต่อเฟรม (คำนวณจาก $1000/60$) สถานการณ์นี้เป็นของสมมติ ไม่ได้บอกว่าเอนจินใดทำงานอย่างไร

บทที่ 9 จบด้วยคำถามว่า เมื่อเมทริกซ์ใหญ่เกินกว่าจะหาค่าลักษณะเฉพาะทั้งหมด การคูณเพียงไม่กี่รอบจะให้ทิศทางอะไรบ้าง บทนี้ตอบจากอีกด้านหนึ่ง คือระบบของควันมีโครงสร้างที่ใช้ได้ แต่ละแถวมีค่าไม่เป็นศูนย์ไม่กี่ตัว และรูปแบบซ้ำกันทุกช่อง ทำให้การแยกตัวประกอบแบบเต็มไม่ใช่ทางเลือกแรก

คำถามของบทนี้คือ ระบบนี้มีโครงสร้างอย่างไร แก้ได้ด้วยวิธีใด และเราจะรู้ได้อย่างไรว่าการวนซ้ำ 20 รอบ *พอ*

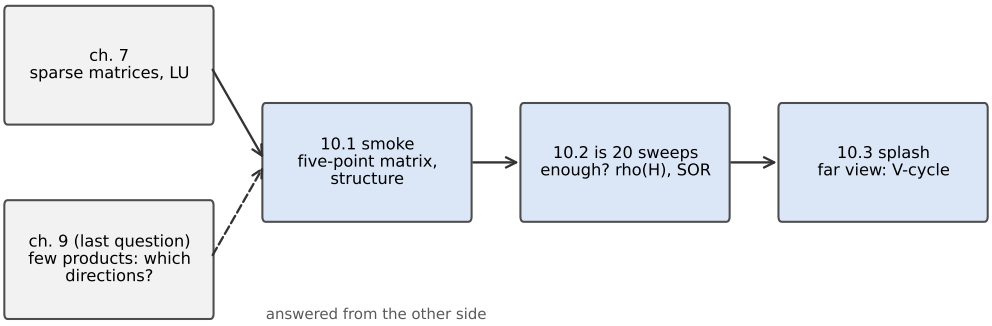
ฉบับร่าง · Preprint · 2026-04-23

ภาพรวมบท

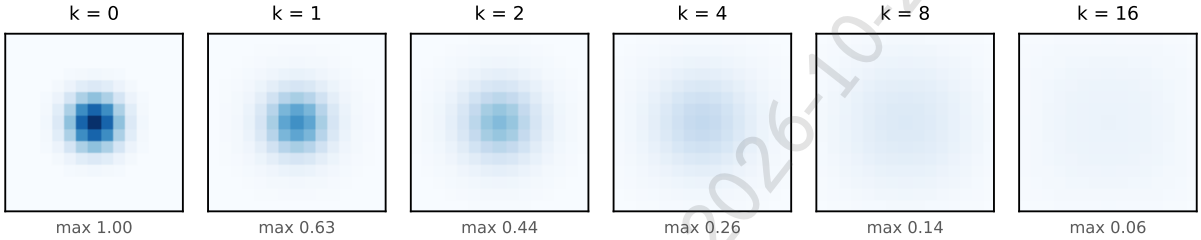
บทนี้มีสามความคิด

- ระบบสมการที่มาจากตารางมักมีแถวละไม่กี่ตัวที่ไม่เป็นศูนย์ จึงเรียกว่า**เมทริกซ์มากเลขศูนย์** (sparse matrix) เก็บและคำนวณได้เล็กกว่าตารางเต็มมาก
- การ**การแยกตัวประกอบ LU** (LU factorization) ให้คำตอบที่แม่น แต่เมื่อระบบใหญ่ หน่วยความจำและการคำนวณอาจเกินกำลังของเครื่อง
- การวนซ้ำไม่ต้องเก็บทั้งตาราง แต่ช้าลงมากเมื่อตารางใหญ่ จึงต้องมีหัวข้อ 10.2 และ 10.3 ต่อไป

เราเริ่มจากควันที่ต้องแก้สมการทุกเฟรม หัวข้อ 10.1 สร้างเมทริกซ์นั้นขึ้นมาและดูโครงสร้างของมัน หัวข้อ 10.2 ถามว่า



รูปที่ 10.1: ลำดับของบทนี้ หัวข้อ 10.1 ถึง 10.3 แต่ละหัวข้อมีคำถามของตัวเอง ลูกศรเส้นทึบคือลำดับการอ่านและสิ่งที่บทที่ 7 ให้ ลูกศรเส้นประคือคำถามท้ายบทที่ 9 ที่บทนี้ตอบจากอีกด้าน (สร้างจาก `numanal/ch10_mg.py`)



รูปที่ 10.2: ควันหนึ่งกลุ่มบนกริด 15×15 หลังการแพร่ทีละขั้นตามระบบ $(I + aA)\mathbf{x}_{\text{ใหม่}} = \mathbf{x}_{\text{เก่า}}$ โดย $a = 1$ เลขที่กำกับคือจำนวนขั้น ตัวเลขใต้ภาพคือค่าสูงสุดของความหนาแน่น (สร้างจาก `numanal/ch10.py`)

20 รอบพอหรือไม่ และหัวข้อ 10.3 ย้ายไปดูอีกด้านของปัญหาเดียวกัน คือน้ำที่กระเซ็นลงสระใหญ่ ซึ่งการวนซ้ำแบบเพื่อนบ้านต้องใช้หลายหมื่นรอบ จึงต้องมองจากกริดหยาบก่อน ภาพด้านล่างแสดงลำดับนี้และสิ่งที่บทที่ 7 กับ 9 ให้บทนี้

10.1 จากควันสู่เมทริกซ์

ภาพรวมก่อนอ่าน

ลองนึกถึงตารางของช่องสี่เหลี่ยม ค่าใหม่ของแต่ละช่องขึ้นกับค่าเดิมของช่องนั้นและค่าของเพื่อนบ้านสี่ช่อง ซึ่งค่าใหม่ของเพื่อนบ้านก็ยังต้องหาเช่นกัน เมื่อเราเขียนความสัมพันธ์นี้ให้ครบทุกช่อง เราจะได้หนึ่งสมการต่อหนึ่งช่อง รวมกันเป็นระบบเชิงเส้นหนึ่งระบบ ระบบนี้คือเมทริกซ์ที่เราจะศึกษา ภาพด้านล่างแสดงควันหนึ่งกลุ่มที่กระจายตัวทีละขั้น โดยทุกขั้นเป็นการแก้ระบบเดียวกัน

ฉบับร่าง · Preprint · 2026-10-23

จากการแพร่ไปสู่ระบบสมการ

ในโค้ดของสแตม ขั้นการแพร่คำนวณค่าของแต่ละช่องใหม่จากค่าเดิมและค่าของเพื่อนบ้านสี่ช่อง ก่อนเขียนสูตรทั่วไป ลองทำกับกริดเล็กที่สุดที่มีความหมายก่อน

ลองเอง: หนึ่งขั้นการแพร่บนกริด 2×2 ด้วยมือ

ให้ $a = 1$ และกริดภายใน 2×2 มีสี่ช่อง เรียงแถวต่อแถวเป็น $x_1 = x_{11}, x_2 = x_{12}, x_3 = x_{21}, x_4 = x_{22}$ ค่าที่ขอบเป็นศูนย์ ดังนั้น x_1 มีเพื่อนบ้านในกริดเพียง x_2 และ x_3 สมการของช่องนี้คือ $(1+4)x_1 - (x_2 + x_3) = x_1^0$ ทำแบบเดียวกันกับอีกสามช่องได้ระบบ

$$\begin{pmatrix} 5 & -1 & -1 & 0 \\ -1 & 5 & 0 & -1 \\ -1 & 0 & 5 & -1 \\ 0 & -1 & -1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

เริ่มจากควันทั้งหมดอยู่ในช่องมุมหนึ่ง คือ $\mathbf{x}^0 = (4, 0, 0, 0)$ ระบบนี้ไม่เปลี่ยนเมื่อสลับ x_2 กับ x_3 และคำตอบมีเพียงคำตอบเดียว ดังนั้น $x_2 = x_3$ ให้ $x_2 = x_3 = y$, $x_1 = p$ และ $x_4 = q$ เหลือสามสมการ

$$5p - 2y = 4, \qquad -p + 5y - q = 0, \qquad -2y + 5q = 0.$$

จากสมการที่สามได้ $q = \frac{2}{5}y$ แทนในสมการที่สองได้ $p = \frac{23}{5}y$ แล้วแทนในสมการแรกได้ $23y - 2y = 4$ จึงได้ $y = \frac{4}{21}$ ดังนั้น

$$\mathbf{x} = \frac{1}{105}\,(92,\ 20,\ 20,\ 8) \approx (0.876,\ 0.190,\ 0.190,\ 0.076).$$

ตรวจได้โดยแทนกลับ เช่นแถวแรก $5 \cdot \frac{92}{105} - 2 \cdot \frac{20}{105} = \frac{420}{105} = 4$
ผลนี้บอกอะไร ค่าสูงสุดลดจาก 4 เหลือ 0.876 และ $\|\mathbf{x}\|_2 = \sqrt{9328}/105 \approx 0.920$ ซึ่งไม่เกิน $\|\mathbf{x}^0\|_2 = 4$ มวลรวมลดจาก 4 เหลือ $\frac{4}{3}$ เพราะขอบดูดมวลไปส่วนหนึ่ง นี่คือเหตุผลที่ก้อนควันจางลงในภาพ

ทำแบบเดียวกันกับทุกช่องของกริดใด ๆ จะได้สมการทั่วไปของช่อง (i, j) ดังนี้

$$(1 + 4a)\, x_{i,j} - a\big(x_{i-1,j} + x_{i+1,j} + x_{i,j-1} + x_{i,j+1}\big) = x^0_{i,j},$$

โดย $a = \Delta t \cdot \mathrm{diff} \cdot N^2$ ตามตัวแปรในโค้ดของบทความ และ x^0 คือค่าก่อนการแพร่ เรากำหนดให้ค่าที่ขอบของกริดเป็นศูนย์เพื่อให้เห็นโครงสร้างชัดขึ้น

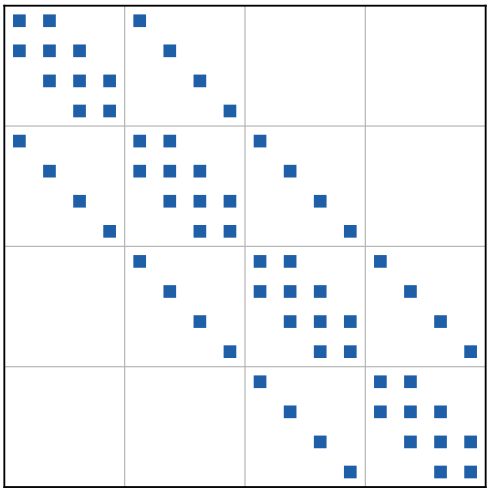
ถ้าเรารวมค่าทุกช่องเป็นเวกเตอร์ $\mathbf{x}$ และให้ A เป็นเมทริกซ์ที่มี 4 บนแนวทแยงและ -1 ที่ช่องเพื่อนบ้าน สมการข้างบนจะเขียนสั้นได้เป็น

$$(I + aA)\,\mathbf{x} = \mathbf{x}^0.$$

ผลนี้ไม่ได้ขึ้นกับกรณีนี้เท่านั้น ค่าลักษณะเฉพาะของ A บนกริด 2×2 คือ $2, 4, 4, 6$ (จากทฤษฎีบท 10.1.2) ดังนั้น $(I + aA)^{-1}$ มีค่าลักษณะเฉพาะ $\frac{1}{1+2a}, \frac{1}{1+4a}, \frac{1}{1+6a}$ และ $\|(I + aA)^{-1}\|_2 = \frac{1}{1+2a} \le 1$ สำหรับทุก $a \ge 0$ ค่าที่ใหญ่ที่สุดนี้ได้จากทิศ $(1, 1, 1, 1)$ ซึ่งหดเป็น $\frac{1}{3}$ เท่าเมื่อ $a = 1$ ดังนั้นไม่ว่าเลือก a เท่าใด ขั้นการแพร่หนึ่งขั้นจะไม่ขยายคำตอบ

ขั้นการปรับความเร็วของบทความต้องแก้**สมการปัวซง** (Poisson equation) ซึ่งมีรูปทั่วไปว่า $-\nabla^2 u = f$ บนพื้นที่ที่มีขอบ ระบบ $(I + aA)\mathbf{x} = \mathbf{x}^0$ ของการแพร่อยู่ในตระกูลเดียวกัน เพราะมีโครงสร้างของเมทริกซ์เหมือนกัน

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 10.3: รูปแบบของ A_4 ซึ่งเป็นกริด 4×4 หรือ 16 ตัวไม่รู้ จุดแต่ละจุดคือสมาชิกที่ไม่เป็นศูนย์ เส้นสีเทาแบ่งเป็นบล็อกขนาด 4×4 บล็อกบนแนวทแยงเป็นเมทริกซ์สามแนวเฉียง และบล็อกนอกแนวทแยงเป็นเมทริกซ์ $-I$ (สร้างจาก `numanal/ch10.py`)

สูตรห้าจุด

สมการปัวซงบนกริดจะเปลี่ยนเป็นระบบเชิงเส้นได้โดยแทนอนุพันธ์อันดับสองด้วย**ผลต่างอันตะ** (finite difference) ซึ่งเป็นผลต่างของค่าที่จุดใกล้กัน ผลที่ได้เรียกว่า**สูตรห้าจุด** (five-point formula) เพราะเชื่อมค่าที่ช่องหนึ่งกับเพื่อนบ้านสี่ช่องรวมเป็นห้าค่า Iserles 2009, §8.2, p. 149

นิยาม 10.1.1: สูตรห้าจุดและเมทริกซ์ของกริด

ให้กริดภายใน $m \times m$ มีระยะห่าง $h = 1/(m+1)$ นับช่องแถวต่อแถวเป็นตำแหน่ง $r = 1, \ldots, n$ โดย $n = m^2$ **เมทริกซ์ของสูตรห้าจุด** (five-point matrix) A_m คือเมทริกซ์ $n \times n$ ที่มี $a_{rr} = 4$ ทุกแถว และ $a_{rs} = -1$ เมื่อช่อง s เป็นเพื่อนบ้านในกริดของช่อง r อื่นเป็นศูนย์ และ $-\nabla^2 u = f$ มีสูตรห้าจุด $A_m \mathbf{u} = h^2 \mathbf{f}$ เมื่อค่าที่ขอบเป็นศูนย์

อ่านแบบเข้าใจง่าย

ทุกแถวของ A_m มีตัวเลข 4 หนึ่งตัวและ -1 ที่เพื่อนบ้านแต่ละช่อง ช่องกลางมีเพื่อนบ้านในกริดครบสี่ช่องจึงมีห้าตัวที่ไม่เป็นศูนย์ ช่องที่อยู่ริมจะมีสี่ตัว และช่องมุมมีสามตัว เพราะค่าที่ขอบถูกนับเป็นศูนย์ เครื่องหมายลบบนนิยามของเรามาจากการเขียน $-\nabla^2 u = f$ ซึ่งทำให้ทุกค่าลักษณะเฉพาะเป็นบวก (ISE เขียน $\nabla^2 u = f$ จึงได้เครื่องหมายตรงข้าม)

ลองเอง: นับตัวเลขที่ไม่เป็นศูนย์

กริด 3×3 มี 9 ช่อง ช่องกลาง $(2,2)$ มีเพื่อนบ้านครบสี่ช่อง จึงมี 5 ตัว ช่องมุม 4 ช่องมีเพื่อนบ้านสองช่อง จึงมีช่องละ 3 ตัว ช่องริมอีก 4 ช่องมีช่องละ 4 ตัว รวมเป็น $4 \cdot 3 + 4 \cdot 4 + 1 \cdot 5 = 33$ ตัว ถ้าใช้สูตรทั่วไป $5m^2 - 4m$ ได้ $5 \cdot 9 - 12 = 33$ เหมือนกัน

ตำราเรียน · Preprint · 2026-10-23

ข้อควรระวัง: ห้าตัวต่อแถว ไม่ได้เท่ากับห้าคูณจำนวนแถว

ถ้าคิดว่าเมทริกซ์มีไม่เกิน $5n$ ตัวที่ไม่เป็นศูนย์ เราจะได้ $5 \cdot 16{,}384 = 81{,}920$ สำหรับกริด 128×128 แต่จำนวนจริงคือ $81{,}408$ ตัว เพราะแถวที่อยู่ริมมีเพื่อนบ้านน้อยกว่า ส่วนต่าง $512 = 4m$ คือจำนวนเพื่อนบ้านที่หายไปที่ขอบ ค่าที่ถูกต้องคือ $5m^2 - 4m$

เมทริกซ์หนึ่งมิติและค่าลักษณะเฉพาะ

ถ้ากริดมีแกนเดียว (เช่นโซ่ของลูกบอลที่ต่อกันด้วยสปริง) เมทริกซ์จะเป็น

$$T_m = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{m \times m},$$

ซึ่งเป็น**เมทริกซ์สามแนวเฉียง** (tridiagonal matrix) เมทริกซ์นี้คือโซ่ลูกบอลระหว่างกำแพงในหัวข้อ 9.1 และเราใช้สูตร $2 - 2\cos\frac{j\pi}{n+1}$ ที่นั่นโดยไม่พิสูจน์ ตอนนี้เรามีที่มาของสูตรนั้นแล้ว

ทฤษฎีบท 10.1.1: ค่าลักษณะเฉพาะของ T_m

ให้ $\theta_j = \dfrac{j\pi}{m+1}$ สำหรับ $j = 1, \ldots, m$ และให้ $\mathbf{v}_j = (\sin i\theta_j)_{i=1}^{m}$ แล้ว

1. ค่าลักษณะเฉพาะของ T_m คือ $\lambda_j = 2 - 2\cos\theta_j$ สำหรับ $j = 1, \ldots, m$ และค่าทั้ง m นี้ต่างกันคู่ละค่า
2. $\mathbf{v}_j$ เป็นเวกเตอร์ลักษณะเฉพาะที่ไม่เป็นศูนย์ของ λ_j

พิสูจน์. **แนวคิด.** ขยายเวกเตอร์ด้วยศูนย์ที่ขอบ แล้วใช้เอกลักษณ์ $\sin(i-1)\theta + \sin(i+1)\theta = 2\cos\theta\,\sin i\theta$ ซึ่งทำให้ทุกแถวของ $T_m\mathbf{v}_j$ เป็นการคูณด้วยค่าคงที่

1. **ขยายด้วยศูนย์.** กำหนด $v_0 = v_{m+1} = 0$ ค่านี้ตรงกับ $\sin 0 = 0$ และ $\sin(m+1)\theta_j = \sin j\pi = 0$ ดังนั้นสมการของทุกแถว $i = 1, \ldots, m$ ใช้รูปเดียวกันได้โดยไม่ต้องแยกกรณีขอบ
2. **คำนวณแถว i.** แถว i ของ $T_m\mathbf{v}_j$ คือ $-\sin(i-1)\theta_j + 2\sin i\theta_j - \sin(i+1)\theta_j$ จากเอกลักษณ์ $\sin(i-1)\theta + \sin(i+1)\theta = 2\cos\theta\,\sin i\theta$ ได้ว่าเท่ากับ $(2 - 2\cos\theta_j)\sin i\theta_j$ ดังนั้น $T_m\mathbf{v}_j = \lambda_j\mathbf{v}_j$ ซึ่งพิสูจน์ส่วน (ii) และการที่ λ_j เป็นค่าลักษณะเฉพาะ
3. **ไม่เป็นศูนย์.** เพราะ $0 < \theta_j < \pi$ จึงได้ $\sin\theta_j > 0$ ดังนั้นสมาชิกแรก $\sin\theta_j$ ของ $\mathbf{v}_j$ ไม่เป็นศูนย์
4. **ต่างกันคู่ละค่า.** บน $(0, \pi)$ ฟังก์ชัน $\cos$ ลดลงอย่างเคร่งครัด ดังนั้น $\theta_1 < \cdots < \theta_m$ ให้ $\lambda_1 < \cdots < \lambda_m$ และ m ค่านี้เป็นค่าลักษณะเฉพาะทั้งหมด เพราะเมทริกซ์ขนาด $m \times m$ มีค่าลักษณะเฉพาะได้ไม่เกิน m ค่า นับตามภาวะรากซ้ำ Iserles 2009, ภาคผนวก A.1.5, pp. 437--438

สรุป: ไซน์ของมุมที่เว้นเท่า ๆ กันคือเวกเตอร์ลักษณะเฉพาะของ T_m และ $\cos$ ที่ลดลงทำให้ค่าลักษณะเฉพาะต่างกันทั้งหมด □

ฉบับร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

ค่า $\lambda_1 = 2 - 2\cos\frac{\pi}{m+1}$ คือค่าลักษณะเฉพาะที่เล็กที่สุด เมื่อ m ใหญ่ค่านี้ใกล้กับ $\pi^2/(m+1)^2$ (เราตรวจด้วยโค้ดแล้ว) นั่นแปลว่า กริดที่ละเอียดขึ้นจะมีค่าลักษณะเฉพาะที่เล็กที่สุดเล็กลงเรื่อย ๆ ซึ่งเป็นเหตุผลที่การวนซ้ำแบบง่ายช้าลงเมื่อกริดใหญ่ เราจะกลับมาที่จุดนี้ในหัวข้อ 10.2

ค่าลักษณะเฉพาะของสูตรห้าจุดในสองมิติ

เมทริกซ์ A_m ของกริดสองมิติเขียนได้เป็น**ผลคูณโครเนกเกอร์** (Kronecker product) $A_m = T_m \otimes I + I \otimes T_m$ เมื่อ I คือเมทริกซ์เอกลักษณ์ $m \times m$ เมทริกซ์นี้จึงเป็นเมทริกซ์สามแนวเฉียงแบบบล็อก ดังรูป 10.3 ผลที่ได้คือค่าลักษณะเฉพาะของ A_m สร้างจากค่าลักษณะเฉพาะของ T_m สองชุด

ทฤษฎีบท 10.1.2: ค่าลักษณะเฉพาะของสูตรห้าจุด

ให้ λ_p เป็นค่าลักษณะเฉพาะของ T_m ตามทฤษฎีบท 10.1.1 แล้ว

1. A_m มีค่าลักษณะเฉพาะ $\mu_{p,q} = \lambda_p + \lambda_q$ สำหรับ $p, q = 1, \ldots, m$ โดยมีเวกเตอร์ลักษณะเฉพาะ $V_{p,q}$ ที่ช่อง (i,j) มีค่า $\sin\frac{ip\pi}{m+1}\,\sin\frac{jq\pi}{m+1}$
2. A_m สมมาตรและบวกแน่นอน กล่าวคือ $\mu_{p,q} > 0$ ทุกค่า
3. **คอนดิชันนัมเบอร์** (condition number)ของ A_m คือ $\kappa(A_m) = \dfrac{1+\cos\frac{\pi}{m+1}}{1-\cos\frac{\pi}{m+1}}$

พิสูจน์. **แนวคิด.** ค่า ของ A_m บน แต่ละ ช่อง เป็น ผล บวก ของ ตัว ดำเนิน การ หนึ่ง มิติ สอง ตัว และ แต่ละ ตัว ทำงาน เหมือน ทฤษฎีบท 10.1.1

1. **ขยายด้วยศูนย์.** ขยาย $V_{p,q}$ ด้วยศูนย์ที่ขอบทั้งสี่ด้าน เช่นเดียวกับข้อ 1 ของทฤษฎีบท 10.1.1
2. **คำนวณช่อง** (i,j). $(A_m V_{p,q})_{ij} = \big(2V_{ij} - V_{i-1,j} - V_{i+1,j}\big) + \big(2V_{ij} - V_{i,j-1} - V_{i,j+1}\big)$ วงเล็บแรกเป็นการทำงานของ T_m ในแกนแรก ซึ่งให้ $\lambda_p V_{ij}$ ตามทฤษฎีบท 10.1.1 และวงเล็บที่สองให้ $\lambda_q V_{ij}$ ดังนั้น $A_m V_{p,q} = (\lambda_p + \lambda_q) V_{p,q}$ ซึ่งพิสูจน์ส่วน (i)
3. **บวกแน่นอน.** A_m สมมาตรเพราะสมาชิก $a_{rs} = a_{sr}$ และ $\lambda_p > 0$ จึงได้ $\mu_{p,q} > 0$ ซึ่งพิสูจน์ส่วน (ii)
4. **ครบทั้งหมด.** ผลคูณภายในของ $V_{p,q}$ กับ $V_{p',q'}$ เท่ากับผลคูณของผลรวมไซน์หนึ่งมิติสองค่า (แบบฝึกหัด 10.1.5) ดังนั้นเวกเตอร์ m^2 ตัวตั้งฉากคู่ละคู่และไม่เป็นศูนย์ จึงเป็นอิสระเชิงเส้น และครบเป็นค่าลักษณะเฉพาะ m^2 ตัวของ A_m
5. **คอนดิชันนัมเบอร์.** เพราะ A_m สมมาตร และ บวก แน่นอน ค่า เอก ฐาน เท่ากับ ค่า ลักษณะ เฉพาะ จึง ได้ $\kappa(A_m) = \mu_{\max}/\mu_{\min} = 2\lambda_m/2\lambda_1$ ตามข้อ 1 และจากการที่ $\cos(\pi - \theta) = -\cos\theta$ ได้ $\lambda_m = 2 - 2\cos\frac{m\pi}{m+1} = 2 + 2\cos\frac{\pi}{m+1}$ ซึ่งให้ส่วน (iii)

สรุป: ผลบวกของค่าลักษณะเฉพาะหนึ่งมิติสองตัวเป็นค่าลักษณะเฉพาะของสูตรห้าจุด และคอนดิชันนัมเบอร์ตามมาจากค่าที่เล็กและใหญ่ที่สุด □

คอนดิชันนัมเบอร์วัดว่าคำตอบเปลี่ยนมากเพียงใดเมื่อข้อมูลเปลี่ยนเล็กน้อย ทฤษฎีบทนี้บอกว่าคอนดิชันนัมเบอร์โตตามขนาดกริด เพราะ $\kappa(A_m) \approx 4(m+1)^2/\pi^2$ เมื่อ m ใหญ่ ซึ่งเราตรวจในแบบฝึกหัด 10.1.4

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 10.1: ค่าใช้จ่ายของกริด 128×128 (ตัวเลขคำนวณตามเกณฑ์ของเรา ไม่ใช่เวลาที่เครื่องใช้จริง)

วิธี	ที่เก็บ (ไบต์)	การดำเนินการ (ลำดับ)
แยกตัวประกอบ LU แบบหนาแน่น	$2^{31} \approx 2.1 \times 10^9$	$\frac{2}{3}n^3 \approx 2.9 \times 10^{12}$
แยกตัวประกอบ LU แบบแถบ	$16{,}908{,}288 \approx 1.7 \times 10^7$	$nb^2 = 2^{28} \approx 2.7 \times 10^8$
หนึ่งรอบเกาส์–ไซเดล	$651{,}264 \approx 6.5 \times 10^5$	$81{,}408 \approx 8.1 \times 10^4$

ค่าใช้จ่ายของสามวิธี

เราจะเปรียบเทียบสามวิธีบนกริด 128×128 ซึ่งมี $n = 16{,}384$ โดยนับจำนวนค่าที่ต้องเก็บและจำนวนการคูณ-บวกแบบคร่าว ๆ ตามเกณฑ์ของเรา

- **การแยกตัวประกอบ LU แบบหนาแน่น**: เก็บ n^2 ค่า และใช้การดำเนินการประมาณ $\frac{2}{3}n^3$ Vorst 2003, §1.1, p. 5
- **การแยกตัวประกอบ LU แบบแถบ**: เมื่อเรียงช่องแถวต่อแถว แถบมีความกว้าง $b = m$ ใช้การดำเนินการประมาณ nb^2 และเก็บเฉพาะแถบล่างได้ประมาณ $n(b+1)$ ค่า Vorst 2003, §1.2, p. 9
- **หนึ่งรอบของเกาส์–ไซเดล**: ใช้หนึ่งการคูณ-บวกต่อค่าที่ไม่เป็นศูนย์หนึ่งค่า และเก็บเฉพาะค่าที่ไม่เป็นศูนย์

ตารางนี้บอกลำดับของค่าใช้จ่าย ไม่ได้บอกเวลาจริง เมื่อนับแบบนี้ หนึ่งรอบเกาส์–ไซเดลถูกกว่าการแยกแบบหนาแน่นราว 3.6×10^7 เท่า และ 20 รอบของสแตมยังถูกกว่าราว 1.8×10^6 เท่า แต่ตารางไม่ได้บอกว่าคำตอบหลังการวนซ้ำแม่นแค่ไหน คำถามนั้นเป็นของหัวข้อ 10.2 ส่วนการแยกตัวประกอบแบบแถบก็ยังมีความกว้างแถบที่โตตามขนาดกริด จึงไม่ได้ทำให้ปัญหาหายไป

ข้อควรระวัง: เมทริกซ์เบาบางอาจไม่เบาบางหลังการแยก

ระหว่างการแยกตัวประกอบ ค่าศูนย์บางตัวจะกลายเป็นค่าที่ไม่เป็นศูนย์ เรียกว่า **ฟิลอิน** (fill-in) Vorst 2003, §1.2, p. 8 ดังนั้นการเก็บเมทริกซ์แบบเบาบางเพียงอย่างเดียวไม่พอ เราต้องคาดการณ์ด้วยว่าตัวประกอบจะเติมเต็มเพิ่มขึ้นเท่าใด

สรุปใจความ

- ขั้นการแพร่ของสแตมเป็นระบบ $(I + aA)\mathbf{x} = \mathbf{x}^0$ และ A คือเมทริกซ์ของสูตรห้าจุด
- เมทริกซ์ของกริด $m \times m$ มีตัวที่ไม่เป็นศูนย์ $5m^2 - 4m$ ตัว ไม่ใช่ $5m^2$ และแต่ละแถวมีไม่เกินห้าตัว
- ค่าลักษณะเฉพาะของ T_m คือ $2 - 2\cos\frac{j\pi}{m+1}$ และของ A_m คือผลบวกของค่าสองตัว ทำให้ A_m บวกแน่นอนและคอนดิชันนัมเบอร์โตตามขนาดกริด
- กลับไปที่ควัน: กริด 128×128 ถ้าแยกแบบหนาแน่นต้องใช้ราว 2 กิกะไบต์ ส่วนหนึ่งรอบเกาส์–ไซเดลใช้ที่เก็บราว 0.65 เมกะไบต์ คำถามที่เหลือคือ 20 รอบพอหรือไม่ ซึ่งเราจะตอบในหัวข้อ 10.2

อ่านต่อ. หนังสือ Iserles 2009, §8.2, pp. 147--150 อธิบายสูตรห้าจุดและการหาคำตอบของกริดแบบสี่เหลี่ยมละเอียดกว่านี้ Grossmann, Roos, and Stynes 2007, §2.4.2, pp. 62--63 วางสมการปัวซงในรูปทั่วไปที่มีค่าสัมประสิทธิ์เปลี่ยนตามตำแหน่ง และ Iserles 2009, §12.4, p. 282 ให้ค่าลักษณะเฉพาะของเมทริกซ์การทำซ้ำจาโคบีบนกริดเดียวกัน ซึ่งเราจะใช้ในหัวข้อ 10.2 ส่วน Vorst 2003, §1.1--1.3, pp. 3--11 อธิบายว่าทำไมระบบจากกริดใหญ่จึงต้องใช้วิธีวนซ้ำแทนการแยกตัวประกอบ

ฉบับร่าง · preprint 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 10.1.1 (★, ทำด้วยมือ) ให้กริดภายใน 3×3 ตามสูตรห้าจุด และให้ A_3 เป็นเมทริกซ์ของสูตรนี้

1. จงเขียนสมการของสูตรห้าจุดที่ช่อง $(1,1)$ และช่อง $(2,2)$ โดยให้มีเฉพาะค่าที่ไม่รู้ และค่าที่ขอบเป็นศูนย์
2. จงนับจำนวนตัวที่ไม่เป็นศูนย์ของ A_3 แล้วตรวจกับสูตร $5m^2 - 4m$
3. จงนับจำนวนตัวที่ไม่เป็นศูนย์ของ A_4 แล้วตรวจกับสูตรเดียวกัน

แบบฝึกหัด 10.1.2 (★, ทำด้วยมือ) ให้ $T_3 = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$

1. จงหา $\operatorname{tr} T_3$ และ $\det T_3$ โดยคำนวณตรง ๆ ไม่ใช้ค่าลักษณะเฉพาะ
2. จงใช้ทฤษฎีบท 10.1.1 หาค่าลักษณะเฉพาะทั้งสาม แล้วตรวจว่าผลบวกเท่ากับ $\operatorname{tr} T_3$ และผลคูณเท่ากับ $\det T_3$

แบบฝึกหัด 10.1.3 (★★, คำนวณ) สำหรับกริด $m \times m$ ที่ $m = 32, 64, 128, 256$ ให้ $n = m^2$

1. จงหาจำนวนไบต์ที่ใช้เก็บตารางเต็ม $n \times n$ ของค่าขนาด 8 ไบต์ แต่ละกรณี
2. จงหาจำนวนตัวที่ไม่เป็นศูนย์ $5m^2 - 4m$ และจำนวนไบต์ของค่าเหล่านั้น (ไม่นับดัชนี) แต่ละกรณี
3. เมื่อ m เพิ่มเป็นสองเท่า อัตราส่วนของไบต์ตารางเต็มเป็นเท่าใด และอัตราส่วนของตัวที่ไม่เป็นศูนย์ประมาณเท่าใด จงอธิบายว่าทำไมสองอัตราส่วนนี้จึงต่างกัน

แบบฝึกหัด 10.1.4 (★★, คำนวณและตีความ) ให้ $c_m = \cos\frac{\pi}{m+1}$ และ $\kappa_m = \dfrac{1 + c_m}{1 - c_m}$ คือคอนดิชันนัมเบอร์ของ A_m ตามทฤษฎีบท 10.1.2

1. จงแสดงว่า $\kappa_3 = 3 + 2\sqrt{2}$
2. จงแสดงว่า $\kappa_m \cdot \dfrac{\pi^2}{4(m+1)^2} \to 1$ เมื่อ $m \to \infty$ โดยใช้ $1 - \cos x \approx x^2/2$ เมื่อ x เล็ก
3. จงคำนวณ κ_{63} และ κ_{127} แล้วหาอัตราส่วน κ_{127}/κ_{63} จงอธิบายว่าทำไมค่านี้ใกล้ 4

แบบฝึกหัด 10.1.5 (★★, พิสูจน์) ให้ $\theta = \pi/(m+1)$ และ $p, q \in \{1, \dots, m\}$ จงแสดงว่า

$$\sum_{i=1}^{m} \sin(ip\theta)\sin(iq\theta) = \begin{cases} \dfrac{m+1}{2}, & p = q, \\[2mm] 0, & p \neq q. \end{cases}$$

ตัวช่วยคิด

ใช้ผลบวกเรขาคณิต $\sum_{j=0}^{m} e^{\mathrm{i}j\varphi}$ โดยให้ $\varphi = k\pi/(m+1)$ เมื่อ $e^{\mathrm{i}(m+1)\varphi} = (-1)^k$ แล้วนำส่วนจริงมาใช้ (ใช้ i สำหรับหน่วยจินตภาพ และ j สำหรับดัชนี)

1. จงใช้ $\sin a \sin b = \frac{1}{2}[\cos(a-b) - \cos(a+b)]$ เขียนผลรวมเป็นผลต่างของผลรวมของ $\cos(ik\theta)$ สองชุด เมื่อ $k = p - q$ และ $k = p + q$

ฉบับร่าง · Preprint · 2026-10-23

2. จงแสดงว่า ถ้า k เป็นจำนวนเต็มที่ $0 < |k| < 2(m+1)$ และ $k \neq 0$ แล้ว $\sum_{i=1}^{m} \cos(ik\theta) = -1$ เมื่อ k เป็นเลขคู่ และเท่ากับ 0 เมื่อ k เป็นเลขคี่

3. จบด้วยการพิจารณากรณี $p = q$ และ $p \neq q$ โดยแยกตามความเท่ากันของภาวะคู่คี่ (ภาวะคู่คี่ของ $p - q$ และ $p + q$ เท่ากันเสมอ)

แบบฝึกหัด 10.1.6 (⋆⋆, พิสูจน์และทดลอง) ให้ $A = A_m$ เป็นเมทริกซ์ของสูตรห้าจุด และให้ $a \geq 0$ ค่าคงที่ $M = I + aA$ คือเมทริกซ์ของขั้นการแพร่หนึ่งขั้นของสแตม

1. จงแสดงว่าค่าลักษณะเฉพาะของ M คือ $1 + a\mu$ เมื่อ μ เป็นค่าลักษณะเฉพาะของ A และ $1 + a\mu \geq 1$ (ใช้ว่า A สมมาตรจึงมีเวกเตอร์ลักษณะเฉพาะตั้งฉากคู่ละคู่ Iserles 2009, ภาคผนวก A.1.5)

2. จงแสดงว่า $\|M^{-1}\|_2 \leq 1$ สำหรับทุก $a \geq 0$ แล้วสรุปว่า $\|\mathbf{x}_{\text{ใหม่}}\|_2 \leq \|\mathbf{x}_{\text{เก่า}}\|_2$ สำหรับทุกขั้นการแพร่ ไม่ว่า a จะใหญ่เพียงใด

3. ทดลองกับ $m = 4$ และ $a \in \{0.1, 1, 50\}$ โดยเริ่มจากเวกเตอร์สุ่มที่ fixed seed แล้วทำห้าขั้น ตรวจว่านอร์มไม่โตขึ้น แล้วอธิบายว่าผลนี้ตรงกับคำว่า *ไม่ระเบิด* ของบทความอย่างไร

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 10.1.1.

1. **ช่อง** $(1,1)$ **(มุม).** มีเพื่อนบ้านในกริดเพียง $(2,1)$ และ $(1,2)$ ค่าที่ขอบเป็นศูนย์จึงหายไป สมการคือ $4u_{11} - u_{21} - u_{12} = h^2 f_{11}$ **ช่อง** $(2,2)$ **(กลาง).** มีเพื่อนบ้านครบสี่ช่อง สมการคือ $4u_{22} - u_{12} - u_{32} - u_{21} - u_{23} = h^2 f_{22}$

2. แถวของมุมมี 3 ตัวที่ไม่เป็นศูนย์ มี 4 แถว ได้ 12 แถวริมมี 4 ตัว มี 4 แถว ได้ 16 แถวกลางมี 5 ตัว มี 1 แถว ได้ 5 รวม $12 + 16 + 5 = 33$ ตรงกับ $5 \cdot 9 - 4 \cdot 3 = 33$

3. กริด 4×4 มีมุม 4 ช่อง ริม 8 ช่อง และกลาง 4 ช่อง จึงได้ $4 \cdot 3 + 8 \cdot 4 + 4 \cdot 5 = 64$ ตรงกับ $5 \cdot 16 - 4 \cdot 4 = 64$

ตรวจด้วยโค้ด: `test_ch10.py` (`test_ex1011_nonzeros_3x3`)

เฉลย 10.1.2.

1. $\operatorname{tr} T_3 = 2 + 2 + 2 = 6$ สำหรับดีเทอร์มิแนนต์ ขยายตามแถวแรก

$$\det T_3 = 2\big(2 \cdot 2 - (-1)(-1)\big) - (-1)\big((-1) \cdot 2 - (-1) \cdot 0\big) = 2 \cdot 3 - 2 = 4.$$

2. ทฤษฎีบท 10.1.1 ให้ $\lambda_j = 2 - 2\cos\frac{j\pi}{4}$ ดังนั้น $\lambda_1 = 2 - \sqrt{2}$, $\lambda_2 = 2$, $\lambda_3 = 2 + \sqrt{2}$ ผลบวกคือ $6 = \operatorname{tr} T_3$ และผลคูณคือ $(2 - \sqrt{2}) \cdot 2 \cdot (2 + \sqrt{2}) = 2\,(4 - 2) = 4 = \det T_3$

ตรวจด้วยโค้ด: `test_ch10.py` (`test_ex1012_t3_trace_det`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 10.1.3.

1. ตารางเต็มมีจำนวนไบต์ $n^2 \cdot 8$ ดังนี้
   $m = 32$: $n = 1{,}024$ ได้ $8{,}388{,}608$ ไบต์ (8 MiB) $m = 64$: $n = 4{,}096$ ได้ $134{,}217{,}728$ ไบต์ (128 MiB)
   $m = 128$: $n = 16{,}384$ ได้ $2^{31} = 2{,}147{,}483{,}648$ ไบต์ (2 GiB) $m = 256$: $n = 65{,}536$ ได้ 2^{35} ไบต์ (32 GiB)

2. จำนวนตัวที่ไม่เป็นศูนย์และไบต์ของค่า (แปดไบต์ต่อตัว) คือ $m = 32$: $4{,}992$ และ $39{,}936$; $m = 64$: $20{,}224$ และ $161{,}792$; $m = 128$: $81{,}408$ และ $651{,}264$; $m = 256$: $326{,}656$ และ $2{,}613{,}248$

3. เมื่อ m เป็นสองเท่า ไบต์ของตารางเต็มเพิ่มเป็น 16 เท่า เพราะ $n^2 = m^4$ ส่วนตัวที่ไม่เป็นศูนย์เพิ่มประมาณ 4 เท่า (อัตราส่วนจริง 4.05, 4.03, 4.01) เพราะจำนวนนี้เป็น $5m^2 - 4m$ ซึ่งเติบโตตาม m^2 เท่านั้น ความต่างนี้คือเหตุผลที่วิธีแบบเบาบางรอดได้เมื่อกริดใหญ่

ตรวจด้วยโค้ด: `test_ch10.py` (`test_ex1013_storage_table`)

เฉลย 10.1.4.

1. $c_3 = \cos\frac{\pi}{4} = \frac{\sqrt{2}}{2}$ ดังนั้น $\kappa_3 = \dfrac{1+\frac{\sqrt{2}}{2}}{1-\frac{\sqrt{2}}{2}} = \dfrac{2+\sqrt{2}}{2-\sqrt{2}} = \dfrac{(2+\sqrt{2})^2}{2} = \dfrac{6+4\sqrt{2}}{2} = 3+2\sqrt{2}$

2. เมื่อ $x = \pi/(m+1)$ มีค่าเล็ก $1 - c_m = 1 - \cos x \approx x^2/2 = \dfrac{\pi^2}{2(m+1)^2}$ และ $1 + c_m \to 2$ ดังนั้น $\kappa_m \approx \dfrac{2}{\pi^2/(2(m+1)^2)} = \dfrac{4(m+1)^2}{\pi^2}$ นั่นคือ $\kappa_m \cdot \dfrac{\pi^2}{4(m+1)^2} \to 1$

3. จากการคำนวณ $\kappa_{63} \approx 1659.38$ และ $\kappa_{127} \approx 6639.52$ ดังนั้น $\kappa_{127}/\kappa_{63} \approx 4.0012$ เหตุผลคือเมื่อ $m+1$ เพิ่มเป็นสองเท่า $(m+1)^2$ เพิ่มเป็นสี่เท่า ค่าที่เกินจาก 4 เพียง 0.03% มาจากพจน์ถัดไปของการประมาณ $1 - \cos x$

ตรวจด้วยโค้ด: `test_ch10.py` (`test_ex1014_condition_127`)

เฉลย 10.1.5.

1. ใช้ $\sin a \sin b = \frac{1}{2}[\cos(a-b) - \cos(a+b)]$ กับ $a = ip\theta$ และ $b = iq\theta$ ได้

$$\sum_{i=1}^{m} \sin(ip\theta)\sin(iq\theta) = \tfrac{1}{2}\Big[\sum_{i=1}^{m}\cos(ik\theta)\Big|_{k=p-q} - \sum_{i=1}^{m}\cos(ik\theta)\Big|_{k=p+q}\Big].$$

2. ให้ $\varphi = k\theta$ โดยที่ $0 < |k| < 2(m+1)$ ดังนั้น $e^{i\varphi} \neq 1$ และผลบวกเรขาคณิต $\sum_{j=0}^{m} e^{ij\varphi} = \dfrac{e^{i(m+1)\varphi} - 1}{e^{i\varphi} - 1}$ โดย $e^{i(m+1)\varphi} = e^{ik\pi} = (-1)^k$ เมื่อ k เป็นเลขคู่ผลบวกเป็นศูนย์ เมื่อ k เป็นเลขคี่ได้ $\dfrac{-2}{e^{i\varphi}-1}$ และ $\operatorname{Re}\dfrac{1}{e^{i\varphi}-1} = -\frac{1}{2}$ (จาก $\frac{1}{e^{i\varphi}-1} = -\frac{1}{2} - \frac{i}{2}\cot\frac{\varphi}{2}$) จึงได้ $\sum_{j=0}^{m}\cos j\varphi = 1$ เมื่อ k คี่ และ 0 เมื่อ k คู่ ตัดพจน์ $j = 0$ (เท่ากับ 1) ออกจะได้ผลสรุปในโจทย์ (ค่า cos ไม่เปลี่ยนเมื่อกลับเครื่องหมายของมุม จึงใช้ได้กับ $k < 0$)

3. กรณี $p = q$: $k = 0$ ได้ $\sum_{i=1}^{m}\cos 0 = m$ และ $k = 2p$ เป็นเลขคู่ได้ -1 ดังนั้นผลรวมคือ $\frac{1}{2}[m - (-1)] = \frac{m+1}{2}$
   กรณี $p \neq q$: $k_1 = p - q$ และ $k_2 = p + q$ ไม่เป็นศูนย์ อยู่ใน $(-2(m+1), 2(m+1))$ และมีภาวะคู่คี่เดียวกัน จึงได้ค่าเท่ากัน ผลต่างจึงเป็น 0

ตรวจด้วยโค้ด: `test_ch10.py` (`test_ex1015_sine_m7`) และ `test_sine_orthogonal`

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 10.1.6.

1. เพราะ A สมมาตร จึงมีฐานเชิงตั้งฉากปรกติของเวกเตอร์ลักษณะเฉพาะ $\mathbf{u}_r$ ที่ $A\mathbf{u}_r = \mu_r \mathbf{u}_r$ (ดู Iserles 2009, ภาคผนวก A.1.5) ดังนั้น $M\mathbf{u}_r = (1 + a\mu_r)\mathbf{u}_r$ และจาก $\mu_r > 0$ (ทฤษฎีบท 10.1.2 ข้อ (ii)) กับ $a \geq 0$ ได้ $1 + a\mu_r \geq 1 > 0$

2. ค่าลักษณะเฉพาะของ M^{-1} คือ $1/(1+a\mu_r) \in (0, 1]$ เมื่อ M^{-1} สมมาตร $\|M^{-1}\|_2$ เท่ากับค่าสัมบูรณ์สูงสุดของค่าลักษณะเฉพาะ จึงได้ $\|M^{-1}\|_2 = \dfrac{1}{1 + a\mu_{\min}} \leq 1$ แล้ว $\|\mathbf{x}_{\text{ใหม่}}\|_2 = \|M^{-1}\mathbf{x}_{\text{เก่า}}\|_2 \leq \|\mathbf{x}_{\text{เก่า}}\|_2/(1+a\mu_{\min}) \leq \|\mathbf{x}_{\text{เก่า}}\|_2$ โดยไม่ขึ้นกับขนาดของ a

3. ที่ $m = 4$ ได้ $\mu_{\min} = 2(2 - 2\cos\frac{\pi}{5}) \approx 0.764$ ดังนั้นที่ $a = 50$ ค่าคงที่ของการลดคือ $1/(1 + 50 \cdot 0.764) \approx 0.026$ ซึ่งเล็กกว่า 1 มาก ทั้งสามค่าของ a ให้นอร์มที่ไม่โตขึ้นทุกขั้น ซึ่งตรงกับคำ *ไม่ระเบิด* ของบทความ เพราะขั้นการแพร่ไม่ขยายคำตอบ จะเลือก a ใหญ่แค่ไหนก็ไม่ทำให้คำตอบพุ่งขึ้น

ตรวจด้วยโค้ด: `test_ch10.py` (`test_ex1016_step`) และ `test_diffusion_norm_bound`

10.2 เฉลี่ยกับเพื่อนบ้านซ้ำ ๆ

สถานการณ์: โค้ดวนซ้ำ 20 รอบ ทำไมพอ

ในบทความของสแตม ขั้นการแพร่และขั้นการปรับความเร็วใช้ตัวแก้แบบเดียวกัน คือการทำซ้ำแบบเกาส์–ไซเดล ซึ่งผู้เขียนบอกว่าเป็นตัวแก้ที่ง่ายที่สุดและใช้ได้ดีในทางปฏิบัติ Stam 2003 โค้ดวนซ้ำแค่ 20 รอบ และบทความไม่ได้อธิบายว่าทำไมเลือกเลข 20

ลองนึกภาพควันที่เราเห็นกระจายสวยงามหลัง 20 รอบ แล้วลองดูอีกด้านหนึ่ง ถ้าค่าที่ทำให้ควันกระจายเร็วขึ้นมาก ภาพนั้นยังดีอยู่หรือไม่ คำถามของหัวข้อนี้คือ หลัง 20 รอบ คำตอบยังคลาดเคลื่อนเท่าไร และขึ้นกับอะไร

ภาพรวมก่อนอ่าน

การวนซ้ำแต่ละรอบคือการแก้ทุกช่องหนึ่งครั้ง (**รอบกวาด** (sweep)) โดยใช้ค่าเพื่อนบ้านที่มีอยู่ตอนนั้น ถ้า $\mathbf{e}_k$ คือความคลาดเคลื่อนหลัง k รอบ รอบถัดไปจะได้ $\mathbf{e}_{k+1} = H\mathbf{e}_k$ โดย H เป็นเมทริกซ์ที่ไม่เปลี่ยนตามรอบ สิ่งที่ต้องรู้จึงเป็นว่า H ทำให้ความคลาดเคลื่อนเล็กลงเร็วแค่ไหน ภาพด้านล่างแสดงความคลาดเคลื่อนหลัง 1, 5, 20 และ 100 รอบบนกริด 15×15 โดยแถวบนคือ $a = 1$ และแถวล่างคือ $a = 10$

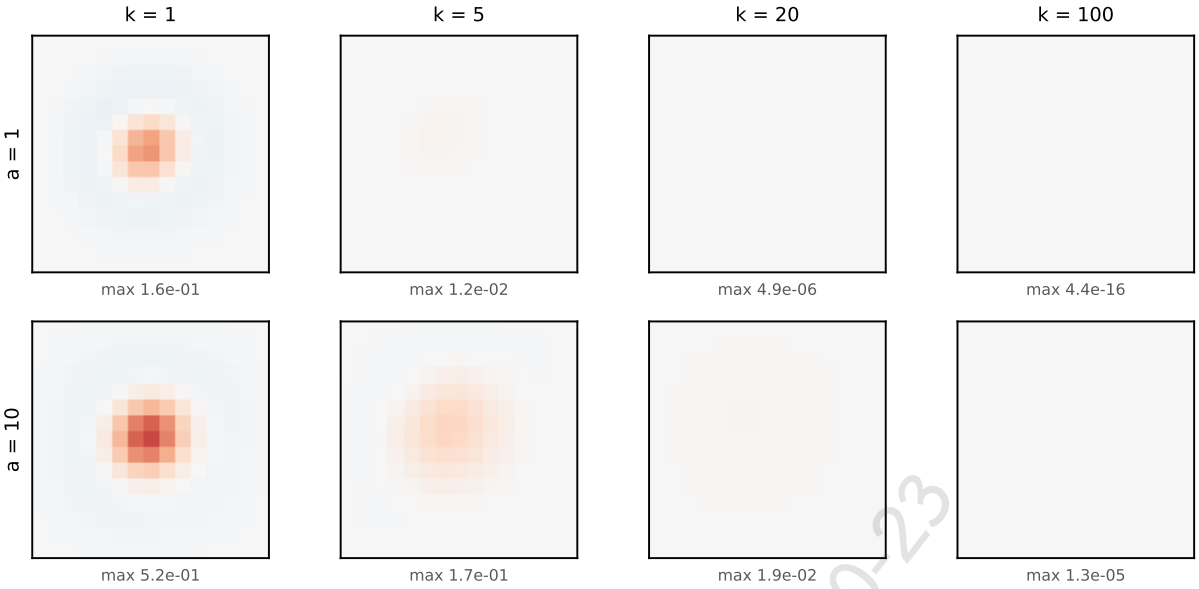
จากเฉลี่ยรอบข้างสู่การแยกเมทริกซ์

ให้ A เป็นเมทริกซ์ $n \times n$ ที่ $a_{ii} \neq 0$ ทุก i เขียนได้เป็น $A = D - L - U$ เมื่อ D เป็นเส้นทแยงมุมของ A และ L, U เป็นลบของส่วนใต้และเหนือเส้นทแยงมุมตามลำดับ สำหรับ A_m ของหัวข้อ 10.1 สมาชิกนอกแนวทแยงคือ -1 ดังนั้น L และ U มีแต่ค่าไม่เป็นลบ

วิธีทั้งสามต่างกันที่วิธีใช้ค่าเพื่อนบ้าน

- **การทำซ้ำแบบจาโคบี** (Jacobi iteration) แก้ทุกช่องพร้อมกันจากค่าเดิมทั้งหมด
- **การทำซ้ำแบบเกาส์–ไซเดล** (Gauss--Seidel iteration) แก้ทีละช่องตามลำดับ และใช้ค่าใหม่ของช่องก่อนหน้าทันที

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 10.4: ความคลาดเคลื่อนของการทำซ้ำแบบเกาส์–ไซเดลกับระบบ $(I+aA)\mathbf{x}=\mathbf{x}^0$ บนกริด 15×15 หลัง k รอบ เริ่มจากค่าปัจจุบันเป็นกระจุกควันจากหัวข้อ 10.1 สีแดงและน้ำเงินแสดงเครื่องหมาย ตัวเลขใต้ภาพคือค่าสัมบูรณ์สูงสุดของความคลาดเคลื่อน (สร้างจาก `numanal/ch10_iter.py`)

- **วิธีผ่อนเกินเรียงลำดับ** (successive over-relaxation, SOR) คือเกาส์–ไซเดลที่ผสมค่าเดิมกับค่าใหม่ด้วยน้ำหนัก ω เมื่อ $\omega>1$ จะเลยค่าใหม่ไปอีกหน่อย

สำหรับ SOR แต่ละช่องคำนวณได้เป็น

$$x_i^{\text{ใหม่}} = (1-\omega)\,x_i^{\text{เก่า}} + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij}x_j^{\text{ใหม่}} - \sum_{j>i} a_{ij}x_j^{\text{เก่า}}\Big)$$

และเมื่อรวมทุกช่องจะได้ $x^{\text{ใหม่}} = Hx^{\text{เก่า}} + v$ โดย

$$H_{\mathrm{J}} = D^{-1}(L+U), \qquad H_{\mathrm{GS}} = (D-L)^{-1}U, \qquad H_\omega = (D-\omega L)^{-1}\big((1-\omega)D+\omega U\big).$$

เกาส์–ไซเดลคือ H_ω เมื่อ $\omega=1$

อ่านแบบเข้าใจง่าย

จาโคบีช้ากว่าเพราะไม่เคยใช้ค่าที่เพิ่งคำนวณได้ในรอบเดียวกัน เกาส์–ไซเดลใช้ค่าใหม่ทันทีจึงมักเร็วกว่า ส่วน SOR เปรียบเหมือนการเดินเลยจุดที่เกาส์–ไซเดลจะไปถึง ถ้าเลยพอดีจะถึงเร็วขึ้นมาก แต่ถ้าเลยมากเกินไปก็จะแกว่งและช้าลงได้ ค่า ω ที่ดีที่สุดจึงเป็นเรื่องสำคัญ (ดูหัวข้อย่อยด้านล่าง)

ลองทำด้วยมือบนกริด 2×2

ลองเอง: สองรอบด้วยมือบนกริด 2×2

ใช้เมทริกซ์ A_2 ของหัวข้อ 10.1 (สี่ช่อง เรียงแถวต่อแถวเป็น $x_1,\dots,x_4$) โดยให้ผลเฉลยแม่นตรง $\mathbf{u}=(2,1,1,0)$ และ $\mathbf{b}=A_2\mathbf{u}=(6,2,2,-2)$ เริ่มจาก $\mathbf{x}^0=\mathbf{0}$

จาโคบี. แต่ละช่องใช้ค่าเดิมทั้งหมด แถวแรกคือ $4x_1 = 6 + x_2 + x_3$ เมื่อค่าเดิมเป็นศูนย์จะได้ $x_1 = 1.5$ ส่วนแถว

ฉบับร่าง Preprint 2026-10-23

อื่นได้ $x_2 = x_3 = 0.5$ และ $x_4 = -0.5$ ดังนั้น $\mathbf{x}^1 = (1.5,\ 0.5,\ 0.5,\ -0.5)$ รอบที่สองใช้ $\mathbf{x}^1$ ทั้งหมด จะได้ $\mathbf{x}^2 = (1.75,\ 0.75,\ 0.75,\ -0.25)$

เกาส์–ไซเดล. รอบแรก $x_1 = \frac{6}{4} = 1.5$ แล้ว $x_2 = \frac{2+x_1+x_4}{4} = \frac{2+1.5+0}{4} = 0.875$ ทันที และ $x_3 = 0.875$ ทันที สุดท้าย $x_4 = \frac{-2+x_2+x_3}{4} = -0.0625$ ได้ $\mathbf{x}^1 = (1.5,\ 0.875,\ 0.875,\ -0.0625)$ รอบที่สองได้ $\mathbf{x}^2 = (1.9375,\ 0.96875,\ 0.96875,\ -0.015625)$

เปรียบเทียบ. ค่าคลาดเคลื่อนสูงสุดหลังสองรอบคือ 0.25 สำหรับจาโคบี และ 0.0625 สำหรับเกาส์–ไซเดล อัตราส่วนคือ $\frac{1}{4}$ ซึ่งเท่ากับกำลังสองของอัตราส่วนของจาโคบี ($\frac{1}{2}$) เราจะพิสูจน์ความสัมพันธ์นี้ในทฤษฎีบท 10.2.5

ตรวจด้วยโค้ด: `test_ch10_iter.py` (`test_hand_example_*`)

การลู่เข้า: เมทริกซ์การทำซ้ำ

นิยาม 10.2.1: การวนซ้ำแบบคงที่และเมทริกซ์การทำซ้ำ

ให้ระบบ $A\mathbf{x} = \mathbf{b}$ ที่ $\mathbf{x}^*$ เป็นคำตอบ วิธีวนซ้ำ **วิธีทำซ้ำเชิงเส้นขั้นเดียวแบบคงที่** (linear one-step stationary scheme) คือ $\mathbf{x}^{k+1} = H\mathbf{x}^k + \mathbf{v}$ เมื่อ H และ $\mathbf{v}$ ไม่ขึ้นกับ k โดยที่ $\mathbf{x}^* = H\mathbf{x}^* + \mathbf{v}$ เราเรียก H นี้ว่า **เมทริกซ์การทำซ้ำ** (iteration matrix) และเรียก $\rho(H) = \max|\lambda|$ ว่า **รัศมีสเปกตรัม** (spectral radius) ของ H

อ่านแบบเข้าใจง่าย

เมทริกซ์การทำซ้ำ (iteration matrix) คือเมทริกซ์ที่คูณค่ารอบก่อนเพื่อได้รอบใหม่ **รัศมีสเปกตรัม** (spectral radius) คือขนาดของค่าลักษณะเฉพาะที่ใหญ่ที่สุด ถ้าค่านี้น้อยกว่า 1 ทุกรอบความคลาดเคลื่อนจะหดลงโดยประมาณ ถ้ามากกว่าหรือเท่ากับ 1 ความคลาดเคลื่อนบางส่วนจะไม่หาย

ทฤษฎีบท 10.2.1: การลู่เข้าของการวนซ้ำแบบคงที่

ให้ $\mathbf{x}^{k+1} = H\mathbf{x}^k + \mathbf{v}$ และ $\mathbf{x}^* = H\mathbf{x}^* + \mathbf{v}$ แล้ว

1. ถ้า $\rho(H) < 1$ จะได้ $\mathbf{x}^k \to \mathbf{x}^*$ สำหรับทุกค่าเริ่มต้น $\mathbf{x}^0$
2. ถ้า $\rho(H) \geq 1$ มีค่าเริ่มต้น $\mathbf{x}^0$ ที่ $\mathbf{x}^k \not\to \mathbf{x}^*$

พิสูจน์. **แนวคิด.** ความคลาดเคลื่อนคูณด้วย H ทุกรอบ ดังนั้นต้องดูว่า H^k เล็กลงหรือไม่ ส่วนที่พิสูจน์เต็มคือกรณีที่มีนอร์มของ H น้อยกว่า 1

1. **สมการของความคลาดเคลื่อน.** ให้ $\mathbf{e}_k = \mathbf{x}^k - \mathbf{x}^*$ เมื่อลบสมการของ $\mathbf{x}^*$ จาก $\mathbf{x}^{k+1} = H\mathbf{x}^k + \mathbf{v}$ จะได้ $\mathbf{e}_{k+1} = H\mathbf{e}_k$ และ $\mathbf{e}_k = H^k\mathbf{e}_0$
2. **กรณีมีนอร์มเล็กกว่า 1.** ถ้า $\|H\| < 1$ สำหรับนอร์มเชิงเหนี่ยวนำบางตัว จะได้ $\|\mathbf{e}_k\| \leq \|H\|^k\|\mathbf{e}_0\| \to 0$ จึงได้ข้อ 1 ในกรณีนี้
3. **กรณีทั่วไป (อ้างผล).** ถ้า $\rho(H) < 1$ จะได้ $H^k \to 0$ (Braess 2007, ทฤษฎีบท 1.3, p. 189) และจึงลู่เข้าสำหรับทุก $\mathbf{e}_0$ ส่วนข้อ 2 พร้อมพิสูจน์อยู่ใน Iserles 2009, Lemma 12.1, p. 252 ดูเพิ่มบทที่ 12 บทตั้ง 12.1 ของ Iserles

สรุป: ความคลาดเคลื่อนถูกคูณด้วย H ทุกรอบ ดังนั้นสิ่งที่ตัดสินคือ $\rho(H)$ ว่าน้อยกว่า 1 หรือไม่ □

ฉบับร่าง · preprint · 2026-10-23

ข้อควรระวัง: $\rho(H) < 1$ ตัดสินการลู่เข้า ไม่ใช่ความเร็วทั้งหมด

ค่า $\rho(H)$ บอกอัตราการลดลงเมื่อวนซ้ำหลายรอบ แต่ในรอบแรก ๆ ความคลาดเคลื่อนอาจลดได้ช้ากว่า ρ^k เพราะ H ไม่จำเป็นต้องมีเวกเตอร์ลักษณะเฉพาะตั้งฉากกัน ดังนั้นการนับจำนวนรอบจริงต้องดูเพิ่ม และในหัวข้อนี้เราใช้ ρ^k เป็นตัวประมาณเท่านั้น

เมื่อ A สมมาตรบวกแน่นอน: ความคลาดเคลื่อนไม่โต

ทฤษฎีบท 10.2.2: เกาส์–ไซเดลและ SOR กับเมทริกซ์สมมาตรบวกแน่นอน

ให้ A เป็นเมทริกซ์สมมาตรบวกแน่นอน แล้ว

1. การทำซ้ำแบบเกาส์–ไซเดลลู่เข้าสำหรับทุกค่าเริ่มต้น
2. วิธีผ่อนเกินเรียงลำดับ (SOR) ลู่เข้าสำหรับทุก $\omega \in (0, 2)$
3. ความคลาดเคลื่อนมีนอร์มพลังงาน $\|\mathbf{e}\|_A = \sqrt{\mathbf{e}^{\top} A \mathbf{e}}$ ไม่โตขึ้นในทุกรอบ คือ $\|\mathbf{e}_{k+1}\|_A \le \|\mathbf{e}_k\|_A$

พิสูจน์. **แนวคิด.** เปลี่ยนมุมมองจากนอร์มมาตรฐานเป็นนอร์มพลังงาน แล้วแสดงว่าการวนซ้ำหนึ่งรอบลดนอร์มนี้เสมอ

1. **รูปของ SOR เป็นการแยกแบบปรกติ.** ให้ $P = \frac{1}{\omega} D - L$ จะได้ $H_\omega = I - P^{-1} A$ (ตรวจได้โดยคูณ $(D - \omega L)$ ออกจากทั้งสองข้าง) สำหรับเกาส์–ไซเดลคือ $\omega = 1$ จึง $P = D - L$
2. **เอกลักษณ์ของพลังงาน.** ให้ $\mathbf{y} = P^{-1} A \mathbf{x}$ จะได้ $H_\omega \mathbf{x} = \mathbf{x} - \mathbf{y}$ และ $A\mathbf{x} = P\mathbf{y}$ ดังนั้น
$$\|H_\omega \mathbf{x}\|_A^2 = \|\mathbf{x}\|_A^2 - 2\mathbf{y}^{\top} P \mathbf{y} + \mathbf{y}^{\top} A \mathbf{y} = \|\mathbf{x}\|_A^2 - \mathbf{y}^{\top} (P + P^{\top} - A) \mathbf{y}.$$
3. **เมทริกซ์ $P + P^{\top} - A$ บวกแน่นอน.** เพราะ A สมมาตร $U = L^{\top}$ จึงได้ $P + P^{\top} - A = \left(\frac{2}{\omega} - 1\right) D$ ซึ่งบวกแน่นอนเมื่อ $0 < \omega < 2$ และเมื่อ $\omega = 1$ ได้ D ซึ่งบวกแน่นอนเสมอ
4. **สรุป.** ถ้า $\mathbf{x} \neq \mathbf{0}$ แล้ว $\mathbf{y} \neq \mathbf{0}$ (เพราะ $A\mathbf{x} \neq \mathbf{0}$) จึงได้ $\|H_\omega \mathbf{x}\|_A < \|\mathbf{x}\|_A$ ดังนั้น $\|H_\omega\|_A < 1$ และ $\rho(H_\omega) \le \|H_\omega\|_A < 1$ จึงได้ข้อ 1 และ 2 จากทฤษฎีบท 10.2.1 ส่วนข้อ 3 เป็นผลโดยตรงจาก $\mathbf{e}_{k+1} = H_\omega \mathbf{e}_k$ และขั้นที่ 2

สรุป: ทุกรอบลดนอร์มพลังงานของความคลาดเคลื่อน ไม่ว่าจะวนกี่รอบ □

อ่านแบบเข้าใจง่าย

ข้อ 3 คือเหตุผลที่ตัวแก้ของสแตม *ไม่ระเบิด* ในความหมายของ **นอร์มพลังงาน** (energy norm) ความคลาดเคลื่อนไม่เคยโตขึ้นในทุกรอบ แต่ข้อนี้ไม่ได้บอกว่าความคลาดเคลื่อนจะเหลือเท่าไรหลัง 20 รอบ ความแม่นของ 20 รอบขึ้นกับ $\rho(H)$ ซึ่งเราจะคำนวณในหัวข้อย่อยสุดท้าย

ข้อควรระวัง: เงื่อนไขพอและเงื่อนไขจำเป็นต่างกัน

ทฤษฎีบท 10.2.2 บอกว่า SPD *เพียงพอ* ให้ SOR ลู่เข้า ส่วนเงื่อนไข *จำเป็น* ของ ω อยู่ในทฤษฎีบท 10.2.4 ถ้า A ไม่สมมาตร ค่า $\omega \in (0, 2)$ อย่างเดียวไม่รับประกันการลู่เข้า ส่วนผลกลับสำหรับเกาส์–ไซเดลคือ เมื่อ A สมมาตรและมีสมาชิกแนวทแยงบวก การลู่เข้าเกิดขึ้นก็ต่อเมื่อ A บวกแน่นอน Atkinson 1989, Thm 8.7, p. 551

ฉบับร่าง Preprint 2026-10-23

เด่นทแยงและขอบเขตของ ω

นิยาม 10.2.2: เมทริกซ์เด่นทแยงอย่างเข้ม

ถ้าในทุกแถวของ A สมาชิกบนแนวทแยงมีขนาดมากกว่าผลรวมขนาดของสมาชิกที่เหลือ คือ

$$|a_{ll}| > \sum_{j \neq l} |a_{lj}| \qquad \text{สำหรับทุก } l,$$

อสมการต้องเคร่งครัดในทุกแถว (มากกว่า ไม่ใช่มากกว่าหรือเท่ากัน) เราเรียก A ว่า**เมทริกซ์เด่นทแยงอย่างเข้ม** (strictly diagonally dominant matrix)

อ่านแบบเข้าใจง่าย

แต่ละแถวต้องมีสมาชิกบนแนวทแยงใหญ่กว่าผลรวมขนาดของสมาชิกอื่นในแถวนั้น เช่นแถวของ A_m ที่ช่องริมมีเพื่อนบ้านสามช่อง และช่องมุมมีเพื่อนบ้านสองช่อง ผลรวมขนาดนอกแนวทแยงจึงเป็น 3 หรือ 2 ซึ่งน้อยกว่า 4 แต่แถวกลางมีผลรวมเป็น 4 พอดี ทำให้ไม่เข้มงวดทุกแถว

ทฤษฎีบท 10.2.3: จาโคบีและเกาส์–ไซเดลกับเมทริกซ์เด่นทแยงอย่างเข้ม

ถ้า A เป็นเมทริกซ์เด่นทแยงอย่างเข้ม แล้วการทำซ้ำแบบจาโคบีและแบบเกาส์–ไซเดลลู่เข้าสำหรับทุกค่าเริ่มต้น

พิสูจน์. **แนวคิด.** นอร์มค่าสัมบูรณ์รวมสูงสุดของแถว (นอร์มอนันต์) ของ H น้อยกว่า 1 สำหรับจาโคบีโดยตรง ส่วนเกาส์–ไซเดลต้องติดตามทีละแถว

1. **จาโคบี.** ให้ $r_l = \sum_{j \neq l} |a_{lj}|/|a_{ll}|$ และ $\mu = \max_l r_l < 1$ แถวของ H_{J} คือ $|h_{lj}| = |a_{lj}|/|a_{ll}|$ เมื่อ $j \neq l$ ดังนั้น $\|H_{\mathrm{J}}\|_\infty = \mu < 1$
2. **เกาส์–ไซเดล: ตัวแปรของความคลาดเคลื่อน.** ให้ $\mathbf{z} = H_{\mathrm{GS}}\mathbf{y}$ จะได้ $z_l = -\sum_{j<l} \frac{a_{lj}}{a_{ll}} z_j - \sum_{j>l} \frac{a_{lj}}{a_{ll}} y_j$ ดังนั้น $|z_l| \le \beta_l \max_{j<l} |z_j| + \alpha_l \|\mathbf{y}\|_\infty$ เมื่อ $\beta_l = \sum_{j<l} |a_{lj}/a_{ll}|$ และ $\alpha_l = \sum_{j>l} |a_{lj}/a_{ll}|$
3. **ขอบเขตที่เท่ากันทุกแถว.** กำหนด $\kappa = \max_l \frac{\alpha_l}{1-\beta_l}$ เพราะ $\alpha_l + \beta_l = r_l < 1$ จึงได้ $\alpha_l < 1 - \beta_l$ และ $\kappa < 1$ เราแสดงโดยเหนี่ยวนำว่า $|z_l| \le \kappa \|\mathbf{y}\|_\infty$ สำหรับทุก l เมื่อ $l = 1$ ได้ $\beta_1 = 0$ และ $|z_1| \le \alpha_1 \|\mathbf{y}\|_\infty \le \kappa \|\mathbf{y}\|_\infty$ เมื่อ $l > 1$ ถ้าข้อนี้จริงสำหรับ $j < l$ จะได้ $|z_l| \le (\beta_l \kappa + \alpha_l)\|\mathbf{y}\|_\infty \le \kappa \|\mathbf{y}\|_\infty$ เพราะ $\alpha_l \le \kappa(1 - \beta_l)$ ดังนั้น $\|H_{\mathrm{GS}}\|_\infty \le \kappa < 1$

สรุป: แถวที่เข้มงวดทำให้นอร์มอนันต์ของ H น้อยกว่า 1 ซึ่งให้ $\rho(H) < 1$ ทั้งสองวิธี □

ข้อควรระวัง: ระบบห้าจุดไม่เข้มงวดทุกแถว

แถวกลางของ A_m มีสมาชิกบนแนวทแยงเท่ากับผลรวมขนาดของเพื่อนบ้าน คือ $4 = 4$ จึงไม่ใช่เมทริกซ์เด่นทแยงอย่างเข้ม ทฤษฎีบท 10.2.3 จึงใช้กับ A_m โดยตรงไม่ได้ ข้อเท็จจริงที่ใกล้เคียงคือ ถ้าเมทริกซ์เป็น**เด่นทแยงแบบลดทอนไม่ได้** (irreducibly diagonally dominant) คือทุกแถวเด่นทแยงและมีอย่างน้อยหนึ่งแถวเข้มงวด จาโคบีและเกาส์–ไซเดลก็ยังลู่เข้า Iserles 2009, Def. 12.2, Thm 12.3, p. 261 ดูเพิ่มบทที่ 12 ทฤษฎีบท 12.3 ของ Iserles เราจะได้ค่าที่แม่นกว่านั้นในทฤษฎีบท 10.2.5

ฉบับร่าง Preprint · 2026-10-23

ทฤษฎีบท 10.2.4: ขอบเขตของ ω

ถ้า $\rho(H_\omega) < 1$ แล้ว $\omega \in (0,2)$ นั่นคือ $\omega \in (0,2)$ เป็นเงื่อนไข *จำเป็น* ของการลู่เข้า และ $\rho(H_\omega) \geq |1-\omega|$ เสมอ

พิสูจน์. **แนวคิด.** ดีเทอร์มิแนนต์ของ H_ω หาได้ง่ายเพราะ $D-\omega L$ และ $(1-\omega)D+\omega U$ เป็นสามเหลี่ยม

1. $D-\omega L$ เป็นสามเหลี่ยมล่างที่มีแนวทแยงเป็น D จึง $\det(D-\omega L) = \det D$
2. $(1-\omega)D+\omega U$ เป็นสามเหลี่ยมบนที่มีแนวทแยงเป็น $(1-\omega)D$ จึง $\det\big((1-\omega)D+\omega U\big) = (1-\omega)^n \det D$
3. ดังนั้น $\det H_\omega = (1-\omega)^n$ ค่าดีเทอร์มิแนนต์เท่ากับผลคูณของค่าลักษณะเฉพาะ $\lambda_1 \cdots \lambda_n$ และ $|\lambda_i| \leq \rho(H_\omega)$ จึงได้ $|1-\omega|^n \leq \rho(H_\omega)^n$ นั่นคือ $|1-\omega| \leq \rho(H_\omega)$

สรุป: ดีเทอร์มิแนนต์ของ H_ω คือ $(1-\omega)^n$ จึงบังคับให้ $|1-\omega|$ ต้องเล็กกว่า 1 □

กริดห้าจุด: ค่าคงที่ที่คำนวณได้เป็นสูตร

ในทฤษฎีบทต่อไปนี้ใช้ $c = \cos\frac{\pi}{m+1}$ และ $\omega_{\text{opt}} = \dfrac{2}{1+\sqrt{1-c^2}}$

ทฤษฎีบท 10.2.5: อัตราการลู่เข้าของ A_m และ ω_{opt}

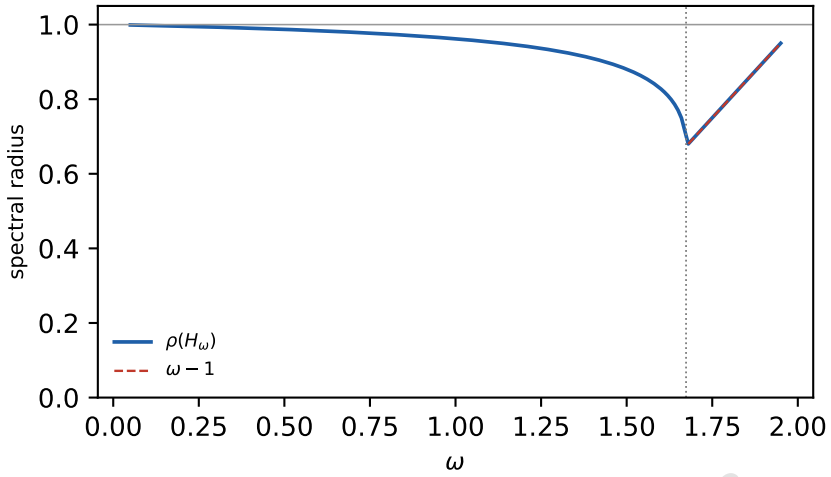
สำหรับ A_m ของหัวข้อ 10.1 และ c ดังข้างบน

1. $\rho(H_{\text{J}}) = c$
2. $\rho(H_{\text{GS}}) = c^2$
3. ค่า $\rho(H_\omega)$ น้อยที่สุดเมื่อ $\omega = \omega_{\text{opt}}$ และมีค่า $\rho(H_{\omega_{\text{opt}}}) = \omega_{\text{opt}} - 1$
4. สำหรับเมทริกซ์การทำซ้ำของ $I + aA_m$ $(a \geq 0)$ จะได้ $\rho(H_{\text{J}}) = \dfrac{4ac}{1+4a}$ และ $\rho(H_{\text{GS}}) = \left(\dfrac{4ac}{1+4a}\right)^2$

พิสูจน์. **แนวคิด.** ข้อ 1 และ 4 คำนวณตรงจากค่าลักษณะเฉพาะในทฤษฎีบท 10.1.2; ข้อ 2 และ 3 อ้างผลของยัง (Young) ซึ่งใช้ได้เพราะ A_m มีเวกเตอร์การจัดลำดับที่เข้ากันได้

1. **จาโคบี (พิสูจน์เต็ม).** เพราะ $D = 4I$ และ $L+U = 4I - A_m$ จึง $H_{\text{J}} = I - \frac{1}{4}A_m$ จากทฤษฎีบท 10.1.2 ค่าลักษณะเฉพาะคือ $1 - \frac{1}{4}(\lambda_p + \lambda_q) = \frac{1}{2}(\cos\theta_p + \cos\theta_q)$ เมื่อ $\theta_p = \frac{p\pi}{m+1}$ เพราะ $|\cos\theta_p| \leq c$ และค่าเท่ากับ c ที่ $p = q = 1$ จึงได้ $\rho(H_{\text{J}}) = c$
2. **เกาส์–ไซเดล (อ้างผล).** กราฟของ A_m มีเวกเตอร์การจัดลำดับ $j(i,j) = i + j$ ที่เข้ากันได้ เพราะเพื่อนบ้านต่างกันหนึ่งขั้น และเพื่อนบ้านที่อยู่ถัดไปในลำดับแถวต่อแถวมีค่า j มากกว่าหนึ่ง จึงใช้ Iserles 2009, Thm 12.8, p. 274 (ข้อ $\omega = 1$) ได้ $\rho(H_{\text{GS}}) = \rho(H_{\text{J}})^2$ เช่นเดียวกับกรณีเส้นสามแนวเฉียงใน Iserles 2009, Cor. 12.8.2, p. 275
3. **SOR (อ้างผลของยัง).** จาก Iserles 2009, Thm 12.10, p. 276 และ (12.47)–(12.48), p. 277 เมื่อค่าลักษณะเฉพาะของ H_{J} เป็นจำนวนจริงและ $\rho(H_{\text{J}}) < 1$ ได้ ω_{opt} ที่ทำให้ $\rho(H_\omega)$ น้อยที่สุด และ $\rho(H_{\omega_{\text{opt}}}) = \omega_{\text{opt}} - 1$ ไอเดียคือ ค่าลักษณะเฉพาะ λ ของ H_ω สัมพันธ์กับค่าลักษณะเฉพาะ μ ของ H_{J} ผ่านสมการ $\lambda + \omega - 1 = \omega\mu\lambda^{1/2}$ ซึ่งเป็นสมการกำลังสองใน $\lambda^{1/2}$ ขนาดของรากทั้งสองลดลงสูงสุดเมื่อ ω ทำให้รากในนั้นเป็นรากคู่กันพอดี

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 10.5: รัศมีสเปกตรัม $\rho(H_\omega)$ ของ SOR บนกริด 15×15 เส้นเต็มคือค่าที่คำนวณจริง เส้นประสีแดงคือ $\omega - 1$ ซึ่งเป็นค่าสำหรับ $\omega \geq \omega_{\text{opt}}$ เส้นจุดแนวตั้งคือ $\omega_{\text{opt}} \approx 1.6735$ (สร้างจาก `numanal/ch10_iter.py`)

ตารางที่ 10.2: อัตราการลู่เข้าของเกาส์–ไซเดลกับ $I + aA$ บนกริด 128×128 คอลัมน์สุดท้ายคือจำนวนรอบที่ทำให้ $\rho^k \leq 10^{-6}$ (ตัวเลขคำนวณตามสูตรของเรา)

a	$\rho(H_{\text{GS}})$	ρ^{20} (หลัง 20 รอบ)	รอบที่ต้องใช้ให้เหลือ 10^{-6}
0.1	0.0816	1.7×10^{-22}	6
1	0.6396	1.3×10^{-4}	31
10	0.9513	0.368	277
100	0.9944	0.894	2473

4. **สมการทั่วไป.** ที่นี่ $D = (1+4a)I$ และ $L+U = a(4I - A_m)$ จึง $H_{\text{J}} = \frac{a}{1+4a}(4I - A_m)$ ค่าลักษณะเฉพาะคือ $\frac{a}{1+4a}(2\cos\theta_p + 2\cos\theta_q)$ ซึ่งมีขนาดสูงสุด $\frac{4ac}{1+4a}$ ส่วนเกาส์–ไซเดลใช้ข้อ 2 แบบเดิม เพราะเวกเตอร์การจัดลำดับไม่เปลี่ยน

สรุป: ค่าคงที่ของกริดห้าจุดทั้งหมดคำนวณจากค่าลักษณะเฉพาะของเส้นสามแนวเฉียง ข้อ 2 และ 3 เป็นผลของยัง □

ผลข้อ 2 อธิบายตัวอย่างด้วยมือ เมื่อ $m = 2$ ได้ $c = \frac{1}{2}$ จึง $\rho(H_{\text{J}}) = \frac{1}{2}$ และ $\rho(H_{\text{GS}}) = \frac{1}{4}$ ค่า $\omega_{\text{opt}} = \frac{2}{1+\sqrt{3}/2} = 8 - 4\sqrt{3} \approx 1.0718$ และ $\rho(H_{\omega_{\text{opt}}}) = 7 - 4\sqrt{3} \approx 0.0718$ เทียบกับ $\rho(H_{\text{GS}}) = 0.25$ เมื่อเทียบอัตราการลดต่อรอบการผ่อนค่า ω ที่ถูกต้องเร็วกว่าเกาส์–ไซเดลราว 3.5 เท่าบนกริดเล็กนี้

กลับไปที่สแตม: 20 รอบพอหรือไม่

สำหรับขั้นการแพร่ของสแตมที่ $m = 128$ ค่า a คือ $\Delta t \cdot \text{diff} \cdot N^2$ ที่แปรตามค่าพารามิเตอร์ ตามทฤษฎีบท 10.2.5 ข้อ 4 รัศมีของเกาส์–ไซเดลคือ $\big(4ac/(1+4a)\big)^2$ ตารางนี้คำนวณตามสูตรนั้นด้วยกริด 128×128

ตารางบอกว่า 20 รอบพอเมื่อ a เล็กหรือปานกลาง แต่เมื่อ a ใหญ่ 20 รอบเหลือความคลาดเคลื่อนมาก ทั้งที่ตัวแก้ยังไม่ระเบิดตามทฤษฎีบท 10.2.2 ข้อ 3 ทั้งสองอย่างเป็นความจริงพร้อมกัน ความเสถียรไม่ได้รับประกันความแม่น

สมุดร่าง · Preprint · 2026-10-23

ข้อควรระวัง: ค่าตกค้างเล็กไม่ได้แปลว่าคำตอบแม่น

ค่า**ค่าตกค้าง** (residual) คือส่วนที่ยังไม่ตรงกับสมการเมื่อแทนคำตอบโดยประมาณ และเขียนได้เป็น $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ และ $\mathbf{r} = -A\mathbf{e}$ เมื่อ $\mathbf{e} = \mathbf{x} - \mathbf{x}^*$ จากนิยาม $\|\mathbf{e}\| \le \|A^{-1}\|\,\|\mathbf{r}\|$ และ $\|\mathbf{b}\| \le \|A\|\,\|\mathbf{x}^*\|$ จึงได้ขอบเขตของ**ค่าคลาดเคลื่อนสัมพัทธ์** (relative error)

$$\frac{\|\mathbf{x} - \mathbf{x}^*\|}{\|\mathbf{x}^*\|} \le \kappa(A)\,\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$$

สำหรับ A_8 ได้ $\kappa(A_8) \approx 32.16$ ดังนั้นค่าตกค้างสัมพัทธ์ 10^{-6} อาจยังปล่อยให้ค่าคลาดเคลื่อนสัมพัทธ์ถึง 3×10^{-5} ได้ ขอบเขตนี้เป็นขอบบน ไม่ใช่ค่าที่เกิดขึ้นเสมอ

สรุปใจความ

- ทั้งสามวิธีเป็นการวนซ้ำแบบคงที่ $\mathbf{x}^{k+1} = H\mathbf{x}^k + \mathbf{v}$ ความคลาดเคลื่อนคูณด้วย H ทุกรอบ สิ่งที่ตัดสินคือ $\rho(H)$
- เมื่อ A สมมาตรบวกแน่นอน เกาส์–ไซเดลลู่เข้า และ SOR ลู่เข้าเมื่อ $0 < \omega < 2$ นอร์มพลังงานของความคลาดเคลื่อนไม่โต ซึ่งคือเหตุผลที่ตัวแก้ไม่ระเบิด
- สำหรับกริดห้าจุด $\rho_{\mathrm{J}} = \cos\frac{\pi}{m+1}$, $\rho_{\mathrm{GS}} = \rho_{\mathrm{J}}^2$ และ SOR ที่ ω_{opt} เร็วกว่าเกาส์–ไซเดลมาก
- กลับไปที่สแตม: 20 รอบพอเมื่อ a เล็ก และไม่พอเมื่อ a ใหญ่ ค่า a ขึ้นกับ Δt, diff และขนาดกริด ดังนั้นการเลือก 20 รอบไม่ได้รับประกันความแม่นทุกกรณี

อ่านต่อ. Iserles 2009, §12.1--12.3, pp. 251--281 มีทฤษฎีของยังและเงื่อนไขของเวกเตอร์การจัดลำดับ ดูเพิ่มบทที่ 12 ทฤษฎีบท 12.8--12.10 ของ Iserles ดูเพิ่มบทที่ 12 ทฤษฎีบท 12.8--12.10 ของ Iserles Braess 2007, §IV.1, pp. 187--195 มี**ทฤษฎีบทออสตรอฟสกี–ไรช์** (Ostrowski--Reich theorem) ที่ให้ความสมมูลสำหรับ SOR กับ SPD Atkinson 1989, §8.6--8.7, pp. 544--557 ให้วิธีทำซ้ำแบบจาโคบีและเกาส์–ไซเดลพร้อมตัวอย่าง และ Grossmann, Roos, and Stynes 2007, Lemma 8.12, p. 517 ให้ผลเกาส์–ไซเดลกับ SPD ด้วยนอร์มพลังงาน วิธีที่เร็วกว่าการวนซ้ำแบบนี้มากคือการมองปัญหาจากกริดหยาบก่อน ซึ่งเป็นหัวข้อ 10.3

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 10.2.1 (⋆, ทำด้วยมือ) ใช้ระบบเดียวกับตัวอย่างด้วยมือ คือ A_2 และ $\mathbf{b} = (6, 2, 2, -2)$ แต่เริ่มจาก $\mathbf{x}^0 = (1, 1, 1, 1)$ ทำหนึ่งรอบของเกาส์–ไซเดล (ใช้ค่าใหม่ทันที) แล้วหาค่าคลาดเคลื่อน $\mathbf{e}^1 = \mathbf{x}^1 - \mathbf{u}$ เมื่อ $\mathbf{u} = (2, 1, 1, 0)$ จงบอกว่าค่าคลาดเคลื่อนสัมบูรณ์สูงสุดเท่าใด และค่านั้นอยู่ที่ช่องใด

แบบฝึกหัด 10.2.2 (⋆⋆, พิสูจน์และคำนวณ) ให้ $A = A_2$ และ $\omega = 2.1$ ตามทฤษฎีบท 10.2.4

1. จงแสดงว่า $\det H_\omega = (1 - \omega)^4$ โดยใช้ว่า $D - \omega L$ และ $(1 - \omega)D + \omega U$ เป็นเมทริกซ์สามเหลี่ยม
2. จงหาขอบเขตล่างของ $\rho(H_{2.1})$
3. จงอธิบายว่าทำไมการทำซ้ำนี้จึงไม่ลู่เข้า

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 10.2.3 (⋆⋆, คำนวณ) ให้ $B = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}$

1. จงตรวจว่า B เป็นเมทริกซ์เด่นทแยงอย่างเข้ม
2. จงหา $\|H_{\mathrm{J}}\|_\infty$ โดยคำนวณ $H_{\mathrm{J}} = D^{-1}(L+U)$ ตรง ๆ
3. จงหา $\rho(H_{\mathrm{J}})$ โดยใช้ค่าลักษณะเฉพาะ แล้วตรวจว่า $\rho(H_{\mathrm{J}}) \le \|H_{\mathrm{J}}\|_\infty$

แบบฝึกหัด 10.2.4 (⋆⋆, พิสูจน์และคำนวณ) ให้ $A = A_2$ และใช้ระบบของตัวอย่างด้วยมือ เริ่มจาก $\mathbf{x}^0 = \mathbf{0}$ ค่าคลาดเคลื่อนหลังหนึ่งรอบของเกาส์–ไซเดลคือ $\mathbf{e}^1 = (-\tfrac{1}{2}, -\tfrac{1}{8}, -\tfrac{1}{8}, -\tfrac{1}{16})$ และ $\mathbf{e}^0 = (-2,-1,-1,0)$

1. จงแสดงว่า $P + P^{\top} - A = D$ เมื่อ $P = D - L$
2. จงคำนวณ $\|\mathbf{e}^0\|_A$ และ $\|\mathbf{e}^1\|_A$ แล้วหาอัตราส่วน ตรวจว่าลดลงตามทฤษฎีบท 10.2.2

ตัวช่วยคิด

สำหรับเวกเตอร์ $\mathbf{e}$ จะได้ $\mathbf{e}^{\top}A\mathbf{e} = 4\sum e_i^2 - 2(e_1e_2 + e_1e_3 + e_2e_4 + e_3e_4)$ สำหรับกริด 2×2

แบบฝึกหัด 10.2.5 (⋆⋆, คำนวณ) ให้ A_m ตามหัวข้อ 10.1 และ $c = \cos\frac{\pi}{m+1}$

1. จงหา $\rho(H_{\mathrm{J}})$ และ $\rho(H_{\mathrm{GS}})$ สำหรับ $m = 2$ และ $m = 3$ เป็นค่าที่แน่นอน
2. สำหรับ $m = 8$ จงหาจำนวนรอบ k ที่น้อยที่สุดที่ทำให้ $\rho^k \le 10^{-6}$ ทั้งของจาโคบีและของเกาส์–ไซเดล แล้วอธิบายว่าทำไมอัตราส่วนของสองจำนวนนี้ใกล้ 2

แบบฝึกหัด 10.2.6 (⋆⋆⋆, คำนวณ) ให้ A_m และ $\omega_{\mathrm{opt}} = \dfrac{2}{1+\sqrt{1-c^2}}$

1. สำหรับ $m = 2$ จงแสดงว่า $\omega_{\mathrm{opt}} = 8 - 4\sqrt{3}$ และ $\rho(H_{\omega_{\mathrm{opt}}}) = 7 - 4\sqrt{3}$
2. สำหรับ $m = 15$ จงหา ω_{opt} และ $\rho(H_{\omega_{\mathrm{opt}}})$ แล้วหาจำนวนรอบที่ทำให้ $\rho^k \le 10^{-6}$ สำหรับ SOR ที่ ω_{opt} เทียบกับเกาส์–ไซเดลและจาโคบี
3. จงคำนวณ $\rho(H_{1.9})$ สำหรับ $m = 15$ แล้วอธิบายว่าเหตุใด $\omega = 1.9$ จึงให้อัตราที่แย่กว่ามากเมื่อเทียบกับ ω_{opt} (ใช้ $\rho(H_\omega) = \omega - 1$ เมื่อ $\omega \ge \omega_{\mathrm{opt}}$ จาก Iserles 2009, (12.48), p. 277)

แบบฝึกหัด 10.2.7 (⋆⋆, คำนวณและตีความ) ให้ $M = I + aA_m$ เป็นเมทริกซ์ของขั้นการแพร่ และ $m = 128$

1. จงตรวจตารางที่ 10.2 สำหรับ $a = 10$ โดยใช้สูตร $\rho(H_{\mathrm{GS}}) = \left(\frac{4ac}{1+4a}\right)^2$
2. สำหรับ $a = 10$ จงคำนวณ ω_{opt} ของ M (ใช้ $\rho(H_{\mathrm{J}}) = \frac{4ac}{1+4a}$) และหาจำนวนรอบที่ทำให้ $\rho^k \le 10^{-6}$ สำหรับ SOR ที่ค่านี้ เทียบกับ 277 รอบของเกาส์–ไซเดล
3. จงแสดงว่าเมื่อ $a \to \infty$ อัตราของเกาส์–ไซเดลเข้าใกล้ c^2 จึงอธิบายได้ว่าทำไมการวนซ้ำ 20 รอบจึงไม่พอเมื่อ a ใหญ่

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 10.2.8 (⋆ ⋆ ⋆, พิสูจน์และคำนวณ) ให้ A เป็นเมทริกซ์สมมาตรบวกแน่นอน $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ คือค่าตกค้าง และ $\mathbf{e} = \mathbf{x} - \mathbf{x}^*$ คือค่าคลาดเคลื่อน โดย $A\mathbf{x}^* = \mathbf{b}$

1. จงแสดงว่า $\mathbf{r} = -A\mathbf{e}$ และ $\dfrac{\|\mathbf{e}\|_2}{\|\mathbf{x}^*\|_2} \le \kappa(A)\dfrac{\|\mathbf{r}\|_2}{\|\mathbf{b}\|_2}$
2. สำหรับ $A = A_8$ จงหา $\kappa(A_8)$ โดยใช้ค่าลักษณะเฉพาะจากทฤษฎีบท 10.1.2 แล้วหาขอบเขตของค่าคลาดเคลื่อนสัมพัทธ์เมื่อค่าตกค้างสัมพัทธ์เท่ากับ 10^{-6}
3. จงแสดงว่าขอบเขตในข้อ 1 เท่ากันได้จริง โดยเลือก $\mathbf{x}^* = \mathbf{v}_{\max}$ (เวกเตอร์ลักษณะเฉพาะหน่วยที่ค่าลักษณะเฉพาะใหญ่ที่สุด) และ $\mathbf{x} = \mathbf{x}^* + \delta\,\mathbf{v}_{\min}$ (เมื่อ $\mathbf{v}_{\min}$ เป็นเวกเตอร์หน่วยที่ค่าลักษณะเฉพาะเล็กที่สุด)

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 10.2.1. เริ่มจาก $\mathbf{x}^0 = (1,1,1,1)$ และ $\mathbf{b} = (6,2,2,-2)$ แต่ละแถวใช้ค่าใหม่ที่เพิ่งคำนวณ

1. แถวแรก $4x_1 = 6 + x_2 + x_3 = 6 + 1 + 1 = 8$ ได้ $x_1 = 2$
2. แถวที่สอง $4x_2 = 2 + x_1 + x_4 = 2 + 2 + 1 = 5$ ได้ $x_2 = 1.25$
3. แถวที่สาม $4x_3 = 2 + x_1 + x_4 = 5$ ได้ $x_3 = 1.25$
4. แถวที่สี่ $4x_4 = -2 + x_2 + x_3 = 0.5$ ได้ $x_4 = 0.125$

ดังนั้น $\mathbf{x}^1 = (2,\ 1.25,\ 1.25,\ 0.125)$ และ $\mathbf{e}^1 = \mathbf{x}^1 - \mathbf{u} = (0,\ 0.25,\ 0.25,\ 0.125)$ ค่าคลาดเคลื่อนสัมบูรณ์สูงสุดคือ 0.25 ที่ช่อง $(1,2)$ และ $(2,1)$
ตรวจด้วยโค้ด: `test_ch10_iter.py` (`test_exercise_start_2x2_one_sweep`)

เฉลย 10.2.2.

1. $D = 4I$ จึง $D - \omega L$ เป็นสามเหลี่ยมล่างที่แนวทแยงเป็น 4 ทั้งสี่ตัว ดังนั้น $\det(D - \omega L) = 4^4 = 256$ และ $(1-\omega)D + \omega U$ เป็นสามเหลี่ยมบนที่แนวทแยงเป็น $4(1-\omega)$ ดังนั้น $\det\big((1-\omega)D + \omega U\big) = 256(1-\omega)^4$ เมื่อหารกันจะได้ $\det H_\omega = (1-\omega)^4$
2. ค่าลักษณะเฉพาะ $\lambda_1, \dots, \lambda_4$ ของ $H_{2.1}$ มีผลคูณ $\lambda_1\lambda_2\lambda_3\lambda_4 = (1-2.1)^4 = (-1.1)^4 = 1.4641$ และ $|\lambda_i| \le \rho$ ดังนั้น $\rho^4 \ge 1.4641$ จึงได้ $\rho \ge 1.1$
3. เพราะ $\rho(H_{2.1}) \ge 1.1 > 1$ ทฤษฎีบท 10.2.1 ข้อ 2 บอกว่ามีค่าเริ่มต้นที่การวนซ้ำไม่ลู่เข้า นอกจากนี้ $\omega = 2.1$ อยู่นอกช่วง $(0,2)$ ซึ่งเป็นเงื่อนไขจำเป็นตามทฤษฎีบท 10.2.4

ตรวจด้วยโค้ด: `test_ch10_iter.py` (`test_ex1022_det_and_bound`)

ฉบับร่าง Preprint · 2026-10-23

เฉลย 10.2.3.

1. แถวแรก $4 > |-1| = 1$ แถวที่สอง $4 > |-1| + |-1| = 2$ แถวที่สาม $4 > |-1| = 1$ ทุกแถวเข้มงวด จึงเป็นเมทริกซ์เด่นทแยงอย่างเข้ม
2. $D = 4I$ และ $L + U = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ จึง $H_{\mathrm{J}} = \frac{1}{4}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ ผลรวมแถวของค่าสัมบูรณ์คือ $\frac{1}{4}$, $\frac{2}{4}$, $\frac{1}{4}$ ดังนั้น $\|H_{\mathrm{J}}\|_\infty = \frac{1}{2}$
3. เมทริกซ์ในวงเล็บคือเมทริกซ์เพื่อนบ้านของเส้นสามจุด ซึ่งมีค่าลักษณะเฉพาะ $0, \pm\sqrt{2}$ (พหุนามลักษณะเฉพาะ $\lambda^3 - 2\lambda = 0$) ดังนั้น $\rho(H_{\mathrm{J}}) = \frac{\sqrt{2}}{4} \approx 0.354$ ซึ่งไม่เกิน $\|H_{\mathrm{J}}\|_\infty = 0.5$ ตามที่ควร เพราะ $\rho(H) \le \|H\|$ สำหรับนอร์มใด ๆ

ตรวจด้วยโค้ด: `test_ch10_iter.py` (`test_ex1023_tridiagonal_dominant`)

เฉลย 10.2.4.

1. $P + P^{\top} - A = (D - L) + (D - L)^{\top} - (D - L - U)$ เมื่อ $U = L^{\top}$ (เพราะ A สมมาตร) จึง $= 2D - L - U - D + L + U = D$
2. สำหรับ $\mathbf{e}^0 = (-2, -1, -1, 0)$ จะได้ $\sum e_i^2 = 6$ และ $e_1e_2 + e_1e_3 + e_2e_4 + e_3e_4 = 2 + 2 + 0 + 0 = 4$ ดังนั้น $\mathbf{e}^{\top}A\mathbf{e} = 4 \cdot 6 - 2 \cdot 4 = 16$ และ $\|\mathbf{e}^0\|_A = 4$ สำหรับ $\mathbf{e}^1 = (-\frac{1}{2}, -\frac{1}{8}, -\frac{1}{8}, -\frac{1}{16})$ จะได้ $\sum e_i^2 = \frac{73}{256}$ และ $e_1e_2 + e_1e_3 + e_2e_4 + e_3e_4 = \frac{1}{16} + \frac{1}{16} + \frac{1}{128} + \frac{1}{128} = \frac{9}{64}$ ดังนั้น $\mathbf{e}^{\top}A\mathbf{e} = \frac{73}{64} - \frac{9}{32} = \frac{55}{64}$ และ $\|\mathbf{e}^1\|_A = \frac{\sqrt{55}}{8} \approx 0.927$ อัตราส่วนคือ $\frac{\sqrt{55}}{32} \approx 0.232 < 1$ ค่าคลาดเคลื่อนในนอร์มพลังงานลดลงในรอบเดียว ตามทฤษฎีบท 10.2.2 ข้อ 3

ตรวจด้วยโค้ด: `test_ch10_iter.py` (`test_ex1024_energy_drop`)

เฉลย 10.2.5.

1. $m = 2$: $c = \cos\frac{\pi}{3} = \frac{1}{2}$ จึง $\rho(H_{\mathrm{J}}) = \frac{1}{2}$ และ $\rho(H_{\mathrm{GS}}) = \frac{1}{4}$ $\quad$ $m = 3$: $c = \cos\frac{\pi}{4} = \frac{\sqrt{2}}{2}$ จึง $\rho(H_{\mathrm{J}}) = \frac{\sqrt{2}}{2} \approx 0.707$ และ $\rho(H_{\mathrm{GS}}) = \frac{1}{2}$ (ตามทฤษฎีบท 10.2.5)
2. สำหรับ $m = 8$: $c = \cos\frac{\pi}{9} \approx 0.93969$ จึง $\rho(H_{\mathrm{J}}) \approx 0.93969$ และ $\ln\rho(H_{\mathrm{J}}) \approx -0.062208$ เมื่อ $\ln 10^{-6} \approx -13.8155$ ได้ $k \ge 13.8155/0.062208 \approx 222.1$ จึง $k_{\mathrm{J}} = 223$ ส่วนเกาส์–ไซเดลมี $\rho = c^2$ จึง $\ln\rho = 2\ln\rho(H_{\mathrm{J}}) \approx -0.124416$ และ $k \ge 111.04$ จึง $k_{\mathrm{GS}} = 112$ อัตราส่วนของจำนวนรอบจริง ๆ คือ $\dfrac{\ln\rho_{\mathrm{J}}}{\ln\rho_{\mathrm{GS}}} = \dfrac{1}{2}$ เสมอ เพราะ $\rho_{\mathrm{GS}} = \rho_{\mathrm{J}}^2$ การปัดขึ้นเป็นจำนวนรอบเต็มทำให้ได้ $223/112 \approx 1.99$

ตรวจด้วยโค้ด: `test_ch10_iter.py` (`test_ex1025_radii_and_sweeps`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 10.2.6.

1. $m = 2$: $c = \frac{1}{2}$ จึง $\omega_{\text{opt}} = \dfrac{2}{1+\sqrt{3}/2} = \dfrac{4}{2+\sqrt{3}} = 4(2-\sqrt{3}) = 8-4\sqrt{3} \approx 1.0718$ และ $\rho(H_{\omega_{\text{opt}}}) = \omega_{\text{opt}} - 1 = 7 - 4\sqrt{3} \approx 0.0718$

2. $m = 15$: $c = \cos\frac{\pi}{16} \approx 0.980785$ จึง $c^2 \approx 0.961940$ และ $\sqrt{1-c^2} \approx 0.195090$ ดังนั้น $\omega_{\text{opt}} \approx 1.673514$ และ $\rho \approx 0.673514$ ใช้ $\ln 10^{-6} \approx -13.8155$ ได้จำนวนรอบ: SOR $k \geq 13.8155/0.39501 \approx 34.98$ จึง 35 รอบ; เกาส์–ไซเดล $\rho = c^2$ จึง $k \geq 356.04$ จึง 357 รอบ; จาโคบี $\rho = c$ จึง $k \geq 712.1$ จึง 713 รอบ

3. เมื่อ $\omega = 1.9 \geq \omega_{\text{opt}}$ ได้ $\rho(H_{1.9}) = 1.9 - 1 = 0.9$ จึงต้องใช้ $k \geq 13.8155/0.10536 \approx 131.1$ จึง 132 รอบ มากกว่าที่ ω_{opt} ให้ราว 3.8 เท่า ค่าที่เลยเกินจุดที่เหมาะสมมากเกินไปทำให้อัตราแย่ลงอย่างรวดเร็ว (ดูรูป 10.5)

ตรวจด้วยโค้ด: `test_ch10_iter.py` (`test_ex1026_sor_optimum_and_19`)

เฉลย 10.2.7.

1. สำหรับ $a = 10$ และ $c = \cos\frac{\pi}{129} \approx 0.999703$ ได้ $\rho(H_{\text{J}}) = \frac{40c}{41} \approx 0.975320$ และ $\rho(H_{\text{GS}}) = \rho(H_{\text{J}})^2 \approx 0.95125$ ตรงกับตาราง

2. จาก $\rho(H_{\text{J}}) = 0.975320$ ได้ $\sqrt{1-\rho_{\text{J}}^2} \approx 0.220800$ จึง $\omega_{\text{opt}} = \dfrac{2}{1.220800} \approx 1.63828$ และ $\rho(H_{\omega_{\text{opt}}}) \approx 0.63828$ จำนวนรอบ $k \geq 13.8155/0.44879 \approx 30.8$ จึง 31 รอบ เทียบกับ 277 รอบของเกาส์–ไซเดลที่ $a = 10$ เดียวกัน (ค่า ω ที่เหมาะสมเปลี่ยนตาม a)

3. เมื่อ $a \to \infty$ จะได้ $\frac{4a}{1+4a} \to 1$ จึง $\rho(H_{\text{J}}) \to c$ และ $\rho(H_{\text{GS}}) \to c^2$ สำหรับ $m = 128$ คือ $c^2 \approx 0.999407$ ดังนั้น $\rho^{20} \approx 0.988$ หมายความว่า 20 รอบลดส่วนที่ซ้ำที่สุดลงเพียงราว 1.2% ตัวแก้จึงยังไม่ระเบิดตามทฤษฎีบท 10.2.2 แต่คำตอบยังไม่แม่นหลัง 20 รอบ

ตรวจด้วยโค้ด: `test_ch10_iter.py` (`test_ex1027_diffusion_sor`)

เฉลย 10.2.8.

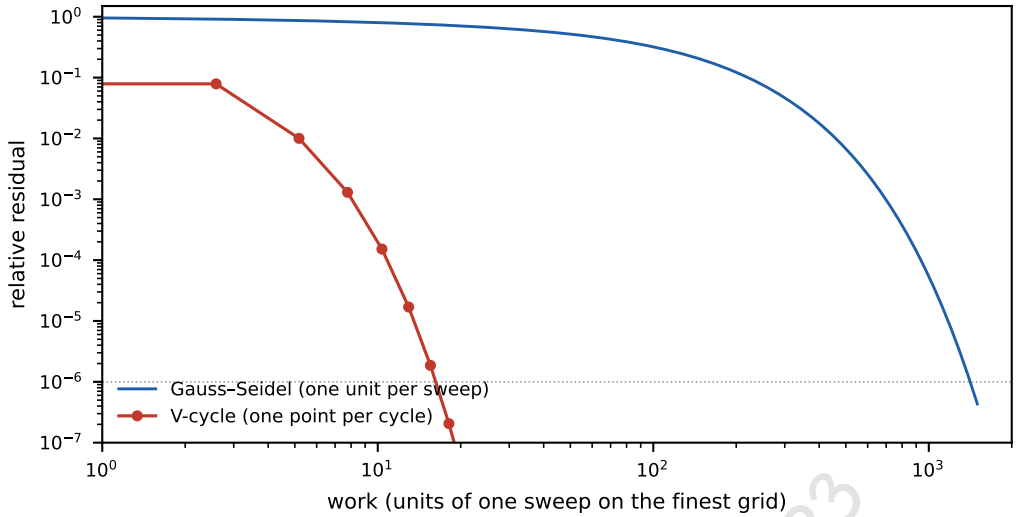
1. จาก $\mathbf{b} = A\mathbf{x}^*$ และ $\mathbf{r} = \mathbf{b} - A\mathbf{x} = A\mathbf{x}^* - A\mathbf{x} = -A\mathbf{e}$ เมื่อ $\mathbf{e} = \mathbf{x} - \mathbf{x}^*$ ได้ $\mathbf{e} = -A^{-1}\mathbf{r}$ จึง $\|\mathbf{e}\| \leq \|A^{-1}\|\|\mathbf{r}\|$ และจาก $\|\mathbf{b}\| \leq \|A\|\|\mathbf{x}^*\|$ ได้ $\frac{1}{\|\mathbf{x}^*\|} \leq \frac{\|A\|}{\|\mathbf{b}\|}$ คูณกันได้ $\dfrac{\|\mathbf{e}\|}{\|\mathbf{x}^*\|} \leq \|A\|\|A^{-1}\|\dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|} = \kappa(A)\dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$

2. A_8 สมมาตรบวกแน่นอน ค่าลักษณะเฉพาะที่เล็กที่สุดและใหญ่ที่สุดคือ $\mu_{\min} = 2\lambda_1$ และ $\mu_{\max} = 2\lambda_8$ ตามทฤษฎีบท 10.1.2 โดย $\lambda_1 = 2 - 2\cos\frac{\pi}{9}$ และ $\lambda_8 = 2 + 2\cos\frac{\pi}{9}$ ดังนั้น $\kappa(A_8) = \dfrac{\mu_{\max}}{\mu_{\min}} = \dfrac{1+c}{1-c}$ ที่ $c = \cos\frac{\pi}{9} \approx 0.939693$ จึง $\kappa(A_8) \approx 32.163$ เมื่อ $\|\mathbf{r}\|/\|\mathbf{b}\| = 10^{-6}$ ได้ขอบเขต $\|\mathbf{e}\|/\|\mathbf{x}^*\| \leq 3.2 \times 10^{-5}$

3. ให้ $\mathbf{x}^* = \mathbf{v}_{\max}$ ซึ่งมี $\|\mathbf{v}_{\max}\| = 1$ และ $\mathbf{b} = \mu_{\max}\mathbf{v}_{\max}$ จึง $\|\mathbf{b}\| = \mu_{\max}$ และ $\|\mathbf{x}^*\| = 1$ ให้ $\mathbf{e} = \delta\mathbf{v}_{\min}$ จึง $\|\mathbf{e}\|/\|\mathbf{x}^*\| = \delta$ และ $\mathbf{r} = -A\mathbf{e} = -\delta\mu_{\min}\mathbf{v}_{\min}$ จึง $\|\mathbf{r}\|/\|\mathbf{b}\| = \delta\mu_{\min}/\mu_{\max} = \delta/\kappa$ ดังนั้น $\|\mathbf{e}\|/\|\mathbf{x}^*\| = \kappa\,\|\mathbf{r}\|/\|\mathbf{b}\|$ เท่ากับขอบเขตพอดี

ตรวจด้วยโค้ด: `test_ch10_iter.py` (`test_ex1028_residual_sharpness`)

สำนักงาน · Preprint · 2026-10-23



รูปที่ 10.6: ค่าตกค้างสัมพัทธ์ $\|\mathbf{r}_k\|_2/\|\mathbf{r}_0\|_2$ ของ $A_{31}\mathbf{x}=\mathbf{1}$ เริ่มจาก $\mathbf{x}_0=\mathbf{0}$ (กริด 31×31 มี 961 ค่าที่ไม่รู้) เส้นสีน้ำเงินคือเกาส์–ไซเดลหนึ่งหน่วยงานต่อรอบ จุดสีแดงคือรอบแบบ V ทีละรอบ เส้นประคือ 10^{-6} (สร้างจาก `numanal/ch10_mg.py`)

10.3 มองจากไกลก่อน

สถานการณ์สมมติ: น้ำกระเซ็นลงสระใหญ่

สมมติว่าในฉากเกมมีสระน้ำที่แบ่งเป็นกริด 256×256 จุดภายใน เมื่อมีคนโยนก้อนหินลงกลางสระ ระลอกน้ำต้องแผ่ไปถึงขอบซึ่งห่างจากจุดกลางราว 128 ช่อง สถานการณ์นี้เป็นของสมมติ ไม่ได้วัดจากสระจริง ตัวเลขในหัวข้อนี้มาจากสูตรปิดของหัวข้อ 10.2 ที่เราคำนวณเอง

ถ้าแก้ระลอกนี้ด้วยเกาส์–ไซเดลจากหัวข้อ 10.2 ต้องใช้ราว 9.2×10^4 รอบ เพื่อให้ความคลาดเคลื่อนลดลง 10^{-6} จาโคบีต้องใช้ราว 1.85×10^5 รอบ และ SOR ที่ $\omega_{\text{opt}}=1.9758$ ต้องใช้ 566 รอบ เราจะดูว่าทำไมจำนวนรอบจึงใหญ่ขนาดนี้ และมีวิธีมองปัญหาจากไกลกว่านั้นหรือไม่

ภาพรวมก่อนอ่าน

เกาส์–ไซเดลลดส่วนที่แกว่งเร็วของความคลาดเคลื่อนได้เร็วมาก แต่ส่วนที่เรียบหายช้า ส่วนเรียบนี้จะดูแกว่งน้อยลงเมื่อมองจากกริดที่หยาบกว่า ความคิดของ**วิธีหลายกริด** (multigrid method) จึงเป็นการสลับกันระหว่างสองกริด คือ วนซ้ำบนกริดละเอียดเพื่อกำจัดส่วนแกว่ง แล้วไปแก้ส่วนเรียบบนกริดหยาบที่มีจุดน้อยกว่า และนำคำตอบกลับลงมา ที่มาของจำนวนรอบเช่น 9.2×10^4 คือ $1-\rho_{\text{GS}}=\sin^2\frac{\pi}{m+1}\approx\frac{\pi^2}{(m+1)^2}$ จึงต้องใช้รอบราว $1.4(m+1)^2$ เมื่อกริดละเอียดเป็นสองเท่า จำนวนรอบจึงเพิ่มราวสี่เท่า รูปที่ 10.6 เทียบค่าตกค้างของเกาส์–ไซเดลกับรอบแบบ V บนกริด 31×31 ตามงานที่ใช้

ฉบับร่าง preprint 2026-10-23

สองส่วนของความคลาดเคลื่อน

ความคลาดเคลื่อนของ A_m เขียนเป็นผลรวมของเวกเตอร์ $V_{p,q}$ จากทฤษฎีบท 10.1.2 โหมด (p,q) มี**เลขคลื่น** (wavenumber) (θ_p,θ_q) ถ้า θ_p หรือ θ_q มีค่าอย่างน้อย $\pi/2$ เราเรียกโหมดนั้นว่า **ความถี่สูง** (high frequency) เพราะมันแกว่งเร็วเทียบกับระยะห่างของกริด ส่วนที่เหลือ คือโหมดที่ทั้งสองทิศน้อยกว่า $\pi/2$ เรียกว่า **ความถี่ต่ำ** (low frequency) หรือส่วนเรียบ

วิธีทำซ้ำที่ลดส่วนความถี่สูงได้ดีเรียกว่า**ตัวปรับเรียบ** (smoother) เกาส์–ไซเดลเป็นตัวปรับเรียบ Iserles 2009, §13.1, pp. 296--297 แต่การพิสูจน์ที่เราทำได้เต็มในหัวข้อนี้เป็นของ**จาโคบีแบบถ่วง** (weighted Jacobi) คือการปรับค่าแต่ละช่องตามจาโคบี แต่เลือกขนาดก้าวด้วยน้ำหนัก ω เขียนเป็น $\mathbf{x}_{\text{ใหม่}} = \mathbf{x}_{\text{เก่า}} + \omega D^{-1}(\mathbf{b} - A\mathbf{x}_{\text{เก่า}})$ เมื่อ $\omega = 1$ ได้จาโคบีธรรมดา สำหรับ A_m เรามี $D = 4I$ จึงได้เมทริกซ์การทำซ้ำ $S = I - \frac{\omega}{4}A_m$

ลองเอง: หนึ่งรอบจาโคบีแบบถ่วงบนโหมดแกว่งของกริด 3×3

ใช้ $m=3$ และ $\omega=\frac{4}{5}$ จาก $D=4I$ ได้ $S = I - \frac{1}{5}A_3$ ให้ความคลาดเคลื่อนเป็นรูปที่สลับเครื่องหมายบนทุกช่องเรียงแถวต่อแถว

$$\mathbf{e} = (1,-1,1,-1,1,-1,1,-1,1)^{\top}$$

ช่องมุมของ $A_3\mathbf{e}$ ได้ $4\cdot 1-(-1-1)=6$ ช่องริมได้ $4\cdot(-1)-(1+1+1)=-7$ และช่องกลางได้ $4\cdot 1-4\cdot(-1)=8$ ดังนั้น $A_3\mathbf{e}=(6,-7,6,-7,8,-7,6,-7,6)^{\top}$ และ

$$S\mathbf{e} = \mathbf{e} - \tfrac{1}{5}A_3\mathbf{e} = \tfrac{1}{5}\,(-1,\ 2,\ -1,\ 2,\ -3,\ 2,\ -1,\ 2,\ -1)^{\top}.$$

หลังหนึ่งรอบ ค่าสัมบูรณ์ของทุกช่องไม่เกิน $\frac{3}{5}$ (ช่องกลางได้ $\frac{3}{5}$ จากเดิม $|e|=1$) ทฤษฎีบทถัดไปให้ขอบนี้สำหรับทุกโหมดความถี่สูงของทุกกริด

ทฤษฎีบท 10.3.1: ตัวปรับเรียบจากจาโคบีแบบถ่วง

ให้ $\omega=\frac{4}{5}$ และ $S=I-\frac{\omega}{4}A_m$ แล้ว

1. $V_{p,q}$ เป็นเวกเตอร์ลักษณะเฉพาะของ S ที่มีค่าลักษณะเฉพาะ $\sigma_{p,q}=\frac{1}{5}+\frac{2}{5}(\cos\theta_p+\cos\theta_q)$
2. ถ้า $\max(\theta_p,\theta_q)\ge\frac{\pi}{2}$ แล้ว $|\sigma_{p,q}|\le\frac{3}{5}$ ไม่ว่า m จะเป็นเท่าใด

พิสูจน์. **แนวคิด.** โหมด $V_{p,q}$ เป็นเวกเตอร์ลักษณะเฉพาะของ A_m อยู่แล้ว ดังนั้นจาโคบีบนแต่ละโหมดเป็นการคูณด้วยตัวเลขเดียว แล้วดูขนาดของตัวเลขนั้น

1. **ค่าลักษณะเฉพาะ.** จากทฤษฎีบท 10.1.2 $A_mV_{p,q}=(\lambda_p+\lambda_q)V_{p,q}$ โดย $\lambda_p+\lambda_q=4-2\cos\theta_p-2\cos\theta_q$ จึง
$$SV_{p,q}=\Big(1-\tfrac{1}{5}\,(4-2\cos\theta_p-2\cos\theta_q)\Big)V_{p,q}=\Big(\tfrac{1}{5}+\tfrac{2}{5}(\cos\theta_p+\cos\theta_q)\Big)V_{p,q}.$$
2. **ขอบเขต.** สมมติ $\theta_p\ge\frac{\pi}{2}$ จะได้ $-1\le\cos\theta_p\le 0$ ถ้า $\theta_q\ge\frac{\pi}{2}$ ด้วย ผลรวมของโคไซน์อยู่ใน $[-2,0]$ จึง $\sigma_{p,q}\in[-\frac{3}{5},\frac{1}{5}]$ ถ้า $\theta_q<\frac{\pi}{2}$ จะได้ $0<\cos\theta_q\le 1$ จึงผลรวมอยู่ใน $(-1,1]$ และ $\sigma_{p,q}\in(-\frac{1}{5},\frac{3}{5}]$ ทั้งสองกรณีได้ $|\sigma_{p,q}|\le\frac{3}{5}$ กรณีที่ θ_q สูงเป็นแบบเดียวกันโดยสลับ p กับ q

สรุป: ค่าลักษณะเฉพาะของจาโคบีแบบถ่วงบนแต่ละโหมดเป็นผลรวมของโคไซน์สองตัว และโหมดที่มีทิศหนึ่งแกว่งเร็วจะถูกลดไม่เกิน $\frac{3}{5}$ เท่า □

ร่างไม่เป็นทางการ · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

ผลนี้บอกว่าไม่ว่ากริดจะละเอียดแค่ไหน โหมดความถี่สูงจะถูกลดลงเหลือไม่เกิน $\frac{3}{5}$ เท่าในหนึ่งรอบ ค่าเข้าใกล้ $\frac{3}{5}$ เมื่อกริดใหญ่ขึ้น (สำหรับ $m = 15$ ค่าสูงสุดคือ 0.5923 ที่ $(p, q) = (8, 1)$) ส่วนโหมดความถี่ต่ำอย่าง $(1, 1)$ ได้ $\sigma \approx 0.985$ เท่านั้น แทบไม่หายในรอบเดียว นี่คือเหตุผลที่ต้องมีกริดหยาบมาช่วย

ข้อควรระวัง: จาโคบีธรรมดาไม่ใช่ตัวปรับเรียบ

เมื่อ $\omega = 1$ จะได้ $\sigma_{p,q} = \frac{1}{2}(\cos\theta_p + \cos\theta_q)$ โหมดที่แกว่งที่สุด (m, m) ได้ $\sigma = -\cos\frac{\pi}{m+1}$ ซึ่งใกล้ -1 มาก จึงแทบไม่ถูกลดเลย ISE แสดงข้อนี้ไว้ในแบบฝึกหัด 13.2 Iserles 2009, §13.1, p. 296 ดังนั้นตัวปรับเรียบต้องเลือกให้ดี การถ่วงน้ำหนักด้วย $\omega = \frac{4}{5}$ ทำให้โหมดแกว่งที่สุดไม่คงอยู่

กริดหยาบ: ทำให้ส่วนเรียบดูแกว่งขึ้น

ให้ $m = 2^s - 1$ และให้กริดหยาบมี $m_c = \frac{m-1}{2}$ จุดต่อด้าน โดยจุดหยาบ (I, J) ตรงกับจุดละเอียด $(2I, 2J)$ กริดทั้งสองจึงซ้อนกันพอดี Iserles 2009, §13.2, pp. 298–300 ตัวดำเนินการที่ย้ายค่าไปมาคือ

นิยาม 10.3.1: การกำกัด การขยาย และเมทริกซ์กาเลอคินของกริดหยาบ

- **เมทริกซ์การกำกัด** (restriction matrix) R ใช้**การกำกัดแบบถ่วงน้ำหนักเต็ม** (full weighting) คือ ค่าที่จุดหยาบเป็นค่าเฉลี่ยถ่วงน้ำหนักของจุดละเอียดรอบ ๆ ตามสเตนซิล $\frac{1}{16}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}$
- **เมทริกซ์การขยาย** (prolongation matrix) P ใช้**การประมาณค่าในช่วงเชิงเส้นคู่** (bilinear interpolation) คือจุดหยาบให้ค่าจุดละเอียดที่ตรงกันด้วยน้ำหนัก 1 จุดขอบข้างด้วย $\frac{1}{2}$ และจุดมุมด้วย $\frac{1}{4}$
- **เมทริกซ์กาเลอคิน** (Galerkin matrix) ของกริดหยาบคือ $A_c = RA_mP$ ซึ่งเป็นเมทริกซ์ของกริดหยาบที่สืบมาจากกริดละเอียด ต่างจากการทำสูตรห้าจุดซ้ำบนกริดหยาบ (rediscretisation) ซึ่งหนังสือ ISE ใช้ เราเลือกนิยามนี้เพราะทำให้ทฤษฎีบท 10.3.2 เป็นจริงแน่นอน

ตัวเลือกนี้ให้ $P = 4R^\top$ เพราะน้ำหนักของ P ที่จุดละเอียดแต่ละชนิดเป็นสี่เท่าของน้ำหนักของ R ที่จุดนั้น จึง $R = \frac{1}{4}P^\top$ ทฤษฎีบทถัดไปอาศัยสองสิ่งนี้ ไม่ต้องทำการคำนวณเพิ่มเติม เราใช้ $A_c = RA_mP$ แทนสูตรห้าจุดบนกริดหยาบ เพราะทำให้ทฤษฎีบทเป็นจริงแน่นอน

ทฤษฎีบท 10.3.2: การแก้ไขบนกริดหยาบ

ให้ $\mathbf{r} = \mathbf{b} - A_m\mathbf{x}$ และให้ $\mathbf{e}_c$ เป็นคำตอบของ $A_c\mathbf{e}_c = R\mathbf{r}$ ตั้งค่า $\mathbf{x}^+ = \mathbf{x} + P\mathbf{e}_c$ และ $\mathbf{r}^+ = \mathbf{b} - A_m\mathbf{x}^+$ แล้ว

1. A_c สมมาตรบวกแน่นอน
2. $\mathbf{r}^+ = (I - A_mPA_c^{-1}R)\mathbf{r}$
3. $R\mathbf{r}^+ = \mathbf{0}$ คือส่วนของค่าตกค้างที่กริดหยาบมองเห็นเป็นศูนย์หลังการแก้ไข

พิสูจน์. **แนวคิด.** การแก้บนกริดหยาบทำให้ค่าตกค้างที่กริดหยาบเห็นหายไป แล้วเขียนผลนี้ด้วยเมทริกซ์

ฉบับร่าง Preprint · 2026-10-23

1. **บวกแน่นอน.** จาก $R = \frac{1}{4}P^\top$ ได้ $A_c = \frac{1}{4}P^\top A_m P$ ซึ่งสมมาตร และถ้า $\mathbf{w} \neq \mathbf{0}$ จะได้ $P\mathbf{w} \neq \mathbf{0}$ เพราะแถวของ P ที่ตรงกับจุดหยาบเป็นเมทริกซ์เอกลักษณ์ ดังนั้น $\mathbf{w}^\top A_c \mathbf{w} = \frac{1}{4}(P\mathbf{w})^\top A_m (P\mathbf{w}) > 0$
2. **ค่าตกค้างใหม่.** $\mathbf{r}^+ = \mathbf{b} - A_m\mathbf{x} - A_m P\mathbf{e}_c = \mathbf{r} - A_m P A_c^{-1} R\mathbf{r}$
3. **หายที่กริดหยาบ.** จากข้อ 2 $R\mathbf{r}^+ = R\mathbf{r} - (RA_mP)A_c^{-1}R\mathbf{r} = R\mathbf{r} - A_cA_c^{-1}R\mathbf{r} = \mathbf{0}$

สรุป: แก้ค่าตกค้างของกริดหยาบให้เป็นศูนย์โดยเลือก $\mathbf{e}_c$ ที่เหมาะสม และการเขียนด้วยเมทริกซ์ได้ผลเดียวกัน □

ข้อควรระวัง: $R\mathbf{r}^+ = \mathbf{0}$ ไม่ได้แปลว่าค่าตกค้างเป็นศูนย์

ทฤษฎีบทนี้บอกเพียงว่าส่วนที่กริดหยาบเห็นหายไป ส่วนที่กริดหยาบมองไม่เห็นซึ่งแกว่งยังคงอยู่ และไม่ได้บอกว่าค่าตกค้างลดลงเท่าใด ดังนั้นต้องใช้ตัวปรับเรียบมาจัดการส่วนแกว่งนี้ และนั่นคือรอบแบบ V ในหัวข้อย่อยถัดไป

ลองเอง: กริด 3×3 เรียบและแกว่ง

ใช้ $m = 3$ จุดภายใน กริดหยาบมีจุดเดียว คือ $(2, 2)$ เรียงแถวต่อแถว P เป็นเวกเตอร์

$$\mathbf{p} = \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{2}, 1, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}\right)^\top, \qquad R = \tfrac{1}{4}\mathbf{p}^\top.$$

จาก $A_3\mathbf{p} = (0, \frac{1}{2}, 0, \frac{1}{2}, 2, \frac{1}{2}, 0, \frac{1}{2}, 0)^\top$ ได้ $\mathbf{p}^\top A_3 \mathbf{p} = 3$ จึง $A_c = \frac{3}{4}$
ส่วนเรียบ. ให้ความคลาดเคลื่อนเป็น $\mathbf{e} = \mathbf{p}$ (รูปเรียบที่มาจากกริดหยาบ) แล้ว $\mathbf{r} = -A_3\mathbf{e}$ และ $R\mathbf{r} = -\frac{3}{4}$ จึง $\mathbf{e}_c = -1$ ดังนั้น $\mathbf{e} + P\mathbf{e}_c = \mathbf{0}$ คือกริดหยาบกำจัดความคลาดเคลื่อนนี้ได้หมดในหนึ่งขั้น
ส่วนแกว่ง. ให้ $\mathbf{e}$ เป็นเวกเตอร์เดิมจากลองเองก่อนทฤษฎีบท 10.3.1 ซึ่งสลับเครื่องหมายระหว่างจุดข้างเคียงทุกคู่ จะได้

$$A_3\mathbf{e} = (6, -7, 6, -7, 8, -7, 6, -7, 6)^\top, \qquad R(A_3\mathbf{e}) = \tfrac{1}{16}\big(4 \cdot 8 + 2 \cdot 4 \cdot (-7) + 4 \cdot 6\big) = 0.$$

กริดหยาบมองไม่เห็นส่วนแกว่งนี้เลย ตัวปรับเรียบต้องกำจัดมันเอง

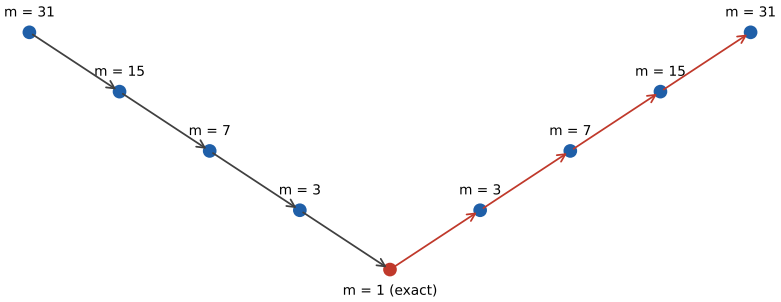
รอบแบบ V และงานที่ใช้

รอบแบบ V (V-cycle) เป็นการเดินไปกลับบนลำดับกริด $m = 2^s - 1,\ 2^{s-1} - 1, \ldots, 1$ แต่ละรอบมีขั้นดังนี้

1. ปรับเรียบบนกริดปัจจุบัน ν_1 รอบ (**การปรับเรียบก่อน** (pre-smoothing)) คือกวาดกริดละเอียดเพื่อลดส่วนแกว่งก่อนย้ายไปกริดหยาบ
2. คำนวณค่าตกค้างแล้วกำกัดลงกริดหยาบด้วย R
3. แก้ส่วนแก้ไขบนกริดหยาบโดยเริ่มจากศูนย์ ด้วยรอบแบบ V ของกริดนั้น จนถึงกริดหนึ่งจุดที่แก้ตรงได้
4. ขยายส่วนแก้ไขขึ้นด้วย P แล้วบวกเข้ากับคำตอบ
5. ปรับเรียบอีก ν_2 รอบ (**การปรับเรียบหลัง** (post-smoothing)) คือกวาดกริดละเอียดอีกครั้งหลังแก้ไขเพื่อเก็บส่วนแกว่งที่การขยายอาจเพิ่มเข้ามา

ขั้นที่ 3 และ 4 คือ**การแก้ไขบนกริดหยาบ** (coarse-grid correction) ถ้าแก้กริดหยาบตรงเสมอ รอบนี้เรียกว่า**วิธีสองกริด** (two-grid method) ซึ่งทฤษฎีบท 10.3.2 บรรยายไว้ รูปที่ 10.7 เป็นเส้นทางของรอบแบบ V บนกริด $31, 15, 7, 3, 1$

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 10.7: เส้นทางของรอบแบบ V ลงด้วยการกำกัด (ลูกศรเทา) และขึ้นด้วยการขยาย (ลูกศรแดง) จุดล่างสุดคือกริดหนึ่งจุดที่แก้ตรง (สร้างจาก `numanal/ch10_mg.py`)

ทฤษฎีบท 10.3.3: งานของรอบแบบ V

ให้ $m = 2^s - 1$ และ $\nu = \nu_1 + \nu_2$ หนึ่งหน่วยงานคือหนึ่งรอบกวาดเกาส์–ไซเดลบนกริดละเอียดที่สุด กริดหนึ่งจุดที่แก้ตรงไม่คิดงาน และ**นับเฉพาะรอบกวาด** งานของการกำกัด การขยาย และการคำนวณค่าตกค้างไม่รวมอยู่ในตัวเลขนี้ แล้วหนึ่งรอบแบบ V ใช้งานไม่เกิน

$$\frac{4}{3}\Big(1+\frac{1}{m}\Big)^2 \nu$$

หน่วยงาน ดังนั้นงานเป็นเชิงเส้นในจำนวนจุด $n = m^2$ เมื่อ $m \geq 3$ ขอบบนนี้ไม่เกิน $\frac{64}{27}\nu$ งานส่วนที่ไม่รวมนี้ก็เป็นสัดส่วนกับจำนวนจุดของแต่ละชั้น จึงยังเป็นเชิงเส้น (นับแบบเดียวกับที่ ISE ใช้ Iserles 2009, §13.2--13.3, p. 306)

พิสูจน์. **แนวคิด.** กริดแต่ละชั้นมีจุดราวหนึ่งในสี่ของชั้นที่ละเอียดกว่า ดังนั้นงานรวมเป็นอนุกรมเรขาคณิต

1. **จำนวนจุดแต่ละชั้น.** ชั้นที่ $l = 0, 1, \ldots$ มี $m_l = 2^{s-l} - 1$ จุดต่อด้าน และมี $n_l = m_l^2 \leq 4^{s-l}$ จุด
2. **รวมงาน.** ชั้นที่ถูกปรับเรียบคือ $l = 0, \ldots, s-2$ (กริดสุดท้ายแก้ตรง) งานในหนึ่งรอบคือ $\nu \sum_l n_l / n_0$ และ

$$\sum_{l=0}^{s-2} \frac{n_l}{n_0} \leq \frac{4^s}{m^2} \sum_{l=0}^{\infty} 4^{-l} = \frac{4}{3}\Big(\frac{m+1}{m}\Big)^2$$

   เพราะ $4^s = (m+1)^2$
3. **ขอบบนคงที่.** สำหรับ $m \geq 3$ จะได้ $(1+\frac{1}{m})^2 \leq \frac{16}{9}$ จึงขอบบนไม่เกิน $\frac{64}{27}\nu$ ซึ่งไม่ขึ้นกับ m และงานของหนึ่งหน่วยคือ n ดังนั้นงานรวมเป็นเชิงเส้นในจำนวนจุด

สรุป: ชั้นกริดลดจำนวนจุดลงสี่เท่า จึงรวมงานเป็นอนุกรมเรขาคณิตที่มีผลรวมคงที่ □

อ่านแบบเข้าใจง่าย

ข้อนี้บอกว่าการทำรอบแบบ V หนึ่งรอบ (ที่ $\nu = 2$) ใช้งานไม่เกิน $\frac{8}{3}\big(1+\frac{1}{m}\big)^2$ หน่วย ซึ่งเข้าใกล้ราว 2.7 รอบเกาส์–ไซเดลเมื่อกริดใหญ่ และไม่เกินราว 4.7 สำหรับทุก $m \geq 3$ สำหรับ $m \geq 15$ งานต่อหนึ่งรอบที่คำนวณจากตารางที่ 10.3 คือราว 2.5 ถึง 2.6 หน่วย ขอบเขตนี้ไม่ได้บอกว่ารอบแบบ V ลดค่าตกค้างได้เท่าใด เรื่องนั้นต้องดูตัวเลข

ตารางที่ 10.3 วัดตัวเลขบนกริดสามขนาดในโค้ดของเรา เพื่อดูอัตราต่อรอบและจำนวนงานจนถึงค่าตกค้างสัมพัทธ์ 10^{-6}

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 10.3: จำนวนรอบที่ทำให้ $\|\mathbf{r}_k\|/\|\mathbf{r}_0\| \le 10^{-6}$ สำหรับ $A_m\mathbf{x} = \mathbf{1}$ จาก $\mathbf{x}_0 = \mathbf{0}$ (ตัวเลขคำนวณตามโค้ดของเรา ไม่ใช่เวลาจริง งานนับเฉพาะรอบกวาด ไม่นับการกำกัดและการขยาย)

m	$n = m^2$	เกาส์–ไซเดล (รอบ)	รอบแบบ V (รอบ)	งานของรอบแบบ V (หน่วย)
7	49	88	6	14.2
15	225	354	6	15.1
31	961	1,414	7	18.1

ในทุกกรณีที่ทดสอบ ค่าตกค้างสัมพัทธ์ลดลงราว 0.03 ถึง 0.13 เท่าต่อรอบแบบ V และงานรวมโตช้ากว่าเกาส์–ไซเดลมาก ผลนี้เป็นการสังเกตจากตัวเลขสามขนาดเท่านั้น ยังไม่ใช่ทฤษฎีบท ทฤษฎีที่ให้อัตราลดคงที่ไม่ขึ้นกับ m อยู่ในเอกสารอ้างอิง Braess 2007, บทที่ V, p. 225 ซึ่งอธิบายว่าวิธีหลายกริดใช้งานเชิงเส้นในจำนวนไม่รู้ ดูเพิ่มบทที่ V ทฤษฎีบท V.3.5 ของ Braess และ Iserles 2009, §13.2--13.3, pp. 298--306 ให้รายละเอียดของการวนซ้ำ ดูเพิ่มบทที่ 13 นิยาม 13.3 ของ Iserles เกี่ยวกับการเริ่มต้นแบบหลายกริดเต็มรูปแบบ

วิธีหลายกริดเต็มรูปและการเริ่มต้นที่ดี

วิธีหลายกริดเต็มรูป (full multigrid) เริ่มจากกริดหยาบที่สุดแล้วขยายขึ้นทีละชั้น โดยทำรอบแบบ V บนแต่ละชั้น ค่าเริ่มต้นที่ได้จึงมีส่วนเรียบถูกแก้ไปแล้ว ISE ประมาณว่าช่วงแรกนี้ใช้งานราวหนึ่งในสามของรอบแบบ V Iserles 2009, §13.3, p. 306 เราไม่ได้วัดตัวเลขนี้เอง จึงรายงานเป็นของ ISE

สัญญาณล่วงหน้า: เชบีเชฟ

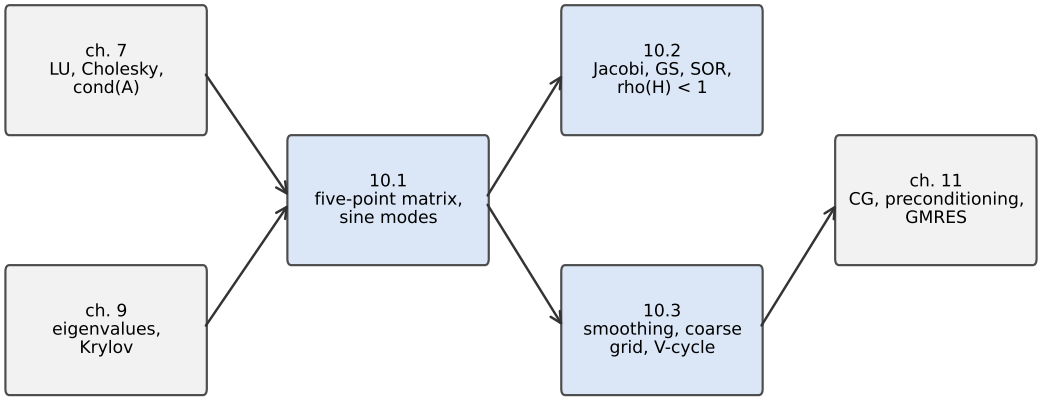
ในหัวข้อ 10.2 เราใช้รัศมีสเปกตรัมวัดอัตรา ถ้ารู้ช่วงของค่าลักษณะเฉพาะ เราเลือกการวนซ้ำที่ดีกว่าได้ เพราะค่าตกค้างหลัง k รอบเป็นพหุนามของ A คูณกับค่าเริ่มต้น Vorst 2003, §3.1, pp. 21--22 ดูเพิ่มบทที่ 3 ข้อเสนอ 3.1 ของ van der Vorst เช่น การวนซ้ำแบบ Richardson ให้ $\mathbf{r}_{i+1} = (I - A)\mathbf{r}_i$ ดังนั้น $\mathbf{r}_k = (I - A)^k\mathbf{r}_0$ ถ้าเลือกพหุนาม p ที่มี $p(0) = 1$ และมีขนาดเล็กที่สุดบนช่วงค่าลักษณะเฉพาะ จะได้**วิธีทำซ้ำเชบีเชฟ** (Chebyshev iterative method) ซึ่งใช้พหุนามเชบีเชฟ (ดูบทที่ 3 สำหรับพหุนามนี้) เราไม่พิสูจน์ในหัวข้อนี้ บทที่ 11 จะใช้ความคิดเดียวกันกับตัวแก้แบบคอนจูเกตกราเดียนต์

สรุปใจความ

- ตัวปรับเรียบอย่างจาโคบีแบบถ่วง $\omega = \frac{4}{5}$ ลดโหมดความถี่สูงทุกโหมดให้ไม่เกิน $\frac{3}{5}$ เท่าในหนึ่งรอบ (ทฤษฎีบท 10.3.1) ส่วนโหมดความถี่ต่ำแทบไม่หาย
- กริดหยาบด้วย $R = \frac{1}{4}P^{\top}$ และ $A_c = RA_mP$ ทำให้ค่าตกค้างที่กริดหยาบเห็นเป็นศูนย์หลังการแก้ไข (ทฤษฎีบท 10.3.2)
- รอบแบบ V ใช้งานไม่เกิน $\frac{4}{3}(1 + \frac{1}{m})^2\nu$ หน่วย ซึ่งเป็นเชิงเส้นในจำนวนจุด (ทฤษฎีบท 10.3.3)
- ตัวเลขของสระ 256×256 ไม่ใช่รูป $2^s - 1$ จึงซ้อนกริดไม่พอดี แบบฝึกหัด 10.3.6 ให้ดูทางแก้

อ่านต่อ. Iserles 2009, §13.1--13.3, pp. 291--308 ให้ทฤษฎีของวิธีหลายกริดละเอียดกว่านี้ Braess 2007, pp. 225--236 ให้ประวัติของวิธีหลายกริดและทฤษฎีการลู่เข้าสำหรับปัญหาเชิงแปรผัน และ Vorst 2003, §3.1, pp. 21--22 ให้กรอบของการวนซ้ำที่มี K ซึ่งเป็นตัวปรับสภาพเงื่อนไขของบทที่ 11

ฉบับร่าง preprint · 2026-10-23



รูปที่ 10.8: แผนผังบทที่ 10 ทั้งบท: หัวข้อ 10.1 ถึง 10.3 และสิ่งที่บทที่ 7 กับ 9 ให้บทนี้ เส้นจากหัวข้อ 10.3 ไปบทที่ 11 คือสิ่งที่บทนี้ให้ต่อ (สร้างจาก `numanal/ch10_mg.py`)

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 10.3.1 (⋆, ทำด้วยมือ) ใช้กริด 3×3 และ P, $R = \frac{1}{4}P^{\top}$, $A_c = \frac{3}{4}$ ของตัวอย่างด้วยมือ ให้ความคลาดเคลื่อน $\mathbf{e}$ มีค่า 1 ที่จุดกลาง $(2,2)$ และ 0 ที่จุดอื่น

1. จงคำนวณ $R\mathbf{r}$ เมื่อ $\mathbf{r} = -A_3\mathbf{e}$
2. จงหา $\mathbf{e}_c$ และ $\mathbf{e}^{+} = \mathbf{e} + P\mathbf{e}_c$
3. จงตรวจว่า $R(A_3\mathbf{e}^{+}) = \mathbf{0}$ และค่าสัมบูรณ์สูงสุดของความคลาดเคลื่อนลดจากเท่าใดเหลือเท่าใด

แบบฝึกหัด 10.3.2 (⋆⋆, พิสูจน์) ให้ A เป็นเมทริกซ์สมมาตรบวกแน่นอนขนาด $m^2 \times m^2$ ให้ P เป็นเมทริกซ์ที่มีอันดับเต็มคอลัมน์ และให้ $R = \frac{1}{4}P^{\top}$ กับ $A_c = RAP$

1. จงแสดงว่า A_c สมมาตรบวกแน่นอน
2. ให้ $\mathbf{e} = P\mathbf{w}$ สำหรับเวกเตอร์ $\mathbf{w}$ ใด ๆ จงแสดงว่าการแก้ไขบนกริดหยาบหนึ่งครั้ง (โดยไม่มีการปรับเรียบ) ให้ $\mathbf{e}^{+} = \mathbf{0}$ เมื่อ $\mathbf{e}^{+} = \mathbf{e} + P\mathbf{e}_c$ และ $A_c\mathbf{e}_c = R(-A\mathbf{e})$

แบบฝึกหัด 10.3.3 (⋆⋆, คำนวณ) ให้ A_{15} และ $\sigma_{p,q}$ ของทฤษฎีบท 10.3.1 สำหรับ $\omega = \frac{4}{5}$ และ $\omega = 1$ (เมื่อ $\omega = 1$ ใช้ $\sigma_{p,q} = \frac{1}{2}(\cos\theta_p + \cos\theta_q)$)

1. สำหรับ $\omega = \frac{4}{5}$ จงหาค่าสัมบูรณ์สูงสุดของ $\sigma_{p,q}$ บนโหมดความถี่สูงทั้งหมด ($\max(\theta_p, \theta_q) \geq \frac{\pi}{2}$) และระบุ (p,q) ที่ได้
2. สำหรับ $\omega = \frac{4}{5}$ จงหา $\sigma_{1,1}$ แล้วอธิบายว่าทำไมโหมดนี้ไม่ถูกลดในรอบเดียว
3. สำหรับ $\omega = 1$ จงหาค่าสัมบูรณ์สูงสุดของ $\sigma_{p,q}$ บนโหมดความถี่สูง แล้วเทียบกับข้อ 1

ฉบับร่าง Preprint · 2026-10-23

แบบฝึกหัด 10.3.4 (⋆⋆, คำนวณและพิสูจน์) ให้ $W(m)$ เป็นงานของรอบแบบ V ที่ $\nu_1 = \nu_2 = 1$ บนกริด $m = 2^s - 1$ ตามทฤษฎีบท 10.3.3

1. จงหา $W(m)$ สำหรับ $m = 3, 7, 15, 31$ แล้วเทียบกับขอบบน $\frac{4}{3}(1 + \frac{1}{m})^2 \cdot 2$
2. จงแสดงว่า $W(m) \to \frac{8}{3}$ เมื่อ $m \to \infty$ โดยใช้ว่า $m_l/m \to 2^{-l}$ และ $\sum_{l \ge 0} 4^{-l} = \frac{4}{3}$

แบบฝึกหัด 10.3.5 (⋆⋆, คำนวณและตีความ) สำหรับ $A_{15}\mathbf{x} = \mathbf{1}$ เริ่มจาก $\mathbf{x}_0 = \mathbf{0}$ และใช้เกณฑ์ $\|\mathbf{r}_k\|_2 / \|\mathbf{r}_0\|_2 \le 10^{-6}$

1. จงหาจำนวนรอบเกาส์–ไซเดลที่ทำให้เกณฑ์นี้เป็นจริง
2. จงหาจำนวนรอบแบบ V ($\nu_1 = \nu_2 = 1$) ที่ทำให้เกณฑ์นี้เป็นจริง และงานรวมเป็นหน่วย
3. จงเทียบกับ 357 รอบของแบบฝึกหัด 10.2.6 ซึ่งเป็นสูตรปิดของความคลาดเคลื่อนของเกาส์–ไซเดลบน A_{15} แล้วอธิบายว่าเหตุใดตัวเลขทั้งสองจึงใกล้กัน

แบบฝึกหัด 10.3.6 (⋆ ⋆ ⋆, โครงงานเล็ก) สระในสถานการณ์สมมติมี 256×256 จุดภายใน ซึ่งไม่เป็นรูป $2^s - 1$ จึงซ้อนกริดไม่พอดี

1. จงหาจำนวนรอบเกาส์–ไซเดลที่ทำให้ $\rho^k \le 10^{-6}$ สำหรับ $m = 255$ และ $m = 511$ โดยใช้ $\rho = \cos^2 \frac{\pi}{m+1}$ แล้วเทียบกับ 92,454 รอบของ $m = 256$
2. จงระบุลำดับของกริดที่ซ้อนกันสำหรับ $m = 255$ และบอกจำนวนชั้น
3. สมมติว่ารอบแบบ V ลดค่าตกค้างราว 10 เท่าต่อรอบ (ค่าที่สังเกตได้ในหัวข้อนี้ ไม่ใช่ทฤษฎี) จงหาจำนวนรอบแบบ V ที่ทำให้ค่าตกค้างต่ำกว่า 10^{-6} และงานรวมโดยทฤษฎีบท 10.3.3 ($\nu = 2$) แล้วเทียบกับข้อ 1

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 10.3.1. ตามตัวอย่างด้วยมือ $\mathbf{p}$ คือ P ของกริด 3×3 และ $R = \frac{1}{4}\mathbf{p}^{\top}$ กับ $A_c = \frac{3}{4}$

1. $A_3\mathbf{e}$ คือคอลัมน์ของจุดกลาง คือ 4 ที่ $(2,2)$, -1 ที่จุดขอบทั้งสี่ และ 0 ที่มุม ดังนั้น $\mathbf{r} = -A_3\mathbf{e}$ มี -4 ที่ $(2,2)$, $+1$ ที่ขอบ และ 0 ที่มุม จึง
$$R\mathbf{r} = \frac{1}{4}\mathbf{p}^{\top}\mathbf{r} = \frac{1}{4}\Big(1 \cdot (-4) + 4 \cdot \tfrac{1}{2} \cdot 1\Big) = \frac{1}{4}(-2) = -\frac{1}{2}$$
2. จาก $A_c\mathbf{e}_c = R\mathbf{r}$ ได้ $\frac{3}{4}\mathbf{e}_c = -\frac{1}{2}$ จึง $\mathbf{e}_c = -\frac{2}{3}$ แล้ว $\mathbf{e}^{+} = \mathbf{e} + \mathbf{p}\,\mathbf{e}_c$ มีค่า $1 - \frac{2}{3} = \frac{1}{3}$ ที่จุดกลาง $-\frac{1}{3}$ ที่ขอบทั้งสี่ และ $-\frac{1}{6}$ ที่มุมทั้งสี่
3. คำนวณ $A_3\mathbf{e}^{+}$: ที่จุดกลางได้ $4 \cdot \frac{1}{3} - 4 \cdot (-\frac{1}{3}) = \frac{8}{3}$; ที่ขอบ $(1,2)$ ได้ $4 \cdot (-\frac{1}{3}) - \big((-\frac{1}{6}) + (-\frac{1}{6}) + \frac{1}{3}\big) = -\frac{4}{3}$; ที่มุม $(1,1)$ ได้ $4 \cdot (-\frac{1}{6}) - \big((-\frac{1}{3}) + (-\frac{1}{3})\big) = 0$ ดังนั้น
$$R(A_3\mathbf{e}^{+}) = \frac{1}{16}\big(4 \cdot \tfrac{8}{3} + 2 \cdot 4 \cdot (-\tfrac{4}{3}) + 4 \cdot 0\big) = \frac{1}{16}\big(\tfrac{32}{3} - \tfrac{32}{3}\big) = 0$$
ค่าสัมบูรณ์สูงสุดของความคลาดเคลื่อนลดจาก 1 เหลือ $\frac{1}{3}$ ตามที่ทฤษฎีบท 10.3.2 ข้อ 3 ทำนาย

ตรวจด้วยโค้ด: `test_ch10_mg.py` (`test_exercise_1031_centre_error`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 10.3.2.

1. $A_c^{\top} = \frac{1}{4}P^{\top}A^{\top}P = \frac{1}{4}P^{\top}AP = A_c$ จึงสมมาตร ถ้า $\mathbf{w} \neq \mathbf{0}$ แล้ว $P\mathbf{w} \neq \mathbf{0}$ เพราะ P มีอันดับเต็มคอลัมน์ จึง
$$\mathbf{w}^{\top}A_c\mathbf{w} = \tfrac{1}{4}(P\mathbf{w})^{\top}A(P\mathbf{w}) > 0$$
เพราะ A บวกแน่นอน ดังนั้น A_c บวกแน่นอน

2. ให้ $\mathbf{r} = -A\mathbf{e} = -AP\mathbf{w}$ จะได้ $R\mathbf{r} = -RAP\mathbf{w} = -A_c\mathbf{w}$ แล้ว $\mathbf{e}_c = A_c^{-1}(-A_c\mathbf{w}) = -\mathbf{w}$ (ใช้ข้อ 1 ว่า A_c ผกผันได้) ดังนั้น
$$\mathbf{e}^{+} = P\mathbf{w} + P\mathbf{e}_c = P\mathbf{w} - P\mathbf{w} = \mathbf{0}$$

ตรวจด้วยโค้ด: `test_ch10_mg.py` (`test_exercise_1032_coarse_step_exact_on_range_of_p`)

เฉลย 10.3.3. สำหรับ $m = 15$ ค่า $\cos\frac{\pi}{16} \approx 0.980785$

1. โหมดความถี่สูงที่ให้ค่าสัมบูรณ์สูงสุดคือ $(p,q) = (8,1)$ และ $(1,8)$ เพราะ $\cos\theta_8 = \cos\frac{\pi}{2} = 0$ จึง
$$\sigma_{8,1} = \tfrac{1}{5} + \tfrac{2}{5}\cos\tfrac{\pi}{16} \approx 0.5923$$
ส่วนโหมดความถี่สูงอื่นทั้งหมดได้ค่าสัมบูรณ์ไม่เกินนี้ ซึ่งเราตรวจด้วยการนับทุกคู่ (p,q) ทฤษฎีบท 10.3.1 ให้ขอบ $\frac{3}{5}$ และค่านี้เข้าใกล้ $\frac{3}{5}$ เมื่อ m ใหญ่

2. $\sigma_{1,1} = \frac{1}{5} + \frac{2}{5}\cdot 2\cos\frac{\pi}{16} \approx 0.9846$ โหมดนี้ไม่ถูกลดในรอบเดียวเพราะ θ_1 เล็กมาก จึง $\cos\theta_1$ ใกล้ 1 และ σ ใกล้ 1 มาก การลดต่อรอบของโหมดนี้จึงแทบไม่มี

3. เมื่อ $\omega = 1$ ค่าสูงสุดของ $|\sigma_{p,q}|$ บนโหมดความถี่สูงอยู่ที่ $(15,15)$ ได้ $|\sigma| = \cos\frac{\pi}{16} \approx 0.9808$ ซึ่งมากกว่าค่า 0.5923 ของ $\omega = \frac{4}{5}$ มาก จึงยืนยันว่าจาโคบีธรรมดาไม่ใช่ตัวปรับเรียบที่ดี

ตรวจด้วยโค้ด: `test_ch10_mg.py` (`test_ex1033_low_and_high_for_omega_08_and_one`)

เฉลย 10.3.4.

1. จาก $W(m) = 2\sum_l n_l/n_0$ โดยรวมเฉพาะชั้นที่ถูกปรับเรียบ ได้
   - $m = 3$: ชั้นเดียว $W = 2\cdot\frac{9}{9} = 2.000$ ขอบบน $\frac{4}{3}\cdot\frac{16}{9}\cdot 2 = \frac{128}{27} \approx 4.74$
   - $m = 7$: ชั้น $7,3$ $W = 2\cdot\frac{49+9}{49} = \frac{116}{49} \approx 2.367$ ขอบบน $\frac{512}{147} \approx 3.48$
   - $m = 15$: ชั้น $15,7,3$ $W = 2\cdot\frac{225+49+9}{225} = \frac{566}{225} \approx 2.516$ ขอบบน $\frac{2048}{675} \approx 3.03$
   - $m = 31$: ชั้น $31,15,7,3$ $W = 2\cdot\frac{961+225+49+9}{961} = \frac{2488}{961} \approx 2.589$ ขอบบน $\frac{8192}{2883} \approx 2.84$

   ค่าทุกตัวอยู่ใต้ขอบบน

2. สำหรับ $m = 2^s - 1$ และ l คงที่ จะได้ $m_l/m = (2^{s-l}-1)/(2^s-1) \to 2^{-l}$ จึง $(m_l/m)^2 \to 4^{-l}$ และ $(m_l/m)^2 \le 4^{-l}(1+\frac{1}{m})^2$ ซึ่งเป็นตัวควบคุมที่ลู่เข้า (ใช้เกณฑ์ของ Weierstrass) ดังนั้น
$$W(m) = 2\sum_{l=0}^{s-2}\Big(\frac{m_l}{m}\Big)^2 \longrightarrow 2\sum_{l\ge 0}4^{-l} = 2\cdot\frac{4}{3} = \frac{8}{3}$$

ตรวจด้วยโค้ด: `test_ch10_mg.py` (`test_ex1034_work_limit_and_bound_m3_m31`)

สมบัติร่าง · Preprint · 2026-10-23

เฉลย 10.3.5.

1. เกาส์–ไซเดลบน A_{15} จาก $\mathbf{x}_0 = \mathbf{0}$ ให้ $\|\mathbf{r}_k\|_2/\|\mathbf{r}_0\|_2 \approx 9.66 \times 10^{-7}$ ที่ $k = 354$ และยังมากกว่า 10^{-6} ที่ $k = 353$ จึงต้องใช้ 354 รอบ

2. รอบแบบ V ($\nu_1 = \nu_2 = 1$) ให้ตัวประกอบประมาณ 0.1 ต่อรอบ จึงต้องใช้ 6 รอบ (ค่าตกค้างหลังรอบที่ 6 ต่ำกว่า 10^{-6}) งานรวมคือ $6 \cdot W(15) = 6 \cdot \frac{566}{225} \approx 15.1$ หน่วย

3. สูตรปิดของแบบฝึกหัด 10.2.6 ให้ 357 รอบสำหรับความคลาดเคลื่อน ซึ่งใกล้กับ 354 มาก ทั้งสองวัดอัตราของโหมดที่ช้าที่สุด ซึ่งครอบงำในระยะยาว ความต่างราว 3 รอบมาจากค่าคงที่หน้าเลขยกกำลังที่ขึ้นกับจุดเริ่มต้นและจากการวัดค่าตกค้างแทนความคลาดเคลื่อน เมื่อเทียบงาน เกาส์–ไซเดลใช้ราว $354/15.1 \approx 23$ เท่าของรอบแบบ V

ตรวจด้วยโค้ด: `test_ch10_mg.py` (`test_ex1035_compare_gs_and_vcycle_m15`)

เฉลย 10.3.6.

1. ใช้ $\rho = \cos^2 \frac{\pi}{m+1}$ และ $k = \lceil \ln 10^{-6} / \ln \rho \rceil$ ได้ $m = 255$: $\rho \approx 0.999\,851$ และ $k = 91{,}736$; $m = 511$: $k = 366{,}948$ จะเห็นว่า $m = 255$ ให้ค่าใกล้ 92,454 ของ $m = 256$ (ต่างกันตามอัตราส่วน $(256/257)^2 \approx 0.992$) และเมื่อ m เพิ่มเป็นสองเท่า จำนวนรอบเพิ่มราว 4 เท่า

2. สระ 256 ช่องที่มีกำแพงทั้งสองด้านมีจุดภายในต่อด้าน $255 = 2^8 - 1$ ซึ่งซ้อนกริดได้พอดี ลำดับของกริดคือ $255, 127, 63, 31, 15, 7, 3, 1$ รวม 8 ชั้น โดยเจ็ดชั้นแรกถูกปรับเรียบ และชั้นสุดท้ายมีจุดเดียวที่แก้ตรงได้

3. หากรอบแบบ V ลดค่าตกค้างราว 10 เท่าต่อรอบ จะต้อง 6 รอบ เพราะ 10^{-6} (ค่าที่สังเกตได้ในหัวข้อนี้ ไม่ใช่ทฤษฎี) งานรวมโดยทฤษฎีบท 10.3.3 คือ $6 \cdot W(255)$ โดย $W(255) \approx 2.656$ (คำนวณด้วยโค้ด) จึงได้ประมาณ 15.9 หน่วย เทียบกับ $91{,}736$ รอบของเกาส์–ไซเดลในข้อ 1 ซึ่งมากกว่าราว 5.8×10^3 เท่า
เราไม่ได้รันรอบแบบ V บน $m = 255$ เพราะเมทริกซ์ตามกริดนี้มี $65{,}025$ แถว จึงเกินกว่าที่โค้ดในบทนี้ใช้แบบเมทริกซ์เต็มได้ ข้อ 3 จึงเป็นการคาดการณ์ที่มีเงื่อนไข

ตรวจด้วยโค้ด: `test_ch10_mg.py` (`test_ex1036_sweeps_for_larger_pools`)

ฉบับร่าง · Preprint · 2026-10-23

สรุปบทที่ 10 ในหน้าเดียว

ตารางนี้เทียบวิธีหลัก โดย $n = m^2$ คือจำนวนค่าที่ไม่รู้

วิธี	ต้องการอะไร	ค่าใช้จ่าย	หมายเหตุ
เกาส์–ไซเดล (10.2)	เฉพาะค่าที่ไม่เป็นศูนย์	ต่อรอบ $\propto$ จำนวนค่าที่ไม่เป็นศูนย์	$\rho_{\mathrm{GS}} = \cos^2 \frac{\pi}{m+1}$
SOR ที่ ω_{opt} (10.2)	เหมือนเกาส์–ไซเดล	ต่อรอบเท่าเกาส์–ไซเดล	$\rho = \omega_{\mathrm{opt}} - 1$
รอบแบบ V (10.3)	กริดหยาบ R, P, A_c	ไม่เกิน $\frac{4}{3}(1 + \frac{1}{m})^2 \nu$	ราว 0.1 ต่อรอบ (ที่สังเกต)

สูตรที่ควรจำ

- คอนดิชันนัมเบอร์ของ A_m: $\kappa(A_m) = \frac{1+\cos\frac{\pi}{m+1}}{1-\cos\frac{\pi}{m+1}}$ (หัวข้อ 10.1)

อ่านต่อ. บทที่ 11 *ลงหุบเขาให้ฉลาดขึ้น* ใช้ปริภูมิไครลอฟที่เห็นในหัวข้อ 9.4 แก้ระบบสมมาตรบวกแน่นอนด้วยคอนจูเกตกราเดียนต์ และใช้ตัวปรับสภาพเงื่อนไข (preconditioner) ซึ่งมาจากแนวคิดของการวนซ้ำแบบมี K ในหัวข้อ 10.3 Vorst 2003, §3.1, pp. 21--22

บทที่ 11

ลงหุบเขาให้ฉลาดขึ้น

สถานการณ์: ลบรอยขีดข่วนบนรูปถ่าย

ลองนึกถึงรูปถ่ายเก่าที่มีรอยขีดข่วนพาดกลางภาพ ถ้ามีรูปอีกใบของฉากเดียวกันที่ไม่มีรอย แต่แสงในรูปนั้นต่างกันเล็กน้อย เราจะเอาเนื้อภาพจากรูปที่สองมาแทนรอยได้อย่างไรโดยไม่เห็นขอบ

การตัดแล้ววางตรง ๆ มักเห็นเส้นขอบ เพราะความสว่างสองฝั่งไม่เท่ากัน ในปี ค.ศ. 2003 เปเรซ เกนเนต และเบลก เสนอวิธีที่ให้เนื้อภาพในบริเวณที่เลือกเป็นคำตอบของสมการปัวซง (Poisson equation) โดยมีค่าที่ขอบมาจากภาพรอบนอก Pérez, Gangnet, and Blake 2003 เมื่อแทนอนุพันธ์ด้วยผลต่างบนพิกเซล จะได้ระบบสมการเชิงเส้นขนาดใหญ่ ซึ่งเอกสารระบุว่าเป็นระบบที่เบาบาง สมมาตร และบวกแน่นอน ผู้เขียนแก้ระบบนี้ด้วยการวนซ้ำแบบเกาส์–ไซเดลกับ SOR หรือวิธีหลายกริด ซึ่งเราเห็นในบทที่ 10 Pérez, Gangnet, and Blake 2003, §2, p. 315

คำถามของบทนี้คือ ระบบแบบนี้แก้ด้วยวิธีอื่นได้หรือไม่ และทำไมวิธีที่ดูฉลาดบางวิธีกลับเดินซิกแซกช้า ๆ ตัวอย่างต่อไปนี้เป็นรูปสังเคราะห์ของเราเอง (สถานการณ์สมมติ) ไม่ใช่รูปจริง

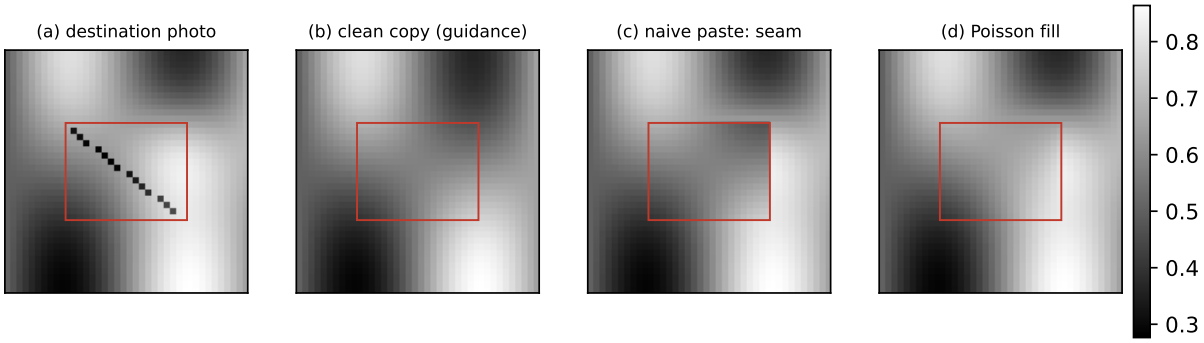
ภาพรวมบท

บทนี้มีสี่ความคิด

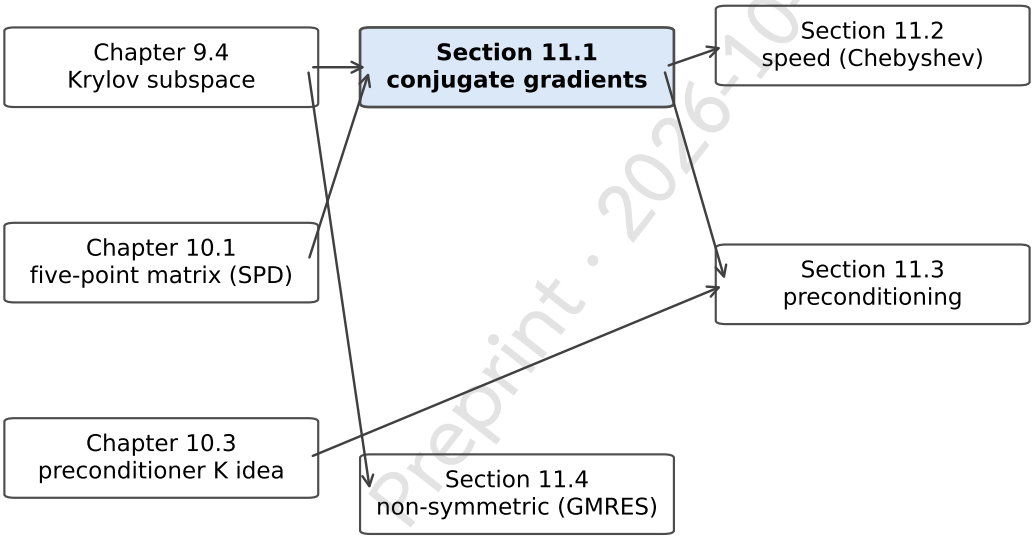
- ถ้า A สมมาตรและบวกแน่นอน การแก้ระบบ $A\mathbf{x} = \mathbf{b}$ เท่ากับการหาจุดต่ำสุดของฟังก์ชันรูปหุบเขา ซึ่งเป็นเรื่องของหัวข้อ 11.1
- วิธีที่เดินลงชันที่สุดจะซิกแซก ส่วน**วิธีเกรเดียนต์สังยุค** (conjugate gradients) เลือกทิศใหม่ที่ไม่ย้อนกลับทิศเดิม และในเลขคณิตแม่นตรงจะจบภายในจำนวนรอบเท่ากับมิติของระบบ
- อัตราที่วิธีเหล่านี้ลู่เข้าขึ้นกับคอนดิชันนัมเบอร์ของระบบ ซึ่งบอกว่าระบบไวต่อการเปลี่ยนแปลงแค่ไหน หัวข้อ 11.2 จะวัดอัตรานี้ ส่วนหัวข้อ 11.3 จะปรับระบบให้ง่ายขึ้นก่อนแก้
- ถ้า A ไม่สมมาตร CG ใช้ไม่ได้โดยตรง หัวข้อ 11.4 ใช้**วิธีค่าตกค้างน้อยสุดนัยทั่วไป** (generalized minimal residual method (GMRES)) ซึ่งเลือกคำตอบที่ค่าตกค้างเล็กที่สุดในปริภูมิไครลอฟ และต้องรีสตาร์ตเมื่อความจำไม่พอ

เริ่มจากรูปถ่ายที่มีรอยขีดข่วน แล้วจะเจออีกด้านหนึ่งของปัญหาเดียวกัน คือผ้าในเกมที่ต้องแก้ระบบหนึ่งครั้งทุกเฟรม (หัวข้อ 11.3)

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 11.1: รูปสังเคราะห์ของเรา 40×40 พิกเซล (ก) รูปปลายทางพร้อมรอยขีดสีดำ กรอบแดงคือบริเวณที่จะเติม ขนาด 16×20 พิกเซล (ข) ภาพต้นแบบที่สะอาดซึ่งเป็นฉากเดียวกันแต่ไม่มีรอย (ค) การวางภาพต้นแบบลงตรง ๆ เห็นรอยต่อชัด (ง) การเติมด้วยสมการปัวซง ค่าต่างสูงสุดที่รอยต่อลดจาก 0.2423 เหลือ 0.0311 (สร้างจาก `numanal/ch11.py`)



ฉบับร่าง · preprint · 2026-10-23

รูปที่ 11.2: แผนผังบทที่ 11 กล่องสีฟ้าคือหัวข้อ 11.1 ลูกศรบอกแนวคิดที่แต่ละหัวข้อใช้ เช่น วิธีเกรเดียนต์สังยุคใช้ปริภูมิย่อยไครลอฟจากหัวข้อ 9.4 และเมทริกซ์ห้าจุดจากหัวข้อ 10.1 (สร้างจาก `numanal/ch11.py`)

11.1 ซิกแซกลงเขา

สถานการณ์สมมติ: นักเดินป่าในหมอก

ลองนึกถึงนักเดินป่าที่ลงเขาในหมอกหนา เขามองเห็นแค่พื้นใต้เท้า จึงเดินไปทางที่ลาดชันที่สุดตรงจุดนั้นทุกก้าว ฟังดูสมเหตุสมผล แต่ถ้าหุบเขายาวและแคบ ทางที่ชันที่สุดจะพาเขาข้ามหุบจากฝั่งหนึ่งไปอีกฝั่งสลับไปมา และได้ความคืบหน้าทีละนิดตามความยาวหุบ

ทั้งที่ก้นหุบอยู่ไม่ไกลนัก คำถามคือ ทำไมทางที่ชันที่สุดจึงซิกแซก และมีทางไหนที่พาถึงก้นหุบได้เร็วกว่า

ภาพรวมก่อนอ่าน

เราจะเปลี่ยนคำถามเรื่องระบบสมการให้เป็นคำถามเรื่องหุบเขา ถ้า A บวกแน่นอน ฟังก์ชัน $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top\mathbf{x}$ เป็นรูปชาม และก้นชามคือคำตอบของ $A\mathbf{x} = \mathbf{b}$ ทิศที่ชันที่สุดของชามคือค่าตกค้าง $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ วิธีลงชันที่สุดเดินตามทิศนี้ ส่วนวิธีเกรเดียนต์สังยุคเลือกทิศใหม่ที่ไม่ย้อนกลับไปทิศเดิม รูป 11.3 แสดงทั้งสองวิธีบนหุบเขาสองมิติ

ลองเอง: หนึ่งก้าวลงชันที่สุดก่อนดูสูตร

ให้ $A = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$, $\mathbf{b} = (1, 2)^\top$ และ $\mathbf{x}_0 = \mathbf{0}$ จากนิยาม $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top\mathbf{x}$ ได้ $f(\mathbf{x}_0) = 0$ และค่าตกค้าง $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0 = (1, 2)^\top$ ซึ่งเป็นทิศที่ชันที่สุดของชามที่จุดนี้

เดินตามเส้น $\mathbf{x}_0 + t\mathbf{r}_0$ แล้วเลือก t ให้ f ต่ำที่สุด ได้ $t = \dfrac{\mathbf{r}_0^\top\mathbf{r}_0}{\mathbf{r}_0^\top A\mathbf{r}_0} = \dfrac{5}{20} = \dfrac{1}{4}$ จึง $\mathbf{x}_1 = (\frac{1}{4}, \frac{1}{2})^\top$ และ

$$f(\mathbf{x}_1) = -\tfrac{5}{8}.$$

คำตอบแม่นตรงคือ $\mathbf{x}^* = (\frac{1}{11}, \frac{7}{11})^\top$ ซึ่งให้ $f(\mathbf{x}^*) = -\frac{15}{22}$ ต่ำกว่า $f(\mathbf{x}_1)$ ผลต่างคือ $\frac{5}{88} = \frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}^*\|_A^2$ ตามทฤษฎีบท 11.1.1

จากระบบเชิงเส้นสู่หุบเขา

ให้ A เป็นเมทริกซ์จริงขนาด $n \times n$ ที่สมมาตร คือ $A^\top = A$ และให้ $\mathbf{b} \in \mathbb{R}^n$ เราจะเรียกฟังก์ชันที่ต่อไปนี้ว่า**ฟังก์ชันกำลังสอง** (quadratic function)ของระบบ

นิยาม 11.1.1: ฟังก์ชันกำลังสองของระบบ $A\mathbf{x} = \mathbf{b}$

ให้ A สมมาตรและ $\mathbf{b} \in \mathbb{R}^n$ นิยาม

$$f(\mathbf{x}) = \tfrac{1}{2}\,\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top\mathbf{x}, \qquad \mathbf{x} \in \mathbb{R}^n.$$

ค่า $\mathbf{r}(\mathbf{x}) = \mathbf{b} - A\mathbf{x}$ เรียกว่า **ค่าตกค้าง** (residual) ที่จุด $\mathbf{x}$ และถ้า A บวกแน่นอน นิยาม **นอร์มพลังงาน** (energy norm) ของเวกเตอร์ $\mathbf{v}$ เป็น $\|\mathbf{v}\|_A = \sqrt{\mathbf{v}^\top A\mathbf{v}}$

อ่านแบบเข้าใจง่าย

ลองดูกรณี $n = 1$ เมื่อ $A = a > 0$ จะได้ $f(x) = \frac{1}{2}ax^2 - bx$ ซึ่งเป็นพาราโบลาคว่ำขึ้น มีจุดต่ำสุดที่ $x = b/a$ และนั่นคือคำตอบของ $ax = b$ ส่วนค่าตกค้าง $r = b - ax$ เป็นความชันของพาราโบลาที่ x ด้วยเครื่องหมายลบ ถ้า x อยู่ทางซ้ายของจุดต่ำสุด r เป็นบวกและชี้ไปทางขวา ซึ่งเป็นทิศที่ f ลดลง

ผลต่อไปนี้บอกว่า การแก้ระบบกับการหาจุดต่ำสุดของชามเป็นปัญหาเดียวกันจริง ๆ

ทฤษฎีบท 11.1.1: ผลเฉลยเป็นจุดต่ำสุดของชาม

ให้ A สมมาตรและบวกแน่นอน และให้ $\mathbf{x}^* = A^{-1}\mathbf{b}$ แล้วสำหรับทุก $\mathbf{x} \in \mathbb{R}^n$

$$f(\mathbf{x}) = f(\mathbf{x}^*) + \tfrac{1}{2}\,\|\mathbf{x} - \mathbf{x}^*\|_A^2.$$

ฉบับร่าง · Preprint · 2026-10-23

ดังนั้น $\mathbf{x}^*$ เป็นจุดต่ำสุดเพียงจุดเดียวของ f

พิสูจน์. **แนวคิด.** เติมกำลังสองให้สมบูรณ์รอบ $\mathbf{x}^*$ แล้วใช้ว่า A บวกแน่นอน

1. **ใช้ $A\mathbf{x}^* = \mathbf{b}$.** เพราะ A สมมาตร $\mathbf{x}^\top A\mathbf{x}^* = \mathbf{x}^\top \mathbf{b}$ ดังนั้น
$$\tfrac{1}{2}(\mathbf{x}-\mathbf{x}^*)^\top A(\mathbf{x}-\mathbf{x}^*) = \tfrac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{x}^\top\mathbf{b} + \tfrac{1}{2}\mathbf{x}^{*\top}A\mathbf{x}^* = f(\mathbf{x}) + \tfrac{1}{2}\mathbf{x}^{*\top}A\mathbf{x}^*.$$
2. **ค่าที่จุด $\mathbf{x}^*$.** จากขั้นที่ 1 เมื่อ $\mathbf{x} = \mathbf{x}^*$ ได้ $f(\mathbf{x}^*) = -\tfrac{1}{2}\mathbf{x}^{*\top}A\mathbf{x}^*$ เมื่อนำมาลบจากขั้นที่ 1 จะได้ $f(\mathbf{x}) - f(\mathbf{x}^*) = \tfrac{1}{2}\|\mathbf{x}-\mathbf{x}^*\|_A^2$ ซึ่งเป็นสูตรที่ต้องการ
3. **จุดต่ำสุดเพียงจุดเดียว.** เพราะ A บวกแน่นอน $\|\mathbf{v}\|_A^2 > 0$ เมื่อ $\mathbf{v} \neq \mathbf{0}$ ดังนั้น $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ และเท่ากันก็ต่อเมื่อ $\mathbf{x} = \mathbf{x}^*$

สรุป: ค่า f ที่สูงกว่าค่าต่ำสุดเท่ากับครึ่งหนึ่งของกำลังสองของระยะในนอร์มพลังงาน □

ผลนี้บอกว่า การลดค่า f คือการลดระยะห่างจากคำตอบในนอร์มพลังงาน ต่อไปเราจะดูว่าการลดนี้เดินอย่างไร

ลงทิศชันที่สุดและการซิกแซก

ความคิดแรกที่นึกได้คือ เดินไปตามทิศที่ f ลดเร็วที่สุด จากสูตรของ f จะได้ว่า $\nabla f(\mathbf{x}) = A\mathbf{x} - \mathbf{b} = -\mathbf{r}$ ดังนั้นทิศที่ชันที่สุดคือทิศของค่าตกค้าง เมื่อเดินไปตามเส้น $\mathbf{x} + t\mathbf{r}$ จะได้

$$f(\mathbf{x} + t\mathbf{r}) = f(\mathbf{x}) - t\,\mathbf{r}^\top\mathbf{r} + \tfrac{1}{2}t^2\,\mathbf{r}^\top A\mathbf{r},$$

ซึ่งเป็นพาราโบลาในตัวแปร t เราเลือกระยะก้าวที่ทำให้ f ต่ำที่สุดบนเส้นนั้น นี่คือ**การค้นหาตามเส้น** (line search)แบบแม่นตรง

นิยาม 11.1.2: วิธีลงชันที่สุด

เริ่มจาก $\mathbf{x}_0$ แล้วทำซ้ำสำหรับ $k = 0, 1, 2, \ldots$

1. คำนวณค่าตกค้าง $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$;
2. ตั้งระยะก้าว $\alpha_k = \dfrac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{r}_k^\top A\mathbf{r}_k}$ ซึ่งทำให้ f ต่ำที่สุดบนเส้น $\mathbf{x}_k + t\mathbf{r}_k$;
3. ปรับตำแหน่ง $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{r}_k$.

อ่านแบบเข้าใจง่าย

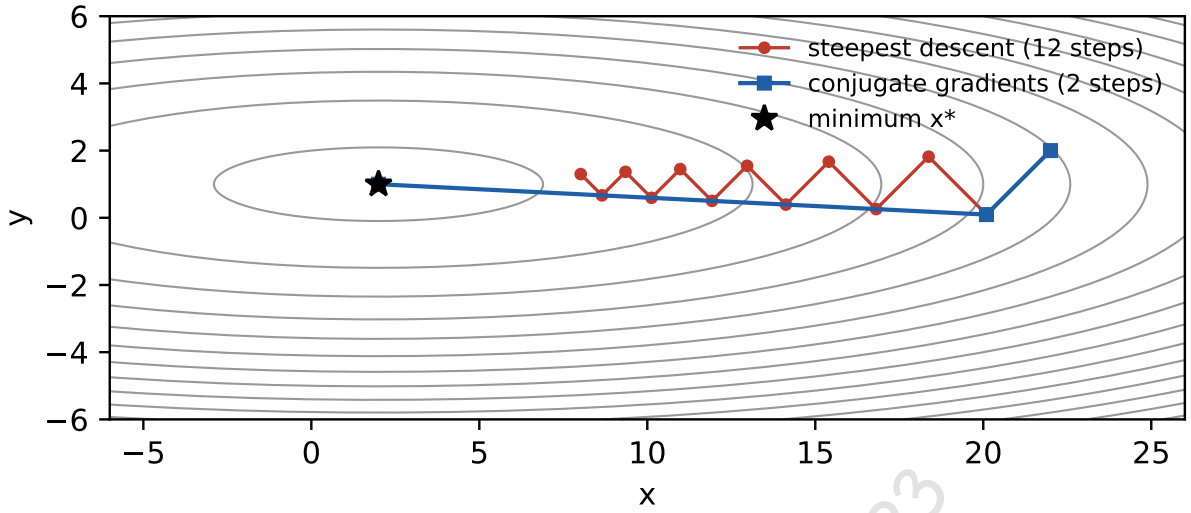
ทุกก้าวของวิธีนี้ลดค่า f ลงจริง (ดูแบบฝึกหัด 11.1.4) แต่ไม่ได้บอกว่าลดเร็วพอ ลองดูว่าก้าวต่อกันเป็นอย่างไร

ทฤษฎีบท 11.1.2: ก้าวติดกันตั้งฉากกันเสมอ

ให้ $\mathbf{r}_k$ และ $\mathbf{r}_{k+1}$ เป็นค่าตกค้างของวิธีลงชันที่สุดสองค่าติดกันที่ $\mathbf{r}_k \neq \mathbf{0}$ แล้ว $\mathbf{r}_{k+1}^\top \mathbf{r}_k = 0$

พิสูจน์. **แนวคิด.** ระยะก้าว α_k ถูกเลือกให้ f ต่ำที่สุดตามทิศ $\mathbf{r}_k$ และนั่นบังคับให้ทิศถัดไปตั้งฉาก

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 11.3: หุบเขา $f(x,y) = \frac{1}{2}(x-2)^2 + 10(y-1)^2 - 12$ ซึ่งเป็นหุบที่ความโค้งในแนวตั้งมากกว่าแนวนอนยี่สิบเท่า เส้นสีเทาคือระดับความสูงเท่ากัน วิธีลงชันที่สุดสิบสองก้าว (สีแดง) ซิกแซกข้ามหุบ ส่วนวิธีเกรเดียนต์สังยุคไปถึงก้นหุบในสองก้าว (สีน้ำเงิน) (สร้างจาก `numanal/ch11.py`)

1. **สูตรของ $\mathbf{r}_{k+1}$.** $\mathbf{r}_{k+1} = \mathbf{b} - A(\mathbf{x}_k + \alpha_k \mathbf{r}_k) = \mathbf{r}_k - \alpha_k A\mathbf{r}_k$
2. **คูณด้วย $\mathbf{r}_k^\top$.** $\mathbf{r}_{k+1}^\top \mathbf{r}_k = \mathbf{r}_k^\top \mathbf{r}_k - \alpha_k\, \mathbf{r}_k^\top A \mathbf{r}_k$
3. **แทน α_k.** ได้ $\mathbf{r}_k^\top \mathbf{r}_k - \mathbf{r}_k^\top \mathbf{r}_k = 0$ จึงสรุปได้

สรุป: ก้าวแต่ละก้าวตั้งฉากกับก้าวก่อนหน้า ในหุบเขายาวสิ่งนี้ทำให้ทิศสลับไปมาข้ามหุบ □

อ่านแบบเข้าใจง่าย

ทั้งหมดนี้อธิบายภาพในรูป 11.3 ก้าวแต่ละก้าวตั้งฉากกับก้าวก่อน ถ้าหุบยาวและแคบ ก้าวที่ตั้งฉากกันจะข้ามหุบไปมาแทนที่จะเดินตามความยาวของหุบ สิ่งนี้เกิดขึ้นกับตัวอย่าง $\mathrm{diag}(1, 20)$ ที่ความคลาดเคลื่อนในนอร์มพลังงานของทุกก้าวเป็น $\frac{19}{21}$ ของก้าวก่อนหน้าพอดี (ดูแบบฝึกหัด 11.1.2 และการทดสอบ)

ทฤษฎีบท 11.1.3: อัตราของวิธีลงชันที่สุด

ให้ A สมมาตรบวกแน่นอนที่มีคอนดิชันนัมเบอร์เชิงสเปกตรัม $\kappa = \lambda_{\max}/\lambda_{\min}$ แล้วสำหรับวิธีลงชันที่สุดที่เขียนในนิยามข้างบน

$$\|\mathbf{x}_k - \mathbf{x}^*\|_A \le \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|_A .$$

หมายเหตุการอ้างผล. เรายังไม่พิสูจน์ทฤษฎีบทนี้ในหัวข้อนี้ การพิสูจน์ใช้อสมการของคันโตโรวิช (Kantorovich inequality) และดูได้ใน Braess 2007, §IV.2, Thm 2.4, p. 199 ดูเพิ่มบทที่ IV ทฤษฎีบท 2.4 ของ Braess ตัวอย่างในตำราเดียวกัน Braess 2007, Example 2.5, p. 199 แสดงว่าอัตรานี้ถูกเข้าได้จริงในกรณีหนึ่ง

ร่างปรับปรุง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

ค่า $\frac{\kappa-1}{\kappa+1}$ ใกล้ $1-\frac{2}{\kappa}$ เมื่อ κ ใหญ่ ถ้า $\kappa = 140.86$ (ค่าของรอยขีดข่วนที่เราจะทำในหัวข้อย่อยท้ายบท ตัวเลขนี้มาจากโค้ดของเรา) ค่านี้เท่ากับ 0.9859 ขอบเขตจึงรับประกันเพียงว่าความคลาดเคลื่อนลดลงอย่างน้อยประมาณ 1.4% ต่อก้าว ซึ่งช้ามากเมื่อหุบยาว

ทิศที่ไม่ซ้ำทาง

ทิศตั้งฉากตามปกติอาจยังซ้ำทางกันในหุบยาว ความคิดของวิธีเกรเดียนต์สังยุคคือเลือกทิศที่ตั้งฉากเมื่อวัดด้วย A แทน

นิยาม 11.1.3: ทิศสังยุคเทียบ A

เวกเตอร์ที่ไม่เป็นศูนย์สองตัว $\mathbf{u}$ และ $\mathbf{v}$ เรียกว่า **สังยุคเทียบ** A (conjugate with respect to A) ถ้า $\mathbf{u}^{\top} A\mathbf{v} = 0$ เมื่อ $A = I$ ผลนี้ตรงกับการตั้งฉากตามปกติ

อ่านแบบเข้าใจง่าย

การวัดมุมด้วย $\mathbf{u}^{\top} A\mathbf{v}$ แทน $\mathbf{u}^{\top}\mathbf{v}$ ทำให้ทิศที่สังยุคกันไม่ขยับกลับไปทางเดิมที่เคยใช้มาแล้วในหุบเขาที่รูปร่างถูกบีบโดย A นี่คือแนวคิดที่ทำให้ทิศไม่วนกลับไปทางเดิม

นิยาม 11.1.4: วิธีเกรเดียนต์สังยุค

ให้ A สมมาตรบวกแน่นอน เริ่มจาก $\mathbf{x}_0$ ตั้ง $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$ และ $\mathbf{p}_0 = \mathbf{r}_0$ แล้วทำซ้ำสำหรับ $k = 0, 1, 2, \ldots$

1. $\alpha_k = \dfrac{\mathbf{r}_k^{\top}\mathbf{r}_k}{\mathbf{p}_k^{\top} A\mathbf{p}_k}$, $\quad \mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$, $\quad \mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k$;
2. $\beta_k = \dfrac{\mathbf{r}_{k+1}^{\top}\mathbf{r}_{k+1}}{\mathbf{r}_k^{\top}\mathbf{r}_k}$ และ $\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_k \mathbf{p}_k$.

อ่านแบบเข้าใจง่าย

ในแต่ละรอบเรามีการคูณ A กับเวกเตอร์หนึ่งครั้ง คือ $A\mathbf{p}_k$ และเก็บเวกเตอร์ไม่กี่ตัวเท่านั้น ไม่ต้องจำทิศทั้งหมดที่ผ่านมา ทิศใหม่ $\mathbf{p}_{k+1}$ คิดจากค่าตกค้างและทิศล่าสุดเพียงทิศเดียว นี่คือเหตุที่สูตรเรียกว่า**สูตรสั้น** (short recurrences)

ลองเอง: สองก้าวด้วยมือบนระบบ 2×2

ให้ $A = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$, $\mathbf{b} = (1, 2)^{\top}$, $\mathbf{x}_0 = \mathbf{0}$ คำตอบแม่นตรงคือ $\mathbf{x}^{*} = (\frac{1}{11}, \frac{7}{11})$ (ตรวจได้ด้วยการแทนกลับ)

รอบที่ 0. $\mathbf{r}_0 = \mathbf{p}_0 = (1, 2)^{\top}$ และ $A\mathbf{p}_0 = (6, 7)^{\top}$ จึง $\mathbf{p}_0^{\top} A\mathbf{p}_0 = 20$ และ $\alpha_0 = \frac{5}{20} = \frac{1}{4}$ ได้ $\mathbf{x}_1 = (\frac{1}{4}, \frac{1}{2})^{\top}$ และ $\mathbf{r}_1 = (-\frac{1}{2}, \frac{1}{4})^{\top}$

ฉบับร่าง · Preprint · 2026-10-23

รอบที่ 1. $\beta_0 = \dfrac{\mathbf{r}_1^\top \mathbf{r}_1}{\mathbf{r}_0^\top \mathbf{r}_0} = \dfrac{5/16}{5} = \dfrac{1}{16}$ ได้ $\mathbf{p}_1 = \mathbf{r}_1 + \beta_0 \mathbf{p}_0 = (-\frac{7}{16}, \frac{3}{8})^\top$ ต่อไป $A\mathbf{p}_1 = (-\frac{11}{8}, \frac{11}{16})^\top$ และ $\mathbf{p}_1^\top A\mathbf{p}_1 = \frac{55}{64}$ จึง $\alpha_1 = \dfrac{5/16}{55/64} = \dfrac{4}{11}$ ดังนั้น

$$\mathbf{x}_2 = \mathbf{x}_1 + \tfrac{4}{11}\mathbf{p}_1 = \left(\tfrac{1}{11}, \tfrac{7}{11}\right)^\top = \mathbf{x}^*.$$

ตรวจสองข้อ: $\mathbf{r}_1^\top \mathbf{r}_0 = -\frac{1}{2} + \frac{1}{2} = 0$ และ $\mathbf{p}_0^\top A\mathbf{p}_1 = -\frac{11}{8} + \frac{11}{8} = 0$ ทั้งสองตรงกับทฤษฎีบท 11.1.4

เทียบกับวิธีลงชันที่สุด. ถ้าเดินด้วยวิธีลงชันที่สุดสองก้าวจาก $\mathbf{x}_0$ จะได้ $\mathbf{x}_2 = (\frac{1}{12}, \frac{7}{12})$ ซึ่งยังไม่ถึง $\mathbf{x}^*$ (ดูการทดสอบ `test_ch11.py`)

ปริภูมิที่เกิดจากเวกเตอร์ $\mathbf{r}_0$ และการคูณด้วย A ซ้ำ ๆ คือ $\operatorname{span}\{\mathbf{r}_0, A\mathbf{r}_0, A^2\mathbf{r}_0, \dots\}$ เรียกว่า **ปริภูมิย่อยไครลอฟ** (Krylov subspace) ทิศและค่าตกค้างของวิธีเกรเดียนต์สังยุคอยู่ในปริภูมิเหล่านี้ที่เติบโตขึ้นทีละมิติ ดังที่ทฤษฎีบทถัดไปบอก

ทฤษฎีบท 11.1.4: สมบัติของวิธีเกรเดียนต์สังยุค

ให้ A สมมาตรบวกแน่นอนและรันวิธีเกรเดียนต์สังยุคในเลขคณิตแม่นตรง สำหรับทุก $k \geq 1$ ที่ $\mathbf{r}_0, \dots, \mathbf{r}_k$ ไม่เป็นศูนย์ ได้ว่า

1. $\mathbf{r}_k^\top \mathbf{r}_i = 0$ และ $\mathbf{r}_k^\top \mathbf{p}_i = 0$ สำหรับ $i < k$;
2. $\mathbf{p}_k^\top A\mathbf{p}_i = 0$ สำหรับ $i < k$;
3. ปริภูมิที่สร้างจาก $\mathbf{r}_0, \dots, \mathbf{r}_{k-1}$ และปริภูมิที่สร้างจาก $\mathbf{p}_0, \dots, \mathbf{p}_{k-1}$ เท่ากับปริภูมิย่อยไครลอฟ

$$\mathcal{K}_k(A, \mathbf{r}_0) = \operatorname{span}\{\mathbf{r}_0, A\mathbf{r}_0, \dots, A^{k-1}\mathbf{r}_0\}.$$

พิสูจน์. **แนวคิด.** ใช้การเหนี่ยวนำ: ทิศเก่าทุกทิศตั้งฉากกับค่าตกค้างใหม่ และค่าตกค้างใหม่ตั้งฉากกับค่าตกค้างเก่า ซึ่งทำให้ทิศใหม่สังยุคกับทิศเก่า

1. **ค่าตกค้างตั้งฉากกับทิศก่อนหน้า.** จากนิยาม $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k$ และ α_k ได้ $\mathbf{r}_{k+1}^\top \mathbf{p}_k = \mathbf{r}_k^\top \mathbf{p}_k - \alpha_k \mathbf{p}_k^\top A\mathbf{p}_k = 0$ โดยใช้ $\mathbf{r}_k^\top \mathbf{p}_k = \mathbf{r}_k^\top \mathbf{r}_k$ ซึ่งได้จาก $\mathbf{p}_k = \mathbf{r}_k + \beta_{k-1}\mathbf{p}_{k-1}$ และข้อนี้ของก้าวก่อนหน้า
2. **ค่าตกค้างตั้งฉากกับทิศเก่าทุกทิศ.** สมมติข้อ 2 ใช้ได้ถึง $k-1$ สำหรับ $i < k$ ใช้ $\mathbf{r}_{l+1} = \mathbf{r}_l - \alpha_l A\mathbf{p}_l$ ย้อนกลับจาก $\mathbf{r}_{i+1}$ ถึง $\mathbf{r}_k$ ได้ $\mathbf{r}_k = \mathbf{r}_{i+1} - \sum_{l=i+1}^{k-1} \alpha_l A\mathbf{p}_l$ เมื่อคูณด้วย $\mathbf{p}_i^\top$ พจน์ผลรวมเป็นศูนย์จากข้อสมมติ และ $\mathbf{r}_{i+1}^\top \mathbf{p}_i = 0$ จากข้อ 1 ดังนั้น $\mathbf{r}_k^\top \mathbf{p}_i = 0$
3. **ค่าตกค้างตั้งฉากกันเอง.** สำหรับ $i \geq 1$ มี $\mathbf{r}_i = \mathbf{p}_i - \beta_{i-1}\mathbf{p}_{i-1}$ จึง $\mathbf{r}_k^\top \mathbf{r}_i = \mathbf{r}_k^\top \mathbf{p}_i - \beta_{i-1}\mathbf{r}_k^\top \mathbf{p}_{i-1} = 0$ จากข้อ 2 และ $\mathbf{r}_0 = \mathbf{p}_0$ จบข้อ 1
4. **ทิศใหม่สังยุคกับทิศเก่า.** ถ้า $i \leq k-2$ จาก $\mathbf{p}_k = \mathbf{r}_k + \beta_{k-1}\mathbf{p}_{k-1}$ จะได้ $\mathbf{p}_k^\top A\mathbf{p}_i = \mathbf{r}_k^\top A\mathbf{p}_i + \beta_{k-1}\mathbf{p}_{k-1}^\top A\mathbf{p}_i$ พจน์หลังเป็นศูนย์จากข้อสมมติ และ $A\mathbf{p}_i = (\mathbf{r}_i - \mathbf{r}_{i+1})/\alpha_i$ จึง $\mathbf{r}_k^\top A\mathbf{p}_i = (\mathbf{r}_k^\top \mathbf{r}_i - \mathbf{r}_k^\top \mathbf{r}_{i+1})/\alpha_i = 0$ เพราะ $i+1 < k$ จากข้อ 3 ส่วน $i = k-1$ เลือก $\beta_{k-1} = -\mathbf{r}_k^\top A\mathbf{p}_{k-1}/\mathbf{p}_{k-1}^\top A\mathbf{p}_{k-1}$ ให้ $\mathbf{p}_k^\top A\mathbf{p}_{k-1} = 0$ และค่านี้ตรงกับสูตรของ β_{k-1} ในนิยาม เพราะ $\mathbf{r}_k^\top A\mathbf{p}_{k-1} = -\mathbf{r}_k^\top \mathbf{r}_k/\alpha_{k-1}$ และ $\mathbf{p}_{k-1}^\top A\mathbf{p}_{k-1} = \mathbf{r}_{k-1}^\top \mathbf{r}_{k-1}/\alpha_{k-1}$ ตามข้อ 1 และ 3
5. **ปริภูมิย่อยไครลอฟ.** สมมติว่า $\mathbf{r}_0, \dots, \mathbf{r}_{k-1}$ และ $\mathbf{p}_0, \dots, \mathbf{p}_{k-1}$ ต่างก็สร้าง $\mathcal{K}_k$ (กรณี $k = 1$ เป็นจริงเพราะ $\mathbf{p}_0 = \mathbf{r}_0$) เพราะ $\mathbf{p}_{k-1} \in \mathcal{K}_k$ จึง $A\mathbf{p}_{k-1} \in \mathcal{K}_{k+1}$ และจาก $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}A\mathbf{p}_{k-1}$ ได้ $\mathbf{r}_k \in \mathcal{K}_{k+1}$ ดัง

ฉบับร่าง · Preprint · 2026-10-23

นั้น $\mathbf{r}_0, \dots, \mathbf{r}_k$ อยู่ใน $\mathcal{K}_{k+1}$ เวกเตอร์ $k+1$ ตัวนี้ไม่เป็นศูนย์และตั้งฉากกันทีละคู่ จึงเป็นอิสระเชิงเส้นและสร้าง $\mathcal{K}_{k+1}$ พอดี ส่วน $\mathbf{p}_k = \mathbf{r}_k + \beta_{k-1}\mathbf{p}_{k-1}$ ทำให้ปริภูมิของทิศทั้งสองเท่ากันด้วย

สรุป: ข้อ 1 ถึง 3 ใช้ตั้งฉากเมื่อวัดแบบปกติ และข้อ 4 ใช้ตั้งฉากเมื่อวัดด้วย A สองส่วนนี้ช่วยกันรักษาให้กันและกันเป็นจริงตลอด □

ทฤษฎีบท 11.1.5: จุดต่ำสุดบนปริภูมิที่ขยาย และการจบภายในจำนวนรอบจำกัด

ให้ A สมมาตรบวกแน่นอนขนาด $n \times n$ และให้ $\mathbf{x}_k$ เป็นค่าจากวิธีเกรเดียนต์สังยุคในเลขคณิตแม่นตรง แล้ว

1. $f(\mathbf{x}_k) \le f(\mathbf{x})$ สำหรับทุก $\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_k(A, \mathbf{r}_0)$ ดังนั้น $\|\mathbf{x}_k - \mathbf{x}^*\|_A$ เป็นค่าที่น้อยที่สุดบนปริภูมิที่ขยาย;
2. มี $m \le n$ ที่ $\mathbf{r}_m = \mathbf{0}$ และ $\mathbf{x}_m = \mathbf{x}^*$

พิสูจน์. **แนวคิด.** ข้อ 1 เป็นการอ่านค่า f บนปริภูมิที่ค่าตกค้างตั้งฉาก ข้อ 2 เป็นการนับจำนวนเวกเตอร์ตั้งฉากกันที่ไม่เป็นศูนย์

1. **ข้อ 1.** ถ้า $\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_k$ จะได้ $\mathbf{d} = \mathbf{x} - \mathbf{x}_k \in \mathcal{K}_k$ เพราะ $\mathbf{x}_k - \mathbf{x}_0 = \sum_{i<k} \alpha_i \mathbf{p}_i$ และทิศ $\mathbf{p}_i$ อยู่ใน $\mathcal{K}_k$ ตามทฤษฎีบท 11.1.4 ข้อ 3 ส่วน $\mathbf{r}_k$ ตั้งฉากกับทุกเวกเตอร์ใน $\mathcal{K}_k$ เพราะตั้งฉากกับ $\mathbf{p}_i$ ทุกตัวที่ $i < k$ (ข้อ 1 ของทฤษฎีบทเดียวกัน) จึงได้ $f(\mathbf{x}_k + \mathbf{d}) = f(\mathbf{x}_k) - \mathbf{r}_k^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top A\mathbf{d} = f(\mathbf{x}_k) + \frac{1}{2}\mathbf{d}^\top A\mathbf{d}$ ซึ่งมากกว่าหรือเท่ากับ $f(\mathbf{x}_k)$ และเท่ากันเมื่อ $\mathbf{d} = \mathbf{0}$
2. **ข้อ 2.** ถ้า $\mathbf{r}_0, \dots, \mathbf{r}_n$ ไม่เป็นศูนย์ทั้งหมด จะได้ $n+1$ เวกเตอร์ที่ตั้งฉากกันทีละคู่และไม่เป็นศูนย์ เวกเตอร์เหล่านี้เป็นอิสระเชิงเส้นจึงมีได้ไม่เกิน n ตัวในปริภูมิ $\mathbb{R}^n$ เกิดข้อขัดแย้ง ดังนั้นมี $m \le n$ ที่ $\mathbf{r}_m = \mathbf{0}$ และจาก $A\mathbf{x}_m = \mathbf{b}$ จึงได้ $\mathbf{x}_m = \mathbf{x}^*$

สรุป: วิธีเกรเดียนต์สังยุคเลือกจุดที่ต่ำที่สุดบนปริภูมิที่ขยายขึ้นทุกรอบ และเมื่อปริภูมิเต็มมิติก็ถึงคำตอบ □

ข้อควรระวัง: จบภายใน n รอบ ใช้ได้เฉพาะเลขคณิตแม่นตรง

ข้อ 2 ของทฤษฎีบท 11.1.5 เป็นจริงในเลขคณิตแม่นตรง บนเครื่องจริงการปัดเศษทำให้ค่าตกค้างตั้งฉากกันไม่สมบูรณ์เมื่อ n ใหญ่ ดังนั้นวิธีนี้ไม่จบที่ n รอบแน่นอน และเราต้องนับรอบจนค่าตกค้างเล็กพอ ดังที่ทำในหัวข้อย่อยท้ายบทนี้ Iserles 2009, p. 315 เตือนเรื่องเดียวกัน

รอยขีดข่วนที่ถูกเติม

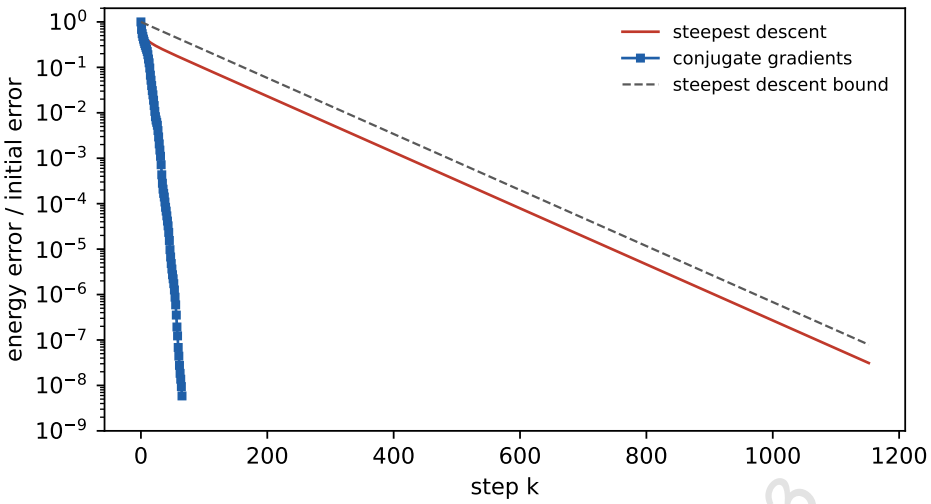
กลับไปที่รูปถ่ายกัน ให้ f^* เป็นรูปปลายทางที่มีแสงต่างจากต้นแบบเล็กน้อย และให้ g เป็นภาพต้นแบบที่สะอาดซึ่งเรารู้จากรูปอีกใบ (ทั้งสองเป็นสถานการณ์สมมติ ค่าตัวเลขมาจากโค้ดของเรา) บริเวณที่จะเติม Ω เป็นสี่เหลี่ยม 16×20 พิกเซล ค่าภาพในรูปปลายทางภายในบริเวณนี้ไม่ได้ใช้ เพราะเราจะแทนที่มัน

แนวคิดของเปเรซและคณะคือ ให้ตัวดำเนินการลาปลาซ Δ ของ f ภายในบริเวณเท่ากับของ g (เพราะใช้เกรเดียนต์ของ g เป็นสนามนำทาง) และให้ค่าที่ขอบเท่ากับ f^* Pérez, Gangnet, and Blake 2003, §3, p. 315 ในภาษาสมการคือ $\Delta f = \Delta g$ ใน Ω และ $f = f^*$ บนขอบ เมื่อดิสครีไทซ์บนพิกเซล สมการของจุด $p \in \Omega$ คือ

$$4f_p - \sum_{q \sim p,\ q \in \Omega} f_q = \sum_{q \sim p,\ q \notin \Omega} f^*_q + \sum_{q \sim p} (g_p - g_q)$$

เมื่อ $q \sim p$ หมายถึงเพื่อนบ้านสี่ทิศของ p ด้านซ้ายคือสูตรห้าจุดของบทที่ 10 และด้านขวาคือค่าที่ทราบจากขอบบวกกับความต่างของภาพต้นแบบ ดังนั้นได้ระบบ $A\mathbf{x} = \mathbf{b}$ โดยที่ A คือเมทริกซ์ห้าจุดของบล็อก 16×20 (ขนาด 320×320) และ $\mathbf{b}$ คือด้านขวานี้ที่คำนวณจากรูป

ฉบับร่าง Preprint 2026-10-23



รูปที่ 11.4: ค่าคลาดเคลื่อนในนอร์มพลังงานเทียบค่าเริ่มต้น (สเกลลอการิทึม) ของการเติมรอยขีดข่วนด้วยสองวิธี เส้นประคือขอบเขตของวิธีลงชันที่สุดจากทฤษฎีบท 11.1.3 ซึ่งลดลงตามอัตรา 0.9859 ต่อก้าว วิธีลงชันที่สุดอยู่ใต้เส้นประนี้ทุกก้าว ส่วนวิธีเกรเดียนต์สังยุคลดลงเร็วกว่ามาก (สร้างจาก `numanal/ch11.py`)

อ่านแบบเข้าใจง่าย

ถ้า $c = f - g$ ในบริเวณ สมการข้างบนกลายเป็น $4c_p - \sum c_q = 0$ โดยมีค่าที่ขอบ $f_q^* - g_q$ นั่นคือ c เป็นฟังก์ชันฮาร์มอนิก **ฟังก์ชันฮาร์มอนิก** (harmonic function) แบบดิสครีต ความต่างของแสงจึงถูกกระจายเรียบ ๆ เข้าไปในบริเวณแทนที่จะกระโดดที่ขอบ ค่าที่ขอบเพียงอย่างเดียวกำหนดค่า c ทั้งหมด

ตัวเลขจากโค้ดของเรา: A มีค่าลักษณะเฉพาะ $4 - 2\cos\frac{i\pi}{17} - 2\cos\frac{j\pi}{21}$ สำหรับ $i = 1, \ldots, 16$ และ $j = 1, \ldots, 20$ (ด้วยวิธีเดียวกับทฤษฎีบท 10.1.2 ของบทที่ 10 ซึ่งใช้กับบล็อกสี่เหลี่ยมได้) ทำให้ $\kappa \approx 140.86$ วิธีลงชันที่สุดต้องใช้ 1152 รอบ ส่วนวิธีเกรเดียนต์สังยุคใช้ 65 รอบ จนค่าตกค้างเหลือ 10^{-8} เท่าของ $\|\mathbf{b}\|$ ผลของ CG ห่างจากคำตอบแม่นตรงที่ได้จากการแก้ระบบโดยตรงไม่เกิน 10^{-8} เราวัดรอยต่อด้วยค่า $|f_p - f_q^*|$ ที่มากที่สุดของคู่พิกเซลติดกันที่ข้ามขอบบริเวณ ค่านี้ลดจาก 0.2423 (วางตรง ๆ) เหลือ 0.0311 หลังการเติม

ข้อควรระวัง: ค่าตกค้างเล็กไม่ได้แปลว่าคำตอบแม่น

การหยุดเมื่อค่าตกค้างเล็กเป็นเพียงการดูผลลัพธ์ ไม่ใช่การรับประกันความถูกต้องของคำตอบ ความสัมพันธ์ระหว่างค่าตกค้างกับความคลาดเคลื่อนขึ้นกับคอนดิชันนัมเบอร์ ถ้า κ ใหญ่ ความคลาดเคลื่อนอาจใหญ่กว่าค่าตกค้างหลายเท่า (ดูการขยายค่าคลาดเคลื่อนในบทที่ 7) เราจึงต้องรู้ κ ก่อนตั้งเกณฑ์หยุด

สรุปใจความ

- ระบบ $A\mathbf{x} = \mathbf{b}$ ที่ A สมมาตรบวกแน่นอนเทียบเท่ากับการหาจุดต่ำสุดของฟังก์ชันกำลังสอง f ซึ่งลดค่าลงตามระยะจากคำตอบในนอร์มพลังงาน
- วิธีลงชันที่สุดเดินตามค่าตกค้างและทุกก้าวตั้งฉากกับก้าวก่อน จึงซิกแซกเมื่อหุบยาว อัตราขึ้นกับ κ
- วิธีเกรเดียนต์สังยุคเลือกทิศที่สังยุคเทียบ A ด้วยสูตรสั้น ได้จุดต่ำสุดบนปริภูมิย่อยไครลอฟที่ขยายทุกรอบ และในเลขคณิตแม่นตรงจบภายใน n รอบ

ฉบับร่าง · Preprint · 2026-10-23

- กลับไปที่รอยขีด: รูป 40×40 ของเราใช้วิธีเกรเดียนต์สังยุค 65 รอบ และวิธีลงชันที่สุดใช้ 1152 รอบที่ระดับ 10^{-8} คำถามที่ค้างคือ เมื่อรอยใหญ่ขึ้นวิธีเกรเดียนต์สังยุคจะช้าลงเพียงใด

อ่านต่อ. หัวข้อ 11.2 จะถามว่า ถ้ารอยขีดข่วนใหญ่ขึ้นสิบเท่า CG ต้องใช้กี่รอบ และทำไมขึ้นกับ κ เราจะเปรียบเทียบการประมาณด้วยพหุนามเชบีเชฟ (Chebyshev polynomial) กับตัวเลขจริง อ่านเพิ่มได้ใน Iserles 2009, §14.2, pp. 312--316 Vorst 2003, §5.1, pp. 37--41 และ Braess 2007, §IV.2--IV.3, pp. 196--204 ดูเพิ่มบทที่ IV ทฤษฎีบท 2.4 และ 3.7 ของ Braess ดูเพิ่มบทที่ 14 ทฤษฎีบท 14.3 ของ Iserles และบทที่ 8 ทฤษฎีบท 8.30--8.31 ของ Grossmann

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 11.1.1 (⋆, ทำด้วยมือ) ให้ $A = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$ และ $\mathbf{b} = (1,2)^{\top}$ จงหาค่า $f(\mathbf{0}) - f(\mathbf{x}^*)$ โดยตรงจากนิยามของ f แล้วตรวจว่าเท่ากับ $\frac{1}{2}\|\mathbf{x}^*\|_A^2$ ตามทฤษฎีบท 11.1.1

แบบฝึกหัด 11.1.2 (⋆, ทำด้วยมือ) ให้ $A = \operatorname{diag}(1,4)$, $\mathbf{b} = \mathbf{0}$ และ $\mathbf{x}_0 = (4,1)^{\top}$ จงทำหนึ่งก้าวของวิธีลงชันที่สุดแล้วหาอัตราส่วนของความคลาดเคลื่อนในนอร์มพลังงานระหว่างสองจุดนี้ เทียบกับ $\frac{a-1}{a+1}$ เมื่อ $a = 4$

แบบฝึกหัด 11.1.3 (⋆⋆, คำนวณ) ให้ $A = \operatorname{diag}(2,1)$, $\mathbf{b} = (2,1)^{\top}$ และ $\mathbf{x}_0 = \mathbf{0}$ ทำวิธีเกรเดียนต์สังยุคสองรอบ จงหา α_0, $\mathbf{x}_1$, β_0, $\mathbf{p}_1$, α_1 และ $\mathbf{x}_2$ แล้วตรวจว่า $\mathbf{x}_2 = \mathbf{x}^*$ ตามที่ทฤษฎีบท 11.1.5 ทำนาย

แบบฝึกหัด 11.1.4 (⋆⋆, พิสูจน์) ให้ A สมมาตรบวกแน่นอน และให้ $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{r}_k$ เป็นก้าวของวิธีลงชันที่สุด จงแสดงว่า

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k) - \frac{1}{2}\,\frac{(\mathbf{r}_k^{\top}\mathbf{r}_k)^2}{\mathbf{r}_k^{\top} A \mathbf{r}_k}$$

แล้วใช้ผลนี้กับทฤษฎีบท 11.1.1 สรุปว่า $\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_A \le \|\mathbf{x}_k - \mathbf{x}^*\|_A$

แบบฝึกหัด 11.1.5 (⋆⋆, คำนวณและตีความ) ใช้ระบบของรอยขีดข่วนในหัวข้อ 11.1 นับจำนวนรอบของวิธีลงชันที่สุดและวิธีเกรเดียนต์สังยุค จนค่าตกค้างเหลือ 10^{-4}, 10^{-8} และ 10^{-12} เท่าของ $\|\mathbf{b}\|$ ตามลำดับ จากนั้น

1. ทำตารางจำนวนรอบทั้งสองวิธี และหาผลต่างของจำนวนรอบเมื่อลดค่าตกค้างจาก 10^{-4} ถึง 10^{-12};
2. ตรวจว่าผลต่างของวิธีลงชันที่สุดใกล้กับ $\dfrac{\ln 10^{8}}{-\ln\!\left(\frac{\kappa-1}{\kappa+1}\right)}$ เมื่อ $\kappa \approx 140.86$ หรือไม่ และอธิบายว่าทำไม;
3. ถ้านักเดินป่าขอให้เราหยุดเมื่อค่าตกค้างลดลงสิบเท่า ควรเลือกวิธีใด เพราะเหตุใด

แบบฝึกหัด 11.1.6 (⋆⋆, ทำด้วยมือ; ดูว่าอะไรผิด) ให้ $A = \operatorname{diag}(1,-1)$ ซึ่งสมมาตรแต่ไม่บวกแน่นอน และให้ $\mathbf{b} = (1,1)^{\top}$

1. จงแสดงว่า $f(0,-t) = -\frac{1}{2}t^2 + t$ และอธิบายว่าเหตุใด f ไม่มีจุดต่ำสุด
2. จงแสดงว่า $\mathbf{p}_0^{\top} A \mathbf{p}_0 = 0$ เมื่อ $\mathbf{x}_0 = \mathbf{0}$ และอธิบายว่าเหตุใดระยะก้าว α_0 ของวิธีเกรเดียนต์สังยุคจึงไม่นิยาม

ฉบับร่าง · Preprint 2026-10-23

แบบฝึกหัด 11.1.7 (⋆⋆⋆, พิสูจน์) ให้ A เป็นเมทริกซ์สมมาตรบวกแน่นอนขนาด 2×2 และ $\mathbf{b}$ เป็นเวกเตอร์ใดก็ได้ จงแสดงว่าวิธีเกรเดียนต์สังยุคหาคำตอบแม่นตรงได้ภายในสองรอบโดยไม่ขึ้นกับ $\mathbf{x}_0$

ตัวช่วยคิด

ใช้ทฤษฎีบท 11.1.4 ข้อ 1 กับ $k = 2$ และ $k = 1$ จะได้ว่า $\mathbf{r}_2$ ตั้งฉากกับทั้ง $\mathbf{r}_0$ และ $\mathbf{r}_1$ ในปริภูมิสองมิติ เวกเตอร์สองตัวที่ตั้งฉากกันและไม่เป็นศูนย์จะสร้างฐาน ลองคิดว่าเวกเตอร์ที่ตั้งฉากกับฐานทั้งสองพร้อมกันเป็นอย่างไร และอย่าลืมกรณี $\mathbf{r}_1 = \mathbf{0}$

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 11.1.1.

1. $f(\mathbf{0}) = 0$ เพราะทุกพจน์มี $\mathbf{x}$ เป็นตัวคูณ
2. แก้ $A\mathbf{x}^* = \mathbf{b}$: $4x_1 + x_2 = 1$ และ $x_1 + 3x_2 = 2$ ได้ $\mathbf{x}^* = (\frac{1}{11}, \frac{7}{11})$ (แทนกลับได้ $(1, 2)$)
3. $f(\mathbf{x}^*) = \frac{1}{2}\mathbf{x}^{*\top}A\mathbf{x}^* - \mathbf{b}^\top\mathbf{x}^* = \frac{1}{2}\mathbf{x}^{*\top}\mathbf{b} - \mathbf{b}^\top\mathbf{x}^* = -\frac{1}{2}\mathbf{b}^\top\mathbf{x}^* = -\frac{1}{2}\left(\frac{1}{11} + \frac{14}{11}\right) = -\frac{15}{22}$
4. ดังนั้น $f(\mathbf{0}) - f(\mathbf{x}^*) = \frac{15}{22}$
5. ตรวจด้วยทฤษฎีบท: $\mathbf{x}^{*\top}A\mathbf{x}^* = \mathbf{x}^{*\top}\mathbf{b} = \frac{15}{11}$ จึง $\frac{1}{2}\|\mathbf{x}^*\|_A^2 = \frac{15}{22}$ ตรงกัน

ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_1_1_energy_identity`)

เฉลย 11.1.2.

1. $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0 = (-4, -4)^\top$ จึง $\mathbf{r}_0^\top\mathbf{r}_0 = 32$ และ $A\mathbf{r}_0 = (-4, -16)^\top$ จึง $\mathbf{r}_0^\top A\mathbf{r}_0 = 16 + 64 = 80$
2. $\alpha_0 = \dfrac{32}{80} = \dfrac{2}{5}$ และ $\mathbf{x}_1 = (4, 1) + \frac{2}{5}(-4, -4) = (\frac{12}{5}, -\frac{3}{5})$
3. $\mathbf{e}_0 = (4, 1)$ มี $\|\mathbf{e}_0\|_A^2 = 16 + 4 = 20$ และ $\mathbf{e}_1 = (\frac{12}{5}, -\frac{3}{5})$ มี $\|\mathbf{e}_1\|_A^2 = \frac{144}{25} + 4 \cdot \frac{9}{25} = \frac{180}{25} = 7.2$
4. อัตราส่วนของนอร์ม $= \sqrt{7.2/20} = \sqrt{0.36} = 0.6$ ตรงกับ $\frac{a-1}{a+1} = \frac{3}{5} = 0.6$ เมื่อ $a = 4$

ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_1_2_one_step_ratio_is_0_6`)

เฉลย 11.1.3.

1. $\mathbf{r}_0 = \mathbf{p}_0 = (2, 1)^\top$ และ $\mathbf{p}_0^\top A\mathbf{p}_0 = 2 \cdot 4 + 1 \cdot 1 = 9$ จึง $\alpha_0 = \dfrac{5}{9}$ และ $\mathbf{x}_1 = \left(\frac{10}{9}, \frac{5}{9}\right)^\top$
2. $\mathbf{r}_1 = \mathbf{b} - A\mathbf{x}_1 = (2 - \frac{20}{9},\ 1 - \frac{5}{9})^\top = (-\frac{2}{9}, \frac{4}{9})^\top$ จึง $\mathbf{r}_1^\top\mathbf{r}_1 = \frac{20}{81}$ และ $\beta_0 = \dfrac{20/81}{5} = \dfrac{4}{81}$
3. $\mathbf{p}_1 = \mathbf{r}_1 + \beta_0\mathbf{p}_0 = \left(-\frac{2}{9} + \frac{8}{81},\ \frac{4}{9} + \frac{4}{81}\right)^\top = \left(-\frac{10}{81}, \frac{40}{81}\right)^\top$
4. $A\mathbf{p}_1 = (-\frac{20}{81}, \frac{40}{81})^\top$ จึง $\mathbf{p}_1^\top A\mathbf{p}_1 = \frac{1800}{6561} = \frac{200}{729}$ และ $\alpha_1 = \dfrac{20/81}{200/729} = \dfrac{9}{10}$
5. $\mathbf{x}_2 = \left(\frac{10}{9} - \frac{9}{10} \cdot \frac{10}{81},\ \frac{5}{9} + \frac{9}{10} \cdot \frac{40}{81}\right)^\top = (1, 1)^\top$ และ $\mathbf{x}^* = A^{-1}\mathbf{b} = (1, 1)^\top$ จึงตรงกัน

ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_1_3_diagonal_two_steps`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 11.1.4.

1. ให้ $\mathbf{r} = \mathbf{r}_k$ และ $\alpha = \alpha_k$ จาก $f(\mathbf{x} + t\mathbf{r}) = f(\mathbf{x}) - t\,\mathbf{r}^\top\mathbf{r} + \frac{1}{2}t^2\mathbf{r}^\top A\mathbf{r}$ (ดูหัวข้อ 11.1)
2. แทน $t = \alpha = \dfrac{\mathbf{r}^\top\mathbf{r}}{\mathbf{r}^\top A\mathbf{r}}$ ได้ $f(\mathbf{x} + \alpha\mathbf{r}) = f(\mathbf{x}) - \dfrac{(\mathbf{r}^\top\mathbf{r})^2}{\mathbf{r}^\top A\mathbf{r}} + \frac{1}{2}\dfrac{(\mathbf{r}^\top\mathbf{r})^2}{\mathbf{r}^\top A\mathbf{r}} = f(\mathbf{x}) - \frac{1}{2}\dfrac{(\mathbf{r}^\top\mathbf{r})^2}{\mathbf{r}^\top A\mathbf{r}}$
3. จากทฤษฎีบท 11.1.1 $f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_A^2$ สำหรับทุก $\mathbf{x}$ ข้อ 2 ทำให้ $f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k)$ จึง $\frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_A^2 \le \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}^*\|_A^2$ ซึ่งให้ผลที่ต้องการ

ตรวจด้วยโค้ด: `test_ch11.py` (`test_steepest_descent_decrease_identity`) และการตรวจแบบสุ่มในเทสต์เดียวกัน

เฉลย 11.1.5.

1. ตารางจำนวนรอบ (นับจากระบบของรอยขีดข่วนด้วยค่าตกค้างเทียบ $\|\mathbf{b}\|$):

ค่าตกค้างเทียบกับ $\|\mathbf{b}\|$	เกรเดียนต์สังยุค	ลงชันที่สุด
10^{-4}	39	503
10^{-8}	65	1152
10^{-12}	86	1800

   ผลต่างเมื่อลดจาก 10^{-4} ถึง 10^{-12} คือ 47 รอบของ CG และ 1297 รอบของวิธีลงชันที่สุด

2. จากทฤษฎีบท 11.1.3 แต่ละก้าวลดความคลาดเคลื่อนของวิธีลงชันที่สุดเป็นอย่างน้อย $\frac{\kappa-1}{\kappa+1} = 0.985902$ เท่า ดังนั้นการลดลง 10^8 เท่าต้องใช้อย่างมาก $\dfrac{\ln 10^8}{-\ln(0.985902)} \approx 1297.4$ รอบ ซึ่งใกล้กับ 1297 ที่นับได้ ค่าที่นับได้จึงสอดคล้องกับอัตราของทฤษฎีบท แต่ตัวเลขนี้เป็นการตรวจแบบคร่าวจากโค้ดของเรา ไม่ใช่การพิสูจน์

3. เลือกวิธีเกรเดียนต์สังยุค จากตารางเพิ่มเติม (ค่าตกค้าง $10^{-4} \to 10^{-5}$) วิธีเกรเดียนต์สังยุคเพิ่ม 8 รอบ ($39 \to 47$) ส่วนวิธีลงชันที่สุดเพิ่ม 163 รอบ ($503 \to 666$) ความต่างนี้จะยิ่งใหญ่ขึ้นเมื่อหุบยาวขึ้น ซึ่งเป็นเรื่องของหัวข้อ 11.2

ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_1_5_counts_for_three_tolerances`)

เฉลย 11.1.6.

1. $f(x, y) = \frac{1}{2}(x^2 - y^2) - x - y$ เมื่อ $x = 0$, $y = -t$ จะได้ $f(0, -t) = -\frac{1}{2}t^2 + t$ ซึ่งลดลงไม่มีที่สิ้นสุดเมื่อ $t \to \infty$ ดังนั้น f ไม่มีค่าต่ำสุด ทฤษฎีบท 11.1.1 ต้องการให้ A บวกแน่นอน แต่ที่นี่ A มีค่าลักษณะเฉพาะ -1 จึงใช้ไม่ได้

2. $\mathbf{r}_0 = \mathbf{p}_0 = \mathbf{b} = (1, 1)^\top$ และ $A\mathbf{p}_0 = (1, -1)^\top$ จึง $\mathbf{p}_0^\top A\mathbf{p}_0 = 1 - 1 = 0$ ดังนั้น $\alpha_0 = \mathbf{r}_0^\top\mathbf{r}_0/0$ ไม่นิยาม วิธีเกรเดียนต์สังยุคต้องใช้ว่าค่านี้ไม่เป็นศูนย์ ซึ่งการที่ A บวกแน่นอนรับประกันเสมอ

ตรวจด้วยโค้ด: `test_ch11.py` (`test_cg_breakdown_on_an_indefinite_matrix` และ `test_indefinite_bowl_has_no_minimum`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 11.1.7.

1. ถ้า $\mathbf{r}_1 = \mathbf{0}$ แล้ว $\mathbf{x}_1 = \mathbf{x}^*$ เสร็จแล้ว สมมติต่อไปว่า $\mathbf{r}_0$ และ $\mathbf{r}_1$ ไม่เป็นศูนย์
2. จากทฤษฎีบท 11.1.4 ข้อ 1 ที่ $k = 1$ ได้ $\mathbf{r}_1^\top \mathbf{r}_0 = 0$ จึงสองเวกเตอร์ที่ไม่เป็นศูนย์และตั้งฉากกันนี้เป็นอิสระเชิงเส้น และสร้าง $\mathbb{R}^2$ ได้
3. จากทฤษฎีบทเดียวกันที่ $k = 2$ ได้ $\mathbf{r}_2^\top \mathbf{r}_0 = \mathbf{r}_2^\top \mathbf{r}_1 = 0$ ถ้าเขียน $\mathbf{r}_2 = c_0\mathbf{r}_0 + c_1\mathbf{r}_1$ แล้ว $\mathbf{r}_2^\top \mathbf{r}_2 = c_0\, \mathbf{r}_0^\top \mathbf{r}_2 + c_1\, \mathbf{r}_1^\top \mathbf{r}_2 = 0$ ดังนั้น $\mathbf{r}_2 = \mathbf{0}$
4. $\mathbf{r}_2 = \mathbf{0}$ หมายความว่า $A\mathbf{x}_2 = \mathbf{b}$ จึง $\mathbf{x}_2 = \mathbf{x}^*$ โดยไม่ขึ้นกับ $\mathbf{x}_0$ และ $\mathbf{b}$

ตรวจด้วยโค้ด: `test_ch11.py` (`test_two_steps_solve_any_two_by_two_system`) ซึ่งสุ่มระบบ 2×2 ยี่สิบชุด

11.2 เร็วแค่ไหน

สถานการณ์สมมติ: รอยขีดข่วนที่ใหญ่ขึ้นสิบเท่า

ในหัวข้อ 11.1 เราเติมรอยขนาด 16×20 พิกเซลในรูป 40×40 พิกเซล ด้วยวิธีเกรเดียนต์สังยุคประมาณ 65 รอบ ลองนึกถึงรูปเดียวกันที่ละเอียดขึ้นสิบเท่าทั้งสองด้าน รอยขีดข่วนยาว 160×200 พิกเซล ตอนนี้มีค่าที่ไม่รู้ 32 000 ค่า มากกว่าเดิมร้อยเท่า

ก่อนรันโค้ด ลองทายว่าเมื่อจำนวนค่าที่ไม่รู้เพิ่มร้อยเท่า จำนวนรอบจะเพิ่มร้อยเท่าด้วยหรือไม่

ตัวเลขทั้งหมดในหัวข้อนี้มาจากโค้ดของเรา ไม่ได้มาจากเกมหรือแอปใด

ภาพรวมก่อนอ่าน

หลัง k รอบ ความคลาดเคลื่อนของวิธีเกรเดียนต์สังยุคเป็นพหุนามของ A คูณกับความคลาดเคลื่อนเริ่มต้น ถ้าเลือกพหุนามที่มีค่าเล็กบนช่วงที่ค่าลักษณะเฉพาะของ A อยู่ ความคลาดเคลื่อนก็จะเล็กลง**พหุนามเชบีเชฟ** (Chebyshev polynomial) ให้ขอบเขตที่ดีแบบหนึ่ง และขอบเขตนี้ขึ้นกับ $\sqrt{\kappa}$ ไม่ใช่กับจำนวนตัวแปรโดยตรง ส่วนต้นทุนต่อรอบยังขึ้นกับจำนวนตัวแปร เพราะแต่ละรอบต้องคูณ A หนึ่งครั้ง

ลองเอง: หนึ่งรอบบนช่วง $[1, 9]$

สมมติค่าลักษณะเฉพาะของ A อยู่ในช่วง $[1, 9]$ ดังนั้น $\kappa = 9$ เราอยากเลือกพหุนามดีกรี 1 รูป $p(t) = 1 - ct$ ที่ทำให้ $|p(t)|$ มีค่าสูงสุดบน $[1, 9]$ น้อยที่สุด เพราะ p เป็นเส้นตรง ค่าสูงสุดของ $|p|$ อยู่ที่ปลายทั้งสอง คือ $|1 - c|$ และ $|1 - 9c|$ ค่าสองค่านี้เท่ากันและเป็นค่าเล็กที่สุดเมื่อ $c = \frac{1}{5}$ จะได้ $|p(1)| = |p(9)| = \frac{4}{5}$ ดังนั้นพหุนามดีกรี 1 ที่ดีที่สุดให้ค่าสูงสุดเหลือ 0.8 ซึ่งเท่ากับ $\frac{\kappa-1}{\kappa+1}$ ตัวเลขนี้เป็นอัตราของวิธีลงชันที่สุดในทฤษฎีบท 11.1.3 ความต่างจะปรากฏเมื่อ $k \geq 2$

ความคลาดเคลื่อนหลังทำ k รอบเป็นพหุนามของ A

ให้ $\mathbf{e}_k = \mathbf{x}_k - \mathbf{x}^*$ เป็นความคลาดเคลื่อนหลัง k รอบ หลัง k รอบ $\mathbf{x}_k - \mathbf{x}_0$ อยู่ในปริภูมิย่อยไครลอฟ ดังนั้นความคลาดเคลื่อนจึงเขียนเป็นพหุนามของ A คูณ $\mathbf{e}_0$ ได้เสมอ คำถามเรื่องจำนวนรอบจึงเปลี่ยนเป็นคำถามว่ามีพหุนามดีกรีต่ำ ๆ ที่เล็กบนค่าลักษณะเฉพาะหรือไม่ ทฤษฎีบทต่อไปทำให้ข้อนี้แม่นยำ

ฉบับร่าง · Preprint · 2026-10-23

ทฤษฎีบท 11.2.1: ความคลาดเคลื่อนเป็นพหุนามของ A

ดูเพิ่มบทที่ 5 ข้อเสนอ 5.3 ของ van der Vorst ให้ A สมมาตรบวกแน่นอน ให้ $\mathbf{e}_0 = \mathbf{x}_0 - \mathbf{x}^*$ และ $\mathbf{e}_k = \mathbf{x}_k - \mathbf{x}^*$ เมื่อ $\mathbf{x}_k$ เป็นค่าจากวิธีเกรเดียนต์สังยุคในเลขคณิตแม่นตรง แล้ว

$$\|\mathbf{e}_k\|_A = \min\bigl\{\|p(A)\mathbf{e}_0\|_A : \ p \text{ เป็นพหุนามดีกรีไม่เกิน } k \text{ และ } p(0)=1\bigr\}.$$

พิสูจน์. **แนวคิด.** ค่า $\mathbf{x}_k - \mathbf{x}_0$ อยู่ในปริภูมิย่อยไครลอฟ และทุกจุดในปริภูมินั้นเขียนได้เป็นพหุนามของ A คูณ $\mathbf{e}_0$

1. **ค่าตกค้างเริ่มต้น.** เพราะ $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0 = A(\mathbf{x}^* - \mathbf{x}_0) = -A\mathbf{e}_0$ ปริภูมิย่อยไครลอฟ $\mathcal{K}_k(A, \mathbf{r}_0)$ จึงเท่ากับ $\operatorname{span}\{A\mathbf{e}_0, \dots, A^k\mathbf{e}_0\}$
2. **ทุกจุดเป็นพหุนาม.** จากทฤษฎีบท 11.1.5 ข้อ 1 ค่า $\mathbf{x}_k$ ทำให้ $\|\mathbf{x} - \mathbf{x}^*\|_A$ น้อยที่สุดเมื่อ $\mathbf{x}$ วิ่งผ่าน $\mathbf{x}_0 + \mathcal{K}_k$ จุดใดก็ตามในนั้นเขียนได้เป็น $\mathbf{x} - \mathbf{x}^* = \mathbf{e}_0 + \sum_{j=1}^{k} c_j A^j \mathbf{e}_0 = p(A)\mathbf{e}_0$ โดย $p(t) = 1 + \sum_{j=1}^{k} c_j t^j$ มีดีกรีไม่เกิน k และ $p(0) = 1$ ในทางกลับกัน พหุนามทุกตัวแบบนี้ให้จุดในปริภูมินั้น จึงสองเซตนี้ตรงกัน
3. **สรุป.** ค่าต่ำสุดของ $\|\mathbf{x} - \mathbf{x}^*\|_A$ บน $\mathbf{x}_0 + \mathcal{K}_k$ คือค่าต่ำสุดของ $\|p(A)\mathbf{e}_0\|_A$ เหนือพหุนามทั้งหมดในข้อ 2

สรุป: วิธีเกรเดียนต์สังยุคเลือกพหุนาม p ที่ดีที่สุดในชั้นที่มี $p(0)=1$ และดีกรีไม่เกิน k เสมอ □

ขอบเขตเชบีเชฟ

พหุนามที่ใช้กันมากที่สุดสำหรับงานนี้คือพหุนามเชบีเชฟ ซึ่งบนช่วง $[-1, 1]$ มีค่าอยู่ระหว่าง -1 ถึง 1 เสมอ และโตเร็วมากเมื่อออกนอกช่วง เราจะย่อช่วงของค่าลักษณะเฉพาะให้ไปอยู่ในช่วง $[-1, 1]$ แล้วพหุนามที่ได้จะเล็กบนค่าลักษณะเฉพาะทั้งหมด

นิยาม 11.2.1: พหุนามเชบีเชฟ

พหุนามเชบีเชฟดีกรี k คือ

$$T_k(x) = \cos\bigl(k \arccos x\bigr) \qquad (-1 \le x \le 1),$$

$$T_k(x) = \tfrac{1}{2}\Bigl[(x + \sqrt{x^2-1})^k + (x - \sqrt{x^2-1})^k\Bigr] \qquad (x \ge 1).$$

อ่านแบบเข้าใจง่าย

สูตรสองบรรทัดนี้เป็นพหุนามตัวเดียวกัน บนช่วง $[-1, 1]$ ค่าของมันอยู่ในช่วง -1 ถึง 1 เสมอ นอกช่วงค่าโตเร็วตามดีกรี ตัวอย่างเช่น เพราะ $T_k(\cos\theta) = \cos k\theta$ จึงได้ $T_1(x) = x$ และ $T_2(x) = 2x^2 - 1$

ทฤษฎีบท 11.2.2: ขอบเขตของวิธีเกรเดียนต์สังยุค

ให้ A สมมาตรบวกแน่นอนที่มีคอนดิชันนัมเบอร์ $\kappa = \lambda_{\max}/\lambda_{\min}$ และให้ $\mathbf{x}_k$ เป็นค่าจากวิธีเกรเดียนต์สังยุคในเลขคณิตแม่นตรง แล้วสำหรับทุก $k \ge 0$

$$\|\mathbf{x}_k - \mathbf{x}^*\|_A \le 2\Bigl(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Bigr)^k \|\mathbf{x}_0 - \mathbf{x}^*\|_A.$$

พิสูจน์. **แนวคิด.** เลือกพหุนามเชบีเชฟที่ย่อ $[\lambda_{\min}, \lambda_{\max}]$ ลงใน $[-1, 1]$ แล้วใช้ทฤษฎีบทก่อนหน้า

ฉบับร่าง · Preprint · 2026-10-25

1. **เลือกพหุนาม.** ให้ $a = \lambda_{\min}$, $b = \lambda_{\max}$, $s(t) = \dfrac{a+b-2t}{b-a}$ เมื่อ $t \in [a,b]$ จะได้ $s(t) \in [-1,1]$ และ $s(0) = \dfrac{a+b}{b-a} = \dfrac{\kappa+1}{\kappa-1} =: z$ นิยาม $p(t) = T_k(s(t))/T_k(z)$ ซึ่งมีดีกรี k และ $p(0) = 1$ (ตรงนี้ $T_k(z) > 0$ เพราะ $z > 1$)

2. **ค่าบนค่าลักษณะเฉพาะ.** ค่าลักษณะเฉพาะทุกค่าของ A อยู่ใน $[a,b]$ ดังนั้น $|T_k(s(\lambda))| \le 1$ และ $|p(\lambda)| \le 1/T_k(z)$

3. **ขนาดของ $p(A)\mathbf{e}_0$.** เขียน $\mathbf{e}_0 = \sum_i c_i \mathbf{v}_i$ โดย $\mathbf{v}_i$ เป็นเวกเตอร์ลักษณะเฉพาะที่ตั้งฉากกัน แล้ว $\|p(A)\mathbf{e}_0\|_A^2 = \sum_i \lambda_i\, p(\lambda_i)^2 c_i^2 \le \dfrac{1}{T_k(z)^2}\,\|\mathbf{e}_0\|_A^2$

4. **ใช้ทฤษฎีบทก่อนหน้า.** จากทฤษฎีบท 11.2.1 $\|\mathbf{e}_k\|_A \le \|p(A)\mathbf{e}_0\|_A \le \dfrac{1}{T_k(z)}\|\mathbf{e}_0\|_A$

5. **ประมาณ $T_k(z)$ จากข้างล่าง.** จาก $z^2-1 = \dfrac{4\kappa}{(\kappa-1)^2}$ ได้ $\sqrt{z^2-1} = \dfrac{2\sqrt{\kappa}}{\kappa-1}$ และ $z+\sqrt{z^2-1} = \dfrac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$ (เพราะ $\kappa+1+2\sqrt{\kappa} = (\sqrt{\kappa}+1)^2$ และ $\kappa-1 = (\sqrt{\kappa}-1)(\sqrt{\kappa}+1)$) ค่าอีกราก $z-\sqrt{z^2-1}$ มีผลคูณกับตัวแรกเท่ากับ 1 จึงไม่ลบ ดังนั้น $T_k(z) \ge \tfrac{1}{2}\big(\tfrac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\big)^k$ และ $\dfrac{1}{T_k(z)} \le 2\Big(\dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^k$

สรุป: ความคลาดเคลื่อนในนอร์มพลังงานลดลงอย่างน้อยเท่ากับ $2\rho^k$ โดย $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ ซึ่งขึ้นกับ $\sqrt{\kappa}$ □

อ่านแบบเข้าใจง่าย

ที่ $\kappa = 140.86$ ของรอยขีดข่วนเรา ได้ $\rho \approx 0.8446$ ซึ่งหมายความว่าขอบเขตนี้เทียบเท่าการลดลงราว 15% ต่อรอบ (เป็นค่าประมาณ) ขอบเขตบอกว่าความคลาดเคลื่อนลดลงเป็น 10^{-8} เท่าภายใน 114 รอบ ตัวเลขนี้เป็นการคำนวณจากสูตรของเรา และโค้ดยืนยันว่า 114 เป็นจำนวนรอบที่ขอบเขตรับประกัน

ขอบเขตนี้ใช้สะดวกเพราะแปลงเป็นจำนวนรอบได้ทันที เนื่องจาก $\ln\dfrac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} = 2\,\operatorname{artanh}\dfrac{1}{\sqrt{\kappa}} \ge \dfrac{2}{\sqrt{\kappa}}$ เมื่อ $\kappa > 1$ จึงได้ $\rho^k \le e^{-2k/\sqrt{\kappa}}$ และ $2\rho^k \le \varepsilon$ เมื่อ $k \ge \tfrac{1}{2}\sqrt{\kappa}\ln\dfrac{2}{\varepsilon}$ นั่นคือจำนวนรอบที่ขึ้นกับ $\sqrt{\kappa}$ เป็นเชิงเส้น ไม่ใช่ κ

เมื่อค่าลักษณะเฉพาะรวมกันเป็นกลุ่ม

ขอบเขตข้างบนรู้เพียงปลายสองข้างของช่วง $[\lambda_{\min}, \lambda_{\max}]$ แต่ถ้าค่าลักษณะเฉพาะรวมกันเป็นกลุ่มไม่กี่กลุ่ม พหุนามดีกรีต่ำสามารถมีรากตรงกลางแต่ละกลุ่มได้ ทำให้ค่าบนกลุ่มนั้นเป็นศูนย์ และความคลาดเคลื่อนหายไปเร็วกว่าขอบเขตมาก กรณีสุดขั้วคือเมื่อมีค่าลักษณะเฉพาะที่ต่างกันเพียง m ค่า

ทฤษฎีบท 11.2.3: จำนวนค่าลักษณะเฉพาะที่ต่างกันกับจำนวนรอบ

ให้ A สมมาตรบวกแน่นอนและมีค่าลักษณะเฉพาะที่ต่างกันเพียง m ค่า แล้วสำหรับทุก $\mathbf{x}_0$ วิธีเกรเดียนต์สังยุคในเลขคณิตแม่นตรงให้ $\mathbf{x}_m = \mathbf{x}^*$

พิสูจน์. **แนวคิด.** พหุนามดีกรี m ที่มีรากที่ค่าลักษณะเฉพาะทุกค่า ทำให้ $p(A)\mathbf{e}_0 = \mathbf{0}$ ทันที

1. **พหุนามที่หายไป.** ให้ $\mu_1, \dots, \mu_m$ เป็นค่าลักษณะเฉพาะที่ต่างกันทั้งหมด และนิยาม $q(t) = \prod_{j=1}^{m}\big(1 - \frac{t}{\mu_j}\big)$ ซึ่งมีดีกรี m และ $q(0) = 1$ ($\mu_j > 0$)

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 11.1: จำนวนรอบของวิธีเกรเดียนต์สังยุคเมื่อเติมรอยขนาด $16s \times 20s$ เกณฑ์ค่าตกค้างคือ $\|\mathbf{r}\| \le 10^{-8}\|\mathbf{b}\|$ และเกณฑ์ค่าคลาดเคลื่อนคือ $\|\mathbf{e}\|_A \le 10^{-8}\|\mathbf{e}_0\|_A$ (สร้างจาก `numanal/ch11.py` และตรวจด้วย `test_ch11.py`)

s	จำนวนค่าที่ไม่รู้	κ	รอบ (ค่าตกค้าง)	รอบ (ค่าคลาดเคลื่อน)	ขอบเขต
1	320	140.86	65	64	114
2	1 280	535.03	124	126	221
3	2 880	1 182.25	181	186	329
4	5 120	2 082.53	240	246	437
5	8 000	3 235.86	297	305	544
6	11 520	4 642.25	355	364	652
8	20 480	8 214.19	470	481	867
10	32 000	12 798.36	582	598	1 082

2. **ค่าที่ได้เป็นศูนย์.** เขียน $\mathbf{e}_0 = \sum_{i=1}^{n} c_i \mathbf{v}_i$ ตามเวกเตอร์ลักษณะเฉพาะ แล้ว $q(A)\mathbf{e}_0 = \sum_i c_i\, q(\lambda_i)\mathbf{v}_i$ ทุก λ_i เป็นหนึ่งใน μ_j จึง $q(\lambda_i) = 0$ และ $q(A)\mathbf{e}_0 = \mathbf{0}$

3. **สรุป.** จากทฤษฎีบท 11.2.1 ได้ $\|\mathbf{e}_m\|_A \le \|q(A)\mathbf{e}_0\|_A = 0$ ดังนั้น $\mathbf{e}_m = \mathbf{0}$

สรุป: จำนวนรอบไม่เกินจำนวนค่าลักษณะเฉพาะที่ต่างกัน ซึ่งอาจน้อยกว่าจำนวนตัวแปรมาก $\square$

ผลนี้อยู่ในรูปแบบแบบฝึกหัดของ Vorst 2003, §5.3, p. 48 ดูเพิ่มบทที่ 5 ข้อเสนอ 5.3 ของ van der Vorst การพิสูจน์ข้างบนเราเขียนเอง ส่วนแนวคิดว่าการรวมกลุ่มของค่าลักษณะเฉพาะช่วยให้ลู่เข้าเร็วขึ้น ตรงกับที่ Atkinson 1989, §8.9, p. 567 กล่าวไว้ว่า เมื่อค่าลักษณะเฉพาะรวมกันรอบค่าจำนวนน้อย ก็ใช้สูตรให้ลู่เข้าได้ภายในไม่กี่รอบ

กลับไปที่รอยที่ใหญ่ขึ้น

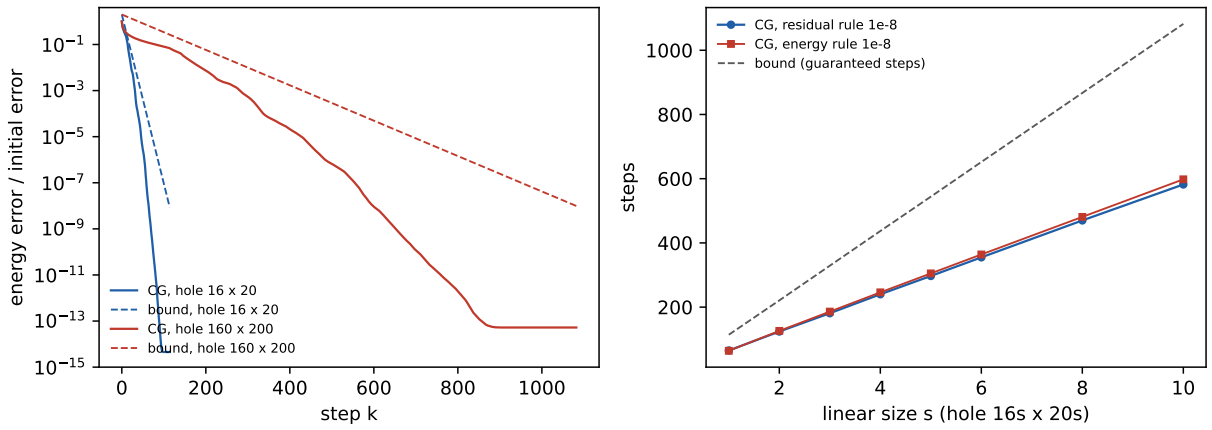
เราทำให้รูปเดิมละเอียดขึ้นได้โดยสุ่มตัวอย่างฉากเดิมที่ $40s \times 40s$ พิกเซล และให้รอยเป็น $16s \times 20s$ พิกเซล (ที่ $s = 1$ ได้ระบบของหัวข้อ 11.1 พอดี) บล็อกห้าจุดนี้มีค่าลักษณะเฉพาะที่รู้รูปแบบปิด ดังนั้น κ ก็คำนวณได้ตรง ๆ และเมื่อ s โตขึ้น κ โตราว s^2 จึง $\sqrt{\kappa}$ โตราว s ตารางต่อไปนี้วัดจำนวนรอบด้วยเกณฑ์สองแบบ เกณฑ์แรกคือค่าตกค้างเทียบ $\|\mathbf{b}\|$ ซึ่งเราใช้ในหัวข้อ 11.1 เกณฑ์ที่สองคือค่าคลาดเคลื่อนในนอร์มพลังงานเทียบค่าเริ่มต้น ซึ่งเป็นปริมาณที่ขอบเขตควบคุม คำตอบแม่นตรงที่ใช้เปรียบเทียบหาได้จากเวกเตอร์ลักษณะเฉพาะของเมทริกซ์หนึ่งมิติในแต่ละทิศ ไม่ได้มาจากการวนซ้ำ CG

ตารางและรูปบอกเรื่องเดียวกันสามข้อ ข้อแรก เมื่อจำนวนค่าที่ไม่รู้เพิ่มร้อยเท่า (320 เป็น $32\,000$) จำนวนรอบเพิ่มราวเก้าเท่า (64 เป็น 598 ตามเกณฑ์ค่าคลาดเคลื่อน) ข้อสอง κ เพิ่มราวเก้าสิบเท่า และ $\sqrt{\kappa}$ เพิ่มราว 9.5 เท่า จำนวนรอบจึงตามรากของ κ ข้อสาม ขอบเขตรับประกันมากกว่าจำนวนที่วัดได้ราว 1.8 เท่าในทุกขนาด เพราะขอบเขตดูเฉพาะปลายสองข้างของช่วงค่าลักษณะเฉพาะ สองเกณฑ์ให้จำนวนรอบต่างกันไม่กี่รอบ เช่น 65 กับ 64 ที่ $s = 1$ เพราะวัดปริมาณต่างกัน

เกร็ดเพิ่มเติม (ผู้เรียบเรียง)

ทฤษฎีบทข้างบนเป็นเลขคณิตแม่นตรง ในเลขคณิตจุดลอยตัว การปัดเศษอาจทำให้ค่าตกค้างเสียการตั้งฉากกันเมื่อ n ใหญ่ CG จึงอาจลู่เข้าช้าลงหรือต้องเริ่มใหม่ ส่วนนี้เกินระดับของหนังสือเล่มนี้ ดู Iserles 2009, p. 315

ฉบับร่าง Preprint 2026-10-23



รูปที่ 11.5: (ก) ค่าคลาดเคลื่อนในนอร์มพลังงานเทียบค่าเริ่มต้นของวิธีเกรเดียนต์สังยุค (เส้นทึบ) กับขอบเขต $2\rho^k$ (เส้นประ) สำหรับรอย 16×20 และ 160×200 (ข) จำนวนรอบเทียบกับขนาดเชิงเส้น s จากสองเกณฑ์ และจำนวนรอบที่ขอบเขตรับประกัน (สร้างจาก `numanal/ch11.py`)

ข้อควรระวัง: ขอบเขตไม่ใช่การทำนายจำนวนรอบ

เลข $1\,082$ ในตารางคือจำนวนรอบที่รับประกันได้ ไม่ใช่จำนวนที่ต้องใช้จริง จำนวนที่วัดได้คือ 598 ขอบเขตยังต้องใช้ κ ซึ่งเราต้องรู้หรือประมาณก่อน ถ้าประมาณ κ ต่ำเกินจริง ขอบเขตที่คำนวณจากค่านั้นจะไม่รับประกันอีกต่อไป

สรุปใจความ

- ความคลาดเคลื่อนหลัง k รอบคือพหุนาม $p(A)$ คูณ $\mathbf{e}_0$ โดย $p(0) = 1$ และวิธีเกรเดียนต์สังยุคเลือกพหุนามที่ดีที่สุดในชั้นนั้น
- ขอบเขตเชบีเชฟ $2\rho^k$ โดย $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ ทำให้ต้องการราว $\frac{1}{2}\sqrt{\kappa}\ln\frac{2}{\varepsilon}$ รอบ
- ถ้าค่าลักษณะเฉพาะที่ต่างกันมีเพียง m ค่า วิธีนี้จบภายใน m รอบ
- กลับไปที่รอย: จำนวนค่าที่ไม่รู้เพิ่มร้อยเท่า แต่รอบเพิ่มราวเก้าเท่า เพราะ $\sqrt{\kappa}$ โตตามขนาดเชิงเส้น และขอบเขตยังอยู่เหนือค่าที่วัดได้เสมอ
- คำถามที่ค้างคือ เราจะทำให้ κ เล็กลงได้อย่างไรก่อนแก้ ซึ่งเป็นเรื่องของหัวข้อ 11.3

อ่านต่อ. หัวข้อ 11.3 จะปรับระบบก่อนแก้ เพื่อให้ κ หรือการรวมกลุ่มของค่าลักษณะเฉพาะดีขึ้น อ่านเพิ่มเกี่ยวกับขอบเขตและการประมาณได้ใน Vorst 2003, §5.3, pp. 47--52 ดูเพิ่มบทที่ 5 ข้อเสนอ 5.3 ของ van der Vorst และ Braess 2007, §IV.3, Thm 3.7, p. 206 ดูเพิ่มบทที่ IV ทฤษฎีบท 3.7 ของ Braess

ฉบับร่าง · Preprint · 2026-10-25

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 11.2.1 (⋆, ทำด้วยมือ) ให้ $\kappa = 9$ จงหา $\rho = \dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ แล้วคำนวณขอบเขต $2\rho^k$ สำหรับ $k = 0, 1, 2, 3$ และหาจำนวนรอบ k ที่น้อยที่สุดซึ่งทำให้ $2\rho^k \le 10^{-3}$

แบบฝึกหัด 11.2.2 (⋆, ทำด้วยมือ) ให้ $\kappa = 9$ และ $z = \dfrac{\kappa+1}{\kappa-1}$

1. จงแสดงว่า $z + \sqrt{z^2-1} = 2$ และ $z - \sqrt{z^2-1} = \tfrac{1}{2}$
2. จงคำนวณ $T_1(z)$ และ $T_2(z)$ จากนิยามของพหุนามเชบีเชฟ แล้วหา $1/T_2(z)$ และเปรียบเทียบกับ $2\rho^2$ จากแบบฝึกหัด 11.2.1

แบบฝึกหัด 11.2.3 (⋆⋆, คำนวณและตีความ) ให้ $A_n = \operatorname{tridiag}(-1, 2, -1)$ ขนาด $n \times n$ ซึ่งเป็นวิธีห้าจุดในหนึ่งมิติ โดยมีค่าลักษณะเฉพาะ $2 - 2\cos\dfrac{j\pi}{n+1}$ สำหรับ $j = 1, \dots, n$

1. จงแสดงว่า $\kappa_n = \cot^2\dfrac{\pi}{2(n+1)}$
2. จงหาจำนวนรอบ k ที่น้อยที่สุดซึ่งทำให้ขอบเขต $2\rho^k \le 10^{-6}$ สำหรับ $n = 10$, 100 และ 1000
3. เมื่อ n เพิ่มเป็นสิบเท่า จำนวนรอบเพิ่มขึ้นประมาณกี่เท่า และสัมพันธ์กับ $\sqrt{\kappa_n}$ อย่างไร

แบบฝึกหัด 11.2.4 (⋆⋆, คำนวณและตีความ; กลับไปที่สถานการณ์) ใช้รอยขนาด $s = 5$ คือ 80×100 พิกเซลในรูป 200×200 พิกเซล

1. ก่อนรันโค้ด ทำนายว่าจำนวนค่าที่ไม่รู้และขอบเขตจำนวนรอบจะเป็นกี่เท่าของกรณี $s = 1$ (ซึ่งมี 320 ค่า และขอบเขต 114 รอบ)
2. ใช้โค้ดหาจำนวนรอบจริงทั้งสองเกณฑ์และค่า κ แล้วเปรียบเทียบกับการทำนายในข้อ 1
3. อธิบายว่าทำไมขอบเขตจึงสูงกว่าจำนวนรอบจริงอยู่ราว 1.8 เท่า

แบบฝึกหัด 11.2.5 (⋆⋆⋆, พิสูจน์) ให้ U เป็นเมทริกซ์ขนาด $n \times r$ และให้ $A = I + UU^{\top}$ จงแสดงว่า วิธีเกรเดียนต์สังยุคให้ $\mathbf{x}_{r+1} = \mathbf{x}^*$ สำหรับทุก $\mathbf{x}_0$ และทุก $\mathbf{b}$

> **ตัวช่วยคิด**
>
> ค่าลักษณะเฉพาะของ $UU^{\top}$ ที่ไม่เป็นศูนย์มีได้ไม่เกิน r ค่า ค่าลักษณะเฉพาะของ A เป็น 1 บวกค่าเหล่านั้น ลองนับจำนวนค่าที่ต่างกันทั้งหมดของ A แล้วใช้ทฤษฎีบท 11.2.3

แบบฝึกหัด 11.2.6 (⋆⋆, ดูว่าอะไรผิด) ให้ A สมมาตรบวกแน่นอนมีคอนดิชันนัมเบอร์ κ ให้ $\mathbf{x}$ เป็นค่าประมาณใด ๆ และ $\mathbf{r} = \mathbf{b} - A\mathbf{x}$

1. จงแสดงว่า $\dfrac{\|\mathbf{x}-\mathbf{x}^*\|}{\|\mathbf{x}^*\|} \le \kappa\,\dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$ โดยใช้ $\mathbf{x}-\mathbf{x}^* = -A^{-1}\mathbf{r}$, $\|A^{-1}\| = 1/\lambda_{\min}$ และ $\|\mathbf{x}^*\| \ge \|\mathbf{b}\|/\lambda_{\max}$
2. ถ้า $\kappa = 10^6$ และค่าตกค้างสัมพัทธ์เป็น 10^{-8} จงหาขอบเขตบนของความคลาดเคลื่อนสัมพัทธ์ และอธิบายว่าเหตุใดค่าตกค้างเล็กจึงไม่รับประกันคำตอบแม่นเสมอไป

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 11.2.7 (⋆ ⋆ ⋆, คำนวณและตีความ; กลับไปที่สถานการณ์) จากตารางของหัวข้อ 11.2 ใช้จำนวนรอบตามเกณฑ์ค่าคลาดเคลื่อนสำหรับ $s = 1$ และ $s = 10$ ให้ α เป็นความชันของ $\log(\text{รอบ})$ เทียบกับ $\log s$ และ β เป็นความชันของ $\log \kappa$ เทียบกับ $\log s$ จากสองจุดนี้

1. คำนวณ α และ β
2. อธิบายว่าทำไม α จึงใกล้ครึ่งหนึ่งของ β
3. ตามสัดส่วนนี้ ถ้าต้องการให้จำนวนรอบลดลงครึ่งหนึ่งที่ขนาดเดิม κ ต้องลดลงราวกี่เท่า (ตอบจากสัดส่วนประมาณ ไม่ต้องรันโค้ด)

ตัวช่วยคิด

ใช้ค่า $\log_{10}$ ของอัตราส่วน $598/64$ และ $12798.36/140.86$ แล้วคิดว่าขอบเขต $\frac{1}{2}\sqrt{\kappa}\ln\frac{2}{\varepsilon}$ ขึ้นกับ κ อย่างไร

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 11.2.1.

1. $\rho = \dfrac{3-1}{3+1} = \dfrac{1}{2}$
2. $2\rho^{k}$ เป็น $2,\ 1,\ \frac{1}{2},\ \frac{1}{4}$ สำหรับ $k = 0, 1, 2, 3$
3. $2 \cdot 2^{-k} \le 10^{-3}$ เทียบเท่า $2^{k-1} \ge 1000$ ซึ่งได้ $k \ge 1 + \log_2 1000 \approx 10.97$ ดังนั้น $k = 11$ ตรวจได้ว่า $2 \cdot 2^{-10} \approx 1.95 \times 10^{-3} > 10^{-3}$ และ $2 \cdot 2^{-11} \approx 9.8 \times 10^{-4} \le 10^{-3}$

ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_2_1_rate_for_kappa_9`)

เฉลย 11.2.2.

1. $z = \dfrac{10}{8} = \dfrac{5}{4}$ จึง $z^2 - 1 = \dfrac{25}{16} - 1 = \dfrac{9}{16}$ และ $\sqrt{z^2-1} = \dfrac{3}{4}$ ดังนั้น $z + \sqrt{z^2-1} = \dfrac{5}{4} + \dfrac{3}{4} = 2$ และ $z - \sqrt{z^2-1} = \dfrac{5}{4} - \dfrac{3}{4} = \dfrac{1}{2}$ (ผลคูณของทั้งสองเท่ากับ 1 ตามที่ควร)
2. เมื่อ $z > 1$ นิยามให้ $T_k(z) = \frac{1}{2}\big(2^k + 2^{-k}\big)$ จึง $T_1(z) = \frac{1}{2}(2 + \frac{1}{2}) = \frac{5}{4}$ ซึ่งตรงกับ z และ $T_2(z) = \frac{1}{2}(4 + \frac{1}{4}) = \frac{17}{8}$ ซึ่งตรงกับ $2z^2 - 1 = \frac{50}{16} - 1 = \frac{17}{8}$ ดังนั้น $1/T_2(z) = \frac{8}{17} \approx 0.471$ ซึ่งไม่เกิน $2\rho^2 = \frac{1}{2}$ ตามทฤษฎีบท

ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_2_2_identity_and_t2_for_kappa_9`)

ร่างเพื่อการอ่าน · Preprint · 2026-10-23

เฉลย 11.2.3.

1. ค่าลักษณะเฉพาะที่เล็กที่สุดคือ $\lambda_{\min} = 2 - 2\cos\dfrac{\pi}{n+1} = 4\sin^2\dfrac{\pi}{2(n+1)}$ และที่ใหญ่ที่สุดคือ $\lambda_{\max} = 2 - 2\cos\dfrac{n\pi}{n+1} = 2 + 2\cos\dfrac{\pi}{n+1} = 4\cos^2\dfrac{\pi}{2(n+1)}$ จึง $\kappa_n = \cot^2\dfrac{\pi}{2(n+1)}$

2. $\kappa_{10} \approx 48.37$, $\kappa_{100} \approx 4133.6$ และ $\kappa_{1000} \approx 406\,095$ เมื่อใช้ $2\rho^k \le 10^{-6}$ ได้ $k = 51, 467$ และ 4623 ตามลำดับ

3. $467/51 \approx 9.2$ และ $4623/467 \approx 9.9$ จำนวนรอบจึงเพิ่มราวสิบเท่า เพราะ $\sqrt{\kappa_n} \approx \dfrac{2(n+1)}{\pi}$ เป็นสัดส่วนกับ n และขอบเขตรับประกันของรอบเป็นสัดส่วนกับ $\sqrt{\kappa}$ (ตรวจค่า: $\sqrt{406\,095} \approx 637.3$ และ $2 \cdot 1001/\pi \approx 637.0$)

ตรวจด้วยโค้ด: `test_ch11.py`
(`test_exercise_11_2_3_one_dimensional_kappa_and_bound_steps`)

เฉลย 11.2.4.

1. จำนวนค่าที่ไม่รู้เพิ่มเป็น $8000/320 = 25$ เท่า เพราะจำนวนค่าเป็นสัดส่วนกับ s^2 เมื่อ κ โตราว s^2 ดังนั้น $\sqrt{\kappa}$ โตราว 5 เท่า และขอบเขตรอบโตราว 5 เท่า (การทำนายนี้เป็นการประมาณ)

2. จากโค้ด $\kappa \approx 3235.86$ ค่าตกค้างต้องใช้ 297 รอบ (กรณี $s = 1$ คือ 65) ค่าคลาดเคลื่อนใช้ 305 รอบ (กรณี $s = 1$ คือ 64) และขอบเขตคือ 544 รอบ (กรณี $s = 1$ คือ 114) อัตราส่วนสามค่านี้คือราว 4.6 ถึง 4.8 เท่า ซึ่งใกล้กับการทำนายข้อ 1 ที่ 5 เท่า

3. ขอบเขตเป็นค่าที่แย่ที่สุดสำหรับทุกค่าความคลาดเคลื่อนเริ่มต้นที่มี κ เดียวกัน และไม่ได้ใช้การกระจายของค่าลักษณะเฉพาะ ดังนั้นจึงสูงกว่าจำนวนรอบที่วัดได้ราว 1.8 เท่า ในทุกขนาดของตาราง (ดูหัวข้อการรวมกลุ่มของค่าลักษณะเฉพาะ)

ตรวจด้วยโค้ด: `test_ch11.py`
(`test_scaled_study_counts` ที่ $s = 5$)

เฉลย 11.2.5.

1. $UU^\top$ มีอันดับไม่เกิน r จึงมีค่าลักษณะเฉพาะที่ไม่เป็นศูนย์ไม่เกิน r ค่า และค่าลักษณะเฉพาะที่เหลือเป็นศูนย์

2. ดังนั้น $A = I + UU^\top$ มีค่าลักษณะเฉพาะไม่เกิน r ค่าที่มาจาก $1 + \sigma$ และอีกหนึ่งค่าคือ 1 รวมไม่เกิน $r + 1$ ค่าที่ต่างกัน

3. A สมมาตรและบวกแน่นอนเพราะ $UU^\top$ ไม่เป็นลบ จากทฤษฎีบท 11.2.3 จึงได้ $\mathbf{x}_{r+1} = \mathbf{x}^*$ สำหรับทุก $\mathbf{x}_0$ และทุก $\mathbf{b}$

ตรวจด้วยโค้ด: `test_ch11.py`
(`test_exercise_11_2_5_low_rank_perturbation_ends_in_r_plus_1_steps`)
ซึ่งสุ่ม $n = 8$ และ $r = 3$ พบว่าค่าตกค้างหลัง 4 รอบเหลือราว 10^{-16} เท่า แต่หลัง 3 รอบยังเหลือราว 0.014 เท่า

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 11.2.6.

1. $\mathbf{x}-\mathbf{x}^* = -A^{-1}\mathbf{r}$ จึง $\|\mathbf{x}-\mathbf{x}^*\| \le \|\mathbf{r}\|/\lambda_{\min}$ และ $\|\mathbf{b}\| = \|A\mathbf{x}^*\| \le \lambda_{\max}\|\mathbf{x}^*\|$ จึง $1/\|\mathbf{x}^*\| \le \lambda_{\max}/\|\mathbf{b}\|$ เมื่อนำมาคูณกันได้ $\dfrac{\|\mathbf{x}-\mathbf{x}^*\|}{\|\mathbf{x}^*\|} \le \dfrac{\|\mathbf{r}\|}{\lambda_{\min}} \cdot \dfrac{\lambda_{\max}}{\|\mathbf{b}\|} = \kappa \dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$
2. ขอบเขตบนคือ $\kappa \cdot 10^{-8} = 10^{-2}$ คือความคลาดเคลื่อนสัมพัทธ์อาจถึง 1% ทั้งที่ค่าตกค้างสัมพัทธ์เล็กถึง 10^{-8} ค่าตกค้างเล็กจึงรับประกันคำตอบแม่นได้เฉพาะเมื่อ κ ไม่ใหญ่มาก

ตรวจด้วยโค้ด: `test_ch11.py`
(`test_exercise_11_2_6_relative_error_at_most_kappa_times_residual`)
ซึ่งสุ่มระบบ 6×6 สองร้อยชุด และพบว่าอัตราส่วนสูงสุดของ “ความคลาดเคลื่อนต่อขอบเขต” คือราว 0.023

เฉลย 11.2.7.

1. จากตาราง $\alpha = \log_{10} \dfrac{598}{64} \approx 0.9705$ และ $\beta = \log_{10} \dfrac{12798.36}{140.86} \approx 1.958$
2. ถ้าจำนวนรอบเป็นสัดส่วนกับ $\sqrt{\kappa}$ แล้วความชันของรอบต่อ s จะเป็นครึ่งหนึ่งของความชันของ κ ค่า $\frac{1}{2}\cdot 1.958 \approx 0.979$ ใกล้กับ 0.9705 ความต่างเล็กน้อยเกิดจากรอบจริงไม่ได้ตามรากของ κ พอดี
3. ถ้ารอบเป็นสัดส่วนกับ $\sqrt{\kappa}$ การลดจำนวนรอบลงครึ่งหนึ่งต้องลด κ ลงราว 4 เท่า (คำตอบนี้มาจากสัดส่วนประมาณ ไม่ได้รันโค้ด)

ตรวจด้วยโค้ด: `test_ch11.py`
(`test_exercise_11_2_7_steps_and_kappa_growth_with_the_linear_size`)

11.3 ปรับโจทย์ให้ง่ายขึ้น

สถานการณ์สมมติ: เปลี่ยนหน่วยของแผนที่

สมมติว่าแผนที่ของนักเดินป่าใช้มาตราส่วนต่างกันในสองแนว แนวตะวันออก–ตกใช้หน่วยหนึ่ง แนวเหนือ–ใต้ใช้อีกหน่วยหนึ่ง ทำให้หุบเขาที่จริงกลมดูยาวและแคบบนแผนที่ ทางที่ชันที่สุดบนแผนที่จึงซิกแซกเหมือนในหัวข้อ 11.1 ถ้าปรับหน่วยของแกนให้ถูกต้อง หุบจะกลมขึ้น และการเดินตรงลงชันที่สุดจะพาไปถึงก้นได้เร็วขึ้น คำถามคือ เราหาการปรับหน่วยแบบนี้ได้โดยไม่รู้คำตอบล่วงหน้าหรือไม่ (ตัวอย่างนี้เป็นสถานการณ์สมมติ)

ภาพรวมก่อนอ่าน

หัวข้อ 11.2 บอกว่า อัตราขึ้นกับคอนดิชันนัมเบอร์ κ ถ้า κ ใหญ่ต้องใช้รอบมาก วิธีที่เราจะใช้คือ เลือกเมทริกซ์ M ที่หน้าตาคล้าย A แต่แก้ระบบด้วย M ได้ง่าย แล้วแก้ระบบบน $M^{-1}A$ แทน ผลคือ κ ลดลงและจำนวนรอบก็ลดลง สิ่งที่ต้องจำคือ ความดีของ M วัดจากค่าลักษณะเฉพาะของ $M^{-1}A$ ไม่ได้วัดจากขนาดของ $A - M$ โดยตรง

หุบที่ถูกเปลี่ยนหน่วย

ให้ M เป็นเมทริกซ์สมมาตรบวกแน่นอน แยกได้เป็น $M = BB^{\top}$ โดย B ไม่เอกฐาน (ทำได้เสมอด้วย**การแยกตัวประกอบโชเลสกี** (Cholesky decomposition)) เมื่อเปลี่ยนตัวแปร $\mathbf{y} = B^{\top}\mathbf{x}$ ระบบ $A\mathbf{x} = \mathbf{b}$ กลายเป็น

$$C\mathbf{y} = \mathbf{h}, \qquad C = B^{-1}AB^{-\top}, \qquad \mathbf{h} = B^{-1}\mathbf{b}.$$

ฉบับร่าง · preprint · 2026-10-23

การเปลี่ยนตัวแปรนี้คือการเปลี่ยนหน่วยของแกน ถ้าเลือก $M = A$ จะได้ $C = I$ หุบกลมพอดีและเดินตรงครั้งเดียวก็ถึง แต่การแยก A เองยากพอ ๆ กับการแก้ระบบเดิม เราจึงต้องการ M ที่อยู่ระหว่างสองขั้วนี้

นิยาม 11.3.1: ตัวปรับสภาพเงื่อนไขและวิธี PCG

ให้ A สมมาตรบวกแน่นอน และให้ M เป็นเมทริกซ์สมมาตรบวกแน่นอนที่เราเลือกเอง โดยให้การแก้ $M\mathbf{z} = \mathbf{r}$ ถูกกว่าการแก้ $A\mathbf{x} = \mathbf{b}$ เรียก M ว่า**ตัวปรับสภาพเงื่อนไข** (preconditioner) ของ A เริ่มจาก $\mathbf{x}_0$ ตั้ง $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$, $\mathbf{z}_0 = M^{-1}\mathbf{r}_0$ และ $\mathbf{p}_0 = \mathbf{z}_0$ แล้วทำซ้ำสำหรับ $k = 0, 1, 2, \ldots$

1. $\alpha_k = \dfrac{\mathbf{r}_k^\top \mathbf{z}_k}{\mathbf{p}_k^\top A\mathbf{p}_k}$, $\quad \mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$, $\quad \mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k$;
2. $\mathbf{z}_{k+1} = M^{-1}\mathbf{r}_{k+1}$, $\quad \beta_k = \dfrac{\mathbf{r}_{k+1}^\top \mathbf{z}_{k+1}}{\mathbf{r}_k^\top \mathbf{z}_k}$ และ $\mathbf{p}_{k+1} = \mathbf{z}_{k+1} + \beta_k \mathbf{p}_k$.

วิธีนี้เรียกว่า**วิธีเกรเดียนต์สังยุคที่ปรับสภาพเงื่อนไข** (preconditioned conjugate gradients) หรือ PCG

อ่านแบบเข้าใจง่าย

ต่อหนึ่งรอบ PCG ทำสองอย่าง คือคูณ A กับเวกเตอร์หนึ่งครั้ง และแก้ระบบ $M\mathbf{z} = \mathbf{r}$ หนึ่งครั้ง ถ้า $M = I$ จะได้ $\mathbf{z}_k = \mathbf{r}_k$ และสูตรทั้งหมดกลับไปเป็นวิธีเกรเดียนต์สังยุคเดิม ค่าใช้จ่ายส่วนเพิ่มจึงอยู่ที่การแก้ระบบกับ M เท่านั้น

ลองเอง: PCG สองรอบบนระบบ 2×2 ด้วยตัวปรับสภาพแบบจาโคบี

ให้ $A = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$, $\mathbf{b} = (4, 3)^\top$, $\mathbf{x}_0 = \mathbf{0}$ และ $M = \operatorname{diag}(4, 3)$ ซึ่งคือแนวทแยงของ A จึงแก้ $M\mathbf{z} = \mathbf{r}$ ได้ทันที

รอบที่ 0. $\mathbf{r}_0 = (4, 3)^\top$ และ $\mathbf{z}_0 = (1, 1)^\top$ จึง $\mathbf{p}_0 = \mathbf{z}_0$ ต่อไป $A\mathbf{p}_0 = (5, 4)^\top$, $\mathbf{r}_0^\top \mathbf{z}_0 = 7$, $\mathbf{p}_0^\top A\mathbf{p}_0 = 9$ และ $\alpha_0 = \frac{7}{9}$ ได้ $\mathbf{x}_1 = (\frac{7}{9}, \frac{7}{9})^\top$ และ $\mathbf{r}_1 = (\frac{1}{9}, -\frac{1}{9})^\top$

รอบที่ 1. $\mathbf{z}_1 = (\frac{1}{36}, -\frac{1}{27})^\top$ และ $\mathbf{r}_1^\top \mathbf{z}_1 = \frac{7}{972}$ จึง $\beta_0 = \dfrac{7/972}{7} = \frac{1}{972}$ ได้ $\mathbf{p}_1 = \mathbf{z}_1 + \beta_0 \mathbf{p}_0 = \frac{7}{972}(4, -5)^\top$ ต่อไป $A\mathbf{p}_1 = \frac{7}{972}(11, -11)^\top$ และ $\mathbf{p}_1^\top A\mathbf{p}_1 = \frac{49 \cdot 99}{972^2}$ จึง $\alpha_1 = \frac{108}{77}$ ดังนั้น

$$\mathbf{x}_2 = \mathbf{x}_1 + \tfrac{108}{77}\,\mathbf{p}_1 = \left(\tfrac{9}{11}, \tfrac{8}{11}\right)^\top = \mathbf{x}^*.$$

ตรวจสองข้อ: $\mathbf{r}_1^\top \mathbf{z}_0 = \frac{1}{9} - \frac{1}{9} = 0$ และ $\mathbf{p}_0^\top A\mathbf{p}_1 = 0$ ระบบนี้มีสองตัวแปร จึงจบในสองรอบ เพราะ PCG คือ CG บนระบบที่เปลี่ยนตัวแปรแล้ว (ทฤษฎีบท 11.3.1)

ทฤษฎีบท 11.3.1: PCG คือ CG บนระบบที่เปลี่ยนตัวแปรแล้ว

ให้ A สมมาตรบวกแน่นอน $M = BB^\top$ โดย B ไม่เอกฐาน และให้ $C = B^{-1}AB^{-\top}$, $\mathbf{h} = B^{-1}\mathbf{b}$ ถ้า $\mathbf{y}_k$ เป็นค่าจากวิธีเกรเดียนต์สังยุคบนระบบ $C\mathbf{y} = \mathbf{h}$ ที่เริ่มจาก $\mathbf{y}_0 = B^\top \mathbf{x}_0$ แล้ว $\mathbf{x}_k = B^{-\top}\mathbf{y}_k$ สำหรับทุก k โดยที่ $\mathbf{x}_k$ เป็นค่าจาก PCG ในนิยามข้างบน (ในเลขคณิตแม่นตรง)

พิสูจน์. **แนวคิด.** ทุกปริมาณของ CG บน C แปลงกลับเป็นปริมาณของ PCG ได้ด้วยการคูณ B หรือ $B^{-\top}$ และ $M^{-1} = B^{-\top}B^{-1}$ เป็นตัวเชื่อมทุกอย่าง

ฉบับร่าง · Preprint · 2026-10-23

1. **ตัวแปรที่แปลงแล้ว.** ให้ $\hat{\mathbf{r}}_k = \mathbf{h} - C\mathbf{y}_k = B^{-1}(\mathbf{b} - A\mathbf{x}_k) = B^{-1}\mathbf{r}_k$ และ $\hat{\mathbf{p}}_k = B^{\top}\mathbf{p}_k$ เมื่อ $k = 0$ ค่าเริ่มต้น $\mathbf{y}_0 = B^{\top}\mathbf{x}_0$ ให้ $C\mathbf{y}_0 = B^{-1}A\mathbf{x}_0$ จึง $\hat{\mathbf{r}}_0 = B^{-1}\mathbf{r}_0$ และจาก $\mathbf{p}_0 = \mathbf{z}_0 = M^{-1}\mathbf{r}_0$ ได้ $\hat{\mathbf{p}}_0 = \hat{\mathbf{r}}_0$ ทั่วไป $\hat{\mathbf{r}}_k^{\top}\hat{\mathbf{r}}_k = \mathbf{r}_k^{\top}B^{-\top}B^{-1}\mathbf{r}_k = \mathbf{r}_k^{\top}\mathbf{z}_k$
2. **ระยะก้าวตรงกัน.** $\hat{\mathbf{p}}_k^{\top}C\hat{\mathbf{p}}_k = \mathbf{p}_k^{\top}BB^{-1}AB^{-\top}B^{\top}\mathbf{p}_k = \mathbf{p}_k^{\top}A\mathbf{p}_k$ ดังนั้นระยะก้าวของ CG คือ $\hat{\alpha}_k = \dfrac{\mathbf{r}_k^{\top}\mathbf{z}_k}{\mathbf{p}_k^{\top}A\mathbf{p}_k} = \alpha_k$
3. **ตำแหน่งและค่าตกค้าง.** คูณ $\mathbf{y}_{k+1} = \mathbf{y}_k + \alpha_k\hat{\mathbf{p}}_k$ ด้วย $B^{-\top}$ ได้ $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{p}_k$ คูณ $\hat{\mathbf{r}}_{k+1} = \hat{\mathbf{r}}_k - \alpha_k C\hat{\mathbf{p}}_k$ ด้วย B ได้ $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k$ เพราะ $BC\hat{\mathbf{p}}_k = A\mathbf{p}_k$
4. **ทิศถัดไป.** จากข้อ 1 $\hat{\beta}_k = \dfrac{\hat{\mathbf{r}}_{k+1}^{\top}\hat{\mathbf{r}}_{k+1}}{\hat{\mathbf{r}}_k^{\top}\hat{\mathbf{r}}_k} = \dfrac{\mathbf{r}_{k+1}^{\top}\mathbf{z}_{k+1}}{\mathbf{r}_k^{\top}\mathbf{z}_k} = \beta_k$ และ $\hat{\mathbf{p}}_{k+1} = \hat{\mathbf{r}}_{k+1} + \beta_k\hat{\mathbf{p}}_k$ คูณด้วย $B^{-\top}$ ได้ $\mathbf{p}_{k+1} = B^{-\top}B^{-1}\mathbf{r}_{k+1} + \beta_k\mathbf{p}_k = \mathbf{z}_{k+1} + \beta_k\mathbf{p}_k$
5. **สรุป.** ทุกสูตรของ PCG ตรงกับ CG บน C ที่แปลงกลับ ดังนั้นเมื่อเริ่มจากค่าเดียวกัน ทั้งสองให้ค่า $\mathbf{x}_k$ เดียวกันทุกรอบ

สรุป: PCG คือการรัน CG บนตัวแปรใหม่ การเลือก M จึงเป็นการเลือกหน่วยของแกนนั่นเอง □

ทฤษฎีบท 11.3.2: อัตราของ PCG

ให้ A และ M สมมาตรบวกแน่นอน และให้ $\kappa' = \lambda_{\max}/\lambda_{\min}$ ของเมทริกซ์ $M^{-1}A$ (ค่าลักษณะเฉพาะทั้งหมดเป็นจำนวนจริงบวก) แล้วสำหรับ PCG ในเลขคณิตแม่นตรง

$$\|\mathbf{x}_k - \mathbf{x}^*\|_A \le 2\left(\frac{\sqrt{\kappa'}-1}{\sqrt{\kappa'}+1}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|_A.$$

พิสูจน์. **แนวคิด.** ทฤษฎีบทเกี่ยวกับ CG ในหัวข้อ 11.2 ใช้กับ C ได้ และนอร์มพลังงานของ A เท่ากับนอร์มของ C หลังเปลี่ยนตัวแปร

1. **ค่าลักษณะเฉพาะ.** $M^{-1}A = B^{-\top}CB^{\top}$ เป็นเมทริกซ์คล้ายกับ C จึงมีค่าลักษณะเฉพาะเดียวกัน และ C สมมาตรบวกแน่นอนเพราะ $\mathbf{v}^{\top}C\mathbf{v} = \mathbf{u}^{\top}A\mathbf{u} > 0$ เมื่อ $\mathbf{u} = B^{-\top}\mathbf{v} \neq \mathbf{0}$ ดังนั้น $\kappa(C) = \kappa'$
2. **นอร์มพลังงานไม่เปลี่ยน.** ให้ $\mathbf{e} = \mathbf{x} - \mathbf{x}^*$ และ $\boldsymbol{\varepsilon} = B^{\top}\mathbf{e}$ จะได้ $\boldsymbol{\varepsilon}^{\top}C\boldsymbol{\varepsilon} = \mathbf{e}^{\top}BB^{-1}AB^{-\top}B^{\top}\mathbf{e} = \mathbf{e}^{\top}A\mathbf{e}$ นั่นคือ $\|\boldsymbol{\varepsilon}\|_C = \|\mathbf{e}\|_A$ และ $\mathbf{y}^* = B^{\top}\mathbf{x}^*$ เป็นคำตอบของ $C\mathbf{y} = \mathbf{h}$
3. **ใช้ทฤษฎีบทเชบีเชฟ.** จากทฤษฎีบท 11.3.1 $\boldsymbol{\varepsilon}_k = B^{\top}\mathbf{e}_k$ คือค่าคลาดเคลื่อนของ CG บน C ทฤษฎีบท 11.2.2 ใช้กับ C ได้ $\|\boldsymbol{\varepsilon}_k\|_C \le 2\rho'^k\|\boldsymbol{\varepsilon}_0\|_C$ เมื่อ $\rho' = \frac{\sqrt{\kappa'}-1}{\sqrt{\kappa'}+1}$
4. **กลับไปที่ $\mathbf{x}$.** จากข้อ 2 ได้ $\|\mathbf{e}_k\|_A = \|\boldsymbol{\varepsilon}_k\|_C$ และ $\|\mathbf{e}_0\|_A = \|\boldsymbol{\varepsilon}_0\|_C$ จึงได้ผลที่ต้องการ

สรุป: ขอบเขตของ CG ใช้กับ PCG ได้ทันที โดยแทน κ ด้วย κ' ของ $M^{-1}A$ □

ข้อควรระวัง: ขอบเขตใช้ κ' ของ $M^{-1}A$ ไม่ใช่ κ ของ A

ถ้า M ถูกเลือกไม่ดี κ' อาจใหญ่กว่า κ ของ A ได้เช่นกัน และถ้า κ' ไม่รู้ล่วงหน้า ขอบเขตก็ให้คำตอบได้เพียงประมาณ การใช้ขอบเขตจึงต้องวัด κ' หรือประมาณค่าให้ดีก่อน

ฉบับร่าง · Preprint · 2026-10-23

ตัวปรับสภาพที่ดีต้องทำสองอย่าง

จากทฤษฎีบทข้างบน M ที่ดีต้องให้ $M^{-1}A$ มีค่าลักษณะเฉพาะรวมกันและ κ' เล็ก และต้องแก้ระบบ $M\mathbf{z} = \mathbf{r}$ ได้ถูกลงทุนไม่มาก สองข้อนี้ขัดกันเสมอ ถ้า $M = A$ จะได้ $\kappa' = 1$ แต่แก้ระบบ $M\mathbf{z} = \mathbf{r}$ ยากเท่าเดิม ถ้า $M = I$ แก้ง่ายที่สุด แต่ไม่ช่วยอะไร วิธีที่ดีอยู่ระหว่างสองขั้วนี้ Iserles บอกว่า ตัวปรับสภาพที่ดีจะ *บีบ* ค่าลักษณะเฉพาะให้รวมกัน Iserles 2009, §14.3, p. 324 และ Ern–Guermond เรียกการเลือกนี้ว่าการประนีประนอมระหว่างต้นทุนต่อรอบกับอัตราการลู่เข้า Ern and Guermond 2004, §9.3.5, p. 411

เกร็ดเพิ่มเติม (ผู้เรียบเรียง): ทำให้ $A - M$ เล็กไม่ใช่เป้าหมาย

การทำให้ $A - M$ เล็กอาจพาไปผิดทาง Iserles 2009, §14.3, p. 323 เพราะถ้าคูณ M ด้วยจำนวนบวกใด ๆ ปริภูมิไครลอฟ $\mathcal{K}_m(AM^{-1}, \mathbf{b})$ ยังเป็นปริภูมิเดิม แต่ค่า $A - M$ เปลี่ยนไป ดังนั้นเราควรดู κ' และต้นทุน ไม่ใช่ความต่างของเมทริกซ์โดยตรง

ตัวปรับสภาพสามแบบ

ต่อไปนี้คือตัวปรับสภาพสามแบบที่ใช้กันมาก เรียงจากง่ายที่สุดไปถึงแม่นกว่า สำหรับทุกแบบ M สมมาตรบวกแน่นอนเมื่อ A เป็นเช่นนั้น

- **ตัวปรับสภาพแบบจาโคบี** (Jacobi) คือ $M = \operatorname{diag}(A)$ การแก้ $M\mathbf{z} = \mathbf{r}$ คือการหารแต่ละสมาชิกด้วยสมาชิกแนวทแยง ค่าใช้จ่ายต่อรอบเพิ่มเพียง $O(n)$

- **SOR สมมาตร** (symmetric SOR, SSOR) ที่ $\omega = 1$ คือ $M = (D + L)D^{-1}(D + L^{\top})$ เมื่อ $A = D + L + L^{\top}$ โดย D เป็นแนวทแยง และ L เป็นส่วนสามเหลี่ยมล่างที่ไม่รวมแนวทแยง Ern and Guermond 2004, §9.3.5, p. 411 การแก้ระบบกับ M ทำได้ด้วยการแทนค่าไปหน้าแล้วถอยหลัง ตัวปรับสภาพแบบนี้ใช้รูปแบบทั่วไปของ SSOR ที่มี ω ได้ด้วย Braess 2007, §IV.4, p. 213

- **การแยกตัวโชเลสกีแบบไม่สมบูรณ์** (IC(0)) คือ $M = LL^{\top}$ โดย L เป็นสามเหลี่ยมล่างที่มีสมาชิกไม่เป็นศูนย์เฉพาะที่ที่ A ไม่เป็นศูนย์ คำนวณดังนี้

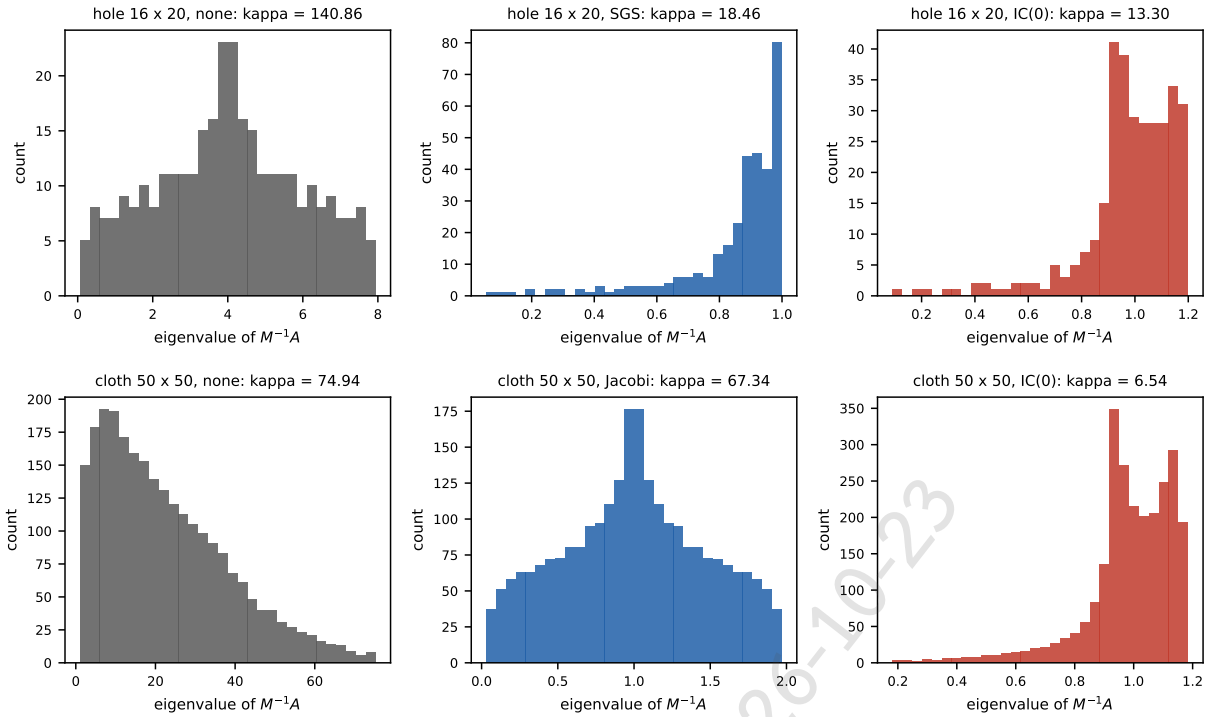
$$L_{ii} = \sqrt{a_{ii} - \sum_{k<i} L_{ik}^2}, \qquad L_{ij} = \frac{1}{L_{jj}}\Big(a_{ij} - \sum_{k<j} L_{ik}L_{jk}\Big) \quad (j < i,\ a_{ij} \neq 0),$$

  และตั้ง $L_{ij} = 0$ เมื่อ $a_{ij} = 0$ การกระทำนี้ตัดสมาชิกที่จะเกิดขึ้นใหม่ (fill-in) ทิ้ง Braess 2007, §IV.4, p. 214 สำหรับห้าจุดบนกริดสี่เหลี่ยม $M = (L_A + D)D^{-1}(L_A^{\top} + D)$ โดย L_A คือส่วนสามเหลี่ยมล่างของ A Vorst 2003, §13.2.1, p. 184 การแก้ระบบกับ M ใช้การแทนค่าสองครั้ง จึงใช้ต้นทุนตามจำนวนสมาชิกที่ไม่เป็นศูนย์ของ L

ข้อควรระวัง: IC(0) ไม่ได้แยก A จริง

$LL^{\top}$ ไม่เท่ากับ A ทุกตำแหน่ง เพราะตัด fill-in ทิ้ง มันจึงเป็นเพียงตัวประมาณ ถ้า A ไม่ใช่เมทริกซ์แบบ M อาจมีตัวหารเป็นศูนย์หรือลบระหว่างทางได้ ซึ่งโค้ดของเราจะหยุดด้วยข้อความเตือน สำหรับ**เมทริกซ์เอ็ม** (M-matrix) คือเมทริกซ์ที่ไม่เอกฐาน มีสมาชิกนอกแนวทแยงไม่เป็นบวก และมีอินเวอร์สที่สมาชิกไม่เป็นลบ การแยกแบบไม่สมบูรณ์มีอยู่เสมอสำหรับเมทริกซ์ชนิดนี้ Vorst 2003, Thm 13.3, p. 181 เมทริกซ์ห้าจุดและ $I + \gamma K$ ของเราเป็นเมทริกซ์เอ็ม และโค้ดไม่พบเหตุนั้น

ฉบับร่าง Preprint · 2026-10-23



รูปที่ 11.6: ฮิสโทแกรมของค่าลักษณะเฉพาะของ $M^{-1}A$ บนสองระบบ บนสุด: รอยขีดข่วนของเรา 16×20 (ไม่มีตัวปรับสภาพ, SOR สมมาตร, IC(0)) ล่างสุด: ผ้าจำลองขนาด 50×50 จุด (ไม่มีตัวปรับสภาพ, จาโคบี, IC(0)) ค่าคอนดิชันนัมเบอร์แสดงในหัวข้อของแต่ละกราฟ ยิ่งค่าลักษณะเฉพาะรวมตัวใกล้กันมากขึ้น κ' ยิ่งเล็ก (สร้างจาก `numanal/ch11.py`)

ตารางที่ 11.2: รอยขีดข่วน 16×20 ของหัวข้อ 11.1 เมื่อเลือกตัวปรับสภาพต่างกัน (สร้างจาก `numanal/ch11.py` และตรวจด้วย `test_ch11.py`)

ตัวปรับสภาพ	รอบ	κ'
ไม่มี (CG)	55	140.86
จาโคบี	55	140.86
SOR สมมาตร	21	18.46
IC(0)	18	13.30

ฉบับร่าง · preprint · 2026-10-23

ในรอยขีดข่วนของเรา ตัวปรับสภาพจาโคบีไม่เปลี่ยนอะไรเลย เพราะสมาชิกแนวทแยงของห้าจุดคือ 4 ทุกตำแหน่ง ดังนั้น $M = 4I$ และ PCG เท่ากับ CG เดิมทุกรอบ ซึ่งยืนยันด้วยการทดสอบ ส่วนนี้เป็นข้อสังเกตที่ทำให้เห็นว่าจาโคบีช่วยได้เฉพาะเมื่อแนวทแยงแปรผัน

การวัดบนรอยขีดข่วนและบนผ้า

ตารางต่อไปนี้ใช้จำนวนรอบจนค่าตกค้างเหลือ 10^{-6} เท่าของ $\|\mathbf{b}\|$ และ κ' ของ $M^{-1}A$ ระบบแรกคือรอยขีดข่วนของหัวข้อ 11.1 (ระบบ 320 ค่าที่ไม่รู้ และเริ่มจาก $\mathbf{x}_0 = \mathbf{0}$)

ผ้าในเกม (ระบบเดิมทุกเฟรม). Baraff และ Witkin ตีพิมพ์ในปี 1998 ว่าการจำลองผ้าที่ก้าวเวลาใหญ่ต้องแก้ระบบเชิงเส้นขนาดใหญ่ทุกก้าว และใช้วิธีเกรเดียนต์สังยุคที่ดัดแปลงเพื่อแก้ระบบนั้น Baraff and Witkin 1998, บทคัดย่อ เราจะไม่ทำซ้ำวิธีดัดแปลงของเขา แต่จะลองกับผ้าจำลองของเราเอง (สถานการณ์สมมติ) ผ้าเป็นตาราง 50×50 จุด จึงมี $2\,500$ ค่าที่ไม่รู้ สปริงระหว่างจุดข้างเคียงแข็งขึ้นจากซ้ายไปขวาจาก 1 ถึง 10 แถวบนสุดยึดกับกรอบ และเมทริกซ์ของทุกเฟรมคือ $A = I + K$ โดย K คือความแข็งของสปริงเหล่านี้ เราแก้ระบบ 60 เฟรม ภาระแปรไปทุกเฟรม และเริ่มจากคำตอบของเฟรมก่อนหน้า (เหมือนที่ตัวจำลองทำ

ตารางที่ 11.3: ผ้าจำลอง 50×50 จุด 60 เฟรม ค่าในวงเล็บคือค่าสูงสุดในทุกเฟรม (สร้างจาก `numanal/ch11.py` และตรวจด้วย `test_ch11.py`)

ตัวปรับสภาพ	κ'	รอบต่อเฟรม (เฉลี่ย)	(สูงสุด)
ไม่มี (CG)	74.94	45.9	(57)
จาโคบี	67.34	46.3	(56)
IC(0)	6.54	13.9	(17)

จริง)

อ่านตารางนี้ได้สองข้อ ข้อแรก จาโคบีแทบไม่ช่วยผ้านี้ (κ' ลดจาก 74.94 เหลือ 67.34) ข้อสอง IC(0) ลดจำนวนรอบเฉลี่ยลงราว 3.3 เท่า โดยที่ κ' ลดจาก 74.94 เหลือ 6.54 ตัวเลขเหล่านี้เป็นของโค้ดของเรากับผ้าจำลองของเรา ไม่ได้บอกว่าเกมหรือเอนจินใดทำแบบนี้จริง

สรุปใจความ

- การเปลี่ยนตัวแปรด้วย $M = BB^{\top}$ ทำให้หุบกลมขึ้น และ PCG คือ CG บนระบบใหม่นั้นทุกประการ
- ตัวปรับสภาพที่ดีต้องให้ค่าลักษณะเฉพาะรวมตัวกัน และแก้ระบบ $M\mathbf{z} = \mathbf{r}$ ได้ถูกลง
- จาโคบีไม่ช่วยระบบห้าจุดของเรา ส่วน SOR สมมาตรและ IC(0) ช่วยมากกว่า และสำหรับผ้าจำลอง IC(0) ช่วยมากที่สุดที่เราทดสอบ
- กลับไปที่ผ้า: IC(0) ลดรอบเฉลี่ยลงราว 3.3 เท่า แต่แต่ละรอบมีต้นทุนมากกว่า และการแยกตัวประกอบต้องทำซ้ำทุกครั้งที่เมทริกซ์เปลี่ยน

อ่านต่อ. หัวข้อ 11.4 จะดูระบบที่ A ไม่สมมาตร ซึ่งใช้ PCG ไม่ได้โดยตรง อ่านเพิ่มเกี่ยวกับตัวปรับสภาพได้ใน Vorst 2003, §13.1--13.2, pp. 173--187 และ Braess 2007, §IV.4, pp. 210--220 และ Ern and Guermond 2004, §9.3.5, pp. 410--412

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 11.3.1 (⋆, ทำด้วยมือ) ให้ $A = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$ และ $M = \operatorname{diag}(4, 3)$

1. จงเขียน $M^{-1}A$ และหาค่าลักษณะเฉพาะของมัน
2. จงหาคอนดิชันนัมเบอร์ κ' ของ $M^{-1}A$ และ κ ของ A แล้วเปรียบเทียบ การปรับสภาพด้วยจาโคบีช่วยมากหรือน้อยในระบบเล็กนี้

แบบฝึกหัด 11.3.2 (⋆⋆, พิสูจน์) ให้ A สมมาตรบวกแน่นอนและ $M = LL^{\top}$ โดย L ไม่เอกฐาน ให้ $C = L^{-1}AL^{-\top}$ จงแสดงว่า C สมมาตรบวกแน่นอน และค่าลักษณะเฉพาะของ $M^{-1}A$ เป็นจำนวนจริงบวกและเท่ากับค่าลักษณะเฉพาะของ C

ตัวช่วยคิด

เขียน $M^{-1}A = L^{-\top}CL^{\top}$ แล้วใช้ว่าเมทริกซ์คล้ายมีค่าลักษณะเฉพาะเดียวกัน

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 11.3.3 (★★, พิสูจน์) ให้ A สมมาตรบวกแน่นอน และให้ $\mathbf{x}_k$ เป็นค่าของ PCG ในนิยามของหัวข้อ 11.3

1. ถ้า $M = A$ จงแสดงว่า $\mathbf{x}_1 = \mathbf{x}^*$ สำหรับทุก $\mathbf{x}_0$
2. ถ้าแทน M ด้วย cM เมื่อ $c > 0$ จงแสดงว่าทุก $\mathbf{x}_k$ ไม่เปลี่ยน

แบบฝึกหัด 11.3.4 (★★, คำนวณและอธิบาย) ให้ $A = \operatorname{tridiag}(-1, 2, -1)$ ขนาด 100×100

1. จงแสดงว่าตัวประกอบ L ของการแยกโชเลสกีเต็มของ A เป็นสามเหลี่ยมล่างแบบแถบคู่ และไม่มีสมาชิกเกิดใหม่ จึงสรุปได้ว่า IC(0) ของ A เท่ากับการแยกโชเลสกีเต็ม
2. จงอธิบายว่าทำไม PCG ที่ใช้ IC(0) ของ A จึงลู่เข้าในหนึ่งรอบ
3. จงอธิบายว่า $M = \operatorname{diag}(A)$ ให้ผลเหมือนวิธีเกรเดียนต์สังยุคเดิมหรือไม่ เพราะเหตุใด

แบบฝึกหัด 11.3.5 (★★★, พิสูจน์) ให้ $A = D + L + L^{\top}$ สมมาตรบวกแน่นอน โดย D เป็นแนวทแยง (มีสมาชิกบวก) และ L เป็นสามเหลี่ยมล่างที่ไม่รวมแนวทแยง ให้ $M = (D + L)D^{-1}(D + L^{\top})$ (SOR สมมาตรที่ $\omega = 1$)

1. จงแสดงว่า $M - A = L D^{-1} L^{\top}$
2. จงแสดงว่า M สมมาตรบวกแน่นอน

> **ตัวช่วยคิด**
>
> กระจายผลคูณของ M แล้วใช้ว่า D^{-1} เป็นบวกแน่นอน

แบบฝึกหัด 11.3.6 (★★★, คำนวณและตีความ; กลับไปที่สถานการณ์) ใช้ระบบรอยขีดข่วน 16×20 ของหัวข้อ 11.1 (ใช้เมทริกซ์และ $\mathbf{b}$ จาก `fill_problem` ของโค้ด) และเริ่มจาก $\mathbf{x}_0 = \mathbf{0}$

1. หาจำนวนรอบจนค่าตกค้างเหลือ $10^{-6}\|\mathbf{b}\|$ และค่า κ' สำหรับ $M = I$, $M = \operatorname{diag}(A)$, SOR สมมาตร และ IC(0)
2. อธิบายว่าทำไมจาโคบีให้ผลเหมือนไม่มีตัวปรับสภาพ
3. ถ้าใช้ขอบเขต $\lceil \frac{1}{2}\sqrt{\kappa'}\ln\frac{2}{\varepsilon} \rceil$ ของทฤษฎีบท 11.3.2 กับ IC(0) และ $\varepsilon = 10^{-6}$ จงหาจำนวนรอบที่ขอบเขตรับประกันและเปรียบเทียบกับข้อ 1 และอธิบายว่าเหตุใดทั้งสองตัวเลขต่างกัน

แบบฝึกหัด 11.3.7 (★★★, ตีความ; กลับไปที่ผ้า) จากตารางผ้าจำลองของหัวข้อ 11.3 (CG เฉลี่ย 45.9 รอบต่อเฟรม และ IC(0) เฉลี่ย 13.9 รอบต่อเฟรม)

1. จงคำนวณอัตราส่วนของจำนวนรอบเฉลี่ย
2. รอบของ IC(0) หนึ่งรอบมีการคูณ A กับเวกเตอร์หนึ่งครั้ง และการแทนค่าไปหน้าและถอยหลังกับ L ประมาณการเปรียบเทียบจำนวนการคูณต่อรอบของ IC(0) กับ CG เดิม จงประมาณว่า IC(0) ยังคุ้มหรือไม่ เมื่อเมทริกซ์ไม่เปลี่ยนระหว่างเฟรม
3. ถ้าเมทริกซ์เปลี่ยนทุกเฟรม ต้องแยกตัวประกอบ IC(0) ใหม่ทุกเฟรม จงอธิบายว่าการคิดต้นทุนเปลี่ยนไปอย่างไร และทำไมการเริ่มจากคำตอบของเฟรมก่อนช่วยลดจำนวนรอบ

ร่างเพื่อการตรวจทาน · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 11.3.1.

1. $M^{-1}A = \operatorname{diag}(\frac{1}{4}, \frac{1}{3}) \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{4} \\ \frac{1}{3} & 1 \end{pmatrix}$ ค่าลักษณะเฉพาะเป็นคำตอบของ $(1-\lambda)^2 = \frac{1}{12}$ คือ $\lambda = 1 \pm \frac{1}{\sqrt{12}}$
2. $\kappa' = \dfrac{1+1/\sqrt{12}}{1-1/\sqrt{12}} = \dfrac{13+4\sqrt{3}}{11} \approx 1.81165$ และ $\kappa = \dfrac{7+\sqrt{5}}{7-\sqrt{5}} = \dfrac{27+7\sqrt{5}}{22} \approx 1.93875$ การปรับด้วยจาโคบีลด κ ลงเล็กน้อยเพราะระบบ 2×2 นี้กลมอยู่แล้ว ค่าของสิ่งที่ปรับได้มีไม่มาก

ตรวจด้วยโค้ด: `test_ch11.py`
(`test_exercise_11_3_1_jacobi_two_by_two`)

เฉลย 11.3.2. ให้ $C = L^{-1}AL^{-\top}$ ตรวจว่า $C^\top = L^{-1}A^\top L^{-\top} = C$ จึงสมมาตร และสำหรับ $\mathbf{v} \neq \mathbf{0}$ ให้ $\mathbf{u} = L^{-\top}\mathbf{v} \neq \mathbf{0}$ จะได้ $\mathbf{v}^\top C\mathbf{v} = \mathbf{u}^\top A\mathbf{u} > 0$ จึงบวกแน่นอน ต่อไป $M^{-1}A = L^{-\top}L^{-1}A = L^{-\top}\big(L^{-1}AL^{-\top}\big)L^\top = L^{-\top}CL^\top$ เป็นเมทริกซ์คล้ายกับ C จึงมีค่าลักษณะเฉพาะเดียวกัน และเป็นจำนวนจริงบวกทั้งหมดเพราะ C บวกแน่นอน

เฉลย 11.3.3.

1. เมื่อ $M = A$ ได้ $\mathbf{z}_0 = A^{-1}\mathbf{r}_0$ และ $\mathbf{p}_0 = \mathbf{z}_0$ จึง $\mathbf{p}_0^\top A\mathbf{p}_0 = \mathbf{r}_0^\top A^{-1}\mathbf{r}_0 = \mathbf{r}_0^\top \mathbf{z}_0$ ดังนั้น $\alpha_0 = 1$ และ $\mathbf{x}_1 = \mathbf{x}_0 + A^{-1}(\mathbf{b} - A\mathbf{x}_0) = A^{-1}\mathbf{b} = \mathbf{x}^*$
2. แทน M ด้วย cM จะได้ $\mathbf{z}_k \mapsto \mathbf{z}_k/c$ ดังนั้น β_k ไม่เปลี่ยน และ $\mathbf{p}_k \mapsto \mathbf{p}_k/c$ โดยอุปนัย เพราะ $\mathbf{p}_{k+1} = \mathbf{z}_{k+1} + \beta_k\mathbf{p}_k$ ตัวเศษของ α_k ลดลง c เท่า และตัวส่วนลดลง c^2 เท่า จึง $\alpha_k \mapsto c\,\alpha_k$ และ $\alpha_k\mathbf{p}_k$ ไม่เปลี่ยน ดังนั้น $\mathbf{x}_k$ และ $\mathbf{r}_k$ ไม่เปลี่ยนทุกรอบ

ตรวจด้วยโค้ด: `test_ch11.py`
(`test_pcg_with_M_equal_to_A_is_one_step`,
`test_scaling_M_does_not_change_the_iterates`)

เฉลย 11.3.4.

1. ถ้า L เป็นแถบคู่ล่าง สมาชิก $L_{i+1,i} = a_{i+1,i}/L_{ii} = -1/L_{ii}$ และ $L_{ii}^2 = a_{ii} - L_{i,i-1}^2$ ไม่มีการคูณใดที่ทำให้เกิดสมาชิกนอกแถบ จึงไม่มี fill-in IC(0) ซึ่งตัดเฉพาะ fill-in ทิ้ง จึงไม่ตัดอะไรเลย และ $LL^\top = A$
2. จาก $M = A$ และข้อ 3 ของฝึกหัด 11.3.3 ได้ $\mathbf{x}_1 = \mathbf{x}^*$ ดังนั้นลู่เข้าในหนึ่งรอบ
3. $\operatorname{diag}(A) = 2I$ จึง $M = 2I$ ซึ่งเป็นค่าคงที่คูณ I ตามข้อ 2 ของฝึกหัด 11.3.3 ดังนั้น PCG เท่ากับวิธีเกรเดียนต์สังยุคเดิมทุกรอบ

ตรวจด้วยโค้ด: `test_ch11.py`
(`test_ic0_is_exact_cholesky_for_a_tridiagonal_matrix_and_one_step`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 11.3.5.

1. กระจายผลคูณ $(D+L)D^{-1}(D+L^\top) = DD^{-1}D + DD^{-1}L^\top + LD^{-1}D + LD^{-1}L^\top = D + L^\top + L + LD^{-1}L^\top$ จึง $M - A = LD^{-1}L^\top$
2. M สมมาตรเพราะ $M^\top = (D+L)D^{-1}(D+L^\top)^\top$ ซึ่งเท่ากับ M เอง สำหรับ $\mathbf{x} \neq \mathbf{0}$ ได้ $\mathbf{x}^\top M\mathbf{x} = \mathbf{x}^\top A\mathbf{x} + \mathbf{w}^\top D^{-1}\mathbf{w}$ เมื่อ $\mathbf{w} = L^\top\mathbf{x}$ เทอมหลังไม่เป็นลบเพราะ D^{-1} มีสมาชิกบวก ดังนั้น $\mathbf{x}^\top M\mathbf{x} \geq \mathbf{x}^\top A\mathbf{x} > 0$

ตรวจด้วยโค้ด: `test_ch11.py` (`test_sgs_matrix_is_symmetric_positive_definite_and_has_the_formula`)

เฉลย 11.3.6.

1. จากตารางในหัวข้อ 11.3: ไม่มีตัวปรับสภาพ 55 รอบ และ $\kappa' = 140.86$; จาโคบี 55 รอบ และ $\kappa' = 140.86$; SOR สมมาตร 21 รอบ และ $\kappa' = 18.46$; IC(0) 18 รอบ และ $\kappa' = 13.30$
2. แนวทแยงของห้าจุดคือ 4 ทุกตำแหน่ง จึง $M = 4I$ และตามข้อ 2 ของฝึกหัด 11.3.3 รอบทุกรอบเหมือนเดิม
3. $\kappa' = 13.30$ ให้ $\frac{1}{2}\sqrt{\kappa'}\ln\frac{2}{10^{-6}} \approx 26.46$ จึงขอบเขตรับประกันคือ 27 รอบ ซึ่งมากกว่า 18 รอบที่วัดได้ เพราะขอบเขตเป็นกรณีเลวร้ายที่สุดสำหรับนอร์มพลังงานของความคลาดเคลื่อน ส่วนการวัดในข้อ 1 ใช้เกณฑ์ค่าตกค้าง ซึ่งคนละปริมาณกัน

ตรวจด้วยโค้ด: `test_ch11.py` (`test_hole_study_steps_and_kappa`, `test_exercise_11_3_6_bound_steps_for_ic0`)

เฉลย 11.3.7.

1. อัตราส่วน $45.9/13.9 \approx 3.30$
2. จากการนับโดยประมาณ CG เดิมใช้ $A\mathbf{p}$ (ประมาณ 5 การคูณต่อแถว) กับการดำเนินการเวกเตอร์ราว 5 ครั้งต่อแถว รวมประมาณ 10 ต่อแถว IC(0) ใช้เพิ่มการแทนค่าไปหน้าและถอยหลังกับ L ซึ่งมีราว 3 สมาชิกต่อแถว จึงเพิ่มราว 6 ต่อแถว รวมประมาณ 17 ต่อแถว คือราว 1.7 เท่า ซึ่งยังน้อยกว่าอัตราส่วนรอบ 3.3 ดังนั้นถ้าเมทริกซ์ไม่เปลี่ยนระหว่างเฟรม IC(0) ยังคุ้ม การนับนี้เป็นการประมาณ ควรวัดเวลาจริงก่อนสรุป
3. เมื่อเมทริกซ์เปลี่ยน ต้นทุนของการแยกตัวประกอบต้องนับเป็นต้นทุนของทุกเฟรม ซึ่งเทียบได้กับไม่กี่รอบของ CG เดิม การเริ่มจากคำตอบของเฟรมก่อนช่วยเพราะคำตอบใหม่อยู่ใกล้คำตอบเก่า ค่าตกค้างเริ่มต้นจึงเล็กลง และต้องใช้รอบน้อยกว่า

11.4 เมื่อภูมิประเทศไม่สมมาตร

สถานการณ์สมมติ: ควันที่ลมพัดไปทางขวา

ลองนึกถึงฉากควันในภาพจำลองที่มีลมพัดจากซ้ายไปขวา ควันไม่ได้ฟุ้งออกทุกทิศเท่ากัน ลมพาควันไปทางหนึ่ง และการแพร่ทำให้ควันกระจายด้วย เมื่อคำนวณความเข้มของควันทีละจุดบนกริด จุดหนึ่งจะรับผลจากเพื่อนบ้านด้านต้นลมกับปลายลมไม่เท่ากัน สมาชิกของเมทริกซ์ที่เชื่อมจุดกับเพื่อนบ้านสองฝั่งจึงต่างกัน ระบบที่ได้จึงไม่สมมาตร

วิธีเกรเดียนต์สังยุคในหัวข้อ 11.1 ใช้ไม่ได้โดยตรง เพราะต้องการให้ A สมมาตรบวกแน่นอน คำถามของหัวข้อนี้คือ จะแก้ระบบที่ไม่สมมาตรได้อย่างไร และควรเลือกวิธีไหน (ตัวอย่างนี้เป็นสถานการณ์สมมติ ไม่ได้บอกว่าเอนจินใดทำเช่นนี้)

ฉบับร่าง · Preprint · 2026-10-23

ภาพรวมก่อนอ่าน

หัวข้อนี้ใช้ความคิดเดิมอีกข้อหนึ่ง คือเลือกคำตอบจากปริภูมิไครลอฟที่ทำให้ค่าตกค้างมีนอร์มเล็กที่สุด ต่างจาก CG ซึ่งทำให้ความคลาดเคลื่อนในนอร์มพลังงานเล็กที่สุด เมื่อไม่มีการสังยุคแบบสั้น ความจำจะโตตามจำนวนรอบ เราจึงรีสตาร์ต ซึ่งต้องแลกกับอัตราการลู่เข้าที่คาดเดายาก

ทำไม CG จึงใช้ไม่ได้ตรง ๆ

ทฤษฎีบทความสัมพันธ์อาร์โนลดีในบทที่ 9 (ทฤษฎีบท 9.4.1) ให้ $A\mathbf{V}_k = \mathbf{V}_{k+1}\bar{H}_k$ เมื่อคอลัมน์ของ $\mathbf{V}_k$ ตั้งฉากกันเป็นฐานของ $\mathcal{K}_k(A,\mathbf{r}_0)$ และ $\bar{H}_k$ เป็นเมทริกซ์เฮสเซนเบิร์กบนขนาด $(k+1)\times k$ ถ้า A สมมาตร $\mathbf{V}_k^{\top}A\mathbf{V}_k$ จะสมมาตรด้วยและเป็นเมทริกซ์สามแนวทแยง นั่นคือเหตุที่ CG เก็บเวกเตอร์เพียงไม่กี่ตัว Vorst 2003, §6.1, p. 65

เมื่อ A ไม่สมมาตร $\bar{H}_k$ เต็มส่วนบน เวกเตอร์ใหม่แต่ละตัวต้องตั้งฉากกับเวกเตอร์เดิมทุกตัว ความจำและงานต่อรอบจึงโตตามจำนวนรอบ ซึ่งเป็นเหตุผลที่เราต้องรีสตาร์ต

นิยาม 11.4.1: ระบบลมพัดบนกริด

ให้ $m \geq 1$ ค่าที่ไม่รู้คือ u_{ij} ที่จุดภายใน $1 \leq i,j \leq m$ บนกริดของสี่เหลี่ยมจัตุรัสหนึ่งหน่วย โดยช่องกริดกว้าง $h = \frac{1}{m+1}$ และ $u = 0$ บนขอบ ให้ c เป็นความเร็วลมในทิศ x และให้ $p = \frac{ch}{2}$ ซึ่งเรียกว่า**เลขเปเคลต์** (Péclet number)ของช่องกริด บอกว่าลมพาควันไปไกลเท่าใดเทียบกับการแพร่ในหนึ่งช่อง เมื่อใช้ผลต่างกลางกับสมการ $-\Delta u + c\,\partial_x u = f$ แล้วคูณด้วย h^2 จะได้สมการที่จุด (i,j)

$$4u_{ij} - u_{i-1,j} - u_{i+1,j} - (1+p)\,u_{i,j-1} + (p-1)\,u_{i,j+1} = h^2 f_{ij}.$$

เรียงค่าที่ไม่รู้ตามแถวทีละแถว จะได้เมทริกซ์ A ขนาด $m^2 \times m^2$ และให้ $\mathbf{b} = \mathbf{1}$ ซึ่งเป็นแหล่งควันที่สม่ำเสมอ

อ่านแบบเข้าใจง่าย

สมาชิกสองตัวที่เกี่ยวกับเพื่อนบ้านซ้ายกับขวาคือ $-(1+p)$ และ $p-1$ ถ้า $p=0$ ทั้งคู่เป็น -1 และ A สมมาตร ซึ่งกลับไปเป็นสูตรห้าจุดของบทที่ 10 เมื่อ $p>0$ สมาชิกสองตัวนี้ไม่เท่ากัน จึงไม่สมมาตร ค่าลักษณะเฉพาะของเมทริกซ์นี้มีสูตรปิด และเป็นจำนวนจริงบวกเมื่อ $0<p<1$ (แบบฝึกหัด 11.4.7)

GMRES: ค่าตกค้างต่ำสุดในปริภูมิไครลอฟ

นิยาม 11.4.2: วิธีค่าตกค้างน้อยสุดนัยทั่วไป

ให้ A ไม่เอกฐาน $\mathbf{x}_0$ เป็นค่าเริ่มต้น และ $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$ ค่าประมาณลำดับที่ k ของ GMRES คือสมาชิก $\mathbf{x}_k$ ของ $\mathbf{x}_0 + \mathcal{K}_k(A,\mathbf{r}_0)$ ที่ทำให้ $\|\mathbf{b} - A\mathbf{x}_k\|_2$ เล็กที่สุด เมื่อทำ m ขั้นแล้วเริ่มรอบใหม่จาก $\mathbf{x}_m$ โดยใช้ $\mathbf{r} = \mathbf{b} - A\mathbf{x}_m$ เป็นค่าตกค้างเริ่มต้น เรียกว่า GMRES(m) ซึ่งเป็น**การรีสตาร์ต** (restart) ถ้าไม่รีสตาร์ตเลยเรียกว่า GMRES เต็มรูป

วิธี GMRES มาจากงานของ Saad and Schultz 1986

ขณะร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

GMRES ดูทุกคำตอบที่สร้างได้จากเวกเตอร์ที่คูณ A มาแล้ว แล้วเลือกอันที่ค่าตกค้างเล็กที่สุด CG เลือกจากปริภูมิเดียวกัน แต่วัดความคลาดเคลื่อนด้วยนอร์มพลังงาน ส่วน GMRES วัดค่าตกค้างด้วยนอร์มยูคลิด

ลองเอง: หนึ่งรอบ แล้วสองรอบบน $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$, $\mathbf{b} = \mathbf{e}_2$

เริ่มจาก $\mathbf{x}_0 = \mathbf{0}$ ได้ $\mathbf{r}_0 = \mathbf{e}_2$ รอบแรกค้นหาคำตอบในเส้น $\mathbf{x}_0 + \mathcal{K}_1 = \{t\,\mathbf{e}_2\}$ เท่านั้น สำหรับ $\mathbf{x} = t\,\mathbf{e}_2$ จะได้ $\mathbf{b} - A\mathbf{x} = (-2t,\, 1-t)^\top$ ซึ่งมีนอร์มกำลังสอง $4t^2 + (1-t)^2$ ค่านี้เล็กที่สุดเมื่อ $t = \frac{1}{5}$ จึง $\mathbf{x}_1 = (0, \frac{1}{5})^\top$ และค่าตกค้างมีนอร์ม $\frac{2}{\sqrt{5}} \approx 0.894$
รอบที่สอง $\mathcal{K}_2 = \operatorname{span}\{\mathbf{e}_2, (2,1)^\top\}$ เป็น $\mathbb{R}^2$ ทั้งหมด จึงได้คำตอบแม่นตรง $\mathbf{x}_2 = (-2, 1)^\top$ ซึ่งตรวจได้โดยแทนค่าในระบบ ทฤษฎีบทถัดไปให้วิธีคำนวณแบบเดียวกันด้วยกำลังสองน้อยที่สุด

ทฤษฎีบท 11.4.1: GMRES คือกำลังสองน้อยที่สุดบนเมทริกซ์เฮสเซนเบิร์ก

ให้ $\mathbf{V}_k$ และ $\bar{H}_k$ มาจากการวนซ้ำอาร์โนลดี k ขั้นจาก $\mathbf{r}_0$ (ไม่มีการหยุดกลางคัน) และให้ $\beta = \|\mathbf{r}_0\|_2$ แล้วสำหรับทุก $\mathbf{y} \in \mathbb{R}^k$

$$\|\mathbf{b} - A(\mathbf{x}_0 + \mathbf{V}_k\mathbf{y})\|_2 = \|\beta\mathbf{e}_1 - \bar{H}_k\mathbf{y}\|_2 .$$

ดังนั้น $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k\mathbf{y}_k$ โดย $\mathbf{y}_k$ เป็นคำตอบของกำลังสองน้อยที่สุด $\min_{\mathbf{y}} \|\beta\mathbf{e}_1 - \bar{H}_k\mathbf{y}\|_2$

พิสูจน์. **แนวคิด.** ค่าตกค้างของ $\mathbf{x}_0 + \mathbf{V}_k\mathbf{y}$ คือเวกเตอร์ฐานตั้งฉากคูณด้วยสัมประสิทธิ์ นอร์มของมันจึงเท่ากับนอร์มของสัมประสิทธิ์

1. **ค่าตกค้างเป็นผลต่าง.** เพราะ $\mathbf{b} - A\mathbf{x}_0 = \mathbf{r}_0$ ได้ $\mathbf{b} - A(\mathbf{x}_0 + \mathbf{V}_k\mathbf{y}) = \mathbf{r}_0 - A\mathbf{V}_k\mathbf{y}$
2. **ใช้ความสัมพันธ์อาร์โนลดี.** $\mathbf{r}_0 = \beta\mathbf{v}_1$ และ $A\mathbf{V}_k = \mathbf{V}_{k+1}\bar{H}_k$ จึง $\mathbf{r}_0 - A\mathbf{V}_k\mathbf{y} = \mathbf{V}_{k+1}(\beta\mathbf{e}_1 - \bar{H}_k\mathbf{y})$
3. **นอร์มไม่เปลี่ยน.** คอลัมน์ของ $\mathbf{V}_{k+1}$ ตั้งฉากเป็นคู่และยาวหนึ่ง จึง $\mathbf{V}_{k+1}^\top\mathbf{V}_{k+1} = I$ และ $\|\mathbf{V}_{k+1}\mathbf{z}\|_2 = \|\mathbf{z}\|_2$ สำหรับทุก $\mathbf{z}$
4. **สรุป.** เมื่อ $\mathbf{y}$ วิ่งผ่าน $\mathbb{R}^k$ คำตอบ $\mathbf{x}_0 + \mathbf{V}_k\mathbf{y}$ วิ่งผ่านทุกสมาชิกของ $\mathbf{x}_0 + \mathcal{K}_k(A, \mathbf{r}_0)$ การเลือก $\mathbf{y}$ ที่ทำกำลังสองน้อยที่สุดจึงเป็นการเลือก GMRES

สรุป: GMRES ทุกรอบคือกำลังสองน้อยที่สุดขนาด $(k+1) \times k$ ซึ่งแก้ได้เร็วเมื่อ k เล็ก □

ทฤษฎีบท 11.4.2: ค่าตกค้างของ GMRES ไม่เพิ่ม

สำหรับ GMRES เต็มรูป $\|\mathbf{b} - A\mathbf{x}_{k+1}\|_2 \le \|\mathbf{b} - A\mathbf{x}_k\|_2$ ทุก k และสำหรับ GMRES(m) ค่าตกค้างไม่เพิ่มทั้งในแต่ละรอบและข้ามรอบ

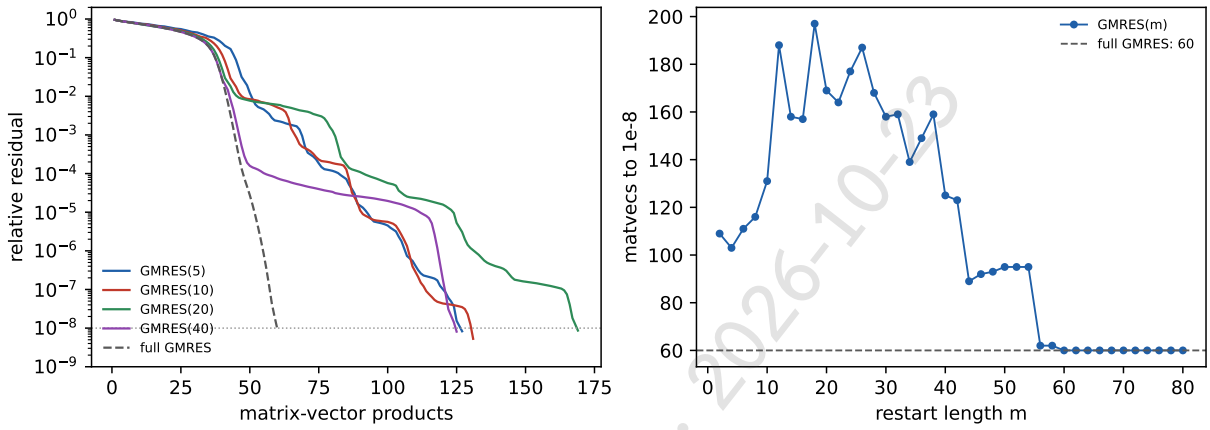
พิสูจน์. **แนวคิด.** ปริภูมิที่ใช้หาค่าต่ำสุดซ้อนกันมากขึ้น ค่าต่ำสุดบนปริภูมิใหญ่จึงไม่มากกว่า

1. **ปริภูมิซ้อนกัน.** $\mathcal{K}_k(A, \mathbf{r}_0) \subseteq \mathcal{K}_{k+1}(A, \mathbf{r}_0)$ เพราะ $\mathcal{K}_{k+1}$ มีทุกเวกเตอร์ของ $\mathcal{K}_k$ เป็นส่วนหนึ่ง ดังนั้น $\mathbf{x}_0 + \mathcal{K}_k \subseteq \mathbf{x}_0 + \mathcal{K}_{k+1}$

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 11.4: GMRES บนระบบลม 900 ค่าที่ไม่รู้ ($p = 0.9$, $\mathbf{b} = \mathbf{1}$) จำนวนการคูณเมทริกซ์กับเวกเตอร์จนค่าตกค้างสัมพัทธ์ไม่เกิน 10^{-8} (สร้างจาก `numanal/ch11.py` และตรวจด้วย `test_ch11.py`)

ความยาวรอบ m	จำนวนการคูณ
5	127
10	131
20	169
40	125
เต็มรูป (ไม่รีสตาร์ต)	60



รูปที่ 11.7: ซ้าย: ค่าตกค้างสัมพัทธ์ของ GMRES(m) ต่อจำนวนการคูณเมทริกซ์กับเวกเตอร์ ระบบลม 900 ค่าที่ไม่รู้ เส้นจุดคือ 10^{-8} ขวา: จำนวนการคูณที่ใช้ถึง 10^{-8} เทียบกับความยาวรอบ m เส้นประคือ GMRES เต็มรูป (สร้างจาก `numanal/ch11.py`)

2. **ค่าต่ำสุดบนเซตใหญ่.** $\mathbf{x}_{k+1}$ ทำให้ค่าตกค้างต่ำสุดบนเซตใหญ่ และ $\mathbf{x}_k$ เป็นสมาชิกของเซตนั้น จึง $\|\mathbf{b} - A\mathbf{x}_{k+1}\|_2 \leq \|\mathbf{b} - A\mathbf{x}_k\|_2$

3. **การรีสตาร์ต.** ปริภูมิของรอบใหม่ของ GMRES(m) มี $\mathbf{x}_m$ เป็นสมาชิก (เมื่อเลือกสัมประสิทธิ์ทั้งหมดเป็นศูนย์) ดังนั้นค่าตกค้างไม่เพิ่มข้ามรอบด้วย

สรุป: ค่าตกค้างไม่เพิ่ม แต่ทฤษฎีบทนี้ไม่ได้บอกว่าลดเร็วแค่ไหน □

ฉบับร่าง · Preprint · 2026-10-23

รีสตาร์ตเมื่อความจำไม่พอ

เมื่อ k โต ต้องเก็บเวกเตอร์ $\mathbf{V}_k$ ทั้งหมด ซึ่งใช้ nk ตัวเลข และการตั้งฉากต่อรอบก็โตตาม k จึงต้องจำกัดความยาวรอบ m VdV เตือนว่าไม่มีกฎง่าย ๆ สำหรับเลือก m และการลู่เข้าอาจเปลี่ยนมากเมื่อ m ต่างกันเพียงเล็กน้อย Vorst 2003, §6.1, p. 66 ตาราง 11.4 แสดงตัวอย่างจากระบบลม 900 ค่าที่ไม่รู้ ($m = 30$, $p = 0.9$, $\mathbf{b} = \mathbf{1}$, $\mathbf{x}_0 = \mathbf{0}$) เรานับการคูณ $A\mathbf{v}$ ทุกขั้นอาร์โนลดี และนับการคำนวณ $\mathbf{b} - A\mathbf{x}$ ทุกการรีสตาร์ต จนค่าตกค้างสัมพัทธ์ไม่เกิน 10^{-8}

เมื่อ m เพิ่มจาก 5 ถึง 20 จำนวนการคูณกลับเพิ่มขึ้น และเมื่อ $m = 40$ ก็ลดลงอีก ดังนั้น GMRES(20) ต้องใช้มากกว่าทั้ง GMRES(5) และ GMRES(40) รูป 11.7 ทางซ้ายแสดงเส้นค่าตกค้างของแต่ละ m และทางขวาแสดงจำนวนการคูณเมื่อ m เปลี่ยนจาก 2 ถึง 80 ตัวเลขเหล่านี้เป็นของโค้ดของเราบนระบบนี้ ไม่ใช่ทฤษฎีบททั่วไป

เมื่อค่าตกค้างไม่ลดเลย

ตัวอย่างต่อไปนี้มีคอนดิชันนัมเบอร์เท่ากับ 1 แต่ GMRES ยังไม่ลดค่าตกค้างเลยจนถึงรอบที่ n ให้ A เป็น**เมทริกซ์เลื่อนวงกลม** (cyclic shift matrix) ซึ่งมีคอลัมน์ที่ j เป็น $\mathbf{e}_{j+1}$ สำหรับ $j < n$ และคอลัมน์สุดท้ายเป็น $\mathbf{e}_1$ (นั่นคือ $A\mathbf{e}_j = \mathbf{e}_{j+1}$ โดยดัชนีบวกหนึ่งแบบวนรอบ)

ทฤษฎีบท 11.4.3: GMRES หยุดนิ่งก่อนรอบที่ n แล้วตอบในรอบที่ n

ให้ $n \geq 2$ A เป็นเมทริกซ์เลื่อนวงกลม $\mathbf{b} = \mathbf{e}_1$ และ $\mathbf{x}_0 = \mathbf{0}$ แล้ว

(i) $\kappa(A) = 1$;

(ii) $\|\mathbf{r}_k\|_2 = 1$ สำหรับ $k = 0, 1, \ldots, n-1$;

(iii) $\mathbf{x}_n = \mathbf{e}_n$ และ $\mathbf{r}_n = \mathbf{0}$.

ดังนั้น GMRES(m) กับ $m < n$ ไม่มีความคืบหน้าเลยในทุกรอบ (ในเลขคณิตแม่นตรง)

พิสูจน์. **แนวคิด.** ทุกคำตอบใน $\mathcal{K}_k$ ส่งเวกเตอร์ไปที่ปริภูมิที่ตั้งฉากกับ $\mathbf{e}_1$ เสมอ จึงแพ้การเลือกศูนย์

1. **คอนดิชันนัมเบอร์.** คอลัมน์ของ A เป็นเวกเตอร์หนึ่งหน่วยตั้งฉากกันเป็นคู่ ดังนั้น $A^\top A = I$ จึง $\|A\|_2 = \|A^{-1}\|_2 = 1$ และ $\kappa(A) = 1$
2. **ปริภูมิไครลอฟ.** $A^j\mathbf{e}_1 = \mathbf{e}_{j+1}$ สำหรับ $0 \leq j \leq n-1$ จึง $\mathcal{K}_k(A, \mathbf{e}_1) = \operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ สำหรับ $k \leq n$
3. **ค่าตกค้างคงที่.** สำหรับ $k \leq n-1$ และ $\mathbf{x} \in \mathcal{K}_k$ จะได้ $A\mathbf{x} \in \operatorname{span}\{\mathbf{e}_2, \ldots, \mathbf{e}_{k+1}\}$ ซึ่งตั้งฉากกับ $\mathbf{e}_1$ ดังนั้น $\|\mathbf{e}_1 - A\mathbf{x}\|_2^2 = 1 + \|A\mathbf{x}\|_2^2 \geq 1$ และเท่ากับ 1 เมื่อ $A\mathbf{x} = \mathbf{0}$ นั่นคือ $\mathbf{x} = \mathbf{0}$ เพราะ A ไม่เอกฐาน ค่าต่ำสุดจึงได้ที่ $\mathbf{x}_k = \mathbf{0}$ และ $\|\mathbf{r}_k\|_2 = 1$
4. **รอบที่ n.** $\mathcal{K}_n = \mathbb{R}^n$ และ $A\mathbf{e}_n = \mathbf{e}_1 = \mathbf{b}$ ดังนั้น $\mathbf{x}_n = \mathbf{e}_n$ ทำให้ค่าตกค้างเป็นศูนย์ และคำตอบนี้เป็นค่าเดียวเพราะ A ไม่เอกฐาน
5. **รีสตาร์ต.** ถ้า $m < n$ ทุกรอบเริ่มจาก $\mathbf{x} = \mathbf{0}$ และปริภูมิของรอบคือ $\operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ ซึ่งเป็นกรณีที่ข้อ 3 ให้ $\mathbf{x} = \mathbf{0}$ อีกครั้ง

สรุป: คอนดิชันนัมเบอร์เท่ากับ 1 ไม่ได้การันตีว่า GMRES จะลู่เข้าเร็ว □

ข้อควรระวัง: $\kappa = 1$ ก็ยังหยุดนิ่งได้

คอนดิชันนัมเบอร์และค่าลักษณะเฉพาะเพียงอย่างเดียวจึงไม่พอที่จะบอกอัตราของ GMRES VdV ใช้ตัวอย่างนี้เป็นแบบฝึกหัด 6.11 และสรุปว่า การลู่เข้าของ GMRES อธิบายด้วยคอนดิชันนัมเบอร์หรือค่าลักษณะเฉพาะเพียงอย่างเดียวไม่ได้ Vorst 2003, §6.2, p. 77 เราพิสูจน์ทฤษฎีบทนี้เองข้างบน ส่วนทฤษฎีบท 6.1 ของ VdV ให้ว่า ถ้า $f_0 \geq f_1 \geq \cdots \geq f_{n-1} > 0$ เป็นลำดับจำนวนจริงบวกที่ไม่เพิ่ม ยาว n พจน์ และ $\lambda_1, \ldots, \lambda_n$ เป็นจำนวนเชิงซ้อนที่ไม่เป็นศูนย์ n ตัว จะมีเมทริกซ์ A ที่มีค่าลักษณะเฉพาะ λ_j และ $\mathbf{b}$ ที่ $\|\mathbf{b}\|_2 = f_0$ ซึ่งทำให้ค่าตกค้างของ GMRES (เริ่มจาก $\mathbf{x}_0 = \mathbf{0}$) มี $\|\mathbf{r}_k\|_2 = f_k$ สำหรับ $k = 0, \ldots, n-1$ (ไม่พิสูจน์ที่นี่) ดูเพิ่มบทที่ 6 ทฤษฎีบท 6.1 ของ van der Vorst

ฉบับร่าง Preprint · 2026-10-23

GMRES เต็มรูปจบภายใน n รอบ

ทฤษฎีบท 11.4.4: GMRES เต็มรูปให้คำตอบแม่นตรงภายใน n รอบ

ให้ A ไม่เอกฐาน แล้วมีจำนวนรอบ $d \leq n$ ที่ GMRES เต็มรูปให้ค่าตกค้าง $\mathbf{r}_d = \mathbf{0}$ (ในเลขคณิตแม่นตรง)

พิสูจน์. **แนวคิด.** ถ้าปริภูมิไครลอฟหยุดขยาย A จะส่งมันเข้าหาตัวมันเอง และเพราะ A ไม่เอกฐาน ค่าตกค้างจึงเป็นผลของ A จากภายในปริภูมินั้น

1. **จุดหยุด.** ให้ d เป็นจำนวนเต็มที่น้อยที่สุดที่ $A^d\mathbf{r}_0 \in \mathcal{K}_d(A, \mathbf{r}_0)$ เวกเตอร์ $\mathbf{r}_0, A\mathbf{r}_0, \ldots, A^{d-1}\mathbf{r}_0$ เป็นอิสระเชิงเส้นตามความน้อยที่สุดของ d ดังนั้น $d \leq n$
2. **A ส่งปริภูมิเข้าหาตัวมันเอง.** $A\mathcal{K}_d \subseteq \mathcal{K}_d$ เพราะ $A(A^j\mathbf{r}_0) = A^{j+1}\mathbf{r}_0$ อยู่ใน $\mathcal{K}_d$ เมื่อ $j+1 < d$ และเมื่อ $j = d-1$ เป็นไปตามข้อ 1 เพราะ A ไม่เอกฐาน จึง $A\mathcal{K}_d = \mathcal{K}_d$
3. **คำตอบอยู่ในปริภูมิ.** เนื่องจาก $\mathbf{r}_0 \in \mathcal{K}_d = A\mathcal{K}_d$ มี $\mathbf{z} \in \mathcal{K}_d$ ที่ $A\mathbf{z} = \mathbf{r}_0$ และ $\mathbf{x}_0 + \mathbf{z}$ ทำให้ค่าตกค้างเป็นศูนย์
4. **สรุป.** GMRES เต็มรูปที่รอบ $d \leq n$ อยู่ในปริภูมิ $\mathbf{x}_0 + \mathcal{K}_d$ ซึ่งมีค่าตกค้างต่ำสุดเป็นศูนย์

สรุป: GMRES เต็มรูปจบภายในมิติของระบบ เหมือน CG ในหัวข้อ 11.1 □

ข้อควรระวัง: จบภายใน n รอบใช้ได้เฉพาะเลขคณิตแม่นตรง

ถ้า $n = 900$ การรอกระทั่ง 900 รอบไม่ใช่ทางเลือกในทางปฏิบัติ ตัวอย่างลมแสดงว่า GMRES เต็มรูปลู่เข้าภายในไม่กี่สิบรอบในเลขคณิตที่มีการปัดเศษ และการสร้างเวกเตอร์ตั้งฉากทั้งหมดก็แพง

วิธีอื่นที่ควรรู้

GMRES ไม่ใช่วิธีเดียวสำหรับระบบที่ไม่สมมาตร ตารางต่อไปนี้สรุปสามวิธีที่ VdV อธิบาย โดยเราไม่พิสูจน์ผลของแต่ละวิธี

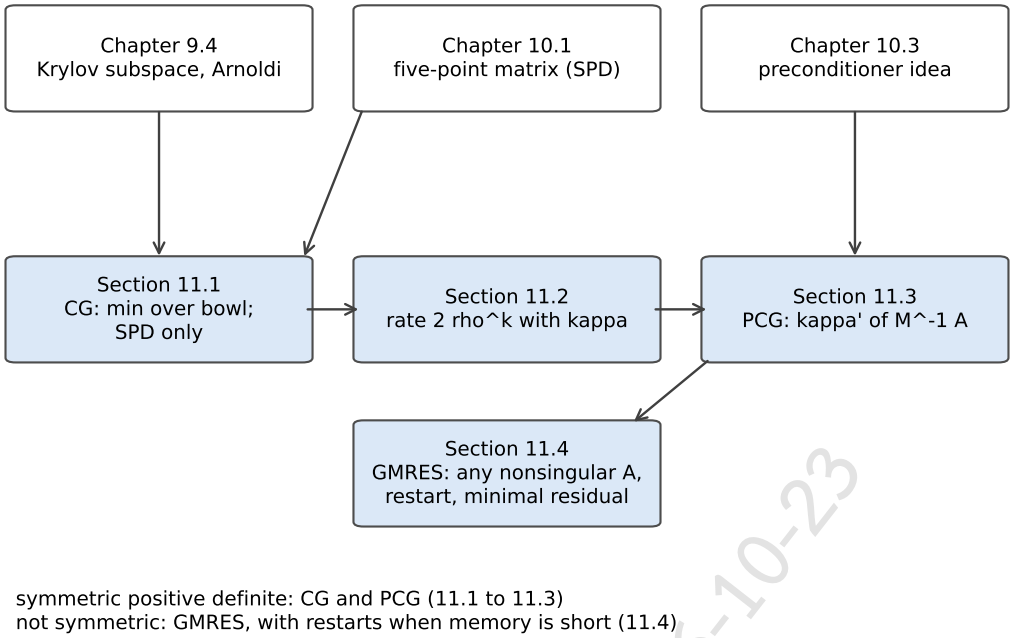
วิธี	ใช้กับ A แบบไหน	ความจำต่อรอบ	หมายเหตุ
MINRES Vorst 2003, §6.4	สมมาตร (อาจ ไม่ บวกแน่นอน)	คงที่	ทำค่าตกค้างให้น้อยที่สุด; ถ้า A บวกแน่นอน CG ดีกว่า
Bi-CG Vorst 2003, §7.1	ไม่เอกฐานใด ๆ	คงที่	ต้องคูณด้วย $A^\top$ ด้วย และอาจล้มเหลวกลางทาง
Bi-CGSTAB Vorst 2003, §9.1	ไม่เอกฐานใด ๆ	คงที่	เลือก พหุ นาม ชนิด $\prod(1 - \omega_j x)$ ให้ค่าตกค้างเรียบขึ้น; สูตรใช้เฉพาะ A

MINRES ใช้ความสัมพันธ์สามพจน์ของบทที่ 9 ได้เพราะ A สมมาตร Bi-CG ใช้ความสัมพันธ์สองชุดที่ต้องมี $A^\top$ และในหนึ่งขั้นต้องคูณเมทริกซ์กับเวกเตอร์ถึงสองครั้ง Vorst 2003, §7.1, pp. 95--96 Bi-CGSTAB ปรับ Bi-CG ให้ค่าตกค้างเรียบขึ้น Vorst 2003, §9.1, pp. 133--134 เราไม่ได้วัดว่าวิธีไหนเร็วกว่าในระบบลม

สรุปใจความ

- GMRES เลือกคำตอบที่ค่าตกค้างเล็กที่สุดในปริภูมิไครลอฟที่เลื่อนมาจาก $\mathbf{x}_0$ ทุกรอบ และคำนวณได้ด้วยกำลังสองน้อยที่สุดบน $\bar{H}_k$ (ทฤษฎีบท 11.4.1)
- ค่าตกค้างไม่เพิ่มทั้งในแต่ละรอบและข้ามรอบ แต่ไม่มีกฎง่าย ๆ สำหรับความยาวรอบ m ที่ทำให้ลู่เข้าเร็วที่สุด

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 11.8: แผนผังบทที่ 11 ทั้งบท หัวข้อ 11.1 ถึง 11.4 และสิ่งที่บทที่ 9 กับ 10 ให้บทนี้ ระบบที่สมมาตรบวกแน่นอนใช้ CG และ PCG ระบบที่ไม่สมมาตรใช้ GMRES (สร้างจาก `numanal/ch11.py`)

- เมทริกซ์เลื่อนวงกลมมี $\kappa = 1$ แต่ GMRES หยุดนิ่งจนถึงรอบที่ n (ทฤษฎีบท 11.4.3)
- GMRES เต็มรูปจบภายใน n รอบในเลขคณิตแม่นตรง (ทฤษฎีบท 11.4.4)
- ถ้า A สมมาตรบวกแน่นอนใช้ CG หรือ PCG ถ้า A สมมาตรแต่ไม่บวกใช้ MINRES และถ้าไม่สมมาตรใช้ GMRES หรือ Bi-CG

อ่านต่อ. หัวข้อ 11.4 อ่านเพิ่มได้ใน Vorst 2003, §6.1--6.2, pp. 65--78 และ Vorst 2003, §6.4, pp. 84--86 ซึ่งให้ GMRES และ MINRES ละเอียดกว่านี้ Vorst 2003, §7.1--7.3, pp. 95--100 ให้ Bi-CG และ QMR และ Vorst 2003, §9.1, pp. 133--137 ให้ Bi-CGSTAB

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 11.4.1 (⋆, ทำด้วยมือ) ให้ $A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ และ $\mathbf{b} = \mathbf{e}_2$ เริ่มจาก $\mathbf{x}_0 = \mathbf{0}$ จงหา $\bar{H}_1$ และ $\mathbf{x}_1$ ของ GMRES จากนั้นหาค่าตกค้างของ $\mathbf{x}_1$ และหา $\mathbf{x}_2$ ของ GMRES เต็มรูป

ตัวช่วยคิด

ทำตามตัวอย่างลองเองในหัวข้อนี้ ขั้นแรก $\mathbf{v}_1 = \mathbf{e}_2$ และ $A\mathbf{v}_1 = (1, 2)^\top$ ส่วนปริภูมิ $\mathcal{K}_2$ จะเป็น $\mathbb{R}^2$ ทั้งหมดเมื่อ $\mathbf{v}_1$ และ $A\mathbf{v}_1$ เป็นอิสระเชิงเส้น

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 11.4.2 (★★, คำนวณและอธิบาย) ให้ $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ และ $\mathbf{b} = \mathbf{e}_1$ เริ่มจาก $\mathbf{x}_0 = \mathbf{0}$

1. จงแสดงว่า GMRES(1) ไม่ขยับจาก $\mathbf{x}_0$ เลยในทุกรอบ
2. จงแสดงว่า GMRES เต็มรูปให้ $\mathbf{x}_2 = (0, -1)^\top$ และอธิบายว่าทำไมการรีสตาร์ตที่ความยาว 1 ทำให้ผลต่างกันมาก

แบบฝึกหัด 11.4.3 (★★, พิสูจน์) ให้ A ไม่เอกฐานและแยกได้ $A = SDS^{-1}$ โดย D เป็นเมทริกซ์แนวทแยงที่มีค่าลักษณะเฉพาะ q ค่าที่ต่างกัน และทุกค่าไม่เป็นศูนย์ จงแสดงว่า GMRES เต็มรูปให้ค่าตกค้างเป็นศูนย์ภายในไม่เกิน q รอบ (ในเลขคณิตแม่นตรง)

ตัวช่วยคิด

ให้ $P(x) = \prod_{j=1}^{q} (1 - x/\lambda_j)$ จงแสดงว่า $P(0) = 1$ และ $P(A) = 0$ จากนั้นใช้ว่า ค่าตกค้างของ GMRES รอบที่ k เขียนเป็น $P_k(A)\mathbf{r}_0$ โดย $P_k(0) = 1$ และ $\deg P_k \le k$

แบบฝึกหัด 11.4.4 (★★, พิสูจน์) ให้ S เป็น**เมทริกซ์สมมาตรเสมือน** (skew-symmetric matrix) คือ $S^\top = -S$ และให้ $\bar{H}_k$ มาจากการวนซ้ำอาร์โนลดีของ S จงแสดงว่า $\bar{H}_k$ มีสามแนวทแยง โดย $h_{jj} = 0$ และ $h_{j-1,j} = -h_{j,j-1}$ ดังนั้นขั้นตอนอาร์โนลดีเหลือสูตรสามพจน์

ตัวช่วยคิด

เขียน $H_k = \mathbf{V}_k^\top S \mathbf{V}_k$ แล้วดูว่า $H_k^\top = -H_k$ และ H_k เป็นเฮสเซนเบิร์กบน

แบบฝึกหัด 11.4.5 (★, คำนวณ) ระบบมี $n = 10^6$ ค่าที่ไม่รู้ และเก็บแต่ละตัวเลขเป็น double ขนาด 8 ไบต์ GMRES เต็มรูปที่ขั้นที่ k ต้องเก็บเวกเตอร์ของ $\mathbf{V}_{k+1}$ ทั้งหมด จงหาว่าเก็บได้กี่ขั้นในหน่วยความจำ 8 GB และ GMRES(50) ต้องใช้หน่วยความจำเท่าใด (นับเฉพาะ $\mathbf{V}$ ไม่นับเวกเตอร์เล็ก ๆ อื่น)

แบบฝึกหัด 11.4.6 (★★, คำนวณและอธิบาย) ใช้โค้ดของเราบนระบบลมขนาด $m = 8$ (64 ค่าที่ไม่รู้) $p = 0.9$ $\mathbf{b} = \mathbf{1}$ และ $\mathbf{x}_0 = \mathbf{0}$ นับจำนวนการคูณเมทริกซ์กับเวกเตอร์จนค่าตกค้างสัมพัทธ์ไม่เกิน 10^{-8} สำหรับ GMRES เต็มรูป และ GMRES(m') เมื่อ $m' = 2, 3, 4, 5, 8$ จงอธิบายว่าทำไม GMRES(8) ต้องใช้มากกว่า GMRES เต็มรูป ทั้งที่รอบของมันยาวถึง 8 ขั้น

แบบฝึกหัด 11.4.7 (★★, พิสูจน์) ให้ $T_c = \operatorname{tridiag}(-1 - p,\ 2,\ -1 + p)$ และ $T_d = \operatorname{tridiag}(-1,\ 2,\ -1)$ ขนาด $m \times m$ จงแสดงว่าเมทริกซ์ลมของหัวข้อนี้เท่ากับ**ผลบวกโครเนกเกอร์** (Kronecker sum) $T_d \otimes I + I \otimes T_c$ (ผลรวมของเมทริกซ์สองตัว ที่แต่ละตัวทำงานบนแกนหนึ่งของกริด) และค่าลักษณะเฉพาะคือ

$$4 - 2\sqrt{1 - p^2}\cos\frac{a\pi}{m+1} - 2\cos\frac{b\pi}{m+1}, \qquad a, b = 1, \dots, m,$$

แล้วอธิบายว่าเมื่อ $0 < p < 1$ ค่าลักษณะเฉพาะทั้งหมดเป็นจำนวนจริงบวก

ตัวช่วยคิด

หาเมทริกซ์แนวทแยงบวก $\mathbf{D}$ ที่ทำให้ $\mathbf{D}^{-1} T_c \mathbf{D}$ สมมาตร โดยเลือก $\delta_j / \delta_{j-1} = \sqrt{(1+p)/(1-p)}$ จากนั้นใช้เวกเตอร์ $\sin(ja\pi/(m+1))$ เป็นเวกเตอร์ลักษณะเฉพาะของ T_d และของเมทริกซ์สมมาตรที่ได้

ฉบับร่าง · Preprint · 2026-10-25

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 11.4.1.

1. $\mathbf{v}_1 = \mathbf{e}_2$ และ $A\mathbf{v}_1 = (1,2)^\top$ จึง $h_{11} = \mathbf{v}_1^\top A\mathbf{v}_1 = 2$ เหลือ $(1,2)^\top - 2\mathbf{e}_2 = (1,0)^\top$ มีนอร์ม 1 ดังนั้น $h_{21} = 1$ และ $\bar{H}_1 = (2,1)^\top$
2. กำลังสองน้อยที่สุด $\min_y |(1,0)^\top - y(2,1)^\top|^2 = (1-2y)^2 + y^2$ ได้ $y = \frac{2}{5}$ จึง $\mathbf{x}_1 = \frac{2}{5}\mathbf{e}_2 = (0, \frac{2}{5})^\top$
3. $A\mathbf{x}_1 = (\frac{2}{5}, \frac{4}{5})^\top$ จึง $\mathbf{b} - A\mathbf{x}_1 = (-\frac{2}{5}, \frac{1}{5})^\top$ มีนอร์ม $\frac{1}{\sqrt{5}}$ ซึ่งเท่ากับนอร์มของ $(\frac{1}{5}, -\frac{2}{5})^\top = \beta\mathbf{e}_1 - \bar{H}_1 y$ ตามทฤษฎีบท 11.4.1
4. $\mathcal{K}_2 = \operatorname{span}\{\mathbf{e}_2, (1,2)^\top\} = \mathbb{R}^2$ ดังนั้น GMRES เต็มรูปได้คำตอบแม่นตรงที่รอบที่ 2 ซึ่งคือคำตอบของ $2x_1 + x_2 = 0$ และ $2x_2 = 1$ ได้ $\mathbf{x}_2 = (-\frac{1}{4}, \frac{1}{2})^\top$ ตรวจได้โดยแทนค่า

ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_4_1_two_by_two_by_hand`)

เฉลย 11.4.2.

1. $\mathbf{r}_0 = \mathbf{e}_1$ และ $\mathbf{v}_1 = \mathbf{e}_1$ เมื่อคูณ $A\mathbf{v}_1 = \mathbf{e}_2$ จึง $h_{11} = 0$ และ $h_{21} = 1$ ดังนั้น $\bar{H}_1 = (0,1)^\top$ กำลังสองน้อยที่สุดคือ $\min_y |\mathbf{e}_1 - y\mathbf{e}_2|^2 = 1 + y^2$ ซึ่งได้ $y = 0$ จึง $\mathbf{x}_1 = \mathbf{0}$ รอบถัดไปเริ่มจาก $\mathbf{0}$ อีกครั้ง จึงทำซ้ำเดิมทุกรอบ เหตุผลคือปริภูมิของ GMRES(1) เป็นเส้นตรงที่มีทิศ $\mathbf{e}_1$ และ $A\mathbf{e}_1 = \mathbf{e}_2$ ตั้งฉากกับทิศนั้น
2. $\mathcal{K}_1 = \operatorname{span}\{\mathbf{e}_1\}$ และ $\mathcal{K}_2 = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2\} = \mathbb{R}^2$ จึงแก้ $A\mathbf{x} = \mathbf{e}_1$ ได้โดยตรง $(-x_2, x_1)^\top = (1,0)^\top$ ให้ $\mathbf{x}_2 = (0,-1)^\top$ การรีสตาร์ตที่ความยาว 1 ตัดปริภูมิสองมิติที่ GMRES เต็มรูปใช้ออกเหลือเส้นเดียว จึงไม่มีการขยับเลย

ตรวจด้วยโค้ด: `test_ch11.py` (`test_rotation_gmres1_stalls_and_full_gmres_solves_exercise_11_4_2`)

เฉลย 11.4.3. ทุกรอบของ GMRES มี $\mathbf{x}_k - \mathbf{x}_0 = \sum_{j<k} c_j A^j \mathbf{r}_0$ จึง $\mathbf{r}_k = P(A)\mathbf{r}_0$ โดย $P(x) = 1 - \sum_{j<k} c_j x^{j+1}$ มี $P(0) = 1$ และ $\deg P \le k$ ในทางกลับกัน ถ้า $P(0) = 1$ และ $\deg P \le k$ ให้ $1 - P(x) = xq(x)$ โดย $\deg q \le k-1$ แล้ว $\mathbf{x} = \mathbf{x}_0 + q(A)\mathbf{r}_0 \in \mathbf{x}_0 + \mathcal{K}_k$ มีค่าตกค้าง $P(A)\mathbf{r}_0$ ดังนั้น GMRES เลือกค่าต่ำสุดของ $\|P(A)\mathbf{r}_0\|_2$ ในบรรดา P เหล่านี้

ให้ $P(x) = \prod_{j=1}^{q}(1 - x/\lambda_j)$ ซึ่ง $P(0) = 1$ และ $\deg P = q$ เพราะ $A = SDS^{-1}$ จึง

$$\prod_{j=1}^{q}(A - \lambda_j I) = S\Big(\prod_{j=1}^{q}(D - \lambda_j I)\Big)S^{-1} = 0,$$

เพราะ ทุก สมาชิก แนว ทแยง ของ D เป็น หนึ่ง ใน λ_j และ ตัวประกอบ ที่ ตรง กัน ทำให้ ผล คูณ เป็น ศูนย์ จึง $P(A) = \prod_j(-\lambda_j^{-1})(A - \lambda_j I) = 0$ ดังนั้น $\|\mathbf{r}_q\|_2 \le \|P(A)\mathbf{r}_0\|_2 = 0$ GMRES เต็มรูปจึงให้คำตอบแม่นตรงภายในไม่เกิน q รอบ

สรุป: จำนวนค่าลักษณะเฉพาะที่ต่างกันคือขอบเขตของจำนวนรอบ ซึ่งอาจน้อยกว่า n มาก ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_4_3_distinct_eigenvalues_give_at_most_m_steps`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 11.4.4. ให้ $H_k = \mathbf{V}_k^\top S \mathbf{V}_k$ ซึ่งเป็นส่วนบน $k \times k$ ของ $\bar{H}_k$ เพราะ $A\mathbf{V}_k = \mathbf{V}_{k+1}\bar{H}_k$ และ $\mathbf{V}_k^\top \mathbf{V}_k = I$ เมื่อ $S^\top = -S$ จะได้ $H_k^\top = \mathbf{V}_k^\top S^\top \mathbf{V}_k = -H_k$ นั่นคือ $h_{ji} = -h_{ij}$ จากนั้น

1. ถ้า $i > j+1$ เฮสเซนเบิร์กทำให้ $h_{ij} = 0$ และความเอียงทำให้ $h_{ji} = 0$ ด้วย จึงเหลือเพียง $|i-j| \le 1$
2. $h_{jj} = -h_{jj}$ จึง $h_{jj} = 0$ และ $h_{j-1,j} = -h_{j,j-1}$

ในการวนซ้ำอาร์โนลดี ขั้นที่ j คือ $\mathbf{w} = S\mathbf{v}_j - \sum_{i \le j} h_{ij}\mathbf{v}_i = S\mathbf{v}_j - h_{j-1,j}\mathbf{v}_{j-1} = S\mathbf{v}_j + h_{j,j-1}\mathbf{v}_{j-1}$ แล้ว $\mathbf{v}_{j+1} = \mathbf{w}/h_{j+1,j}$ จึงเหลือสูตรสามพจน์

สรุป: การเป็นเมทริกซ์สมมาตรเสมือนทำให้อาร์โนลดีกลายเป็นการเวียนเกิดสามพจน์แบบเดียวกับลานซอส ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_4_4_skew_arnoldi_is_tridiagonal_and_skew`)

เฉลย 11.4.5. แต่ละเวกเตอร์มี 10^6 ตัวเลขขนาด 8 ไบต์ จึงใช้ 8×10^6 ไบต์ หรือ 8 MB 8 GB เท่ากับ 8×10^9 ไบต์ จึงเก็บได้ $\dfrac{8 \times 10^9}{8 \times 10^6} = 1000$ เวกเตอร์ ดังนั้น GMRES เต็มรูปเก็บได้ราว 1000 ขั้นก่อนหน่วยความจำเต็ม GMRES(50) เก็บ $\mathbf{V}_{51}$ จึงใช้ $51 \times 8\,\text{MB} = 408\,\text{MB}$ ซึ่งไม่รวมเมทริกซ์ $\bar{H}$ ที่เล็กมาก ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_4_5_memory_arithmetic`)

เฉลย 11.4.6. จำนวนการคูณเมทริกซ์กับเวกเตอร์ที่โค้ดนับ คือ GMRES เต็มรูป 22 ครั้ง GMRES(2) 49 ครั้ง GMRES(3) 45 ครั้ง GMRES(4) 54 ครั้ง GMRES(5) 47 ครั้ง และ GMRES(8) 43 ครั้ง

- GMRES เต็มรูปเก็บปริภูมิไครลอฟเดียวไว้ตลอดและลู่เข้าที่ขั้นที่ 22
- GMRES(8) ทำ 8 ขั้นแล้วเริ่มรอบใหม่ เพราะที่ขั้นที่ 8 ยังไม่ลู่เข้า (เต็มรูปยังต้องถึงขั้นที่ 22) การเริ่มใหม่ทิ้งปริภูมิที่สร้างมาแล้ว และทุกการรีสตาร์ตเสียการคูณ A หนึ่งครั้งเพื่อคำนวณค่าตกค้าง ทั้งสองอย่างนี้ทำให้ต้องใช้ขั้นมากกว่า

ตัวเลขเหล่านี้เป็นของโค้ดบนกริดนี้ โดยไม่มีกฎทั่วไปว่าความยาวรอบใดดีที่สุด ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_4_6_eight_by_eight_counts`)

เฉลย 11.4.7.

1. **ทำให้สมมาตรด้วยเมทริกซ์แนวทแยงบวก.** ให้ $\delta_0 = 1$ และ $\delta_j = \delta_{j-1}\sqrt{(1+p)/(1-p)}$ และให้ $\mathbf{D} = \operatorname{diag}(\delta_1, \ldots, \delta_m)$ สมาชิกนอกแนวทแยงของ $\mathbf{D}^{-1}T_c\mathbf{D}$ คือ $(\mathbf{D}^{-1}T_c\mathbf{D})_{j,j-1} = (-1-p)\dfrac{\delta_{j-1}}{\delta_j} = -(1+p)\sqrt{\dfrac{1-p}{1+p}} = -\sqrt{1-p^2}$ และ $(\mathbf{D}^{-1}T_c\mathbf{D})_{j-1,j} = (-1+p)\dfrac{\delta_j}{\delta_{j-1}} = -(1-p)\sqrt{\dfrac{1+p}{1-p}} = -\sqrt{1-p^2}$ จึงได้ $\mathbf{D}^{-1}T_c\mathbf{D} = \operatorname{tridiag}(-s, 2, -s) =: S_c$ โดย $s = \sqrt{1-p^2}$ และ $T_c = \mathbf{D}S_c\mathbf{D}^{-1}$
2. **เวกเตอร์ลักษณะเฉพาะของ T_d และ S_c.** ให้ $\theta = \dfrac{a\pi}{m+1}$ และ $w_j = \sin(j\theta)$ เมื่อ $w_0 = w_{m+1} = 0$ จากเอกลักษณ์ $\sin((j-1)\theta) + \sin((j+1)\theta) = 2\cos\theta\sin(j\theta)$ ได้ $(S_c\mathbf{w})_j = (2 - 2s\cos\theta)w_j$ ดังนั้น S_c มีค่าลักษณะเฉพาะ $2 - 2s\cos\theta_a$ และ T_d (กรณี $s=1$) มี $2 - 2\cos\theta_b$ ซึ่งเวกเตอร์ของทั้งคู่คือ $\mathbf{w}$ ที่ดัชนี a หรือ b ตามลำดับ
3. **ผลบวกโครเนกเกอร์.** สมาชิกของ $T_d \otimes I + I \otimes T_c$ ที่ตำแหน่ง (i,j) คือ 4 บนแนวทแยง -1 ที่ $(i \pm 1, j)$ และ $-1 \mp p$ ที่ $(i, j \mp 1)$ ตามลำดับ ซึ่งตรงกับเมทริกซ์ของหัวข้อนี้ (ดัชนี (i,j) เป็น $i \cdot m + j$) ให้ $\mathbf{w}_b$ เป็นเวกเตอร์ของ T_d และ $\mathbf{D}\mathbf{w}_a$ เป็นเวกเตอร์ของ T_c แล้ว $(\mathbf{w}_b \otimes \mathbf{D}\mathbf{w}_a)$ เป็นเวกเตอร์ลักษณะเฉพาะที่มีค่า $\lambda_b + \mu_a = 4 - 2\cos\frac{b\pi}{m+1} - 2\sqrt{1-p^2}\cos\frac{a\pi}{m+1}$ ทั้ง m^2 เวกเตอร์เป็นอิสระเชิงเส้นเพราะเป็นผลคูณของฐาน จึงได้ค่าลักษณะเฉพาะทั้งหมด

ม.ปชีร์ง · Preprint 2026-10-23

4. **บวกแน่นอน.** เมื่อ $0 < p < 1$ จะได้ $0 < s < 1$ และ $\lambda \geq (2 - 2s) + 0 > 0$ เพราะ $\cos \leq 1$ ทั้งสองพจน์ จึงทุกค่าลักษณะเฉพาะเป็นจำนวนจริงบวก

สรุป: เมทริกซ์ลมไม่สมมาตรแต่มีค่าลักษณะเฉพาะจริงบวก ทั้งหมดได้จากการเปลี่ยนฐานด้วย $\mathbf{D}$ และผลบวกโครเนกเกอร์ ตรวจด้วยโค้ด: `test_ch11.py` (`test_exercise_11_4_7_closed_form_matches_numpy_for_m6`)

สรุปบทที่ 11 ในหน้าเดียว

ตารางนี้เทียบวิธีหลัก โดย n คือจำนวนค่าที่ไม่รู้ และ κ คือคอนดิชันนัมเบอร์

วิธี	ต้องการอะไรจาก A	ความจำต่อรอบ	ค่าที่ทำให้น้อยที่สุด
CG (11.1)	สมมาตรบวกแน่นอน	คงที่	ความคลาดเคลื่อนในนอร์มพลังงาน
PCG (11.3)	สมมาตร บวก แน่นอน และ M	คงที่	เหมือน CG บนระบบที่เปลี่ยนตัวแปร
MINRES (11.4)	สมมาตร	คงที่	ค่าตกค้าง
GMRES(m) (11.4)	ไม่เอกฐาน	โต ตาม จำนวน รอบ จนถึง m	ค่าตกค้างในแต่ละรอบ
Bi-CG และ Bi-CGSTAB (11.4)	ไม่เอกฐาน	คงที่	ไม่รับประกัน

สูตรที่ควรจำ: CG ลดความคลาดเคลื่อนในนอร์มพลังงานด้วยอัตรา $2\rho^k$ โดย $\rho = \dfrac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$ (หัวข้อ 11.2)

อ่านต่อ. บทที่ 12 และ 14 จะพบขั้นตอนโดยนัยที่ต้องแก้ระบบเชิงเส้นในแต่ละก้าวเวลา และเครื่องมือของบทนี้ใช้กับระบบเหล่านั้นได้ ส่วนรายละเอียดของ MINRES, Bi-CG และ Bi-CGSTAB ดูที่ Vorst 2003, §6.4, §7.1, §9.1 ตามที่อ้างไว้ในหัวข้อ 11.4

ฉบับร่าง · Preprint · 2026-10-25

ภาค V

สิ่งที่เปลี่ยนไปตามเวลา

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 12

ก้าวไปตามเวลา

สถานการณ์: สปริงที่พลังงานเพิ่มขึ้น

ลองนึกถึงลูกบอลที่ผูกกับสปริงในฉากหนึ่ง เราดึงลูกบอลออกไปแล้วปล่อยมือ ตามสามัญสำนึก สปริงควรแกว่งไปมาแล้วค่อย ๆ หยุดนิ่งที่จุดกลาง เพราะมีแรงหน่วงคอยกินพลังงานไปทีละน้อย
ในคอมพิวเตอร์เราไม่ได้เห็นการแกว่งแบบต่อเนื่อง เราดูสปริงทีละช่วงเวลาสั้น ๆ ที่เรียกว่าขั้นเวลา แล้วคำนวณตำแหน่งและความเร็วของช่วงถัดไปจากค่าของช่วงนี้ เอกสารของ Fiedler ยกตัวอย่างสปริงมวล 1 กิโลกรัม ค่าคงที่ของสปริง $k = 15$ แรงหน่วง $b = 0.1$ และเริ่มที่ 1000 เมตร เมื่อเดินเวลาทีละ $1/100$ วินาทีด้วยวิธีที่ง่ายที่สุด บทความนั้นรายงานว่าระบบนี้ไม่เสถียร และพลังงานของมันเพิ่มขึ้นตามเวลา Fiedler 2004b เอกสารของ Box2D ระบุว่าขั้นเวลา $1/60$ วินาทีมักให้การจำลองคุณภาพดี Catto 2025
ตัวเลขในภาพต่อไปนี้คำนวณจากพารามิเตอร์ของ Fiedler ด้วยโค้ดของเรา (สถานการณ์สมมติ ไม่ได้บอกว่าโปรแกรมใดทำเช่นนี้) สปริงที่ควรหยุดกลับโตขึ้นจนมีค่า $|x|$ เกิน 12,000 เมตรภายใน 100 วินาที คำถามของบทนี้คือ ทำไมขั้นเวลาเล็ก ๆ จึงทำให้สปริงโต และเราจะเดินเวลาอย่างไรให้สปริงหยุดนิ่งตามที่ควร

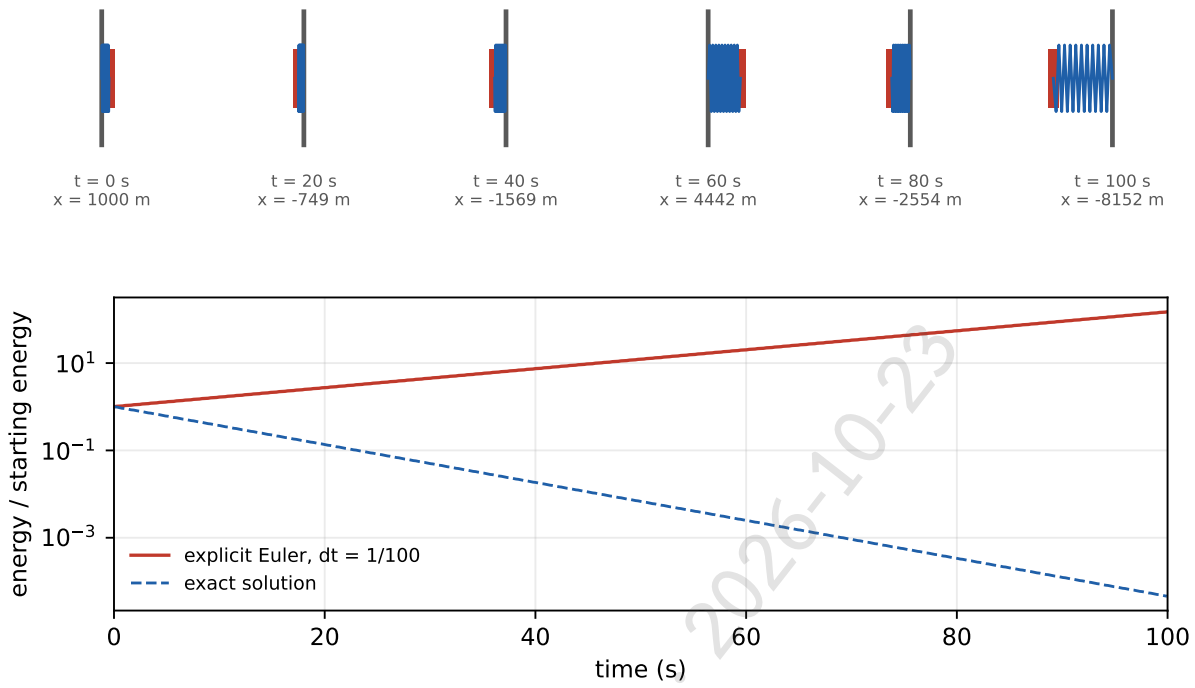
ภาพรวมบท

บทนี้มีสามหัวข้อ เริ่มจากวิธีที่ง่ายที่สุด แล้วดูว่าต้องแก้อะไร

- หัวข้อ 12.1 คือการเดินตามลูกศรทีละก้าวด้วย**วิธีออยเลอร์** (Euler method) และพิสูจน์ว่าเมื่อขั้นสั้นลง คำตอบจะเข้าใกล้คำตอบจริงด้วยอัตราเท่าไร
- หัวข้อ 12.2 ทำให้ก้าวนิ่งขึ้นด้วยการอ่านลูกศรในจุดที่ใกล้เป้าหมายมากขึ้น และดูอีกด้านหนึ่งของปัญหาเดียวกัน คือการข้ามกำแพงบางระหว่างสองเฟรม ซึ่งไม่เกี่ยวกับการระเบิดเลย
- หัวข้อ 12.3 ใช้ตำแหน่งสองจุดก่อนหน้าแทนการเก็บความเร็ว และดูว่าวิธีแบบหลายขั้นต้องระวังอะไร

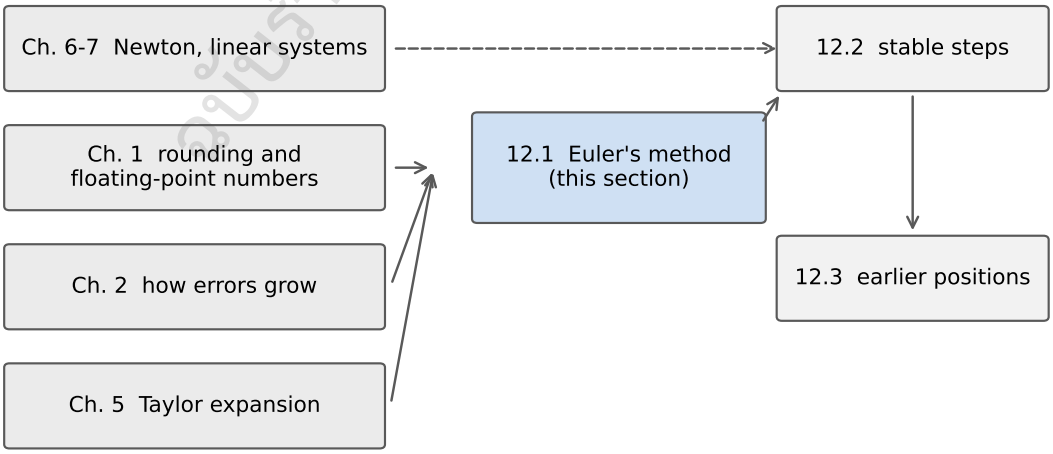
เริ่มจากสปริงที่พลังงานเพิ่มขึ้น แล้วจะเจออีกด้านหนึ่งของปัญหาเดียวกันในหัวข้อ 12.2 คือกำแพงที่ขั้นใหญ่ข้ามไปได้โดยไม่มีอะไรระเบิด

ฉบับร่าง · Preprint 2026-10-23



รูปที่ 12.1: สปริงตามพารามิเตอร์ของ Fiedler ($m = 1$, $k = 15$, $b = 0.1$, เริ่มที่ 1000 เมตร) คำนวณด้วยวิธีออยเลอร์ชัดแจ้งขั้น $1/100$ วินาที ด้านบนคือหกเฟรมที่เวลา $0, 20, \ldots, 100$ วินาที ตำแหน่งมีเครื่องหมาย ด้านล่างคือพลังงานเทียบกับพลังงานเริ่มต้น ของการคำนวณ (เส้นทึบ) และของคำตอบจริง (เส้นประ) (สร้างจาก `numanal/ch12.py`)

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 12.2: แผนผังบทที่ 12 กล่องสีฟ้าคือหัวข้อ 12.1 ลูกศรทึบคือแนวคิดที่หัวข้อหนึ่งใช้ ลูกศรประคือการใช้ในหัวข้อ 12.2 เพราะขั้นโดยนัยต้องแก้สมการทุกก้าว (สร้างจาก `numanal/ch12.py`)

12.1 ตามลูกศรไปทีละก้าว

สถานการณ์: วัตถุที่ถูกผลักด้วยแรงคงที่

ลองดูวัตถุมวล 1 กิโลกรัม ตั้งนิ่งที่จุดกำเนิด แล้วมีแรงคงที่ 10 นิวตันดันไปทางขวา ความเร่งจึงคงที่ 10 เมตรต่อวินาทีกำลังสอง และหลัง 10 วินาทีวัตถุจะอยู่ที่ $\frac{1}{2} \cdot 10 \cdot 10^2 = 500$ เมตร นี่คือคำตอบแม่นตรง

ในโปรแกรมจำลอง เรามักไม่มีสูตรคำตอบล่วงหน้า เรามีแค่กฎที่บอกว่าความเร็วเปลี่ยนเท่าไรในแต่ละช่วงเวลา บทความนั้นใช้ตัวอย่างนี้ทดสอบวิธีที่ง่ายที่สุด คือเดินทีละ 1 วินาที แล้วพบว่าได้ตำแหน่ง 450 เมตร ไม่ใช่ 500 เมตร Fiedler 2004b

ทำไมคำตอบจึงผิดไปเท่านี้ และถ้าเดินเวลาให้เล็กลง คำตอบจะเข้าใกล้ 500 เมตรเร็วแค่ไหน

ภาพรวมก่อนอ่าน

เราจะเปลี่ยนคำถามเรื่องการเคลื่อนที่ให้เป็นคำถามเรื่องลูกศร ทุกจุดมีลูกศรบอกว่าคำตอบควรเปลี่ยนไปทางไหนและเร็วแค่ไหน วิธีออยเลอร์อ่านลูกศรที่จุดเริ่ม แล้วเดินไปตามทิศนั้นหนึ่งขั้น จากนั้นอ่านลูกศรใหม่ ความผิดพลาดเล็ก ๆ ในแต่ละขั้นจะสะสม และเราจะพิสูจน์ว่าเมื่อขั้นสั้นลงครึ่งหนึ่ง ความผิดพลาดรวมจะลดลงประมาณครึ่งหนึ่ง

ลองเอง: สองก้าวด้วยมือ

ใช้กฎที่พูดเป็นภาษาคนก่อน ในแต่ละก้าว ตำแหน่งใหม่คือตำแหน่งเดิมบวก (ความเร็วเดิม $\times\, h$) และความเร็วใหม่คือความเร็วเดิมบวก (ความเร่ง $\times\, h$) ให้ $h = 1$ ความเร่ง 10 และเริ่มที่ $x_0 = 0$ กับ $v_0 = 0$

ก้าวที่ 1. $x_1 = 0 + 0 \cdot 1 = 0$ และ $v_1 = 0 + 10 \cdot 1 = 10$

ก้าวที่ 2. $x_2 = x_1 + v_1 \cdot 1 = 0 + 10 = 10$ และ $v_2 = v_1 + 10 \cdot 1 = 20$

สังเกตว่า x_2 ใช้ความเร็ว v_1 ของก้าวแรก ไม่ใช่ v_2 วิธีนี้อ่านลูกศรที่จุดเริ่มของแต่ละก้าว และไม่รู้ว่าลูกศรเปลี่ยนไปในก้าวนั้นเท่าไร คำตอบจริงที่ $t = 2$ วินาทีคือ 20 เมตร แต่ได้ 10 เมตร

สมการที่บอกทิศทาง

ให้ $\mathbf{y}(t)$ เป็นสถานะของระบบที่เวลา t เช่นของวัตถุเราใช้ $\mathbf{y} = (x, v)$ ซึ่งประกอบด้วยตำแหน่งและความเร็ว แล้วอนุพันธ์ $\mathbf{y}'(t)$ เป็นฟังก์ชันของเวลาและของสถานะเอง สมการที่บอกว่าอนุพันธ์ของคำตอบเท่ากับอะไรแบบนี้เรียกว่า **สมการเชิงอนุพันธ์สามัญ** (ordinary differential equation (ODE)) ถ้ารู้สถานะที่จุดเริ่มด้วย เราจะได้**ปัญหาค่าเริ่มต้น** (initial value problem) นิยามต่อไปนี้ตรงกับที่ใช้กันทั่วไป Iserles 2009, §1.1, p. 3 Atkinson 1989, §6.1, p. 336

นิยาม 12.1.1: ปัญหาค่าเริ่มต้นและเงื่อนไขลิพชิทซ์

ให้ $f : [t_0, t_0 + T] \times \mathbb{R}^d \to \mathbb{R}^d$ และ $\mathbf{y}_0 \in \mathbb{R}^d$ ปัญหาค่าเริ่มต้นคือการหาฟังก์ชัน $\mathbf{y}(t)$ ที่

$$\mathbf{y}'(t) = f\big(t, \mathbf{y}(t)\big), \qquad \mathbf{y}(t_0) = \mathbf{y}_0.$$

เราบอกว่า f เป็นไปตาม**เงื่อนไขลิพชิทซ์** (Lipschitz condition) ในตัวแปร $\mathbf{y}$ ถ้ามีค่าคงที่ $L \geq 0$ ซึ่ง

$$\|f(t, \mathbf{x}) - f(t, \mathbf{y})\| \leq L \, \|\mathbf{x} - \mathbf{y}\|$$

สำหรับทุก $t \in [t_0, t_0 + T]$ และทุก $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

ลงมือทำ · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

เงื่อนไขลิพชิทซ์บอกว่า ถ้าสองสถานะอยู่ใกล้กัน ลูกศรที่จุดทั้งสองต่างกันไม่เกิน L คูณระยะห่างระหว่างสองจุดนั้น ของวัตถุเรามี $f(t,\mathbf{y}) = (v, 10)$ ซึ่งต่างกันเพียงส่วนแรก v จึงได้ $L = 1$ ตัวอย่างที่ไม่ผ่าน: $f(t,y) = y^2$ บนทั้งเส้นจริง เพราะลูกศรยิ่งอยู่ไกลก็ยิ่งชันขึ้นโดยไม่มีขอบเขต (แบบฝึกหัด 12.1.3)

ทฤษฎีบท 12.1.1: การมีคำตอบและคำตอบเดียว

ถ้า f ต่อเนื่องบน $[t_0, t_0+T] \times \mathbb{R}^d$ และเป็นไปตามเงื่อนไขลิพชิทซ์ในตัวแปร $\mathbf{y}$ แล้วปัญหาค่าเริ่มต้นมีคำตอบ $\mathbf{y}(t)$ บน $[t_0, t_0+T]$ เพียงคำตอบเดียว

ที่มา. ผลนี้คือทฤษฎีบทของปีการ์ด–ลินเดอเลิฟ เราไม่พิสูจน์ซ้ำที่นี่ ดูได้ใน Iserles 2009, ภาคผนวก A.2.3, pp. 445--446 ดูเพิ่ม ภาคผนวก A.2.3 ทฤษฎีบท A.2.3.1--A.2.3.2 ของ Iserles และ Atkinson 1989, §6.1, Thm 6.1--6.2, pp. 336--337

เดินตามลูกศรหนึ่งก้าว

ความคิดของวิธีออยเลอร์เรียบง่าย ถ้ารู้สถานะที่เวลา t_k ให้ถือว่าลูกศรไม่เปลี่ยนในช่วง h แล้วเดินไปตามลูกศรนั้น เราเลือกขนาดขั้น $h > 0$ และตั้ง $t_k = t_0 + kh$

นิยาม 12.1.2: วิธีออยเลอร์

ให้ $\mathbf{y}_0$ เป็นค่าเริ่มต้น แล้วคำนวณสำหรับ $k = 0, 1, 2, \dots$

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h\, f(t_k, \mathbf{y}_k), \qquad t_k = t_0 + kh.$$

วิธีนี้คือ**วิธีออยเลอร์** (Euler method) และเป็น**วิธีชัดแจ้ง** (explicit method) เพราะค่าใหม่คำนวณได้ตรง ๆ จากค่าเดิม

อ่านแบบเข้าใจง่าย

วิธีออยเลอร์เดินเป็นเส้นตรงสั้น ๆ หลายช่วงแทนเส้นโค้งจริง ความผิดพลาดที่เกิดในหนึ่งขั้นเรียกว่า **ค่าคลาดเคลื่อนตัดปลายเฉพาะที่** (local truncation error) ส่วนความผิดพลาดที่สะสมมาถึงปลายทางเรียกว่า **ค่าคลาดเคลื่อนวงกว้าง** (global error) ในตัวอย่างข้างบนค่าคลาดเคลื่อนวงกว้างที่ $t = 2$ คือ 10 เมตร

ก้าวเล็กลงแล้วดีขึ้นแค่ไหน

เราทดสอบกับโจทย์ $y' = y$, $y(0) = 1$ ซึ่งคำตอบจริงคือ e^t ที่ $t = 1$ ค่าจริงคือ $e \approx 2.718282$ ตารางนี้ได้จากโค้ดของเรา โดยวัดความคลาดเคลื่อน $|e - y_k|$ ที่ $t = 1$

เมื่อ h ลดครึ่งหนึ่ง ความคลาดเคลื่อนลดลงใกล้ครึ่งหนึ่ง ตารางแสดงอัตราส่วนเข้าหา 2 เราเรียกว่าวิธีนี้มี**อันดับของวิธี** (order of a method) เท่ากับ 1 ซึ่งหมายความว่าความคลาดเคลื่อนวงกว้างลดตามกำลังหนึ่งของ h ตารางบอกเพียงเท่านี้ ส่วนข้อพิสูจน์ถัดไปจะบอกว่าทำไม

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 12.1: ความคลาดเคลื่อนของวิธีออยเลอร์สำหรับ $y' = y$ ที่ $t = 1$ คอลัมน์ขวาคือความคลาดเคลื่อนของแถวก่อนหารด้วยแถวนี้ (สร้างจาก `numanal/ch12.py` และตรวจด้วย `test_ch12.py`)

h	ความคลาดเคลื่อน	อัตราส่วน
$1/2$	0.468282	—
$1/4$	0.276876	1.6913
$1/8$	0.152497	1.8156
$1/16$	0.080353	1.8978
$1/32$	0.041292	1.9460

นิยาม 12.1.3: การลู่เข้า

เราบอกว่าวิธีเชิงตัวเลขมี**การลู่เข้า** (convergence) ถ้าสำหรับทุกปัญหาค่าเริ่มต้นที่ f เป็นไปตามเงื่อนไขลิพชิทซ์ และทุก $t^* > t_0$

$$\lim_{h \to 0^+} \max_{k:\, t_k \le t^*} \|\mathbf{y}_k - \mathbf{y}(t_k)\| = 0$$

แนวคิดนี้มาจาก Iserles 2009, §1.1, p. 6 ซึ่งเป็นเกณฑ์ที่เราต้องการพิสูจน์ให้วิธีออยเลอร์

ทฤษฎีบท 12.1.2: ความคลาดเคลื่อนของวิธีออยเลอร์

ให้ f เป็นไปตามเงื่อนไขลิพชิทซ์ด้วยค่าคงที่ $L > 0$ ให้คำตอบจริง $\mathbf{y}$ หาอนุพันธ์อันดับสองได้ต่อเนื่องบน $[t_0, t_0 + T]$ โดย $M = \max \|\mathbf{y}''(t)\|$ และให้ $\mathbf{y}_k$ เป็นผลของวิธีออยเลอร์โดยเริ่มที่ $\mathbf{y}_0 = \mathbf{y}(t_0)$ แล้วสำหรับทุก k ที่ $t_k \le t_0 + T$

$$\|\mathbf{y}(t_k) - \mathbf{y}_k\| \le \frac{Mh}{2L}\bigl(e^{L(t_k - t_0)} - 1\bigr).$$

ดังนั้นวิธีออยเลอร์ลู่เข้า และความคลาดเคลื่อนวางกว้างเป็น $O(h)$ คืออันดับ 1 Iserles 2009, §1.2, p. 8 ดูเพิ่มบทที่ 1 ทฤษฎีบท 1.1 ของ Iserles

พิสูจน์. **แนวคิด.** ประมาณความคลาดเคลื่อนของหนึ่งก้าวด้วยสูตรเทย์เลอร์ แล้วปล่อยให้ความคลาดเคลื่อนนั้นส่งต่อไปตามลำดับก้าว โดยใช้เงื่อนไขลิพชิทซ์ บทพิสูจน์นี้เราเขียนเองจากแนวคิดของ Iserles 2009, §1.1, Thm 1.1, p. 6

1. **ค่าคลาดเคลื่อนตัดปลายเฉพาะที่.** ให้ $\boldsymbol{\ell}_{k+1} = \mathbf{y}(t_{k+1}) - \mathbf{y}(t_k) - h\,f(t_k, \mathbf{y}(t_k))$ จากสูตรเทย์เลอร์ที่มีส่วนเหลือเป็นปริพันธ์ $\mathbf{y}(t_{k+1}) - \mathbf{y}(t_k) - h\,\mathbf{y}'(t_k) = \int_{t_k}^{t_{k+1}} (t_{k+1} - s)\,\mathbf{y}''(s)\,ds$ และ $\mathbf{y}'(t_k) = f(t_k, \mathbf{y}(t_k))$ ดังนั้น $\|\boldsymbol{\ell}_{k+1}\| \le M \int_{t_k}^{t_{k+1}} (t_{k+1} - s)\,ds = \tfrac{1}{2} M h^2$
2. **สมการของความผิดพลาด.** ให้ $\boldsymbol{\varepsilon}_k = \mathbf{y}(t_k) - \mathbf{y}_k$ นำสมการของวิธีออยเลอร์ลบออกจากคำนิยามของ $\boldsymbol{\ell}_{k+1}$ จะได้ $\boldsymbol{\varepsilon}_{k+1} = \boldsymbol{\varepsilon}_k + h\bigl[f(t_k, \mathbf{y}(t_k)) - f(t_k, \mathbf{y}_k)\bigr] + \boldsymbol{\ell}_{k+1}$
3. **ประมาณนอร์ม.** ใช้เงื่อนไขลิพชิทซ์กับพจน์กลางและข้อ 1 จะได้ $\|\boldsymbol{\varepsilon}_{k+1}\| \le (1 + hL)\|\boldsymbol{\varepsilon}_k\| + \tfrac{1}{2} M h^2$
4. **ทำซ้ำตามลำดับก้าว.** เพราะ $\boldsymbol{\varepsilon}_0 = \mathbf{0}$ ทำซ้ำข้อ 3 ได้ $\|\boldsymbol{\varepsilon}_k\| \le \tfrac{1}{2} M h^2 \sum_{j=0}^{k-1} (1+hL)^j = \frac{Mh}{2L}\bigl((1+hL)^k - 1\bigr)$
5. **ใช้ $1 + x \le e^x$.** จาก $(1 + hL)^k \le e^{khL} = e^{L(t_k - t_0)}$ ได้ขอบเขตที่ต้องการ

สรุป: ความคลาดเคลื่อนของหนึ่งก้าวเป็นกำลังสองของ h แต่จำนวนก้าวเพิ่มเป็นสัดส่วนกับ $1/h$ ผลรวมจึงเป็น $O(h)$ □

ฉบับร่าง · preprint · 2026-10-23

ข้อควรระวัง: ขอบเขตถูกต้องแต่ไม่มีประโยชน์เมื่อเวลายาว

ทฤษฎีบทบอกว่าความคลาดเคลื่อนลู่เข้าเมื่อ $h \to 0$ แต่ปัจจัย $e^{L(t_k - t_0)}$ โตเร็วมากเมื่อ L หรือเวลาใหญ่ ลองดูสปริงของเรา เมทริกซ์ $\begin{pmatrix} 0 & 1 \\ -15 & -0.1 \end{pmatrix}$ มีนอร์มแบบยุคลิด 15.0003 ซึ่งใช้เป็นค่าคงที่ลิพชิทซ์ได้ เมื่อ $T = 100$ วินาที $LT \approx 1500$ และ $e^{1500} \approx 10^{651}$ ขอบเขตนี้จึงถูกจริงแต่บอกอะไรไม่ได้ที่เวลาเช่นนี้ การที่สปริงโตไม่ได้ขัดกับทฤษฎีบท ทฤษฎีบทนี้พูดถึงขั้น h ที่เล็กลงเรื่อย ๆ โดยเวลาคงที่เท่านั้น ส่วนว่าขั้นที่ $h = 1/100$ เพียงพอหรือไม่เป็นเรื่องของเสถียรภาพ ซึ่งจะมาในหัวข้อ 12.2

เกร็ดเพิ่มเติม (ผู้เรียบเรียง): เกณฑ์รากของสมการกำลังสองจริง

ให้ $T, D \in \mathbb{R}$ และ $p(\lambda) = \lambda^2 - T\lambda + D$ ทั้งสองรากของ p มีขนาดน้อยกว่า 1 ก็ต่อเมื่อ $|D| < 1$ และ $|T| < 1 + D$ เกณฑ์นี้ใช้ในแบบฝึกหัด 12.1.7 และในทฤษฎีบท 12.2.4

การพิสูจน์. $p(1) = 1 - T + D$ และ $p(-1) = 1 + T + D$ จึง $|T| < 1 + D$ เทียบเท่ากับ $p(1) > 0$ และ $p(-1) > 0$ ส่วนตรง: ถ้ารากคือ λ_1, λ_2 แล้ว $D = \lambda_1\lambda_2$ จึง $|D| < 1$ ถ้ารากเป็นคู่สังยุคเชิงซ้อน $p(1) = |1 - \lambda_1|^2 > 0$ และ $p(-1) = |1 + \lambda_1|^2 > 0$ ถ้ารากเป็นจำนวนจริงในช่วง $(-1, 1)$ ทั้งสองตัวประกอบของ $p(1) = (1-\lambda_1)(1-\lambda_2)$ เป็นบวก และของ $p(-1) = (1+\lambda_1)(1+\lambda_2)$ ก็เป็นบวก ส่วนกลับ: ให้ $|D| < 1$ และ $|T| < 1+D$ ถ้ารากเป็นคู่สังยุคเชิงซ้อน $|\lambda|^2 = D < 1$ ถ้ารากเป็นจำนวนจริง $r_1 \le r_2$ เพราะ $|T| < 1+D < 2$ จุดกึ่งกลาง $T/2$ อยู่ใน $(-1, 1)$ ถ้า ± 1 อยู่ในช่วง $[r_1, r_2]$ จะได้ $p(\pm 1) \le 0$ ซึ่งขัดกับ $p(\pm 1) > 0$ ดังนั้น $[r_1, r_2]$ ไม่มี ± 1 จึงอยู่ทั้งหมดใน $(-\infty, -1)$, $(-1, 1)$ หรือ $(1, \infty)$ และเพราะจุดกึ่งกลางอยู่ใน $(-1, 1)$ จึงต้องอยู่ใน $(-1, 1)$

สรุปใจความ

- วิธีออยเลอร์อ่านลูกศรที่จุดเริ่มของทุกก้าว แล้วเดินไปตามทิศนั้นระยะ h: $\mathbf{y}_{k+1} = \mathbf{y}_k + h\, f(t_k, \mathbf{y}_k)$
- เงื่อนไขลิพชิทซ์ทำให้ปัญหาค่าเริ่มต้นมีคำตอบเพียงคำตอบเดียว และทำให้ความผิดพลาดที่เกิดในแต่ละก้าวส่งต่อไปได้อย่างควบคุมได้
- ความคลาดเคลื่อนตัดปลายเฉพาะที่เป็น $O(h^2)$ แต่จำนวนก้าวเป็น $O(1/h)$ ผลรวมจึงเป็น $O(h)$ คืออันดับ 1 เมื่อ h ลดครึ่งหนึ่ง ความคลาดเคลื่อนลดลงประมาณครึ่งหนึ่ง
- กลับไปที่สปริง: ทฤษฎีบทบอกว่าลู่เข้า แต่ไม่ได้บอกว่าขั้น $1/100$ จะเสถียร คำถามนั้นเป็นของหัวข้อ 12.2

อ่านต่อ. หัวข้อ 12.2 จะถามว่าทำไมสปริงเดิมโตได้ทั้งที่ผลพิสูจน์บอกว่าลู่เข้า และจะทำให้ก้าวนิ่งขึ้นด้วยวิธีใด ดูเพิ่มเติมเรื่องวิธีออยเลอร์และอันดับได้ที่ Iserles 2009, §1.1--1.2, pp. 3--8 และ Atkinson 1989, §6.2, pp. 341--346

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 12.1.1 (⋆, ทำด้วยมือ) ให้ $y' = -2y$ และ $y(0) = 1$ ทำวิธีออยเลอร์สองก้าวด้วยขนาดขั้น $h = \frac{1}{4}$ จงหา y_1 และ y_2 แล้วเทียบ y_2 กับคำตอบจริง $y(t) = e^{-2t}$ ที่ $t = \frac{1}{2}$

แบบฝึกหัด 12.1.2 (⋆, คำนวณ) ใช้วัตถุมวล 1 กิโลกรัมที่มีแรงคงที่ 10 นิวตันของหัวข้อ 12.1 จงหาตำแหน่งที่ $t = 10$ วินาที เมื่อใช้ขนาดขั้น $h = \frac{1}{2}$ วินาที เทียบกับความคลาดเคลื่อน 50 เมตรที่ได้เมื่อ $h = 1$ ความคลาดเคลื่อนลดลงเท่าไร

แบบฝึกหัด 12.1.3 (⋆⋆, พิสูจน์) ให้ $g(y) = y^2$ แสดงว่า

1. บนช่วง $[0, 1]$ มี $|g(x) - g(y)| \le 2|x - y|$ จึง g เป็นไปตามเงื่อนไขลิพชิทซ์ด้วย $L = 2$ (ใบ้: $x^2 - y^2 = (x + y)(x - y)$)
2. บนช่วง $[0, R]$ ค่าที่มากที่สุดของ $|g(x) - g(y)|/|x - y|$ คือ $2R$ จึงไม่มีค่าคงที่ L เดียวที่ใช้ได้สำหรับทุก R

แบบฝึกหัด 12.1.4 (⋆⋆, คำนวณ) ให้ $y' = y$ และ $y(0) = 1$ ใช้วิธีออยเลอร์กับ $h = \frac{1}{64}$ แล้วหาความคลาดเคลื่อนที่ $t = 1$ และอัตราส่วนกับแถว $h = \frac{1}{32}$ ในตารางหัวข้อ 12.1 ตรวจว่าอัตราส่วนยังเข้าหา 2 ด้วยการคำนวณหนึ่งแถว และอธิบายว่าทำไมจึงเป็นเช่นนั้น (ใบ้: ค่าของวิธีออยเลอร์บนโจทย์นี้คือ $(1 + h)^{1/h}$ เมื่อ $t = 1$)

แบบฝึกหัด 12.1.5 (⋆ ⋆ ⋆, พิสูจน์) ให้สปริงไม่มีแรงหน่วง คือ $m = 1$ และ $b = 0$ ดังนั้น $x' = v$, $v' = -kx$ และวิธีออยเลอร์หนึ่งก้าวเป็น $\mathbf{y}_{n+1} = M_h \mathbf{y}_n$ เมื่อ $M_h = \begin{pmatrix} 1 & h \\ -kh & 1 \end{pmatrix}$

1. แสดงว่า $\det M_h = 1 + kh^2$ และ $\operatorname{tr} M_h = 2$
2. แสดงว่าค่าลักษณะเฉพาะทั้งสองมีขนาด $\sqrt{1 + kh^2}$ ซึ่งมากกว่า 1 เมื่อ $h > 0$
3. คำนวณขนาดนี้เมื่อ $k = 15$ และ $h = 0.01$ แล้วอธิบายว่าทำไมในภาพของเรา ตัวเลขโตขึ้นทุกก้าวแม้ไม่มีแรงหน่วง

แบบฝึกหัด 12.1.6 (⋆⋆, พิสูจน์) สำหรับวัตถุแรงคงที่ 10 นิวตัน ให้ $v_n = 10nh$ และ x_n เป็นตำแหน่งหลัง n ก้าว

1. แสดงด้วยการอุปนัยว่า $x_n = 5h^2 n(n - 1)$
2. ที่ $t = 10$ วินาที ให้ $n = 10/h$ จงแสดงว่าความคลาดเคลื่อนเท่ากับ $50h$ พอดี
3. อธิบายว่าผลนี้สอดคล้องกับทฤษฎีบทในหัวข้อ 12.1 อย่างไร

แบบฝึกหัด 12.1.7 (⋆ ⋆ ⋆, โครงงานเล็ก) ให้สปริงมีแรงหน่วง $m = 1$, $k = 15$, $b = 0.1$ และใช้วิธีออยเลอร์ชัดแจ้งกับขนาดขั้น h เมทริกซ์หนึ่งขั้นคือ $M_h = \begin{pmatrix} 1 & h \\ -kh & 1 - bh \end{pmatrix}$

1. แสดงว่า $\det M_h = 1 - bh + kh^2$ และ $\operatorname{tr} M_h = 2 - bh$
2. แสดงว่าค่าลักษณะเฉพาะทั้งสองอยู่ในวงกลมหนึ่งหน่วย ก็ต่อเมื่อ $0 < h < b/k$ (ใบ้: ใช้เกณฑ์รากจริงในเกร็ดท้ายหัวข้อ 12.1 เมื่อ $T, D \in \mathbb{R}$ รากทั้งสองของ $\lambda^2 - T\lambda + D = 0$ มีขนาดน้อยกว่า 1 ก็ต่อเมื่อ $|D| < 1$ และ $|T| < 1 + D$)
3. หาค่า b/k และอธิบายว่าทำไม $h = 1/100$ ของ Fiedler อยู่นอกช่วงนี้ ส่วน $h = 1/200$ อยู่ในช่วง
4. ตรวจว่าการคำนวณสูงสุดของ $|x|$ บน $[0, 100]$ วินาทีเป็นอย่างไรเมื่อ $h = 1/200$ และ $h = 1/100$

ฉบับร่าง · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 12.1.1.

1. ที่นี่ $f(t,y) = -2y$ และ $h = \frac{1}{4}$ เริ่มที่ $y_0 = 1$
2. ก้าวที่ 1: $y_1 = 1 + \frac{1}{4}(-2 \cdot 1) = \frac{1}{2}$
3. ก้าวที่ 2: $y_2 = \frac{1}{2} + \frac{1}{4}\big(-2 \cdot \frac{1}{2}\big) = \frac{1}{4}$
4. คำตอบจริงที่ $t = \frac{1}{2}$ คือ $e^{-1} \approx 0.36788$ วิธีให้ 0.25 ซึ่งต่ำกว่าจริงมาก เพราะ $h = \frac{1}{4}$ ยังใหญ่สำหรับสมการที่มีอัตราการลดลง 2

ตรวจด้วยโค้ด: `test_ch12.py` (`test_euler_one_step_by_hand`)

เฉลย 12.1.2.

1. สูตรคือ $x_{n+1} = x_n + h v_n$ และ $v_{n+1} = v_n + 10h$ โดย $v_0 = 0$ ดังนั้น $v_n = 10nh$
2. $x_n = \sum_{j=0}^{n-1} h\, v_j = \sum_{j=0}^{n-1} 10jh^2 = 5h^2 n(n-1)$
3. เมื่อ $h = \frac{1}{2}$ และ $t = 10$ จะได้ $n = 20$ จึง $x_{20} = 5 \cdot \frac{1}{4} \cdot 20 \cdot 19 = 475$ เมตร
4. ความคลาดเคลื่อนคือ $500 - 475 = 25$ เมตร เป็นครึ่งหนึ่งของ 50 เมตรที่ $h = 1$ สอดคล้องกับอันดับ 1

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_1_2_half_step_gives_475_m`)

เฉลย 12.1.3.

1. สำหรับ $x, y \in [0,1]$ ได้ $|x^2 - y^2| = |x+y|\,|x-y| \le 2|x-y|$ เพราะ $x + y \le 2$ จึงใช้ $L = 2$ ได้ และ $L = 2$ ดีที่สุดเพราะเมื่อ $x = 1$ และ $y \to 1$ อัตราส่วนเข้าหา 2
2. สำหรับ $x \neq y$ ใน $[0, R]$ อัตราส่วนเท่ากับ $x + y$ ค่าสูงสุดคือ $2R$ ซึ่งเข้าใกล้เมื่อ $x, y \to R$ ดังนั้น L ที่ใช้ได้ต้องมากกว่าหรือเท่ากับ $2R$ สำหรับทุก R จึงไม่มีค่าเดียว

ตัวเลขจากโค้ดที่ตาข่ายระยะ 0.005 บน $[0,1]$ คือ 1.995 และบน $[0,10]$ คือ 19.95 ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_1_3_square_is_lipschitz_on_unit_interval` และ `test_exercise_12_1_3_square_is_not_lipschitz_on_a_long_interval`)

เฉลย 12.1.4.

1. ตามวิธีออยเลอร์ $y_{k+1} = (1+h)y_k$ จึง $y_n = (1+h)^n$ เมื่อ $n = 1/h$ ได้ $y = (1+h)^{1/h}$ สำหรับ $h = \frac{1}{64}$ คือ $(65/64)^{64}$
2. คำตอบจริงคือ $e \approx 2.718282$ ความคลาดเคลื่อนคือ $e - (65/64)^{64} \approx 0.020937$
3. อัตราส่วนกับแถว $h = \frac{1}{32}$ คือ $0.041292 / 0.020937 \approx 1.9722$ เข้าใกล้ 2 ตามที่ตารางคาดไว้
4. เหตุผล: $(1+h)^{1/h} = e^{\ln(1+h)/h}$ และ $\ln(1+h)/h = 1 - \frac{h}{2} + O(h^2)$ ดังนั้นความคลาดเคลื่อนเป็นประมาณ $\frac{e}{2}h$ ซึ่งเป็นกำลังหนึ่งของ h

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_1_4_next_row_h_one_over_64` และ `test_model_error_against_closed_form`)

ฉบับร่าง Preprint · 2026-10-23

เฉลย 12.1.5.

1. $\det M_h = 1 \cdot 1 - h \cdot (-kh) = 1 + kh^2$ และ $\operatorname{tr} M_h = 1 + 1 = 2$
2. ค่าลักษณะเฉพาะเป็นรากของ $\lambda^2 - 2\lambda + (1 + kh^2) = 0$ คือ $\lambda = 1 \pm i\sqrt{kh^2}$ ขนาดคือ $|\lambda| = \sqrt{1 + kh^2}$ เมื่อ $h > 0$
3. เมื่อ $k = 15$ และ $h = 0.01$ ได้ $|\lambda| = \sqrt{1.0015} \approx 1.000750$
4. เพราะ $\mathbf{y}_n = M_h^n \mathbf{y}_0$ เมื่อ $|\lambda| > 1$ ส่วนประกอบของคำตอบในทิศของเวกเตอร์ลักษณะเฉพาะจะโตตามกำลัง $|\lambda|^n$ ทุกก้าว ไม่มีแรงหน่วงมาช่วยหยุด

ตรวจ ด้วย โค้ด: `test_ch12.py` (`test_exercise_12_1_5_undamped_modulus_is_sqrt_one_plus_kh2`)

เฉลย 12.1.6.

1. ฐาน $n = 0$: $x_0 = 0 = 5h^2 \cdot 0 \cdot (-1)$ จริง
2. สมมติว่าจริงถึง n จะได้ $x_{n+1} = x_n + hv_n = 5h^2 n(n-1) + h \cdot 10nh = 5h^2(n^2 - n + 2n) = 5h^2(n+1)n$ ซึ่งเป็นสูตรของ $n + 1$
3. ที่ $t = 10$ ได้ $n = 10/h$ แล้ว $x_n = 5h^2 \cdot \frac{10}{h} \cdot \frac{10-h}{h} = 50(10 - h) = 500 - 50h$ ดังนั้นความคลาดเคลื่อนคือ $50h$ พอดี
4. ความคลาดเคลื่อนเป็นเชิงเส้นใน h ตรงกับอันดับ 1 ในทฤษฎีบท ค่าคงที่ $50 = 5 \cdot 10$ ไม่ขึ้นกับ h จึงเป็นค่าที่เห็นได้จากการคำนวณโดยตรง

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_1_6_error_at_10_s_is_50h` และ `test_closed_form_position_after_n_steps`)

เฉลย 12.1.7.

1. $M_h = \begin{pmatrix} 1 & h \\ -kh & 1 - bh \end{pmatrix}$ จึง $\det M_h = (1 - bh) + kh^2$ และ $\operatorname{tr} M_h = 2 - bh$
2. สำหรับ $\lambda^2 - T\lambda + D = 0$ ที่มี T, D เป็นจำนวนจริง ทั้งสองรากอยู่ในวงกลมหนึ่งหน่วยก็ต่อเมื่อ $|D| < 1$ และ $|T| < 1 + D$ (เกร็ดท้ายหัวข้อ 12.1) เมื่อ $D = 1 - bh + kh^2$ และ $T = 2 - bh$ ได้ $1 + D - T = kh^2 > 0$ จึง $T < 1 + D$ เสมอ และ $1 + D + T = 4 - 2bh + kh^2$ มีดิสคริมิแนนต์ $4b^2 - 16k < 0$ จึงเป็นบวกเสมอ นั่นคือ $-T < 1 + D$ เสมอ ส่วน $D + 1 = 2 - bh + kh^2$ มีดิสคริมิแนนต์ $b^2 - 8k < 0$ จึง $D > -1$ เสมอ เหลือ $D < 1$ คือ $kh^2 < bh$ ซึ่งเทียบเท่ากับ $0 < h < b/k$ จึงได้เงื่อนไขที่ต้องการ
3. $b/k = 0.1/15 = 1/150$ เมื่อ $h = 1/100 > 1/150$ จึงอยู่นอกช่วง และผลการคำนวณสอดคล้องกับ Fiedler ที่ระบุว่าขั้น $1/100$ ไม่เสถียร ส่วน $h = 1/200 < 1/150$ อยู่ในช่วง
4. บน $[0, 100]$ วินาที เมื่อ $h = 1/200$ ค่าสูงสุดของ $|x|$ คือ 1000.0 ซึ่งเป็นค่าเริ่มต้นเอง และตำแหน่งที่ 100 วินาที คือ -181.3 เมตร เมื่อ $h = 1/100$ ค่าสูงสุดของ $|x|$ คือ $12\,108.7$ เมตร (สองค่านี้คือผลของเราที่พารามิเตอร์ของ Fiedler) ที่ขอบ $h = 1/150$ มี $|\lambda| = 1$ ผลจึงไม่ลดลงและไม่โตแบบเอ็กซ์โพเนนเชียล ค่าสูงสุดเกิน 1000 เพียงเล็กน้อยคือ 1000.08 จากการปรับตัวชั่วคราว

ฉบับร่าง · Preprint · 2026-10-23

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_1_7_threshold_is_b_over_k`, `test_exercise_12_1_7_determinant_and_trace`, `test_exercise_12_1_7_step_200_stays_bounded_over_100_s`, `test_exercise_12_1_7_step_150_overshoots_slightly`)

12.2 ก้าวให้นิ่งขึ้น

สถานการณ์: สปริงเดิม แต่เรียงคำสั่งใหม่

ลองเปลี่ยนโปรแกรมของสปริงเพียงบรรทัดเดียว เดิมเราเลื่อนตำแหน่งด้วยความเร็วเก่าก่อน แล้วค่อยเปลี่ยนความเร็ว ตอนนี้เปลี่ยนความเร็วก่อน แล้วเลื่อนตำแหน่งด้วยความเร็วใหม่ การแก้เพียงเท่านี้ผลต่างมาก ที่ขั้น $1/100$ วินาที สปริงเดิมที่โตจนเกิน 12,000 เมตรจะหยุดโต และค่า $|x|$ หลังเริ่มต้นไม่เกิน 998.5 เมตร (ค่าจากการคำนวณของเรา) บทความที่อ้างถึงนี้แนะนำวิธีนี้สำหรับสปริงตัวเดียวกัน Fiedler 2004b

เอกสารของ Box2D ระบุว่าขั้นเวลา $1/60$ วินาทีมักให้การจำลองคุณภาพสูง และโดยทั่วไปควรใช้ $1/30$ วินาทีหรือเล็กกว่า ส่วนจำนวนขั้นย่อยที่แนะนำคือ 4 Catto 2025 คำถามของหัวข้อนี้คือ วิธีไหนนิ่งจริง ทำไม และนิ่งแล้วจะไม่มีปัญหาเหลืออยู่หรือไม่

ภาพรวมก่อนอ่าน

ทำให้ก้าวนิ่งขึ้นได้สองทาง ทางแรกคืออ่านลูกศรที่ปลายก้าวแทนที่จุดเริ่ม ซึ่งต้องแก้สมการทุกก้าว ทางที่สองคือเรียงลำดับการอัปเดต ให้ความเร็วใหม่ช่วยเลื่อนตำแหน่ง ซึ่งไม่ต้องแก้สมการ ทั้งสองทางมีราคาต่างกันและผลต่างกัน เราจะเทียบด้วยลูกตุ้มและสปริง แล้วปิดหัวข้อด้วยอีกด้านหนึ่งของปัญหาเดียวกัน คือกำแพงบางที่ก้าวใหญ่ข้ามได้โดยไม่มีอะไรระเบิด

ลองเอง: หนึ่งก้าวด้วยมือ

ให้ $y' = -2y$, $y_0 = 1$ และ $h = \frac{1}{2}$ ถ้าอ่านลูกศรที่จุดเริ่ม ได้ $y_1 = 1 + \frac{1}{2}(-2) = 0$

ถ้าอ่านลูกศรที่ปลายก้าว ต้องแก้ $y_1 = 1 + \frac{1}{2}(-2y_1)$ จึงได้ $2y_1 = 1$ คือ $y_1 = \frac{1}{2}$ คำตอบจริงที่ $t = \frac{1}{2}$ คือ $e^{-1} \approx 0.368$ ค่า $\frac{1}{2}$ ใกล้คำตอบจริงกว่าค่า 0 ของการอ่านที่จุดเริ่ม

อ่านลูกศรที่ปลายก้าว

ในวิธีออยเลอร์ชัดแจ้ง ค่าใหม่คำนวณจากลูกศรที่จุดเริ่ม ลองเปลี่ยนเป็นอ่านลูกศรที่จุดสิ้นสุดแทน

นิยาม 12.2.1: วิธีออยเลอร์แบบโดยปริยาย

ให้ $\mathbf{y}_{k+1}$ เป็นค่าที่สอดคล้องกับสมการ

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h\, f(t_{k+1}, \mathbf{y}_{k+1}).$$

วิธีนี้เป็น**วิธีแบบโดยปริยาย** (implicit method) เพราะ $\mathbf{y}_{k+1}$ ปรากฏทั้งสองข้างของสมการ จึงต้องแก้สมการเพื่อหาค่าใหม่

ฉบับร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

เรายังไม่รู้ $\mathbf{y}_{k+1}$ จึงต้องหาค่าที่ทำให้สมการเป็นจริง ราคาของการทำแต่ละก้าวจึงสูงกว่าวิธีออยเลอร์ชัดแจ้ง ตัวอย่างที่ไม่ยาก: สำหรับ $y' = -2y$ ค่า $y_{k+1} = y_k/(1+2h)$ ซึ่งลดลงทุกก้าวไม่ว่า h จะเท่าไร (แบบฝึกหัด 12.2.2)

ทฤษฎีบท 12.2.1: การหาค่าใหม่ของวิธีแบบโดยปริยาย

ให้ f เป็นไปตามเงื่อนไขลิพชิทซ์ด้วยค่าคงที่ L และ $hL < 1$ แล้วสมการของวิธีออยเลอร์แบบโดยปริยายมีคำตอบ $\mathbf{w}$ เพียงคำตอบเดียว และการวนซ้ำ $\mathbf{z}_{j+1} = \mathbf{y}_k + h\,f(t_{k+1}, \mathbf{z}_j)$ จากค่าเริ่มต้น $\mathbf{z}_0$ ใด ๆ ทำให้

$$\|\mathbf{z}_j - \mathbf{w}\| \le (hL)^j \|\mathbf{z}_0 - \mathbf{w}\|.$$

พิสูจน์. ดูเพิ่มบทที่ 7 ทฤษฎีบท 7.1 ของ Iserles สำหรับทฤษฎีบทจุดตรึงที่ใช้ตรงนี้
แนวคิด. การวนซ้ำนี้เป็นการหดตัว แผนที่ $G(\mathbf{z}) = \mathbf{y}_k + hf(t_{k+1}, \mathbf{z})$ ทำให้ระยะห่างของสองจุดหดลงตัวคูณ hL จึงมี**จุดตรึง** (fixed point) เพียงจุดเดียว

1. **การหดตัว.** สำหรับ $\mathbf{x}, \mathbf{z}$ ใด ๆ $\|G(\mathbf{x}) - G(\mathbf{z})\| = h\|f(t_{k+1}, \mathbf{x}) - f(t_{k+1}, \mathbf{z})\| \le hL\|\mathbf{x} - \mathbf{z}\|$
2. **มีจุดตรึงเพียงจุดเดียว.** การมีอยู่ของจุดตรึงได้จาก**ทฤษฎีบทจุดตรึงของบานาค** (Banach fixed-point theorem) ซึ่งเราใช้โดยไม่พิสูจน์ซ้ำ ถ้ามีสองจุด $\mathbf{w}, \mathbf{w}'$ จะได้ $\|\mathbf{w} - \mathbf{w}'\| \le hL\|\mathbf{w} - \mathbf{w}'\|$ จึง $\mathbf{w} = \mathbf{w}'$ เพราะ $hL < 1$
3. **อัตราการลู่เข้า.** เพราะ $\mathbf{w} = G(\mathbf{w})$ จึง $\mathbf{z}_{j+1} - \mathbf{w} = G(\mathbf{z}_j) - G(\mathbf{w})$ และจากข้อ 1 ได้ $\|\mathbf{z}_{j+1} - \mathbf{w}\| \le hL\|\mathbf{z}_j - \mathbf{w}\|$ ทำซ้ำได้ผลที่ต้องการ

สรุป: การหาค่าใหม่ของวิธีแบบโดยปริยายคือการหาจุดตรึงของการหดตัว ทำได้ด้วยการวนซ้ำเมื่อ $hL < 1$ และในโค้ดของบทนี้เราใช้วิธีนิวตันแทน ซึ่งอธิบายในหัวข้อ 6--7 □

ค่าเฉลี่ยของสองปลาย

ถ้าใช้ความชันเฉลี่ยของจุดเริ่มและจุดสิ้นสุด เราจะได้วิธีที่แม่นกว่า

นิยาม 12.2.2: กฎเชิงสี่เหลี่ยมคางหมู

ให้ $\mathbf{y}_0$ เป็นค่าเริ่มต้น แล้วคำนวณ

$$\mathbf{y}_{k+1} = \mathbf{y}_k + \tfrac{h}{2}\big[f(t_k, \mathbf{y}_k) + f(t_{k+1}, \mathbf{y}_{k+1})\big].$$

วิธีนี้คือ**กฎเชิงสี่เหลี่ยมคางหมู** (trapezoidal rule) และเป็นวิธีแบบโดยปริยาย เพราะ $\mathbf{y}_{k+1}$ ปรากฏทั้งสองข้าง

วิธีออยเลอร์และกฎนี้เป็นสองกรณีของรูปทั่วไป $\mathbf{y}_{k+1} = \mathbf{y}_k + h[\theta f(t_k, \mathbf{y}_k) + (1-\theta) f(t_{k+1}, \mathbf{y}_{k+1})]$ เมื่อ $\theta = 1$ และ $\theta = \tfrac{1}{2}$ ตามลำดับ เรียกรูปทั่วไปนี้ว่า**วิธีทีตา** (theta method) Iserles 2009, §1.4, p. 13

ทฤษฎีบท 12.2.2: อันดับ 2 ของกฎเชิงสี่เหลี่ยมคางหมู

ให้ f เป็นไปตามเงื่อนไขลิพชิทซ์ด้วย $L > 0$ ให้คำตอบจริง $\mathbf{y}$ หาอนุพันธ์อันดับสามได้ต่อเนื่องบน $[t_0, t_0 + T]$ โดย $M_3 = \max \|\mathbf{y}'''(t)\|$ และ $hL \le 1$ ให้ $\mathbf{y}_k$ เป็นผลของกฎเชิงสี่เหลี่ยมคางหมู โดย $\mathbf{y}_0 = \mathbf{y}(t_0)$ แล้วสำหรับทุก k

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 12.2: ความคลาดเคลื่อนที่ $t = 1$ ของวิธีออยเลอร์ (จากตาราง 12.1) เทียบกับกฎเชิงสี่เหลี่ยมคางหมูสำหรับ $y' = y$, $y(0) = 1$ คอลัมน์ขวาคืออัตราส่วนของแถวก่อนหารด้วยแถวนี้ของกฎ (สร้างจาก `numanal/ch12.py` และตรวจด้วย `test_ch12.py`)

h	ออยเลอร์	กฎเชิงสี่เหลี่ยมคางหมู	อัตราส่วนของกฎ
$1/2$	0.468282	0.059496	—
$1/4$	0.276876	0.014330	4.1520
$1/8$	0.152497	0.003550	4.0364
$1/16$	0.080353	0.000886	4.0090
$1/32$	0.041292	0.000221	4.0022

ที่ $t_k \le t_0 + T$

$$\|\mathbf{y}(t_k) - \mathbf{y}_k\| \le \frac{M_3 h^2}{12L}\big(e^{2L(t_k - t_0)} - 1\big).$$

ดังนั้นความคลาดเคลื่อนวงกว้างเป็น $O(h^2)$ คืออันดับ 2 Iserles 2009, §1.3, Thm 1.2, p. 9 ดูเพิ่มบทที่ 1 ทฤษฎีบท 1.2 ของ Iserles

พิสูจน์. **แนวคิด.** ค่าคลาดเคลื่อนตัดปลายของหนึ่งก้าวเป็น h^3 แทน h^2 ของวิธีออยเลอร์ ส่วนการส่งต่อความผิดพลาดใช้ความเป็นการหดตัวเช่นเดิม บทพิสูจน์นี้เราเขียนเองตามแบบทฤษฎีบท 12.1.2

1. **ค่าคลาดเคลื่อนตัดปลายเฉพาะที่.** ให้ $\boldsymbol{\ell}_{k+1} = \mathbf{y}(t_{k+1}) - \mathbf{y}(t_k) - \frac{h}{2}[\mathbf{y}'(t_k) + \mathbf{y}'(t_{k+1})]$ และตั้ง $\varphi(s) = \mathbf{y}'(t_k + s)$ เมื่อ $0 \le s \le h$ จะได้ $\boldsymbol{\ell}_{k+1} = \int_0^h \varphi(s)\,ds - \frac{h}{2}[\varphi(0) + \varphi(h)]$ การหาปริพันธ์ทีละส่วนสองครั้งได้เอกลักษณ์
$$\int_0^h \varphi\,ds - \tfrac{h}{2}[\varphi(0)+\varphi(h)] = -\tfrac{1}{2}\int_0^h s(h-s)\varphi''(s)\,ds$$
ดังนั้น $\boldsymbol{\ell}_{k+1} = -\frac{1}{2}\int_0^h s(h-s)\,\mathbf{y}'''(t_k+s)\,ds$ และ $\|\boldsymbol{\ell}_{k+1}\| \le \frac{M_3}{2}\int_0^h s(h-s)\,ds = \frac{M_3 h^3}{12}$

2. **สมการของความผิดพลาด.** ให้ $\boldsymbol{\varepsilon}_k = \mathbf{y}(t_k) - \mathbf{y}_k$ นำสมการของวิธีลบออกจากนิยามของ $\boldsymbol{\ell}_{k+1}$ จะได้ $\boldsymbol{\varepsilon}_{k+1} = \boldsymbol{\varepsilon}_k + \frac{h}{2}\big[f(t_k,\mathbf{y}(t_k)) - f(t_k,\mathbf{y}_k)\big] + \frac{h}{2}\big[f(t_{k+1},\mathbf{y}(t_{k+1})) - f(t_{k+1},\mathbf{y}_{k+1})\big] + \boldsymbol{\ell}_{k+1}$

3. **ประมาณนอร์ม.** ใช้เงื่อนไขลิพชิทซ์กับพจน์สองพจน์ จะได้ $\|\boldsymbol{\varepsilon}_{k+1}\| \le \|\boldsymbol{\varepsilon}_k\| + \frac{hL}{2}\|\boldsymbol{\varepsilon}_k\| + \frac{hL}{2}\|\boldsymbol{\varepsilon}_{k+1}\| + \|\boldsymbol{\ell}_{k+1}\|$ ซึ่งจัดได้เป็น $\big(1 - \frac{hL}{2}\big)\|\boldsymbol{\varepsilon}_{k+1}\| \le \big(1 + \frac{hL}{2}\big)\|\boldsymbol{\varepsilon}_k\| + \|\boldsymbol{\ell}_{k+1}\|$ และ $1 - \frac{hL}{2} > 0$

4. **ทำซ้ำตามลำดับก้าว.** ตั้ง $q = \dfrac{1 + hL/2}{1 - hL/2}$ และ $r = \dfrac{M_3 h^3}{12(1 - hL/2)}$ จะได้ $\|\boldsymbol{\varepsilon}_{k+1}\| \le q\|\boldsymbol{\varepsilon}_k\| + r$ เพราะ $\boldsymbol{\varepsilon}_0 = \mathbf{0}$ จึง $\|\boldsymbol{\varepsilon}_k\| \le r\,\dfrac{q^k - 1}{q - 1}$ และ $q - 1 = \dfrac{hL}{1 - hL/2}$ ดังนั้น $\dfrac{r}{q-1} = \dfrac{M_3 h^2}{12L}$

5. **ใช้ $hL \le 1$ และ $1 + x \le e^x$.** เพราะ $1 - \frac{hL}{2} \ge \frac{1}{2}$ จึง $q - 1 \le 2hL$ และ $q^k \le e^{k(q-1)} \le e^{2khL} = e^{2L(t_k - t_0)}$ ซึ่งให้ขอบเขตที่ต้องการ

สรุป: ค่าคลาดเคลื่อนของหนึ่งก้าวเป็น h^3 และมีประมาณ $1/h$ ก้าว ผลรวมจึงเป็น h^2 เร็วกว่าวิธีออยเลอร์หนึ่งอันดับ □

ตารางบอกว่าเมื่อ h ลดครึ่งหนึ่ง ความคลาดเคลื่อนของกฎเชิงสี่เหลี่ยมคางหมูลดลงประมาณ $4 = 2^2$ เท่า จึงเป็นอันดับ 2 ส่วนของวิธีออยเลอร์ยังลดลงเพียงประมาณ 2 เท่า เหตุผลเชิงคำนวณอยู่ในแบบฝึกหัด 12.2.4

ฉบับร่าง · Preprint · 2026-10-23

ก้าวกึ่งโดยปริยาย: สลับบรรทัดเดียว

ลองกลับไปดูสปริงที่ทำให้ระเบิด วิธีออยเลอร์ชัดแจ้งเลื่อนตำแหน่งด้วยความเร็วเก่า แต่ถ้าเราคำนวณความเร็วใหม่ก่อนแล้วใช้ความเร็วใหม่เลื่อนตำแหน่ง จะได้วิธีที่ไม่ต้องแก้สมการ

ลองเอง: หนึ่งก้าวของสปริงที่ถูกผลัก

ให้สปริงของหัวข้อ 12.1 ($m = 1$, $k = 15$, $b = 0.1$) เริ่มที่ $x_0 = 1000$ และ $v_0 = 0$ ขั้น $h = \frac{1}{100}$ ความเร่งที่จุดเริ่มคือ $a = -15 \cdot 1000 - 0.1 \cdot 0 = -15000$
วิธีชัดแจ้งเลื่อนตำแหน่งด้วยความเร็วเก่า $x_1 = x_0 + h\,v_0 = 1000$ แล้วเปลี่ยนความเร็ว $v_1 = v_0 + h\,a = -150$
วิธีซิมเพลกติกเปลี่ยนความเร็วก่อน $v_1 = -150$ แล้วเลื่อนตำแหน่งด้วยความเร็วใหม่ $x_1 = x_0 + h\,v_1 = 1000 + \frac{1}{100}(-150) = 998.5$

นิยาม 12.2.3: วิธีออยเลอร์ซิมเพลกติก

สำหรับระบบ $x' = v$ และ $v' = a(x, v)$ วิธีนี้คือ

$$v_{k+1} = v_k + h\,a(x_k, v_k), \qquad x_{k+1} = x_k + h\,v_{k+1}.$$

ชื่อภาษาไทยที่ใช้คือ**วิธีออยเลอร์ซิมเพลกติก** (symplectic Euler method) ส่วนบทความที่อ้างถึงนี้เรียกวิธีนี้ว่า semi-implicit Euler Fiedler 2004b

เกร็ดเพิ่มเติม (ผู้เรียบเรียง): ทำไมจึงมีคำว่า “กึ่ง”

วิธีนี้ยังเป็นวิธีชัดแจ้งตามนิยามของหัวข้อ 12.1 เพราะค่าใหม่คำนวณได้ตรง ๆ โดยไม่ต้องแก้สมการ คำว่า "กึ่ง" หมายถึงการที่ตำแหน่งใช้ความเร็วใหม่ ซึ่งเป็นการอ่านลูกศรที่ปลายก้าวบางส่วน เราจะไม่อ้างสมบัติซิมเพลกติกของวิธีนี้ในบทนี้ เราดูเพียงผลที่วัดได้บนสปริงและลูกตุ้ม

ทฤษฎีบท 12.2.3: การแกว่งที่ไม่มีแรงหน่วง

ให้ $x'' = -\omega^2 x$ โดย $\omega > 0$ และ $h > 0$ ให้ λ เป็นค่าลักษณะเฉพาะของการก้าวหนึ่งครั้งบน (x, v) แล้ว

1. วิธีออยเลอร์ชัดแจ้งให้ $|\lambda| = \sqrt{1 + \omega^2 h^2}$ เสมอ;
2. วิธีออยเลอร์ซิมเพลกติกให้ $|\lambda| = 1$ ก็ต่อเมื่อ $\omega h \le 2$ เมื่อ $\omega h \le 2$ มุมการหมุนต่อก้าวคือ $\theta = 2\arcsin(\omega h/2)$ และเมื่อ $0 < \omega h < 2$ คำตอบถูกจำกัดขอบเขต;
3. กฎเชิงสี่เหลี่ยมคางหมูให้ $|\lambda| = 1$ เสมอ และมุมการหมุนต่อก้าวคือ $\theta = 2\arctan(\omega h/2)$.

คำตอบจริงหมุนมุม ωh ต่อก้าว

พิสูจน์. **แนวคิด.** ทุกวิธีเป็นเมทริกซ์ 2×2 ที่มีดีเทอร์มิแนนต์และรอยที่คำนวณได้ตรง ๆ ส่วนกฎเชิงสี่เหลี่ยมคางหมูเปลี่ยนค่าลักษณะเฉพาะของ A โดยตรงเป็นค่าที่ขนาดหนึ่ง

1. **วิธีชัดแจ้ง.** เมทริกซ์คือ $\begin{pmatrix} 1 & h \\ -\omega^2 h & 1 \end{pmatrix}$ จึง $\det = 1 + \omega^2 h^2$ และ $\mathrm{tr} = 2$ ค่าลักษณะเฉพาะเป็นรากของ $\lambda^2 -$

ฉบับร่าง · Preprint · 2026-10-23

$2\lambda + (1+\omega^2h^2) = 0$ จึงมีขนาด $\sqrt{1+\omega^2h^2}$

2. **วิธีซิมเพลกติก.** เมทริกซ์คือ $\begin{pmatrix} 1-\omega^2h^2 & h \\ -\omega^2 h & 1 \end{pmatrix}$ จึง $\det = 1$ และ $\operatorname{tr} = 2-\omega^2h^2 =: T$ ถ้า $-2 < T < 2$ รากเป็นคู่เชิงซ้อนสังยุคที่ผลคูณเป็น 1 จึงมีขนาด 1 ถ้า $T = -2$ รากซ้ำเป็น -1 มีขนาด 1 ถ้า $T < -2$ รากเป็นจำนวนจริงที่ผลคูณเป็น 1 ขนาดของรากหนึ่งจึงมากกว่า 1 และ $T \geq -2$ เท่ากับ $\omega h \leq 2$ เขียน $2\cos\theta = T$ จะได้ $\sin^2(\theta/2) = \omega^2h^2/4$ จึง $\theta = 2\arcsin(\omega h/2)$ สำหรับ $0 < \omega h < 2$ ตรวจได้ว่า $I = \omega^2x^2 + v^2 - \omega^2 h\, xv$ ไม่เปลี่ยนตามก้าว (แทน $v' = v - \omega^2 hx$ และ $x' = x + hv'$) และ $\omega^2 h\,|xv| \leq \tfrac{1}{2}\omega h(\omega^2x^2+v^2)$ จึง $I \geq (1-\tfrac{1}{2}\omega h)(\omega^2x^2+v^2)$ ดังนั้น $\omega^2 x_k^2 + v_k^2 \leq I_0/(1-\tfrac{1}{2}\omega h)$ ทุก k คือคำตอบถูกจำกัดขอบเขต

3. **กฎเชิงสี่เหลี่ยมคางหมู.** ให้ $A = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}$ เมทริกซ์หนึ่งก้าวคือ $(I - \tfrac{h}{2}A)^{-1}(I + \tfrac{h}{2}A)$ ถ้า $Av = i\omega v$ จะได้ค่าลักษณะเฉพาะ $\dfrac{1 + i\omega h/2}{1 - i\omega h/2}$ ซึ่งมีขนาด 1 และมุม $2\arctan(\omega h/2)$ (อีกค่าเป็นสังยุค)

สรุป: วิธีออยเลอร์ชัดแจ้งโตทุกก้าว วิธีซิมเพลกติกมีขนาดค่าลักษณะเฉพาะเท่ากับ 1 เมื่อ $\omega h \leq 2$ และคำตอบถูกจำกัดขอบเขตเมื่อ $\omega h < 2$ แต่มุมเกินจริงเล็กน้อย และกฎเชิงสี่เหลี่ยมคางหมูไม่โตและมุมขาดเล็กน้อย □

ข้อควรระวัง: ที่ $\omega h = 2$ คำตอบอาจโตเชิงเส้น

ที่ $\omega h = 2$ ค่าลักษณะเฉพาะมีขนาด 1 แต่เป็นรากซ้ำที่ -1 และเมทริกซ์ไม่ทแยงมุม ลองดู $\omega = 1$ และ $h = 2$ จากค่าเริ่มต้น $(x_0, v_0) = (1, 0)$ จะได้ $v_1 = -2$, $x_1 = -3$, $v_2 = 4$, $x_2 = 5$ ซึ่งเป็น $x_k = (-1)^k(2k+1)$ ดังนั้น $|x_k|/k \to 2$ คำตอบโตเชิงเส้น จึงไม่ถูกจำกัดขอบเขต ส่วนที่ $\omega h = 1.99$ ขอบเขตจากการพิสูจน์ข้างบนให้ $|x| \leq \sqrt{200} \approx 14.1$ เมื่อ $\omega = 1$ และ $(x_0, v_0) = (1, 0)$

ตัวเลขจากทฤษฎีบทนี้ที่ $\omega^2 = 15$ และ $h = 0.1$ คือ $\omega h = 0.3873$ มุมต่อก้าวของคำตอบจริงคือ 0.38730 วิธีซิมเพลกติกได้ 0.38976 และกฎเชิงสี่เหลี่ยมคางหมูได้ 0.38256 มุมของวิธีซิมเพลกติกต่างจากค่าจริงประมาณ 0.6% และของกฎเชิงสี่เหลี่ยมคางหมูประมาณ 1.2% แต่ทิศของความคลาดเคลื่อนต่างกัน

ลูกตุ้มที่เก็บพลังงาน

ลองดูลูกตุ้ม $\theta'' = -\sin\theta$ ที่ปล่อยจาก 1 เรเดียนโดยนิ่ง ๆ พลังงานต่อหน่วยมวลคือ $\tfrac{1}{2}\theta'^2 + (1-\cos\theta)$ ซึ่งคงที่บนคำตอบจริง ภาพที่ 12.3 แสดงความคลาดเคลื่อนสัมพัทธ์ของพลังงานเมื่อใช้ขั้นเวลา 0.1 วินาทีตลอด 100 วินาที

ผลนี้สอดคล้องกับทฤษฎีบทข้างต้น วิธีชัดแจ้งทำให้พลังงานโต วิธีซิมเพลกติกไม่โตขึ้นเรื่อย ๆ แต่ยังแกว่งจากพลังงานเริ่มต้น ส่วนกฎเชิงสี่เหลี่ยมคางหมูแม่นที่สุดในกรณีนี้ โดยแลกกับการต้องแก้สมการทุกก้าว

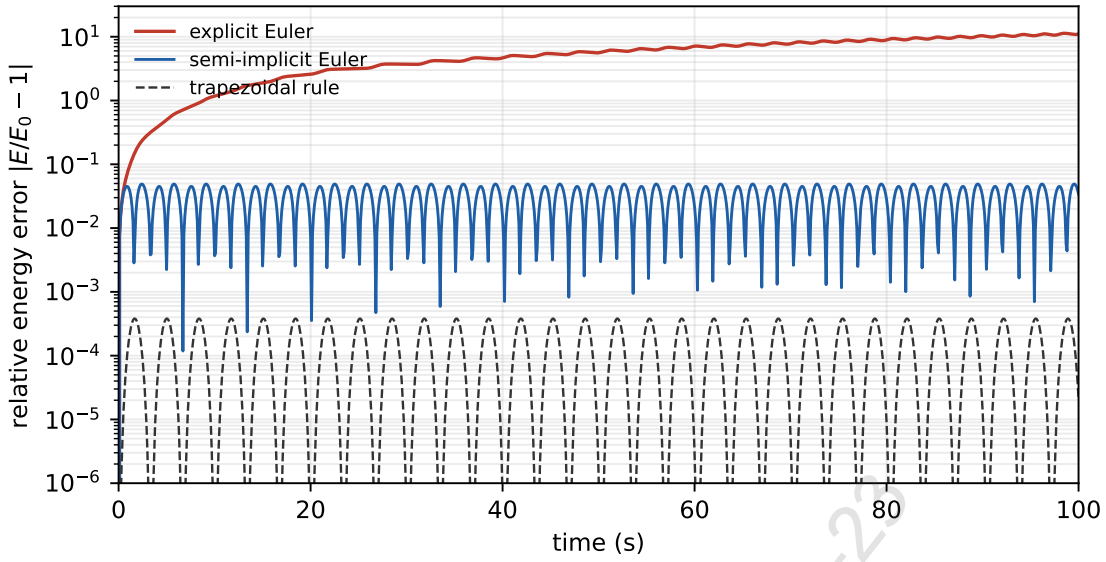
สปริงที่มีแรงหน่วงกับขั้นที่ใหญ่ขึ้น

กลับมาที่สปริงของหัวข้อ 12.1 ($m = 1$, $k = 15$, $b = 0.1$) หนึ่งก้าวของวิธีซิมเพลกติกบนสปริงนี้เป็นเมทริกซ์

$$\begin{pmatrix} 1 - h^2k & h(1-hb) \\ -hk & 1 - hb \end{pmatrix} \begin{pmatrix} x \\ v \end{pmatrix}$$

ดีเทอร์มิแนนต์คือ $1 - hb$ และรอยคือ $2 - hb - h^2k$

ฉบับร่าง preprint · 2026-10-23



รูปที่ 12.3: ความคลาดเคลื่อนสัมพัทธ์ของพลังงาน $|E/E_0 - 1|$ ของลูกตุ้มที่ $\theta(0) = 1$ และ $\theta'(0) = 0$ ด้วยขั้นเวลา 0.1 วินาที เส้นสีแดงคือวิธีออยเลอร์ชัดแจ้งที่โตขึ้นจนพลังงานเป็น 12 เท่าที่ 100 วินาที เส้นสีน้ำเงินคือวิธีซิมเพลกติกที่แกว่งรอบ 1 ไม่เกินประมาณ 5% เส้นประคือกฎเชิงสี่เหลี่ยมคางหมูที่ต่ำกว่า 4×10^{-4} (สร้างจาก `numanal/ch12.py`)

ทฤษฎีบท 12.2.4: ขั้นที่ใหญ่ที่สุดสำหรับสปริงที่มีแรงหน่วง

วิธีออยเลอร์ซิมเพลกติกนิ่งบนสปริงนี้ (รัศมีสเปกตรัมน้อยกว่า 1) ก็ต่อเมื่อ $kh^2 + 2bh < 4$ คือ $0 < h < h^*$ โดย

$$h^* = \frac{-b + \sqrt{b^2 + 4k}}{k}.$$

สำหรับ $k = 15$ และ $b = 0.1$ ได้ $h^* \approx 0.50977$ ส่วนวิธีออยเลอร์ชัดแจ้งนิ่งเมื่อ $h < b/k = 1/150$ (หัวข้อ 12.1)

พิสูจน์. **แนวคิด.** ค่าลักษณะเฉพาะเป็นรากของ $\lambda^2 - T\lambda + D = 0$ ซึ่งนิ่งก็ต่อเมื่อ $|D| < 1$ และ $|T| < 1 + D$ เกณฑ์รากของสมการกำลังสองจริงที่พิสูจน์ในเกร็ดท้ายหัวข้อ 12.1 ซึ่งใช้ในแบบฝึกหัด 12.1.7 ด้วย

1. **ค่าของ D และ T.** $D = 1 - hb$ และ $T = 2 - hb - h^2k$ จึง $1 + D - T = h^2k > 0$ เสมอ
2. **เงื่อนไข $|D| < 1$.** เพราะ $h, b > 0$ ได้ $D < 1$ และ $D > -1$ เทียบเท่า $bh < 2$ ซึ่งตามมาจากข้อ 3
3. **เงื่อนไข $|T| < 1 + D$.** $T < 1 + D$ จากข้อ 1 ส่วน $-T < 1 + D$ เทียบเท่ากับ $-2 + hb + h^2k < 2 - hb$ คือ $h^2k + 2hb < 4$
4. **รวมเงื่อนไข.** ได้ $kh^2 + 2bh < 4$ ซึ่งให้ $h < h^*$ โดยสูตรกำลังสอง และทำให้ $bh < 2$ ด้วย

สรุป: สำหรับ $k = 15$ และ $b = 0.1$ วิธีซิมเพลกติกนิ่งได้ไกลกว่าวิธีชัดแจ้งประมาณ 76 เท่า □

ข้อควรระวัง: นิ่งไม่ได้แปลว่าแม่น

ที่ขั้น $1/100$ วินาที รัศมีสเปกตรัมของวิธีซิมเพลกติกคือ 0.9995 ซึ่งนิ่ง แต่ที่ $t = 100$ วินาทีค่า x คือ -3.76 เมตร ขณะที่คำตอบจริงคือ -4.51 เมตร ความนิ่งบอกว่าค่าไม่โตขึ้นเรื่อย ๆ แต่ไม่ได้บอกว่าเฟสถูกต้อง เฟสต่างจากค่าจริง ซึ่ง

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 12.3: สัดส่วนของการผ่านกำแพงที่ถูกจับได้ ที่ความเร็วต่าง ๆ $w = 0.10$ เมตร $\Delta t = 1/60$ วินาที (สร้างจาก `numanal/ch12.py` และตรวจด้วย `test_ch12.py`)

ความเร็ว (m/s)	ระยะต่อเฟรม $v\Delta t$ (m)	จากสูตร	จากการจำลอง
3	0.050	1.00	1.0000
6	0.100	1.00	1.0000
12	0.200	0.50	0.5019
20	0.333	0.30	0.2994
30	0.500	0.20	0.2018
60	1.000	0.10	0.1026
300	5.000	0.02	0.0215

เห็นได้จากมุมต่อก้าวในทฤษฎีบท 12.2.3

อีกด้านหนึ่งของปัญหาเดียวกัน: กำแพงบาง

นี่คืออีกด้านหนึ่งของปัญหาเดียวกัน ขั้นที่นิ่งแล้วยังพลาดเหตุการณ์ได้ ลองนึกถึงลูกกระสุนที่พุ่งเร็วผ่านกำแพงบางระหว่างสองเฟรม มันอาจข้ามกำแพงไปโดยโปรแกรมไม่เคยเห็นมันอยู่ในกำแพงเลย ไม่มีอะไรระเบิด ค่าทุกค่ายังเล็กและนิ่ง ปัญหาอยู่ที่การ**ชักตัวอย่าง** (sampling) คือโปรแกรมดูตำแหน่งเฉพาะเวลา $k\Delta t$ เท่านั้น

ให้จุดเคลื่อนที่ด้วยความเร็วคงที่ v ผ่านกำแพงที่ตำแหน่ง $[0, w]$ โปรแกรมดูตำแหน่งทุก Δt โดยเริ่มที่เวลาสุ่ม ถ้าระยะที่เลื่อนต่อเฟรม $v\Delta t$ ใหญ่กว่า w ความน่าจะเป็นที่มีเฟรมตกในกำแพงคือ $w/(v\Delta t)$ ถ้าเล็กกว่า w ทุกเฟรมจะจับได้ จึงได้

$$\text{สัดส่วนที่จับได้} = \min\!\left\{1,\ \frac{w}{v\,\Delta t}\right\}.$$

ที่ $w = 0.10$ เมตร และ $\Delta t = 1/60$ วินาที ได้ตารางต่อไปนี้ (สูตรกับการจำลองด้วยเฟสสุ่ม 20 000 ค่าและเมล็ด 3)

สังเกตว่าการทะลุผ่าน**การทะลุผ่าน** (tunnelling) ไม่ใช่ความไม่เสถียร ไม่มีค่าใดโต ไม่มีค่าใดกระเพื่อม เหตุการณ์เพียงเกิดระหว่างเฟรมแล้วไม่มีใครวัด ส่วนของ x ที่ผ่านกำแพงจึงไม่ได้ถูกคำนวณ และการทำให้ขั้นเล็กลงช่วยได้ก็เพราะทำให้ $v\Delta t \le w$ เท่านั้น

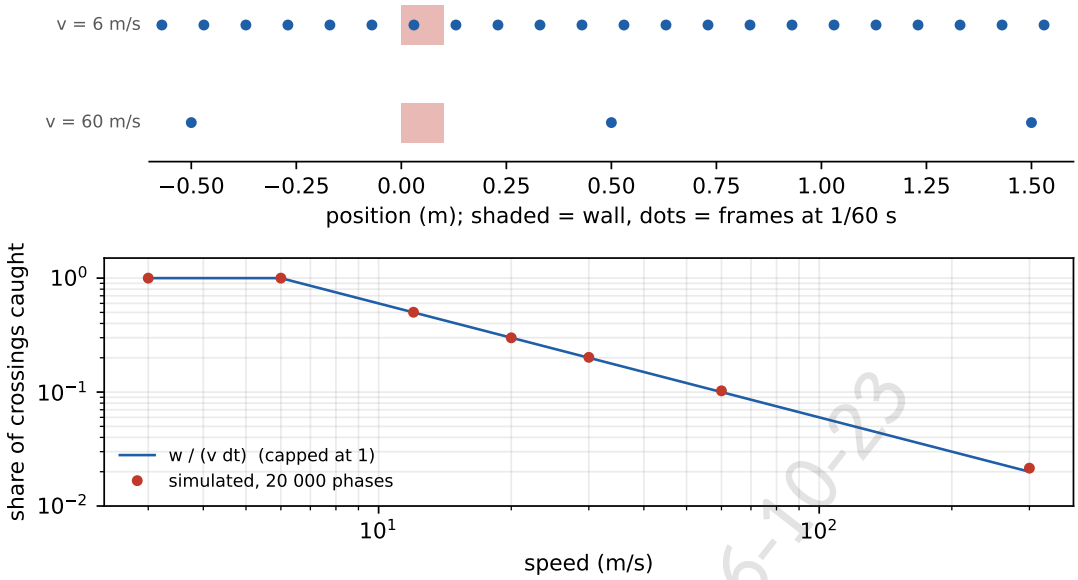
นักวิ่งที่ 6 m/s มีรัศมีตัว 0.25 เมตร จะทะลุกำแพงนี้ไม่ได้ เพราะจุดศูนย์กลางจะต้องอยู่ในช่วงยาว $w + 2r = 0.6$ เมตร ซึ่งยาวกว่าระยะต่อเฟรม 0.1 เมตร จุดศูนย์กลางจึงตกในช่วงนี้อย่างน้อยหนึ่งเฟรมเสมอ

ทางแก้ที่ตรงกับเหตุผลนี้คือดูส่วนของเส้นระหว่างตำแหน่งสองเฟรมแทนที่จะดูจุดเดียว เอกสารของ Box2D ระบุว่า โดยค่าเริ่มต้น Box2D ใช้**การตรวจการชนแบบต่อเนื่อง** (continuous collision detection) (CCD) เพื่อกันวัตถุที่เคลื่อนที่ทะลุวัตถุนิ่ง Catto 2025 เราไม่ได้อ้างถึงเครื่องยนต์อื่นในบทนี้

ข้อควรระวัง: ตรวจหลังก้าวไม่พอ

การตรวจเฉพาะตำแหน่งที่เฟรมจะพลาดได้เสมอเมื่อระยะที่วัตถุเลื่อนต่อเฟรมมากกว่าความหนาของกำแพง การเปลี่ยนไปใช้วิธีที่นิ่งกว่า เช่นวิธีซิมเพลกติก ไม่ช่วยเรื่องนี้ เพราะเหตุการณ์ไม่ได้ขึ้นกับความนิ่งของการคำนวณ มันขึ้นกับระยะที่วัตถุเลื่อนระหว่างการตรวจ

ร่างเรียน Preprint · 2026-10-23



รูปที่ 12.4: บนคือจุดที่ตรวจทุก $1/60$ วินาทีที่ความเร็ว 6 m/s (ตกในกำแพงจึงจับได้) และ 60 m/s (กระโดดข้ามกำแพง) บนพื้นที่แรเงาคือกำแพง ล่างคือสัดส่วนที่จับได้เทียบกับความเร็ว จุดคือการจำลอง เส้นคือสูตร (สร้างจาก `numanal/ch12.py`)

สรุปใจความ

- วิธีแบบโดยปริยายอ่านลูกศรที่ปลายก้าว ต้องแก้สมการทุกก้าว แต่เมื่อ $hL < 1$ การแก้นี้ทำได้ด้วยการหดตัว
- กฎเชิงสี่เหลี่ยมคางหมูมีอันดับ 2 เพราะค่าคลาดเคลื่อนตัดปลายของหนึ่งก้าวเป็น h^3
- วิธีออยเลอร์ชิมเพลกติกไม่ต้องแก้สมการ และบนสปริงนี้นิ่งเมื่อ $h < h^{*} \approx 0.5098$ ซึ่งใหญ่กว่า $1/150$ ของวิธีชัดแจ้งประมาณ 76 เท่า
- การ ข้าม กำแพง บาง เป็น ความ ผิด พลาด จาก การ ชัก ตัวอย่าง ไม่ใช่ ความ ไม่ เสถียร สัดส่วน ที่ จับ ได้ คือ $\min\{1, w/(v\Delta t)\}$

อ่านต่อ. ดูเพิ่มเติมเรื่องกฎเชิงสี่เหลี่ยมคางหมูและวิธีทีตาได้ที่ Iserles 2009, §1.3--1.4, pp. 8--18 หัวข้อ 12.3 จะใช้ตำแหน่งสองจุดก่อนหน้าแทนการเก็บความเร็ว

ร่างสมุดบันทึก · Preprint · 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 12.2.1 (⋆, ทำด้วยมือ) ให้ $y' = -2y$ และ $y(0) = 1$ ทำหนึ่งก้าวด้วย $h = \frac{1}{2}$ ทั้งวิธีออยเลอร์ชัดแจ้งและวิธีออยเลอร์แบบโดยปริยาย แล้วหาว่าค่าใดใกล้ $y(\frac{1}{2}) = e^{-1}$ มากกว่า

แบบฝึกหัด 12.2.2 (⋆⋆, พิสูจน์) ให้ $y' = -2y$

1. แสดงว่าวิธีออยเลอร์แบบโดยปริยายให้ $y_{k+1} = y_k/(1+2h)$ และถ้า $y_k > 0$ แล้ว $0 < y_{k+1} < y_k$ สำหรับทุก $h > 0$
2. แสดงว่าวิธีออยเลอร์ชัดแจ้งให้ $y_{k+1} = (1-2h)y_k$ และที่ $h = \frac{3}{2}$ ได้ $|y_{k+1}| = 2|y_k|$
3. เทียบข้อ 1 กับข้อ 2 แล้วอธิบายว่าทำไมการเลือกขนาดขั้นจึงไม่ทำให้วิธีแบบโดยปริยายโตในกรณีนี้

แบบฝึกหัด 12.2.3 (⋆ ⋆ ⋆, พิสูจน์) ให้ g มีอนุพันธ์อันดับสองต่อเนื่องบน $[0, h]$

1. การหาปริพันธ์ทีละส่วนสองครั้งเพื่อแสดงว่า
$$\int_0^h g\,ds - \tfrac{h}{2}\big(g(0)+g(h)\big) = -\tfrac{1}{2}\int_0^h s(h-s)\,g''(s)\,ds$$
(ใบ้: เริ่มจาก $\int_0^h s(h-s)g''(s)\,ds$ แล้วหาปริพันธ์ทีละส่วนทีละครั้ง)
2. ตรวจด้วย $g(s) = s^3$ และ $h = 0.3$ ว่าทั้งสองข้างเท่ากับ $-h^4/4$
3. ใช้ข้อ 1 กับ $g = \varphi$ เมื่อ $\varphi(s) = \mathbf{y}'(t_k + s)$ เพื่อแสดงว่าค่าคลาดเคลื่อนตัดปลายของกฎเชิงสี่เหลี่ยมคางหมูหนึ่งก้าวมีขนาดไม่เกิน $M_3 h^3/12$ เมื่อ $M_3 = \max \|\mathbf{y}'''\|$

แบบฝึกหัด 12.2.4 (⋆⋆, คำนวณ) ให้ $y' = y$ และ $y(0) = 1$ กฎเชิงสี่เหลี่ยมคางหมูหนึ่งก้าวคือ $y_{k+1} = \dfrac{1+h/2}{1-h/2}\,y_k$

1. หาความคลาดเคลื่อนที่ $t = 1$ เมื่อ $h = \frac{1}{64}$ และอัตราส่วนกับแถว $h = \frac{1}{32}$ ในตารางหัวข้อ 12.2
2. แสดงว่า $\ln \dfrac{1+h/2}{1-h/2} = h + \dfrac{h^3}{12} + O(h^5)$ และใช้ผลนี้อธิบายว่าความคลาดเคลื่อนมีค่าประมาณ $\dfrac{e}{12}h^2$

แบบฝึกหัด 12.2.5 (⋆ ⋆ ⋆, พิสูจน์) ให้ $x'' = -\omega^2 x$ เขียนเป็นระบบ (x, v) และให้ $\omega h > 0$ วิธีออยเลอร์ซิมเพลกติกมีเมทริกซ์หนึ่งก้าวคือ $\begin{pmatrix} 1-\omega^2 h^2 & h \\ -\omega^2 h & 1 \end{pmatrix}$

1. แสดงว่า $\det = 1$ และ $\mathrm{tr} = 2 - \omega^2 h^2$
2. แสดงว่าค่าลักษณะเฉพาะทั้งสองมีขนาด 1 ก็ต่อเมื่อ $\omega h \le 2$
3. เมื่อ $\omega h < 2$ แสดงว่ามุม θ ของค่าลักษณะเฉพาะ $e^{\pm i\theta}$ คือ $\theta = 2\arcsin(\omega h/2)$ และ $\theta > \omega h$

แบบฝึกหัด 12.2.6 (⋆⋆, คำนวณ) ให้สปริงของหัวข้อ 12.1 ($m = 1$, $k = 15$, $b = 0.1$) และใช้วิธีออยเลอร์ซิมเพลกติก

1. หาขนาดขั้น h^* ที่ใหญ่ที่สุดซึ่งยังนิ่ง จากสมการ $kh^2 + 2bh = 4$
2. เทียบ h^* กับขั้น $b/k = 1/150$ ของวิธีออยเลอร์ชัดแจ้ง ว่าใหญ่กว่ากี่เท่า
3. ตรวจว่า $h = \frac{1}{100}$ และ $h = \frac{1}{2}$ นิ่ง แต่ $h = 0.52$ ไม่นิ่ง โดยดูรัศมีสเปกตรัมของเมทริกซ์หนึ่งก้าว

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด 12.2.7 (⋆ ⋆ ⋆, โครงงานเล็ก) กำแพงหนา $w = 0.10$ เมตร โปรแกรมตรวจตำแหน่งทุก $\Delta t = 1/60$ วินาที จุดเคลื่อนที่ด้วยความเร็ว v คงที่ และสัดส่วนที่ถูกจับได้คือ $\min\{1,\, w/(v\Delta t)\}$

1. ที่ความเร็ว 12 m/s สัดส่วนที่จับได้คือเท่าไร และที่ความเร็วเท่าใดสัดส่วนนี้เหลือ 0.1
2. ความเร็ว 60 m/s จะถูกจับได้ทุกครั้งเมื่อ Δt เล็กเพียงใด (ตอบเป็นค่าของ Δt)
3. นักวิ่ง 6 m/s มีรัศมีตัว 0.25 เมตร จะทะลุกำแพงนี้ได้หรือไม่ อธิบายด้วยช่วง $w + 2r$

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 12.2.1.

1. วิธีชัดแจ้ง: $y_1 = 1 + \frac{1}{2}(-2 \cdot 1) = 0$
2. วิธีแบบโดยปริยาย: $y_1 = 1 + \frac{1}{2}(-2y_1)$ จึง $2y_1 = 1$ คือ $y_1 = \frac{1}{2}$
3. คำตอบจริง $e^{-1} \approx 0.368$ ระยะห่างของวิธีชัดแจ้งคือ 0.368 และของวิธีแบบโดยปริยายคือ 0.132 จึงค่าของวิธีแบบโดยปริยายใกล้กว่า

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_2_1_one_step_explicit_and_implicit`)

เฉลย 12.2.2.

1. สูตรแบบโดยปริยายคือ $y_{k+1} = y_k + h(-2y_{k+1})$ จึง $(1 + 2h)y_{k+1} = y_k$ คือ $y_{k+1} = y_k/(1 + 2h)$ เพราะ $1 + 2h > 1$ ถ้า $y_k > 0$ แล้ว $0 < y_{k+1} < y_k$ สำหรับทุก $h > 0$
2. สูตรชัดแจ้งคือ $y_{k+1} = y_k + h(-2y_k) = (1 - 2h)y_k$ เมื่อ $h = \frac{3}{2}$ ได้ $1 - 2h = -2$ จึง $|y_{k+1}| = 2|y_k|$
3. ตัวคูณของวิธีแบบโดยปริยายคือ $1/(1 + 2h)$ ซึ่งมีขนาดน้อยกว่า 1 สำหรับทุก $h > 0$ ส่วนตัวคูณของวิธีชัดแจ้งคือ $1 - 2h$ ซึ่งมีขนาดมากกว่า 1 เมื่อ $h > 1$ ดังนั้นวิธีแบบโดยปริยายไม่โตสำหรับทุกขนาดขั้นในตัวอย่างนี้

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_2_2_implicit_decay_for_any_step`)

เฉลย 12.2.3.

1. ให้ $I = \int_0^h s(h-s)g''(s)\,ds$ การหาปริพันธ์ทีละส่วนครั้งแรกด้วย $u = s(h-s)$ จะได้ $I = \big[s(h-s)g'\big]_0^h - \int_0^h (h-2s)g'\,ds = -\int_0^h (h-2s)g'\,ds$ หาปริพันธ์อีกครั้งได้ $\int_0^h (h-2s)g'\,ds = \big[(h-2s)g\big]_0^h + 2\int_0^h g\,ds = -h\big(g(0)+g(h)\big) + 2\int_0^h g\,ds$ ดังนั้น $I = h\big(g(0)+g(h)\big) - 2\int_0^h g\,ds$ ซึ่งจัดได้เป็นผลที่ต้องการ
2. ด้านซ้าย $\dfrac{h^4}{4} - \dfrac{h}{2}\cdot h^3 = -\dfrac{h^4}{4}$ ด้านขวา $-\dfrac{1}{2}\int_0^h 6s^2(h-s)\,ds = -3\Big(\dfrac{h^4}{3} - \dfrac{h^4}{4}\Big) = -\dfrac{h^4}{4}$ ทั้งสองข้างเท่ากันและเท่ากับ -0.002025 เมื่อ $h = 0.3$
3. ให้ $\varphi(s) = \mathbf{y}'(t_k + s)$ จะได้ $\int_0^h \varphi\,ds = \mathbf{y}(t_{k+1}) - \mathbf{y}(t_k)$ และ $\varphi(0), \varphi(h)$ คือ $\mathbf{y}'(t_k)$ และ $\mathbf{y}'(t_{k+1})$ ดังนั้นค่าคลาดเคลื่อนตัดปลายคือ $-\frac{1}{2}\int_0^h s(h-s)\mathbf{y}'''(t_k+s)\,ds$ และ $\|\cdot\| \le \frac{M_3}{2}\int_0^h s(h-s)\,ds = \frac{M_3}{2}\cdot\frac{h^3}{6} = \frac{M_3 h^3}{12}$

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_2_3_peano_identity_for_the_cubic` และ `test_trapezoid_local_error_obeys_the_bound_at_h_tenth`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 12.2.4.

1. เมื่อ $h = \frac{1}{64}$ ได้ $\dfrac{1+h/2}{1-h/2} = \dfrac{129}{127}$ และ $y_{64} = (129/127)^{64} \approx 2.718337$ ความคลาดเคลื่อนคือ $e - y_{64} \approx -5.5306 \times 10^{-5}$ (ขนาด 5.5306×10^{-5}) อัตราส่วนกับแถว $h = \frac{1}{32}$ คือ $0.00022126/0.000055306 \approx 4.0006$

2. ให้ $x = h/2$ จะได้ $\ln\dfrac{1+x}{1-x} = 2\big(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots\big)$ เมื่อแทน $x = h/2$ ได้ $h + \frac{h^3}{12} + O(h^5)$ ดังนั้น $\frac{1}{h}\ln\frac{1+h/2}{1-h/2} = 1 + \frac{h^2}{12} + O(h^4)$ และ $y_n = \exp\big(n \ln\frac{1+h/2}{1-h/2}\big)$ ที่ $n = 1/h$ จึงเท่ากับ $e\big(1 + \frac{h^2}{12} + O(h^4)\big)$ ความคลาดเคลื่อนมีขนาด $\frac{e}{12}h^2$ โดยวิธีให้ค่าสูงกว่าจริง ซึ่งอธิบายอัตราส่วนประมาณ 4 ในตาราง

ตรวจด้วยโค้ด: `test_ch12.py` (`test_trapezoid_error_matches_closed_form` และ `test_trapezoid_order_table_ratios_approach_four`)

เฉลย 12.2.5.

1. $\det = (1-\omega^2h^2)\cdot 1 - h\cdot(-\omega^2 h) = 1$ และ $\mathrm{tr} = (1-\omega^2h^2) + 1 = 2 - \omega^2h^2$

2. ค่าลักษณะเฉพาะเป็นรากของ $\lambda^2 - T\lambda + 1 = 0$ โดย $T = 2 - \omega^2h^2 \le 2$ เสมอ ถ้า $-2 < T < 2$ รากเป็นคู่สังยุคเชิงซ้อนที่ผลคูณ 1 จึงมีขนาด 1 ถ้า $T = -2$ (คือ $\omega h = 2$) รากซ้ำคือ -1 มีขนาด 1 แต่เมทริกซ์ไม่ทแยงมุม ดูข้อควรระวังในหัวข้อ 12.2 ถ้า $T < -2$ รากเป็นจำนวนจริงสองค่า r และ $1/r$ โดย $|r| \neq 1$ จึงมีรากที่ขนาดมากกว่า 1 และ $T \ge -2$ เทียบเท่า $\omega^2h^2 \le 4$ คือ $\omega h \le 2$

3. เมื่อ $\omega h < 2$ ได้ $2\cos\theta = T = 2 - \omega^2h^2$ จึง $1 - \cos\theta = \frac{1}{2}\omega^2h^2$ คือ $2\sin^2\frac{\theta}{2} = \frac{1}{2}\omega^2h^2$ ดังนั้น $\sin\frac{\theta}{2} = \frac{\omega h}{2}$ และ $\theta = 2\arcsin\frac{\omega h}{2}$ เพราะ $\arcsin y > y$ สำหรับ $0 < y < 1$ จึง $\theta > \omega h$

ตรวจด้วยโค้ด: `test_ch12.py` (`test_undamped_eigenvalues_on_unit_circle_below_two_over_omega` และ `test_undamped_phase_angles_match_the_eigenvalues`)

เฉลย 12.2.6.

1. จาก $kh^2 + 2bh - 4 = 0$ ได้ $h^* = \dfrac{-b + \sqrt{b^2 + 4k}}{k} = \dfrac{-0.1 + \sqrt{60.01}}{15} \approx 0.50977$

2. $h^* \cdot 150 \approx 76.47$ ดังนั้นวิธีซิมเพลกติกใช้ขั้นใหญ่กว่าวิธีชัดแจ้งประมาณ 76 เท่า

3. รัศมีสเปกตรัมคือ $\rho(0.01) = 0.999500$ (นิ่ง) $\rho(0.5) = 0.974679$ (นิ่ง) และ $\rho(0.52) = 1.457629$ (ไม่นิ่ง) ซึ่งสอดคล้องกับ $h^* \approx 0.50977$ เพราะ $0.5 < h^* < 0.52$

ตรวจด้วยโค้ด: `test_ch12.py` (`test_semi_implicit_threshold_is_the_root_of_kh2_plus_2bh_minus_4` และ `test_semi_implicit_spectral_radius_at_threshold_and_around`)

สมบัติ · Preprint · 2026-10-23

เฉลย 12.2.7.

1. ที่ $v = 12$ ได้ $v\Delta t = 0.2$ จึงสัดส่วนคือ $\min\{1, 0.1/0.2\} = 0.5$ สัดส่วนเหลือ 0.1 เมื่อ $v\Delta t = w/0.1 = 1$ เมตร คือ $v = 60$ m/s
2. ต้องให้ $v\Delta t \le w$ คือ $\Delta t \le w/v = 0.1/60 = 1/600$ วินาที
3. ช่วงที่จุดศูนย์กลางต้องอยู่คือ $w + 2r = 0.6$ เมตร ซึ่งยาวกว่าระยะต่อเฟรม 0.1 เมตร ดังนั้นจุดศูนย์กลางจะตกในช่วงนี้อย่างน้อยหนึ่งเฟรมเสมอ นักวิ่งจึงทะลุไม่ได้ที่ความเร็วนี้

ตรวจ ด้วย โค้ด: `test_ch12.py` (`test_tunnelling_exact_share_at_the_sample_speeds`, `test_exercise_12_2_7_frame_time_for_always_caught` และ `test_runner_cannot_tunnel_through_the_wall`)

12.3 ใช้ตำแหน่งก่อนหน้า

สถานการณ์: เชือกในเกมที่ไม่เก็บความเร็ว

ระบบอนุภาคทั่วไปมักเก็บสองอย่างให้ทุกอนุภาค คือตำแหน่งและความเร็ว แต่ Thomas Jakobsen ซึ่งพัฒนาฟิสิกส์ของเกม Hitman: Codename 47 เลือกเก็บเพียง *ตำแหน่งปัจจุบัน* กับ *ตำแหน่งก่อนหน้า* โดยไม่เก็บความเร็วเลย แล้วคำนวณตำแหน่งถัดไปจากสองค่านี้กับความเร่ง Jakobsen 2001, §2 วิธีนี้เรียกตามชื่อ Verlet ซึ่งใช้ในบทความปี 1967 เรื่องการจำลองของเหลวจากโมเลกุล Verlet 1967

ความเร็วจริง ๆ ไม่ได้หายไป มันซ่อนอยู่ในผลต่างของตำแหน่งสองค่า แต่คำถามคือ วิธีที่ใช้ค่าก่อนหน้าหลายค่าแบบนี้เป็นวิธีที่ดีหรือไม่ และต้องระวังอะไร

ในสถานการณ์สมมติของเรา เชือกยาว 10 จุดตรึงปลายหนึ่งไว้ที่ความสูง 1 เมตร แล้วปล่อยให้ตกลงมาจากแนวนอน ถ้าใช้วิธีออยเลอร์ชัดแจ้งที่ขั้น 0.01 วินาที พลังงานของเชือกจะโตเป็นประมาณ 1.4×10^{13} เท่าในสิบวินาที ส่วนวิธีเวอร์เลตพลังงานคลาดเคลื่อนไม่ถึงหนึ่งในพันส่วน (ตัวเลขจากโค้ดของเรา)

ภาพรวมก่อนอ่าน

ถ้าไม่เก็บความเร็ว ความเร็วยังอยู่ได้ในความต่างของตำแหน่งสองค่า เราจะเรียกวิธีแบบนี้ว่า*วิธีหลายขั้น* แล้วดูว่าอันดับของมันอ่านได้จากพหุนามสองตัว ต่อด้วยเงื่อนไขที่ทำให้วิธีลู่เข้าจริง ซึ่งมีมากกว่าแค่อันดับ

ตำแหน่งสองจุดก็พอ

ถ้าต้องการความเร่งจากตำแหน่งสามจุดที่ห่างกันเท่า ๆ กัน ใช้ผลต่างสองชั้นตรงกลาง $\dfrac{x(t+h) - 2x(t) + x(t-h)}{h^2}$ ซึ่งใกล้ $x''(t)$ โดยความคลาดเคลื่อน $\frac{h^2}{12}x^{(4)}(t)$ แทนความเร่ง x'' ด้วยความเร่งที่รู้ $a(x)$ แล้วแก้หาตำแหน่งถัดไป ก็จะได้วิธีเวอร์เลต

ลองเอง: ตกลงมาด้วยความเร่งคงที่

ให้ $a = -10, h = \frac{1}{2}, x_0 = 0, v_0 = 0$ ก่อนก้าวแรกต้องมีตำแหน่งก่อนหน้า $x_{-1} = 0 - 0 + \frac{1}{2}\cdot\frac{1}{4}\cdot(-10) = -\frac{5}{4}$ ซึ่งได้จากการถอยหนึ่งก้าวของเทย์เลอร์ (นิยามด้านล่าง) แล้วทุกก้าวใช้ตำแหน่งปัจจุบันและตำแหน่งก่อนหน้า

$$x_1 = 2\cdot 0 - \left(-\tfrac{5}{4}\right) + \tfrac{1}{4}(-10) = -\tfrac{5}{4}, \qquad x_2 = 2\left(-\tfrac{5}{4}\right) - 0 + \tfrac{1}{4}(-10) = -5,$$

ฉบับร่าง · Preprint · 2026-10-23

$$x_3 = 2(-5) - \left(-\tfrac{5}{4}\right) + \tfrac{1}{4}(-10) = -\tfrac{45}{4}.$$

ค่าเหล่านี้ตรงกับ $-5t^2$ ที่ $t = \frac{1}{2}, 1, \frac{3}{2}$ พอดี ไม่ใช่เรื่องบังเอิญ เหตุผลอยู่ในแบบฝึกหัด 12.3.2

นิยาม 12.3.1: วิธีเวอร์เลต

ให้ $x'' = a(x)$ โดยมีตำแหน่งเริ่ม x_0 และความเร็วเริ่ม v_0 ตำแหน่งก่อนหน้า x_{-1} เริ่มจากสูตรเทย์เลอร์ย้อนหลัง

$$x_{-1} = x_0 - h\,v_0 + \tfrac{1}{2}h^2\,a(x_0),$$

แล้วคำนวณสำหรับ $k = 0, 1, 2, \ldots$

$$x_{k+1} = 2x_k - x_{k-1} + h^2\,a(x_k).$$

ความเร็วไม่ถูกเก็บ เมื่อต้องการอ่านค่าใช้ $\frac{x_{k+1}-x_{k-1}}{2h}$ วิธีนี้คือ**วิธีเวอร์เลต** (Verlet integration) และเป็นวิธีชัดแจ้ง เพราะ x_{k+1} คำนวณได้ตรง ๆ จาก x_k และ x_{k-1}

อ่านแบบเข้าใจง่าย

สูตรนี้บอกว่า ตำแหน่งใหม่คือตำแหน่งปัจจุบัน บวกระยะที่เคลื่อนไปในก้าวที่แล้ว ซึ่งก็คือความเร็วโดยนัย $x_k - x_{k-1}$ บวกผลของความเร่งคูณ h^2 เพราะความเร็วซ่อนอยู่ในตำแหน่งสองค่าที่เก็บไว้แล้ว จึงไม่ต้องเก็บความเร็วแยก

ข้อควรระวัง: ต้องมีค่าก่อนหน้าสองค่าที่ถูกต้อง

ถ้าใส่ $x_{-1} = x_0$ แทนสูตรเทย์เลอร์ในตัวอย่างข้างบน จะได้ $x_1 = x_0 + h^2 a(x_0) = -\frac{5}{2}$ แทน $-\frac{5}{4}$ ความผิดพลาดนี้เกิดตั้งแต่ก้าวแรก และไม่หายไปตามเวลา ดูเหตุผลในหัวข้อ *Verlet บนสมการลำดับสอง* ด้านล่าง

ฉบับร่าง · Preprint · 2026-10-23

วิธีหลายขั้นและอันดับจากพหุนามสองตัว

วิธีออยเลอร์ใช้ค่าเดียวเพื่อหาค่าถัดไป แต่เรามักมีค่าเก่าหลายค่าอยู่แล้ว วิธีที่ใช้ค่าเก่าเหล่านี้เรียกว่าวิธีหลายขั้น

นิยาม 12.3.2: วิธีหลายขั้นเชิงเส้น

ให้ $s \geq 1$ และค่าคงที่ $\alpha_0, \ldots, \alpha_s$ และ $\beta_0, \ldots, \beta_s$ โดย $\alpha_s = 1$ วิธี**วิธีหลายขั้นเชิงเส้น** (linear multistep method) คือการคำนวณ

$$\sum_{j=0}^{s} \alpha_j\, y_{k+j} = h \sum_{j=0}^{s} \beta_j\, f(t_{k+j}, y_{k+j}).$$

ถ้า $\beta_s = 0$ วิธีนี้เป็นแบบชัดแจ้ง ถ้า $\beta_s \neq 0$ เป็นแบบโดยปริยาย กำหนดพหุนาม $\rho(w) = \sum_j \alpha_j w^j$ และ $\sigma(w) = \sum_j \beta_j w^j$

อ่านแบบเข้าใจง่าย

วิธีออยเลอร์ในหัวข้อ 12.1 คือกรณี $s = 1$ ที่ $\alpha = (-1, 1)$ และ $\beta = (1, 0)$ ตัวอย่างที่สำคัญคือ**วิธีอดัมส์-บาชฟอร์ท** (Adams--Bashforth method) สองขั้น $y_{k+2} = y_{k+1} + h\left(\frac{3}{2}f_{k+1} - \frac{1}{2}f_k\right)$ โดย $f_k = f(t_k, y_k)$ ซึ่งมี

$\alpha = (0, -1, 1)$ และ $\beta = (-\frac{1}{2}, \frac{3}{2}, 0)$ วิธีนี้ใช้ความชันที่คำนวณไว้แล้วซ้ำ จึงไม่ต้องคำนวณ f ใหม่ทุกก้าว

เราจะนิยาม*อันดับ* ของวิธีหลายขั้นให้เหมือนหัวข้อ 12.1 คือ ดูความคลาดเคลื่อนตัดปลายเฉพาะที่

$$\psi(t, y) = \sum_{j} \alpha_j \, y(t + jh) - h \sum_{j} \beta_j \, y'(t + jh)$$

ของคำตอบจริง y ที่เรียบพอ วิธีมีอันดับ p ถ้า $\psi = O(h^{p+1})$ ทฤษฎีบทต่อไปบอกว่าเราอ่านอันดับได้จากค่าคงที่สองชุดที่คำนวณได้ตรง ๆ

ทฤษฎีบท 12.3.1: อันดับจากค่าคงที่ของเทย์เลอร์

ให้ $C_0 = \sum_j \alpha_j$ และสำหรับ $q \geq 1$

$$C_q = \sum_{j} \alpha_j \frac{j^q}{q!} - \sum_{j} \beta_j \frac{j^{q-1}}{(q-1)!}.$$

สำหรับจำนวนเต็ม $p \geq 0$ ข้อความต่อไปนี้เทียบเท่ากัน

1. วิธีมีอันดับอย่างน้อย p คือ $\psi = O(h^{p+1})$ สำหรับทุก y ที่เรียบพอ;
2. $C_0 = C_1 = \cdots = C_p = 0$;
3. $\rho(e^z) - z\,\sigma(e^z) = O(z^{p+1})$ เมื่อ $z \to 0$.

อันดับเป็น p พอดีเมื่อ $C_{p+1} \neq 0$

พิสูจน์. **แนวคิด.** กระจาย $y(t + jh)$ และ $y'(t + jh)$ เป็นอนุกรมเทย์เลอร์ จัดพจน์ตามกำลังของ h แล้วพจน์ที่ h^q จะมีสัมประสิทธิ์เป็น C_q ส่วนสมมูลกับ (3) เป็นการเห็นอนุกรมเดียวกันในตัวแปร $z = h$ (บทพิสูจน์นี้เราเขียนเองตามแนวคิดของ Iserles 2009, §2.2, Thm 2.1, pp. 22--23)

1. **การกระจาย.** สำหรับ y ที่วิเคราะห์ได้ $y(t+jh) = \sum_q \frac{(jh)^q}{q!} y^{(q)}(t)$ และ $h\,y'(t+jh) = \sum_{q \geq 1} \frac{j^{q-1} h^q}{(q-1)!} y^{(q)}(t)$ แทนลงในนิยาม จะได้ $\psi(t, y) = \sum_{q \geq 0} C_q \, h^q \, y^{(q)}(t)$
2. **(1) ให้ (2).** เลือก $y(t) = t^q$ ที่ $t = 0$ จากข้อ 1 จะได้ $\psi = C_q \, q! \, h^q$ เมื่อ $\psi = O(h^{p+1})$ และ $q \leq p$ ต้องได้ $C_q = 0$
3. **(2) ให้ (3) และกลับกัน.** จากข้อ 1 และการแทน $z = h$ ได้ $\rho(e^z) - z\sigma(e^z) = \sum_q C_q z^q$ ฟังก์ชันนี้เป็น $O(z^{p+1})$ ก็ต่อเมื่อสัมประสิทธิ์ของ $z^0, \ldots, z^p$ เป็นศูนย์
4. **(2) ให้ (1).** ถ้า $C_0 = \cdots = C_p = 0$ จากข้อ 1 จะได้ $\psi = \sum_{q > p} C_q h^q y^{(q)}(t) = O(h^{p+1})$ เมื่ออนุพันธ์ของ y ถูกกั้นขอบ

สรุป: อันดับอ่านได้จากพหุนาม $\rho(e^z) - z\sigma(e^z)$ โดยตรง ไม่ต้องพึ่งการแก้สมการใด ๆ □

ตัวอย่าง: สำหรับอดัมส์-บาชฟอร์ทสองขั้น $C_0 = C_1 = C_2 = 0$ และ $C_3 = \frac{5}{12}$ จึงมีอันดับ 2 ตรงกับการขยาย $\rho(w) - \sigma(w) \ln w = \frac{5}{12}(w-1)^3 + O((w-1)^4)$ ที่แสดงไว้ใน Iserles 2009, §2.2, p. 23 แบบฝึกหัด 12.3.3 ให้ตรวจเอง

ร่างสมบูรณ์ · Preprint · 2026-10-23

ตารางที่ 12.4: ความคลาดเคลื่อนของอดัมส์-บาชฟอร์ทสองขั้นสำหรับ $y' = y$, $y(0) = 1$ ที่ $t = 1$ ใช้ค่าเริ่มต้น $y_1 = e^h$ (ค่าแม่นตรง) เพื่อให้เห็นผลของวิธีเอง คอลัมน์ขวาคืออัตราส่วนของแถวก่อนหารด้วยแถวนี้ (สร้างจาก `numanal/ch12.py` และตรวจด้วย `test_ch12.py`)

h	ความคลาดเคลื่อน	อัตราส่วน
$1/2$	0.083020	—
$1/4$	0.042404	1.9578
$1/8$	0.014066	3.0147
$1/16$	0.003973	3.5406
$1/32$	0.001050	3.7833

ตารางนี้ที่ $h = \frac{1}{32}$ ให้ความคลาดเคลื่อน 0.00105 ซึ่งเล็กกว่าของวิธีออยเลอร์ในตาราง 12.1 (0.041292) เกือบ 40 เท่า อัตราส่วนยังห่างจาก 4 เล็กน้อยที่ h เท่านี้ เพราะพจน์ถัดไปในอนุกรมของความคลาดเคลื่อนยังมีผล

เงื่อนไขราก: อันดับอย่างเดียวไม่พอ

เราต้องการให้วิธีลู่เข้า นั่นคือ คำตอบของวิธีเข้าใกล้คำตอบจริงเมื่อ $h \to 0$ ทฤษฎีบทต่อไปบอกว่าอันดับเพียงอย่างเดียวยังไม่พอ

นิยาม 12.3.3: เงื่อนไขราก

พหุนาม ρ เป็นไปตาม**เงื่อนไขราก** (root condition) ถ้าทุกรากของ ρ มีขนาดไม่เกิน 1 และรากที่มีขนาดเท่ากับ 1 เป็นรากเดี่ยว

ทฤษฎีบท 12.3.2: ทฤษฎีบทสมมูลของดาห์ลควิสต์

ให้ f เป็นไปตามเงื่อนไขลิพชิทซ์ และให้ความคลาดเคลื่อนของค่าเริ่มต้น $y_1, \dots, y_{s-1}$ มีค่าเป็นศูนย์เมื่อ $h \to 0^+$ วิธีหลายขั้นเชิงเส้นลู่เข้า ก็ต่อเมื่อมีอันดับอย่างน้อย 1 และ ρ เป็นไปตามเงื่อนไขราก

ทฤษฎีบทนี้ยกมาโดยไม่พิสูจน์ ดู Iserles 2009, §2.2, Thm 2.2, p. 25 ดูเพิ่มบทที่ 2 ทฤษฎีบท 2.2 ของ Iserles และ Atkinson 1989, §6.8, pp. 394--410

อ่านแบบเข้าใจง่าย

ทั้งสองส่วนต้องมีครบ ส่วนอันดับบอกว่าแต่ละก้าวผิดน้อยแค่ไหน ส่วนเงื่อนไขรากบอกว่าความผิดพลาดเหล่านั้นจะไม่โตขึ้นเมื่อสะสมไปหลายก้าว วิธีอดัมส์ทุกขั้นมี $\rho(w) = w^{s-1}(w - 1)$ ซึ่งมีรากที่ 0 และ 1 เท่านั้น จึงผ่านเงื่อนไขนี้

ข้อควรระวัง: อันดับสูงไม่ได้แปลว่าลู่เข้า

ลองดูวิธีของ Iserles (2.12): $y_{k+2} - 3y_{k+1} + 2y_k = h\left(\frac{13}{12} f_{k+2} - \frac{5}{3} f_{k+1} - \frac{5}{12} f_k\right)$ วิธีนี้มีอันดับ 2 (แบบฝึกหัด 12.3.5) แต่ $\rho(w) = (w - 1)(w - 2)$ มีรากที่ 2 ซึ่งผิดเงื่อนไขราก ลองดูกรณี $f = 0$ ผลเฉลยจริงคือ 0 ให้ค่าเริ่มต้น $y_0 = 0$ และ $y_1 = \varepsilon$ เมื่อ $\varepsilon = h^2$ ซึ่งเข้าใกล้ 0 ค่าที่ $t = 1$ คือ 10.23 เมื่อ $h = 0.1$ และ 2621.44 เมื่อ $h = 0.05$ ค่ายิ่งใหญ่ขึ้นเมื่อ h เล็กลง ตามที่ทฤษฎีบทบอก Iserles 2009, pp. 23, 25

ร่างต้นฉบับ · Preprint · 2026-10-23

Verlet บนสมการลำดับสอง

เขียนวิธีเวอร์เลตให้เห็นเป็นวิธีหลายขั้นได้ดังนี้ สำหรับ $x'' = F(x)$ ได้ $x_{k+1} - 2x_k + x_{k-1} = h^2 F(x_k)$ ซึ่งมี $\alpha = (1, -2, 1)$ และ $\beta = (0, 1, 0)$ เมื่อเขียนแบบลำดับสองนี้ $\rho(w) = (w-1)^2$ ซึ่งมีรากซ้ำที่ 1 และไม่ผ่านเงื่อนไขราก ทฤษฎีบท 12.3.2 เขียนไว้สำหรับสมการลำดับหนึ่ง $y' = f$ เราจึงไม่สรุปการลู่เข้าของเวอร์เลตจากทฤษฎีบทนั้นโดยตรง

ผลที่พิสูจน์ได้ในที่นี้คือขนาดของความคลาดเคลื่อนในหนึ่งก้าว

ทฤษฎีบท 12.3.3: ความคลาดเคลื่อนตัดปลายเฉพาะที่ของวิธีเวอร์เลต

ให้ x หาอนุพันธ์ได้ต่อเนื่องถึงอันดับหก แล้ว

$$x(t+h) - 2x(t) + x(t-h) - h^2 x''(t) = \frac{h^4}{12} x^{(4)}(t) + O(h^6).$$

ดังนั้นความคลาดเคลื่อนตัดปลายเฉพาะที่ของหนึ่งก้าวของวิธีเวอร์เลตเป็น $O(h^4)$

พิสูจน์. **แนวคิด.** เทย์เลอร์สองด้านของ $x(t \pm h)$ มีพจน์คี่หักล้างกัน เหลือพจน์คู่ซึ่งเท่ากับ $h^2 x''$ บวก h^4 คูณอนุพันธ์อันดับสี่

1. **กระจาย.** $x(t \pm h) = x \pm hx' + \frac{1}{2}h^2 x'' \pm \frac{1}{6}h^3 x''' + \frac{1}{24}h^4 x^{(4)} \pm \frac{1}{120}h^5 x^{(5)} + O(h^6)$
2. **รวม.** บวกสองด้านแล้วลบ $2x(t)$ พจน์คี่หักล้างกัน จะได้ $x(t+h) - 2x(t) + x(t-h) = h^2 x''(t) + \frac{h^4}{12} x^{(4)}(t) + O(h^6)$

สรุป: ความคลาดเคลื่อนตัดปลายเฉพาะที่เป็น h^4 สำหรับสมการลำดับสอง ซึ่งตรงกับอันดับ 2 ตามแบบแผนที่ใช้ได้กับวิธีของเรา □

แบบแผนนี้หมายความว่า วิธีเวอร์เลตมีอันดับ 2 ในความหมายของความคลาดเคลื่อนตัดปลายเฉพาะที่ การพิสูจน์ความคลาดเคลื่อนวงกว้างต้องใช้เสถียรภาพอีกชั้นซึ่งเราไม่ทำในบทนี้ สิ่งที่เรามีต่อไปคือตัวเลขจากเชือก

เชือกที่ตกลงมา: พลังงานที่ไม่โต

ลองดูเชือกในสถานการณ์สมมติ มี 10 จุด จุดแรกตรึงที่ความสูง 1 เมตร สปริงระหว่างจุดข้างเคียงมีค่าคงที่ $k = 100$ นิวตันต่อเมตร ความยาวพัก 0.1 เมตร มวลจุดละ 1 กิโลกรัม $g = 9.8$ เชือกเริ่มนอนนิ่งในแนวนอน แล้วปล่อยให้ตก พลังงานรวมคือพลังงานจลน์บวกพลังงานโน้มถ่วงบวกพลังงานสปริง ซึ่งที่เวลาเริ่มต้นเท่ากับ 98 จูล

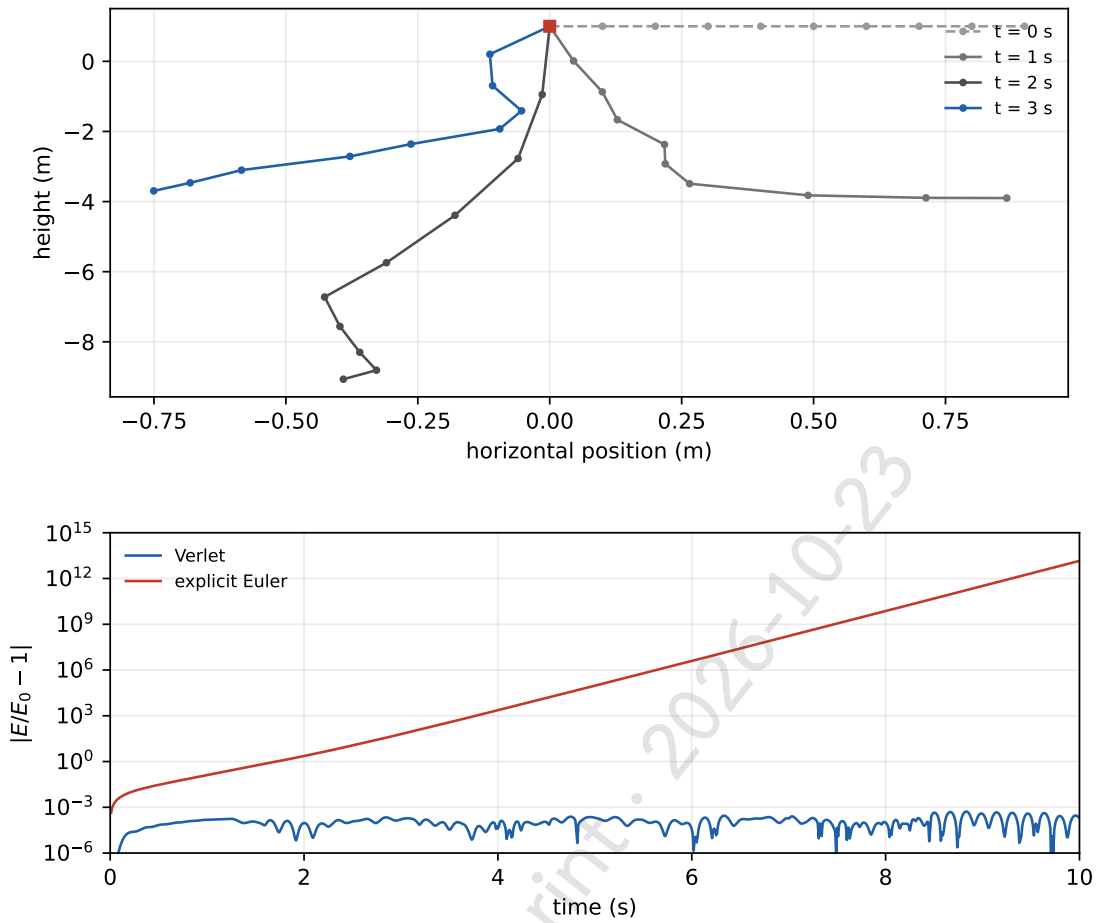
ที่ขั้น $h = 0.01$ วินาทีตลอด 10 วินาที วิธีเวอร์เลตให้ $\max |E/E_0 - 1| \approx 5.3 \times 10^{-4}$ และที่ $t = 10$ วินาทีพลังงานเหลือ 0.99978 เท่าของค่าเริ่มต้น ส่วนวิธีออยเลอร์ชัดแจ้งที่ขั้นเดียวกันให้ $E/E_0 \approx 1.13$ ที่ $t = 1$, $\approx 9.4 \times 10^4$ ที่ $t = 5$ และ $\approx 1.4 \times 10^{13}$ ที่ $t = 10$ ความถี่เชิงมุมสูงสุดของเชือกประมาณ 19.7 เรเดียนต่อวินาที ดังนั้น $\omega h \approx 0.20$ ซึ่งอยู่ในช่วงที่วิธีเวอร์เลตนิ่ง

ผลนี้เป็นตัวเลขจากเชือกนี้ ไม่ใช่การพิสูจน์ว่าวิธีเวอร์เลตไม่มีวันเกิดการเพี้ยน บทความนั้นเตือนว่า วิธีนี้อาจทำให้พลังงานหายไปบ้าง (dissipate) และไม่แม่นเสมอไป Jakobsen 2001, §2 ในการทดสอบนี้พลังงานลดลงเพียงประมาณ 0.02% ที่ $t = 10$ วินาที ซึ่งเข้ากับคำเตือนนั้น

สรุปใจความ

- วิธีเวอร์เลตไม่เก็บความเร็ว: $x_{k+1} = 2x_k - x_{k-1} + h^2 a(x_k)$ ความเร็วอ่านได้จาก $\frac{x_{k+1} - x_{k-1}}{2h}$ และต้องมีค่าเริ่มต้นสองค่าที่ถูกต้อง

ร่างแรก · Preprint · 2026-10-23

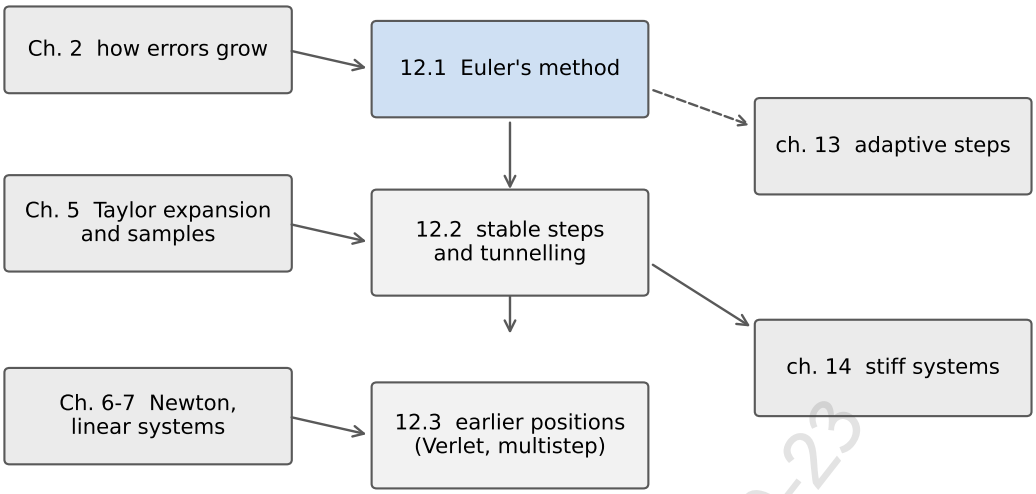


รูปที่ 12.5: บนคือเชือกตั้งแต่เริ่ม (เส้นประ) และที่ $1, 2, 3$ วินาที คำนวณด้วยวิธีเวอร์เลตที่ขั้น 0.01 วินาที ล่างคือ $|E/E_0 - 1|$ ของวิธีเวอร์เลต (สีน้ำเงิน) และวิธีออยเลอร์ชัดแจ้ง (สีแดง) เทียบกับเวลา เส้นสองเส้นอยู่บนแกนลอการิทึม (สร้างจาก `numanal/ch12.py`)

- อันดับของวิธีหลายขั้นอ่านจาก C_q หรือจาก $\rho(e^z) - z\sigma(e^z)$ อดัมส์-บาชฟอร์ทสองขั้นมีอันดับ 2 โดย $C_3 = \frac{5}{12}$
- การลู่เข้า = อันดับอย่างน้อย 1 และเงื่อนไขราก (ทฤษฎีบทสมมูลของดาห์ลควิสต์) วิธี (2.12) มีอันดับ 2 แต่ไม่ลู่เข้า
- เวอร์เลตมีอันดับ 2 ในความหมายของความคลาดเคลื่อนตัดปลายเฉพาะที่ $O(h^4)$ และรากซ้ำทำให้ทฤษฎีบทดาห์ลควิสต์ใช้ตรง ๆ ไม่ได้
- บนเชือกที่ตกลงมา เวอร์เลตคงพลังงานไว้ในช่วงประมาณ 5×10^{-4} ส่วนวิธีออยเลอร์ชัดแจ้งโตขึ้นเป็นประมาณ 10^{13} เท่า

อ่านต่อ. รายละเอียดของวิธีหลายขั้นอยู่ในบทที่ 2 ของ Iserles 2009, §2.1--2.3, pp. 19--32 และ Atkinson 1989, §6.3, p. 357, §6.7, pp. 381--393, §6.8, pp. 394--410 บทความต้นฉบับเรื่องการจำลองโมเลกุลดู Verlet 1967 และบทความเรื่องฟิสิกส์ตัวละครดู Jakobsen 2001

ฉบับร่าง · preprint · 2026-10-23



รูปที่ 12.6: แผนผังบทที่ 12 ทั้งบท หัวข้อ 12.1 ถึง 12.3 และสิ่งที่บทก่อนให้ วิธีออยเลอร์เป็นฐานของหัวข้อ 12.2 และ 12.3 หัวข้อ 12.2 ให้ขั้นโดยนัยที่บทที่ 14 ต้องใช้ ส่วนบทที่ 13 ขยายความคิดของหัวข้อ 12.1 (สร้างจาก `numanal/ch12.py`)

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 12.3.1 (⋆, ทำด้วยมือ) ให้ $x'' = -10$ โดย $x_0 = 0$, $v_0 = 0$ และ $h = \frac{1}{5}$ ใช้วิธีเวอร์เลตหา x_{-1}, x_1, x_2 และ x_3 แล้วเทียบแต่ละค่ากับคำตอบจริง $x(t) = -5t^2$ ที่ $t = \frac{1}{5}, \frac{2}{5}, \frac{3}{5}$

แบบฝึกหัด 12.3.2 (⋆⋆, พิสูจน์) ให้ $x(t) = \alpha + \beta t + \gamma t^2$ โดย α, β, γ เป็นค่าคงที่ และ $x'' = 2\gamma$ คงที่ ให้ $v_0 = x'(0) = \beta$

1. แสดงว่า $x(t+h) - 2x(t) + x(t-h) = h^2 x''(t)$ สำหรับทุก t และ h
2. แสดงว่าค่าเริ่มต้นของวิธีเวอร์เลตให้ $x_{-1} = x(-h)$ และ $x_0 = x(0)$
3. ใช้การเหนี่ยวนำแสดงว่า $x_k = x(kh)$ สำหรับทุก $k \geq 0$

แบบฝึกหัด 12.3.3 (⋆⋆, คำนวณ) ให้วิธีอดัมส์-บาชฟอร์ทสองขั้น $\alpha = (0, -1, 1)$ และ $\beta = (-\frac{1}{2}, \frac{3}{2}, 0)$

1. คำนวณ C_0, C_1, C_2 และ C_3 จากนิยามในทฤษฎีบท 12.3.1 แล้วสรุปอันดับ
2. คำนวณ C_3 แล้วแสดงว่าค่าคลาดเคลื่อนตัดปลายเฉพาะที่เป็น $\frac{5}{12}h^3 y'''(t) + O(h^4)$
3. สำหรับวิธีออยเลอร์ ($\alpha = (-1, 1)$, $\beta = (1, 0)$) คำนวณ C_2 และสรุปอันดับ

แบบฝึกหัด 12.3.4 (⋆⋆⋆, พิสูจน์) ให้ $\rho(w) = \sum_j \alpha_j w^j$ และ $\sigma(w) = \sum_j \beta_j w^j$

1. แสดงว่า $\rho(e^z) = \sum_{q \geq 0} z^q \sum_j \alpha_j \dfrac{j^q}{q!}$ และ $z\,\sigma(e^z) = \sum_{q \geq 1} z^q \sum_j \beta_j \dfrac{j^{q-1}}{(q-1)!}$
2. ใช้ข้อ 1 แสดงว่า $\rho(e^z) - z\,\sigma(e^z) = \sum_{q \geq 0} C_q z^q$ โดย C_q เป็นค่าคงที่ในทฤษฎีบท 12.3.1
3. ตรวจด้วยวิธีออยเลอร์ว่า $\rho(e^z) - z\,\sigma(e^z) = \frac{1}{2}z^2 + \frac{1}{6}z^3 + \cdots$ เมื่อ z เล็ก

แบบฝึกหัด 12.3.5 (⋆ ⋆ ⋆, คำนวณและพิสูจน์) ให้วิธีของ Iserles (2.12) คือ $y_{k+2} - 3y_{k+1} + 2y_k = h\big(\frac{13}{12}f_{k+2} - \frac{5}{3}f_{k+1} - \frac{5}{12}f_k\big)$

1. แสดงว่า $C_0 = C_1 = C_2 = 0$ และ $C_3 = -\frac{1}{2}$ ดังนั้นอันดับของวิธีคือ 2
2. แสดงว่า $\rho(w) = (w-1)(w-2)$ และหารากของ ρ ว่ามีรากที่มีขนาดมากกว่า 1 หรือไม่
3. ให้ $f = 0$, $y_0 = 0$ และ $y_1 = \varepsilon$ ใช้การเหนี่ยวนำแสดงว่า $y_k = \varepsilon(2^k - 1)$
4. ที่ $t = 1$ ให้ $\varepsilon = h^2$ และใช้ $n = 1/h$ ก้าว คำนวณ y_n เมื่อ $h = \frac{1}{10}$ และ $h = \frac{1}{20}$ แล้วอธิบายว่าทำไมค่าไม่เข้าใกล้ 0

แบบฝึกหัด 12.3.6 (⋆⋆, คำนวณ) ให้สปริง $x'' = -\omega^2 x$ โดย $\omega^2 = 15$ และใช้วิธีเวอร์เลต

1. แสดงว่าพหุนามลักษณะเฉพาะของการวนซ้ำคือ $\zeta^2 - (2 - \omega^2 h^2)\zeta + 1 = 0$
2. แสดงว่าถ้า $\omega h \le 2$ ทั้งสองรากมีขนาด 1 และถ้า $\omega h > 2$ มีรากหนึ่งที่มีขนาดมากกว่า 1
3. หา $h^* = 2/\omega$ แล้วตรวจว่า $h = 0.5$ นิ่ง แต่ $h = 0.52$ ไม่นิ่ง โดยดูรัศมีสเปกตรัม

แบบฝึกหัด 12.3.7 (⋆ ⋆ ⋆, โครงงานเล็ก) ใช้เชือกของหัวข้อ 12.3 (10 จุด ปลายหนึ่งตรึงที่ความสูง 1 เมตร สปริง $k = 100$ ความยาวพัก 0.1 เมตร มวลจุดละ 1 กิโลกรัม $g = 9.8$) รันถึง $t = 10$ วินาทีด้วยขั้น $h = 0.01$ และ $h = 0.002$

1. หา $|E/E_0 - 1|$ ที่ $t = 10$ ของวิธีเวอร์เลตทั้งสองขั้น และเทียบว่าค่าที่ขั้นเล็กกว่าน้อยกว่ากี่เท่า
2. วิธีออยเลอร์ชัดแจ้งที่ $h = 0.002$ ได้ E/E_0 ที่ $t = 10$ เท่าใด และยังโตอยู่หรือไม่
3. ตัวคูณของสปริงเดี่ยว $x'' = -15x$ ของวิธีออยเลอร์ชัดแจ้งคือ $|\lambda|^2 = 1 + 15h^2$ จงอธิบายว่าทำไมขั้นที่เล็กลงไม่ทำให้วิธีนี้นิ่งแน่นอน

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 12.3.1. เมื่อ $h = \frac{1}{5}$ และ $a = -10$ ค่าเริ่มต้น $x_{-1} = 0 - 0 + \frac{1}{2}\cdot\frac{1}{25}\cdot(-10) = -\frac{1}{5}$ จากนั้น

$$x_1 = 2\cdot 0 - \big(-\tfrac{1}{5}\big) + \tfrac{1}{25}(-10) = -\tfrac{1}{5}, \qquad x_2 = 2\big(-\tfrac{1}{5}\big) - 0 + \tfrac{1}{25}(-10) = -\tfrac{4}{5},$$

$$x_3 = 2\big(-\tfrac{4}{5}\big) - \big(-\tfrac{1}{5}\big) + \tfrac{1}{25}(-10) = -\tfrac{9}{5}.$$

คำตอบจริง $-5t^2$ ที่ $t = \frac{1}{5}, \frac{2}{5}, \frac{3}{5}$ คือ $-\frac{1}{5}, -\frac{4}{5}, -\frac{9}{5}$ ตรงกันทุกค่า ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_3_1_verlet_hand_values_are_exact_for_constant_acceleration`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 12.3.2.

1. ค่าของ $x(t\pm h) = \alpha+\beta(t\pm h)+\gamma(t\pm h)^2$ เมื่อรวมกันจะได้ $x(t+h)+x(t-h) = 2\alpha+2\beta t+2\gamma t^2+2\gamma h^2$ และ $2x(t) = 2\alpha + 2\beta t + 2\gamma t^2$ ดังนั้น $x(t+h) - 2x(t) + x(t-h) = 2\gamma h^2 = h^2 x''(t)$

2. ค่าเริ่มต้นคือ $x_{-1} = x_0 - hv_0 + \frac{1}{2}h^2 a(x_0) = \alpha - h\beta + \frac{1}{2}h^2 \cdot 2\gamma = \alpha - h\beta + \gamma h^2 = x(-h)$ และ $x_0 = \alpha = x(0)$

3. สมมติ $x_{k-1} = x((k-1)h)$ และ $x_k = x(kh)$ แล้ว $x_{k+1} = 2x_k - x_{k-1} + h^2 a(x_k) = x((k+1)h)$ โดยข้อ 1 ซึ่งทำให้ได้ผลที่ต้องการ

ดังนั้นวิธีเวอร์เลตแม่นตรงสำหรับพหุนามกำลังสองทุกตัว และความคลาดเคลื่อนตัดปลายเฉพาะที่เป็นศูนย์ ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_3_2_verlet_is_exact_for_every_quadratic`)

เฉลย 12.3.3.

1. จาก $\alpha = (0,-1,1)$ และ $\beta = (-\frac{1}{2},\frac{3}{2},0)$ ได้ $C_0 = 0-1+1 = 0$ $C_1 = \sum_j \alpha_j j - \sum_j \beta_j = (-1+2) - (-\frac{1}{2}+\frac{3}{2}) = 1-1 = 0$ $C_2 = \frac{1}{2}\sum_j \alpha_j j^2 - \sum_j \beta_j j = \frac{1}{2}(-1+4) - \frac{3}{2} = 0$ จึงมีอันดับอย่างน้อย 2 และ $C_3 = \frac{1}{6}\sum_j \alpha_j j^3 - \frac{1}{2}\sum_j \beta_j j^2 = \frac{1}{6}(-1+8) - \frac{1}{2}\cdot\frac{3}{2} = \frac{7}{6} - \frac{3}{4} = \frac{5}{12} \neq 0$ ดังนั้นอันดับคือ 2 พอดี

2. จากการกระจายในพิสูจน์ของทฤษฎีบท 12.3.1 ได้ $\psi = \sum_q C_q h^q y^{(q)}(t)$ และเมื่อ $C_0 = C_1 = C_2 = 0$ เหลือ $\psi = \frac{5}{12}h^3 y'''(t) + O(h^4)$

3. วิธีออยเลอร์มี $C_0 = -1+1 = 0$ $C_1 = \sum_j \alpha_j j - \sum_j \beta_j = 1-1 = 0$ และ $C_2 = \frac{1}{2}\sum_j \alpha_j j^2 - \sum_j \beta_j j = \frac{1}{2}\cdot 1 - 0 = \frac{1}{2} \neq 0$ จึงมีอันดับ 1

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_3_3_adams_bashforth2_constants`)

เฉลย 12.3.4.

1. ใช้อนุกรม $e^{jz} = \sum_q \dfrac{(jz)^q}{q!}$ แล้วสลับลำดับการรวม จะได้ $\rho(e^z) = \sum_j \alpha_j \sum_q \dfrac{j^q z^q}{q!} = \sum_q z^q \sum_j \alpha_j \dfrac{j^q}{q!}$ และ $z\sigma(e^z) = \sum_j \beta_j\, z \sum_r \dfrac{j^r z^r}{r!} = \sum_{q\geq 1} z^q \sum_j \beta_j \dfrac{j^{q-1}}{(q-1)!}$ โดยแทน $q = r+1$

2. ลบสองสมการในข้อ 1 ได้พจน์ z^0 เท่ากับ $\sum_j \alpha_j = C_0$ และพจน์ z^q เท่ากับ C_q สำหรับ $q \geq 1$ ตามนิยาม

3. สำหรับออยเลอร์ $\rho(w) = w - 1$ และ $\sigma(w) = 1$ จึง $\rho(e^z) - z\sigma(e^z) = e^z - 1 - z = \frac{1}{2}z^2 + \frac{1}{6}z^3 + \cdots$ ซึ่งตรงกับ $C_2 = \frac{1}{2}$ และ $C_3 = \frac{1}{6}$ ที่ได้ในเฉลย 12.3.3 ข้อ 3

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_3_4_order_identity_rho_sigma`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 12.3.5.

1. ใช้ $\alpha = (2, -3, 1)$ และ $\beta = (-\frac{5}{12}, -\frac{5}{3}, \frac{13}{12})$ จะได้ $\sum_j \alpha_j = 0$ $\sum_j \alpha_j j = -3 + 2 = -1$ และ $\sum_j \beta_j = \frac{-5-20+13}{12} = -1$ ดังนั้น $C_1 = -1 - (-1) = 0$ ส่วน $\sum_j \alpha_j j^2 = -3 + 4 = 1$ และ $\sum_j \beta_j j = -\frac{5}{3} + \frac{13}{6} = \frac{1}{2}$ จึง $C_2 = \frac{1}{2} \cdot 1 - \frac{1}{2} = 0$ สุดท้าย $\sum_j \alpha_j j^3 = -3 + 8 = 5$ และ $\sum_j \beta_j j^2 = -\frac{5}{3} + \frac{13}{3} = \frac{8}{3}$ จึง $C_3 = \frac{5}{6} - \frac{1}{2} \cdot \frac{8}{3} = \frac{5}{6} - \frac{4}{3} = -\frac{1}{2}$ อันดับคือ 2

2. $\rho(w) = 2 - 3w + w^2$ เมื่อจัดเป็น $w^2 - 3w + 2 = (w-1)(w-2)$ รากคือ 1 และ 2 รากที่ 2 มีขนาดมากกว่า 1 จึงไม่ผ่านเงื่อนไขราก

3. เมื่อ $n = 0$ และ $n = 1$ เป็นจริงโดยนิยาม สมมติว่า $y_k = \varepsilon(2^k - 1)$ และ $y_{k+1} = \varepsilon(2^{k+1} - 1)$ แล้ว $y_{k+2} = 3y_{k+1} - 2y_k = \varepsilon\big(3 \cdot 2^{k+1} - 3 - 2^{k+1} + 2\big) = \varepsilon(2^{k+2} - 1)$

4. เมื่อ $h = \frac{1}{10}$ มี $n = 10$ และ $\varepsilon = 0.01$ ได้ $y_{10} = 0.01 \cdot 1023 = 10.23$ เมื่อ $h = \frac{1}{20}$ มี $n = 20$ และ $\varepsilon = 0.0025$ ได้ $y_{20} = 0.0025 \cdot 1048575 \approx 2621.44$ ค่าโตเร็วกว่าการลดของ h เพราะ 2^n เพิ่มเมื่อ $n = 1/h$ เพิ่ม ขณะที่ความคลาดเคลื่อนเริ่มต้นเล็กลงเพียงเป็น h^2 ทฤษฎีบท 12.3.2 บอกว่าวิธีนี้ไม่ลู่เข้า

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_3_5_iserles_212_order_two_root_two`, `test_exercise_12_3_5_iserles_212_closed_form_on_zero` และ `test_exercise_12_3_5_iserles_212_grows_with_starting_error_h_squared`)

เฉลย 12.3.6.

1. ถ้า $x_k = \zeta^k$ แทนลงในสมการ $x_{k+1} = (2 - \omega^2 h^2) x_k - x_{k-1}$ จะได้ $\zeta^2 = (2 - \omega^2 h^2)\zeta - 1$ คือ $\zeta^2 - (2 - \omega^2 h^2)\zeta + 1 = 0$

2. ให้ $T = 2 - \omega^2 h^2$ ผลคูณของรากเป็น 1 เสมอ และ $T \le 2$ เสมอ ถ้า $-2 < T < 2$ ดิสคริมิแนนต์ $T^2 - 4 < 0$ รากเป็นคู่เชิงซ้อนสังยุคที่มีขนาด $\sqrt{1} = 1$ ถ้า $T = -2$ รากซ้ำคือ -1 มีขนาด 1 ถ้า $T < -2$ รากเป็นจำนวนจริงที่ผลคูณเป็น 1 จึงมีรากหนึ่งที่มีขนาดมากกว่า 1 และ $T \ge -2$ เทียบเท่ากับ $\omega h \le 2$

3. $h^* = 2/\sqrt{15} \approx 0.51640$ เมื่อ $h = 0.5$ ได้ $\omega h \approx 1.9365 \le 2$ รัศมีสเปกตรัมคือ 1 จึงนิ่ง เมื่อ $h = 0.52$ ได้ $\omega h \approx 2.0141$ และ $T \approx -2.0560$ รากคือ $\frac{1}{2}(T \pm \sqrt{T^2 - 4}) \approx -1.2663$ และ -0.7898 รัศมีสเปกตรัมคือ 1.2663 จึงไม่นิ่ง

ตรวจ ด้วย โค้ด: `test_ch12.py` (`test_exercise_12_3_6_threshold_for_the_spring_is_two_over_omega`)

เฉลย 12.3.7.

1. ที่ $h = 0.01$ ความคลาดเคลื่อนสัมพัทธ์ของพลังงานที่ $t = 10$ คือ $|0.999776 - 1| \approx 2.24 \times 10^{-4}$ ที่ $h = 0.002$ คือ $|0.999991 - 1| \approx 8.5 \times 10^{-6}$ ขั้นที่เล็กลง 5 เท่าทำให้ค่าลดลงประมาณ 26 เท่า (ราว 5^2)

2. วิธีออยเลอร์ชัดแจ้งที่ $h = 0.002$ ให้ $E/E_0 \approx 6.35$ เมื่อ $t = 10$ วินาที พลังงานยังโตอยู่ แม้จะช้ากว่าที่ $h = 0.01$ มาก

3. ถ้า $|\lambda|^2 = 1 + 15h^2$ แล้วตัวคูณของพลังงานหลัง $n = 10/h$ ก้าวคือ $|\lambda|^{2n} = (1 + 15h^2)^{10/h}$ ซึ่งมากกว่า 1 เสมอสำหรับทุก $h > 0$ และที่ $h = 0.002$ มีค่าประมาณ 1.35 ขนาดขั้นที่เล็กลงลดตัวคูณต่อก้าวแต่ไม่ได้ทำให้ตัวคูณรวมไม่เกิน 1 วิธีชัดแจ้งจึงไม่นิ่งสำหรับทุกขนาดขั้น (ตรงกับผล $|\lambda|^2 = 1 + \omega^2 h^2 > 1$ ในทฤษฎีบทการแกว่งที่ไม่มีแรงหน่วงของหัวข้อ 12.2) เชือกเองขยายตัวเร็วกว่านี้เพราะมีโหมดที่มีความถี่สูงกว่า

ฉบับร่าง · Preprint · 2026-10-23

ตรวจด้วยโค้ด: `test_ch12.py` (`test_exercise_12_3_7_strand_with_half_the_step` และ `test_strand_verlet_energy_drift_stays_below_one_thousandth_for_10_s`)

สรุปบทที่ 12 ในหน้าเดียว

ตารางนี้เทียบวิธีหลักของบท โดย h คือขนาดขั้นเวลา และ $f_k = f(t_k, y_k)$

วิธี	ค่าที่ใช้ต่อก้าว	อันดับ	ต้นทุนต่อก้าว	สิ่งที่ต้องระวัง
ออยเลอร์ชัดแจ้ง (12.1)	y_k	1	ประเมิน f หนึ่งครั้ง	นิ่งบนสปริงของ Fiedler เมื่อ $h < \frac{1}{150}$
ออยเลอร์แบบโดยปริยาย (12.2)	y_k และ y_{k+1}	1	แก้สมการหนึ่งครั้ง	ไม่โตกับ $y' = -2y$ แต่ความแม่นเป็นอันดับ 1
กฎ เชิง สี่เหลี่ยม คางหมู (12.2)	y_k และ y_{k+1}	2	แก้สมการหนึ่งครั้ง	แม่นกว่า แต่ต้องแก้สมการ
ออย เลอ ร์ ชิ ม เพล ก ติก (12.2)	x_k, v_k	1	ประเมิน a หนึ่งครั้ง	นิ่งบนสปริงเมื่อ $h < h^* \approx 0.5098$
เวอร์เลต (12.3)	x_k, x_{k-1}	2	ประเมิน a หนึ่งครั้ง	ต้องค่าเริ่มต้นสองค่า รากซ้ำที่ 1 (ตัดปลาย $O(h^4)$)
อดัมส์-บาชฟอร์ทสองขั้น (12.3)	y_k, y_{k+1}, f_k, f_{k+1}	2	ประเมิน f หนึ่งครั้ง (ใช้ค่าเก่าซ้ำ)	ผ่านเงื่อนไขราก

สูตรที่ควรจำ: ออยเลอร์คือ $y_{k+1} = y_k + h f_k$ กฎเชิงสี่เหลี่ยมคางหมูคือ $y_{k+1} = y_k + \frac{h}{2}(f_k + f_{k+1})$ และวิธีเวอร์เลตคือ $x_{k+1} = 2x_k - x_{k-1} + h^2 a(x_k)$ ลู่เข้าเมื่อมีอันดับอย่างน้อย 1 และผ่านเงื่อนไขราก (หัวข้อ 12.3) และการข้ามกำแพงบางมีสัดส่วนที่จับได้ $\min\{1,\, w/(v\Delta t)\}$ (หัวข้อ 12.2)

อ่านต่อ. บทที่ 13 ขยายความคิดของวิธีออยเลอร์ให้ใช้หลายจุดต่อหนึ่งก้าว และปรับขนาดขั้นอัตโนมัติ บทที่ 14 พบระบบที่แข็งเกินไป ซึ่งวิธีโดยปริยายและวิธีหลายขั้นแบบโดยปริยายจำเป็นต้องใช้ ดูรายละเอียดวิธีหลายขั้นได้ใน Iserles 2009, §2.1--2.3, pp. 19--32

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 13

ก้าวให้ฉลาดขึ้น

สถานการณ์สมมติ: ยานอวกาศในเกมที่โคจรรอบดาวเคราะห์

ลองนึกถึงเกมอวกาศที่ยานของเรากำลังโคจรรอบดาวเคราะห์ วงโคจรเป็นรูปวงรีแบน ๆ ตอนที่ยานอยู่ไกลจากดาว ยานจะเคลื่อนช้า เหมือนลอยเอื่อย ๆ แต่พอเข้าใกล้ดาว ยานจะพุ่งผ่านไปอย่างรวดเร็ว ทั้งสองช่วงอยู่ในวงโคจรเดียวกัน เกมคำนวณฟิสิกส์ทีละเฟรม สมมติว่าเฟรมละ $1/60$ วินาที และในแต่ละเฟรมมันอ่านแรงดึงดูดที่จุดเริ่มเพียงจุดเดียว แล้วเดินไปตามทิศนั้น นี่คือวิธีออยเลอร์ของบทที่ 12 ยานของเราเป็นยานที่เราตั้งขึ้นเอง ตัวเลขทั้งหมดในบทนี้มาจากโค้ดของเรา และไม่ได้บอกว่าเกมใดทำเช่นนี้

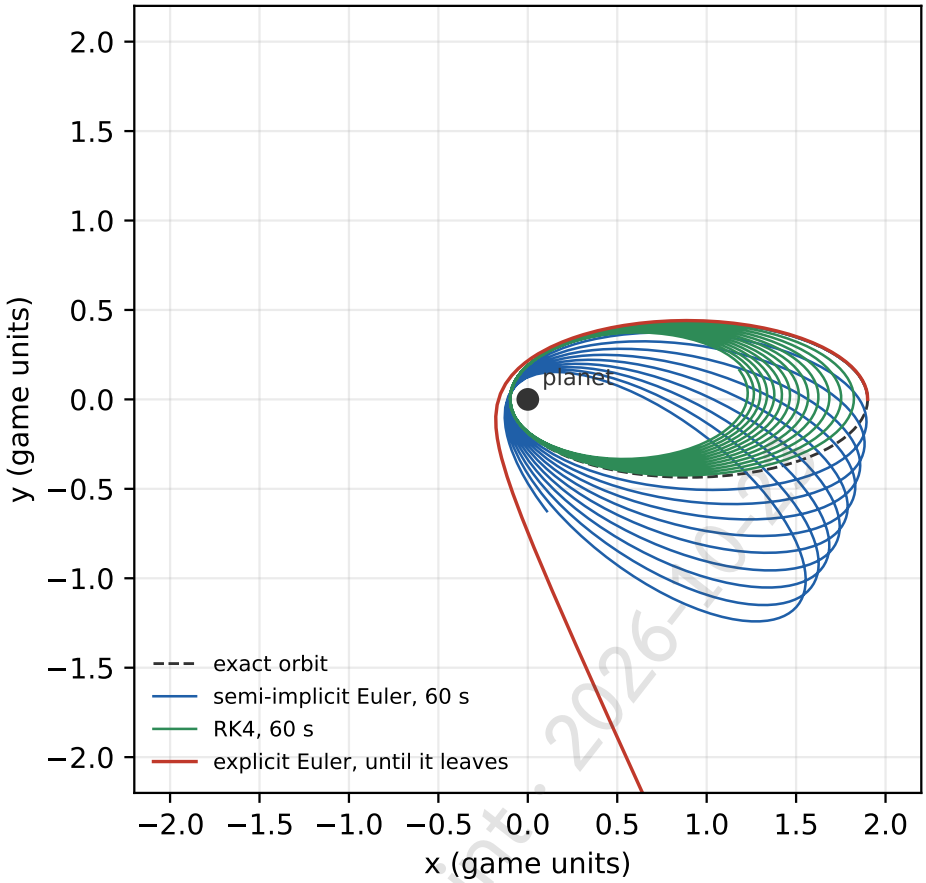
ถ้ารัน 60 วินาทีของเกม (ราว 9.5 รอบ) ยานที่ใช้วิธีออยเลอร์จะพุ่งออกไปไกลจนเลยกรอบภาพที่ $r = 3$ หน่วยในเวลา 4.4 วินาที ส่วนยานที่ใช้วิธีรุงเงอ-คุททาอันดับสี่ยังโคจรอยู่ใกล้ดาว แต่พลังงานของมันค่อย ๆ ลดลงจนวงโคจรเล็กลง คำถามของบทนี้จึงมีสองข้อ ทำไมการอ่านลูกศรเพียงจุดเดียวในแต่ละก้าวจึงพลาดตอนที่ยานเข้าใกล้ดาว และเราจะอ่านลูกศรหลายจุดในก้าวเดียวได้อย่างไร

ภาพรวมบท

บทที่ 12 ทำให้เราเดินทีละก้าวด้วยลูกศรจุดเดียว และเห็นว่าความผิดพลาดสะสมได้เร็ว บทนี้มีสองหัวข้อ

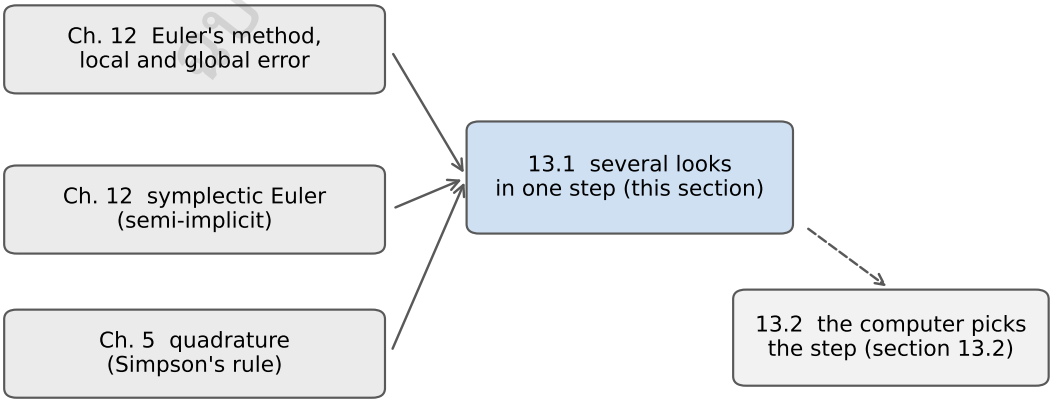
- หัวข้อ 13.1 ลองหลายจุดต่อหนึ่งก้าว เราดูว่าวิธีรุงเงอ-คุททาอ่านลูกศรหลายจุดในก้าวเดียวอย่างไร ทำไมอันดับจึงสูงขึ้น และทำไมอันดับสูงกว่าไม่ได้ช่วยเสมอไป
- หัวข้อ 13.2 ให้เครื่องเลือกขนาดก้าวเอง เราประมาณความผิดพลาดของแต่ละก้าว แล้วเลือกก้าวที่สั้นลงเมื่อยานเข้าใกล้ดาว

ฉบับร่าง Preprint 2026-10-23

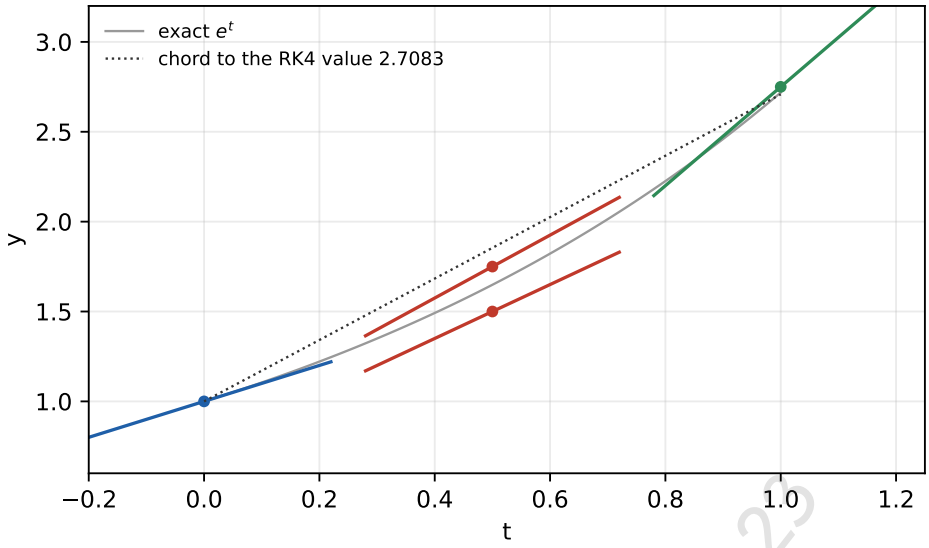


รูปที่ 13.1: วงโคจรของยานสมมติ ($e = 0.9$, $\mu = 1$, หน่วยเวลาเป็นวินาทีของเกม) ที่ 60 ก้าวต่อวินาที เส้นประคือวงโคจรจริง เส้นแดงคือวิธีออยเลอร์จนถึงจุดที่ออกนอกกรอบ เส้นน้ำเงินคือวิธีออยเลอร์ซิมเพลกติก และเส้นเขียวคือ RK4 ทั้ง 60 วินาที (สร้างจาก `numanal/ch13.py`)

ฉบับร่าง · Preprint, 2026-10-27



รูปที่ 13.2: แผนผังบทที่ 13 กล่องสีฟ้าคือหัวข้อ 13.1 ลูกศรทึบคือแนวคิดที่หัวข้อหนึ่งใช้ ลูกศรประคือแนวคิดที่หัวข้อ 13.2 ต่อยอดจากหัวข้อ 13.1 (สร้างจาก `numanal/ch13.py`)



รูปที่ 13.3: หนึ่งก้าวของ RK4 สำหรับ $y' = y$ จาก $(0, 1)$ ด้วย $h = 1$ (เพื่อให้เห็นชัด) จุดสี่จุดคือชั้นของวิธี และแต่ละเส้นคือความชันที่อ่านได้ที่จุดนั้น ค่าถ่วงน้ำหนักของความชันทั้งสี่รวมกันให้ $y_1 = 65/24 \approx 2.7083$ ใกล้ $e \approx 2.7183$ กว่าค่า 2 ที่ได้จากออยเลอร์ (สร้างจาก `numanal/ch13.py`)

13.1 ลองหลายจุดต่อหนึ่งก้าว

สถานการณ์: หนึ่งนาทีของเกมที่ 60 ก้าวต่อวินาที

ยานสามลำเริ่มที่จุดเดียวกันและเดินเวลาเท่ากัน ลำแรกใช้วิธีออยเลอร์ ลำที่สองใช้วิธีออยเลอร์ซิมเพลกติก ลำที่สามใช้ RK4 ซึ่งเป็นวิธีที่มีอันดับสี่ ลำแรกพุ่งออกนอกกรอบในเวลา 4.4 วินาที ลำที่สองและสามยังโคจรอยู่เมื่อครบ 60 วินาที แต่ห่างจากตำแหน่งจริงต่างกัน คือ 0.43 และ 0.91 หน่วย

ทำไมอันดับที่สูงกว่าจึงให้ผลที่แย่กว่าในกรณีนี้ และอันดับนั้นหมายถึงอะไรกันแน่ คำตอบเริ่มจากการอ่านลูกศรมากกว่าหนึ่งจุดในก้าวเดียว

ภาพรวมก่อนอ่าน

วิธีออยเลอร์อ่านลูกศรที่จุดเริ่มเพียงจุดเดียวแล้วเดินไปตามทิศนั้นตลอดก้าว ถ้าลูกศรเปลี่ยนมากในก้าวนั้น ความผิดพลาดก็มาก ความคิดของวิธีรุงเงอ-คุททาคือ อ่านลูกศรที่จุดอื่นในก้าวด้วย แล้วถ่วงน้ำหนักค่าที่อ่านได้ให้ได้ทิศเฉลี่ยของเส้นโค้งในก้าวนั้น

จากปริพันธ์สู่ลูกศรหลายจุด

ให้ $\mathbf{y}$ เป็นผลเฉลยจริงของ $\mathbf{y}' = f(t, \mathbf{y})$ เมื่อเดินจาก t_n ถึง $t_n + h$ เราได้

$$\mathbf{y}(t_n + h) = \mathbf{y}(t_n) + \int_{t_n}^{t_n+h} f\big(\tau, \mathbf{y}(\tau)\big)\, d\tau.$$

ถ้าแทนปริพันธ์ด้วย **ควอดราเจอร์** (quadrature) ที่จุดซ้ายสุดของช่วง เราจะได้วิธีออยเลอร์ ซึ่งมีรูป $\mathbf{y}_{n+1} = \mathbf{y}_n + h f(t_n, \mathbf{y}_n)$ ถ้าอยากใช้จุดกลางแทน ปัญหาคือเราไม่รู้ค่า $\mathbf{y}$ ที่จุดกลางนั้น ทางออกคือประมาณค่านั้นจากก้าวที่เดินมาแล้ว แล้วอ่านลูกศรที่จุดประมาณนั้น วิธีรุงเงอ-คุททาทำอย่างนี้ซ้ำ ๆ ในก้าวเดียว

ร่างปรับปรุง · Preprint · 2026-10-23

นิยาม 13.1.1: วิธีรุงเงอ-คุททาแบบชัดแจ้ง

ให้จำนวนชั้น ν จุดเวลา $c_1 = 0, c_2, \ldots, c_\nu$ เมทริกซ์ $A = (a_{ij})$ ที่มีสมาชิกเฉพาะเมื่อ $j < i$ และน้ำหนัก $b_1, \ldots, b_\nu$ วิธี**วิธีรุงเงอ-คุททา** (Runge--Kutta method)แบบ**ชัดแจ้ง** (explicit method) เดินหนึ่งก้าวจาก $(t_n, \mathbf{y}_n)$ โดย

$$\mathbf{k}_i = f\Big(t_n + c_i h,\ \mathbf{y}_n + h\sum_{j<i} a_{ij}\mathbf{k}_j\Big), \qquad i = 1, \ldots, \nu, \qquad \mathbf{y}_{n+1} = \mathbf{y}_n + h\sum_{i=1}^{\nu} b_i \mathbf{k}_i.$$

เราเรียก $\mathbf{k}_i$ ว่า**ความชันของชั้น** (stage slope) ซึ่งเป็นค่า f ที่จุดของชั้น i บอกว่าผลเฉลยเปลี่ยนเร็วแค่ไหนตรงจุดนั้น และเรียก $c_i h$ ว่าจุดเวลาของชั้น i เราใช้เงื่อนไข $c_i = \sum_{j<i} a_{ij}$ ตลอดบทเป็น**ข้อตกลง**ของบท ไม่ใช่เงื่อนไขที่จำเป็นต่อการมีอันดับหนึ่ง อันดับหนึ่งต้องการเพียง $\sum_i b_i = 1$ ดูทฤษฎีบท 13.1.1 และกล่องข้อควรระวังหลังทฤษฎีบทนั้น

อ่านแบบเข้าใจง่าย

แต่ละชั้นเปรียบเหมือนการหยุดมองลูกศรที่จุดหนึ่ง ชั้นแรกหยุดที่จุดเริ่ม ชั้นต่อไปหยุดที่จุดที่เดินประมาณมาแล้ว และไม่ต้องแก้สมการ เพราะ A เป็นสามเหลี่ยมล่างเคร่งครัด จึงเป็น**วิธีชัดแจ้ง** สุดท้ายเราถ่วงน้ำหนักความชันทั้งหมดด้วย b_i แล้วเดินไปตามค่าเฉลี่ยนั้น ตาราง c, A และ b เก็บไว้ในตารางเดียว เรียกว่า **ตารางบุชเชอร์** (Butcher tableau) ตัวอย่างเช่น RK4 เขียนได้เป็น

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

ลองเอง: หนึ่งก้าวด้วยมือ

ให้ $y' = y$, $y(0) = 1$ และ $h = \frac{1}{2}$ คำตอบจริงที่ $t = \frac{1}{2}$ คือ $e^{1/2} \approx 1.6487213$

วิธีออยเลอร์ ได้ $y_1 = 1 + \frac{1}{2}\cdot 1 = 1.5$

วิธีจุดกึ่งกลาง ($c_2 = \frac{1}{2}$, $a_{21} = \frac{1}{2}$, $b = (0, 1)$) ความชันชั้นที่สองอ่านที่ $\xi_2 = 1 + \frac{1}{2}\cdot\frac{1}{2}\cdot 1 = 1.25$ จึงได้ $\mathbf{k}_2 = 1.25$ และ $y_1 = 1 + \frac{1}{2}\cdot 1.25 = 1.625$

วิธีของฮอยน์ (สองชั้นอันดับสอง, $c_2 = 1$, $a_{21} = 1$, $b = (\frac{1}{2}, \frac{1}{2})$) ได้ $\xi_2 = 1.5$ และ $y_1 = 1 + \frac{1}{2}\cdot\frac{1}{2}(1 + 1.5) = 1.625$

RK4 ความชันทั้งสี่คือ $\mathbf{k}_1 = 1$, $\mathbf{k}_2 = 1.25$, $\mathbf{k}_3 = 1.3125$, $\mathbf{k}_4 = 1.65625$ แล้ว $y_1 = 1 + \frac{1}{12}(1 + 2\cdot 1.25 + 2\cdot 1.3125 + 1.65625) = 1.6484375$

ความคลาดเคลื่อนคือ 0.149 สำหรับออยเลอร์ 0.024 สำหรับจุดกึ่งกลางและฮอยน์ และ 0.00028 สำหรับ RK4

อ่านแบบเข้าใจง่าย: ทำไมวิธีกึ่งกลางกับฮอยน์ได้ค่าเท่ากัน

สำหรับ $y' = y$ วิธีทั้งสองได้ $1 + h + \frac{1}{2}h^2$ เสมอ นี่คือพหุนามเทย์เลอร์ถึงพจน์ที่สอง จึงไม่แปลกที่ทั้งสองเท่ากัน ความต่างจะปรากฏเมื่อลูกศรเป็นฟังก์ชันของ y ที่ซับซ้อนขึ้น ซึ่งเราจะเห็นในการคำนวณวงโคจร

ร่างปรับปรุง · Preprint · 2026-04-23

เกร็ดเพิ่มเติม

ถ้า f ไม่ขึ้นกับ $\mathbf{y}$ เช่น $y' = f(t)$ แล้ว $\mathbf{k}_i = f(t_n + c_i h)$ และ RK4 ให้

$$y_{n+1} - y_n = \frac{h}{6}\Big(f(t_n) + 4f\big(t_n + \tfrac{h}{2}\big) + f(t_n + h)\Big),$$

ซึ่งคือ **กฎของซิมป์สัน** (Simpson's rule) ในบทที่ 5 ดังนั้นวิธีรุงเงอ-คุททาเป็นการแทนปริพันธ์ด้วยควอดราเจอร์ที่ค่อย ๆ ซับซ้อนขึ้น ส่วนวิธีออยเลอร์คือกฎจุดซ้ายสุด

เงื่อนไขอันดับ

ตามที่เห็นในหัวข้อ 12.1 วิธีมี**อันดับของวิธี** (order of a method) p เมื่อ**ค่าคลาดเคลื่อนตัดปลายเฉพาะที่** (local truncation error) เป็น $O(h^{p+1})$ เราต้องการรู้ว่าสัมประสิทธิ์ c, A, b ต้องเป็นอย่างไร การขยายทุกอย่างเป็นอนุกรมเทย์เลอร์ให้คำตอบและสำหรับอันดับไม่เกินสาม การขยายแบบนี้ยังไม่ซับซ้อนมาก

ทฤษฎีบท 13.1.1: เงื่อนไขอันดับถึงอันดับสาม

ให้วิธีรุงเงอ-คุททาชัดแจ้งที่ใช้ข้อตกลงของบท $c_i = \sum_{j<i} a_{ij}$ สำหรับทุก i และให้วิธีมีอันดับอย่างน้อย p สำหรับทุกปัญหาที่หาอนุพันธ์ได้ต่อเนื่องพอ โดย $p \le 3$ แล้ว

$$\begin{aligned} p \ge 1: &\quad \sum_i b_i = 1, \\ p \ge 2: &\quad \sum_i b_i c_i = \tfrac{1}{2}, \\ p \ge 3: &\quad \sum_i b_i c_i^2 = \tfrac{1}{3} \quad \text{และ} \quad \sum_{i,j} b_i a_{ij} c_j = \tfrac{1}{6}. \end{aligned}$$

ผลกลับ (เงื่อนไขเหล่านี้ทำให้วิธีมีอันดับนั้น) เป็นจริงด้วย เราไม่พิสูจน์ส่วนนั้น Iserles 2009, §3.2, pp. 39--40 แสดงรายการเงื่อนไขของ $\nu = 3$ และอ้างโดยไม่พิสูจน์ว่าการทดสอบกับสมการสเกลาร์อิสระจากเวลาใช้ได้สำหรับ $p \le 3$ การพิสูจน์ทั่วไปอยู่ในทฤษฎีต้นไม้ของบุชเชอร์ ดูเพิ่มใน Marlis Hochbruck, Numerische Mathematik, บทที่ 8 ทฤษฎีบท 8.16 หน้า 172 Hochbruck 2016

พิสูจน์. **แนวคิด.** เลือกปัญหาที่ง่ายที่สุดที่ยังบอกเงื่อนไขได้ครบ คือ $y' = f(y)$ สเกลาร์ที่ไม่ขึ้นกับเวลา แล้วเทียบสัมประสิทธิ์ของ h ทีละกำลัง ทั้งสองข้างใช้ค่า f, f' และ f'' ที่ y_n เท่านั้น

1. ผลเฉลยจริงมี $y' = f$, $y'' = f'f$ และ $y''' = f''f^2 + f'^2 f$ โดยกฎลูกโซ่ ดังนั้น
$$y(t_n + h) = y_n + hf + \tfrac{1}{2}h^2 f'f + \tfrac{1}{6}h^3\big(f''f^2 + f'^2 f\big) + O(h^4).$$
2. ให้ $\xi_i - y_n = h\sum_{j<i} a_{ij}\mathbf{k}_j$ เนื่องจาก $\mathbf{k}_j = f + O(h)$ เราได้ $\xi_i - y_n = hc_i f + h^2 f'f \sum_{j<i} a_{ij}c_j + O(h^3)$ จากนั้นขยาย $\mathbf{k}_i = f(\xi_i)$ ถึงกำลังสองของ $\xi_i - y_n$:
$$\mathbf{k}_i = f + hc_i f'f + h^2\Big(f'^2 f \sum_{j<i} a_{ij}c_j + \tfrac{1}{2}c_i^2 f''f^2\Big) + O(h^3).$$
3. แทนลงในสูตร $y_{n+1} = y_n + h\sum_i b_i \mathbf{k}_i$:
$$y_{n+1} = y_n + h\Big(\sum_i b_i\Big)f + h^2\Big(\sum_i b_i c_i\Big)f'f + h^3\Big[\Big(\sum_{i,j} b_i a_{ij} c_j\Big)f'^2 f + \tfrac{1}{2}\Big(\sum_i b_i c_i^2\Big)f''f^2\Big] + O(h^4).$$

ฉบับร่าง Preprint · 2026-10-23

4. ผลต่างระหว่างผลเฉลยจริงกับวิธีคือ

$$h\big(1-\textstyle\sum b_i\big)f + h^2\big(\tfrac{1}{2}-\textstyle\sum b_i c_i\big)f'f + h^3\Big[\big(\tfrac{1}{6}-\textstyle\sum b_i a_{ij} c_j\big)f'^2 f + \tfrac{1}{2}\big(\tfrac{1}{3}-\textstyle\sum b_i c_i^2\big)f''f^2\Big] + O(h^4).$$

5. เราต้องการให้ผลต่างนี้เป็น $O(h^{p+1})$ สำหรับทุกปัญหาที่ราบรื่น เลือกพหุนาม $f(y)=\alpha+\beta(y-y_n)+\frac{1}{2}\gamma(y-y_n)^2$ ซึ่งให้ $f=\alpha$, $f'=\beta$, $f''=\gamma$ ที่ y_n โดย α,β,γ เป็นจำนวนจริงใด ๆ จึงได้เงื่อนไขว่า $(1-\sum b_i)\alpha=0$ เมื่อ $\alpha=1$ ได้ $\sum b_i=1$ เมื่อ $p\ge 1$ เมื่อ $p\ge 2$ ได้ $(\frac{1}{2}-\sum b_i c_i)\alpha\beta=0$ จึง $\sum b_i c_i=\frac{1}{2}$ และเมื่อ $p\ge 3$ ได้ $(\frac{1}{6}-\sum b_i a_{ij}c_j)\beta^2\alpha+\frac{1}{2}(\frac{1}{3}-\sum b_i c_i^2)\gamma\alpha^2=0$ โดยใช้ $\alpha=\beta=1,\gamma=0$ จะได้เงื่อนไขแรก และใช้ $\alpha=1$, $\beta=0,\gamma=1$ จะได้เงื่อนไขที่สอง

สรุป: การขยายเทย์เลอร์ของวิธีและของผลเฉลยจริงที่จุดเดียวกันให้เงื่อนไขอันดับที่ต้องการโดยตรง □

ข้อควรระวัง: เราพิสูจน์เฉพาะด้านจำเป็น

เราใช้เพียงปัญหาที่ไม่ขึ้นกับเวลาในการพิสูจน์ ซึ่งเพียงพอสำหรับด้านจำเป็น เพราะทฤษฎีบทกำหนดให้วิธีมีอันดับสำหรับทุกปัญหา รวมทั้งปัญหาแบบนี้ ส่วนผลกลับเราอ้างโดยไม่พิสูจน์ซ้ำ

ข้อควรระวัง: ข้อตกลงแถวไม่ใช่เงื่อนไขของอันดับหนึ่ง

ลองวิธีสองชั้นที่มี $A=0$, $c=(0,1)$ และ $b=(\frac{1}{2},\frac{1}{2})$ ผลของหนึ่งก้าวคือ $\mathbf{y}_{n+1}=\mathbf{y}_n+\frac{h}{2}\big(f(t_n,\mathbf{y}_n)+f(t_n+h,\mathbf{y}_n)\big)$ วิธีนี้มี $\sum_i b_i=1$ จึงมีอันดับหนึ่ง แม้ $c_2=1\neq 0=a_{21}$ ซึ่งฝ่าฝืนข้อตกลง ค่าคลาดเคลื่อนเฉพาะที่คือ $\frac{1}{2}h^2 f_{\mathbf{y}}f+O(h^3)$ จึงไม่มีอันดับสอง แม้จะมี $\sum_i b_i c_i=\frac{1}{2}$ เพราะเงื่อนไขอันดับสองก็อาศัยข้อตกลงเช่นกัน ดู `test_ch13.py`

ทฤษฎีบท 13.1.2: วิธีจุดกึ่งกลางและวิธีของฮอยน์มีอันดับสอง

ให้ $f(t,\mathbf{y})$ มีอนุพันธ์ย่อยอันดับสองต่อเนื่อง และให้วิธีสองชั้นที่มี $c_2=a_{21}=c$, $b_1+b_2=1$ และ $b_2 c=\frac{1}{2}$ วิธีนี้มีค่าคลาดเคลื่อนตัดปลายเฉพาะที่เป็น $O(h^3)$ ดังนั้นมีอันดับสอง วิธีจุดกึ่งกลาง ($c=\frac{1}{2}$, $b=(0,1)$) และวิธีของฮอยน์ ($c=1$, $b=(\frac{1}{2},\frac{1}{2})$) เป็นกรณีเช่นนี้

พิสูจน์. **แนวคิด.** เทียบพจน์ของ h ถึง h^2 ทั้งสองข้าง ผลเฉลยจริงกับวิธีตรงกันเมื่อมีเงื่อนไขสามข้อนี้

1. ที่ $(t_n,\mathbf{y}_n)$ ให้ $\mathbf{y}''=f_t+f_{\mathbf{y}}f$ โดยกฎลูกโซ่ เทย์เลอร์ให้ $\mathbf{y}(t_n+h)=\mathbf{y}_n+hf+\frac{1}{2}h^2(f_t+f_{\mathbf{y}}f)+O(h^3)$
2. วิธีมี $\xi_2=\mathbf{y}_n+chf$ จึง $\mathbf{y}_{n+1}=\mathbf{y}_n+h\big[b_1 f+b_2 f(t_n+ch,\ \mathbf{y}_n+chf)\big]$
3. ขยาย $f(t_n+ch,\mathbf{y}_n+chf)=f+ch(f_t+f_{\mathbf{y}}f)+O(h^2)$ เพราะ f มีอนุพันธ์อันดับสองต่อเนื่อง
4. แทนลงไป $\mathbf{y}_{n+1}=\mathbf{y}_n+h(b_1+b_2)f+h^2 b_2 c(f_t+f_{\mathbf{y}}f)+O(h^3)$ ใช้ $b_1+b_2=1$ และ $b_2c=\frac{1}{2}$ ได้ $\mathbf{y}_{n+1}=\mathbf{y}_n+hf+\frac{1}{2}h^2(f_t+f_{\mathbf{y}}f)+O(h^3)$
5. ลบกับข้อ 1 ได้ค่าคลาดเคลื่อนตัดปลายเฉพาะที่เป็น $O(h^3)$

สรุป: วิธีออยเลอร์มี $b_1=1$ จึงไม่มีพจน์ $h^2 b_2 c$ และได้อันดับหนึ่งเท่านั้น □

ฉบับร่าง Preprint · 2026-10-23

ตารางที่ 13.1: ความคลาดเคลื่อนที่ $t = 1$ ของ $y' = y$ เมื่อใช้ n ก้าวเท่า ๆ กัน และอัตราส่วนของความคลาดเคลื่อนแถวก่อนกับแถวนี้ (สร้างจาก `numanal/ch13.py` และตรวจด้วย `test_ch13.py`)

n	ออยเลอร์	อัตราส่วน	ฮอยน์	อัตราส่วน	RK4	อัตราส่วน
4	0.276876	---	0.0234261	---	7.18893×10^{-5}	---
8	0.152497	1.8156	0.00644059	3.6373	4.98404×10^{-6}	14.4239
16	0.0803533	1.8978	0.00168831	3.8148	3.28118×10^{-7}	15.1898
32	0.0412917	1.9460	0.000432154	3.9067	2.10479×10^{-8}	15.5892
64	0.0209369	1.9722	0.000109317	3.9532	1.33272×10^{-9}	15.7931

ทฤษฎีบท 13.1.3: วิธีอันดับสี่ของ RK4

วิธีรุงเงอ-คุททาที่มีตาราง RK4 ตามนิยาม วิธีรุงเงอ-คุททาแบบชัดแจ้ง มีค่าคลาดเคลื่อนตัดปลายเฉพาะที่เป็น $O(h^5)$ สำหรับปัญหาที่ราบรื่นพอ ดังนั้นมีอันดับสี่

ที่มา. เรายอมรับผลนี้โดยไม่พิสูจน์ซ้ำ ดูได้ใน Iserles 2009, §3.2, p. 41 ซึ่งแสดงตารางนี้เป็นวิธีชัดแจ้งสี่ชั้นอันดับสี่ที่รู้จักกันดีที่สุด การพิสูจน์ต้องขยายเทย์เลอร์ถึงพจน์ h^4 และเราไม่ทำที่นี่ ในโค้ดเราตรวจว่าเงื่อนไขอันดับสามและสี่ทุกข้อเป็นศูนย์บนตารางนี้ และอัตราส่วนความคลาดเคลื่อนในตาราง 13.1 เข้าหา 16

เกร็ดเพิ่มเติม

บทความที่เราอ้างถึงในบทที่ 12 ระบุว่า RK4 เป็นวิธีอันดับสี่สำหรับหาผลเฉลยเชิงตัวเลข Fiedler 2004b ซึ่งตรงกับผลข้างบน

ตาราง 13.1 ให้ความรู้สึกที่ชัด เมื่อ n เพิ่มเป็นสองเท่า ความคลาดเคลื่อนของออยเลอร์ลดลงราวครึ่งหนึ่ง (อันดับหนึ่ง) ของฮอยน์ลดลงราวหนึ่งในสี่ (อันดับสอง) และของ RK4 ลดลงราวหนึ่งในสิบหก (อันดับสี่) อัตราส่วนยังไม่เข้าค่าเต็มที่ที่ $n = 8$ แต่ใกล้ค่านั้นขึ้นทุกแถว

สรุปใจความ

- วิธีรุงเงอ-คุททาอ่านลูกศรหลายจุดในก้าวเดียว แล้วถ่วงน้ำหนักด้วย b โดยไม่ต้องแก้สมการ เมื่อ A เป็นสามเหลี่ยมล่างเคร่งครัด
- $\sum b_i = 1$ ทำให้มีอันดับหนึ่ง และ $\sum b_i c_i = \tfrac{1}{2}$ ทำให้มีอันดับสอง (ภายใต้ข้อตกลง $c_i = \sum_j a_{ij}$) อีกสองเงื่อนไขทำให้มีอันดับสาม
- RK4 มีอันดับสี่ และเมื่อ h เล็กพอ ความคลาดเคลื่อนลดลงราว 16 เท่าเมื่อ h ลดครึ่งหนึ่ง

วงโคจรของยานอวกาศ

ตอนนี้เราเอาวิธีไปใช้กับยานของเรา ยานเคลื่อนที่ตามสมการ $\mathbf{r}'' = -\mu \mathbf{r}/|\mathbf{r}|^3$ ที่เรียกว่า**ปัญหาเคปเลอร์** (Kepler problem) ซึ่งอธิบายวัตถุเบาที่ถูกดึงเข้าหาดาวด้วยแรงโน้มถ่วง เราตั้ง $\mu = 1$ และหน่วยเวลาเป็นวินาทีของเกม ตัวแปรสถานะเป็น $\mathbf{y} = (x, y, v_x, v_y)$

ยานเริ่มที่**จุดไกลดาวที่สุด** (apoapsis) ซึ่งอยู่ที่ระยะ $a(1 + e) = 1.9$ เมื่อ $a = 1$ และ $e = 0.9$ ค่า e นี้คือ**ความเยื้องศูนย์กลาง** (eccentricity) ซึ่งบอกว่าวงรีแบนเพียงใด (0 คือวงกลม) ความเร็วที่จุดนั้นได้จากพลังงานของวงรี คือ $v =$

ร่างต้น · Preprint · 2026-10-25

ตารางที่ 13.2: ยานที่ 60 ก้าวต่อวินาที เป็นเวลา 60 วินาที (ราว 9.5 รอบ) พลังงานที่ค่าจริงคือ -0.5 และระยะห่างวัดจากตำแหน่งจริงที่ $t = 60$ (สร้างจาก `numanal/ch13.py`)

วิธี	พลังงานที่ $t = 60$	พลังงานต่ำสุด–สูงสุดที่เห็น	ระยะห่างที่ $t = 60$
ออยเลอร์	$+2.2936$	-0.5366 ถึง $+2.4256$	122.33 (ออกนอกกรอบที่ 4.4 วินาที)
ออยเลอร์ซิมเพลกติก	-0.4721	-1.2831 ถึง $+0.4983$	0.4263
RK4	-0.7515	-0.7515 ถึง -0.4999	0.9083

ตารางที่ 13.3: ระยะห่างจากตำแหน่งจริงที่ $t = 60$ เมื่อเปลี่ยนขนาดก้าว ที่ $h = 1/60$ วิธีออยเลอร์ซิมเพลกติกดีกว่า RK4 และที่ $h = 1/480$ RK4 ดีกว่า (สร้างจาก `numanal/ch13.py`)

h	ออยเลอร์	ออยเลอร์ซิมเพลกติก	RK4
$1/60$	122.33	0.4263	0.9083
$1/120$	103.38	0.1110	0.6405
$1/240$	79.96	0.02742	0.02856
$1/480$	53.90	0.006653	0.000914
$1/960$	24.23	0.001638	2.90×10^{-5}

$\sqrt{(1-e)/(1+e)} \approx 0.2294$ **พลังงานต่อหน่วยมวล** (energy per unit mass) คือผลรวมของพลังงานจลน์และพลังงานศักย์ต่อหนึ่งหน่วยมวล $E = \tfrac{1}{2}v^2 - \mu/r$ และวงโคจรจริงมี $E = -\tfrac{1}{2}$ เสมอ คาบของวงโคจรคือ 2π วินาที

คำตอบจริงหาได้จากสมการของเคปเลอร์ ให้ $M = \pi + t$ (เพราะเราเริ่มที่จุดไกลสุด และคาบคือ 2π) แล้วแก้ $E - e\sin E = M$ หา E จากนั้น $x = -a(\cos E - e)$ และ $y = -a\sqrt{1-e^2}\sin E$ ตาราง 13.2 และ 13.3 เทียบกับคำตอบนี้

วิธีออยเลอร์ซิมเพลกติกของหัวข้อ 12.2 ใช้กับยานได้โดยปรับความเร็วก่อนด้วยตำแหน่งเดิม แล้วเลื่อนตำแหน่งด้วยความเร็วใหม่ ส่วนวิธีที่สามคือ RK4 ใช้กับสมการทั้งหมด เราวัดผลที่ 60 ก้าวต่อวินาที คือ $h = 1/60$

ออยเลอร์พุ่งออกไปจนรัศมีเกิน 120 หน่วย พลังงานกลายเป็นบวก จึงหลุดจากดาว ออยเลอร์ซิมเพลกติกยังอยู่ในวงโคจร แต่จุดไกลดาวที่สุดของวงรีหมุนไปประมาณ $-4.14°$ ต่อรอบ จากการวัดของเรา ตามภาพ 13.1 (เส้นสีน้ำเงิน) พลังงานที่คำนวณจากค่าที่เก็บในแต่ละก้าวแกว่งขึ้นลงรอบ -0.5 ไม่ได้ไหลไปทางเดียว และสิ้นสุดที่ -0.4721 ส่วน RK4 ยังโคจรอยู่ใกล้ดาว แต่พลังงานลดลงโดยรวมจาก -0.5 เป็น -0.7515 วงโคจรเล็กลง ทำให้ระยะห่างจากตำแหน่งจริงเพิ่มขึ้น

ตาราง 13.3 บอกเรื่องสำคัญ อันดับสี่ของ RK4 ต้องรอให้ก้าวเล็กพอ ที่ $h = 1/240$ วิธีออยเลอร์ซิมเพลกติกยังแม่นกว่าเล็กน้อย แต่ที่ $h = 1/480$ RK4 แม่นกว่าชัดเจน จุดที่สลับกันอยู่ระหว่าง $1/240$ กับ $1/480$ ซึ่งสั้นกว่าที่ 60 ก้าวต่อวินาทีถึงแปดเท่า

ข้อควรระวัง: อันดับสูงไม่ได้ชนะเสมอ

ทฤษฎีบทอันดับบอกพฤติกรรมเมื่อ $h \to 0$ ไม่ได้บอกว่าที่ h ของเราแล้วจะดีที่สุด ที่ 30 ก้าวต่อวินาที RK4 พังจนพลังงานกลายเป็น $+34.47$ ในขณะที่วิธีออยเลอร์ซิมเพลกติกยังมีพลังงานเป็นลบ จึงยังโคจรอยู่ (ดูแบบฝึกหัด 13.1.7) ก้าวที่ $1/30$ วินาทีนี้ยาวกว่าเวลาที่ยานผ่านดาว ซึ่งราว $r/v \approx 0.1/\sqrt{19} \approx 0.023$ วินาที และช่วงนั้นคือช่วงที่ลูกศรเปลี่ยนเร็วที่สุด

ค่าพลังงานที่คำนวณจากค่าที่เก็บในแต่ละก้าวไม่ใช่พลังงานจริงของวิธี จึงอย่าใช้เกณฑ์นี้อย่างเดียวตัดสินวิธี ควรดูระยะห่างจากคำตอบจริงและดูการลดลงของความคลาดเคลื่อนเมื่อ h ลดลงด้วย

ฉบับร่าง · Preprint · 2026-05-23

สรุปใจความของหัวข้อ 13.1: กลับไปที่ยาน

- ออยเลอร์พุ่งออกนอกวงโคจรเพราะอ่านลูกศรเพียงจุดเดียวตอนที่ยานอยู่ใกล้ดาว ซึ่งลูกศรเปลี่ยนเร็ว
- ออยเลอร์ซิมเพลกติกยังอยู่ในวงโคจร แต่จุดไกลดาวที่สุดหมุนไปเรื่อย ๆ RK4 มีอันดับสูงกว่า แต่ที่ 60 ก้าวต่อวินาทีพลังงานลดลง
- RK4 แม่นกว่าเมื่อ h ถึง $1/480$ ซึ่งสั้นกว่าที่ 60 ก้าวต่อวินาทีถึงแปดเท่า ทางออกที่หัวข้อ 13.2 จะพิจารณาคือการเลือกก้าวให้เครื่องเอง

อ่านต่อ

ดู Iserles 2009, §3.2--3.3, pp. 38--41 สำหรับวิธีรุงเงอ-คุททาแบบชัดแจ้งและแบบโดยปริยาย และ Atkinson 1989, §6.10, pp. 418--432 สำหรับวิธีเดินทีละก้าวทั่วไป

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 13.1.1 (⋆, ทำด้วยมือ) ให้ $y' = -2y$ และ $y(0) = 1$ ใช้ขนาดขั้น $h = \frac{1}{4}$ คำนวณ y_1 ด้วยวิธีออยเลอร์และด้วยวิธีของฮอยน์ แล้วเทียบแต่ละค่ากับ $e^{-1/4}$ วิธีใดให้ค่าที่ใกล้กว่า

แบบฝึกหัด 13.1.2 (⋆, ทำด้วยมือ) ให้ $y' = y$ และ $y(0) = 1$ ใช้วิธี RK4 หนึ่งก้าวด้วย $h = \frac{1}{4}$ แสดงว่าค่าที่ได้เท่ากับพหุนามเทย์เลอร์ $1 + h + \frac{h^2}{2} + \frac{h^3}{6} + \frac{h^4}{24}$ แล้วหาเศษส่วนที่ตรงกับค่านั้น

แบบฝึกหัด 13.1.3 (⋆⋆, พิสูจน์) ให้วิธีรุงเงอ-คุททาชัดแจ้งที่มี $c_i = \sum_{j<i} a_{ij}$ พิจารณาปัญหา $y' = 1$, $y(0) = 0$ แสดงว่าหลังหนึ่งก้าว ความคลาดเคลื่อนเท่ากับ $h\big(1 - \sum_i b_i\big)$ แล้วสรุปว่าถ้า $\sum_i b_i \neq 1$ วิธีนี้ไม่มีอันดับหนึ่ง

แบบฝึกหัด 13.1.4 (⋆⋆, พิสูจน์) ให้วิธีรุงเงอ-คุททาชัดแจ้งเดียวกับข้อ 13.1.3 พิจารณาปัญหา $y' = t$, $y(0) = 0$ โดยเริ่มที่ $t_0 = 0$ แสดงว่าหลังหนึ่งก้าว ความคลาดเคลื่อนเท่ากับ $h^2\big(\frac{1}{2} - \sum_i b_i c_i\big)$ แล้วอธิบายว่าเหตุใดเงื่อนไข $\sum_i b_i c_i = \frac{1}{2}$ จึงจำเป็นต่ออันดับสอง

แบบฝึกหัด 13.1.5 (⋆⋆, คำนวณ) ขยายตาราง 13.1 ไปที่ $n = 128$ สำหรับ $y' = y$ บน $[0, 1]$ คำนวณความคลาดเคลื่อนของ RK4 และอัตราส่วนกับ $n = 64$ อัตราส่วนนั้นเข้าใกล้ 16 มากขึ้นหรือไม่ และเหตุใดทฤษฎีบทวิธีอันดับสี่จึงทำนายค่าประมาณนี้ได้

แบบฝึกหัด 13.1.6 (⋆⋆⋆, คำนวณและตีความ) จากตาราง 13.3 ที่ $h = 1/240$ และ $h = 1/480$ วิธีใดแม่นกว่าเมื่อ $t = 60$ อธิบายว่าเหตุใดอันดับสี่ของ RK4 ยังไม่เห็นผลที่ 60 ก้าวต่อวินาที

ตัวช่วยคิด

เทียบขนาดก้าวกับเวลาที่ยานผ่านดาว ซึ่งประมาณ $r/v \approx 0.1/\sqrt{19} \approx 0.023$ วินาที นับว่าในช่วงนั้นมีกี่ก้าวที่ $h = 1/60, 1/240$ และ $1/480$

ฉบับร่าง · Preprint · 2026-10-29

แบบฝึกหัด 13.1.7 (⋆ ⋆ ⋆, คำนวณและตีความ) สมมติว่าเกมเปลี่ยนเป็น 30 ก้าวต่อวินาที คือ $h = 1/30$ คำนวณระยะห่างจากตำแหน่งจริงและพลังงานที่ $t = 60$ สำหรับวิธีออยเลอร์ ออยเลอร์ซิมเพลกติก และ RK4 วิธีใดยังโคจรอยู่ วิธีใดพุ่งออกไป และผลนี้เกี่ยวกับข้อสรุปในหัวข้อ 13.1 อย่างไร

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 13.1.1.

1. ที่นี่ $f(t, y) = -2y$ และ $h = \frac{1}{4}$ เริ่มที่ $y_0 = 1$
2. ออยเลอร์: $y_1 = 1 + \frac{1}{4}(-2 \cdot 1) = \frac{1}{2} = 0.5$
3. ฮอยน์: $\mathbf{k}_1 = f(0, 1) = -2$ และ $\xi_2 = 1 + \frac{1}{4} \cdot (-2) = \frac{1}{2}$ จึง $\mathbf{k}_2 = f(\frac{1}{4}, \frac{1}{2}) = -1$ แล้ว $y_1 = 1 + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2) = 1 + \frac{1}{8}(-3) = \frac{5}{8} = 0.625$
4. คำตอบจริงคือ $e^{-1/4} \approx 0.778801$ ความคลาดเคลื่อนของออยเลอร์คือ 0.2788 และของฮอยน์คือ 0.1538 วิธีของฮอยน์ใกล้กว่า

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_131_heun_by_hand`)

เฉลย 13.1.2.

1. ที่นี่ $f(t, y) = y$ และ $h = \frac{1}{4}$ ดังนั้น $\mathbf{k}_1 = 1$, $\mathbf{k}_2 = 1 + \frac{1}{8} = \frac{9}{8}$, $\mathbf{k}_3 = 1 + \frac{1}{8} \cdot \frac{9}{8} = \frac{73}{64}$ และ $\mathbf{k}_4 = 1 + \frac{1}{4} \cdot \frac{73}{64} = \frac{329}{256}$
2. แทนลงในสูตร $y_1 = 1 + \frac{h}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4) = 1 + \frac{1}{24} \cdot \frac{1745}{256} = \frac{7889}{6144} \approx 1.2840169$
3. พหุนามเทย์เลอร์ที่ $h = \frac{1}{4}$ คือ $1 + \frac{1}{4} + \frac{1}{32} + \frac{1}{384} + \frac{1}{6144} = \frac{7889}{6144}$ ซึ่งตรงกับผลข้างบนพอดี

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_132_rk4_quarter_step`)

เฉลย 13.1.3.

1. สำหรับ $y' = 1$ ค่า f ไม่ขึ้นกับ y ดังนั้น $\mathbf{k}_i = 1$ ทุกชั้น
2. วิธีให้ $y_1 = 0 + h \sum_i b_i \cdot 1 = h \sum_i b_i$ แต่ผลเฉลยจริงคือ $y(h) = h$
3. ความคลาดเคลื่อนจึงเท่ากับ $h(1 - \sum_i b_i)$ ซึ่งเป็น $O(h^2)$ ได้ก็ต่อเมื่อ $\sum_i b_i = 1$ ดังนั้นถ้า $\sum_i b_i \neq 1$ วิธีนี้มีความคลาดเคลื่อนตัดปลายเฉพาะที่เป็น $O(h)$ เท่านั้น จึงไม่มีอันดับหนึ่ง

ตรวจด้วยโค้ด: ทำด้วยมือ และตรวจด้วยการสุ่มตัวอย่างที่ตั้งค่าเมล็ดสุ่มคงที่ `test_ch13.py` (`test_exercise_133_sum_b_is_necessary`)

สัมบัติ · Preprint · 2026-10-23

เฉลย 13.1.4.

1. สำหรับ $y' = t$ ค่า f ไม่ขึ้นกับ y จึง $\mathbf{k}_i = f(c_i h) = c_i h$
2. วิธีให้ $y_1 = h\sum_i b_i c_i h = h^2 \sum_i b_i c_i$ ส่วนผลเฉลยจริงคือ $y(h) = \frac{1}{2}h^2$
3. ความคลาดเคลื่อนคือ $\frac{1}{2}h^2 - h^2\sum_i b_i c_i = h^2\big(\frac{1}{2} - \sum_i b_i c_i\big)$
4. ความคลาดเคลื่อนนี้เป็น $O(h^3)$ ได้ก็ต่อเมื่อ $\sum_i b_i c_i = \frac{1}{2}$ ซึ่งคือเงื่อนไขอันดับสอง ไม่มีวิธีใดที่ได้อันดับสองโดยไม่มีเงื่อนไขนี้

ตรวจด้วยโค้ด: ทำด้วยมือ และตรวจด้วยการสุ่มตัวอย่างที่ตั้งค่าเมล็ดสุ่มคงที่ `test_ch13.py` (`test_exercise_134_sum_bc_is_necessary`)

เฉลย 13.1.5.

1. ที่ $n = 128$ ก้าวมีขนาด $h = 1/128$ คำตอบของ RK4 ที่ $t = 1$ ให้ความคลาดเคลื่อน 8.384093×10^{-11}
2. ที่ $n = 64$ ความคลาดเคลื่อนคือ 1.332722×10^{-9} อัตราส่วนคือ $1.332722 \times 10^{-9} / 8.384093 \times 10^{-11} \approx 15.896$ ซึ่งสูงกว่า 15.79 ในตาราง 13.1
3. อัตราส่วนเข้าใกล้ 16 เมื่อ h เล็กลง เพราะทฤษฎีบทวิธีอันดับสี่ให้ความคลาดเคลื่อนเป็น Ch^4 บวกพจน์ที่เล็กกว่าอัตราส่วนของ h สองค่าที่ต่างกันสองเท่าจึงเข้าหา $2^4 = 16$
4. เพื่อเทียบ วิธีออยเลอร์ที่ $n = 64$ ถึง 128 ได้อัตราส่วน 1.9859 และฮอยน์ได้ 3.9766 ซึ่งเข้าหา 2 และ 4 ตามลำดับ

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_135_model_order_at_128`)

เฉลย 13.1.6.

1. ที่ $h = 1/240$ ระยะห่างของออยเลอร์ซิมเพลกติกคือ 0.02742 ส่วน RK4 คือ 0.02856 จึงวิธีซิมเพลกติกแม่นกว่าเล็กน้อย
2. ที่ $h = 1/480$ ระยะห่างของออยเลอร์ซิมเพลกติกคือ 0.006653 ส่วน RK4 คือ 0.000914 จึง RK4 แม่นกว่า
3. ช่วงผ่านดาวใช้เวลาราว 0.023 วินาที ที่ $h = 1/60$ มีราว 1.4 ก้าวในช่วงนั้น ที่ $h = 1/240$ มีราว 5.5 ก้าว และที่ $h = 1/480$ มีราว 11 ก้าว (ตัวเลขเหล่านี้เป็นการประมาณ)
4. อันดับสี่บอกว่าความคลาดเคลื่อนลดเร็วเมื่อ h เล็กพอที่จะเห็นลูกศรเปลี่ยนในช่วงผ่านดาว ที่ 60 ก้าวต่อวินาที มีก้าวเพียงราว 1.4 ก้าวในช่วงนั้น ทำให้ผลของอันดับยังไม่ปรากฏ

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_136_semi_beats_rk4_at_240`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 13.1.7.

1. ที่ $h = 1/30$ เราคำนวณ $n = 1800$ ก้าว และเทียบกับตำแหน่งจริงที่ $t = 60$
2. ออยเลอร์: ระยะห่าง 135.45 และพลังงาน $+2.8245$ วงโคจรหลุดออกไป
3. ออยเลอร์ซิมเพลกติก: ระยะห่าง 7.87 และพลังงาน -0.1121 พลังงานยังเป็นลบ จึงยังโคจรอยู่ แต่ตำแหน่งห่างจากจริงมาก
4. RK4: ระยะห่าง 409.27 และพลังงาน $+34.4708$ วงโคจรพังจนหลุดออกไปไกลมาก
5. ที่ $h = 1/30$ ก้าวยาวกว่าเวลาที่ยานผ่านดาว (≈ 0.023 วินาที) ซึ่งเป็นช่วงที่ลูกศรเปลี่ยนเร็ว ในการคำนวณของเรา RK4 มีพลังงานเพิ่มขึ้นจนผิดปกติ ส่วนวิธีซิมเพลกติกยังคงอยู่ในวงโคจร เราไม่ได้วิเคราะห์สาเหตุเชิงทฤษฎีของการพังนี้ในบทนี้
6. ผลนี้ยืนยันคำเตือนในหัวข้อ 13.1 คืออันดับสูงไม่ได้ชนะเสมอ ในการทดลองนี้ ที่ขนาดก้าวใหญ่ วิธีที่มีอันดับต่ำกว่าแต่รักษาโครงสร้างดีกว่าให้ผลที่ดีกว่า

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_137_thirty_steps_per_second`)

13.2 ให้เครื่องเลือกขนาดก้าวเอง

สถานการณ์: ยานเดิม แต่ก้าวสั้นลงเมื่อเข้าใกล้ดาว

ยานลำเดิมจากหัวข้อ 13.1 แต่คราวนี้เกมไม่ใช้ก้าวคงที่ เกมเลือกก้าวสั้นเมื่อยานเข้าใกล้ดาว และยาวขึ้นเมื่อยานอยู่ไกล ในหัวข้อ 13.1 เราเห็นว่าก้าว $1/60$ วินาทีพลาดตอนที่ยานผ่านดาว ส่วนก้าว $1/480$ วินาทีแม่นกว่ามาก แต่ 60 วินาทีด้วยก้าวนั้นต้องคำนวณ $28\,800$ ก้าว ทั้งที่ยานส่วนใหญ่เคลื่อนช้า
คำถามของหัวข้อนี้คือ เครื่องจะรู้ได้อย่างไรว่าก้าวไหนสั้นพอ ทั้งที่ไม่รู้คำตอบจริง

ภาพรวมก่อนอ่าน

ในแต่ละก้าวเดินหนึ่งก้าวยาว h และสองก้าวครึ่งยาว $h/2$ จากจุดเดิม ผลต่างของสองค่านี้บอกความผิดพลาดของก้าวได้โดยไม่ต้องรู้คำตอบจริง ค่าที่ยอมรับได้เรียกว่า **ความคลาดเคลื่อนยินยอม** (tolerance)

เดินหนึ่งก้าวหรือสองครึ่งก้าว

ลองเริ่มจากจุดเดียวกันสองทาง ทางแรกเดินก้าวเดียวยาว h ได้ค่า $\mathbf{u}$ ทางที่สองเดินสองก้าวยาว $h/2$ ได้ค่า $\mathbf{v}$ หนังสือเรียกวิธีนี้ว่า ``one step, two half-steps" และถือว่าเป็นวิธีพื้นฐานของการควบคุมความคลาดเคลื่อน Iserles 2009, บทที่ 6, หมายเหตุ, p. 119

ลองเอง: ค่าประมาณบนกระดาษ

ให้ $y' = y$, $y(0) = 1$ ใช้วิธีจุดกึ่งกลาง (ดูหัวข้อ 13.1) กับ $h = \frac{1}{2}$ ก้าวเดียวได้ $\mathbf{u} = 1 + \frac{1}{2} \cdot \frac{5}{4} = \frac{13}{8}$ ส่วนสองครึ่งก้าวยาว $\frac{1}{4}$ ได้ $\mathbf{v} = \left(1 + \frac{1}{4} \cdot \frac{9}{8}\right)^2 = \left(\frac{41}{32}\right)^2 = \frac{1681}{1024} \approx 1.6416016$ ตัวประมาณที่นิยามต่อไปคือ $(\mathbf{v} - \mathbf{u})/3$ ซึ่งได้

$$\frac{\mathbf{v} - \mathbf{u}}{3} = \frac{17}{3072} \approx 0.0055339.$$

ฉบับร่าง · Preprint · 2026-10-23

ค่าคลาดเคลื่อนจริงของ $\mathbf{v}$ คือ $e^{1/2} - \mathbf{v} \approx 0.0071197$ ตัวประมาณนี้จึงต่ำกว่าจริงราว 22 เปอร์เซ็นต์ แต่ถ้าเอาไปบวกกับ $\mathbf{v}$ จะได้ $\mathbf{v} + \frac{\mathbf{v}-\mathbf{u}}{3} \approx 1.6471354$ ซึ่งใกล้ $e^{1/2} \approx 1.6487213$ กว่า $\mathbf{v}$ มาก (ตัวเลขทั้งหมดเป็นการคำนวณของเราเอง)

นิยาม 13.2.1: การประมาณค่าคลาดเคลื่อนแบบหนึ่งก้าวกับสองครึ่งก้าว

ให้ ψ เป็นวิธีหนึ่งก้าวอันดับ p และให้ $\mathbf{y}_n$ เป็นค่าที่จุด t_n กำหนด

$$\mathbf{u} = \psi(t_n, \mathbf{y}_n; h), \qquad \mathbf{v} = \psi\big(t_n + \tfrac{h}{2},\ \psi(t_n, \mathbf{y}_n; \tfrac{h}{2});\ \tfrac{h}{2}\big)$$

คือผลก้าวเดียวและผลสองครึ่งก้าว **ตัวประมาณค่าคลาดเคลื่อน** (error estimate) ของ $\mathbf{v}$ คือ

$$\mathbf{e}_h = \frac{\mathbf{v} - \mathbf{u}}{2^p - 1}.$$

อ่านแบบเข้าใจง่าย

ก้าวเดียวพลาดมากกว่าสองครึ่งก้าวราว 2^p เท่า ดังนั้นผลต่าง $\mathbf{v} - \mathbf{u}$ คือส่วนที่ $\mathbf{v}$ พลาด คูณด้วย $2^p - 1$ เราหารด้วย $2^p - 1$ ก็ได้ขนาดความพลาดของ $\mathbf{v}$ เอง โดยไม่ต้องรู้คำตอบจริง

ทฤษฎีบท 13.2.1: ผลประมาณค่าคลาดเคลื่อนของครึ่งก้าว

สมมติว่า ψ มีอันดับ p และผลเฉลย $\mathbf{y}$ หาอนุพันธ์ได้ต่อเนื่องพอ (ถึงอันดับ $p+2$) ค่าคลาดเคลื่อนเฉพาะที่ของก้าวเดียวมีรูป $\mathbf{y}(t_n + h) - \mathbf{u} = C(t_n, \mathbf{y}_n)h^{p+1} + O(h^{p+2})$ โดย C หาอนุพันธ์ได้ต่อเนื่องใกล้ผลเฉลย และ $\partial\psi/\partial\mathbf{y} = I + O(h)$ ใกล้ผลเฉลย (เป็นจริงสำหรับวิธีรุงเง-คุททาเมื่อ f หาอนุพันธ์ได้ต่อเนื่อง) ให้ h เล็กพอให้ทุกข้อนี้ใช้ได้ แล้ว

$$\mathbf{y}(t_n + h) - \mathbf{v} = \mathbf{e}_h + O(h^{p+2}).$$

พิสูจน์. **แนวคิด.** เทียบค่าคลาดเคลื่อนของก้าวเดียวกับของสองครึ่งก้าว แล้วกำจัด C ออกจากผลต่างของสองค่า

1. ก้าวเดียวมีค่าคลาดเคลื่อน $Ch^{p+1} + O(h^{p+2})$ ตามสมมติฐาน
2. ก้าวครึ่งแต่ละก้าวมีค่าคลาดเคลื่อนเฉพาะที่ $C(h/2)^{p+1}$ และก้าวครึ่งที่สองเริ่มจากค่าที่ติดค่าคลาดเคลื่อนของก้าวแรกมาด้วย ในขั้นนี้เราอ้างว่าค่าคลาดเคลื่อนส่งต่อด้วยตัวคูณ $\partial\psi/\partial\mathbf{y} = I + O(h)$ และค่า C ที่จุดกลางต่างจากจุดเริ่มเพียง $O(h)$ ซึ่งเราไม่พิสูจน์ที่นี่ ดังนั้น $\mathbf{y}(t_n + h) - \mathbf{v} = 2C(h/2)^{p+1} + O(h^{p+2}) = 2^{-p}Ch^{p+1} + O(h^{p+2})$
3. ลบสองสมการ: $\mathbf{v} - \mathbf{u} = (\mathbf{y} - \mathbf{u}) - (\mathbf{y} - \mathbf{v}) = (1 - 2^{-p})Ch^{p+1} + O(h^{p+2})$
4. ดังนั้น $Ch^{p+1} = \dfrac{\mathbf{v} - \mathbf{u}}{1 - 2^{-p}} + O(h^{p+2})$
5. แทนลงในข้อ 2: $\mathbf{y}(t_n + h) - \mathbf{v} = 2^{-p}\dfrac{\mathbf{v} - \mathbf{u}}{1 - 2^{-p}} + O(h^{p+2}) = \dfrac{\mathbf{v} - \mathbf{u}}{2^p - 1} + O(h^{p+2})$

สรุป: ผลต่างของสองค่าบอกค่าคลาดเคลื่อนของครึ่งก้าวได้ เพราะทั้งสองมีรูปกำลังของ h เหมือนกัน □

ฉบับร่าง · Preprint · 2026-10-23

เกร็ดเพิ่มเติม

หนังสือ ISE ระบุว่าวิธีหนึ่งก้าวกับสองครึ่งก้าวด้อยกว่าวิธีในหัวข้อนี้เมื่อใช้ควบคุมความคลาดเคลื่อนเพียงอย่างเดียว สำหรับวิธีหลายขั้น ISE ใช้ **กลวิธีมิลน์** (Milne device) ซึ่งเทียบวิธีอันดับเดียวกันสองวิธีแล้วใช้ค่าคงที่ความคลาดเคลื่อนของทั้งสองเป็นตัวคูณ Iserles 2009, §6.2, pp. 107--109 เกณฑ์ที่ใช้คู่กับกลวิธีนี้เป็นเกณฑ์ต่อหนึ่งหน่วยเวลา $\|\kappa\| \le h\delta$ ซึ่งเข้มขึ้นเมื่อก้าวเล็กลง

ค่าที่ยอมรับได้และการเลือกก้าวถัดไป

ต้องตกลงว่า "มากเกินไป" คืออะไร เราเลือกความคลาดเคลื่อนยินยอม (tolerance) $\delta > 0$ ก้าวนั้นยอมรับได้เมื่อ $E = \|\kappa\|_2$ ซึ่งเป็นขนาดของค่าประมาณ κ ไม่เกิน δ ถ้าเกิน เราปฏิเสธก้าวนั้น ถอยกลับไปจุดเดิม แล้วลองก้าวที่สั้นลง Iserles 2009, §6.2, p. 109 ค่า κ ที่ใช้ได้จะมาจากคู่ฝังตัวที่กล่าวต่อไป

นิยาม 13.2.2: กฎการเลือกขนาดก้าวถัดไป

ให้ก้าวยาว h มีค่าประมาณ κ (เวกเตอร์เมื่อระบบมีหลายตัวแปร) และค่าที่ยอมรับได้ δ ให้ $E = \|\kappa\|_2$ คือนอร์มยูคลิดของ κ (ค่าสัมบูรณ์เมื่อสเกลาร์) ก้าวถัดไปคือ

$$h_{\text{ใหม่}} = h \cdot \min\Big\{5,\ \max\Big\{0.2,\ 0.9\Big(\frac{\delta}{E}\Big)^{1/(p+1)}\Big\}\Big\}$$

เมื่อ $E > 0$ และ $h_{\text{ใหม่}} = 5h$ เมื่อ $E = 0$ (ตรงกับขีดจำกัดบน เพราะ δ/E ไม่มีที่สิ้นสุด) เมื่อ p คืออันดับของค่าที่ถูกประมาณ (ในคู่ฝังตัวของเราคือ $p = 2$) ตัวเลข 0.9 คือตัวสำรองความปลอดภัย ส่วน 0.2 และ 5 คือขีดจำกัดการเปลี่ยนแปลงในหนึ่งก้าว ซึ่งเป็นค่าที่เราเลือกเองในบทนี้

อ่านแบบเข้าใจง่าย

ถ้าค่าคลาดเคลื่อนเป็น Ch^{p+1} เมื่อเปลี่ยน h เป็น $h_{\text{ใหม่}}$ ค่าประมาณ E จะเป็น $E(h_{\text{ใหม่}}/h)^{p+1}$ เราอยากให้มันเท่ากับ δ พอดี สูตรนี้จึงให้ $h_{\text{ใหม่}}$ ที่ทำอย่างนั้น ส่วนตัวสำรอง 0.9 เผื่อความไม่แม่นของ κ เอง

ข้อควรระวัง: ค่าที่ยอมรับได้ต่อก้าวไม่ใช่ความคลาดเคลื่อนรวม

เกณฑ์ $E \le \delta$ ตรวจทีละก้าว ถ้าทุกก้าวที่ยอมรับมี E ไม่เกิน δ ค่าคลาดเคลื่อนรวมก็ยังสะสมได้มาก ดูตารางที่ 13.4 ที่ $\delta = 10^{-3}$ ระยะห่างหลัง 60 วินาทีเป็น 0.5533 ทั้งที่ทุกก้าวที่ยอมรับมี E ไม่เกิน 10^{-3} เกณฑ์ต่อหนึ่งหน่วยเวลาของ ISE คำนึงถึงการสะสมนี้บ้าง Iserles 2009, §6.2, p. 108

คู่ฝัง: ค่าประมาณเกือบไม่เสียเพิ่ม

การประมาณแบบหนึ่งก้าวกับสองครึ่งก้าวต้องเดินสามก้าวเพื่อได้ผลจริงหนึ่งก้าว ราคาจึงแพงเกินไปสำหรับใช้ทุกก้าว วิธีที่ใช้จริงคือ **คู่วิธีรุงเงอ-คุททาแบบฝังตัว** (embedded Runge--Kutta pair) ซึ่งรันวิธีสองวิธีด้วยชั้นร่วมกัน วิธีอันดับต่ำกว่าให้ค่า $\mathbf{y}_{n+1}$ ที่เราใช้ประมาณความคลาดเคลื่อน วิธีอันดับสูงกว่าให้ค่า $\mathbf{x}_{n+1}$ ที่เราเก็บไว้ เพราะแม่นกว่า Iserles 2009, §6.3, p. 116

ร่างเพื่อการอ่าน · Preprint · 2026-10-23

นิยาม 13.2.3: คู่วิธีรุงเงอ-คุททาแบบฝังตัวอันดับ 2 และ 3

ให้ชั้นสามชั้น

$$\mathbf{k}_1 = f(t_n, \mathbf{y}_n), \quad \mathbf{k}_2 = f\big(t_n + \tfrac{2}{3}h,\ \mathbf{y}_n + \tfrac{2}{3}h\mathbf{k}_1\big), \quad \mathbf{k}_3 = f\big(t_n + \tfrac{2}{3}h,\ \mathbf{y}_n + \tfrac{2}{3}h\mathbf{k}_2\big),$$

และตั้ง

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\big(\tfrac{1}{4}\mathbf{k}_1 + \tfrac{3}{4}\mathbf{k}_2\big), \qquad \mathbf{x}_{n+1} = \mathbf{y}_n + h\big(\tfrac{1}{4}\mathbf{k}_1 + \tfrac{3}{8}\mathbf{k}_2 + \tfrac{3}{8}\mathbf{k}_3\big).$$

ตารางบุชเชอร์ของคู่นี้คือ

$$\begin{array}{c|ccc} 0 & & & \\ \frac{2}{3} & \frac{2}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & \frac{3}{4} & 0 \\ & \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{array}$$

แถวล่างสองแถวคือน้ำหนักของ $\mathbf{y}_{n+1}$ และ $\mathbf{x}_{n+1}$ ตามลำดับ และ $\kappa = \mathbf{y}_{n+1} - \mathbf{x}_{n+1}$ คือค่าประมาณที่เราใช้

ทฤษฎีบท 13.2.2: อันดับและค่าประมาณของคู่ฝังตัว

สำหรับคู่ในนิยามคู่ฝังตัวอันดับ 2 และ 3 ได้ว่า

1. $\mathbf{y}_{n+1}$ มีอันดับ 2 และ $\mathbf{x}_{n+1}$ มีอันดับ 3;
2. $\kappa = \mathbf{y}_{n+1} - \mathbf{x}_{n+1} = h\frac{3}{8}(\mathbf{k}_2 - \mathbf{k}_3)$;
3. ถ้าผลเฉลยราบรื่นพอ แล้ว $\mathbf{y}(t_n + h) - \mathbf{y}_{n+1} = -\kappa + O(h^4)$ ดังนั้น $\|\kappa\|$ เป็นตัวประมาณค่าคลาดเคลื่อนเฉพาะที่ของ $\mathbf{y}_{n+1}$ ซึ่งเป็น $O(h^3)$

พิสูจน์. **แนวคิด.** ตรวจเงื่อนไขอันดับของแต่ละวิธี แล้วลบสองสูตรที่ใช้ชั้นเดียวกัน

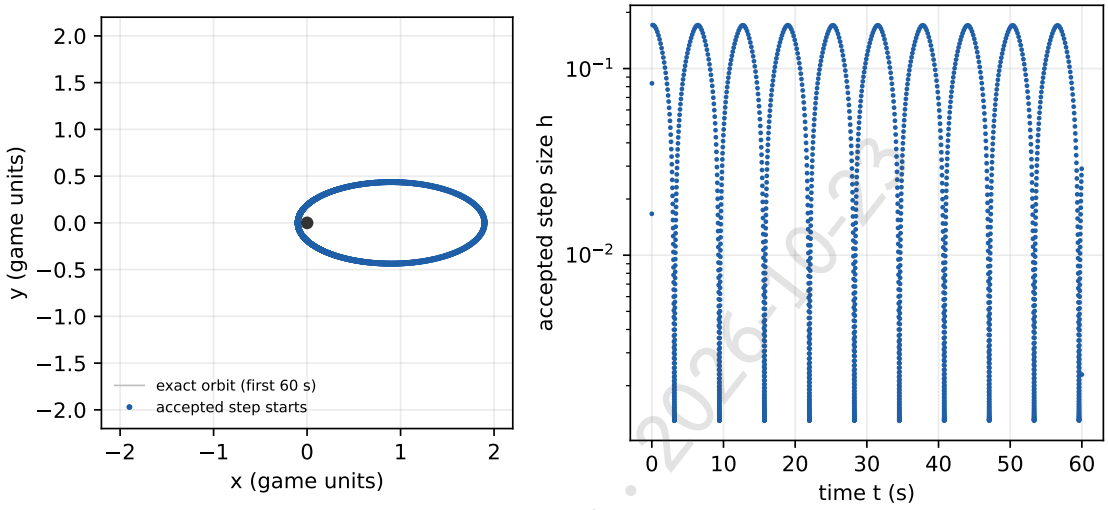
1. ทั้งสองวิธีมี $c_2 = a_{21} = \frac{2}{3}$ และ $c_3 = a_{31} + a_{32} = \frac{2}{3}$ ตามเงื่อนไขของทฤษฎีบทเงื่อนไขอันดับถึงอันดับสาม (หัวข้อ 13.1) สำหรับ $\mathbf{y}_{n+1}$ น้ำหนักให้ $\sum b_i = 1$ และ $\sum b_i c_i = \frac{1}{2}$ จึงได้อันดับ 2 แต่ $\sum b_i a_{ij} c_j = 0$ ไม่เท่ากับ $\frac{1}{6}$ จึงไม่ได้อันดับ 3 สำหรับ $\mathbf{x}_{n+1}$ น้ำหนักให้ $\sum b_i = 1$, $\sum b_i c_i = \frac{1}{2}$, $\sum b_i c_i^2 = \frac{1}{3}$ และ $\sum b_i a_{ij} c_j = \frac{1}{6}$ จึงได้อันดับ 3 ด้านพอของทฤษฎีบทนั้นเราอ้างจากหนังสือโดยไม่พิสูจน์ซ้ำ
2. ลบสองสูตร: $\mathbf{y}_{n+1} - \mathbf{x}_{n+1} = h\big[(\frac{3}{4} - \frac{3}{8})\mathbf{k}_2 - \frac{3}{8}\mathbf{k}_3\big] = h\frac{3}{8}(\mathbf{k}_2 - \mathbf{k}_3)$
3. เขียน $\mathbf{y}(t_n + h) - \mathbf{y}_{n+1} = \big(\mathbf{y}(t_n + h) - \mathbf{x}_{n+1}\big) + \big(\mathbf{x}_{n+1} - \mathbf{y}_{n+1}\big)$ ข้อแรกเป็น $O(h^4)$ เพราะ $\mathbf{x}_{n+1}$ มีอันดับ 3 ข้อหลังคือ $-\kappa$

สรุป: ผลต่างของสองวิธีที่ใช้ชั้นร่วมกันคือค่าประมาณที่ราคาเกือบไม่เพิ่ม และอันดับของแต่ละวิธีมาจากเงื่อนไขของหัวข้อ 13.1 □

ร่างปรับปรุง · Preprint · 2026-10-23

ตารางที่ 13.4: วงโคจร 60 วินาทีเมื่อเลือกก้าวเองด้วยคู่ฝังตัว เทียบกับ RK4 ที่ 60 ก้าวต่อวินาที (3600 ก้าว 14400 ชั้น ระยะห่าง 0.9083) (สร้างจาก `numanal/ch13.py` และตรวจด้วย `test_ch13.py`)

δ	ก้าวที่ยอมรับ	ก้าวที่ปฏิเสธ	h น้อยสุด	h มากสุด	ชั้น	ระยะห่างที่ $t = 60$
10^{-3}	811	269	0.002807	0.3684	3240	0.5533
10^{-4}	1648	0	0.001299	0.1710	4944	0.1298
10^{-5}	3552	0	0.000603	0.07939	10656	0.01304
10^{-6}	7656	0	0.000280	0.03685	22968	0.001297



รูปที่ 13.4: วงโคจรที่เลือกก้าวเองด้วย $\delta = 10^{-4}$ ใน 60 วินาทีแรก จุดแต่ละจุดคือจุดเริ่มของก้าวที่ยอมรับ จุดชิดกันมากใกล้ดาว ทางขวาคือขนาดก้าวที่ยอมรับเทียบกับเวลา แกนตั้งเป็นลอการิทึม (สร้างจาก `numanal/ch13.py`)

ข้อควรระวัง: ตัวคูณ h ที่หายไปในหนังสือ

ISE พิมพ์ค่าประมาณของคู่นี้โดยไม่มีตัวคูณ h Iserles 2009, §6.3, p. 116 แต่จากทฤษฎีบท $\mathbf{y}_{n+1} - \mathbf{x}_{n+1}$ มีตัวคูณ h อยู่ ถ้าละ h ไป ค่าประมาณจะผิดด้วยตัวคูณ h ในโปรแกรมของเราจึงเก็บ h ไว้ และการทดสอบยืนยันสูตรนี้

วงโคจรที่เลือกก้าวเอง

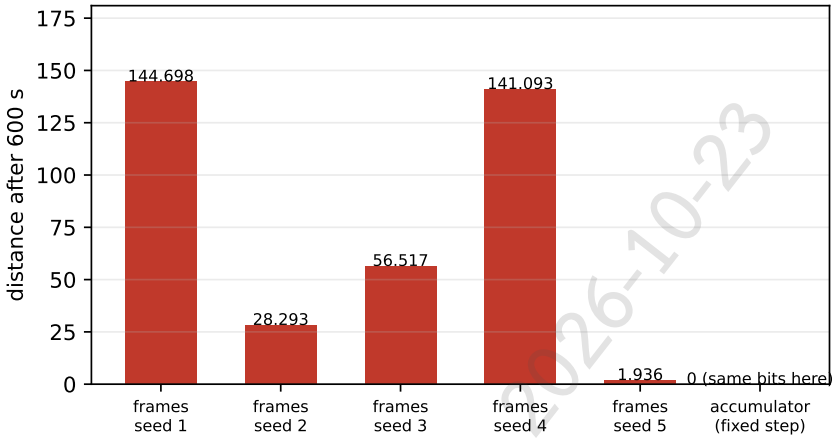
ตอนนี้เอาคู่ฝังตัวไปใช้กับยาน เริ่มที่จุดไกลดาวที่สุดเหมือนหัวข้อ 13.1 และก้าวเริ่มต้น $1/60$ วินาที ตารางที่ 13.4 สรุปผลสำหรับค่า δ สี่ค่า และภาพที่ 13.4 แสดงจุดเริ่มของทุกก้าวที่ยอมรับ

ที่ $\delta = 10^{-4}$ ก้าวที่สั้นที่สุดยาวราว 0.0013 วินาทีตอนยานผ่านดาว และยาวที่สุดราว 0.17 วินาทีตอนยานอยู่ไกล ต่างกันราว 130 เท่า วิธีนี้ยังไม่ได้เท่ากับก้าว $1/480$ ทั้งหมด แต่ใช้ชั้นน้อยกว่า RK4 ที่ 60 ก้าวต่อวินาที ($14\,400$ ชั้น) และให้ระยะห่างเล็กกว่าชัดเจน

ฉบับร่าง · Preprint · 2026-10-23

ตารางที่ 13.5: การวิ่ง 600 วินาทีเมื่อเฟรมยาว 40 ถึง 60 tick เทียบกับการวิ่งแบบคงที่ $1/60$ วินาที ระยะไกลสุดจากดาวของการวิ่งแบบคงที่คือ 1.902 (สร้างจาก `numanal/ch13.py` และตรวจด้วย `test_ch13.py`)

เมล็ด	จำนวนเฟรม	ระยะห่างจากการวิ่งแบบคงที่ที่ $t = 600$	ระยะไกลสุดจากดาว
1	36037	144.698	145.808
2	36016	28.293	26.704
3	36029	56.517	57.167
4	36015	141.093	142.172
5	36006	1.936	13.482



รูปที่ 13.5: ระยะห่างจากการวิ่งแบบคงที่หลัง 600 วินาที สำหรับเฟรมแบบสุ่มห้าเมล็ด และตัวสะสมเวลาหนึ่งแท่ง (แท่งสุดท้าย) ซึ่งในการทดสอบของเราให้ผลเหมือนกันทุกบิต บนโปรแกรมและเครื่องเดียวกัน (สร้างจาก `numanal/ch13.py`)

สำนักงาน preprint · 2026-10-23

อีกด้านหนึ่งของปัญหาเดียวกัน: รันซ้ำต้องได้ผลเหมือนกัน

สถานการณ์สมมติ: ผู้เล่นสองคนบนเครื่องต่างกัน

เกมออนไลน์ที่ผู้เล่นทุกคนเห็นยานเดียวกัน ต้องได้ผลเหมือนกันทุกเครื่อง และการเล่นซ้ำของเกมเดิมก็ต้องเหมือนกันด้วย ลองนึกถึงเครื่องหนึ่งที่ทำงานได้เฟรมละราว $1/50$ วินาที และอีกเครื่องราว $1/75$ วินาที เวลาของเฟรมเปลี่ยนไปตามภาระของเครื่อง เมื่อครบ 10 นาทีของเกม ยานของทั้งสองเครื่องจะยังอยู่ที่เดิมหรือไม่

บทความ Fiedler 2004a เริ่มจากประโยคที่ตรงกับปัญหานี้ คือ ``the behavior of your physics simulation depends on the delta time you pass in'' หมายความว่าผลของการจำลองฟิสิกส์ขึ้นกับช่วงเวลาที่ส่งเข้าไปในแต่ละเฟรม บทความแนะนำวิธีที่ง่ายที่สุดคือใช้ช่วงเวลาคงที่ เช่น $1/60$ วินาที และอธิบายว่าเวลาที่เหลือจากเฟรมหนึ่งถูกส่งต่อไปยังเฟรมถัดไปผ่าน **ตัวสะสมเวลา** (accumulator) ซึ่งเป็นตัวนับที่เก็บเวลาที่ยังไม่ได้ใช้ไว้ให้เฟรมถัดไป เวลานั้นไม่ถูกทิ้ง

เราทดสอบด้วยยานวงโคจรเดิม โดยใช้หน่วย tick ซึ่งเท่ากับ $1/3000$ วินาที เพื่อให้นับเวลาเป็นจำนวนเต็ม การวิ่งแบบคงที่มี $36\,000$ เฟรม เท่ากับ $1/60$ วินาทีทุกเฟรม ส่วนเมล็ดสุ่มทั้งห้าให้เฟรมยาว 40 ถึง 60 tick (ราว $1/75$ ถึง $1/50$ วินาที) ตารางที่ 13.5 แสดงผลหลังครบ 600 วินาที

ผลของทั้งห้าเมล็ดไม่เหมือนการวิ่งแบบคงที่เลย ระยะไกลสุดจากดาวอยู่ที่ 13 ถึง 146 หน่วย ทั้งที่การวิ่งแบบคงที่ไม่เคยไปไกลเกิน 1.9 หน่วย ความแปรผันที่แคบกว่ามากก็ทำให้ผลต่างออกไปแล้ว เช่น เมื่อเฟรมอยู่ระหว่าง 50 ถึง 51 tick (ประมาณ 2 เปอร์เซ็นต์) เมล็ด 1 ถึง 3 มีระยะไกลสุดราว 2.07 ถึง 2.30 หน่วย (ตัวเลขทั้งหมดเป็นการคำนวณของเราเอง ไม่ได้บอกว่าเกมใดทำเช่นนี้)

วิธีแรก: ก้าวคงที่กับตัวสะสมเวลา. ให้ก้าวคงที่ $1/60$ วินาทีเสมอ และเก็บเวลาที่เหลือในตัวสะสมเวลา ทั้งห้าเมล็ดได้ $36\,000$ ก้าวเท่ากันและเหลือเวลา 0 tick ในการทดสอบของเรา ผลเหมือนกันทุกบิตเมื่อรันในโปรแกรมเดียวกันบนเครื่องเดียวกัน การทดสอบนี้เปลี่ยนเฉพาะตารางเฟรม ไม่ได้ทดสอบข้ามแพลตฟอร์ม คอมไพเลอร์ หรือการคำนวณทศนิยมแบบอื่น ก้าวคงที่อย่างเดียวจึงไม่รับประกันความเหมือนกันทุกบิตข้ามเครื่อง

ก้าวย่อยแบบปรับตัวในเฟรมคงที่. ทางประนีประนอมคือยังใช้ตัวสะสมนับเฟรมคงที่ $1/60$ วินาที แต่ภายในแต่ละเฟรมเดินด้วยคู่ฝังตัวเป็น **ก้าวย่อย** (substep) ซึ่งคือก้าวที่สั้นกว่าเฟรมหนึ่งและเดินหลายครั้งในเฟรมเดียว โดยปรับขนาดเอง ก้าวย่อยเหล่านี้ขึ้นกับสถานะตอนเริ่มเฟรมและขนาดก้าวที่ค้างไว้ เท่านั้น ไม่ขึ้นกับความยาวจริงของเฟรม ด้วย $\delta = 10^{-6}$ ใช้ $92\,003$ ก้าวย่อย หลัง 600 วินาทีห่างจากคำตอบจริง 0.2767 หน่วย ส่วนการวิ่งแบบคงที่ 60 ก้าวต่อวินาทีห่าง 1.9800 หน่วย ในการทดสอบของเรา เมล็ด 1 และ 2 ให้บิตเดียวกันทุกประการ (โปรแกรมและเครื่องเดียวกัน)

ข้อควรระวัง: อย่าให้เวลาของเฟรมเป็นตัวกำหนดก้าว

ปัญหาไม่ได้มาจากการปัดเศษอย่างเดียว ผลของวิธีขึ้นกับว่าเราแบ่งเวลาเป็นก้าวอย่างไร แม้แต่ละก้าวจะแม่น ก้าวที่มาจากเวลาของเฟรมก็ทำให้การแบ่งเวลาต่างกันในแต่ละเครื่อง และการเลือกก้าวแบบปรับตัวที่ใช้เวลาของเฟรมเป็นตัวตั้งต้นก็รับปัญหานี้ไปด้วย ตัวสะสมเวลาหรือก้าวย่อยในเฟรมคงที่ช่วยได้เพราะทำให้การแบ่งเวลาไม่ขึ้นกับเฟรมจริง (ในการทดสอบของเรา)

สรุปใจความของหัวข้อ 13.2: กลับไปที่ยาน

- เราประมาณค่าคลาดเคลื่อนของก้าวได้ด้วยการเดินหนึ่งก้าวกับสองครึ่งก้าว หรือด้วยคู่ฝังตัวที่ราคาเกือบไม่เพิ่มค่าประมาณ κ ใช้ตัดสินว่าจะยอมรับหรือปฏิเสธก้าว
- เกณฑ์ต่อก้าวไม่ใช่การรับประกันความคลาดเคลื่อนรวม
- การรันซ้ำให้ผลเดียวกันในโปรแกรมและเครื่องเดียวกันต้องไม่ให้เวลาของเฟรมกำหนดก้าว ตัวสะสมเวลากับก้าวย่อยในเฟรมคงที่ทำได้ (ข้ามเครื่องยังไม่ได้ทดสอบ)

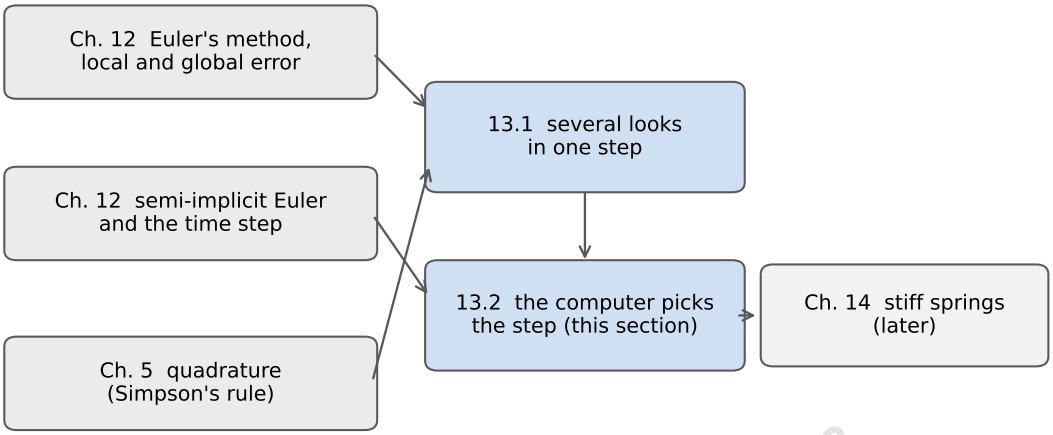
อ่านต่อ. บทที่ 5 ของ Iserles 2009, บทที่ 5, p. 73 ว่าด้วยการหาปริพันธ์เชิงตัวเลขเชิงเรขาคณิต ซึ่งตั้งใจรักษาสมบัติบางอย่างของผลเฉลยจริงไว้ ส่วนการควบคุมค่าคลาดเคลื่อนของวิธีหลายขั้นอยู่ใน Iserles 2009, §6.2, pp. 107--112

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 13.2.1 (★, ทำด้วยมือ) ให้ $y' = y$ และ $y(0) = 1$ ใช้วิธีจุดกึ่งกลางกับ $h = \frac{1}{2}$ คำนวณผลก้าวเดียว $\mathbf{u}$ และผลสองครึ่งก้าว $\mathbf{v}$ (แต่ละครึ่งก้าวยาว $\frac{1}{4}$) แล้วหาตัวประมาณค่าคลาดเคลื่อน $(\mathbf{v} - \mathbf{u})/3$ เทียบกับค่าคลาดเคลื่อนจริง $e^{1/2} - \mathbf{v}$

แบบฝึกหัด 13.2.2 (★★, คำนวณ) ทำซ้ำข้อ 13.2.1 กับ $h = \frac{1}{4}$ และ $h = \frac{1}{8}$ หาอัตราส่วนของตัวประมาณค่าคลาดเคลื่อนต่อค่าคลาดเคลื่อนจริงของ $\mathbf{v}$ ทั้งสองค่า อัตราส่วนเข้าใกล้ 1 เมื่อ h เล็กลงหรือไม่ และส่วนที่ยังห่างจาก 1 ลดลงเร็วเท่าใดเมื่อ h ลดครึ่งหนึ่ง

ฉบับร่าง · preprint · 2026-10-23



รูปที่ 13.6: แผนผังบทที่ 13 ทั้งบท หัวข้อ 13.1 และ 13.2 และสิ่งที่บทก่อนให้ วิธีออยเลอร์และวิธีออยเลอร์ซิมเพลกติกในบทที่ 12 รวมถึงกฎซิมป์สันในบทที่ 5 เป็นฐาน บทที่ 14 ต้องการก้าวที่เข้มขึ้น (สร้างจาก `numanal/ch13.py`)

แบบฝึกหัด 13.2.3 (⋆⋆, พิสูจน์) ให้คู่ฝังตัวในหัวข้อนี้ ลองพิสูจน์ว่า $\kappa = \mathbf{y}_{n+1} - \mathbf{x}_{n+1} = h\frac{3}{8}(\mathbf{k}_2 - \mathbf{k}_3)$ แล้วใช้ผลนี้กับ $y' = -y$, $y(0) = 2$, $h = \frac{1}{4}$ หาค่า κ

แบบฝึกหัด 13.2.4 (⋆, ทำด้วยมือ) ก้าวยาว $h = 0.2$ ให้ตัวประมาณค่าคลาดเคลื่อน $\kappa = 4\times10^{-4}$ และเรายอมรับความคลาดเคลื่อนได้ $\delta = 10^{-4}$ ก้าวนี้ถูกยอมรับหรือถูกปฏิเสธ และถ้าถูกปฏิเสธ ก้าวที่ลองใหม่ยาวเท่าใด ถ้า $\kappa = 0.5 \times 10^{-4}$ แทนล่ะ ก้าวถัดไปยาวเท่าใด

แบบฝึกหัด 13.2.5 (⋆⋆⋆, คำนวณและตีความ) จากตาราง 13.4 เมื่อ δ ลดจาก 10^{-5} เป็น 10^{-6} ระยะห่างที่ $t = 60$ ลดลงกี่เท่า และจำนวนขั้นเพิ่มขึ้นกี่เท่า ต่อมาเทียบกับ RK4 ที่ $h = 1/480$ ในตาราง 13.3 ซึ่งใช้ $28\,800$ ก้าว วิธีใดใช้ขั้นน้อยกว่า และวิธีใดแม่นกว่า

แบบฝึกหัด 13.2.6 (⋆⋆⋆, โครงการเล็ก)

ตัวช่วยคิด

นึกถึงช่วงเฟรมที่ไม่เท่ากันว่าส่งผลต่อก้าวของแต่ละเครื่องอย่างไร แล้วถามว่าการนับก้าวจากเวลาที่สะสมจะทำให้ทุกเครื่องทำจำนวนก้าวเท่ากันได้หรือไม่

โครงการนี้กลับไปที่เกมที่ทุกเครื่องต้องได้ผลเดียวกัน ให้หน่วยเวลา 1 tick เท่ากับ $1/3000$ วินาที ใช้ช่วงเฟรมสุ่มจาก 49 ถึง 51 tick (ราว $\pm4\%$ จาก $1/60$ วินาที) และใช้เมล็ดสุ่ม 1 ถึง 5 ทำดังนี้ (ก) หาระยะที่ไกลที่สุดจากดาวในสิบนาทีของแต่ละเมล็ด (ข) หาระยะห่างจากการวิ่งแบบคงที่ที่ $t = 600$ วินาที (ค) เขียนย่อหน้าสั้น ๆ ว่าเกมควรเลือกการคำนวณแบบใด เพื่อให้ผลเดียวกันทุกเครื่อง โดยอ้างถึงตัวสะสมเวลาและก้าวย่อยที่กล่าวในหัวข้อนี้

ร่างพิมพ์ · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 13.2.1.

1. ผลก้าวเดียวของจุดกึ่งกลาง: $\mathbf{k}_1 = f(0,1) = 1$ และ $\mathbf{k}_2 = f(\tfrac{1}{4},\, 1 + \tfrac{1}{4}\cdot 1) = \tfrac{5}{4}$ จึงได้ $\mathbf{u} = 1 + \tfrac{1}{2}\cdot\tfrac{5}{4} = \tfrac{13}{8} = 1.625$
2. ครึ่งก้าวแรกยาว $\tfrac{1}{4}$ จาก $(0,1)$: $\mathbf{k}_2 = f(\tfrac{1}{8},\, 1+\tfrac{1}{8}) = \tfrac{9}{8}$ จึงได้ $1 + \tfrac{1}{4}\cdot\tfrac{9}{8} = \tfrac{41}{32}$
3. ครึ่งก้าวที่สองจาก $\tfrac{41}{32}$: $\mathbf{k}_2 = \tfrac{41}{32}\cdot\tfrac{9}{8} = \tfrac{369}{256}$ จึงได้ $\mathbf{v} = \tfrac{41}{32} + \tfrac{1}{4}\cdot\tfrac{369}{256} = \tfrac{1681}{1024} \approx 1.6416016$
4. ตัวประมาณค่าคลาดเคลื่อน $\mathbf{e}_h = \dfrac{\mathbf{v}-\mathbf{u}}{2^2-1} = \dfrac{1}{3}\cdot\dfrac{17}{1024} = \dfrac{17}{3072} \approx 0.0055339$
5. ค่าคลาดเคลื่อนจริงของ $\mathbf{v}$ คือ $e^{1/2} - \mathbf{v} \approx 1.6487213 - 1.6416016 = 0.0071197$ ตัวประมาณต่ำกว่าจริงราว 22 เปอร์เซ็นต์ ซึ่งเป็นผลของการที่ $h = \tfrac{1}{2}$ ยังไม่เล็กพอ

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_1321_midpoint_by_hand`)

เฉลย 13.2.2.

1. ที่ $h = \tfrac{1}{4}$: ผลก้าวเดียว $\mathbf{u} = 1 + \tfrac{1}{4} + \tfrac{1}{2}(\tfrac{1}{4})^2 = \tfrac{41}{32} = 1.28125$ ผลสองครึ่งก้าวยาว $\tfrac{1}{8}$ คือ $\mathbf{v} = \big(1 + \tfrac{1}{8} + \tfrac{1}{2}(\tfrac{1}{8})^2\big)^2 = \big(\tfrac{145}{128}\big)^2 \approx 1.2832642$ ตัวประมาณคือ $(\mathbf{v}-\mathbf{u})/3 \approx 0.00067139$ และค่าจริง $e^{1/4} - \mathbf{v} \approx 0.00076126$ อัตราส่วนคือ 0.8819
2. ที่ $h = \tfrac{1}{8}$: $\mathbf{u} = \tfrac{145}{128} \approx 1.1328125$ และ $\mathbf{v} = \big(\tfrac{545}{512}\big)^2 \approx 1.1330605$ ตัวประมาณคือ 0.00008265 และค่าจริง $e^{1/8} - \mathbf{v} \approx 0.00008800$ อัตราส่วนคือ 0.9392
3. อัตราส่วนเข้าใกล้ 1 ส่วนที่ห่างจาก 1 คือ 0.2227, 0.1181 และ 0.0608 เมื่อ $h = \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}$ ลดลงราวครึ่งทุกครั้งที่ h ลดครึ่ง เพราะส่วนเกินเป็น $O(h)$ เทียบกับค่าหลัก

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_1322_ratio_towards_one`)

เฉลย 13.2.3.

1. อันดับสองคือ $\mathbf{y}_{n+1} = \mathbf{y}_n + h\big(\tfrac{1}{4}\mathbf{k}_1 + \tfrac{3}{4}\mathbf{k}_2\big)$ อันดับสามคือ $\mathbf{x}_{n+1}$ ที่มี $\tfrac{3}{8}\mathbf{k}_2 + \tfrac{3}{8}\mathbf{k}_3$ แทน
2. ลบกัน: $\mathbf{y}_{n+1} - \mathbf{x}_{n+1} = h\big[(\tfrac{3}{4} - \tfrac{3}{8})\mathbf{k}_2 - \tfrac{3}{8}\mathbf{k}_3\big] = h\tfrac{3}{8}(\mathbf{k}_2 - \mathbf{k}_3)$ ซึ่งเป็นสิ่งที่ต้องการ
3. สำหรับ $y' = -y$: $\mathbf{k}_1 = -2$, $\mathbf{k}_2 = -(2 + \tfrac{1}{6}\cdot(-2)) = -\tfrac{5}{3}$ และ $\mathbf{k}_3 = -(2 + \tfrac{1}{6}\cdot(-\tfrac{5}{3})) = -\tfrac{31}{18}$
4. $\mathbf{k}_2 - \mathbf{k}_3 = -\tfrac{5}{3} + \tfrac{31}{18} = \tfrac{1}{18}$ จึงได้ $\kappa = \tfrac{1}{4}\cdot\tfrac{3}{8}\cdot\tfrac{1}{18} = \tfrac{1}{192} \approx 0.0052083$

ตรวจด้วยโค้ด: ทำด้วยมือ และตรวจการเท่ากันของผลต่างด้วยการสุ่มตัวอย่างที่ตั้งค่าเมล็ดสุ่มคงที่ `test_ch13.py` (`test_exercise_1323_kappa_identity_on_a_linear_problem`)

เฉลย 13.2.4.

1. เมื่อ $\kappa = 4\times 10^{-4}$ และ $\delta = 10^{-4}$ จะได้ $\kappa > \delta$ ก้าวนี้ถูกปฏิเสธ
2. ก้าวที่ลองใหม่คือ $0.9\cdot 0.2\cdot\big(\tfrac{\delta}{\kappa}\big)^{1/3} = 0.18\cdot\big(\tfrac{1}{4}\big)^{1/3} \approx 0.1134$
3. เมื่อ $\kappa = 0.5\times 10^{-4}$ ก้าวนี้ยอมรับได้ เพราะ $\kappa < \delta$ ก้าวถัดไปยาวขึ้นเป็น $0.18\cdot 2^{1/3} \approx 0.2268$

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_1324_step_rule`)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 13.2.5.

1. ที่ $\delta = 10^{-5}$ ระยะห่างคือ 0.013044 ใช้ $10\,656$ ชั้น ที่ $\delta = 10^{-6}$ ระยะห่างคือ 0.001297 ใช้ $22\,968$ ชั้น
2. ระยะห่างลดลงราว $0.013044/0.001297 \approx 10.1$ เท่า ส่วนจำนวนชั้นเพิ่มขึ้นราว $22\,968/10\,656 \approx 2.16$ เท่า ในการทดลองนี้ระยะห่างลดตามสัดส่วนของ δ
3. RK4 ที่ $h = 1/480$ ใช้ $28\,800 \times 4 = 115\,200$ ชั้น และได้ระยะห่าง 0.000914 เทียบกับ $\delta = 10^{-6}$ ที่ใช้ $22\,968$ ชั้น ซึ่งน้อยกว่าราว 5 เท่า ระยะห่างที่ได้ 0.001297 มากกว่า RK4 ราว 1.4 เท่า
4. วิธีปรับขนาดก้าวประหยัดชั้น แต่เกณฑ์ต่อก้าวไม่รับประกันความคลาดเคลื่อนรวมโดยตรง

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_1325_orbit_table_at_two_tolerances`)

เฉลย 13.2.6.

1. ระยะห่างจากดาวที่ไกลที่สุดในสิบนาทีของเมล็ด 1 ถึง 5 คือ 5.405, 2.008, 3.657, 3.028 และ 4.177 หน่วย ส่วนการวิ่งแบบคงที่ได้ 1.902 หน่วย
2. ระยะห่างจากการวิ่งแบบคงที่ที่ $t = 600$ วินาทีคือ 5.646, 2.969, 4.468, 3.879 และ 4.469 หน่วย
3. เมื่อใช้ตัวสะสมเวลา เกมนับก้าวคงที่ $1/60$ วินาทีจากเวลาที่สะสมได้ ทุกเครื่องจึงทำ $36\,000$ ก้าวเท่ากันทุกประการ และในการทดสอบของเรา (โปรแกรมและเครื่องเดียวกัน) ผลเหมือนกันทุกบิต โดยที่ก้าวไม่ได้ยาวเท่ากับเวลาของเฟรม
4. หากต้องการก้าวปรับตัว ให้ปรับตัวภายในเฟรม $1/60$ วินาทีคงที่ โดยก้าวย่อยขึ้นกับสถานะตอนเริ่มเฟรมเท่านั้น

ตรวจด้วยโค้ด: `test_ch13.py` (`test_exercise_1326_jitter_returns_to_the_frame_story`)

สรุปบทที่ 13 ในหน้าเดียว

ตารางเทียบวิธีหลักของบท โดย h คือขนาดก้าว

วิธี	ค่าที่ใช้ต่อก้าว	อันดับ	ตัวประมาณค่าคลาดเคลื่อน	สิ่งที่ต้องระวัง
ออยเลอร์ (12.1)	f_k หนึ่งชั้น	1	ไม่มี	อ่านลูกศรจุดเดียว
จุดกึ่งกลางและฮอยน์ (13.1)	2 ชั้น	2	ไม่มี	$\sum b_i c_i = \frac{1}{2}$ จำเป็น (ภายใต้ข้อตกลงแถว)
RK4 (13.1)	4 ชั้น	4	ไม่มี	ก้าวใหญ่เกินไปพังได้
หนึ่ง ก้าว กับ สอง ครึ่ง ก้าว (13.2)	ก้าว h และ $2 \times h/2$	p	$\mathbf{e}_h$	เดินสามก้าวต่อหนึ่งก้าวจริง
คู่ฝังตัว 2 และ 3 (13.2)	3 ชั้น	2 (ส่งออก 3)	κ	เกณฑ์ ต่อ ก้าว ไม่ใช่ ความคลาดเคลื่อนรวม

สูตรที่ควรจำ: กฎก้าวคือ $h_{\text{ใหม่}} = h \cdot \min\{5, \max\{0.2,\, 0.9(\delta/E)^{1/(p+1)}\}\}$ เมื่อ $E = \|\kappa\|_2$ และ $E = 0$ ให้ $5h$ $\mathbf{e}_h = \dfrac{\mathbf{v} - \mathbf{u}}{2^p - 1}$ และสำหรับคู่ 2 และ 3 $\kappa = h\frac{3}{8}(\mathbf{k}_2 - \mathbf{k}_3)$

อ่านต่อ. บทที่ 14 พบสมการที่แข็งมาก ซึ่งต้องใช้วิธีโดยนัย ดูการควบคุมค่าคลาดเคลื่อนได้ใน Iserles 2009, §6.1--6.3, pp. 105--118

ฉบร่าง · Preprint · 2026-10-23

บทที่ 14

เมื่อสปริงแข็งเกินไป

สถานการณ์สมมติ: หุ่นที่สั่นเพราะข้อต่อแข็งเกินไป

ลองนึกถึงตัวละครในเกมที่เป็นหุ่นประกอบจากท่อนร่างต่อกันด้วยข้อต่อ เรียกว่า **หุ่นแร็กดอลล์** (ragdoll) คือหุ่นที่ปล่อยให้ฟิสิกส์ควบคุมการล้มและการกระดอนของทุกท่อน เมื่อหุ่นล้มลง ข้อต่อแต่ละข้อทำตัวเหมือนสปริง ถ้าสปริงนุ่ม หุ่นก็ขยับอย่างเป็นธรรมชาติ แต่ถ้าเราตั้งให้ข้อต่อแข็งมาก หุ่นอาจสั่นหรือกระเด็นออกไป ทั้งที่สมการที่เราเขียนลงไปนั้นถูกต้อง

สมมติว่าเกมคำนวณ 60 ครั้งต่อวินาที (ก้าวละ $1/60$ วินาที) และแขนของหุ่นมีสปริงที่แข็งมาก $k = 10^6$ นิวตันต่อเมตร โดยมวลของแขนเป็น 1 กิโลกรัม ตัวเลขจากโค้ดของเราบอกว่าแขนจะกระเด็นออกนอกภาพภายในสี่ก้าว แต่ถ้าแบ่งแต่ละเฟรมเป็นสิบ**ก้าวย่อย** (sub-step) คือคำนวณก้าวเล็ก ๆ หลายก้าวในเฟรมเดียว แขนจะแกว่งนิ่ง ๆ ตามที่ควรเป็น (ตัวเลขทั้งหมดในบทนี้มาจากโค้ดของเรา และไม่ได้บอกว่าเกมใดคำนวณแบบนี้)

คำถามของบทนี้จึงเป็นว่า ทำไมก้าวต้องเล็กจนทำอะไรไม่ได้ ทั้งที่เราไม่ได้ต้องการความแม่นยำมากเลย และทำไมผลเฉลยจริงที่นิ่งและเรียบ จึงยังทำให้วิธีคำนวณระเบิดได้

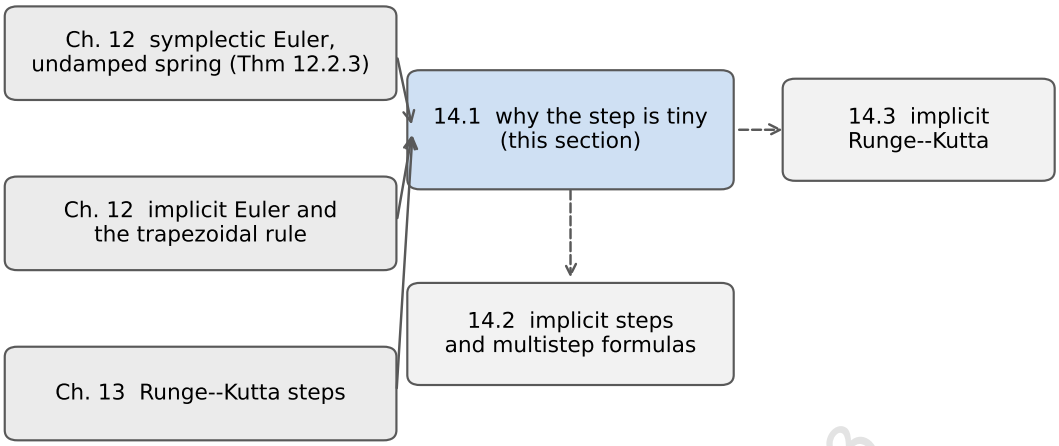
ภาพรวมบท

บทที่ 12 และ 13 ทำให้เราเห็นว่าก้าวที่ใหญ่เกินไปทำให้ความผิดพลาดโตขึ้น บทนี้ต่อไปอีกขั้น ความผิดพลาดอาจไม่ได้โตเพราะวิธีแม่นน้อย แต่โตเพราะ**ความไม่เสถียร** ซึ่งเป็นอีกเรื่องหนึ่ง บทนี้มีสามหัวข้อ

- หัวข้อ 14.1 ทำไมก้าวต้องเล็กจนทำอะไรไม่ได้ เราดูสปริงแข็ง โดเมนเสถียรภาพ และคำว่า A-เสถียร
- หัวข้อ 14.2 ก้าวแบบโดยปริยายและ**สูตร BDF** (backward differentiation formula) ซึ่งเป็นวิธีหลายขั้นที่ใช้ค่าเก่าหลายค่า เราดูวิธีที่ไม่ระเบิดแม้ก้าวใหญ่ และราคาที่ต้องจ่ายคือการแก้สมการทุกก้าวกับพลังงานที่หายไป
- หัวข้อ 14.3 **วิธีรุงเงอ-คุททาแบบโดยปริยาย** (implicit Runge–Kutta, IRK) คือวิธีที่ค่าแต่ละชั้นพึ่งพากันได้ เราดูวิธีที่ไม่ระเบิดกับระบบแข็ง

เริ่มจากหุ่นที่สั่นในหัวข้อนี้ แล้วจะเจออีกด้านหนึ่งของปัญหาเดียวกันในหัวข้อ 14.2 คือมือที่กดลงบนกำแพงแข็ง ซึ่งเป็นสปริงแข็งอีกแบบหนึ่ง

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 14.1: แผนผังบทที่ 14 กล่องสีฟ้าคือหัวข้อ 14.1 ลูกศรทึบคือแนวคิดที่หัวข้อหนึ่งใช้ ลูกศรประคือแนวคิดที่ต่อจากหัวข้อก่อนหน้า (สร้างจาก `numanal/ch14.py`)

14.1 ทำไมก้าวต้องเล็กจนทำอะไรไม่ได้

สถานการณ์สมมติ: ข้อต่อหนึ่งข้อของหุ่นบนจอ

เริ่มจากข้อต่อหนึ่งข้อของหุ่นบนจอ ซึ่งเราจำลองเป็นสปริงที่ง่ายที่สุด คือมวล 1 กิโลกรัมติดอยู่กับสปริงที่มีค่าความแข็ง k นิวตันต่อเมตร โดยไม่มีแรงหน่วง (แบบจำลองนี้เป็นของเรา ไม่ได้บอกว่าเอนจินใดใช้) ความถี่เชิงมุมของการแกว่งคือ $\omega = \sqrt{k}$ ถ้า $k = 10^2$, 10^4 และ 10^6 ความถี่จะเป็น 10, 100 และ 1000 เรเดียนต่อวินาที

เมื่อเราเดินทีละก้าว **ขนาดก้าววิกฤต** (ค่าที่แยกระหว่างรอดกับไม่รอด) คือเท่าไร ตารางด้านล่างบอกว่าสำหรับวิธีที่เราจะเลือก ขนาดก้าววิกฤตคือ 0.2, 0.02 และ 0.002 วินาที เมื่อสปริงแข็งขึ้นร้อยเท่า ขนาดก้าววิกฤตเล็กลงสิบเท่า นี่คือสิ่งที่ต้องอธิบาย

ภาพรวมก่อนอ่าน

ก้าวของวิธีเชิงตัวเลขมีข้อจำกัดสองอย่าง อย่างแรกคือ**ความแม่น** ความผิดพลาดในแต่ละก้าวต้องไม่มาก อย่างที่สองคือ**เสถียรภาพ** ความผิดพลาดต้องไม่โตขึ้นเรื่อย ๆ จนกลบคำตอบจริง ในสปริงแข็งข้อจำกัดที่สองคือตัวที่บังคับ และสิ่งที่กำหนดมันคือความถี่ที่สูงที่สุดของระบบ ไม่ใช่รูปร่างของคำตอบ

สปริงที่ไม่มีแรงหน่วง

เราจะใช้วิธีที่ทำให้เห็นง่ายที่สุด คือวิธีออยเลอร์ซิมเพลกติกที่เราเรียนในบทที่ 12 สำหรับระบบ $x'' = -\omega^2 x$ วิธีนี้เดินหนึ่งก้าวดังนี้ ความเร็วใหม่คำนวณจากตำแหน่งเดิม แล้วตำแหน่งใหม่คำนวณจากความเร็วใหม่

$$v_{j+1} = v_j - h\,\omega^2 x_j, \qquad x_{j+1} = x_j + h\,v_{j+1}.$$

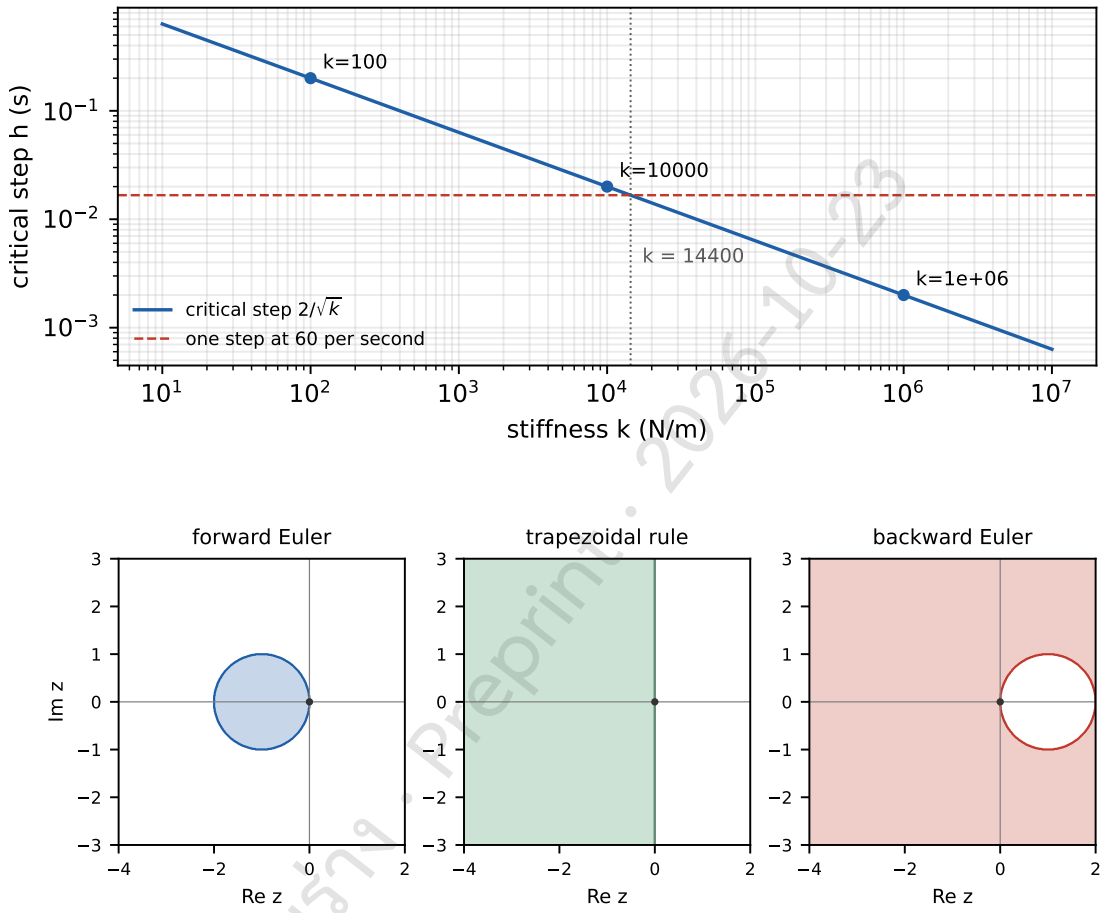
ไม่มีสมการที่ต้องแก้ในแต่ละก้าว ดังนั้นเป็นวิธีชัดแจ้ง ทฤษฎีบท 12.2.3 บอกว่าเมื่อ $\omega h < 2$ คำตอบถูกจำกัดขอบเขต ค่า h ที่ใช้ในหนึ่งก้าวเรียกว่า **ขนาดก้าว** (step size) ดังนั้น **ขนาดก้าววิกฤต** (critical step size) ของสปริงนี้คือ

$$h_{\text{crit}} = \frac{2}{\omega} = \frac{2}{\sqrt{k}}.$$

ฉบับร่าง · preprint · 2026-10-23

ตารางที่ 14.1: สปริงมวล 1 กิโลกรัมที่ไม่มีแรงหน่วง กับวิธีออยเลอร์ซิมเพลกติก ค่า ωh คำนวณที่ 60 ก้าวต่อวินาที (สร้างจาก `numanal/ch14.py` และตรวจด้วย `test_ch14.py`)

k (N/m)	ω (rad/s)	h_{crit} (s)	ωh ที่ 60 ต่อวินาที	เสถียรที่ 60 ต่อวินาที
10^2	10	0.2	0.1667	ใช่
10^4	100	0.02	1.6667	ใช่
10^6	1000	0.002	16.6667	ไม่



รูปที่ 14.2: บน: ขนาดก้าววิกฤตของสปริงมวล 1 กิโลกรัม $h_{\text{crit}} = 2/\sqrt{k}$ (รอดเมื่อ h น้อยกว่าค่านี้) เทียบกับความแข็ง k เส้นประคือหนึ่งก้าวที่ 60 ต่อวินาที และเส้นจุดคือ $k = 14400$ ล่าง: โดเมนเสถียรภาพของออยเลอร์ไปข้างหน้า (ดิสก์ที่มีจุดศูนย์กลางที่ -1) ของกฎเชิงสี่เหลี่ยมคางหมู (ซีกซ้ายของระนาบ) และของออยเลอร์ถอยหลัง (นอกดิสก์ที่มีจุดศูนย์กลางที่ 1) พื้นที่สีคือค่า z ที่ทำให้ $|R(z)| < 1$ (สร้างจาก `numanal/ch14.py`)

ค่านี้เป็นขีดแบ่ง ไม่ใช่ค่าที่ปลอดภัยที่สุด: ถ้า $h < h_{\text{crit}}$ คำตอบถูกจำกัดขอบเขต แต่ที่ $h = h_{\text{crit}}$ พอดี เมทริกซ์การก้าวมีค่าลักษณะเฉพาะซ้ำ -1 และคำตอบโตเป็นเส้นตรง จึงไม่ถูกจำกัดขอบเขต ตารางที่ 14.1 แสดงค่าเหล่านี้สำหรับสปริงทั้งสามตัว และแสดงค่า ωh ที่อัตราสมมติ 60 ก้าวต่อวินาที

ค่าขีดวิกฤตของ k ที่ 60 ก้าวต่อวินาทีคือ $k < (2 \cdot 60)^2 = 14400$ นิวตันต่อเมตร สปริงที่ $k \geq 14400$ ไม่รอดด้วยก้าว $1/60$ วินาที เช่น $k = 10^6$ ในตาราง ที่ก้าว $1/60$ วินาที สปริง $k = 10^4$ แกว่งแรงขึ้นจาก 0.01 เมตรเป็นราว 0.0181 เมตร แต่ยังไม่ระเบิด เมื่อก้าวเท่ากับ 0.99 เท่าของ h_{crit} ระยะแกว่งสูงสุดจากตำแหน่งสมดุล (**แอมพลิจูด**, amplitude) ขึ้นเป็น 0.0709 เมตรจากค่าเริ่ม 0.01 เมตร ทั้งสองค่าคำนวณจากโค้ดของเรา

ฉบับร่าง · Preprint · 2026-10-23

สมการทดสอบและโดเมนเสถียรภาพ

ทำไมต้องใช้สมการง่าย ๆ เช่น $y' = \lambda y$ ทั้งที่เราไม่ต้องใช้วิธีเชิงตัวเลขกับมันเลย เหตุผลคือ สำหรับระบบเชิงเส้นที่มีค่าลักษณะเฉพาะ λ แต่ละตัว การดูว่าวิธีรอดหรือไม่กับ $y' = \lambda y$ บอกเราได้ตรง ๆ ว่าวิธีนั้นรอดกับโหมดนั้นหรือเปล่า ซึ่งเป็นสิ่งที่ต้องการที่สุดสำหรับสปริงและข้อต่อ

นิยาม 14.1.1: โดเมนเสถียรภาพเชิงเส้นของวิธีเชิงตัวเลข

ให้วิธีเชิงตัวเลขใช้กับสมการทดสอบเชิงเส้น $y' = \lambda y$ โดย $y(0) = 1$ และขนาดก้าว $h > 0$ **โดเมนเสถียรภาพเชิงเส้น** (linear stability domain) ของวิธีนี้คือเซตของจำนวนเชิงซ้อน $z = h\lambda$ ที่ทำให้ $y_n \to 0$ เมื่อ $n \to \infty$ Iserles 2009, §4.2, p. 57

อ่านแบบเข้าใจง่าย

ถ้าผลเฉลยจริง $e^{\lambda t}$ สลายไปเพราะ $\operatorname{Re}\lambda < 0$ เราอยากให้ผลของวิธีก็สลายไปด้วย โดเมนเสถียรภาพคือส่วนของระนาบที่ $h\lambda$ ทำให้เป็นจริง จุดที่อยู่ในโดเมนหมายความว่าวิธีนี้ไม่ระเบิดกับโหมดนั้น ส่วนจุดที่อยู่นอกโดเมนหมายความว่าโหมดนั้นจะโตแม้ว่าผลจริงจะสลายไป

ทฤษฎีบท 14.1.1: โดเมนเสถียรภาพของวิธีออยเลอร์

วิธีออยเลอร์ใช้กับ $y' = \lambda y$ ได้ $y_{n+1} = (1 + z)\, y_n$ โดย $z = h\lambda$ และโดเมนเสถียรภาพของวิธีนี้คือ $D_{\mathrm{E}} = \{z \in \mathbb{C} : \ |1 + z| < 1\}$ ฟังก์ชัน $R(z) = 1 + z$ เรียกว่า**ฟังก์ชันเสถียรภาพ** (stability function)

พิสูจน์ (ของเราเอง).

1. จากสูตรออยเลอร์ $y_{n+1} = y_n + h\lambda y_n = (1 + z)\, y_n$ ดังนั้น $y_n = (1 + z)^n$ เมื่อ $y_0 = 1$ โดยการเหนี่ยวนำตรง ๆ
2. ถ้า $|1 + z| < 1$ ดังนั้น $|y_n| = |1 + z|^n \to 0$ จึงอยู่ในโดเมน
3. ถ้า $|1 + z| = 1$ ดังนั้น $|y_n| = 1$ ทุก n จึงไม่ไปที่ศูนย์
4. ถ้า $|1 + z| > 1$ ดังนั้น $|y_n| \to \infty$ จึงไม่อยู่ในโดเมน

สรุป: วิธีออยเลอร์รอดกับโหมดหนึ่งก็ต่อเมื่อ $|1 + z| < 1$ ซึ่งคือดิสก์รัศมี 1 ที่มีจุดศูนย์กลางที่ -1

ลองเอง: หนึ่งก้าวด้วยมือ

ให้ $y' = -100y$ และ $y(0) = 1$ ถ้า $h = 0.025$ ดังนั้น $z = -2.5$ และ $y_1 = 1+z = -1.5$ ถ้า $h = 0.015$ ดังนั้น $z = -1.5$ และ $y_1 = -0.5$ เครื่องหมายสลับและค่าลดลง จึงรอด ส่วนค่าแรกไม่รอด เพราะ $|1+z| = 1.5 > 1$ ขีดจำกัดที่แท้จริงคือ $h < 2/100 = 0.02$ เดาก่อนว่าที่ $h = 0.0199$ กับ $h = 0.0201$ อะไรจะเกิดขึ้น

เกร็ดเพิ่มเติม

ISE ใช้ $\lambda \in \mathbb{C}$ สำหรับสมการทดสอบ และเขียนว่าสำหรับระบบเชิงเส้นของวิธีออยเลอร์ขนาดก้าวถูกกำหนดโดยองค์ประกอบที่แข็งที่สุดของระบบ Iserles 2009, §4.2, p. 57 เราจะพิสูจน์ผลนี้สำหรับวิธีออยเลอร์ซิมเพลกติกในทฤษฎีบท 14.1.3 โดยไม่ต้องพึ่งหนังสือ

ฉบับร่าง Preprint · 2026-10-23

นิยาม 14.1.2: วิธี A-เสถียร

วิธีเชิงตัวเลขเรียกว่า **A-เสถียร** (A-stable) เมื่อ $\mathbb{C}_- := \{z \in \mathbb{C} : \operatorname{Re} z < 0\} \subseteq D$ นั่นคือวิธีรอดกับทุกโหมดที่สลายไป โดยไม่ว่า h จะใหญ่เพียงใด Iserles 2009, §4.2, p. 59

ทฤษฎีบท 14.1.2: กฎเชิงสี่เหลี่ยมคางหมูเป็น A-เสถียร

กฎเชิงสี่เหลี่ยมคางหมูใช้กับ $y' = \lambda y$ ได้ $y_{n+1} = R(z)\, y_n$ โดย $R(z) = \dfrac{1 + z/2}{1 - z/2}$ และโดเมนเสถียรภาพของวิธีนี้คือ $\{z :\ \operatorname{Re} z < 0\}$ ดังนั้นวิธีนี้เป็น A-เสถียร

พิสูจน์ (ของเราเอง).

1. สูตรของกฎเชิงสี่เหลี่ยมคางหมูคือ $y_{n+1} = y_n + \frac{h}{2}(\lambda y_{n+1} + \lambda y_n)$ จัดใหม่ได้ $(1 - \frac{z}{2})y_{n+1} = (1 + \frac{z}{2})y_n$ เมื่อ $z \neq 2$
2. ดังนั้น $|y_n| \to 0$ ก็ต่อเมื่อ $|R(z)| < 1$ นั่นคือ $|1 + \frac{z}{2}| < |1 - \frac{z}{2}|$
3. เขียน $z = x + iy$ จากเอกลักษณ์ $|a + b|^2 - |a - b|^2 = 4\operatorname{Re}(a\bar{b})$ โดยใส่ $a = 1$ และ $b = z/2$ ได้ $|1 + \frac{z}{2}|^2 - |1 - \frac{z}{2}|^2 = 2x$
4. ดังนั้น $|R(z)| < 1$ ก็ต่อเมื่อ $x < 0$ และถ้า $x = 0$ จะได้ $|R(z)| = 1$ จึงไม่อยู่ในโดเมน

สรุป: กฎเชิงสี่เหลี่ยมคางหมูรอดกับทุกโหมดที่สลายไป โดยไม่มีข้อจำกัดของขนาดก้าว

อ่านแบบเข้าใจง่าย

วิธีออยเลอร์ไปข้างหน้ามีโดเมนเป็นดิสก์เล็กที่อยู่ซ้ายมือ ส่วนกฎเชิงสี่เหลี่ยมคางหมูมีโดเมนเป็นครึ่งระนาบซ้ายทั้งหมด จึงรอดกับโหมดที่สลายไปทุกโหมด ไม่ว่า h จะเท่าไร เพราะนี่คือ A-เสถียร ถ้าเราเลือกก้าวด้วยความแม่นยำอย่างเดียว วิธีนี้จึงไม่ต้องกลัวว่าก้าวจะใหญ่เกินไปเพราะความไม่เสถียร ดูรูปที่ 14.2 ประกอบ

ข้อควรระวัง: เสถียรไม่ได้แปลว่าตรงความจริง

A-เสถียรบอกเพียงว่าไม่ระเบิด ไม่ได้บอกว่าคำตอบแม่น ลองดูกฎเชิงสี่เหลี่ยมคางหมูกับ $y' = -100y$ และ $h = 0.5$ เราได้ $R(-50) = -\frac{24}{26} = -\frac{12}{13}$ ดังนั้น $y_1 = -0.923$ ค่านี้เล็กกว่า 1 จึงไม่ระเบิด แต่คำตอบจริง e^{-50} เกือบเป็นศูนย์ และค่าที่ได้มีเครื่องหมายผิดในก้าวแรก ความแม่นยำต้องการก้าวที่เล็กพอ ซึ่งเป็นอีกเรื่องหนึ่ง

สมการตัวตึงและอัตราส่วนความตึง

ตอนนี้ลองดูระบบที่มีโหมดหลายความถี่พร้อมกัน ซึ่งเป็นกรณีที่ใกล้ตัวที่สุดของหุ่น สมการเหล่านี้เรียกว่า**สมการตัวตึง** (stiff equation) คำนิยามนี้ไม่ใช่นิยามทางคณิตศาสตร์ที่เข้มงวด แต่เป็นความหมายเชิงปฏิบัติ (Iserles 2009, §4.1, p. 56; Atkinson 1989, §6.9, p. 410)

ร่างต้น Preprint · 2026-10-23

นิยาม 14.1.3: สมการตัวตึง (ความหมายเชิงปฏิบัติ)

เรียกระบบ $y' = f(t, y)$ ว่าสมการตัวตึงเมื่อวิธีบางวิธีต้องเดินก้าวเล็กมากเพียงเพื่อกันคำตอบไม่พัง ทั้งที่คำตอบจริงอาจเรียบและนิ่งมาก สำหรับระบบเชิงเส้น $y' = \Lambda y$ ระบบจะตึงเมื่อมีค่าลักษณะเฉพาะที่มีส่วนจริงลบและขนาดใหญ่มากปนอยู่

นิยาม 14.1.4: อัตราส่วนความตึง

สำหรับระบบเชิงเส้น $y' = \Lambda y$ **อัตราส่วนความตึง** (stiffness ratio) วัดว่าโหมดที่เร็วที่สุดเร็วกว่าโหมดที่ช้าที่สุดกี่เท่า โดยดูเฉพาะขนาดของค่าลักษณะเฉพาะ λ_j คือ $\max_j |\lambda_j| / \min_j |\lambda_j|$ ภายใต้สมมติฐาน $\min_j |\lambda_j| > 0$ (ถ้ามี $\lambda_j = 0$ ตัวหารเป็นศูนย์ อัตราส่วนนี้จึงนิยามไม่ได้) Iserles 2009, §4.1, p. 56

ลองเอง: ระบบสองท่อนที่ต่อกันด้วยข้อต่อแข็ง

ให้หุ่นมีสองท่อน มวลท่อนละ 1 กิโลกรัม ท่อนแรกติดลำตัวด้วยข้อต่ออ่อน $k_1 = 10^2$ และท่อนที่สองต่อกับท่อนแรกด้วยข้อต่อแข็ง $k_2 = 10^6$ **เมทริกซ์สติฟเนส** (stiffness matrix) ของระบบนี้คือ $K = \begin{pmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{pmatrix}$ ซึ่งบอกว่าแรงจากสปริงตอบสนองต่อการเลื่อนของมวลแต่ละก้อนอย่างไร ค่าลักษณะเฉพาะ μ ของ K ต้องสอดคล้องกับ $\mu^2 - (k_1 + 2k_2)\mu + k_1 k_2 = 0$ ดังนั้น $\mu_{\text{low}} = 49.99875$ และ $\mu_{\text{high}} = 2000050.00125$ เมื่อใช้ $\omega = \sqrt{\mu}$ ความถี่คือ 7.071 และ 1414.231 เรเดียนต่อวินาที อัตราส่วนความตึงคือ 200.0 (ค่าลักษณะเฉพาะของระบบลำดับหนึ่งคือ $\pm i\omega$ จึงได้ $\omega_{\text{high}}/\omega_{\text{low}}$)

ทฤษฎีบท 14.1.3: ขนาดก้าวของระบบมวลติดสปริงถูกกำหนดโดยโหมดที่แข็งที่สุด

ให้ K เป็นเมทริกซ์สมมาตรบวกแน่นอนขนาด $d \times d$ ที่มีค่าลักษณะเฉพาะ $\mu_{\max}$ มากที่สุด และมวลทุกก้อนเป็น 1 กิโลกรัม สำหรับระบบ $x'' = -Kx$ การจำลองด้วยวิธีออยเลอร์ซิมเพลกติกมีคำตอบถูกจำกัดขอบเขตทุกคำตอบ เมื่อ $h\sqrt{\mu_{\max}} < 2$ และมีคำตอบที่ไม่ถูกจำกัดขอบเขตเมื่อ $h\sqrt{\mu_{\max}} > 2$

พิสูจน์ (ของเราเอง).

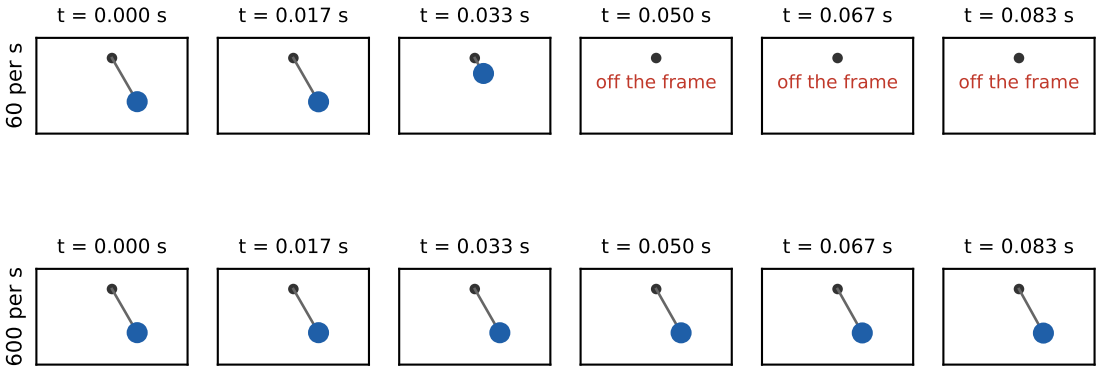
1. ก่อนอื่นดูกรณี $d = 1$ โดยใช้ $\omega^2 = \mu$ และ $c = \omega h$ จากสูตรการก้าว $x_{j+1} = (1 - c^2)x_j + h\,v_j$ และ $v_{j+1} = -\omega^2 h\,x_j + v_j$ ดังนั้นเมทริกซ์การก้าวคือ $M = \begin{pmatrix} 1 - c^2 & h \\ -\omega^2 h & 1 \end{pmatrix}$ มี $\operatorname{tr} M = 2 - c^2$ และ $\det M = 1$

2. ค่าลักษณะเฉพาะของ M เป็นรากของ $\lambda^2 - (2 - c^2)\lambda + 1 = 0$ ถ้า $c < 2$ จะได้ $|2 - c^2| < 2$ รากเป็นคู่สังยุคเชิงซ้อนที่มีขนาด 1 จึงคำตอบถูกจำกัดขอบเขต

3. ถ้า $c > 2$ จะได้ $2 - c^2 < -2$ รากเป็นจำนวนจริงลบสองค่าที่ผลคูณเป็น 1 และต่างกัน จึงมีรากที่มีขนาดมากกว่า 1 คำตอบจะโตขึ้นเรื่อย ๆ สำหรับค่าเริ่มต้นส่วนใหญ่ ส่วนกรณี $c = 2$ ดูข้อควรระวังในทฤษฎีบท 12.2.3

4. สำหรับ $d > 1$ เขียน $K = Q\Lambda Q^{\mathrm{T}}$ โดย Q เป็น**เมทริกซ์เชิงตั้งฉาก** (orthogonal matrix) (หมุนหรือสะท้อนเวกเตอร์โดยไม่เปลี่ยนความยาว) และ $\Lambda = \operatorname{diag}(\mu_1, \ldots, \mu_d)$ การเปลี่ยนตัวแปร $y = Q^{\mathrm{T}}x$ ไม่เปลี่ยนการก้าว เพราะการก้าวเป็นเชิงเส้นและ Q ตั้งฉาก ระบบจึงแยกเป็น d สปริงเดี่ยวที่มีความถี่ $\sqrt{\mu_j}$

ฉบับร่าง · Preprint 2026-10-23



รูปที่ 14.3: แขนลูกตุ้มแข็งที่ $k = 10^6$ นิวตันต่อเมตร แถวบนคือ 60 ก้าวต่อวินาที ที่ $t = 0.050$ วินาที (ก้าวที่ 3) แขนอยู่ไกลถึงราว 180 เมตร และที่ $t = 0.067$ วินาที (ก้าวที่ 4) ไกลถึงราว 4.9×10^4 เมตร จึงหลุดจากภาพ แถวล่างคือ 600 ก้าวต่อวินาที (ก้าวย่อยสิบก้าวต่อเฟรม) แกว่งนิ่งตามเดิม เฟรมทั้งหกใช้เวลาเดียวกัน (สร้างจาก `numanal/ch14.py`)

5. ทุกโหมดถูกจำกัดขอบเขตก็ต่อเมื่อ $h\sqrt{\mu_j} < 2$ ทุก j ซึ่งเทียบเท่ากับ $h\sqrt{\mu_{\max}} < 2$ เพราะ $\mu_{\max}$ มากที่สุด เมื่อ $h\sqrt{\mu_{\max}} > 2$ ให้เลือกค่าเริ่มต้นในทิศของเวกเตอร์ลักษณะเฉพาะของ $\mu_{\max}$ โหมดนั้นจะโตขึ้นเรื่อย ๆ จึงไม่ถูกจำกัดขอบเขต

สรุป: โหมดที่แข็งที่สุดตั้งเพดานให้ก้าว โหมดอื่นที่นิ่มกว่าไม่ช่วยอะไร

ในโลกจริง: ปฏิกิริยาเคมีที่เร็วและช้าปนกัน

ในห้องทดลองเคมี ปฏิกิริยาหนึ่งอาจจบภายในเสี้ยววินาทีขณะที่อีกปฏิกิริยาหนึ่งใช้เวลานานมาก เมื่อสองเรื่องนี้ปนอยู่ในสมการเดียว สมการนั้นจะตึง หนังสือ ISE บอกว่าอัตราส่วนราว 10^{17} พบได้บ่อยในจลนศาสตร์เคมี Iserles 2009, §4.1, p. 56 ในหุ่นของเราอัตราส่วนเพียง 200 แต่กลไกเดียวกันคือเรื่องเร็วกับเรื่องช้าอยู่ในระบบเดียวกัน

เกร็ดเพิ่มเติม

บทความปี 1952 ใน *PNAS* มีชื่อเรื่องว่า *Integration of Stiff Equations* Curtiss and Hirschfelder 1952 เราบอกได้เพียงว่าคำว่า stiff อยู่ในชื่อเรื่องนี้ ไม่ได้บอกว่าเป็นครั้งแรกที่ใช้คำนี้ และเรายังไม่ได้อ่านเนื้อหาบทความ

ฉบับร่าง · Preprint · 2026-10-23

หุ่นแร็กดอลล์ที่สั่น

เอาทฤษฎีกลับไปที่หุ่น แขนหนึ่งท่อนเป็นลูกตุ้มที่มวล 1 กิโลกรัม ห้อยจากจุดหมุนด้วยแท่งที่เป็นสปริงแข็งมาก ความยาวพักคือ 1 เมตร ค่าความแข็ง $k = 10^6$ นิวตันต่อเมตร เริ่มแกว่งที่มุม 30 องศา และมีแรงโน้มถ่วง $g = 9.8$ เมตรต่อวินาทีกำลังสอง เราจำลองด้วยวิธีออยเลอร์ซิมเพลกติกโดยใช้แรงรัศมีจากสปริง

ที่ 60 ก้าวต่อวินาที $\omega h = 1000/60 = 16.7$ เกินค่าขีดจำกัด 2 มาก แขนจึงกระเด็นภายในสี่ก้าว ที่ 600 ก้าวต่อวินาที $\omega h = 1.67 < 2$ และตลอด 1 วินาทีระยะจากจุดหมุนเบี่ยงเบนจากความยาวพักไม่เกิน 2.8×10^{-5} เมตร ทั้งสองค่ามาจากโค้ดของเรา

ข้อควรระวัง: ก้าวเล็กเพราะเสถียรภาพ ไม่ใช่เพราะความแม่น

ISE เขียนว่าในตัวอย่างสมการแข็งของตน ขนาดก้าวที่ถูกกดลงไม่ได้มาจากข้อพิจารณาเรื่องความคลาดเคลื่อน แต่มาจากความไม่เสถียร Iserles 2009, §4.1, p. 55 ดังนั้นการลดความผิดพลาดของคำตอบอย่างเดียวจะไม่ช่วยให้แขนนิ่ง ต้องเลือกวิธีหรือก้าวที่เสถียรสำหรับโหมดที่แข็งที่สุด

ในโลกจริง: ก้าวย่อยในไลบรารีฟิสิกส์

เอกสารของ Box2D เรื่องการจำลอง ระบุว่าข้อต่อที่เป็นโซ่ยืดน้อยลงเมื่อมีก้าวย่อยมากขึ้น และแนะนำให้ใช้ 4 ก้าวย่อย Catto 2025 เอกสารเดียวกันระบุว่ามีการชนแบบไม่ยืดหยุ่นเมื่อความเร็วการชนต่ำ เพื่อป้องกันการสั่น (jitter) Catto 2025 เรานำเฉพาะข้อความเหล่านี้มาใช้ ไม่ได้บอกว่าไลบรารีใช้วิธีเชิงตัวเลขแบบใด

ในโลกจริง: ผ้าที่คำนวณด้วยก้าวใหญ่

บทคัดย่อของเอกสารเรื่องการจำลองผ้าขนาดใหญ่ระบุว่า ขนาดก้าวเป็นข้อจำกัดหลักของการจำลองผ้า และวิธีแบบโดยปริยายที่แก้ระบบเชิงเส้นขนาดใหญ่ทำให้ก้าวใหญ่ได้ Baraff and Witkin 1998 เราอ้างเพียงสิ่งที่บทคัดย่อบอก หัวข้อ 14.2 จะกลับมาที่ข้อนี้

สรุปใจความของหัวข้อ 14.1: กลับไปที่หุ่น

- ก้าวถูกจำกัดสองแบบ คือความแม่นและเสถียรภาพ สปริงแข็งทำให้เสถียรภาพเป็นตัวจำกัด
- สมการ $y' = \lambda y$ บอกได้ว่าวิธีรอดหรือไม่ ออยเลอร์ไปข้างหน้ารอดเมื่อ $|1 + z| < 1$ และไม่รอดบนแกนจินตภาพ จึงใช้ออยเลอร์ไปข้างหน้ากับสปริงที่ไม่มีแรงหน่วงโดยตรงไม่ได้
- กฎเชิงสี่เหลี่ยมคางหมูเป็น A-เสถียร จึงไม่มีข้อจำกัดของขนาดก้าวเพราะความไม่เสถียร แต่ยังต้องเลือกก้าวให้แม่นพอ
- ก้าวที่เสถียรถูกกำหนดโดยโหมดที่แข็งที่สุด คือ $h\sqrt{\mu_{\max}} < 2$

กลับไปที่หุ่น ข้อต่อที่แข็ง 10^6 บังคับให้ก้าวต้องเล็กกว่า $1/60$ วินาทีหลายเท่า แต่การแบ่งเป็นก้าวย่อยสิบก้าวต่อเฟรมทำให้ $\omega h = 1.67$ ซึ่งอยู่ในช่วงที่เสถียร

อ่านต่อ. ตัวอย่างสมการแข็งที่ละเอียดกว่าอยู่ใน Iserles 2009, §4.1--4.2, pp. 53--59 A-เสถียรของวิธีรุงเงอ-คุททาอยู่ใน Iserles 2009, §4.3, pp. 59--63 ส่วนสมการแข็งทั่วไปและวิธีเส้นอยู่ใน Atkinson 1989, §6.9, pp. 409--418 หัวข้อ 14.2 จะบอกราคาของวิธีแบบโดยปริยายเพื่อแลกกับการไม่ระเบิด

ฉบับร่าง · Preprint · 2026-10-2

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 14.1.1 (⋆, ทำด้วยมือ) สปริงมวล 1 กิโลกรัมมีค่าความแข็ง $k = 3600$ นิวตันต่อเมตร หาความถี่เชิงมุม ω และขนาดก้าววิกฤตของวิธีออยเลอร์ชิมเพลกติก แล้วตรวจว่าที่ 60 ก้าวต่อวินาที ωh อยู่ในช่วงที่เสถียรหรือไม่

แบบฝึกหัด 14.1.2 (⋆, ทำด้วยมือ) ให้ $y' = -50y$ และ $y(0) = 1$ ใช้วิธีออยเลอร์กับขนาดก้าว $h = 0.05$ คำนวณ y_1 และ y_2 แล้วหาขนาดก้าววิกฤตที่วิธีนี้แยกระหว่างรอดกับไม่รอด

แบบฝึกหัด 14.1.3 (⋆⋆, พิสูจน์) แสดงว่าวิธีออยเลอร์ถอยหลังมีฟังก์ชันเสถียรภาพ $R(z) = 1/(1 - z)$ และเป็น A-เสถียร ให้ใช้เอกลักษณ์ $|1 - z|^2 = 1 - 2\,\mathrm{Re}\, z + |z|^2$

แบบฝึกหัด 14.1.4 (⋆⋆, คำนวณ) ให้ $y' = -100y$ และ $y(0) = 1$ คำนวณ y_1 และ y_2 ของกฎเชิงสี่เหลี่ยมคางหมูเมื่อ $h = 0.1$ และเมื่อ $h = 0.5$ เปรียบเทียบกับ e^{-10} และ e^{-50} อธิบายว่าเหตุใดทั้งสองกรณีไม่ระเบิด แต่ยังไม่แม่นยำ

แบบฝึกหัด 14.1.5 (⋆⋆, คำนวณและตีความ) ทำซ้ำการคำนวณของหัวข้อ 14.1 โดยให้ข้อต่ออ่อน $k_1 = 10^2$ คงเดิม แต่ข้อต่อแข็ง k_2 เปลี่ยนเป็น 10^4 มวลท่อนละ 1 กิโลกรัม หาอัตราส่วนความถี่และขนาดก้าววิกฤตของวิธีออยเลอร์ชิมเพลกติก แล้วบอกว่าก้าวใหญ่ขึ้นกี่เท่าเมื่อเทียบกับ $k_2 = 10^6$

แบบฝึกหัด 14.1.6 (⋆⋆, ตีความ) ข้อต่อของหุ่นมีค่าความแข็ง $k = 10^5$ นิวตันต่อเมตร มวล 1 กิโลกรัม สมมติให้เกมคำนวณ 60 เฟรมต่อวินาที และแบ่งแต่ละเฟรมเป็นก้าวย่อยที่เท่ากัน ต้องมีก้าวย่อยอย่างน้อยกี่ก้าวต่อเฟรมเพื่อให้วิธีออยเลอร์ชิมเพลกติกเสถียร ตรวจว่าสองก้าวย่อยไม่พอ แต่สามก้าวย่อยพอ

แบบฝึกหัด 14.1.7 (⋆ ⋆ ⋆, พิสูจน์) ให้ $K = \mathrm{diag}(4, 9, 100)$ สำหรับระบบ $x'' = -Kx$ โดยมวลทุกก้อนเป็น 1 กิโลกรัม

1. เขียนระบบนี้เป็นสามสมการแยกจากกัน และหาความถี่ของแต่ละสมการ
2. หาขนาดก้าววิกฤตของวิธีออยเลอร์ชิมเพลกติกที่ทำให้ทุกโหมดถูกจำกัดขอบเขต และอธิบายว่าเหตุใดโหมดที่แข็งที่สุดเป็นตัวกำหนด
3. แสดงว่าเมื่อ $c = 2.1$ สมการ $\lambda^2 - (2 - c^2)\lambda + 1 = 0$ มีรากที่มีขนาดมากกว่า 1 แล้วอธิบายว่าเหตุใดเมื่อ $h = 0.21$ คำตอบของโหมดที่แข็งที่สุดจึงโตขึ้น

คำใบ้.

ตัวช่วยคิด

ข้อ 3 ใช้เมทริกซ์การก้าว $M = \begin{pmatrix} 1 - c^2 & h \\ -\omega^2 h & 1 \end{pmatrix}$ ของทฤษฎีบท 14.1.3 และหาค่าของ $2 - c^2$ เมื่อ $c = 2.1$

ฉบับร่าง · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 14.1.1.

1. ความถี่คือ $\omega = \sqrt{k/m} = \sqrt{3600} = 60$ เรเดียนต่อวินาที
2. ขนาดก้าววิกฤตคือ $h_{\text{crit}} = 2/\omega = 2/60 = \frac{1}{30} \approx 0.0333$ วินาที
3. ที่ 60 ก้าวต่อวินาที $h = \frac{1}{60}$ ดังนั้น $\omega h = 60/60 = 1 < 2$ จึงอยู่ในช่วงที่เสถียร และ $h = \frac{1}{60} < \frac{1}{30}$ สอดคล้องกับข้อ 2

เฉลย 14.1.2.

1. ค่า $z = h\lambda = 0.05 \cdot (-50) = -2.5$
2. จากสูตรออยเลอร์ $y_{n+1} = (1+z)y_n$ จึงได้ $y_1 = (1-2.5)\cdot 1 = -1.5$ และ $y_2 = (-1.5)^2 = 2.25$ ค่าสลับเครื่องหมายและขนาดเพิ่มขึ้น
3. วิธีรอดเมื่อ $|1+z| < 1$ นั่นคือ $-1 < 1 - 50h < 1$ อสมการทางขวาให้ $h > 0$ ส่วนทางซ้ายให้ $50h < 2$ ดังนั้น $0 < h < 0.04$ ขนาดก้าววิกฤตคือ 0.04

เฉลย 14.1.3.

1. วิธีออยเลอร์ถอยหลังเขียนเป็น $y_{n+1} = y_n + z\,y_{n+1}$ เมื่อ $z = h\lambda$ จัดใหม่ได้ $(1-z)y_{n+1} = y_n$ จึง $R(z) = 1/(1-z)$ เมื่อ $z \neq 1$
2. $|R(z)| < 1$ ก็ต่อเมื่อ $|1-z| > 1$ เขียน $z = x + iy$ จะได้ $|1-z|^2 = (1-x)^2 + y^2 = 1 - 2x + x^2 + y^2 = 1 - 2x + |z|^2$
3. ถ้า $x = \operatorname{Re} z < 0$ แล้ว $-2x > 0$ ดังนั้น $|1-z|^2 > 1$ จึง $|R(z)| < 1$ ทุก $z \in \mathbb{C}_-$ นั่นคือ $\mathbb{C}_- \subseteq D$ วิธีนี้เป็น A-เสถียร
4. หมายเหตุ: บนแกนจินตภาพ $x = 0$ ได้ $|1-z|^2 = 1 + y^2 > 1$ เมื่อ $y \neq 0$ จึงรอดด้วยเช่นกัน

เฉลย 14.1.4.

1. เมื่อ $h = 0.1$ ได้ $z = -10$ และ $R(-10) = \dfrac{1-5}{1+5} = -\dfrac{2}{3}$ จึง $y_1 = -\frac{2}{3} \approx -0.6667$ และ $y_2 = \left(-\frac{2}{3}\right)^2 = \frac{4}{9} \approx 0.4444$
2. เมื่อ $h = 0.5$ ได้ $z = -50$ และ $R(-50) = \dfrac{1-25}{1+25} = -\dfrac{12}{13}$ จึง $y_1 = -\frac{12}{13} \approx -0.9231$ และ $y_2 = \frac{144}{169} \approx 0.8521$
3. คำตอบจริงคือ $e^{-10} \approx 4.54 \times 10^{-5}$ และ $e^{-50} \approx 1.9 \times 10^{-22}$ ทั้งสองกรณี $|R(z)| < 1$ จึงไม่ระเบิด (เพราะกฎเชิงสี่เหลี่ยมคางหมูเป็น A-เสถียร) แต่ค่าที่ได้ห่างจากคำตอบจริงมาก และเครื่องหมายของ y_1 ผิดทั้งสองกรณี เพราะคำตอบจริงเป็นบวก สรุปว่าเสถียรภาพกับความแม่นยำเป็นเรื่องแยกกัน

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 14.1.5.

1. $T = k_1 + 2k_2 = 100 + 20000 = 20100$ และ $k_1 k_2 = 10^6$ จึง
$$\sqrt{T^2 - 4k_1 k_2} = \sqrt{404010000 - 4000000} = \sqrt{400010000} \approx 20000.25$$
2. ค่าลักษณะเฉพาะคือ $\mu_{\text{low}} = \frac{1}{2}(20100 - 20000.25) \approx 49.875$ และ $\mu_{\text{high}} = \frac{1}{2}(20100 + 20000.25) \approx 20050.125$
3. อัตราส่วนความตึงคือ $\sqrt{\mu_{\text{high}}/\mu_{\text{low}}} = \sqrt{402.0} \approx 20.05$
4. ขนาดก้าววิกฤตคือ $2/\sqrt{\mu_{\text{high}}} = 2/141.60 \approx 0.014124$ วินาที
5. เมื่อเทียบกับ $k_2 = 10^6$ ซึ่งได้ 0.0014142 วินาที ก้าวใหญ่ขึ้นราว 10 เท่า เพราะ k_2 เล็กลง 100 เท่า ความถี่สูงสุดจึงเล็กลง $\sqrt{100} = 10$ เท่า

เฉลย 14.1.6.

1. ความถี่คือ $\omega = \sqrt{10^5} \approx 316.23$ เรเดียนต่อวินาที
2. เงื่อนไขเสถียรคือ $\omega h < 2$ จึง $h < 2/316.23 \approx 0.0063246$ วินาที นั่นคือต้องมีมากกว่า $1/0.0063246 \approx 158.1$ ก้าวต่อวินาที และต่อเฟรม 60 เฟรมต่อวินาทีต้องมีมากกว่า $158.1/60 \approx 2.64$ ก้าวย่อยต่อเฟรม
3. ตรวจสองก้าวย่อย $h = 1/120$ ได้ $\omega h = 2.635 > 2$ ไม่เสถียร ส่วนสามก้าวย่อย $h = 1/180$ ได้ $\omega h = 1.757 < 2$ เสถียร
4. ในโค้ดของเรา สามก้าวย่อยรันสามพันก้าวได้ค่าสูงสุดของ $|x|$ ประมาณ 0.0209 เมตร จากค่าเริ่ม 0.01 เมตร ส่วนสองก้าวย่อยทำให้ค่าล้น

เฉลย 14.1.7.

1. เมทริกซ์เป็นทแยงมุม จึงได้ $x_1'' = -4x_1$, $x_2'' = -9x_2$, $x_3'' = -100x_3$ ความถี่คือ $\omega_1 = 2$, $\omega_2 = 3$ และ $\omega_3 = 10$ แต่ละสมการเป็นสปริงเดี่ยวที่แยกจากกัน
2. สำหรับสปริงเดี่ยวที่มีความถี่ ω_j วิธีออยเลอร์ชิมเพลกติกถูกจำกัดขอบเขตเมื่อ $h\omega_j < 2$ (ทฤษฎีบท 14.1.3 กรณี $d = 1$) จึงได้ $h < 1$, $h < \frac{2}{3}$ และ $h < 0.2$ ขนาดก้าววิกฤตคือ $0.2 = 2/10$ ซึ่งมาจากโหมดที่แข็งที่สุด $\omega_3 = 10$ โหมดที่นิ่มกว่าอนุญาตให้ก้าวใหญ่กว่า แต่ไม่ได้ช่วยเมื่อมีโหมดแข็งอยู่ในระบบ
3. ถ้า $c = 2.1$ ได้ $2 - c^2 = 2 - 4.41 = -2.41$ สมการคือ $\lambda^2 + 2.41\lambda + 1 = 0$ มีรากคือ $\lambda = \frac{1}{2}\big(-2.41 \pm \sqrt{2.41^2 - 4}\big) = \frac{1}{2}(-2.41 \pm 1.3446)$ รากหนึ่งคือ -1.8773 ซึ่งมีขนาดมากกว่า 1 เมื่อ $h = 0.21$ โหมด $\omega_3 = 10$ มี $c = \omega_3 h = 2.1$ ดังนั้นขนาดของคำตอบจะคูณด้วยประมาณ 1.877 ทุกก้าว จึงโตขึ้น ส่วนโหมดอื่นยังไม่โต แต่โหมดนี้ทำให้ระบบทั้งหมดไม่ถูกจำกัดขอบเขต

ฉบับร่าง · Preprint · 2026-10-25

14.2 ก้าวแบบโดยปริยายและสูตร BDF

สถานการณ์สมมติ: หุ่นที่ไม่กระเด็นแต่หยุดนิ่งเร็วเกินไป

ในหัวข้อ 14.1 ข้อต่อแข็งทำให้หุ่นกระเด็น ทางออกหนึ่งคือเดินก้าวเล็กลง อีกทางคือเลือกวิธีที่ไม่ระเบิดแม้ก้าวใหญ่ เช่น **วิธีออยเลอร์ถอยหลัง** (backward Euler method) ซึ่งเป็น**วิธีแบบโดยปริยาย** (implicit method) คือค่าใหม่ปรากฏทั้งสองข้างของสูตร ทุกก้าวจึงต้องแก้สมการ เมื่อใช้วิธีนี้กับสปริงตัวเดียวที่ 60 ก้าวต่อวินาที สปริงไม่ระเบิด แต่การแกว่งหายไปเร็วกว่าที่ผลเฉลยจริงควรเป็นมาก ถ้ามองเป็นหุ่น นี่คือความเหนียวเหนอะ (sticky) ที่ทำให้หุ่นดูเหมือนติดอยู่กับที่ ตัวเลขในหัวข้อนี้มาจากสปริงตัวเดียวและมือที่กดกำแพง ไม่ใช่ลูกตุ้มทั้งตัว

ภาพรวมก่อนอ่าน

วิธีแบบโดยปริยายบางวิธีไม่ระเบิดเพราะใช้ค่าใหม่ช่วยคำนวณตัวมันเอง ราคาที่ต้องจ่ายคือทุกก้าวต้องแก้สมการ และในสปริงแต่ละก้าวพลังงานจะลดลง ส่วน**สูตร BDF** (backward differentiation formula (BDF)) เป็นวิธีหลายขั้นแบบโดยปริยายที่ใช้ค่าเก่าหลายค่า ซึ่งเมื่อ ωh เล็กจะเสียพลังงานน้อยกว่ามาก แต่มีเพดานว่าวิธีหลายขั้นที่ไม่ระเบิดมีอันดับสูงได้เท่าไร

วิธีออยเลอร์ถอยหลัง: แก้สมการทุกก้าว

วิธีออยเลอร์ถอยหลังเขียนได้เป็น

$$y_{n+1} = y_n + h\, f(t_{n+1}, y_{n+1}).$$

นี่คือวิธีออยเลอร์แบบโดยปริยายที่เราเรียนในบทที่ 12 ค่า y_{n+1} อยู่ทั้งสองข้าง เราจึงหาค่าโดยแก้สมการ ถ้า $f(t,y) = \lambda y$ การแก้เป็นเพียงการหาร ได้ $y_{n+1} = y_n/(1-z)$ เมื่อ $z = h\lambda$ นั่นคือ $R(z) = 1/(1-z)$ ซึ่งเราเคยเห็นในแบบฝึกหัด 14.1.3 แล้วว่าเป็น A-เสถียร ถ้า f ไม่เป็นเชิงเส้น ทุกก้าวต้องใช้**วิธีนิวตัน-ราฟสัน** (Newton–Raphson method) ช่วยหาค่าใหม่ (Iserles 2009, §7.2, pp. 127--130) ซึ่งทำให้หนึ่งก้าวมีต้นทุนมากกว่าวิธีชัดแจ้ง

สำหรับสปริงเชิงเส้น ผลของการแก้สมการเขียนได้เป็นสูตรปิด และทำให้เห็นว่าพลังงานลดลงอย่างไร

ทฤษฎีบท 14.2.1: พลังงานของสปริงลดลงทุกก้าวของวิธีออยเลอร์ถอยหลัง

ให้ $\omega > 0$ และ $h > 0$ ใช้วิธีออยเลอร์ถอยหลังกับระบบ $x' = v$ และ $v' = -\omega^2 x$ ให้ $E_j = \omega^2 x_j^2 + v_j^2$ ซึ่งเป็นสองเท่าของพลังงานต่อหน่วยมวล เมื่อไม่มีข้อจำกัดของ h จะได้

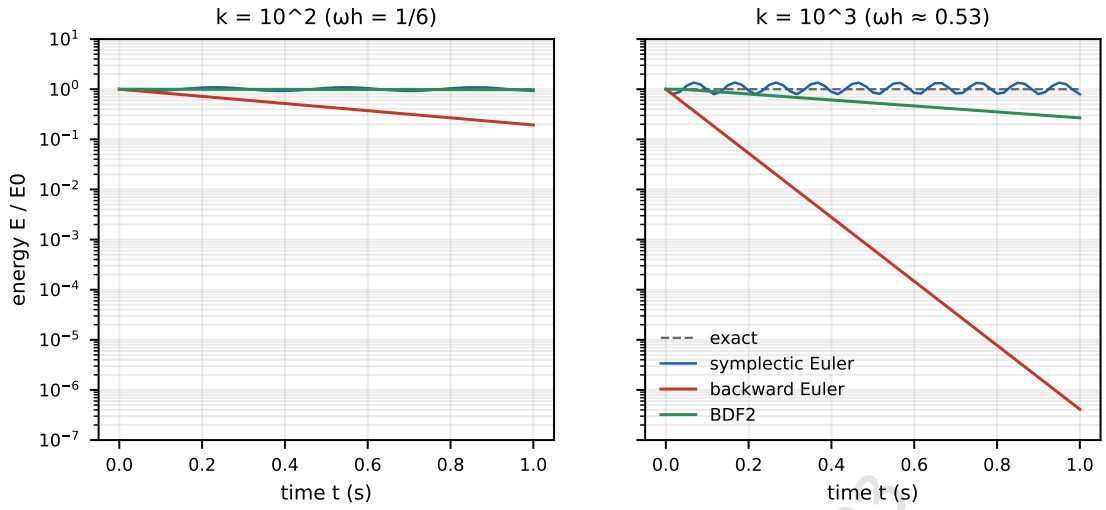
$$E_{j+1} = \frac{E_j}{1 + (\omega h)^2} \qquad \text{สำหรับทุก } j \geq 0.$$

ดังนั้น $E_j = E_0\big(1 + (\omega h)^2\big)^{-j}$ และ E_j ไม่มีวันเพิ่มขึ้น

พิสูจน์ (ของเราเอง).

1. วิธีออยเลอร์ถอยหลังให้ $x_{j+1} = x_j + h v_{j+1}$ และ $v_{j+1} = v_j - h\omega^2 x_{j+1}$ ให้ $s = 1 + (\omega h)^2$ แทน x_{j+1} ลงในสมการที่สองได้ $v_{j+1} = v_j - h\omega^2(x_j + h v_{j+1})$ จึง $s\, v_{j+1} = v_j - h\omega^2 x_j$
2. ดังนั้น $v_{j+1} = (v_j - h\omega^2 x_j)/s$ และ $x_{j+1} = x_j + h\, v_{j+1} = (x_j + h v_j)/s$ โดยใช้ $s - h^2\omega^2 = 1$
3. ยกกำลังสองแล้วรวมกัน ได้ $\omega^2(x_j + h v_j)^2 + (v_j - h\omega^2 x_j)^2 = (1 + h^2\omega^2)(\omega^2 x_j^2 + v_j^2) = s\, E_j$ พจน์กลาง $\pm 2h\omega^2 x_j v_j$ หักล้างกันพอดี

ฉบับร่าง · 2026-10-23



รูปที่ 14.4: พลังงานสัมพัทธ์ E/E_0 ของสปริงมวล 1 กิโลกรัมในหนึ่งวินาทีที่ 60 ก้าวต่อวินาที ซ้าย $k = 10^2$ ($\omega h = 1/6$) ขวา $k = 10^3$ ($\omega h \approx 0.53$) เส้นประคือผลเฉลยจริงที่คงที่ที่ 1 เส้นสีฟ้าคือวิธีออยเลอร์ซิมเพลกติกของหัวข้อ 14.1 ซึ่งแกว่งรอบ 1 แต่ไม่สลายไป เส้นสีแดงคือวิธีออยเลอร์ถอยหลัง และเส้นสีเขียวคือสูตร BDF2 ที่กล่าวถัดไป ใช้ค่าเริ่มต้น $x = 1$ และ $v = 0$ (สร้างจาก `numanal/ch14.py`)

4. หารด้วย s^2 ได้ $\omega^2 x_{j+1}^2 + v_{j+1}^2 = E_j/s$ ซึ่งคือ $E_{j+1} = E_j/s$

สรุป: วิธีออยเลอร์ถอยหลังไม่ระเบิดกับสปริงใดเลย แต่ทุกก้าวคูณพลังงานด้วย $1/(1+(\omega h)^2)$

อ่านแบบเข้าใจง่าย

ให้ $\theta = \omega h$ คือความถี่คูณขนาดก้าว ทุกก้าวพลังงานถูกคูณด้วย $1/(1+\theta^2)$ เมื่อ θ เล็ก ตัวคูณนี้ใกล้ $1-\theta^2$ จึงเสียพลังงานราว θ^2 ต่อเฟรม ที่ $\theta = 1/6$ เสียราว 2.7% ต่อเฟรม และหลังหนึ่งวินาทีเหลือ 19% ถ้า θ ใหญ่ เช่น 1.67 พลังงานเหลือเพียง 26% ในเฟรมเดียว

ในรูปที่ 14.4 แผงซ้ายหลังหนึ่งวินาทีวิธีออยเลอร์ถอยหลังเหลือพลังงานเพียง 0.19 ส่วนแผงขวาเหลือ 4×10^{-7} ในขณะที่ผลเฉลยจริงคงที่ที่ 1 นี่คือราคาของความไม่ระเบิด

สูตร BDF: ใช้ค่าเก่าหลายค่า

วิธีหลายขั้น (multistep method) ใช้ค่าผลเฉลยหลายค่าที่ผ่านมา ไม่ใช่แค่ค่าล่าสุดค่าเดียว Iserles 2009, §2.2, p. 21 เขียนวิธี s ขั้นทั่วไปเป็น

$$\sum_{m=0}^{s} a_m\, y_{n+m} = h \sum_{m=0}^{s} b_m\, f(t_{n+m}, y_{n+m}),$$

โดย a_m และ b_m เป็นค่าคงที่ที่ไม่ขึ้นกับ h และ n ตามธรรมเนียมตั้ง $a_s = 1$ วิธีเป็นชัดแจ้งเมื่อ $b_s = 0$ และเป็นโดยปริยายเมื่อ $b_s \neq 0$ เราเขียน $\rho(w) = \sum_m a_m w^m$ และ $\sigma(w) = \sum_m b_m w^m$

นิยาม 14.2.1: สูตร BDF

วิธีหลายขั้นที่มี s ขั้นและมีอันดับ s เรียกว่า**สูตร BDF** (backward differentiation formula (BDF)) เมื่อ $\sigma(w) = \beta w^s$ สำหรับค่าคงที่ $\beta \neq 0$ Iserles 2009, §2.3, p. 27

ฉบับร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

สูตร BDF ประมาณความชันที่จุดใหม่ด้วยอนุพันธ์ของพหุนามที่ผ่านค่าเก่า s ค่าและค่าใหม่ เพราะ σ มีพจน์เดียวที่จุดใหม่ จึงมีความชันที่จุดใหม่เพียงจุดเดียว ซึ่งอยู่ในวิธีแบบโดยปริยาย

ISE ให้สูตรของ ρ และ β สำหรับ BDF ทุกขั้น Iserles 2009, §2.3, p. 27

$$\beta = \Big(\sum_{m=1}^{s} \tfrac{1}{m}\Big)^{-1}, \qquad \rho(w) = \beta \sum_{m=1}^{s} \tfrac{1}{m}\, w^{s-m}(w-1)^{m}.$$

เมื่อ $s = 1$ ได้วิธีออยเลอร์ถอยหลัง เมื่อ $s = 2$ ได้ $\beta = \tfrac{2}{3}$ และ

$$y_{n+2} - \tfrac{4}{3} y_{n+1} + \tfrac{1}{3} y_n = \tfrac{2}{3}\, h\, f(t_{n+2}, y_{n+2}),$$

ซึ่งเรียกว่า BDF2 ใช้ค่าเก่าสองค่า คือ y_n และ y_{n+1} และค่าใหม่หนึ่งค่า

ลองเอง: ออยเลอร์ถอยหลังกับก้าวใหญ่

ให้ $y' = -100y$ และ $y(0) = 1$ ถ้า $h = 0.5$ ดังนั้น $z = -50$ และ $y_1 = 1/(1+50) = 1/51 \approx 0.0196$ ค่านี้เป็นบวกและเล็ก ต่างจากกฎเชิงสี่เหลี่ยมคางหมูในหัวข้อ 14.1 ซึ่งได้ -0.923 และเครื่องหมายผิด ลองดูว่าเหตุใด ตัวส่วน $1 - z$ จึงมีส่วนจริงเป็นบวกเสมอเมื่อ $\operatorname{Re} z < 0$

เสถียรภาพของสูตร BDF

เมื่อใช้สูตร BDF กับ $y' = \lambda y$ ผลเฉลยมีรูป w^n โดย w เป็นรากของพหุนามลักษณะเฉพาะ

$$\eta(z, w) = \rho(w) - z\,\sigma(w)$$

Iserles 2009, §4.4, pp. 64--65 ISE แสดงว่าวิธีหลายขั้นเป็น A-เสถียรก็ต่อเมื่อรากทุกตัวของ $\eta(z, \cdot)$ มีขนาดน้อยกว่า 1 สำหรับทุก z ในระนาบซ้าย (Lemma 4.7)

ทฤษฎีบท 14.2.2: สูตร BDF สองขั้นเป็น A-เสถียร

สำหรับ BDF2 พหุนามลักษณะเฉพาะคือ $\eta(z, w) = \big(1 - \tfrac{2}{3}z\big)w^2 - \tfrac{4}{3}w + \tfrac{1}{3}$ และวิธีนี้เป็น A-เสถียร Iserles 2009, §4.4, p. 67

แนวคิดของการพิสูจน์ (อ้างอิง). ISE ใช้เกณฑ์ของ Lemma 4.8 (Iserles 2009, §4.4, p. 67) คือ $b_2 > 0$ (ที่นี่ $b_2 = \tfrac{2}{3}$) และรากของ $\eta(it, \cdot)$ ทุกตัวต้องมีขนาดไม่เกิน 1 สำหรับทุก $t \in \mathbb{R}$ เพื่อตรวจข้อหลัง ISE ใช้เกณฑ์ของคอห์น–ชูร์ (Lemma 4.9) กับพหุนามกำลังสอง $\alpha w^2 + \beta w + \gamma$ โดย $\alpha = 1 - \tfrac{2}{3}it$, $\beta = -\tfrac{4}{3}$ และ $\gamma = \tfrac{1}{3}$ เงื่อนไขที่ต้องตรวจมีสองข้อ ข้อแรกคือ $|\alpha|^2 - |\gamma|^2 = \tfrac{4}{9}(2 + t^2) > 0$ ซึ่งทำให้ $|\alpha| \geq |\gamma|$ ข้อที่สองคือ $\big||\alpha|^2 - |\gamma|^2\big| \geq |\alpha\bar{\beta} - \beta\bar{\gamma}|$ ซึ่งเทียบเท่ากับ $(|\alpha|^2 - |\gamma|^2)^2 - |\alpha\bar{\beta} - \beta\bar{\gamma}|^2 = \tfrac{16}{81}t^4 \geq 0$ ทุก t เงื่อนไขพิเศษ $\alpha = \gamma \neq 0$ ไม่เกิดขึ้นเพราะ $\alpha \neq \gamma$ ทุก t ดังนั้นรากทั้งสองอยู่ในจานหนึ่งหน่วยปิดทุก t การตรวจที่ $z = -10$ ในแบบฝึกหัด 14.2.4 (รากมีขนาด $1/\sqrt{23} \approx 0.2085$) เป็นตัวอย่างหนึ่งจุด ไม่ใช่การพิสูจน์ A-เสถียรทั้งระนาบ

ฉบับร่าง Preprint · 2026-10-23

ตารางที่ 14.2: ความลึกของมือที่กดกำแพง ($F = 100$ นิวตัน เริ่มที่ $x = 0$ และ $v = 0$) เทียบวิธีซิมเพลกติกที่ 60 ก้าวต่อวินาที วิธีซิมเพลกติกที่แบ่งเฟรมเป็น 10 ก้าวย่อย และวิธีออยเลอร์ถอยหลังที่ 60 ก้าวต่อวินาที (สร้างจาก `numanal/ch14.py`)

k (N/m)	ความลึกสถิต F/k	ซิมเพลกติก 60 ต่อวินาที	ซิมเพลกติก 10 ก้าวย่อย	ถอยหลัง 60 ต่อวินาที
10^4	1 cm	สูงสุด 0.0281 m	สูงสุด 0.0200 m	สูงสุด 0.0114 m
10^5	1 mm	ระเบิดที่เฟรม 5	สูงสุด 0.00204 m	สูงสุด 0.00103 m
10^6	0.1 mm	ระเบิดที่เฟรม 3	สูงสุด 0.000281 m	สูงสุด 0.000100 m

เกร็ดเพิ่มเติม: นิยามของ A-เสถียรสำหรับวิธีหลายขั้น

ISE ต้องการให้วิธีหลายขั้นนิ่งสำหรับค่าเริ่มต้น $y_1, \ldots, y_{s-1}$ ทุกค่า ไม่ใช่แค่ค่าที่ดี เพราะค่าเหล่านี้ต้องคำนวณด้วยวิธีอื่น และการรบกวนเล็กน้อยอาจพาผลเฉลยออกไปสู่ความไม่เสถียร Iserles 2009, §4.4, p. 63 นิยามในหัวข้อนี้จึงเป็นนิยามที่เข้มกว่าของหัวข้อ 14.1

ทฤษฎีบท 14.2.3: เงื่อนไขรากและขีดจำกัดของสูตร BDF

พหุนาม ρ ของสูตร BDF s ขั้นปฏิบัติตามเงื่อนไขราก (รากทุกตัวมีขนาดไม่เกิน 1 และรากที่มีขนาด 1 เป็นรากเดี่ยว) และวิธี BDF ตามสูตรพื้นฐานลู่เข้า ก็ต่อเมื่อ $1 \le s \le 6$ Iserles 2009, §2.3, p. 28

ทฤษฎีบทนี้ไม่ได้พิสูจน์ที่นี่ แต่การคำนวณของเราตรงกัน รากของ ρ มีขนาดสูงสุดเป็น 1 สำหรับ $s \le 6$ และเป็น 1.0222 เมื่อ $s = 7$ ซึ่งผิดเงื่อนไขราก สูตร BDF ตั้งแต่ $s = 7$ จึงไม่ลู่เข้า แม้เดินก้าวเล็กลงเรื่อย ๆ

ทฤษฎีบท 14.2.4: ขีดกั้นที่สองของดาห์ลควิสต์

อันดับของวิธีหลายขั้นเชิงเส้น (2.8) ที่เป็น A-เสถียรไม่เกิน 2 Iserles 2009, §4.4, p. 67

เราจะไม่พิสูจน์ทฤษฎีบทนี้ที่นี่ ผู้อ่านดูการพิสูจน์ได้ใน ISE Iserles 2009, §4.4, p. 67 ผลที่ตามมาคือ สูตร BDF ที่มีอันดับ $s \ge 3$ ไม่เป็น A-เสถียร ส่วน BDF1 และ BDF2 เป็น A-เสถียร ISE แสดงรูปของ BDF อันดับ 2 ถึง 5 และระบุว่ามีเพียงอันดับ 2 เท่านั้นที่เป็น A-เสถียร Iserles 2009, §4.4, p. 66

ข้อควรระวัง: A-เสถียรไม่ได้แปลว่าแม่น

BDF2 เสียพลังงานน้อยกว่าวิธีออยเลอร์ถอยหลังมาก แต่ก็ยังเสียอยู่ ที่ $k = 10^2$ ค่า BDF2 ที่หนึ่งวินาทีคือ 0.980 และที่ $k = 10^3$ คือ 0.268 เทียบกับ 1 ที่ถูกต้อง การเลือกวิธีที่ไม่ระเบิดจึงไม่ได้ทำให้คำตอบแม่นไปด้วยเสมอไป

ราคาที่ต้องจ่าย: มือที่กดกำแพง

สถานการณ์สมมติ: ลองนึกถึงตัวละครในเกมที่เอามือกดลงบนกำแพงแข็ง แล้วมือไม่ทะลุเข้าไปในกำแพง (สถานการณ์นี้เป็นของเรา ไม่ได้มาจากเกมใดโดยเฉพาะ) เพื่อคำนวณ เราจำลองกำแพงเป็นสปริงแข็งเชิงเส้น ซึ่งเรียกว่า**วิธีลงโทษ** (penalty method) ให้มือมวล 1 กิโลกรัมถูกผลักเข้ากำแพงด้วยแรงคงที่ $F = 100$ นิวตัน ถ้า x คือความลึกที่มือเข้าไปในกำแพง จะได้ $x'' = (F - kx)/m$ โดย $m = 1$ กิโลกรัม ความลึกสถิตคือ F/k ตารางที่ 14.2 เทียบสามวิธี ทุกค่าเป็นสูงสุดของความลึกหรือค่าสุดท้ายจากโค้ดของเรา

ตามทฤษฎีบท 14.1.3 วิธีซิมเพลกติกที่ 60 ก้าวต่อวินาทีถูกจำกัดขอบเขตเมื่อ $k < 14\,400$ ที่ $k = 10^4$ มือแกว่งรอบความลึกสถิตอย่างนิ่ง แต่ความลึกสูงสุดคือ 0.028 เมตร ซึ่งเกินค่าที่ถูกต้อง 0.02 เมตรราว 40% เพราะ $\omega h = 5/3$ อยู่ใกล้ขีดจำกัด

ฉบับร่าง · Preprint · 2026-10-23

เมื่อ $k = 14\,400$ พอดี เมทริกซ์การก้าวมีค่าลักษณะเฉพาะซ้ำที่ -1 (ร่องรอย -2 และดีเทอร์มิแนนต์ 1 ตามทฤษฎีบท 14.1.3) ทำให้การแกว่งโตเป็นเส้นตรงไม่ใช่เอ็กซ์โพเนนเชียล ได้ความลึกเพิ่มทีละ $1/36$ เมตรทุกสองเฟรม จนได้ 8.33 เมตรหลัง 600 เฟรม และที่ $k = 2 \times 10^4$ กับ 10^5 มือระเบิดที่เฟรมที่ 11 และ 5

วิธีซิมเพลกติกกับก้าวย่อยสิบก้าวต่อเฟรมทำให้ทุกค่ารอด แต่ที่ $k = 10^6$ ค่า $\omega \Delta t = 1.67$ ทำให้ความลึกเกินไปราว 40% จึงรอดแต่ไม่แม่น ส่วนวิธีออยเลอร์ถอยหลังที่ 60 ก้าวต่อวินาทีไม่ระเบิดกับ k ใดเลย ความลึกสูงสุดเกินความลึกสถิตเพียง 14%, 3% และ 0.36% แล้วลู่เข้า F/k มือจึงหยุดที่กำแพงโดยเกินลงไปเล็กน้อยครั้งเดียวแล้วลู่เข้า สำหรับหุ่นที่ต้องแกว่งตามข้อต่อ นี่คือพลังงานที่หายไปซึ่งเราอธิบายไว้ข้างบน

ข้อควรระวัง: แบบจำลองนี้เป็นของเรา

สปริงในแบบจำลองนี้ดึงมือได้ทั้งสองทาง ไม่มีการหลุดจากกำแพง ไม่มีแรงเสียดทาน และไม่มีแรงหน่วง ตัวเลขทั้งหมดมาจากโค้ดของเรา ไม่ได้บอกว่าเอนจินใดใช้วิธีใด

การเลือกวิธีคือการเลือกว่าจะจ่ายที่ใด วิธีชัดแจ้งจ่ายด้วยก้าวที่เล็กลง หรือยอมให้ระเบิด วิธีแบบโดยปริยายจ่ายด้วยพลังงานที่หายไป สำหรับสมการเชิงเส้นอย่างสปริงนี้ การแก้สมการทุกก้าวทำได้ด้วยสูตรปิด แต่กำแพงที่ไม่เป็นเชิงเส้นจะต้องใช้วิธีนิวตันในทุกก้าว

สรุปใจความของหัวข้อ 14.2: ราคาของความไม่ระเบิด

- วิธีออยเลอร์ถอยหลังไม่ระเบิดกับสปริงทุกตัว และคูณพลังงานด้วย $1/(1 + (\omega h)^2)$ ทุกก้าว
- วิธีแบบโดยปริยายต้องแก้สมการทุกก้าว สำหรับสมการเชิงเส้นทำได้ด้วยสูตรปิด
- สูตร BDF2 เป็น A-เสถียรและเสียพลังงานน้อยกว่าออยเลอร์ถอยหลังเมื่อ ωh เล็ก แต่ยังเสียอยู่
- ขีดกั้นที่สองของดาห์ลควิสต์บอกว่า วิธีหลายขั้นที่ A-เสถียรมีอันดับไม่เกิน 2 ดังนั้น BDF ที่ $s \geq 3$ ไม่เป็น A-เสถียร และ BDF ที่ $s \geq 7$ ไม่ลู่เข้า
- มือกดกำแพงแสดงทางเลือกเดียวกัน วิธีชัดแจ้งต้องการก้าวย่อยหรือยอมระเบิด ส่วนวิธีแบบโดยปริยายไม่ระเบิด แต่ทำให้มือหยุดที่กำแพงเร็วและเกินลงไปเพียงเล็กน้อย

อ่านต่อ. หัวข้อ 14.3 ใช้วิธีรุงเง-คุททาแบบโดยปริยาย ซึ่งบางแบบเป็น A-เสถียรที่อันดับสูงกว่า 2 ได้ Iserles 2009, §4.3, p. 63 ส่วนวิธีนิวตันสำหรับสมการไม่เชิงเส้นอยู่ใน Iserles 2009, §7.2, pp. 127--130 และการควบคุมค่าคลาดเคลื่อนของวิธีหลายขั้นอยู่ใน Iserles 2009, §6.2, pp. 107--112

ฉบับร่าง. Preprint · 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 14.2.1 (★, ทำด้วยมือ) ให้ $y' = -50y$ และ $y(0) = 1$ ใช้วิธีออยเลอร์ถอยหลังกับขนาดก้าว $h = 0.05$ คำนวณ y_1 และ y_2 แล้วเทียบกับผลของวิธีออยเลอร์ไปข้างหน้าในแบบฝึกหัด 14.1.2

แบบฝึกหัด 14.2.2 (★, ทำด้วยมือ) ใช้สูตรในทฤษฎีบทของ Iserles 2009, §2.3, p. 27 กับ $s = 2$ หา β และ $\rho(w)$ แล้วตรวจว่า $\rho(1) = 0$ และ $\rho'(1) = \beta$

แบบฝึกหัด 14.2.3 (★★, คำนวณ) สปริงมวล 1 กิโลกรัมมี $k = 100$ นิวตันต่อเมตร เริ่มที่ $x = 1$ และ $v = 0$ ทำด้วยวิธีออยเลอร์ถอยหลังที่ 60 ก้าวต่อวินาที ใช้ทฤษฎีบท 14.2.1 หาจำนวนเฟรมที่น้อยที่สุดซึ่งทำให้พลังงานเหลือไม่เกิน 1% ของพลังงานเริ่มต้น

แบบฝึกหัด 14.2.4 (★★, ทำด้วยมือ) สูตร BDF2 ให้ $\eta(z, w) = \left(1 - \tfrac{2}{3}z\right)w^2 - \tfrac{4}{3}w + \tfrac{1}{3}$ เมื่อ $z = -10$ จงหารากทั้งสองของ $\eta(-10, w) = 0$ และแสดงว่าขนาดของทั้งสองราก (มีค่าเท่ากัน) น้อยกว่า 1

แบบฝึกหัด 14.2.5 (★★, คำนวณ) วิธีอดัมส์-บาชฟอร์ทสองขั้น (ชัดแจ้ง) คือ $y_{n+2} - y_{n+1} = h\left(\tfrac{3}{2}f_{n+1} - \tfrac{1}{2}f_n\right)$ ใช้กับ $y' = \lambda y$ ได้สมการ $w^2 - (1 + \tfrac{3}{2}z)w + \tfrac{1}{2}z = 0$ เมื่อ $z = h\lambda$ หารากเมื่อ $z = -3$ แล้วอธิบายด้วยข้อความในหัวข้อนี้ว่าเหตุใดวิธีนี้จึงไม่เป็น A-เสถียร

แบบฝึกหัด 14.2.6 (★★★, พิสูจน์และคำนวณ)

1. ใช้ทฤษฎีบท 14.2.4 แสดงว่าวิธีหลายขั้นเชิงเส้นที่มีอันดับ 3 ไม่เป็น A-เสถียร
2. สูตร BDF3 ของ Iserles 2009, §2.3, p. 27 มี $\beta = \tfrac{6}{11}$ และ $\rho(w) = w^3 - \tfrac{18}{11}w^2 + \tfrac{9}{11}w - \tfrac{2}{11}$ เมื่อ $z = -10^{-4} + 0.5i$ ให้หารากของ $\eta(z, w) = \rho(w) - z\beta w^3$ ด้วยโปรแกรม แล้วตรวจว่ามีรากที่มีขนาดมากกว่า 1

แบบฝึกหัด 14.2.7 (★★★, คำนวณและตีความ) มือมวล 1 กิโลกรัมกดกำแพงด้วยแรงคงที่ $F = 100$ นิวตัน โดยกำแพงเป็นสปริงแข็งที่ $k = 10^5$ นิวตันต่อเมตร ความลึกที่เข้าไปในกำแพงคือ x และ $m\,x'' = F - kx$ เริ่มที่ $x = 0$ และ $v = 0$ ใช้วิธีออยเลอร์ถอยหลังที่ 60 ก้าวต่อวินาที (หนึ่งก้าวต่อเฟรม)

1. ความลึกสถิต F/k คือเท่าไร
2. คำนวณ x_1 และ x_2 ด้วยมือ โดยใช้สูตรในคำใบ้
3. ความลึก x_2 เกินความลึกสถิตหรือไม่ และเกินกี่เปอร์เซ็นต์

คำใบ้.

> **ตัวช่วยคิด**
>
> ข้อ 2 ใช้ $v_{j+1}\left(1 + h^2 k/m\right) = v_j + hF/m - (hk/m)x_j$ และ $x_{j+1} = x_j + h v_{j+1}$ ซึ่งเป็นสูตรของหัวข้อ 14.2 โดยตรง ข้อ 3 เทียบกับ F/k

ฉบับร่าง · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 14.2.1.

1. ค่า $z = h\lambda = 0.05 \cdot (-50) = -2.5$ วิธีออยเลอร์ถอยหลังให้ $y_{n+1} = y_n/(1-z)$ จึงหารด้วย 3.5 ทุกก้าว
2. $y_1 = 1/3.5 = \frac{2}{7} \approx 0.2857$ และ $y_2 = (\frac{2}{7})^2 = \frac{4}{49} \approx 0.0816$
3. ค่าไม่เปลี่ยนเครื่องหมายและลดลงทุกก้าว ต่างจากวิธีออยเลอร์ไปข้างหน้าในแบบฝึกหัด 14.1.2 ที่ให้ -1.5 และ 2.25 แต่ค่าก็ยังไม่แม่น เพราะผลเฉลยจริงที่ $t = 0.05$ คือ $e^{-2.5} \approx 0.0821$ ทำให้ y_1 คลาดเคลื่อนมาก ซึ่งตรงกับข้อควรระวังในหัวข้อ 14.1

เฉลย 14.2.2.

1. เมื่อ $s = 2$ ได้ $\sum_{m=1}^{2} \frac{1}{m} = 1 + \frac{1}{2} = \frac{3}{2}$ จึง $\beta = \frac{2}{3}$
2. จากสูตร $\rho(w) = \beta \sum_{m=1}^{2} \frac{1}{m} w^{2-m}(w-1)^m$ ได้

$$\rho(w) = \tfrac{2}{3}\Big[w(w-1) + \tfrac{1}{2}(w-1)^2\Big] = \tfrac{2}{3}\Big[\tfrac{3}{2}w^2 - 2w + \tfrac{1}{2}\Big] = w^2 - \tfrac{4}{3}w + \tfrac{1}{3}$$

3. $\rho(1) = 1 - \frac{4}{3} + \frac{1}{3} = 0$ และ $\rho'(w) = 2w - \frac{4}{3}$ จึง $\rho'(1) = \frac{2}{3} = \beta$ เมื่อ $\sigma(w) = \beta w^2$ ได้ $\sigma(1) = \beta$ ดังนั้น $\rho(1) = 0$ และ $\rho'(1) = \sigma(1)$ ซึ่งเป็นเงื่อนไขที่ทำให้วิธีนี้ต้องกัน (consistent) และมีอันดับอย่างน้อยหนึ่ง

เฉลย 14.2.3.

1. ที่ $x_0 = 1$ และ $v_0 = 0$ ได้ $E_0 = \omega^2 x_0^2 + v_0^2 = \omega^2 = 100$ และ $\theta = \omega h = 10 \cdot \frac{1}{60} = \frac{1}{6}$
2. ทฤษฎีบท 14.2.1 ให้ $E_n = E_0(1+\theta^2)^{-n}$ ดังนั้นตัวคูณต่อเฟรมคือ $\dfrac{1}{1+1/36} = \dfrac{36}{37} \approx 0.972973$
3. ต้องการ $(36/37)^n \le 0.01$ ซึ่งเทียบเท่ากับ $n \ge \dfrac{\ln 0.01}{\ln(36/37)} = \dfrac{-4.60517}{-0.027399} \approx 168.07$
4. ตรวจ $n = 168$ ได้ $(36/37)^{168} \approx 0.010021 > 0.01$ และ $n = 169$ ได้ $(36/37)^{169} \approx 0.009751 \le 0.01$ จึงต้องใช้ 169 เฟรม หรือประมาณ 2.8 วินาที

เฉลย 14.2.4.

1. เมื่อ $z = -10$ ได้ $1 - \frac{2}{3}z = 1 + \frac{20}{3} = \frac{23}{3}$ จึง $\eta(-10, w) = \frac{23}{3}w^2 - \frac{4}{3}w + \frac{1}{3} = 0$ คูณด้วย 3 ได้ $23w^2 - 4w + 1 = 0$
2. ดิสคริมิแนนต์คือ $4^2 - 4 \cdot 23 \cdot 1 = 16 - 92 = -76 < 0$ จึงมีรากเชิงซ้อนคู่สังยุค $w = \dfrac{2 \pm i\sqrt{19}}{23}$
3. ขนาดของทั้งสองรากเท่ากันและมีค่า $|w|^2 = \dfrac{4+19}{23^2} = \dfrac{23}{529} = \dfrac{1}{23}$ ดังนั้น $|w| = \dfrac{1}{\sqrt{23}} \approx 0.2085 < 1$

ฉบับร่าง · Preprint 2026-10-23

เฉลย 14.2.5.

1. แทน $f_n = \lambda y_n$ และ $y_n = w^n$ ลงในสูตรได้ $w^2 - w - z\left(\frac{3}{2}w - \frac{1}{2}\right) = 0$ จึง $w^2 - (1 + \frac{3}{2}z)w + \frac{1}{2}z = 0$
2. เมื่อ $z = -3$ ได้ $w^2 + \frac{7}{2}w - \frac{3}{2} = 0$ มีรากคือ $w = \dfrac{-\frac{7}{2} \pm \sqrt{\frac{49}{4} + 6}}{2} = \dfrac{-3.5 \pm 4.2720}{2}$ นั่นคือ $w \approx 0.3860$ และ $w \approx -3.8860$
3. รากหนึ่งมีขนาด $3.886 > 1$ และ $z = -3$ อยู่ในระนาบซ้าย ดังนั้นวิธีนี้ไม่เป็น A-เสถียร เหตุผลทั่วไปคือ วิธีหลายขั้นที่เป็น A-เสถียรต้องมี $b_s > 0$ (ISE Lemma 4.8, p. 67) แต่วิธีอดัมส์-บาชฟอร์ทเป็นวิธีชัดแจ้งจึงมี $b_2 = 0$

เฉลย 14.2.6.

1. ทฤษฎีบท 14.2.4 บอกว่าอันดับสูงสุดของวิธีหลายขั้นเชิงเส้นที่ A-เสถียรคือ 2 ดังนั้นวิธีที่มีอันดับ 3 ต้องไม่เป็น A-เสถียร มิฉะนั้นจะมีอันดับสูงกว่า 2 ซึ่งขัดแย้งกับทฤษฎีบท
2. สำหรับ BDF3 ได้ $\beta = \frac{6}{11}$ และ $\eta(z, w) = \rho(w) - z\beta w^3$ เมื่อ $z = -10^{-4} + 0.5i$ รากที่มีขนาดใหญ่ที่สุดของ $\eta(z, \cdot)$ จากโปรแกรมของเราคือประมาณ 1.0109 ซึ่งมากกว่า 1 ตรงกับข้อ 1 การหารากของพหุนามกำลังสามที่มีค่าเชิงซ้อนด้วยมือทำได้ยาก จึงใช้โปรแกรมแทน

เฉลย 14.2.7.

1. ความลึกสถิตคือ $F/k = 100/10^5 = 10^{-3}$ เมตร หรือ 1 มิลลิเมตร
2. ที่ $h = \frac{1}{60}$ และ $m = 1$ ได้ $hF/m = \frac{5}{3}, h^2k/m = \frac{10^5}{3600} = \frac{250}{9}$ และ $hk/m = \frac{10^5}{60}$ ดังนั้น $1 + h^2k/m = \frac{259}{9}$ ตามสูตร
$$v_{j+1} = \frac{v_j + hF/m - (hk/m)x_j}{1 + h^2k/m}, \qquad x_{j+1} = x_j + hv_{j+1}$$
   - เฟรมแรก ($x_0 = 0$, $v_0 = 0$): $v_1 = \dfrac{5/3}{259/9} = \dfrac{15}{259} \approx 0.057915$ และ $x_1 = \dfrac{v_1}{60} = \dfrac{1}{1036} \approx 0.00096525$ เมตร
   - เฟรมที่สอง: $(hk/m)x_1 = \dfrac{10^5}{60} \cdot \dfrac{1}{1036} \approx 1.608752$
   - จึง $v_2 = \dfrac{0.057915 + 1.666667 - 1.608752}{259/9} \approx \dfrac{0.115830}{28.778} \approx 0.0040248$ และ $x_2 = x_1 + \dfrac{v_2}{60} \approx 0.0010323$ เมตร
3. ความลึก $x_2 \approx 1.0323$ มิลลิเมตร เกินความลึกสถิต 1 มิลลิเมตรอยู่ 0.0323 มิลลิเมตร หรือประมาณ 3.2% เป็นการเกินเล็กน้อยและไม่ระเบิด จากโค้ดของเรา ค่าสูงสุดของความลึกเกิดที่เฟรมที่สองนี้ และความลึกลู่เข้าสู่ 1 มิลลิเมตรในภายหลัง

ร่าง Preprint 2026-10-25

14.3 วิธีรุงเงอ-คุททาแบบโดยปริยาย

สถานการณ์สมมติ: โซ่ของมวลที่ต่อด้วยข้อต่อแข็ง

แขนของหุ่นที่มีข้อต่อหลายข้อเป็นโซ่ของมวลที่ต่อกันเป็นแถว เราลองดูโซ่สี่มวล มวลละ 1 กิโลกรัม มวลแรกผูกกับลำตัวด้วยสปริงอ่อน $k_0 = 10^2$ นิวตันต่อเมตร และข้อต่อระหว่างมวลข้างเคียงเป็นสปริงแข็ง $k = 10^6$ นิวตันต่อเมตร โซ่นี้มีความถี่ตั้งแต่ประมาณ 5 ถึง 1848 เรเดียนต่อวินาที ถ้าคำนวณ 60 ครั้งต่อวินาที (อัตราสมมติ) ก็ต้องรับมือทั้งสองปลาย วิธี RK4 แบบชัดแจ้งระเบิดในก้าวแรก วิธีออยเลอร์ถอยหลังทำให้พลังงานหายไปเกือบหมด ส่วนวิธีเกาส์ไม่ระเบิดและรักษาพลังงานไว้ แต่ตำแหน่งของมวลแรกที่หนึ่งวินาทียังคลาดเคลื่อนราว 21% (ตัวเลขทั้งหมดมาจากโค้ดของเรา โซ่นี้เป็นของเราเอง)

ภาพรวมก่อนอ่าน

วิธีชัดแจ้งต้องการก้าวสั้นกว่า $2\sqrt{2}/\omega_{\max}$ วิธีแบบโดยปริยายที่เป็น A-เสถียรรอดกับโหมดที่สลายไปโดยไม่จำกัดก้าว แต่ทุกก้าวต้องแก้ระบบสมการ บทนี้ศึกษาเฉพาะวิธีเกาส์–เลอฌ็องดร์ วิธีเรเดา (Radau) ซึ่งเป็น IRK อีกตระกูลหนึ่งอยู่นอกขอบเขตของบทนี้ ISE ใช้วิธี Radau IA เป็นตัวอย่างในแบบฝึกหัด 5.3 Iserles 2009, p. 100

ตารางบุชเชอร์และวิธีแบบโดยปริยาย

ในบทที่ 13 วิธีรุงเงอ-คุททาแบบชัดแจ้งคำนวณแต่ละขั้นจากขั้นก่อนหน้า แบบโดยปริยายยอมให้ขั้นพึ่งพากันได้ทุกทาง ดังนั้นต้องแก้ระบบสมการพร้อมกัน Iserles 2009, §3.3, pp. 41--42

นิยาม 14.3.1: วิธีรุงเงอ-คุททาแบบโดยปริยาย

ให้ $A = (a_{ji})$ เป็นเมทริกซ์จริงขนาด $\nu \times \nu$ และ $b = (b_1, \dots, b_\nu)$ เราใช้ข้อตกลงแถว $c_j = \sum_i a_{ji}$ ซึ่งเป็นข้อตกลงที่ทำให้สูตรสะดวก ไม่ใช่เงื่อนไขที่ต้องมี **วิธีรุงเงอ-คุททาแบบโดยปริยาย** (implicit Runge–Kutta (IRK)) คือสูตร

$$\xi_j = y_n + h \sum_{i=1}^{\nu} a_{ji}\, f(t_n + c_i h, \xi_i), \qquad j = 1, \dots, \nu,$$

$$y_{n+1} = y_n + h \sum_{j=1}^{\nu} b_j\, f(t_n + c_j h, \xi_j).$$

Iserles 2009, §3.3, p. 41

ตารางบุชเชอร์ (Butcher tableau) คือตารางที่เก็บ c, A และ b ของวิธีไว้ในที่เดียว ตัวอย่างที่เราจะใช้ตลอดหัวข้อนี้คือวิธีเกาส์สองชั้น ซึ่งมีตารางดังนี้ Iserles 2009, §3.4, p. 47

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

อ่านแบบเข้าใจง่าย

วิธีชัดแจ้งคือกรณี $a_{ji} = 0$ ทุกครั้งที่ $i \geq j$ เราจึงคำนวณ ξ_1, ξ_2, ...ได้ตามลำดับ ส่วนวิธีแบบโดยปริยายมี $a_{ji} \neq 0$ ได้ทั้งเหนือและใต้แนวทแยง ทำให้ ξ_j ทุกตัวต้องแก้พร้อมกัน จึงเป็นระบบ ν ชั้นในทุกก้าว

ฉบับร่าง Preprint · 2026-10-23

นิยาม 14.3.2: วิธีเกาส์–เลอฌ็องดร์

วิธีคอลโลเคชัน (collocation method) คือการหาพหุนามดีกรี ν ที่ผ่านค่าเริ่มต้นและตรงกับ ODE ที่จุด $c_1, \dots, c_\nu$ แล้วใช้พหุนามนั้นก้าวไปข้างหน้า วิธีนี้เป็นวิธี IRK เสมอ Iserles 2009, §3.4, Lemma 3.5, p. 43 ถ้าจุด c_j เป็นราก ของพหุนามเชิงตั้งฉาก (บนช่วง $[0,1]$ ด้วยน้ำหนักหนึ่ง) ระดับ ν วิธีคอลโลเคชันนั้นเรียกว่า **วิธีเกาส์–เลอฌ็องดร์** (Gauss–Legendre method) และจุดเหล่านี้คือ**จุดเกาส์** (Gauss points) Iserles 2009, §3.4, p. 47

ทฤษฎีบท 14.3.1: อันดับของวิธีเกาส์–เลอฌ็องดร์

ให้ $\nu \geq 1$ และให้ $c_1, \dots, c_\nu$ เป็นรากที่ต่างกันทั้งหมดของพหุนามเชิงตั้งฉาก $\tilde{P}_\nu$ ระดับ ν บนช่วง $[0,1]$ เทียบกับน้ำหนัก $\omega(t) \equiv 1$ เมื่อนั้นวิธีเกาส์–เลอฌ็องดร์ ν ชั้นมีอันดับ 2ν Iserles 2009, §3.4, Cor. 3.7.1, p. 47

อ้างอิง. ไม่พิสูจน์ที่นี่ ISE ใช้ทฤษฎีบท 3.7 Iserles 2009, §3.4, p. 46 โดยจุดเกาส์ได้ $m = \nu$

สำหรับ $\nu = 1$ ได้ $c_1 = \frac{1}{2}$ และตารางคือ $\frac{1}{2} \mid \frac{1}{2}$ กับ $b_1 = 1$ นี่คือ**กฎจุดกึ่งกลางแบบโดยปริยาย** (implicit midpoint rule) Iserles 2009, §3.4, p. 47 สำหรับ $\nu = 2$ จุดเกาส์คือ $\frac{1}{2} \mp \frac{\sqrt{3}}{6}$ ตามตารางด้านบน

ฟังก์ชันเสถียรภาพของวิธีแบบโดยปริยาย

เราใช้สมการทดสอบ $y' = \lambda y$ เหมือนหัวข้อ 14.1 ตัวคูณต่อก้าวคือฟังก์ชันเสถียรภาพ $R(z)$ ที่เราเจอแล้ว

ทฤษฎีบท 14.3.2: ฟังก์ชันเสถียรภาพของวิธีรุงเงอ-คุททาแบบโดยปริยาย

ให้วิธีรุงเงอ-คุททาแบบโดยปริยายที่มี ν ชั้น ตามนิยาม 14.3.1 ใช้กับสมการทดสอบ $y' = \lambda y$ โดย $\lambda \in \mathbb{C}$, $y(0) = 1$ และขนาดก้าว $h > 0$ ให้ $z = h\lambda$ และสมมติว่า $I - zA$ ไม่เป็นเอกฐาน เมื่อนั้น $y_n = R(z)^n$ สำหรับทุก $n \geq 0$ โดย

$$R(z) = 1 + z\, b^{\top}(I - zA)^{-1}\mathbf{1}, \qquad \mathbf{1} = (1, \dots, 1)^{\top}.$$

พิสูจน์ (ของเราเอง).

1. ใส่ $f(t,y) = \lambda y$ ลงในนิยาม แล้วคูณด้วย h จะได้ $\xi_j = y_n + z\sum_i a_{ji}\xi_i$ ดังนั้น $(I - zA)\xi = \mathbf{1}\, y_n$ เมื่อ $\xi = (\xi_1, \dots, \xi_\nu)^{\top}$
2. เพราะ $I - zA$ ไม่เป็นเอกฐาน เราได้ $\xi = y_n(I - zA)^{-1}\mathbf{1}$
3. แทนลงในสูตร $y_{n+1} = y_n + h\lambda\sum_j b_j\xi_j$ ได้ $y_{n+1} = y_n + z\, b^{\top}\xi = R(z)\, y_n$
4. เมื่อ $y_0 = 1$ โดยการเหนี่ยวนำ $y_n = R(z)^n$

สรุป: วิธี IRK ทำให้สมการทดสอบเป็นลำดับเรขาคณิตที่อัตราส่วน $R(z)$ เหมือนวิธีออยเลอร์ในหัวข้อ 14.1

เกร็ดเพิ่มเติม

ISE แสดงว่า R เป็นฟังก์ชันตรรกยะ และถ้าวิธีเป็นชัดแจ้ง R เป็นพหุนาม Iserles 2009, §4.3, Lemma 4.1, p. 60 ไม่มีวิธีรุงเงอ-คุททาแบบชัดแจ้งใดเป็น A-เสถียร Iserles 2009, §4.3, Cor., p. 60

ฉบับร่าง · Preprint 2026-10-23

วิธีเกาส์สองชั้น: หา R ด้วยมือ

ตารางของวิธีเกาส์สองชั้นมีโครงสร้างดี ทำให้หา R ได้ด้วยมือ เราเรียกตัวเศษและตัวส่วนว่า

$$N(z) = 1 + \tfrac{1}{2}z + \tfrac{1}{12}z^2, \qquad D(z) = 1 - \tfrac{1}{2}z + \tfrac{1}{12}z^2 = \tfrac{1}{12}\big(z^2 - 6z + 12\big).$$

ทฤษฎีบท 14.3.3: ฟังก์ชันเสถียรภาพของวิธีเกาส์สองชั้น

สำหรับวิธีเกาส์–เลอฌ็องดร์สองชั้น (ตารางด้านบน) ฟังก์ชันเสถียรภาพคือ $R(z) = N(z)/D(z)$ ทุก $z \in \mathbb{C}$ ที่ $D(z) \neq 0$ ซึ่งเทียบเท่ากับ $z \neq 3 \pm i\sqrt{3}$ Iserles 2009, §4.3, p. 62

พิสูจน์ (ของเราเอง).

1. ให้ $\alpha = \frac{1}{4}$, $\beta = \frac{1}{4} - \frac{\sqrt{3}}{6}$ และ $\gamma = \frac{1}{4} + \frac{\sqrt{3}}{6}$ จะได้ $\beta\gamma = \frac{1}{16} - \frac{1}{12} = -\frac{1}{48}$ และ $I - zA = \begin{pmatrix} 1 - \alpha z & -\beta z \\ -\gamma z & 1 - \alpha z \end{pmatrix}$ มีดีเทอร์มิแนนต์ $(1 - \alpha z)^2 - \beta\gamma z^2 = 1 - \frac{1}{2}z + \frac{1}{12}z^2 = D(z)$
2. สูตรผกผันของเมทริกซ์ 2×2 ให้ $\xi_1 = \dfrac{1 - \alpha z + \beta z}{D(z)}$ และ $\xi_2 = \dfrac{1 - \alpha z + \gamma z}{D(z)}$
3. $b^{\top}(I - zA)^{-1}\mathbf{1} = \frac{1}{2}(\xi_1 + \xi_2) = \dfrac{2 - 2\alpha z + (\beta + \gamma)z}{2D(z)}$ เพราะ $\beta + \gamma = \frac{1}{2}$ และ $2\alpha = \frac{1}{2}$ ตัวเศษจึงเป็น 2 ได้ $\dfrac{1}{D(z)}$
4. ดังนั้น $R(z) = 1 + \dfrac{z}{D(z)} = \dfrac{D(z) + z}{D(z)} = \dfrac{N(z)}{D(z)}$

สรุป: สำหรับวิธีเกาส์สองชั้น R คือ N/D และตัวส่วนไม่เป็นศูนย์ในครึ่งระนาบซ้าย

เกร็ดเพิ่มเติม

N/D คือ **ตัวประมาณปาเดอ** (Padé approximant) อันดับ $[2/2]$ ของ e^z ISE ให้สูตรเดียวกันที่หน้า 62 ผลการคำนวณของเราตรงกับตัวประมาณปาเดอ $[\nu/\nu]$ สำหรับ $\nu = 1, 2, 3$ ที่ตัวอย่างสุ่มทั้งหมด

ทฤษฎีบท 14.3.4: วิธีเกาส์สองชั้นเป็น A-เสถียร

ให้ $z \in \mathbb{C}$ ที่ $\operatorname{Re} z < 0$ เมื่อนั้น $|R(z)| < 1$ นอกจากนี้ $|R(it)| = 1$ สำหรับทุก $t \in \mathbb{R}$ ดังนั้นวิธีเกาส์สองชั้นเป็น A-เสถียร

ร่างปริญ · Preprint · 2026-10-23

พิสูจน์ (ของเราเอง).

1. ให้ $x = \operatorname{Re} z$ รากของ D คือ $3 \pm i\sqrt{3}$ ซึ่งมีส่วนจริง $3 > 0$ จึง $D(z) \neq 0$ เมื่อ $x \leq 0$
2. เขียน $N = A + B$ โดย $A = 1 + \frac{1}{12}z^2$ และ $B = \frac{1}{2}z$ จะได้ $D = A - B$
3. ใช้เอกลักษณ์ $|a+b|^2 - |a-b|^2 = 4\operatorname{Re}(a\bar{b})$ ได้ $|N|^2 - |D|^2 = 4\operatorname{Re}\big((1+\frac{1}{12}z^2)\frac{1}{2}\bar{z}\big) = 2x + \frac{1}{6}\operatorname{Re}(z|z|^2) = x\big(2 + \frac{1}{6}|z|^2\big)$
4. ถ้า $x < 0$ วงเล็บเป็นบวก จึง $|N|^2 - |D|^2 < 0$ นั่นคือ $|R(z)| = |N|/|D| < 1$
5. ถ้า $z = it$ ค่าสัมประสิทธิ์เป็นจำนวนจริง ดังนั้น $D(it) = N(-it) = \overline{N(it)}$ จึง $|R(it)| = 1$

สรุป: วิธีเกาส์สองขั้นรอดกับทุกโหมดที่สลายไปโดยไม่จำกัดก้าว และไม่ทำให้โหมดแกว่งโตหรือหายบนแกนจินตภาพ

ทฤษฎีบท 14.3.5: วิธีเกาส์–เลอฌ็องดร์ทุกขั้นเป็น A-เสถียร

ให้ $\nu \geq 1$ วิธีเกาส์–เลอฌ็องดร์ ν ขั้น (นิยาม 14.3.2) เป็น A-เสถียร Iserles 2009, §4.3, Cor., p. 63

แนวคิด. วิธีเกาส์ ν ขั้นมีอันดับ 2ν จึงให้ R เป็นตัวประมาณปาเดอ $\hat{r}_{\nu/\nu}$ ซึ่งเป็น A-เสถียร (ISE Thm 4.5 และ 4.6) Iserles 2009, §4.3, p. 63 เราไม่พิสูจน์รายละเอียดที่นี่

โซ่มวลที่แข็งและวิธีแบบโดยปริยาย

ให้ $x \in \mathbb{R}^4$ เป็นการเลื่อนของมวลทั้งสี่ แล้วสมการของโซ่คือ $x'' = -Kx$ โดย

$$K = \begin{pmatrix} k_0 + k & -k & 0 & 0 \\ -k & 2k & -k & 0 \\ 0 & -k & 2k & -k \\ 0 & 0 & -k & k \end{pmatrix}, \qquad k_0 = 10^2, \quad k = 10^6.$$

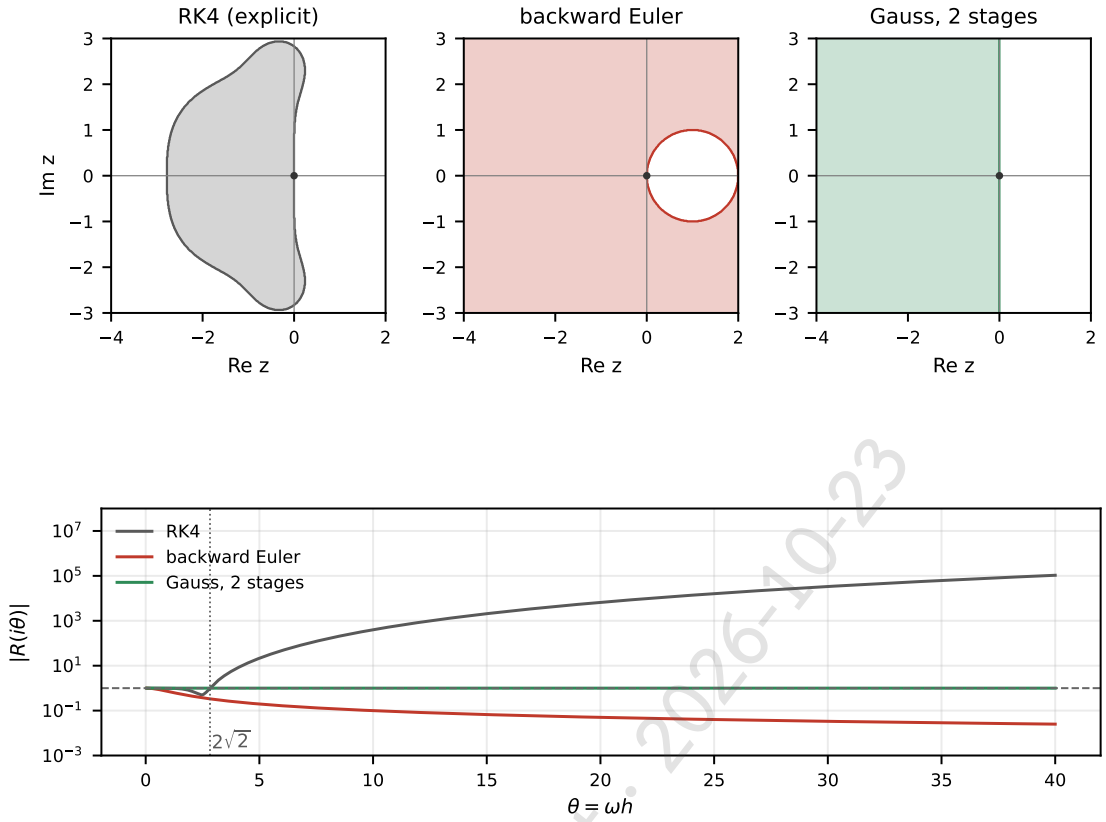
ค่าลักษณะเฉพาะของ K คือ ω^2 ได้ความถี่ $5.000,\ 765.4,\ 1414.2$ และ 1847.8 เรเดียนต่อวินาที เมื่อเขียนเป็นระบบลำดับหนึ่ง $u = (x, v)$ ได้ $u' = Mu$ โดย $M = \begin{pmatrix} 0 & I \\ -K & 0 \end{pmatrix}$ ค่าลักษณะเฉพาะของ M เป็น $\pm i\omega$ ดังนั้นอัตราส่วนความตึง (ตามนิยามในหัวข้อ 14.1) ของโซ่คือ $\omega_{\max}/\omega_{\min} \approx 369.6$ (ตามนิยามในหัวข้อ 14.1)

เมื่อใช้วิธี IRK กับ $u' = Mu$ ค่าขั้นต้องแก้ระบบเชิงเส้นขนาด $\nu d \times \nu d$ โดย $d = 8$ และ $u_{n+1} = u_n + h\sum_j b_j M\xi_j$ วิธีนี้เป็นของเรา ตามนิยาม 14.3.1 ที่ใช้กับระบบเชิงเส้น

ตาราง 14.3 แยกผลของแต่ละโหมดที่ 60 ก้าวต่อวินาที มุมคือค่า $\arg R(i\theta)$ ที่วิธีหมุนโหมดต่อหนึ่งก้าว และมุมจริงคือ θ ปรับให้อยู่ในช่วง $(-\pi, \pi]$

โหมดช้าที่ $\omega = 5.000$ ถูกต้องทุกประการ แต่ที่โหมดแข็ง $\omega = 1847.8$ วิธีเกาส์หมุนไปเพียง -0.3896 เรเดียนต่อก้าว ขณะที่ผลเฉลยจริงหมุน -0.6199 เรเดียน วิธีเกาส์จึงแกว่งโหมดแข็งด้วยความถี่ที่ผิด แม้ขนาดของมันจะคงที่เพราะ $|R(i\theta)| = 1$ ตลอด

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 14.5: บน: โดเมนเสถียรภาพ $\{z : |R(z)| < 1\}$ ของ RK4 (ชัดแจ้ง) ออยเลอร์ถอยหลัง และวิธีเกาส์สองชั้น พื้นที่สีคือ $|R| < 1$ ล่าง: $|R(i\theta)|$ บนแกนจินตภาพเทียบกับ $\theta = \omega h$ เส้นประคือ 1 และเส้นจุดคือ $2\sqrt{2}$ (สร้างจาก `numanal/ch14.py`)

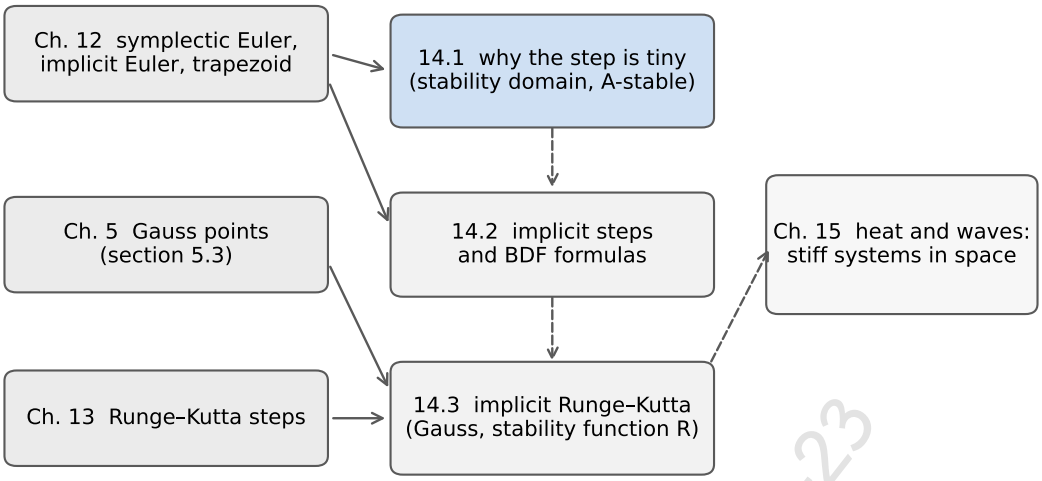
ร่างหนังสือ · Preprint · 2026-10-23

ตารางที่ 14.3: โซ่สี่มวลที่ 60 ก้าวต่อวินาที แยกตามโหมด $\theta = \omega/60$ มุมของวิธีเกาส์ต่างจากมุมจริงที่โหมดแข็ง (สร้างจาก `numanal/ch14.py`และตรวจด้วย `test_ch14.py`)

ω (rad/s)	θ	มุมของเกาส์	มุมจริง	$\lvert R\rvert$ ถอยหลัง	$\lvert R\rvert$ RK4
5.000	0.0833	0.0833	0.0833	0.9965	1.0000
765.4	12.757	−0.9397	0.1902	0.0782	1076
1414.2	23.570	−0.5091	−1.5624	0.0424	1.277×10^{4}
1847.8	30.796	−0.3896	−0.6199	0.0325	3.732×10^{4}

ตารางที่ 14.4: ผลของโซ่สี่มวลหลังหนึ่งวินาทีที่ 60 ก้าวต่อวินาที ผลเฉลยจริง $x_1 = 0.005265$ เมตร (สร้างจาก `numanal/ch14.py`และตรวจด้วย `test_ch14.py`)

วิธี	E/E_0	x_1 (m)	สังเกต
RK4 (ชัดแจ้ง)	ระเบิด	ไม่มีค่า	สถานะพุ่งเป็น 2.3×10^{4} ในก้าวแรก
ออยเลอร์ถอยหลัง	4.1×10^{-6}	0.000553	คลาดเคลื่อน -89%
เกาส์สองชั้น	$1 + 2.5 \times 10^{-14}$	0.006393	คลาดเคลื่อน $+21\%$



รูปที่ 14.6: แผนผังบทที่ 14 ทั้งบท ลูกศรทึบคือแนวคิดที่หัวข้อหนึ่งใช้ ลูกศรประคือแนวคิดที่ต่อจากหัวข้อก่อนหน้า (สร้างจาก `numanal/ch14.py`)

ผลที่ 60 ก้าวต่อวินาที

ตาราง 14.4 เทียบผลหลังหนึ่งวินาทีที่ค่าเริ่มต้น $x = (0.01, 0, 0, 0)$ และ $v = 0$ โดย $E = v^\top v + x^\top K x$ เริ่มที่ 100.01

วิธีเกาส์รักษาพลังงานไว้ได้ถึงระดับการปัดเศษ (ราว 10^{-14}) ส่วนออยเลอร์ถอยหลังทำให้พลังงานหายไปเกือบหมด แต่คำตอบของวิธีเกาส์ที่หนึ่งวินาทียังผิด เพราะโหมดแข็งมีเฟสผิดทุกก้าวและทบกันไปในค่า x_1

ข้อควรระวัง: A-เสถียรไม่ได้แปลว่าแม่น

ที่ 120 ก้าวต่อวินาที x_1 หลังหนึ่งวินาทีคลาดเคลื่อนราว -192% ที่ 600 ก้าวต่อวินาทีราว -173% และที่ 6000 ก้าวต่อวินาที (ค่า $\omega_{\max} h \approx 0.31$) เหลือ 0.27% ผลเหล่านี้เป็นของเรา ความไม่ระเบิดไม่ได้ทำให้ผลแม่น

ราคาของการก้าวหนึ่งครั้ง

วิธีชัดแจ้งไม่แก้ระบบเต็ม ออยเลอร์ถอยหลังแก้ระบบ $d \times d$ ต่อก้าว และวิธีเกาส์ ν ชั้นแก้ระบบ $\nu d \times \nu d$ ต่อก้าว ในโซ่นี้ $d = 8$ จึงได้ 8×8 และ 16×16 ตามลำดับ ราคาจึงเพิ่มตามจำนวนชั้น ISE เห็นว่าเกาส์สามชั้นน่าจะเป็นขนาดสูงสุดที่คุ้มราคา Iserles 2009, §3.4, p. 47

สรุปใจความของหัวข้อ 14.3: เมื่อวิธีต้องแก้ระบบทุกก้าว

- วิธี IRK เขียนด้วยตารางบุชเชอร์ (c, A, b) และค่าชั้นพึ่งพากันได้ทุกทาง
- วิธีเกาส์สองชั้นมีอันดับ 4 และ $|R| < 1$ ในครึ่งระนาบซ้าย จึงไม่ระเบิดกับโซ่ที่ 60 ก้าวต่อวินาที
- แต่โหมดแข็งมีเฟสผิด ตำแหน่งที่หนึ่งวินาทีจึงคลาดเคลื่อนราว 21%

ฉบับร่าง · Preprint · 2026-10-23

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 14.3.1 (⋆, ทำด้วยมือ) ใช้กฎจุดกึ่งกลางแบบโดยปริยาย (ตารางบุชเชอร์ $c_1 = \frac{1}{2}$, $a_{11} = \frac{1}{2}$, $b_1 = 1$) กับ $y' = \lambda y$ หาฟังก์ชันเสถียรภาพ $R(z)$ แล้วคำนวณ $R(-10)$

แบบฝึกหัด 14.3.2 (⋆⋆, ทำด้วยมือ) สำหรับวิธีเกาส์สองชั้น จงคำนวณ $\det(I - zA)$ และแสดงว่าเท่ากับ $1 - \frac{1}{2}z + \frac{1}{12}z^2$ จากนั้นหา $R(-10)$ จากทฤษฎีบท 14.3.3

แบบฝึกหัด 14.3.3 (⋆⋆, ทำด้วยมือ) จากฟังก์ชันเสถียรภาพ $R = N/D$ ของวิธีเกาส์สองชั้น จงคำนวณ $|N(i)|^2$ และ $|D(i)|^2$ แล้วแสดงว่า $|R(i)| = 1$

แบบฝึกหัด 14.3.4 (⋆⋆, ทำด้วยมือและโปรแกรม) วิธี RK4 ชัดแจ้งมีฟังก์ชันเสถียรภาพ $R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$

1. แสดงด้วยมือว่า $|R(2\sqrt{2}\,i)| = 1$
2. ด้วยโปรแกรมของบทนี้ ตรวจว่า $|R(i\theta)| \le 1$ เมื่อ $\theta \le 2\sqrt{2}$ และ $|R(i\theta)| > 1$ เมื่อ $\theta = 2.9$

แบบฝึกหัด 14.3.5 (⋆⋆⋆, พิสูจน์) ให้ $R(z) = N(z)/D(z)$ ของวิธีเกาส์สองชั้น และให้ $S = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix}$ ซึ่งเป็นเมทริกซ์ปฏิสมมาตร

1. แสดงว่า $R(-z)R(z) = 1$ เมื่อ $N(z) \neq 0$ และ $D(z) \neq 0$
2. อธิบายว่าเหตุใด $R(hS)^{\top} = R(-hS)$ และใช้ข้อ 1 แสดงว่า $R(hS)$ เป็นเมทริกซ์เชิงตั้งฉาก
3. จงสรุปว่าเมื่อใช้วิธีเกาส์สองชั้นกับ $u' = Su$ ความยาว $\|u_n\|$ ไม่เปลี่ยนในแต่ละก้าว (แนวทาง: แยก S เป็นค่าลักษณะเฉพาะ $\pm i\omega$ บนจำนวนเชิงซ้อน)

แบบฝึกหัด 14.3.6 (⋆⋆, คำนวณ) โซ่สี่มวลในหัวข้อนี้มีความถี่สูงสุดและต่ำสุด $\omega_{\max} = 1847.8$ และ $\omega_{\min} = 5.000$ เรเดียนต่อวินาที (ค่าจากโค้ดของเรา)

1. หาอัตราส่วนความตึงของโซ่
2. หาก้าวสูงสุด h ที่ RK4 ยังรอดกับโหมดที่เร็วที่สุด และแปลงเป็นจำนวนก้าวต่อวินาที

แบบฝึกหัด 14.3.7 (⋆⋆⋆, คำนวณด้วยโปรแกรมและตีความ) โซ่เดิม เริ่มที่ $x = (0.01, 0, 0, 0)$ และ $v = 0$ ใช้วิธีเกาส์สองชั้นที่ 60 ก้าวต่อวินาที รวม 60 ก้าว (หนึ่งวินาที)

1. หาค่า x_1 หลังหนึ่งวินาที แล้วเทียบกับผลเฉลยจริง 0.0052651 เมตร จงหาค่าคลาดเคลื่อนสัมพัทธ์
2. พลังงาน E เปลี่ยนไปเท่าไรเมื่อเทียบกับค่าเริ่มต้น
3. อธิบายด้วยข้อความในหัวข้อนี้ว่าทำไมวิธีรอดแต่ค่า x_1 ยังผิด

ฉบับร่าง · Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 14.3.1.

1. มีหนึ่งชั้น $\xi = y_n + z\cdot\frac{1}{2}\xi$ จึง $\xi = y_n/(1-\frac{1}{2}z)$ และ $y_{n+1} = y_n + z\,b_1\xi = y_n\bigl(1+\frac{z}{1-z/2}\bigr)$ ดังนั้น
$$R(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}.$$
2. เมื่อ $z=-10$ ได้ $R(-10) = \dfrac{1-5}{1+5} = -\dfrac{2}{3}$ ขนาด $\frac{2}{3}<1$ จึงรอด

เฉลย 14.3.2.

1. จาก $A = \begin{pmatrix} \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4}\end{pmatrix}$ จะได้
$$\det(I-zA) = \bigl(1-\tfrac{z}{4}\bigr)^2 - z^2\Bigl(\tfrac{1}{16}-\tfrac{1}{12}\Bigr) = 1-\tfrac{1}{2}z+\tfrac{1}{16}z^2+\tfrac{1}{48}z^2 = 1-\tfrac{1}{2}z+\tfrac{1}{12}z^2$$
เพราะ $\bigl(\frac{1}{4}-\frac{\sqrt{3}}{6}\bigr)\bigl(\frac{1}{4}+\frac{\sqrt{3}}{6}\bigr) = \frac{1}{16}-\frac{1}{12} = -\frac{1}{48}$
2. เมื่อ $z=-10$ ได้ $\det(I+10A) = 1+5+\frac{100}{12} = \frac{43}{3}$ ซึ่งไม่เป็นศูนย์
3. จากทฤษฎีบท 14.3.3 $R(z) = 1 + \dfrac{z}{D(z)}$ ดังนั้น $R(-10) = 1 - \dfrac{10}{43/3} = 1-\dfrac{30}{43} = \dfrac{13}{43} \approx 0.3023$

เฉลย 14.3.3.

1. $N(i) = 1+\frac{1}{2}i+\frac{1}{12}i^2 = \frac{11}{12}+\frac{1}{2}i$ จึง $|N(i)|^2 = \frac{121}{144}+\frac{36}{144} = \frac{157}{144}$
2. $D(i) = 1-\frac{1}{2}i+\frac{1}{12}i^2 = \frac{11}{12}-\frac{1}{2}i$ จึง $|D(i)|^2 = \frac{157}{144}$ เช่นกัน
3. ดังนั้น $|R(i)| = |N(i)|/|D(i)| = 1$ ค่านี้อยู่บนขอบของโดเมน จึงไม่ใช่ $|R|<1$

เฉลย 14.3.4.

1. ที่ $z = 2\sqrt{2}\,i$ ได้ $z^2 = -8$, $z^3 = -16\sqrt{2}\,i$ และ $z^4 = 64$ จึง
$$R(z) = 1 + 2\sqrt{2}\,i - 4 - \tfrac{8\sqrt{2}}{3}\,i + \tfrac{8}{3} = -\tfrac{1}{3} - \tfrac{2\sqrt{2}}{3}\,i$$
และ $|R|^2 = \frac{1}{9}+\frac{8}{9} = 1$ ดังนั้น $|R(2\sqrt{2}\,i)| = 1$
2. ด้วยโปรแกรม ได้ขอบเขตจริงของ θ เป็น 2.82842712 (ตรงกับ $2\sqrt{2}$) และ $|R| \approx 1.19$ ที่ $\theta = 2.9$ ผลนี้เป็นของเราจากโค้ด

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 14.3.5.

1. เพราะ $N(-z) = D(z)$ และ $D(-z) = N(z)$ จึง $R(-z) = \dfrac{N(-z)}{D(-z)} = \dfrac{D(z)}{N(z)}$ และ $R(-z)R(z) = \dfrac{D(z)}{N(z)} \cdot \dfrac{N(z)}{D(z)} = 1$
2. ค่าสัมประสิทธิ์ของ R จริงและ $(hS)^\top = -hS$ เพราะ S เป็นปฏิสมมาตร จึง $R(hS)^\top = R(-hS)$ ตามข้อ 1 ได้ $R(hS)^\top R(hS) = I$ เมทริกซ์นี้จึงตั้งฉาก
3. แยก $S = V\Lambda V^{-1}$ เหนือจำนวนเชิงซ้อนโดย $\Lambda = \mathrm{diag}(i\omega, -i\omega)$ วิธีเกาส์ในแต่ละพิกัดลักษณะเฉพาะเป็นสมการทดสอบด้วย $z = \pm ih\omega$ ดังนั้น $u_{n+1} = R(hS)u_n$ และจากข้อ 2 ความยาว $\|u_{n+1}\| = \|u_n\|$ ทุกก้าว

เฉลย 14.3.6.

1. อัตราส่วนความตึงคือ $\omega_{\max}/\omega_{\min} = 1847.8/5.000 \approx 369.6$
2. ตามผลในหัวข้อนี้และแบบฝึกหัด 14.3.4 RK4 รอดกับโหมดบนแกนจินตภาพก็ต่อเมื่อ $h\omega_{\max} \le 2\sqrt{2}$ จึง $h_{\mathrm{crit}} = \dfrac{2\sqrt{2}}{1847.8} \approx 1.531 \times 10^{-3}$ วินาที ซึ่งเท่ากับ $1/h_{\mathrm{crit}} \approx 653.3$ ก้าวต่อวินาที

เฉลย 14.3.7.

1. จากโค้ดของเรา $x_1(1) = 0.0063935$ เมตร ผลเฉลยจริงคือ 0.0052651 เมตร ค่าคลาดเคลื่อนสัมพัทธ์คือ $\dfrac{0.0063935}{0.0052651} - 1 \approx +0.214$ หรือประมาณ $+21\%$
2. จากโค้ด $E(1)/E(0) - 1 \approx 2.5 \times 10^{-14}$ ซึ่งเป็นความคลาดเคลื่อนระดับการปัดเศษ
3. วิธีเกาส์มี $|R(i\theta)| = 1$ จึงไม่โตไม่หาย และพลังงานคงที่ (ข้อ 2) แต่ตาราง 14.3 แสดงว่าโหมดแข็ง $\omega = 1847.8$ หมุนผิด (-0.3896 เรเดียนต่อก้าว แทน -0.6199) ค่าเริ่มต้นมีส่วนในโหมดแข็งราว 0.0065, 0.0050 และ 0.0027 เมตร (ตามโค้ด) จึงทำให้ x_1 ผิดแม้พลังงานถูก

สรุปบทที่ 14 ในหน้าเดียว

วิธี	ค่าที่ใช้ต่อก้าว	อันดับ	ฟังก์ชันเสถียรภาพ $R(z)$	สิ่งที่ต้องระวัง
ออยเลอร์ซิมเพลกติก (14.1)	ค่าเดิม, ความเร็วใหม่	1	ดูหัวข้อ 14.1	$\omega h < 2$ สำหรับสปริง
กฎเชิงสี่เหลี่ยมคางหมู (14.1)	โดยปริยาย	2	$\dfrac{1+z/2}{1-z/2}$	A-เสถียร
ออยเลอร์ถอยหลัง (14.2)	โดยปริยาย	1	$\dfrac{1}{1-z}$	A-เสถียร แต่ลดพลังงาน
BDF2 (14.2)	โดยปริยาย, 2 ค่าเก่า	2	จากพหุนาม η	A-เสถียร; BDF ที่ $s \ge 3$ ไม่ A-เสถียร
RK4 (13.1)	ชัดแจ้ง, 4 ชั้น	4	$1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}$	$\omega h < 2\sqrt{2}$ บนแกนจินตภาพ
เกาส์สองชั้น (14.3)	โดยปริยาย, ระบบ $2d$	4	$\dfrac{1+z/2+z^2/12}{1-z/2+z^2/12}$	A-เสถียร, $\lvert R(it)\rvert = 1$

สูตรที่ควรจำ: $h < 2/\omega$ สำหรับสปริงซิมเพลกติก; $h\sqrt{\mu_{\max}} < 2$ สำหรับ $x'' = -Kx$; $R(z) = 1 + z\,b^\top(I - zA)^{-1}\mathbf{1}$; และ $h\omega_{\max} < 2\sqrt{2}$ สำหรับ RK4 บนแกนจินตภาพ

อ่านต่อ. บทที่ 15 ใช้วิธีเส้นกับสมการความร้อนและคลื่น ซึ่งให้ระบบสมการแข็ง ดู Iserles 2009, §3.3--3.4, pp. 41--52 สำหรับวิธี IRK ทั่วไป และ Atkinson 1989, §6.9, pp. 409--418 สำหรับสมการแข็งและวิธีเส้น

ฉบับร่าง · Preprint · 2026-10-23

ภาค VI

ข้างหน้า

ฉบับร่าง · Preprint · 2026-10-23

บทที่ 15

คลื่นบนน้ำและความร้อนที่แผ่ออก

สถานการณ์สมมติ: แตะผิวน้ำหนึ่งครั้งแล้วริ้วคลื่นกระจายออก

ลองนึกถึงฉากน้ำในเกมที่เราแตะผิวน้ำหนึ่งครั้ง แล้วเห็นริ้วคลื่นวงกลมกระจายออกไปทุกทิศ ชนขอบแอ่งแล้วสะท้อนกลับ และค่อย ๆ จางลง คอมพิวเตอร์ไม่ได้วาดริ้วคลื่นนี้จากรูปที่จำไว้ มันเก็บความสูงของผิวน้ำที่จุดกริดหลายพันจุด แล้วคำนวณความสูงใหม่ทุกเฟรม

ในบทนี้เราใช้ตารางขนาด 41×41 จุดภายใน โดยขอบของแอ่งนิ่งเสมอ และตัวเลขทั้งหมดมาจากโค้ดของเรา ไม่ได้บอกว่าเกมใดคำนวณแบบนี้

ถ้าเลือกก้าวเวลาเล็กพอ ริ้วคลื่นจะกระจายและจางอย่างเรียบร้อย แต่ถ้าเพิ่มก้าวเวลาเพียงเล็กน้อย ริ้วคลื่นจะโตขึ้นจนระเบิดทั้งจอภายในห้าสิบก้าว ทั้งที่สมการที่เราเขียนลงไปถูกต้องทุกตัว

คำถามของบทนี้จึงเป็นว่า ก้าวเวลาต้องเล็กเท่าไรริ้วคลื่นจึงไม่ระเบิด เกณฑ์นี้ขึ้นกับขนาดของตารางอย่างไร และเมื่อเราหันไปดูความร้อนที่แผ่ออกและภาพที่เบลอ สิ่งเดียวกันจะปรากฏในรูปอื่นอย่างไร

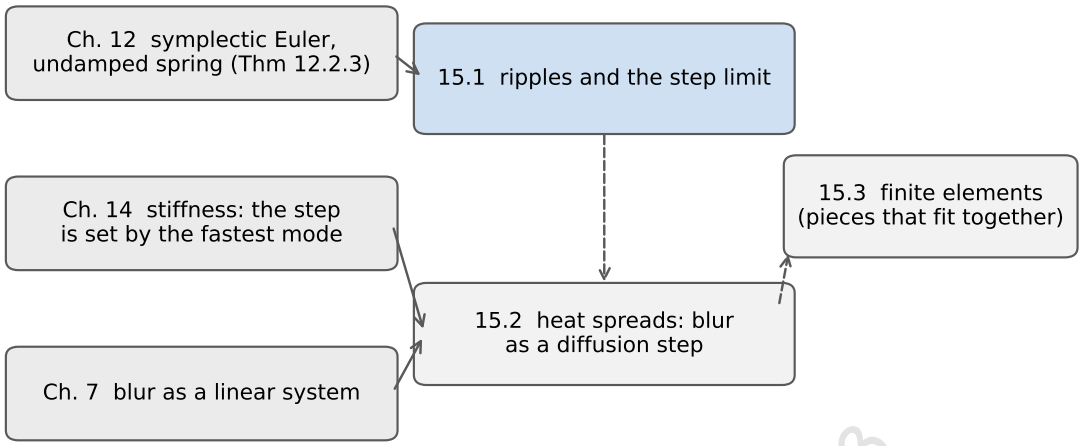
ภาพรวมบท

บทที่ 12 ถึง 14 ดูก้าวเวลาของสปริงและหุ่น ซึ่งมีตัวแปรหนึ่งคือเวลา บทนี้เปลี่ยนไปดูสิ่งที่กระจายอยู่ในที่ว่าง ได้แก่ผิวน้ำและความร้อน ตัวแปรมีทั้งเวลาและตำแหน่ง ทำให้ก้าวของเวลาและก้าวของตำแหน่งต้องเข้ากันพอดี บทนี้มีสามหัวข้อ

- หัวข้อ 15.1 ริ้วคลื่นและก้าวเวลาที่ปลอดภัย เราดูสมการคลื่นและวิธีลีปฟรอก และเกณฑ์ที่ทำให้ก้าวใหญ่เกินไประเบิด
- หัวข้อ 15.2 ความร้อนและภาพเบลอ เราดูว่าการเบลอภาพเป็นการปล่อยให้ความร้อนแผ่ออกไประยะหนึ่ง
- หัวข้อ 15.3 ชิ้นเล็ก ๆ ประกอบกัน เราดูว่าการแบ่งพื้นที่เป็นชิ้นเล็ก ๆ ทำให้แก้ปัญหาได้อย่างไร

เริ่มจากริ้วคลื่นในหัวข้อ 15.1 แล้วจะเจออีกด้านหนึ่งของเรื่องเดียวกันในหัวข้อ 15.2 คือความร้อนที่แผ่ออกซึ่งทำให้ภาพเบลอ

ฉบับร่าง Preprint 2026-10-23



รูปที่ 15.1: แผนผังบทที่ 15 กล่องสีฟ้าคือหัวข้อ 15.1 ลูกศรทึบคือแนวคิดที่หัวข้อหนึ่งใช้ ลูกศรประคือแนวคิดที่ต่อจากหัวข้อก่อนหน้า (สร้างจาก `numanal/ch15.py`)

15.1 ริ้วคลื่นและก้าวเวลาที่ปลอดภัย

สถานการณ์สมมติ: แตะหนึ่งครั้ง แล้วทำไมริ้วคลื่นระเบิด

เราจะดูผิวน้ำบนตารางขนาด 41×41 จุด โดยขอบแอ่งนิ่ง คือความสูงที่ขอบเป็นศูนย์ตลอดเวลา เรียกเงื่อนไขนี้ว่า **เงื่อนไขขอบแบบดีรีเคล** (Dirichlet boundary condition) ที่เวลาศูนย์เราแตะตรงกลางแอ่งด้วยรูปโค้งเรียบที่สูงหนึ่งหน่วย แล้วปล่อยให้ผิวน้ำเคลื่อนที่เอง

แอ่งนี้เป็นแบบจำลองที่เราสร้างขึ้นเอง ไม่ได้วัดจากน้ำจริง และไม่ได้คัดลอกมาจากเกมหรือเอนจินใด ลองนึกภาพอ่างเล็ก ๆ ที่ริ้วคลื่นเดินทางไปถึงขอบแล้วสะท้อนกลับ

ถ้าเลือกก้าวเวลาจนอัตราส่วนก้าวต่อกริด λ (นิยามในหัวข้อนี้) เป็น 0.6 ค่าสูงสุดของ $|h|$ หลัง 300 ก้าวยังเป็นเพียง 0.2729 แต่ที่ $\lambda = 0.8$ ค่าสูงสุดหลัง 10 ก้าวยังเป็น 0.1755 และที่ก้าวที่ 50 ค่าก็เกิน 10^6 (ตัวเลขทั้งหมดมาจากโค้ดของเรา)

คำถามคือ ถ้าเราเลือกก้าวเวลาต่างกันเล็กน้อย ริ้วคลื่นจะเปลี่ยนไปอย่างไรเมื่อเวลาผ่านไป และทำไมก้าวใหญ่ขึ้นนิดเดียวจึงทำให้ค่าโตขึ้นทุกเฟรม

ภาพรวมก่อนอ่าน

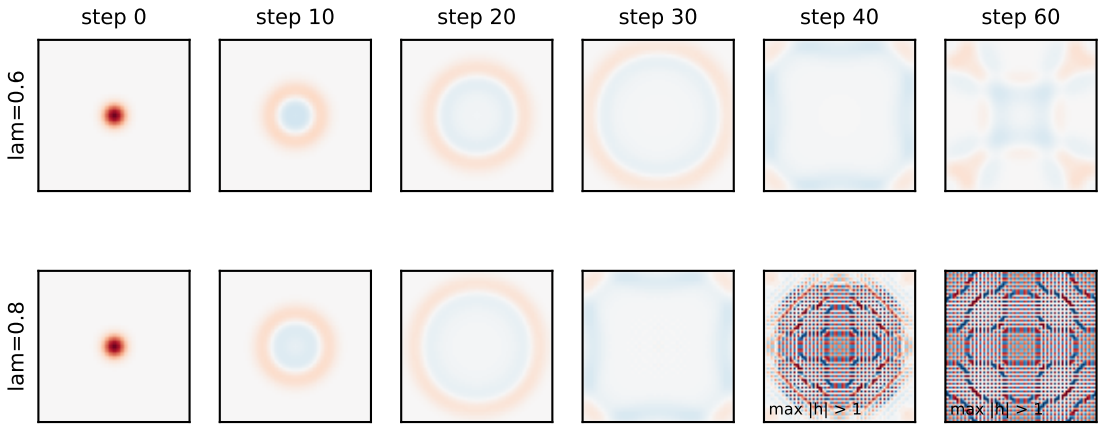
สมการคลื่นบอกว่า ความสูงที่จุดหนึ่งเร่งขึ้นหรือลงตามความโค้งรอบจุดนั้น ถ้าเราแทนอนุพันธ์ด้วยผลต่างบนกริด จะได้สูตรก้าวที่คำนวณได้ตรง ๆ ปัญหาคือก้าวเวลาใหญ่เกินไปทำให้ค่าเก่าของจุดข้างเคียงมีผลต่อค่าใหม่มากเกินกว่าที่สูตรจะรับไหว เกณฑ์ที่เราจะหาคือ ก้าวเวลาต้องสั้นเท่าไรเมื่อขนาดของกริดและจำนวนมิติเป็นค่าหนึ่ง

จากการแตะหนึ่งครั้งถึงสมการคลื่น

ให้ $h(x, y, t)$ เป็นความสูงของผิวน้ำเหนือระดับนิ่ง เมื่อคลื่นมีแอมพลิจูดเล็ก แบบจำลองที่ง่ายที่สุดคือ

$$h_{tt} = c^2 \left(h_{xx} + h_{yy}\right),$$

ฉบับร่าง preprint · 2026-10-23



รูปที่ 15.2: ผิวน้ำหกเฟรมที่ก้าวที่ 0, 10, 20, 30, 40 และ 60 แถวบนคือ $\lambda = 0.6$ ซึ่งริ้วคลื่นกระจายและสะท้อนจากขอบอย่างเรียบร้อย แถวล่างคือ $\lambda = 0.8$ ซึ่งเริ่มเหมือนกันแต่โตจนเกินจอ สีถูกตัดที่ ± 1 และช่องที่มี $\max|h| > 1$ ถูกบอกไว้ (สร้างจาก `numanal/ch15.py`)

โดย c คือความเร็วของคลื่น สมการนี้เรียกว่า**สมการคลื่น** (wave equation) ISE เริ่มจากสมการการพาในหนึ่งมิติ แล้วได้สมการคลื่นจากระบบของการพาสองสมการ Iserles 2009, §17.1, pp. 387--388 ที่นี่เราใช้เวอร์ชันสองมิติซึ่งเป็นการขยายของเรา

เพื่อคำนวณบนตาราง เราแทนค่า h ที่จุด $(i\Delta x, j\Delta x)$ และเวลา $n\Delta t$ ด้วย h_{ij}^{n} และให้ $\lambda = c\,\Delta t/\Delta x$ เรียกว่า**ตัวเลขคูแรนต์** (Courant number) (ISE ใช้ชื่อนี้สำหรับอัตราส่วนเดียวกันในสมการการพา Iserles 2009, §17.2, p. 391) ตัวเลข λ เป็นตัวที่เราปรับได้ ดังนั้นในบทนี้เราตั้ง $c = 1$ และ $\Delta x = 1$ ซึ่งทำให้ λ เท่ากับก้าวเวลาในหน่วยที่เหมาะสม

วิธีลีปฟรอก (**วิธีลีปฟรอก** (leapfrog scheme)) ใช้ค่าความสูงสองก้าวล่าสุดเพื่อหาค่าก้าวใหม่ สูตรในสองมิติคือ

$$h_{ij}^{n+1} = 2h_{ij}^{n} - h_{ij}^{n-1} + \lambda^2\,(Lh^{n})_{ij}, \qquad (Lh)_{ij} = h_{i+1,j} + h_{i-1,j} + h_{i,j+1} + h_{i,j-1} - 4h_{ij}. \tag{15.1}$$

ตัว L คือ**ตัวดำเนินการลาปลาซแบบห้าจุด** (five-point Laplace operator) ซึ่งเป็นผลรวมของเพื่อนบ้านสี่จุดลบ $4h_{ij}$ ขอบของตารางถือว่าเป็นศูนย์ ตามเงื่อนไขขอบแบบดีรีเคล สูตรนี้ต้องการค่าสองค่าเริ่มต้น ในที่นี้ความเร็วเริ่มต้นเป็นศูนย์ เราจึงใช้**ก้าวแรกแบบเทย์เลอร์** (Taylor start) ซึ่งเป็นการกระจายเทย์เลอร์ถึงพจน์ที่สองของเวลา

$$h^{1} = h^{0} + \tfrac{1}{2}\lambda^2\,Lh^{0}.$$

ความผิดพลาดเฉพาะก้าวแรกเป็น $O(\Delta t^{3})$ ซึ่งไม่ทำลายอันดับสองของวิธีทั้งหมด

ลองเอง: หนึ่งจุดบนตาราง 3×3

ให้ตารางมีจุดภายใน 3×3 จุด ความสูงเริ่มต้นเป็นหนึ่งที่จุดกลาง และเป็นศูนย์ที่จุดอื่นทั้งหมด ถ้า $\lambda^2 = 0.4$ จะได้ h^{1} ที่จุดกลางเท่ากับ $1 - 2\lambda^2 = 0.2$ และที่จุดข้างเคียงเท่ากับ $\tfrac{1}{2}\lambda^2 = 0.2$ ต่อไปที่ก้าวที่สอง จุดกลางได้ $2(0.2) - 1 + 0 = -0.6$ เพราะค่า Lh^{1} ที่จุดกลางเป็นศูนย์ และจุดข้างเคียงได้ 0.16 ลองเดาก่อนว่าถ้า $\lambda^2 = 0.6$ จุดกลางจะเป็นบวกหรือลบในก้าวที่สอง แล้วตรวจด้วยสูตรเดียวกัน (ตัวเลขทั้งหมดตรงกับโค้ดของเรา)

ก้าวเวลาที่ปลอดภัยของสมการคลื่น

เพื่อตอบว่าก้าวต้องเล็กเท่าไร เราเขียนสูตร (15.1) ให้เป็นระบบเวกเตอร์ก่อน ให้ u^{m} เป็นเวกเตอร์ของค่า h ทุกจุดที่ก้าวที่ m ให้ $\Omega = -L$ ซึ่งเป็นเมทริกซ์สมมาตร และให้ $\tau = \lambda$ ดังนั้น $\tau^2\Omega = -\lambda^2 L$ และสูตรคือ

$$u^{m+1} = 2u^{m} - u^{m-1} - \tau^2\Omega u^{m}.$$

ฉบับร่าง · Preprint · 2026-10-23

เมทริกซ์ $-L$ มีค่าลักษณะเฉพาะเป็นบวกทั้งหมดเมื่อขอบเป็นศูนย์ ซึ่งเราจะเห็นในทฤษฎีบท 15.1.2 ทฤษฎีบทต่อไปนี้ไม่ขึ้นกับตารางเจาะจง จึงบอกเกณฑ์สำหรับเมทริกซ์สมมาตรใด ๆ

ทฤษฎีบท 15.1.1: เกณฑ์เสถียรภาพของวิธีลีปฟรอก

ให้ Ω เป็นเมทริกซ์จริงสมมาตรที่มีค่าลักษณะเฉพาะเป็นบวกทั้งหมด และให้ $\tau > 0$ วิธีลีปฟรอก $u^{m+1} - 2u^m + u^{m-1} = -\tau^2 \Omega u^m$ มีคำตอบถูกจำกัดขอบเขตสำหรับค่าเริ่มต้น u^0, u^1 ทุกคู่ก็ต่อเมื่อ $\tau\sqrt{\lambda_{\max}(\Omega)} < 2$ ถ้า $\tau\sqrt{\lambda_{\max}(\Omega)} = 2$ จะมีค่าเริ่มต้นที่ทำให้คำตอบโตแบบเชิงเส้น และถ้ามากกว่า 2 จะมีค่าเริ่มต้นที่ทำให้คำตอบโตแบบเอกซ์โพเนนเชียล

พิสูจน์ (ของเราเอง).

1. เนื่องจาก Ω สมมาตร มีการแยกตัว $\Omega = Q\,\mathrm{diag}(\omega_1^2, \ldots, \omega_d^2)\,Q^{\mathsf{T}}$ โดย Q เป็น**เมทริกซ์เชิงตั้งฉาก** (orthogonal matrix) และ $\omega_k > 0$ ให้ $w^m = Q^{\mathsf{T}} u^m$ เนื่องจาก Q ไม่เปลี่ยนนอร์มยูคลิด $\|w^m\| = \|u^m\|$ ดังนั้นพอพิสูจน์สำหรับ w^m

2. สมการของแต่ละองค์ประกอบ w_k เป็นอิสระจากกัน คือ $w_k^{m+1} = (2 - s_k^2) w_k^m - w_k^{m-1}$ โดย $s_k = \tau\omega_k$ สมการลักษณะเฉพาะ $z^2 - (2 - s_k^2)z + 1 = 0$ มีผลคูณของรากเท่ากับ 1 และผลบวกเท่ากับ $2 - s_k^2$

3. ถ้า $0 < s_k < 2$ แล้ว $-2 < 2 - s_k^2 < 2$ รากจึงเป็นจำนวนเชิงซ้อนคู่สังยุคที่มีผลคูณ 1 ดังนั้นทั้งสองรากมีขนาดหนึ่งและต่างกัน คำตอบทั่วไปคือ $w_k^m = 2\,\mathrm{Re}(\alpha_k r_k^m)$ เมื่อ $|r_k| = 1$ จึงมี $|w_k^m| \le 2|\alpha_k|$ สำหรับทุก m ทุกโหมดถูกจำกัดขอบเขต และ $\|w^m\|$ ก็ถูกจำกัดด้วย ผลรวมของจำนวนโหมดจำกัด

4. ถ้า $s_k = 2$ สมการลักษณะเฉพาะคือ $(z+1)^2 = 0$ ดังนั้นคำตอบคือ $(-1)^m(A + Bm)$ เลือกค่าเริ่มต้น $(w_k^0, w_k^1) = (0, 1)$ จะได้ $w_k^m = (-1)^{m+1} m$ ซึ่งไม่ถูกจำกัดขอบเขต เลือกค่าเริ่มต้นเป็นเวกเตอร์ลักษณะเฉพาะที่เข้าโหมด k จะได้ $\|u^m\| = m$

5. ถ้า $s_k > 2$ สมการลักษณะเฉพาะมีรากจริงสองราก r_1, r_2 ที่มีผลคูณ 1 และผลบวก $2 - s_k^2 < -2$ จึงทั้งคู่เป็นลบและหนึ่งในนั้นมีขนาดมากกว่า 1 ให้ $|r_1| > 1$ เลือกค่าเริ่มต้น $w_k^0 = 1$ และ $w_k^1 = r_1$ จะได้ $w_k^m = r_1^m$ ซึ่งโตแบบเอกซ์โพเนนเชียล

สรุป: คำตอบถูกจำกัดขอบเขตสำหรับทุกค่าเริ่มต้นก็ต่อเมื่อทุกโหมดมี $s_k < 2$ นั่นคือ $\tau\sqrt{\lambda_{\max}(\Omega)} < 2$

อ่านแบบเข้าใจง่าย

ความถี่ ω_k ของแต่ละโหมดคูณกับก้าว τ ต้องต่ำกว่าสอง ถ้าเกิน วิธีจะกระโดดข้ามจุดที่ควรอยู่และผลจะสะสมจนโต เกณฑ์นี้ใช้กับทุกโหมด ดังนั้นโหมดที่ถี่ที่สุดเป็นตัวกำหนดก้าวเสมอ

ตารางขนาด 41 คูณ 41 และเกณฑ์ที่คมชัด

บนตารางที่มี $n \times n$ จุดภายในและขอบตรึง ค่าลักษณะเฉพาะของ $-L$ เป็นของชุดรูปคลื่นไซน์ ซึ่งเราจะใช้คำนวณเกณฑ์โดยตรง

ฉบับร่าง Preprint 2026-10-23

ทฤษฎีบท 15.1.2: เกณฑ์ก้าวบนตารางที่ขอบตรึง

ให้ตารางมี $n \times n$ จุดภายใน ขอบตรึงเป็นศูนย์ และให้ $\lambda = c\,\Delta t/\Delta x$ วิธีลีปฟรอกของสมการคลื่นสองมิติบนตารางนี้เสถียร (คำตอบถูกจำกัดขอบเขตสำหรับค่าเริ่มต้นทุกคู่) ก็ต่อเมื่อ

$$\lambda < \lambda_n^{(2)} := \frac{1}{\sqrt{2}\,\cos\!\big(\pi/(2(n+1))\big)}.$$

ในมิติเดียวเกณฑ์คือ $\lambda < \lambda_n^{(1)} := 1/\cos\!\big(\pi/(2(n+1))\big)$ เมื่อ $n \to \infty$ เกณฑ์สองมิติเข้าหา $1/\sqrt{2}$ และเกณฑ์มิติเดียวเข้าหา 1

พิสูจน์ (ของเราเอง).

1. ให้ $\theta_p = p\pi/(n+1)$ และ $v_{pq}(i,j) = \sin(i\theta_p)\sin(j\theta_q)$ สำหรับ $p,q \in \{1,\dots,n\}$ ค่าเป็นศูนย์เมื่อ $i = 0$ หรือ $i = n+1$ เพราะ $\sin(p\pi) = 0$ ดังนั้นสมการขอบตรึงสอดคล้องกับ v_{pq}

2. จากเอกลักษณ์ $\sin((i+1)\theta) + \sin((i-1)\theta) = 2\cos\theta\,\sin(i\theta)$ ได้ $Lv_{pq} = (2\cos\theta_p + 2\cos\theta_q - 4)v_{pq}$ และ $2 - 2\cos\theta = 4\sin^2(\theta/2)$ จึงได้ $-Lv_{pq} = 4\big(\sin^2\tfrac{\theta_p}{2} + \sin^2\tfrac{\theta_q}{2}\big)v_{pq}$

3. เวกเตอร์ v_{pq} มี n^2 ตัว และตั้งฉากกันเป็นคู่ เพราะเป็นฐานของไซน์ จึงเป็นชุดลักษณะเฉพาะครบ ค่าที่ใหญ่ที่สุดเกิดที่ $p = q = n$ คือ $8\sin^2\!\big(\tfrac{n\pi}{2(n+1)}\big) = 8\cos^2\!\big(\tfrac{\pi}{2(n+1)}\big)$

4. ใช้ทฤษฎีบท 15.1.1 กับ $\Omega = -L$ และ $\tau = \lambda$ (ในหน่วยที่ $\Delta x = 1$ และ $c = 1$) ได้ $\tau\sqrt{\lambda_{\max}(\Omega)} = \lambda\sqrt{8}\cos\!\big(\tfrac{\pi}{2(n+1)}\big)$ เกณฑ์ < 2 จึงเทียบเท่ากับ $\lambda < \lambda_n^{(2)}$

5. ในมิติเดียว $-L$ มีค่าลักษณะเฉพาะที่ใหญ่ที่สุด $4\cos^2\!\big(\tfrac{\pi}{2(n+1)}\big)$ การแทนค่าเดียวกันให้ $\lambda \cdot 2\cos\!\big(\tfrac{\pi}{2(n+1)}\big) < 2$ ซึ่งคือ $\lambda < \lambda_n^{(1)}$ (เปลี่ยนเฉพาะจำนวนมิติ)

สรุป: เกณฑ์ขึ้นกับจำนวนจุด และเข้าหาค่าลิมิตเมื่อตารางใหญ่ขึ้น

เกร็ดเพิ่มเติม

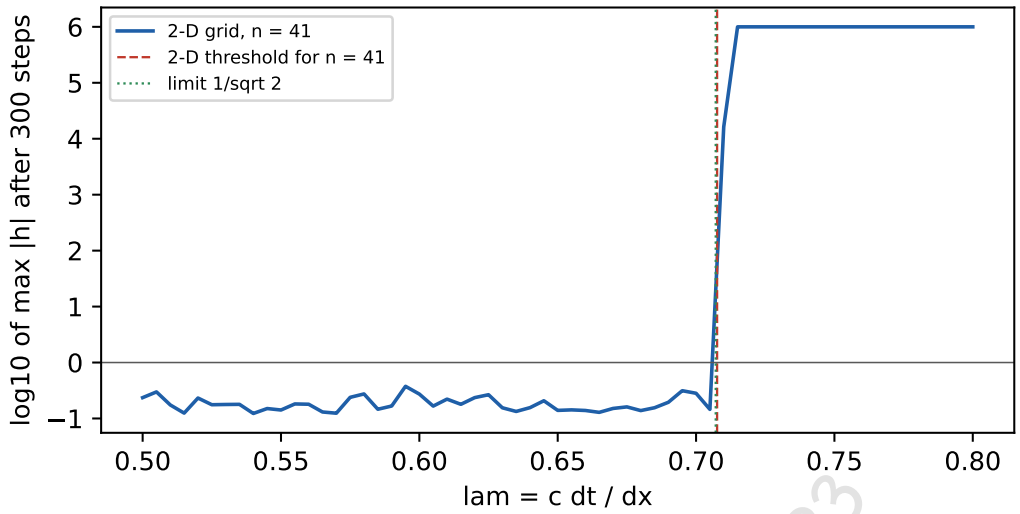
ค่า $\lambda_n^{(2)}$ มากกว่า $1/\sqrt{2}$ เสมอ และลดลงเมื่อ n เพิ่มขึ้น ดังนั้น $\lambda = 1/\sqrt{2}$ รอดบนทุกขนาดของตาราง ส่วนค่าที่มากกว่า $1/\sqrt{2}$ จะระเบิดบนตารางที่ใหญ่พอ ถ้าต้องการให้รอดบนทุกตาราง ต้องเลือก $\lambda \le 1/\sqrt{2}$ เท่านั้น

ตาราง 41×41 ของเรามี $\lambda_{41}^{(2)} = 0.707602$ ซึ่งต่างจาก $1/\sqrt{2} = 0.707107$ เพียงเล็กน้อย ทดลองกับโค้ดได้ดังนี้ ที่ $\lambda = 0.7075$ คำตอบถูกจำกัดขอบเขตตลอด 600 ก้าว ที่ $\lambda = 0.708$ ซึ่งสูงกว่าเกณฑ์ประมาณ 0.0004 ค่าสูงสุดหลัง 600 ก้าวยังเป็นเพียง 2.035 และโตจนเกินภายใน 1500 ก้าว ที่ $\lambda = 0.7085$ ระเบิดภายใน 600 ก้าว (ตัวเลขทั้งหมดมาจากโค้ดของเรา) จะเห็นว่าการโตใกล้เกณฑ์ช้ามาก จึงแยกยากด้วยตาเปล่า

ข้อควรระวัง: กฎมิติเดียวใช้กับสองมิติไม่ได้

ค่า $\lambda = 0.8$ เดียวกันให้ผลต่างกันตามจำนวนมิติ ในมิติเดียว (ตาราง 41 จุดตามแถวกลาง) คำตอบยังถูกจำกัดขอบเขต โดยค่าสูงสุดเป็น 0.3818 หลัง 600 ก้าว แต่บนตาราง 41×41 ค่าเดียวกันระเบิดภายใน 50 ก้าว ขีดจำกัด 1 ของมิติเดียวจึงไม่บอกเกณฑ์ของสองมิติ ต้องใช้ $1/\sqrt{2}$

ฉบับร่าง Preprint · 2026-10-23



รูปที่ 15.3: ลอการิทึมฐาน 10 ของค่า $\max|h|$ หลัง 300 ก้าว เทียบกับ λ บนตาราง 41×41 เส้นประคือเกณฑ์ $\lambda_{41}^{(2)}$ และเส้นจุดคือ $1/\sqrt{2}$ ค่าที่ระเบิดถูกวาดที่ 10^6 (สร้างจาก `numanal/ch15.py`)

การพาไปตามกระแสและการทำอัปวินด์

สมการคลื่นมีอีกมุมหนึ่งคือ**สมการการพา** (advection equation) $u_t + u_x = 0$ ซึ่งกราฟของ u เลื่อนไปทางขวาด้วยความเร็วหนึ่งโดยไม่เปลี่ยนรูปร่าง ค่าใหม่ที่จุดหนึ่งควรขึ้นกับค่าทางซ้ายซึ่งเป็นต้นทางของข้อมูล การเลือกใช้จุดฝั่งต้นทางนี้เรียกว่า**การทำอัปวินด์** (upwinding) ตัวอย่างที่ง่ายที่สุดคือ

$$v_j^{n+1} = \mu\, v_{j-1}^{n} + (1-\mu)\, v_j^{n}, \qquad \mu = \frac{\Delta t}{\Delta x}.$$

ค่าใหม่ของจุด j ขึ้นกับจุด j และ $j-1$ เท่านั้น ISE แสดงว่าวิธีนี้เสถียรเมื่อ $0 < \mu \le 1$ Iserles 2009, §17.2, p. 392 เมื่อ $\mu = 1$ ทุกค่าเลื่อนไปหนึ่งช่องพอดีในแต่ละก้าว จึงเป็นการเลื่อนที่แม่นทุกประการ (ตรวจได้ในแบบฝึกหัด 15.1.4)

ข้อควรระวัง: ค่าลักษณะเฉพาะไม่พอ

ถ้าเมทริกซ์ของวิธีอัปวินด์มี $\mu = 1.5$ ค่าลักษณะเฉพาะทุกตัวเท่ากับ -0.5 ซึ่งอยู่ในวงกลมหนึ่งหน่วย จึงดูเหมือนว่ารอด แต่นอร์มของกำลังของเมทริกซ์โตมาก ค่าสูงสุดของ $\|A^n\|_2$ เมื่อ $n \le 4d$ คือ 4076 ที่ $d = 10$ และ 1.7×10^8 ที่ $d = 20$ ค่าเหล่านี้คำนวณโดยโค้ดของเรา ISE ชี้ว่าการใช้ค่าลักษณะเฉพาะพิสูจน์เสถียรภาพใช้ได้เฉพาะเมื่อเมทริกซ์เป็น**เมทริกซ์ปรกติ** (normal matrix) Iserles 2009, §17.2, p. 393 เมทริกซ์อัปวินด์ไม่ปรกติ ดังนั้นต้องดูนอร์มโดยตรง

ขอบเขตการพึ่งพาและเงื่อนไข CFL

เหตุผลที่ก้าวใหญ่ระเบิดมีชื่อเรียกว่า **เงื่อนไข CFL** (Courant–Friedrichs–Lewy condition) ตามบทความต้นทางของ Courant, Friedrichs และ Lewy Courant, Friedrichs, and Lewy 1928 เราเริ่มจากแนวคิดของ**ขอบเขตการพึ่งพา** (domain of dependence)

นิยาม 15.1.1: ขอบเขตการพึ่งพาเชิงตัวเลข

สำหรับวิธีชัดแจ้งที่คำนวณค่าที่จุด (x_J, t_N) จากค่าที่จุดก่อนหน้า ขอบเขตการพึ่งพาเชิงตัวเลขของจุดนั้นคือชุดของจุดเริ่มต้นทั้งหมดที่ค่าของวิธีอ่าน เมื่อเราย้อนกลับไปทีละก้าวจนถึงเวลาศูนย์

ฉบับร่าง · Preprint · 2026-10-23

อ่านแบบเข้าใจง่าย

ลองนึกว่าเราต้องตอบค่าที่จุดหนึ่งในอนาคต เราย้อนไปดูว่าต้องใช้ค่าเริ่มต้นตรงไหนบ้าง ขอบเขตการพึ่งพาของคำตอบจริงคือจุดเริ่มต้นที่ส่งผลถึงจุดนั้นจริง ๆ ส่วนขอบเขตเชิงตัวเลขคือจุดที่วิธีอ่าน ถ้าจุดจริงอยู่นอกขอบเขตเชิงตัวเลข วิธีไม่มีทางรู้ค่าจริงของจุดนั้น

สำหรับสมการการพา ขอบเขตจริงของ $(\bar{x}, \bar{t})$ คือจุดเดียว $\bar{x} - \bar{t}$ สำหรับสมการคลื่นที่มี $c = 1$ ขอบเขตจริงคือ**กรวยมงจ์** (Monge cone) ซึ่งเป็นสามเหลี่ยมในระนาบ (x, t) ที่มียอดอยู่ที่จุดนั้น และขอบทั้งสองข้างเอียงด้วยความเร็วคลื่น 1 Iserles 2009, §17.4, p. 411 ISE เรียกเงื่อนไขว่า CFL เป็นเงื่อนไขจำเป็นของเสถียรภาพของวิธีชัดแจ้ง Iserles 2009, §17.4, p. 412 เราพิสูจน์ผลที่แข็งกว่าเรื่องการลู่เข้าได้เองสำหรับวิธีอัปวินด์

ทฤษฎีบท 15.1.3: เงื่อนไข CFL จำเป็นต่อการลู่เข้าของวิธีอัปวินด์

ให้ $\mu > 1$ คงที่ ให้ $T > 0$ และให้ g เป็นฟังก์ชันเรียบที่หายไปเมื่อ $x \geq 0$ และมี $g(T/\mu - T) \neq 0$ พิจารณาปัญหาการพา $u_t + u_x = 0$ บนเส้นจริงที่มีค่าเริ่มต้น g และวิธีอัปวินด์บนตารางที่ $\Delta x = T/(\mu N)$ และ $\Delta t = \mu \Delta x$ สำหรับทุก $N \geq 1$ ค่าที่ได้ของวิธีที่จุด $x = T/\mu$ เมื่อถึงเวลา T เป็นศูนย์เสมอ แต่คำตอบจริงคือ $g(T/\mu - T) \neq 0$ ดังนั้นความผิดพลาดเท่ากับ $|g(T/\mu - T)|$ สำหรับทุก N และไม่ลดลงเมื่อ N โต

พิสูจน์ (ของเราเอง).

1. จากสูตรอัปวินด์ ค่า v_j^n ขึ้นกับ v_{j-1}^{n-1} และ v_j^{n-1} โดยการเหนี่ยวนำ v_j^n ขึ้นกับค่าเริ่มต้น v_i^0 เฉพาะเมื่อ $j - n \leq i \leq j$ ดังนั้น v_N^N ขึ้นกับ $v_0^0, \ldots, v_N^0$ เท่านั้น
2. ค่าเริ่มต้น $v_i^0 = g(i\Delta x)$ เป็นศูนย์สำหรับ $i \geq 0$ เพราะ g หายไปเมื่อ $x \geq 0$ ดังนั้น $v_N^N = 0$ สำหรับทุก N
3. คำตอบจริงคือ $u(x, t) = g(x - t)$ เพราะการแทนลงในสมการการพาได้ศูนย์ และตรงตามค่าเริ่มต้น ที่จุด $x = T/\mu$ และเวลา T จึงได้ $u = g(T/\mu - T)$ ซึ่งไม่เป็นศูนย์
4. ความผิดพลาดที่จุดนั้นเท่ากับ $|g(T/\mu - T)|$ ไม่ว่า N จะเป็นเท่าไร ดังนั้นวิธีไม่ลู่เข้าที่จุดนี้

สรุป: เมื่อ $\mu > 1$ ขอบเขตเชิงตัวเลขของวิธีอัปวินด์ไม่ครอบจุดจริงที่ต้องการ

หมายเหตุ ผลนี้ไม่ได้บอกว่าเมื่อ $\mu \leq 1$ วิธีจะลู่เข้า ซึ่งเป็นคนละเรื่อง แนวคิดเดียวกันใช้กับวิธีลีปฟรอกของสมการคลื่น ซึ่ง ISE ระบุเงื่อนไข $\mu \leq 1$ ไว้ Iserles 2009, §17.4, pp. 411--413 แต่เราไม่ได้พิสูจน์ที่นี่ ที่ความถี่ศูนย์ ($\theta = 0$) สมการลักษณะเฉพาะของลีปฟรอกมีรากซ้ำที่ 1 จึงทำให้คำตอบโตแบบเชิงเส้นตามเวลาสำหรับทุก μ ดังนั้นข้อความของ ISE ว่าวิธีเสถียรทุก $0 < \mu \leq 1$ ใช้ได้เฉพาะช่วงเวลาจำกัด และเรื่องนี้ไม่เกิดบนกริดขอบตรึงของเรา เพราะ Ω เป็นบวกแน่นอนตามทฤษฎีบท 15.1.1

ข้อควรระวัง: ก้าวใหญ่ระเบิด ไม่ได้แค่ไม่แม่น

ก้าวใหญ่ที่ทำให้ระเบิดไม่ได้แค่ทำให้คำตอบผิดเล็กน้อย ที่ $\lambda = 0.8$ ค่าความสูงสูงสุดที่ก้าวที่ 10 ยังประมาณ 0.18 แต่ที่ก้าวที่ 50 เกิน 10^6 ดังนั้นการตรวจว่าริ้วคลื่นดูเรียบหรือไม่ ไม่ได้บอกว่าก้าวนั้นปลอดภัย ต้องตรวจเกณฑ์ของทฤษฎีบท 15.1.2 ด้วย

ฉบับร่าง · Preprint · 2026-10-23

ไปต่อ: วิธีกาเลอคินแบบไม่ต่อเนื่อง

อีกแนวทางของการแบ่งพื้นที่เป็นชิ้นเล็ก ๆ คือ**วิธีกาเลอคินแบบไม่ต่อเนื่อง** (discontinuous Galerkin method) ซึ่งฟังก์ชันคำตอบในแต่ละชิ้นเป็นพหุนาม และยอมให้คำตอบกระโดดที่รอยต่อได้ ผู้ที่อยากอ่านต่อเริ่มจากบทที่ 1--2 ของ Hesthaven and Warburton 2008, บทที่ 1--2, pp. 1--42 ส่วนหัวข้อ 15.3 ของบทนี้ให้ภาพพื้นฐานของการประกอบชิ้นเล็ก ๆ ก่อน

สรุปใจความของหัวข้อ 15.1: กลับไปที่ผิวน้ำ

- ริ้วคลื่นเป็นไปตามสมการคลื่น และวิธีลีปฟรอกเดินก้าวด้วยค่าของสองก้าวก่อนหน้า
- คำตอบถูกจำกัดขอบเขตสำหรับทุกค่าเริ่มต้นก็ต่อเมื่อ $\tau\sqrt{\lambda_{\max}(\Omega)} < 2$ โหมดที่ถี่ที่สุดเป็นตัวกำหนดเสมอ
- บนตารางสองมิติที่ขอบตรึง 41×41 เกณฑ์คือ $\lambda < 0.707602$ และขีดจำกัดคือ $1/\sqrt{2}$ กฎของมิติเดียวใช้ไม่ได้
- ก้าวใหญ่ระเบิดได้ภายในไม่กี่สิบก้าว ตัวอย่างของเรา $\lambda = 0.8$ ระเบิดที่ก้าวที่ 50
- วิธีอัปวินด์ที่ $\mu > 1$ ไม่ลู่เข้า เพราะขอบเขตเชิงตัวเลขไม่ครอบจุดจริงที่ต้องการ

อ่านต่อ. สมการคลื่นและสมการการพาอยู่ใน Iserles 2009, §17.1--17.4, pp. 387--412 เรื่องการทำอัปวินด์และเสถียรภาพของผลต่างอยู่ใน Iserles 2009, §17.2, pp. 391--398 ส่วนวิธีกาเลอคินแบบไม่ต่อเนื่องอยู่ใน Hesthaven and Warburton 2008, บทที่ 1--2, pp. 1--42 หัวข้อ 15.2 จะใช้ก้าวแบบโดยปริยายจากบทที่ 14 กับความร้อนที่แผ่ออก

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 15.1.1 (⋆, ทำด้วยมือ) ตารางมิติเดียวมี 9 จุดภายใน ขอบตรึงเป็นศูนย์ และตั้ง $c = 1$ กับ $\Delta x = 1$ หา $\lambda_9^{(1)}$ จากทฤษฎีบท 15.1.2 แล้วตรวจว่า $\lambda = 1$ รอด และ $\lambda = 1.02$ ระเบิด ใช้ค่าเริ่มต้นเป็นรูปโค้งสมมาตรที่ตรงกลางและความเร็วเริ่มต้นเป็นศูนย์ บอกด้วยว่าเหตุใดตารางควรมีจำนวนจุดเป็นเลขคี่

แบบฝึกหัด 15.1.2 (⋆, ทำด้วยมือ) ตารางสองมิติมีจุดภายใน 2×2 จุด ขอบตรึงเป็นศูนย์ และ $\Delta x = 1$

1. หาค่าลักษณะเฉพาะทั้งสี่ตัวของเมทริกซ์ $-L$ ขนาด 4×4
2. หา $\lambda_2^{(2)}$ จากทฤษฎีบท 15.1.2 และเขียนคำตอบโดยไม่มีรากในตัวส่วน
3. ดูจากค่าที่ได้ว่า $\lambda = 0.8$ จะรอดบนตารางนี้หรือไม่

แบบฝึกหัด 15.1.3 (⋆⋆, คำนวณ) ใช้ตารางสามจุดต่อด้าน (จุดภายใน 3×3) ค่าเริ่มต้นเป็นหนึ่งที่จุดกลางและศูนย์ที่จุดอื่น ความเร็วเริ่มต้นเป็นศูนย์ และให้ $\lambda^2 = 0.6$

1. คำนวณค่าที่จุดกลางและจุดข้างเคียงหนึ่งจุดของ h^1 และ h^2 ด้วยมือ
2. ตรวจว่า $\lambda = \sqrt{0.6}$ เกินเกณฑ์ $\lambda_3^{(2)}$ ของตารางนี้หรือไม่
3. ใช้โค้ดหาก้าวแรกที่ $\max|h| > 10$ บอกด้วยว่าผลนี้สอดคล้องกับข้อ 2 อย่างไร

ฉบับร่าง Preprint, 2026-10-23

แบบฝึกหัด 15.1.4 (⋆⋆, พิสูจน์) พิจารณาวิธีอัปวินด์ $v_j^{n+1} = \mu v_{j-1}^n + (1-\mu)v_j^n$ บนเส้นจริง

1. แสดงว่าเมื่อ $\mu = 1$ จะได้ $v_j^n = v_{j-n}^0$ ทุก n และ j (ใช้การเหนี่ยวนำ)
2. เมื่อ $0 \le \mu \le 1$ ให้ A เป็นเมทริกซ์ $d \times d$ ของวิธีนี้ที่มีค่าศูนย์ทางซ้ายของจุดแรก แสดงว่าผลบวกของแถวและผลบวกของหลักมีค่าไม่เกิน 1 ทั้งคู่ แล้วใช้ $\|B\|_2^2 \le \|B\|_1 \|B\|_\infty$ เพื่อสรุปว่า $\|A^n\|_2 \le 1$ ทุก n
3. เมื่อ $\mu = 1.5$ คำนวณ $\|A\|_\infty$ แล้วอธิบายว่าเหตุใดข้อ 2 ใช้ไม่ได้ในกรณีนี้

แบบฝึกหัด 15.1.5 (⋆⋆⋆, พิสูจน์) ให้ $x_{m+1} = (2 - s^2)x_m - x_{m-1}$ เป็นสูตรของวิธีลีปฟรอกสำหรับ $x'' = -\omega^2 x$ เมื่อ $s = \omega \Delta t$

1. เมื่อ $0 < s < 2$ แสดงว่าทั้งสองรากของ $z^2 - (2 - s^2)z + 1 = 0$ มีขนาดเท่ากับหนึ่ง และคำตอบทุกคำตอบถูกจำกัดขอบเขต
2. เมื่อ $s = 2$ แสดงว่าค่าเริ่มต้น $x_0 = 0$ และ $x_1 = 1$ ให้ $x_m = (-1)^{m+1} m$
3. เมื่อ $s = 2.1$ หาสองรากของสมการลักษณะเฉพาะ และให้ r_1 เป็นรากที่มีขนาดมากกว่า 1 แสดงว่าค่าเริ่มต้น $x_0 = 1$ กับ $x_1 = r_1$ ทำให้ x_m โตแบบเอกซ์โพเนนเชียล

คำใบ้. ผลคูณของสองรากเท่ากับ 1 และผลบวกเท่ากับ $2 - s^2$ เมื่อรากเป็นจำนวนจริงลบและมีขนาดไม่เท่ากัน รากหนึ่งต้องมีขนาดมากกว่า 1

แบบฝึกหัด 15.1.6 (⋆⋆, คำนวณและตีความ) ตารางสามมิติมีจุดภายใน $n \times n \times n$ จุด ขอบตรึงเป็นศูนย์ และ $\Delta x = 1$, $c = 1$ ให้ $-L$ เป็นตัวดำเนินการลาปลาซแบบเจ็ดจุดที่มีเครื่องหมายกลับ

1. แสดงว่าค่าลักษณะเฉพาะที่ใหญ่ที่สุดของ $-L$ คือ $12\cos^2\big(\pi/(2(n+1))\big)$ โดยใช้ผลรวมของตัวดำเนินการแบบ Kronecker จากมิติเดียว
2. หาเกณฑ์ $\lambda_n^{(3)} = 1/(\sqrt{3}\cos(\pi/(2(n+1))))$ สำหรับ $n = 4$ และ $n = 10$ แล้วบอกว่าเกณฑ์นี้เข้าหาค่าใด
3. เมื่อ $\lambda = 0.6$ หาจำนวนจุดภายในต่อด้านที่มากที่สุดที่ยังรอด

คำใบ้. ค่าลักษณะเฉพาะของผลบวก Kronecker เท่ากับผลบวกของค่าลักษณะเฉพาะจากแต่ละมิติ

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 15.1.1.

1. ค่าลักษณะเฉพาะที่ใหญ่ที่สุดของ $-L$ บนตาราง 9 จุดคือ $4\cos^2(\pi/20)$ เกณฑ์ของทฤษฎีบท 15.1.2 คือ $\lambda_9^{(1)} = 1/\cos(\pi/20) = 1/0.987688 \approx 1.012465$
2. เมื่อ $\lambda = 1$ ซึ่งน้อยกว่าเกณฑ์ คำตอบถูกจำกัดขอบเขต ในการรันของเรา ค่าสูงสุดหลัง 3000 ก้าวเท่ากับ 1
3. เมื่อ $\lambda = 1.02$ ซึ่งมากกว่าเกณฑ์ คำตอบโตจนเกิน 10^6 ที่ก้าวที่ 89
4. ควรใช้จำนวนจุดเป็นเลขคี่ เพราะโหมดที่ถี่ที่สุดสมมาตรรอบจุดกลางก็ต่อเมื่อ n เป็นเลขคี่ ค่าเริ่มต้นสมมาตรจึงกระตุ้นโหมดนั้นได้ ถ้า n เป็นเลขคู่ การรันจะดูเหมือนรอดแม้ λ เกินเกณฑ์

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 15.1.2.

1. จากสูตร $4(\sin^2\frac{p\pi}{6}+\sin^2\frac{q\pi}{6})$ ได้ $(1,1)\mapsto 2$, $(1,2),(2,1)\mapsto 4$, $(2,2)\mapsto 6$ จึงได้ $2,4,4,6$ (ตรงกับเมทริกซ์ 4×4 โดยตรง)
2. $\lambda_2^{(2)} = 1/(\sqrt{2}\cos(\pi/6)) = 1/(\sqrt{2}\cdot\sqrt{3}/2) = 2/\sqrt{6} = \sqrt{6}/3 \approx 0.8165$
3. $\lambda = 0.8 < 0.8165$ จึงรอดบนตารางนี้ ตามทฤษฎีบท 15.1.2

เฉลย 15.1.3.

1. จาก $h^1 = h^0 + \frac{1}{2}\lambda^2 L h^0$ ได้ $h^1_{\text{กลาง}} = -0.2$ และ $h^1_{\text{ข้างเคียง}} = 0.3$ ต่อไป $h^2_{\text{กลาง}} = 2(-0.2)-1+0.6\cdot 2 = -0.2$ และ $h^2_{\text{ข้างเคียง}} = 2(0.3)+0.6(-1.4) = -0.24$ โดยที่ Lh^1 ที่จุดข้างเคียงคือ $-0.2-4(0.3) = -1.4$
2. $\lambda = \sqrt{0.6} \approx 0.7746$ และ $\lambda_3^{(2)} = 0.7654$ จึง λ เกินเกณฑ์
3. โค้ดให้ก้าวแรกที่ $\max|h| > 10$ คือก้าวที่ 14 ซึ่งสอดคล้องกับข้อ 2 เพราะค่าที่เกินเกณฑ์โตแบบเอกซ์โพเนนเชียล และผ่านค่า 10 ได้ในไม่กี่สิบก้าว

เฉลย 15.1.4.

1. เมื่อ $n=0$ ข้อความเป็นจริง สมมติว่าเป็นจริงถึง n แล้ว ดังนั้น $v_j^{n+1} = \mu v_{j-1}^n + (1-\mu)v_j^n$ เมื่อ $\mu = 1$ เหลือ $v_j^{n+1} = v_{j-1}^n = v_{j-n-1}^0$ ตามสมมติฐาน
2. ทุกสัมประสิทธิ์ของ A ไม่เป็นลบเมื่อ $0 \le \mu \le 1$ ผลบวกของแถวและของหลักไม่เกิน 1 (แถวแรกและหลักสุดท้ายได้ $1-\mu \le 1$ ที่เหลือได้ 1) จึง $\|A\|_1, \|A\|_\infty \le 1$ จาก $\|B\|_2^2 = \rho(B^{\mathsf{T}}B) \le \|B\|_1\|B\|_\infty$ ได้ $\|A^n\|_2 \le 1$
3. เมื่อ $\mu = 1.5$ แถวทั่วไปมีผลบวกของค่าสัมบูรณ์ $|1-\mu|+\mu = 0.5+1.5 = 2$ ดังนั้น $\|A\|_\infty = 2 > 1$ ข้อ 2 ใช้ไม่ได้ เพราะขั้นตอนนั้นพึ่งทั้งความไม่เป็นลบของสัมประสิทธิ์และผลบวกหนึ่ง ซึ่งเป็นจริงเพียงเมื่อ $0 \le \mu \le 1$

เฉลย 15.1.5.

1. เมื่อ $0 < s < 2$ จะได้ $-2 < 2-s^2 < 2$ ดังนั้นสมการ $z^2-(2-s^2)z+1=0$ ดิสคริมิแนนต์เป็นลบ รากจึงเป็นคู่สังยุค r และ $\bar{r}$ โดย $|r|^2 = r\bar{r} = 1$ คำตอบทั่วไปคือ $x_m = \alpha r^m + \bar{\alpha}\bar{r}^m = 2\,\mathrm{Re}(\alpha r^m)$ จึงมี $|x_m| \le 2|\alpha|$ ทุก m
2. เมื่อ $s=2$ สมการลักษณะเฉพาะคือ $(z+1)^2=0$ รากซ้ำคือ -1 และคำตอบทั่วไปคือ $x_m = (-1)^m(A+Bm)$ จาก $x_0=0$ ได้ $A=0$ และ $x_1 = -(A+B) = 1$ ได้ $B=-1$ จึง $x_m = (-1)^{m+1}m$
3. เมื่อ $s = 2.1$ จะได้ $2-s^2 = -2.41$ และรากคือ $r_1 \approx -1.8773$ กับ $r_2 \approx -0.5327$ ซึ่งผลคูณ 1 (ตรวจได้จาก $r_1 r_2 \approx 1.0000$) รากที่มีขนาดมากกว่า 1 คือ r_1 ถ้าเลือก $x_0 = 1$ และ $x_1 = r_1$ แล้วคำตอบคือ $x_m = r_1^m$ ซึ่งมีขนาดเพิ่มขึ้นทุกก้าว

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 15.1.6.

1. ค่าลักษณะเฉพาะของ $-L_1$ คือ $4\sin^2(k\pi/(2(n+1)))$ และของ $-L_3 = T\otimes I\otimes I + I\otimes T\otimes I + I\otimes I\otimes T$ เป็นผลบวกจากทั้งสามมิติ จึงใหญ่ที่สุดเป็น $12\cos^2(\pi/(2(n+1)))$

2. เกณฑ์คือ $\lambda < 1/(\sqrt{3}\cos(\pi/(2(n+1))))$ ได้ 0.607062 เมื่อ $n=4$ และ 0.583287 เมื่อ $n=10$ เข้าหา $1/\sqrt{3}\approx 0.577350$

3. $\lambda = 0.6 < \lambda_n^{(3)}$ ก็ต่อเมื่อ $\cos(\pi/(2(n+1))) < 1/(0.6\sqrt{3}) \approx 0.96225$ ซึ่งให้ $n+1 \le 5$ เมื่อ $n=4$ ได้ $\lambda_4^{(3)} \approx 0.607062 > 0.6$ จึงรอด และเมื่อ $n=5$ ได้ $\lambda_5^{(3)} \approx 0.597717 < 0.6$ จึงไม่รอด ดังนั้นจำนวนจุดภายในต่อด้านที่มากที่สุดคือ 4

15.2 ความร้อนและภาพเบลอ

สถานการณ์สมมติ: เบลองง่าย แต่ทำให้คมคืนยาก

ลองนึกถึงแอปแต่งรูปที่มีแถบเลื่อน "เบลอ" เมื่อลากแถบไปทางขวา ภาพจะนุ่มขึ้นเรื่อย ๆ จุดสว่างเล็ก ๆ กลายเป็นวงจางและหายไปในพื้นหลัง การเบลอทำได้ทันทีและทำซ้ำได้ตามใจ แต่ถ้ากลับแถบเลื่อนเพื่อให้ได้ภาพคมเดิมคืนมา ภาพที่ได้อาจมีจุดรบกวนขึ้นมาได้มาก

ในแอปแบบนี้ การเบลอหนึ่งครั้งทำให้แต่ละพิกเซลรับค่าจากเพื่อนบ้านไปนิดหนึ่ง เมื่อลากแถบซ้ำหลายครั้ง ผลจะแผ่ไปไกลขึ้นทีละน้อย

แอปนี้เป็นตัวอย่างสมมติ เราไม่ได้บอกว่าแอปใดทำงานอย่างไร และไม่ได้อ้างว่าแอปใดใช้สมการความร้อน สิ่งที่เราจะดูคือ การเบลอคล้ายกับการไหลของความร้อน ถ้าเป็นเช่นนั้น ก้าวเวลาของการคำนวณต้องเล็กแค่ไหนจึงไม่ให้ค่าระเบิด และทำไมการกลับทางจึงยาก ตัวเลขทั้งหมดมาจากโค้ดของเรา

ภาพรวมก่อนอ่าน

ความร้อนไหลจากจุดที่ร้อนกว่าไปยังจุดที่เย็นกว่า จนความต่างนั้นเรียบลงเอง การเบลอภาพทำแบบเดียวกันกับความเข้มของพิกเซล แต่ละจุดรับค่าจากเพื่อนบ้านไปนิดหนึ่ง การเบลอหนึ่งรอบจึงเป็นความร้อนที่แผ่ออกไประยะสั้น ๆ ในหัวข้อนี้เราดูสามเรื่อง คือวิธีคำนวณการแผ่ออกบนตาราง เกณฑ์ของก้าวเวลาที่ทำให้คำตอบไม่ระเบิด และเหตุที่การย้อนเวลาทำได้ยาก

จากความร้อนถึงสมการความร้อน

ให้ $u(x,t)$ เป็นค่าที่ต้องการ ณ ตำแหน่ง x และเวลา t เช่นอุณหภูมิของแท่งโลหะ หรือความเข้มของพิกเซลบนแถวหนึ่ง **สมการความร้อน** (heat equation) คือ

$$u_t = u_{xx}, \qquad 0 \le x \le 1, \quad t \ge 0,$$

โดยมีค่าเริ่มต้น $u(x,0) = g(x)$ และค่าที่ขอบ ISE ใช้สมการนี้เป็นตัวอย่างแรกของสมการเชิงวิวัฒนาการ และบอกว่ามันใช้บรรยายการแพร่ของความร้อน รวมถึงการประมวลผลภาพ Iserles 2009, §16.1, p. 350

ในสองมิติ เราเขียนสั้นว่า $u_t = \Delta u$ โดย Δ คือผลรวมของอนุพันธ์อันดับสองสองทิศ ความหมายง่าย ๆ คือ ค่าที่จุดหนึ่งลดลงถ้าสูงกว่าค่าเฉลี่ยของเพื่อนบ้าน และเพิ่มขึ้นถ้าต่ำกว่า จุดร้อนจึงเย็นลงและจุดเย็นจึงอุ่นขึ้น จนทั้งแผ่นเรียบลง

ฉบับร่าง Preprint · 2026-10-23

แบ่งตำแหน่งเป็นจุด แต่เก็บเวลาไว้

วิธีแรกที่เราใช้คือแบ่งเฉพาะตำแหน่ง แล้วปล่อยให้เวลาต่อเนื่อง เรียกว่า **การทำให้เป็นวิยุตบางส่วน** (semi-discretization) ตัวอย่างเช่นแบ่งแกน x เป็นจุด $x_j = jh$ แล้วให้ $v_j(t) \approx u(x_j, t)$ การแทนอนุพันธ์อันดับสองด้วยผลต่างกลางทำให้ได้

$$v_j'(t) = \frac{v_{j-1}(t) - 2v_j(t) + v_{j+1}(t)}{h^2},$$

ซึ่งเป็นระบบของ**สมการเชิงอนุพันธ์สามัญ** (ordinary differential equation) เราแก้ระบบนี้ได้ด้วยวิธีของบทที่ 1 ถึง 7 และ 14 ISE เรียกแนวทางนี้ว่า SD และบอกว่าในตำราเก่าเรียกว่า **วิธีเส้น** (method of lines) แต่ ISE ไม่ใช้ชื่อนั้น Iserles 2009, §16.2, pp. 360--361 การแก้ระบบด้วยวิธีออยเลอร์บนเวลาจะได้วิธีเต็มรูปแบบที่ทั้งตำแหน่งและเวลาถูกแยกออก

ก้าวแบบชัดแจ้งและตัวเลขคูแรนต์

ถ้าใช้วิธีออยเลอร์ไปข้างหน้ากับทุกจุดบนตารางสองมิติ จะได้สูตรต่อไปนี้

นิยาม 15.2.1: วิธีชัดแจ้งของสมการความร้อนบนตารางสองมิติ

ให้ $h = \Delta x$ และ $\mu = \Delta t / h^2$ **วิธีชัดแจ้ง** (explicit) ของสมการความร้อนคือ

$$u_{ij}^{m+1} = u_{ij}^m + \mu\bigl(u_{i+1,j}^m + u_{i-1,j}^m + u_{i,j+1}^m + u_{i,j-1}^m - 4u_{ij}^m\bigr).$$

อัตราส่วน μ เรียกว่า**ตัวเลขคูแรนต์** (Courant number) ตามที่ ISE ใช้กับสมการความร้อน Iserles 2009, §16.1, p. 351

อ่านแบบเข้าใจง่าย

ค่าใหม่ที่จุดหนึ่งเท่ากับค่าเก่า บวกส่วนของความต่างจากเพื่อนบ้านที่ถูกคูณด้วย μ ยิ่ง μ ใหญ่ ความต่างนั้นยิ่งมีผล วิธีนี้คำนวณค่าใหม่ได้ตรง ๆ จากค่าเก่า ไม่ต้องแก้สมการใดในแต่ละก้าว จึงเป็นวิธีชัดแจ้ง ตรงข้ามกับวิธีแบบโดยปริยายที่เราจะดูต่อไป

ข้อควรระวัง: ชื่อเดียวกัน อัตราส่วนต่างกัน

ในหัวข้อ 15.1 เราเรียก $\lambda = c\,\Delta t/\Delta x$ ว่าตัวเลขคูแรนต์ ซึ่งเป็นชื่อที่ ISE ใช้กับสมการการพา ที่นี่ $\mu = \Delta t/(\Delta x)^2$ ก็เรียกว่าตัวเลขคูแรนต์ตามที่ ISE ใช้กับสมการความร้อน สองค่านี้เป็นอัตราส่วนคนละชนิด เกณฑ์ของหัวข้อหนึ่งจึงใช้กับอีกหัวข้อหนึ่งไม่ได้ ให้ดูจากสูตรเสมอ

ลองเอง: แถวห้าจุดทำก้าวเดียว

ให้แถวมีห้าจุด ค่าเริ่มต้นเป็นหนึ่งที่จุดกลางและศูนย์ที่จุดอื่น และค่านอกแถวเป็นศูนย์ ทำหนึ่งก้าว

1. ถ้า $\mu = \frac{1}{4}$ จุดกลางได้ $1 - 2 \cdot \frac{1}{4} = \frac{1}{2}$ และจุดข้างเคียงได้ $\frac{1}{4}$
2. ถ้า $\mu = 0.6$ จุดกลางได้ $1 - 2(0.6) = -0.2$ และจุดข้างเคียงได้ 0.6 ค่ากลางกลายเป็นลบเพราะก้าวใหญ่เกินจนค่าข้ามไปอีกด้านของค่าเฉลี่ยของเพื่อนบ้าน

ร่างเพื่อตรวจ · Preprint · 2026-10-23

การเบลอคือความร้อนที่แผ่ออก

ย้อนกลับไปที่หัวข้อ 7.2 ของบทที่ 7 ซึ่งแต่ละจุดของแถวกลายเป็นค่าเฉลี่ยถ่วงน้ำหนักของตัวมันเองกับเพื่อนบ้านสองข้าง โดยใช้น้ำหนัก $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ จุดนอกแถวนับเป็นศูนย์ การเบลอหนึ่งรอบนี้เหมือนกับก้าวแบบชัดแจ้งในนิยาม 15.2.1 บนแถวหนึ่งมิติที่ $\mu = \frac{1}{4}$ ทุกประการ ดังที่เราพิสูจน์ต่อไปนี้

ทฤษฎีบท 15.2.1: การเบลอหนึ่งรอบคือความร้อนหนึ่งก้าว

บนเส้นจำนวนเต็มที่ไม่มีขอบ ให้ $y_j^{(k)}$ เป็นค่าที่จุด j หลังการเบลอ k รอบ โดยเริ่มจากจุดเดียว $y_j^{(0)} = 1$ เมื่อ $j = 0$ และ $y_j^{(0)} = 0$ เมื่อ $j \neq 0$ แล้ว

(ก) หนึ่งรอบเท่ากับหนึ่งก้าวของวิธีชัดแจ้งที่ $\mu = \frac{1}{4}$

(ข) $\sum_j y_j^{(k)} = 1$ และ $\sum_j j\, y_j^{(k)} = 0$

(ค) $\sum_j j^2\, y_j^{(k)} = \frac{k}{2}$

ข้อ (ข) และ (ค) ยังเป็นจริงบนแถว $1, \ldots, n$ ที่มีขอบตรึงเป็นศูนย์ ถ้าจุดเริ่มต้นอยู่ที่ p โดยที่ $k+1 \le p \le n-k$

พิสูจน์ (ของเราเอง).

1. (ก) เขียนการเบลอหนึ่งรอบเป็น $y'_j = \frac{1}{4}y_{j-1} + \frac{1}{2}y_j + \frac{1}{4}y_{j+1}$ เมื่อจัดใหม่จะได้ $y'_j = y_j + \frac{1}{4}(y_{j-1} - 2y_j + y_{j+1})$ ซึ่งคือวิธีชัดแจ้งที่ $\mu = \frac{1}{4}$ ด้วยผลต่างกลางแบบสามจุด

2. (ข) เลื่อนดัชนีในผลรวม: $\sum_j y'_j = \frac{1}{4}\sum_j y_j + \frac{1}{2}\sum_j y_j + \frac{1}{4}\sum_j y_j = \sum_j y_j$ จึงรักษาผลรวม ส่วนค่าเฉลี่ย $\sum_j j y'_j = \sum_i y_i\big[\frac{1}{4}(i+1) + \frac{1}{2}i + \frac{1}{4}(i-1)\big] = \sum_i i\, y_i$ จึงรักษาค่าเฉลี่ยด้วย เริ่มจาก $\sum_j j\, y_j = 0$ จึงได้ผลตามที่ต้องการ

3. (ค) ให้ $M_k = \sum_j j^2 y_j^{(k)}$ จากการเลื่อนดัชนีเดียวกันได้ $M_{k+1} = \sum_i y_i\big[\frac{1}{4}(i+1)^2 + \frac{1}{2}i^2 + \frac{1}{4}(i-1)^2\big] = \sum_i y_i\,(i^2 + \frac{1}{2}) = M_k + \frac{1}{2}\sum_i y_i = M_k + \frac{1}{2}$ เพราะ $\sum_i y_i = 1$ เมื่อ $M_0 = 0$ จึงได้ $M_k = k/2$

4. (ง) บนแถวที่มีขอบ หลังทุกรอบจนถึงรอบที่ k ค่าที่ไม่เป็นศูนย์อยู่ในช่วง $[p-k,\, p+k] \subseteq [1, n]$ เพราะ $k+1 \le p \le n-k$ จึงไม่มีการแตะขอบ และข้อ (ข) (ค) ยังใช้ได้เหมือนบนเส้นที่ไม่มีขอบ

สรุป: การเบลอ k รอบคือความร้อน k ก้าวที่ $\mu = \frac{1}{4}$ และความแปรปรวนของผลคือ $k/2$

หัวข้อ 7.2 ให้ความกว้างการเบลอ $\sigma = \sqrt{k/2}$ พิกเซลหลัง k รอบ ผลข้อ (ค) ตรงกับนั้นพอดี ในทางกลับกัน สมการความร้อนหนึ่งมิติที่เริ่มจากจุดเดียวมีผลเฉลยเป็นเกาส์เซียนซึ่งความแปรปรวนคือ $2t$ (เป็นข้อเท็จจริงมาตรฐานที่ตรวจได้โดยการหาอนุพันธ์ ไม่ได้ใช้ในการพิสูจน์ของเรา) เมื่อเวลา t มีค่า $k/4$ จึงได้ $2t = k/2$ ตรงกัน ดังนั้นการเบลอสามรอบคือความร้อนที่เวลา $\frac{3}{4}$

ผลนี้ไม่ได้เท่ากันพอดี ค่ากลางหลังสามรอบคือ $\binom{6}{3}/4^3 = 0.3125$ ขณะที่เกาส์เซียนที่มีความแปรปรวน $\frac{3}{2}$ มีค่ากลาง 0.3257 ผลของการเบลอเป็นการแจกแจงทวินาม ซึ่งใกล้เกาส์เซียนเมื่อ k ใหญ่ แต่ไม่ใช่เกาส์เซียนพอดี

เกณฑ์ของก้าวแบบชัดแจ้ง

เราจะถามว่า μ ต้องเล็กเท่าไรจึงไม่ระเบิด เช่นเดียวกับหัวข้อ 15.1 เราใช้ค่าลักษณะเฉพาะ

ฉบับร่าง Preprint 2026-10-23

ทฤษฎีบท 15.2.2: เกณฑ์ของวิธีชัดแจ้งบนตารางที่ขอบตรึง

ให้ตาราง $n \times n$ จุดภายในมีขอบตรึงเป็นศูนย์ วิธีชัดแจ้งในนิยาม 15.2.1 ถูกจำกัดขอบเขต (คำตอบถูกจำกัดขอบเขตสำหรับค่าเริ่มต้นทุกค่า) ก็ต่อเมื่อ

$$\mu \le \mu_n := \frac{1}{4\cos^2\!\big(\pi/(2(n+1))\big)}.$$

สำหรับแถวหนึ่งมิติของ n จุด เกณฑ์คือ $\mu \le 1/\big(2\cos^2(\pi/(2(n+1)))\big)$ เมื่อ $n \to \infty$ เกณฑ์สองมิติเข้าหา $\frac{1}{4}$ และเกณฑ์หนึ่งมิติเข้าหา $\frac{1}{2}$

พิสูจน์ (ของเราเอง).

1. ให้ $\theta_p = p\pi/(n+1)$ และ $v_{pq}(i,j) = \sin(i\theta_p)\sin(j\theta_q)$ สำหรับ $p,q \in \{1,\dots,n\}$ ตามที่ทำในทฤษฎีบท 15.1.2 เวกเตอร์เหล่านี้ตั้งฉากกันเป็นคู่ มี n^2 ตัว จึงเป็นฐาน และ $-Lv_{pq} = \kappa_{pq} v_{pq}$ เมื่อ $\kappa_{pq} = 4\big(\sin^2\frac{p\pi}{2(n+1)} + \sin^2\frac{q\pi}{2(n+1)}\big)$
2. วิธีชัดแจ้งเป็นเชิงเส้น ดังนั้นในแต่ละโหมดของฐานหนึ่งก้าวคูณด้วยตัวคูณ $\rho_{pq} = 1 - \mu\kappa_{pq}$ นั่นคือ $u^m = \sum_{p,q} c_{pq}\rho_{pq}^m v_{pq}$ และเพราะเวกเตอร์ตั้งฉากกัน $\|u^m\|_2^2 = \sum_{p,q} |c_{pq}|^2 \rho_{pq}^{2m}\|v_{pq}\|_2^2$
3. ค่า κ_{pq} ใหญ่ที่สุดที่ $p=q=n$ คือ $\kappa_{\max} = 8\cos^2(\pi/(2(n+1))) = 2/\mu_n$ เพราะ $\sin\frac{n\pi}{2(n+1)} = \cos\frac{\pi}{2(n+1)}$
4. ถ้า $\mu \le \mu_n$ แล้ว $\mu\kappa_{pq} \le \mu\kappa_{\max} \le 2$ ดังนั้น $\rho_{pq} \in [-1,1)$ จึง $|\rho_{pq}| \le 1$ ทุกโหมด และ $\|u^m\|_2 \le \|u^0\|_2$ คำตอบจึงถูกจำกัดขอบเขต
5. ถ้า $\mu > \mu_n$ เลือก $u^0 = v_{nn}$ จะได้ $\rho_{nn} = 1 - \mu\kappa_{\max} < -1$ และ $u^m = \rho_{nn}^m v_{nn}$ โตแบบเอกซ์โพเนนเชียล
6. สำหรับแถวหนึ่งมิติใช้ $\kappa_p = 4\sin^2\frac{p\pi}{2(n+1)}$ และ $\kappa_{\max} = 4\cos^2\frac{\pi}{2(n+1)}$ ซึ่งให้เกณฑ์ $1/(2\cos^2(\pi/(2(n+1))))$

สรุป: วิธีชัดแจ้งรอดบนตาราง $n \times n$ ก็ต่อเมื่อ $\mu \le \mu_n$ โดยที่ μ_n ลดลงเข้าหา $\frac{1}{4}$

ตาราง 41×41 ของเรามี $\mu_{41} = 0.250350$ ทดลองกับโค้ดได้ดังนี้ ที่ $\mu = 0.2$ ค่าสูงสุดของ $|u|$ ลดลงจาก 1 ที่ก้าวศูนย์เหลือ 0.3305 ที่ก้าวที่ 20 และ 0.04753 ที่ก้าวที่ 200 ที่ $\mu = 0.3$ ค่าสูงสุดเกิน 1 ตั้งแต่ก้าวที่ 112 และเกิน 10^6 ตั้งแต่ก้าวที่ 157 (ตัวเลขทั้งหมดมาจากโค้ดของเรา) ภาพเปรียบเทียบอยู่ในรูปที่ 15.4

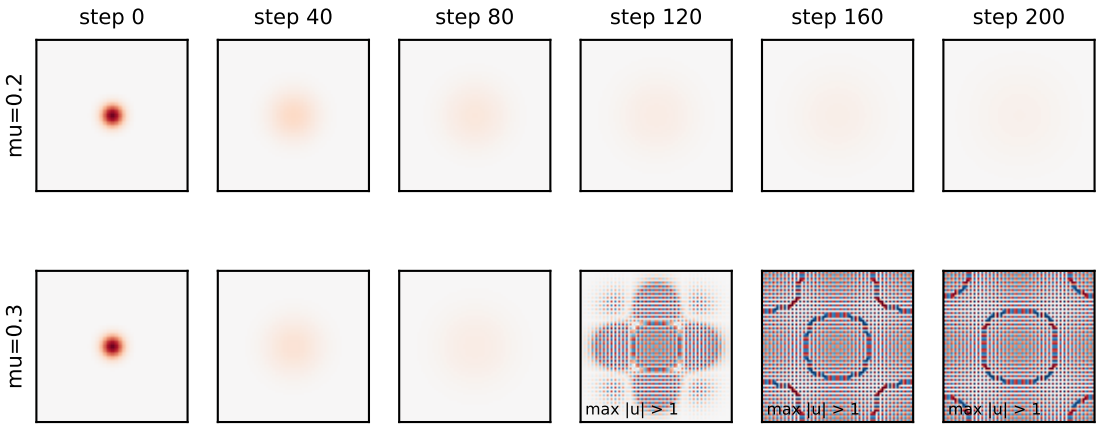
ข้อควรระวัง: ค่าที่ใกล้ $\frac{1}{4}$ ระเบิดช้า

ที่ $\mu = 0.26$ ซึ่งมากกว่าเกณฑ์ของตาราง 41×41 เพียงเล็กน้อย คำตอบยังเล็กมากในช่วงแรก และค่าสูงสุดเกิน 1 ครั้งแรกที่ก้าวที่ 560 เท่านั้น จึงดูเหมือนเรียบร้อยเป็นเวลานาน การดูภาพช่วงแรกไม่ได้บอกว่าก้าวนั้นปลอดภัย ต้องตรวจเกณฑ์ของทฤษฎีบท 15.2.2 ด้วย

เกร็ดเพิ่มเติม (ผู้เรียบเรียง)

ค่า $\mu = \frac{1}{4}$ รอดบนทุกตาราง เพราะ $\mu_n > \frac{1}{4}$ เสมอ แต่ค่า μ ที่มากกว่า $\frac{1}{4}$ เล็กน้อยอาจรอดบนตารางเล็กและระเบิดบนตารางใหญ่ เช่น 0.2503 รอดถึง $n = 44$ เท่านั้น (แบบฝึกหัด 15.2.5) ค่า $\frac{1}{2}$ ที่ ISE ให้สำหรับหนึ่งมิติ Iserles 2009, §16.5, p. 376 เป็นกรณีหนึ่งมิติ ISE บอกว่าการขยายไปหลายมิติทำได้ Iserles 2009, §16.5, p. 377 แต่สูตร μ_n สำหรับสองมิติในที่นี้เป็นการคำนวณของเราเอง

ร่างปริญญา · Preprint · 2026-10-23



รูปที่ 15.4: ค่า u บนตาราง 41×41 ที่ก้าว $0, 40, 80, 120, 160$ และ 200 แถวบนคือ $\mu = 0.2$ ซึ่งเรียบลงและเกือบหายไป แถวล่างคือ $\mu = 0.3$ ซึ่งเริ่มเหมือนกันแต่โตเกินจอ สีถูกตัดที่ ± 1 และช่องที่มี $\max |u| > 1$ ถูกบอกไว้ (สร้างจาก `numanal/ch15.py`)

วิธีแบบโดยปริยายไม่มีเกณฑ์ก้าว

เมื่อก้าวใหญ่กว่าเกณฑ์ วิธีชัดแจ้งจะระเบิด วิธีที่หลบปัญหานี้ได้คือ**วิธีแบบโดยปริยาย** (implicit method) เช่น**วิธีออยเลอร์ถอยหลัง** (backward Euler method) ซึ่งใช้ค่าใหม่ในการคำนวณ จึงได้ระบบสมการเชิงเส้นในแต่ละก้าว

$$(I - \mu L)\, u^{m+1} = u^m.$$

การแก้ระบบนี้ทุกก้าวมีราคา ดูราคาที่ต้องจ่ายในหัวข้อ 14.2 ของบทที่ 14

ทฤษฎีบท 15.2.3: วิธีออยเลอร์ถอยหลังไม่ระเบิดสำหรับทุก μ

บนตาราง $n \times n$ ที่ขอบตรึง ให้ $\mu > 0$ ใด ๆ วิธีออยเลอร์ถอยหลังให้ $\|u^{m+1}\|_2 \le \|u^m\|_2$ ทุก m ดังนั้นคำตอบถูกจำกัดขอบเขตด้วย $\|u^0\|_2$ สำหรับทุกค่าเริ่มต้นและทุก $\mu > 0$

พิสูจน์ (ของเราเอง).

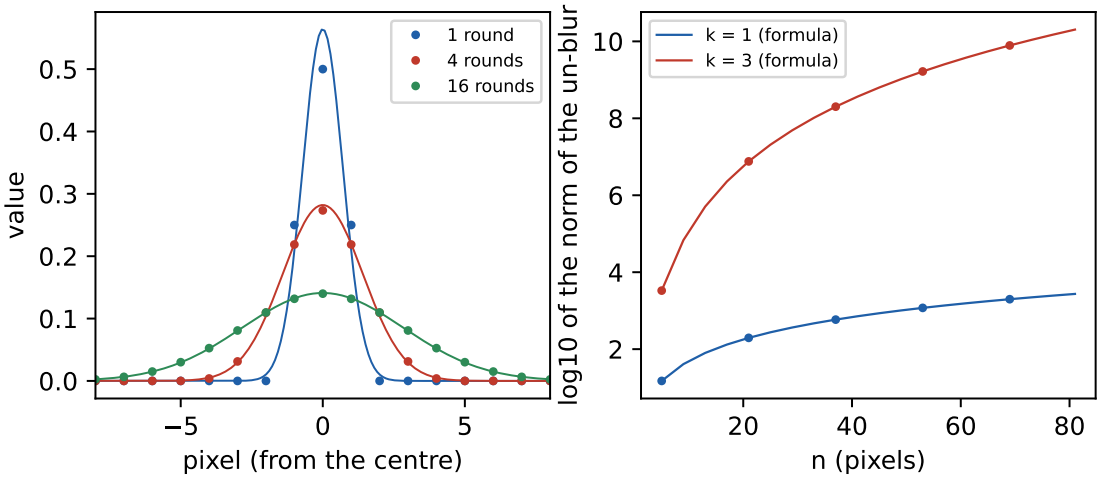
1. ใช้ฐานโหมด v_{pq} จากพิสูจน์ทฤษฎีบท 15.2.2 เมื่อ $\Omega = -L$ ได้ $I - \mu L = I + \mu\Omega$ ซึ่งเป็นเมทริกซ์สมมาตรที่มีค่าลักษณะเฉพาะ $1 + \mu\kappa_{pq}$ และ $\kappa_{pq} > 0$
2. ตัวผกผัน $(I + \mu\Omega)^{-1}$ มีค่าลักษณะเฉพาะ $1/(1 + \mu\kappa_{pq}) \in (0, 1)$ ในฐานเดียวกัน จึง $\|(I + \mu\Omega)^{-1}\|_2 < 1$
3. จาก $u^{m+1} = (I + \mu\Omega)^{-1} u^m$ ได้ $\|u^{m+1}\|_2 \le \|u^m\|_2$ ทุกก้าว

สรุป: วิธีออยเลอร์ถอยหลังไม่ระเบิดเพราะทุกโหมดถูกทำให้เล็กลง แม้ μ จะใหญ่มาก
การไม่ระเบิดไม่ได้หมายความว่าแม่น ความแม่นยำของวิธีนี้เป็นอีกเรื่องหนึ่งที่เราไม่ได้วัดในหัวข้อนี้

กลับเวลาไม่ได้: การเบลอย้อนกลับ

ถ้าการเบลอ k รอบคือการคูณด้วย B^k เมื่อ $B = I + \frac{1}{4}L$ บนแถวหนึ่งมิติ การคืนภาพคือการคูณด้วย B^{-k} ซึ่งเทียบเท่ากับการย้อนเวลาของความร้อน เราถามว่า B^{-k} ขยายข้อผิดพลาดได้มากแค่ไหน

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 15.5: ซ้าย: จุดเดียวหลังการเบลอ 1, 4 และ 16 รอบ (จุดคือค่าจากโค้ดของเรา เส้นคือเกาส์เซียนที่มีความแปรปรวน $k/2$) ขวา: $\log_{10}$ ของ $\|B^{-k}\|_2$ เทียบกับ n สำหรับ $k = 1$ และ 3 (จุดคือค่าจากเมทริกซ์ เส้นคือสูตรในทฤษฎีบท 15.2.4) (สร้างจาก `numanal/ch15.py`)

ทฤษฎีบท 15.2.4: ค่าของการย้อนการเบลอ

ให้แถวมี n จุดที่ขอบตรึง และให้ $B = I + \frac{1}{4}L$ เป็นการเบลอหนึ่งรอบ สำหรับ $k \geq 1$

$$\|B^{-k}\|_2 = \frac{1}{\sin^{2k}\big(\pi/(2(n+1))\big)}.$$

ดังนั้น $\|B^{-k}\|_2 \to \infty$ เมื่อ $n \to \infty$ การย้อนการเบลอ k รอบจึงไม่มีขอบเขตที่ไม่ขึ้นกับขนาดตาราง

พิสูจน์ (ของเราเอง).

1. ใช้ฐานไซน์จากทฤษฎีบท 15.1.2 L มีค่าลักษณะเฉพาะ $-4\sin^2\frac{p\pi}{2(n+1)}$ ดังนั้น B มีค่าลักษณะเฉพาะ $1 - \sin^2\frac{p\pi}{2(n+1)} = \cos^2\frac{p\pi}{2(n+1)}$ สำหรับ $p = 1, \ldots, n$
2. B เป็นเมทริกซ์สมมาตร จึง $\|B^{-k}\|_2 = 1/\lambda_{\min}(B)^k$ โดย $\lambda_{\min}(B)$ คือค่าลักษณะเฉพาะที่เล็กที่สุด ซึ่งเกิดที่ $p = n$: $\cos^2\frac{n\pi}{2(n+1)} = \sin^2\frac{\pi}{2(n+1)}$
3. เมื่อ $n \to \infty$ ได้ $\sin(\pi/(2(n+1))) \to 0$ ดังนั้น $\|B^{-k}\|_2 \to \infty$

สรุป: ค่าลักษณะเฉพาะเล็กที่สุดของการเบลอคือ $\sin^2(\pi/(2(n+1)))$ และมันเล็กลงเมื่อตารางใหญ่ขึ้น

ผลนี้ยืนยันสิ่งที่บทที่ 7 เห็นด้วยโค้ดว่าการคืนภาพเป็นระบบที่ไวต่อข้อมูลมาก ที่ $n = 41$ ได้ $\|B^{-1}\|_2 \approx 715.26$ และ $\|B^{-3}\|_2 \approx 3.66 \times 10^8$ ถ้าสัญญาณรบกวนมีนอร์มยูคลิดเพียง $1/255$ ข้อผิดพลาดในทิศทางที่แย่ที่สุดอาจสูงถึงราว 2.8 สำหรับ $k = 1$ และราว 1.4×10^6 สำหรับ $k = 3$ ภาพที่คืนมาด้วยสัญญาณรบกวนเล็กขนาดนี้จึงไม่มีประโยชน์ รูปที่ 15.5 ซ้ายแสดงการกระจายของจุดเดียวหลังการเบลอ และขวาแสดงนอร์มของการย้อนที่โตตาม n

ข้อควรระวัง: ย้อนเวลาไม่ใช่แค่ยาก แต่ไม่ถูกรูปแบบ

เกณฑ์ของทฤษฎีบท 15.2.4 บอกว่าความผิดเล็กในข้อมูลขยายได้ไม่มีขอบเขตเมื่อตารางใหญ่ขึ้น ปัญหาแบบนี้เรียกว่า**ปัญหาที่ไม่ถูกรูปแบบ** (ill-posed problem) ISE ต้องการให้ปัญหาการเปลี่ยนแปลงตามเวลาเป็น**ถูกรูปแบบ** (well

สัมมนา · Preprint · 2026-10-23

posed) ในทฤษฎีบท Lax Iserles 2009, §16.2, p. 360 การย้อนเวลาของสมการความร้อนจึงอยู่นอกกรอบของทฤษฎีบทนี้

สรุปใจความของหัวข้อ 15.2: ความร้อนที่แผ่ออกและภาพเบลอ

- การเบลอหนึ่งรอบคือความร้อนหนึ่งก้าวที่ $\mu = \frac{1}{4}$ (ตัวเลขคูแรนต์ตามนิยาม) และ k รอบมีความแปรปรวน $k/2$ พิกเซล^2
- วิธีชัดแจ้งบนตาราง $n \times n$ ขอบตรึงรอดก็ต่อเมื่อ $\mu \le \mu_n$ ขีดจำกัดคือ $\frac{1}{4}$ (สองมิติ) และ $\frac{1}{2}$ (หนึ่งมิติ)
- วิธีออยเลอร์ถอยหลังไม่ระเบิดสำหรับทุก μ แต่ต้องแก้ระบบทุกก้าว
- การย้อนการเบลอขยายความผิดด้วย $1/\sin^{2k}(\pi/(2(n+1)))$ ซึ่งโตไม่มีขอบเขตเมื่อตารางใหญ่ขึ้น
- คำว่าตัวเลขคูแรนต์ในหัวข้อ 15.1 และ 15.2 เป็นอัตราส่วนคนละชนิด

อ่านต่อ. ISE บรรยายสมการความร้อนและวิธีวิเคราะห์เสถียรภาพไว้ใน Iserles 2009, §16.1--16.5, pp. 349--378 เกณฑ์ของวิธีชัดแจ้งด้วยค่าลักษณะเฉพาะอยู่ใน Iserles 2009, §16.4, p. 368 หัวข้อ 15.3 จะดูการแบ่งพื้นที่เป็นชิ้นเล็ก ๆ แบบฟังก์ชันรูปหมวก

แบบฝึกหัด 15.2.1 (⋆, ทำด้วยมือ) แถวหนึ่งมิติมี 3 จุดภายใน ขอบตรึงเป็นศูนย์ ค่าเริ่มต้นเป็น $u^0 = (0,\ 1,\ 0)$ และ $\mu = 0.6$

1. ทำวิธีชัดแจ้งหนึ่งก้าวด้วยมือ และหา u^1
2. หา $\mu_3^{(1)}$ จากทฤษฎีบท 15.2.2 และบอกว่า $\mu = 0.6$ รอดบนตาราง 3 จุดหรือไม่
3. ทำข้อ 2 กับตาราง 2 จุด
4. จากผลข้อ 2 และ 3 บอกว่า $\mu = 0.6$ ระเบิดบนทุกตารางที่มี 3 จุดขึ้นไปหรือไม่ ตอบด้วยอสมการ

แบบฝึกหัด 15.2.2 (⋆, ทำด้วยมือ) ตารางสองมิติมี 2×2 จุดภายใน ขอบตรึงเป็นศูนย์ และ $h = \Delta x = 1$

1. หาค่าลักษณะเฉพาะทั้งสี่ตัวของ $-L$ จากสูตรรูปคลื่นไซน์ในทฤษฎีบท 15.1.2
2. หา $\mu_2^{(2)}$ จากทฤษฎีบท 15.2.2 และตรวจว่าเป็น $1/3$ พอดี
3. เมื่อ $\mu = 0.4$ ให้ u^0 มีแถวแรกเป็น $(1,\ -1)$ และแถวที่สองเป็น $(-1,\ 1)$ (ค่าที่จุดภายในตามรูปตารางหมากรุก และศูนย์นอกตาราง) แสดงว่า $u^1 = -1.4\,u^0$ และอธิบายว่าเหตุใดคำตอบโตขึ้นทุกก้าว
4. เทียบ $1/3$ กับค่าลิมิต $1/4$ ของทฤษฎีบท 15.2.2 และบอกว่าตารางใหญ่ขึ้นทำให้เกณฑ์เปลี่ยนอย่างไร

แบบฝึกหัด 15.2.3 (⋆⋆, คำนวณ) แถวของห้าจุด เริ่มจากค่าหนึ่งที่จุดกลางและศูนย์ที่จุดอื่น ทำการเบลอสองรอบตามข้อความในหัวข้อ 15.2 (น้ำหนัก $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$)

1. หาค่าทั้งห้าจุดหลังรอบที่สองด้วยมือ และตรวจว่าผลรวมเป็นหนึ่ง
2. หาความแปรปรวนของการกระจายนี้ (หน่วยพิกเซล^2) และเทียบกับ $k/2$ เมื่อ $k = 2$
3. เทียบค่ากลาง $3/8$ กับค่ากลางของเกาส์เซียนที่มีความแปรปรวนเท่ากัน
4. บอกว่าการเบลอสองรอบเทียบเท่ากับความร้อนที่ $\mu = \frac{1}{4}$ กี่ก้าว และเวลาเท่ากับเท่าไร

ตัวอย่าง · Preprint · 2026-10-23

แบบฝึกหัด 15.2.4 (⋆⋆⋆, พิสูจน์) วิธีแครงก์–นิโคลสันบนตาราง $n \times n$ ขอบตรึง เขียนเป็น $u^{m+1} = u^m + \frac{\mu}{2}\big(Lu^{m+1} + Lu^m\big)$

1. จัดรูปเป็น $(I + \frac{\mu}{2}\Omega)u^{m+1} = (I - \frac{\mu}{2}\Omega)u^m$ เมื่อ $\Omega = -L$ และแสดงว่าตัวคูณในแต่ละโหมดของทฤษฎีบท 15.1.2 คือ $g = (1 - \frac{1}{2}\mu\kappa)/(1 + \frac{1}{2}\mu\kappa)$ เมื่อ $\kappa > 0$ เป็นค่าลักษณะเฉพาะของ Ω
2. แสดงว่า $|g| \le 1$ สำหรับทุก $\mu > 0$ และทุก $\kappa > 0$ จึงได้ว่าวิธีนี้ถูกจำกัดขอบเขตสำหรับทุก μ
3. เมื่อ $\mu\kappa \to \infty$ แสดงว่า $g \to -1$ ไม่ใช่ 0 อธิบายว่าความต่างนี้มีผลอย่างไรต่อโหมดที่แข็งที่สุด เทียบกับวิธีออยเลอร์ถอยหลังในทฤษฎีบท 15.2.3

แบบฝึกหัด 15.2.5 (⋆⋆, คำนวณและตีความ) ให้ $\mu = 0.2503$ บนตารางสองมิติ $n \times n$ ที่ขอบตรึง

1. ใช้ทฤษฎีบท 15.2.2 หาจำนวนจุดภายในต่อด้าน n ที่มากที่สุดที่ $\mu = 0.2503$ ยังรอด แสดงการจัดรูปด้วยมือ
2. ตรวจผลข้อ 1 ด้วยโค้ดที่ $n = 44$ และ $n = 45$
3. บอกว่าเหตุใดค่า μ ที่ต่างจาก $1/4$ เพียง 0.0003 จึงทำให้ผลต่างกันระหว่างตารางขนาดต่างกัน

แบบฝึกหัด 15.2.6 (⋆⋆⋆, คำนวณและตีความ) ให้แถวของ 41 พิกเซลขอบตรึง และให้ $B = I + \frac{1}{4}L$ เป็นการเบลอหนึ่งรอบ

1. ใช้ทฤษฎีบท 15.2.4 หา $\|B^{-1}\|_2$ และ $\|B^{-3}\|_2$
2. สมมติว่าข้อมูลที่เบลอแล้วมีสัญญาณรบกวนที่มีนอร์มยูคลิด $1/255$ ประมาณขนาดข้อผิดพลาดสูงสุดของภาพที่คืนมาสำหรับ $k = 1$ และ $k = 3$ โดยใช้นอร์มข้อ 1
3. บอกว่าเหตุใดการคืนภาพจากการเบลอสามรอบจึงไม่ได้ผล แม้สัญญาณรบกวนจะเล็กมาก

คำใบ้. ข้อ 1 ของ 15.2.1 ใช้ตัวคูณ $1 - 2\mu$ ที่จุดกลาง และ μ ที่จุดข้างเคียง ข้อ 2 ของ 15.2.2 ใช้ค่าลักษณะเฉพาะที่ใหญ่ที่สุดของ $-L$ ข้อ 1 ของ 15.2.4 ใช้ $\frac{1}{2}\mu\kappa$ ในตัวส่วนและตัวเศษ ข้อ 2 ของ 15.2.6 ใช้ $\|B^{-k}\|_2$ เป็นตัวคูณที่ใหญ่ที่สุดของการขยายสัญญาณรบกวน

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 15.2.1.

1. จุดที่ 1 ได้ $0 + 0.6\,(0 - 0 + 1) = 0.6$ จุดกลางได้ $1 + 0.6\,(0 - 2 + 0) = 1 - 1.2 = -0.2$ และจุดที่ 3 ได้ $0 + 0.6\,(1 - 0 + 0) = 0.6$ ดังนั้น $u^1 = (0.6,\ -0.2,\ 0.6)$
2. จาก $\mu_n^{(1)} = 1/(2\cos^2(\pi/(2(n+1))))$ ได้ $\mu_3^{(1)} = 1/(2\cos^2(\pi/8)) = 1/(1 + \frac{\sqrt{2}}{2}) = 2 - \sqrt{2} \approx 0.585786$ เพราะ $0.6 > 2 - \sqrt{2}$ จึงระเบิด ค่าเริ่มต้น $(0, 1, 0)$ มีส่วนประกอบในโหมดที่ถี่ที่สุด ซึ่งโตตามตัวคูณ $1 - 0.6 \cdot 4\cos^2(\pi/8)$ ที่มีขนาดมากกว่าหนึ่ง
3. $\mu_2^{(1)} = 1/(2\cos^2(\pi/6)) = 1/(2 \cdot \frac{3}{4}) = \frac{2}{3} \approx 0.6667$ และ $0.6 < \frac{2}{3}$ จึงรอดบนตาราง 2 จุด
4. เพราะ $\cos^2(\pi/(2(n+1)))$ เพิ่มขึ้นเมื่อ n เพิ่ม $\mu_n^{(1)}$ จึงลดลงเข้าหา $\frac{1}{2}$ สำหรับ $n \ge 3$ ได้ $\mu_n^{(1)} \le \mu_3^{(1)} = 2 - \sqrt{2} < 0.6$ ดังนั้น $\mu = 0.6$ ระเบิดบนทุกตารางที่มี 3 จุดขึ้นไป (ค่าเริ่มต้นนี้มีส่วนประกอบในทุกโหมด)

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 15.2.2.

1. จาก $\kappa_{pq} = 4(\sin^2 \frac{p\pi}{6} + \sin^2 \frac{q\pi}{6})$ โดยที่ $\sin^2 \frac{\pi}{6} = \frac{1}{4}$ และ $\sin^2 \frac{2\pi}{6} = \frac{3}{4}$ ได้ $(1,1) \mapsto 2$, $(1,2)$ และ $(2,1) \mapsto 4$, $(2,2) \mapsto 6$ จึงได้ค่าลักษณะเฉพาะ $2, 4, 4, 6$ (ตรงกับ $4I - A$ เมื่อ A เป็นเมทริกซ์ประชิดของวัฏจักรสี่จุด ซึ่งมีค่าลักษณะเฉพาะ $2, 0, 0, -2$)

2. $\mu_2^{(2)} = 2/\kappa_{\max} = 2/6 = \frac{1}{3}$ และจากสูตรปิด $1/(4\cos^2(\pi/6)) = 1/(4 \cdot \frac{3}{4}) = \frac{1}{3}$ พอดี

3. ให้ u^0 เป็นตารางหมากรุก ที่จุด $(1,1)$ ได้ $(Lu^0)_{11} = (-1) + (-1) + 0 + 0 - 4 \cdot 1 = -6$ และที่จุด $(1,2)$ ได้ $(Lu^0)_{12} = 1 + 1 + 0 + 0 - 4 \cdot (-1) = 6$ ดังนั้น $Lu^0 = -6u^0$ และ $u^1 = u^0 + 0.4(-6)u^0 = -1.4\,u^0$ เมื่อทำซ้ำ $u^m = (-1.4)^m u^0$ ขนาดโตขึ้น 1.4 เท่าทุกก้าว จึงไม่ถูกจำกัดขอบเขต

4. $1/3$ ของตาราง 2×2 มากกว่า $1/4$ เพราะตารางหยาบกว่า เกณฑ์ $\mu_n^{(2)}$ ลดลงเมื่อ n เพิ่มขึ้น (จากสูตรของเรา $\mu_3^{(2)} \approx 0.2929$, $\mu_5^{(2)} \approx 0.2679$ และ $\mu_9^{(2)} \approx 0.2563$) และเข้าหา $\frac{1}{4}$ เมื่อตารางใหญ่ขึ้น

เฉลย 15.2.3.

1. รอบแรกได้ $(0, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0)$ รอบที่สองได้ $(\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16})$ ผลรวมคือ $\frac{1}{16} + \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = 1$ (จุดที่ 2 ของรอบที่สองคือ $\frac{1}{4} \cdot 0 + \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{4}$ และจุดที่ 1 คือ $\frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$)

2. ความแปรปรวนคือ $\sum_j j^2 y_j = 2\,(1 \cdot \frac{1}{4} + 4 \cdot \frac{1}{16}) = 2 \cdot \frac{1}{2} = 1$ ตรงกับ $k/2 = 1$ เมื่อ $k = 2$

3. เกาส์เซียนที่มีความแปรปรวน 1 มีค่ากลาง $1/\sqrt{2\pi} \approx 0.398942$ ต่างจาก $\frac{3}{8} = 0.375$ อยู่ประมาณ 0.024 (ราว 6 เปอร์เซ็นต์) ค่ากลางของการเบลอคือ $\binom{4}{2}/4^2 = \frac{3}{8}$ ซึ่งเป็นสัมประสิทธิ์ทวินาม

4. การเบลอหนึ่งรอบเท่ากับความร้อนหนึ่งก้าวที่ $\mu = \frac{1}{4}$ (ทฤษฎีบท 15.2.1 ข้อ (ก)) ดังนั้นสองรอบเท่ากับสองก้าว และเวลาคือ $2 \cdot \frac{1}{4} = \frac{1}{2}$ (เมื่อ $\Delta x = 1$) ความแปรปรวน $2t = 1$ ตรงกับข้อ 2

เฉลย 15.2.4.

1. จาก $u^{m+1} - \frac{\mu}{2} L u^{m+1} = u^m + \frac{\mu}{2} L u^m$ และ $\Omega = -L$ ได้ $(I + \frac{\mu}{2}\Omega)u^{m+1} = (I - \frac{\mu}{2}\Omega)u^m$ ในโหมด v ที่ $\Omega v = \kappa v$ ได้ $(1 + \frac{1}{2}\mu\kappa)\rho = 1 - \frac{1}{2}\mu\kappa$ จึง $\rho = g = (1 - \frac{1}{2}\mu\kappa)/(1 + \frac{1}{2}\mu\kappa)$

2. เมื่อ $a = \frac{1}{2}\mu\kappa \geq 0$ จะได้ $|1 - a| \leq 1 + a$ จึง $|g| = |1 - a|/(1 + a) \leq 1$ โหมดของ Ω ตั้งฉากกันและครบฐาน จึง $\|u^m\|_2 \leq \|u^0\|_2$ สำหรับทุก m และทุก $\mu > 0$

3. เมื่อ $\mu\kappa \to \infty$ ได้ $g \to -1$ ไม่ใช่ 0 โหมดแข็งจึงไม่หายไป แต่สลับเครื่องหมายด้วยขนาดเกือบเท่าเดิม ตรงกันข้ามกับวิธีออยเลอร์ถอยหลัง ซึ่งมี $g = 1/(1 + \mu\kappa) \to 0$ และทำให้โหมดแข็งหายไปเร็ว (ทฤษฎีบท 15.2.3) วิธีแครงก์–นิโคลสันจึงรอดทุกก้าวแต่ยังปล่อยให้โหมดแข็งสั่นอยู่ ซึ่งสอดคล้องกับ A-เสถียร (ทฤษฎีบท 14.1.2)

ร่างก่อน · Preprint · 2026-10-23

เฉลย 15.2.5.

1. $\mu_n \geq 0.2503$ ก็ต่อเมื่อ $\cos^2(\pi/(2(n+1))) \leq 1/(4 \cdot 0.2503) = 1/1.0012 \approx 0.998802$ ก็ต่อเมื่อ $\cos(\pi/(2(n+1))) \leq 0.999401$ ก็ต่อเมื่อ $\pi/(2(n+1)) \geq \arccos(0.999401) \approx 0.034612$ ก็ต่อเมื่อ $n+1 \leq \pi/(2 \cdot 0.034612) \approx 45.40$ ดังนั้น $n \leq 44$

2. โค้ดให้ $\mu_{44} \approx 0.2503049 \geq 0.2503$ และ $\mu_{45} \approx 0.2502917 < 0.2503$ ตรงกับข้อ 1

3. $\mu_n - \frac{1}{4} \approx \pi^2/(16(n+1)^2)$ ดังนั้นช่องว่างระหว่าง μ_n กับ $\frac{1}{4}$ เล็กลงตามกำลังสองของ $n+1$ เมื่อ $\mu - \frac{1}{4} = 0.0003$ ตารางที่ใหญ่กว่าราว 45 จุดต่อด้านจะเกินช่องว่างนี้ จึงระเบิด ตารางเล็กกว่ายังรอด ค่า μ ที่ใกล้ $\frac{1}{4}$ จึงรอดหรือระเบิดขึ้นกับขนาดตาราง

เฉลย 15.2.6.

1. จากทฤษฎีบท 15.2.4 $\|B^{-1}\|_2 = 1/\sin^2(\pi/84) \approx 715.26$ และ $\|B^{-3}\|_2 = 1/\sin^6(\pi/84) \approx 3.659 \times 10^8$ (ค่าเหล่านี้ตรงกับการคำนวณเมทริกซ์ผกผันด้วยโค้ดของเรา)

2. ข้อผิดพลาดสูงสุดไม่เกิน $\|B^{-k}\|_2 \cdot \|e\|_2$ และในทิศทางที่แย่ที่สุดเท่ากับค่านี้ ดังนั้นสำหรับ $k = 1$ ได้ราว $715.26/255 \approx 2.8$ และสำหรับ $k = 3$ ได้ราว $3.659 \times 10^8/255 \approx 1.43 \times 10^6$

3. ค่าของ $\|B^{-3}\|_2$ ใหญ่มากจนสัญญาณรบกวนระดับหนึ่งในหนึ่งร้อยเล็กน้อยกลายเป็นข้อผิดพลาดนับล้านเท่า ภาพที่คืนมาจึงถูกครอบงำด้วยสัญญาณรบกวนในโหมดที่สั่นที่สุด และค่านี้โตขึ้นอีกเมื่อตารางใหญ่ขึ้น การคืนภาพด้วยการกลับตัวดำเนินการตรง ๆ จึงใช้ไม่ได้

15.3 ชิ้นเล็ก ๆ ประกอบกัน

สถานการณ์สมมติ: ผ้าที่ถูกกดตรงกลาง

ลองนึกถึงผ้าในเกมที่ยึดไว้ที่ขอบสองด้าน แล้วมีของหนักมากดลงตรงกลาง ผ้าจะท้องลงเป็นโค้ง รูปโค้งขึ้นกับตำแหน่งที่ยึดและแรงที่กด แต่คอมพิวเตอร์ไม่ได้เก็บรูปโค้งทั้งผืนไว้ มันเก็บค่าที่จุดจำนวนจำกัด งานจำลองผ้าของ Baraff และ Witkin ใช้เมชของสามเหลี่ยมแทนผ้า Baraff and Witkin 1998 คือแบ่งผ้าเป็นชิ้นเล็ก ๆ แล้วคิดทีละชิ้น
ในหัวข้อนี้เราใช้เชือกหนึ่งมิติแทนผ้า เชือกยึดสองปลายและรับแรงกระจาย f เราจะถามว่า ถ้าแบ่งเชือกเป็นชิ้นเล็ก ๆ ค่าที่ได้จะใกล้ผลเฉลยจริงแค่ไหน และทำไมค่าที่จุดแบ่งจึงแม่นเป็นพิเศษ ผ้าในสถานการณ์นี้เป็นเรื่องสมมติ เราไม่ได้บอกว่าเกมใดคำนวณผ้าแบบนี้ ตัวเลขทั้งหมดมาจากโค้ดของเรา

ภาพรวมก่อนอ่าน

วิธีสมาชิกจำกัด (finite element method) หาผลเฉลยประมาณโดยแบ่งโดเมนเป็นชิ้นเล็ก ๆ แล้วใช้ฟังก์ชันง่าย ๆ ทีละชิ้น ชิ้นเล็กแต่ละชิ้นเรียกว่า**ไฟไนต์เอลิเมนต์** (finite element) ในมิติเดียวคือช่วงย่อยหนึ่งช่วงของโดเมน เราเปลี่ยนสมการเป็นรูปอ่อน เลือกฟังก์ชันรูปหมวกที่ต่อเนื่อง แล้วประกอบเป็นระบบสมการเดียว ผลสำคัญคือบทตั้งเซอา ซึ่งบอกว่าผลเฉลยกาเลอคินดีที่สุดในนอร์มพลังงานเมื่อเทียบกับปริภูมิประมาณที่เลือก

ร่างปรับปรุง · Preprint 2026-10-23

จากสมการเชือกถึงรูปอ่อน

เชือกที่ยึดสองปลายและมีความตึงหนึ่งหน่วย ยุบเป็น $u(x)$ เมื่อมีแรงกระจาย $f(x)$ ต่อหน่วยความยาว สมการคือ

$$-u'' = f \quad \text{บน } (0,1), \qquad u(0) = u(1) = 0,$$

โดย f เป็นฟังก์ชันต่อเนื่องบน $[0,1]$ สมการนี้ต้องการอนุพันธ์อันดับสองทุกจุด วิธีที่ลดความต้องการนี้คือคูณสมการด้วย**ฟังก์ชันทดสอบ** (test function) v ที่เป็นศูนย์ที่ปลาย แล้วหาปริพันธ์ จากนั้นหาปริพันธ์ทีละส่วนเพื่อย้ายอนุพันธ์ไปหา v

นิยาม 15.3.1: รูปอ่อนของปัญหาเชือก

ให้ V เป็นเซตของฟังก์ชันจริง v บน $[0,1]$ ที่ต่อเนื่อง มีอนุพันธ์ยกเว้นที่จุดจำนวนจำกัด อนุพันธ์นั้นยกกำลังสองแล้วหาปริพันธ์ได้ และ $v(0) = v(1) = 0$ ให้

$$a(v,w) = \int_0^1 v'(x)\, w'(x)\, dx, \qquad F(w) = \int_0^1 f(x)\, w(x)\, dx.$$

ผลเฉลยแบบอ่อน (weak solution) ของปัญหาคือ $u \in V$ ที่

$$a(u,v) = F(v) \qquad \text{สำหรับทุก } v \in V.$$

ฟังก์ชัน a ป้อนฟังก์ชันสองตัวเข้าไปแล้วได้จำนวนจริงออกมา เช่น

$$a(x,\, x^2) = \int_0^1 1 \cdot 2x\, dx = 1,$$

จึงเป็น**รูปแบบเชิงเส้นคู่** (bilinear form) ส่วน F ป้อนฟังก์ชันหนึ่งตัวแล้วได้จำนวนจริง และเป็นเชิงเส้นในฟังก์ชันนั้น

อ่านแบบเข้าใจง่าย

สมการ $-u'' = f$ ต้องเป็นจริงทุกจุด แต่รูปอ่อนขอแค่ว่าผลคูณกับทุกฟังก์ชันทดสอบเท่ากัน ซึ่งอ่อนกว่า และผลเฉลยเรียบก็ยังเป็นผลเฉลยเดิม

ทฤษฎีบท 15.3.1: ผลเฉลยแบบฉบับคือผลเฉลยแบบอ่อน

ถ้า f ต่อเนื่องบน $[0,1]$ แล้วมีฟังก์ชัน $u \in C^2[0,1]$ ที่ $-u'' = f$ และ $u(0) = u(1) = 0$ ฟังก์ชันนี้เป็นผลเฉลยแบบอ่อน และไม่มีผลเฉลยแบบอ่อนอื่น

ฉบับร่าง · Preprint · 2026-10-23

พิสูจน์ (ของเราเอง).

1. (การมีอยู่) ให้ $A = \int_0^1 (1-t)f(t)\,dt$ และ $u(x) = \int_0^x \left(A - \int_0^s f(t)\,dt\right) ds$ จะได้ $u'(x) = A - \int_0^x f$ และ $u'' = -f$ จึง $u \in C^2$ ที่ $x = 0$ ได้ $u(0) = 0$ และ $u(1) = A - \int_0^1 \int_0^s f(t)\,dt\,ds = A - \int_0^1 (1-t)f(t)\,dt = 0$ ตามการสลับลำดับปริพันธ์

2. (ตรงกับรูปอ่อน) สำหรับ $v \in V$ หาปริพันธ์ทีละส่วนบนแต่ละช่วงระหว่างจุดที่อนุพันธ์ไม่มี พจน์ที่จุดแบ่งหักล้างกันเพราะ u' และ v ต่อเนื่อง และพจน์ที่ปลายเป็นศูนย์เพราะ $v(0) = v(1) = 0$ จึง $\int_0^1 u'v'\,dx = -\int_0^1 u''v\,dx = \int_0^1 fv\,dx = F(v)$

3. (ความมีเพียงหนึ่ง) ถ้า u_1, u_2 เป็นผลเฉลยแบบอ่อนทั้งคู่ ให้ $w = u_1 - u_2 \in V$ จะได้ $a(w, v) = 0$ สำหรับทุก $v \in V$ เลือก $v = w$ ได้ $\int_0^1 (w')^2\,dx = 0$ ดังนั้น $w' = 0$ ยกเว้นจุดจำนวนจำกัด และ w เป็นค่าคงที่ เพราะ w ต่อเนื่อง และ $w(0) = 0$ จึง $w \equiv 0$

สรุป: สมการคลาสสิกคูณด้วยฟังก์ชันทดสอบแล้วหาปริพันธ์ให้รูปอ่อน และรูปอ่อนมีผลเฉลยเดียว

ISE สร้างทฤษฎีของรูปอ่อนเต็มรูปแบบด้วย**ปริภูมิโซโบเลฟ** (Sobolev space) ซึ่งจำเป็นเมื่อผลเฉลยไม่เรียบพอ Iserles 2009, §9.2, p. 184 กรณีของเราไม่ต้องใช้ทฤษฎีนั้น เพราะผลเฉลยเรียบ

ฟังก์ชันรูปหมวกและเมทริกซ์สติฟเนส

ให้ m เป็นจำนวนจุดภายใน $h = \frac{1}{m+1}$ และ $x_k = kh$ สำหรับ $k = 0, 1, \ldots, m+1$

นิยาม 15.3.2: ฟังก์ชันรูปหมวก

ฟังก์ชันรูปหมวก (hat function) φ_k สำหรับ $k = 1, \ldots, m$ คือ

$$\varphi_k(x) = \max\left\{0,\ 1 - \left|\frac{x}{h} - k\right|\right\}.$$

ดังนั้น $\varphi_k(x_j) = \delta_{kj}$ ที่ทุกจุดกริดเมื่อ $j = 0, \ldots, m+1$ คือ φ_k เป็นหนึ่งที่จุด x_k และเป็นศูนย์ที่จุดกริดอื่น

อ่านแบบเข้าใจง่าย

φ_k เป็นเส้นตรงเป็นท่อนสองท่อน ยอดสูง 1 ที่จุด x_k และเป็นศูนย์ที่จุดกริดอื่นทุกจุด ระหว่างจุดกริดติดกันมันเป็นเส้นตรง ISE เขียนรูปเดียวกันด้วย $\psi(x) = (1 - |x|)_+$ โดย $\varphi_k(x) = \psi(x/h - k)$ Iserles 2009, §9.1, p. 176

รูปที่ 15.6 ซ้ายแสดงรูปหมวกสี่ตัวเมื่อ $m = 4$ ขวาเทียบผลเฉลยกาเลอคินกับผลเฉลยจริงของปัญหาตัวอย่างที่ทำในหัวข้อนี้

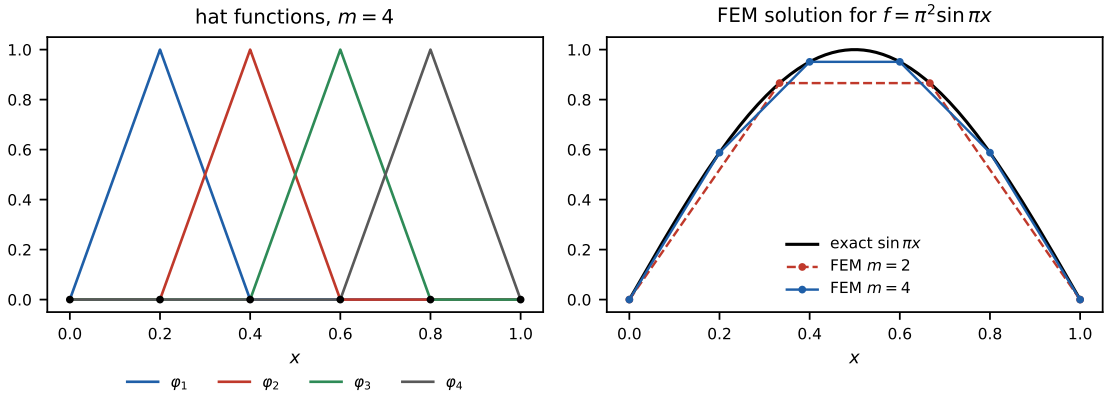
ฟังก์ชันรูปหมวกแต่ละตัวเป็นเส้นตรงเป็นท่อนบนช่วงสองช่วงที่ติดกัน และเป็นศูนย์นอกช่วงนั้น ฟังก์ชันสองตัวที่ดัชนีต่างกันเกินหนึ่งจึงไม่ซ้อนทับกันเป็นช่วง และทุกตัวเป็นศูนย์ที่ปลายทั้งสอง เมื่อรวมกันแล้ว $V_h = \operatorname{span}\{\varphi_1, \ldots, \varphi_m\}$ มีสมาชิกที่ต่อเนื่องและเป็นศูนย์ที่ 0 และ 1 ทุกตัว อนุพันธ์ของมันเป็นค่าคงที่ทีละช่วง จึงได้ $V_h \subset V$ เมื่อปริภูมิประมาณอยู่ในปริภูมิ V เดิมทั้งหมด เราเรียก V_h ว่า**ปริภูมิไฟไนต์เอลิเมนต์แบบคอนฟอร์มมิง** (conforming finite element space) ซึ่งหมายความว่าคำตอบประมาณไม่มีรอยหักที่จุดกริด และวิธีนี้เป็น**วิธีแบบคอนฟอร์มมิง** (conforming method)

เราหา $u_h = \sum_{k=1}^m c_k \varphi_k$ โดยให้สมการแปรผันเป็นจริงเพียงกับฟังก์ชันทดสอบในปริภูมิ V_h วิธีนี้คือ**วิธีกาเลอคิน** (Galerkin method)

$$a(u_h, \varphi_l) = F(\varphi_l), \qquad l = 1, \ldots, m.$$

สมการเหล่านี้เรียกว่า**สมการกาเลอคิน** (Galerkin equations) ISE ใช้ชื่อเดียวกัน Iserles 2009, §9.1, p. 172

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 15.6: ซ้าย: ฟังก์ชันรูปหมวก $\varphi_1, \ldots, \varphi_4$ เมื่อ $m = 4$ จุดดำคือจุดกริด ขวา: ผลเฉลยจริง $\sin \pi x$ (เส้นดำ) และผลเฉลยกาเลอคินของ $-u'' = \pi^2 \sin \pi x$ เมื่อ $m = 2$ (เส้นประ) กับ $m = 4$ (เส้นสี) ค่า FEM ที่จุดกริดทับกับผลเฉลยจริงตามทฤษฎีบท 15.3.5 (สร้างจาก `numanal/ch15.py`)

ทฤษฎีบท 15.3.2: ระบบสมการกาเลอคินของเชือก

ให้ K เป็นเมทริกซ์ $m \times m$ ที่มี $K_{lk} = a(\varphi_k, \varphi_l)$ (ผลคูณภายในเชิงพลังงานของรูปหมวกคู่นั้น) และให้ $b_l = F(\varphi_l)$ (แรง f ที่ถ่วงด้วยรูปหมวก φ_l) เรียก b ว่า**เวกเตอร์ภาระ** (load vector) แล้ว

(ก) $K = \frac{1}{h} T$ โดย $T = \operatorname{tridiag}(-1, 2, -1)$ เรียก K ว่า**เมทริกซ์สติฟเนส** (stiffness matrix)

(ข) สมการกาเลอคินเป็น $Kc = b$ และมีผลเฉลยเดียว

(ค) $u_h(x_k) = c_k$ สำหรับ $k = 1, \ldots, m$

พิสูจน์ (ของเราเอง).

1. อนุพันธ์ของ φ_k คือ $\frac{1}{h}$ บน (x_{k-1}, x_k) และ $-\frac{1}{h}$ บน (x_k, x_{k+1}) นอกนั้นเป็นศูนย์
2. ดังนั้น $K_{kk} = \int \varphi_k'^2 \, dx = \frac{2h}{h^2} = \frac{2}{h}$ และ $K_{k,k+1} = \int_{x_k}^{x_{k+1}} \left(-\frac{1}{h}\right)\frac{1}{h} \, dx = -\frac{1}{h}$ ตัวรองรับของ φ_k กับ φ_l ไม่ซ้อนกันเป็นช่วงเมื่อ $|k - l| \geq 2$ จึง $K_{lk} = 0$ ได้ $K = \frac{1}{h} T$ ซึ่งคือข้อ (ก)
3. สำหรับข้อ (ข) ให้ $u_h = \sum c_k \varphi_k$ ถ้า $c \neq 0$ แล้ว $c^{\mathsf{T}} K c = \int_0^1 (u_h')^2 \, dx > 0$ เพราะถ้า $u_h' \equiv 0$ จะได้ $u_h \equiv u_h(0) = 0$ และ $c_k = u_h(x_k)$ จะเป็นศูนย์หมด ดังนั้น K เป็นบวกแน่นอนจึงไม่เอกฐาน และระบบมีผลเฉลยเดียว
4. ข้อ (ค) ตามจาก $\varphi_k(x_j) = \delta_{kj}$

เมทริกซ์ T นี้คือเมทริกซ์สามแนวเฉียงที่สูตรผลต่างกลางของหัวข้อ 10.1 คูณด้วย $\frac{1}{h^2}$ ส่วนที่ต่างจากวิธีนั้นมีเพียงตัวคูณและด้านขวาของสมการ

ลองเอง: หนึ่งจุดภายในและแรงคงที่

ให้ $m = 1$ ($h = \frac{1}{2}$) และ $f = 2$ ผลเฉลยจริงคือ $u(x) = x(1 - x)$ หา K, b_1 และ c_1 แล้วหา $u_h(\frac{1}{4})$ เทียบกับ $u(\frac{1}{4}) = \frac{3}{16}$ (คำตอบ: $K = 4$, $b_1 = 1$, $c_1 = \frac{1}{4}$, $u_h(\frac{1}{4}) = \frac{1}{8}$ ค่าคลาดเคลื่อน $\frac{1}{16}$ ที่จุดที่ไม่ใช่จุดกริด) จากนั้นลอง

ฉบับร่าง · Preprint · 2026-10-25

ทำ $m = 3$ และตรวจว่า $c = (\frac{3}{16}, \frac{1}{4}, \frac{3}{16})$ ให้ค่าที่โหนดตรงกับ u พอดี

บทตั้งเซอา: ดีที่สุดในนอร์มพลังงาน

สมมติว่า V เป็นปริภูมิเชิงเส้นที่มีนอร์ม $\|\cdot\|$ และ a เป็นรูปแบบเชิงเส้นคู่ที่มีขอบเขต $|a(v,w)| \le M\|v\|\,\|w\|$ และ**โคเออร์ซีฟ** (coercive) $a(v,v) \ge \alpha\|v\|^2$ โดย $M, \alpha > 0$ สมมติฐานสองข้อนี้ต้องตรวจก่อนใช้บทตั้ง ไม่ใช่สมบัติที่มีในทุก a

ทฤษฎีบท 15.3.3: บทตั้งเซอา

ให้ $V_h \subset V$ เป็นปริภูมิจำกัดมิติ ให้ $u \in V$ เป็นผลเฉลยของ $a(u,v) = F(v)$ ทุก $v \in V$ และให้ $u_h \in V_h$ เป็นผลเฉลยของ $a(u_h, v_h) = F(v_h)$ ทุก $v_h \in V_h$ แล้ว u_h มีเพียงหนึ่งเดียว และ

$$\|u - u_h\| \le \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|.$$

พิสูจน์ (ของเราเอง).

1. (ความตั้งฉากแบบกาเลอคิน) สำหรับ $v_h \in V_h \subset V$ ได้ $a(u - u_h, v_h) = F(v_h) - F(v_h) = 0$
2. (มีเพียงหนึ่ง) ถ้า $w \in V_h$ และ $a(w, v_h) = 0$ ทุก $v_h \in V_h$ เลือก $v_h = w$ จะได้ $\alpha\|w\|^2 \le 0$ จึง $w = 0$ ระบบเชิงเส้นของสัมประสิทธิ์ของ u_h จึงไม่เอกฐาน และมีผลเฉลยเดียว
3. ให้ $e = u - u_h$ และ $v_h \in V_h$ ใด ๆ เขียน $e = (u - v_h) + (v_h - u_h)$ โดย $v_h - u_h \in V_h$ จากข้อ 1 ได้ $a(e, v_h - u_h) = 0$ จึง
$$\alpha\|e\|^2 \le a(e,e) = a(e, u - v_h) \le M\|e\|\,\|u - v_h\|.$$
4. หารด้วย $\|e\|$ (ถ้า $\|e\| = 0$ ไม่มีอะไรต้องพิสูจน์) ได้ $\|e\| \le \frac{M}{\alpha}\|u - v_h\|$ สำหรับทุก v_h จึงได้ infimum

สรุป: ผลเฉลยกาเลอคินมีค่าคลาดเคลื่อนไม่เกินค่าคงที่คูณค่าคลาดเคลื่อนของการประมาณที่ดีที่สุดในปริภูมิประมาณ

ในปัญหาของเรา $\|v\|_E = \|v'\|_{L^2}$ เป็น**นอร์มพลังงาน** (energy norm) บน V เพราะถ้า $v' = 0$ และ $v(0) = 0$ จะได้ $v \equiv 0$ และ $a(v,v) = \|v\|_E^2$ จึงโคเออร์ซีฟด้วย $\alpha = 1$ อีกทั้ง $|a(v,w)| \le \|v\|_E\|w\|_E$ จาก Cauchy--Schwarz จึงได้ $M = 1$ ดังนั้น

$$\|u' - u_h'\|_{L^2} \le \inf_{v_h \in V_h} \|u' - v_h'\|_{L^2}.$$

ค่าคงที่ของเซอาเท่ากับหนึ่งในนอร์มนี้ บทตั้งทั่วไปอยู่ใน Ern and Guermond 2004, Lemma 2.28, p. 96

นิยาม 15.3.3: ฟังก์ชันประมาณค่าในช่วงเชิงเส้นชิ้น

ถ้า u ต่อเนื่องบน $[0,1]$ และเป็นศูนย์ที่ปลาย ให้

$$I_h u = \sum_{k=1}^{m} u(x_k)\,\varphi_k,$$

ซึ่งเป็นฟังก์ชันเส้นตรงเป็นท่อนที่ผ่านค่าของ u ที่จุดกริดทุกจุด

ฉบับร่าง Preprint, 2026-10-23

ทฤษฎีบท 15.3.4: การประมาณในช่วงเชิงเส้นชิ้นในนอร์มพลังงาน

ถ้า $u \in C^2[0,1]$ และ $u(0) = u(1) = 0$ แล้ว

$$\|u' - (I_h u)'\|_{L^2(0,1)} \le h\,\|u''\|_{L^2(0,1)}.$$

พิสูจน์ (ของเราเอง).

1. ให้ $e = u - I_h u$ บนช่วง $I = [x_{k-1}, x_k]$ แต่ละช่วง $e(x_{k-1}) = e(x_k) = 0$ ตามนิยาม จึงมี $\xi \in I$ ที่ $e'(\xi) = 0$ จากทฤษฎีบทของโรล
2. บน I ฟังก์ชัน $(I_h u)'' = 0$ จึง $e'' = u''$ และ $e'(x) = \int_\xi^x u''(t)\,dt$ สำหรับ $x \in I$
3. จากอสมการโคชี--ชวาร์ซ $|e'(x)|^2 \le |x-\xi| \int_I (u'')^2\,dt \le h \int_I (u'')^2\,dt$ จึง $\int_I (e')^2\,dx \le h^2 \int_I (u'')^2\,dx$
4. รวมทุกช่วง I ได้ผลที่ต้องการ

สรุป: ผลต่างของความชันมีนอร์ม L^2 ไม่เกิน h คูณนอร์ม L^2 ของ u''

ทฤษฎีบท 15.3.5: ที่โหนดแม่นและขอบเขตพลังงาน

ให้ f ต่อเนื่องบน $[0,1]$ ให้ u เป็นผลเฉลยแบบฉบับของ $-u'' = f$ ที่เป็นศูนย์ที่ปลาย และให้ $u_h = \sum c_k \varphi_k$ เป็นผลเฉลยกาเลอคินที่ใช้ภาระแม่น $b_l = F(\varphi_l)$ แล้ว

(ก) $u_h = I_h u$ นั่นคือ $c_k = u(x_k)$ ทุก k

(ข) $\|u' - u_h'\|_{L^2(0,1)} \le h\,\|f\|_{L^2(0,1)}$

พิสูจน์ (ของเราเอง).

1. ให้ $e = u - I_h u$ สำหรับ l ใด ๆ บนช่วง $[x_{l-1}, x_l]$ อนุพันธ์ของ φ_l เป็น $\frac{1}{h}$ และบนช่วง $[x_l, x_{l+1}]$ เป็น $-\frac{1}{h}$ จึง
$$a(e, \varphi_l) = \frac{1}{h}\big(e(x_l) - e(x_{l-1})\big) - \frac{1}{h}\big(e(x_{l+1}) - e(x_l)\big) = 0,$$
เพราะ e เป็นศูนย์ที่ทุกจุดกริด
2. ผลเฉลย u เป็นคำตอบของสมการแปรผัน จึง $F(\varphi_l) = a(u, \varphi_l)$ ดังนั้น $a(I_h u, \varphi_l) = a(u, \varphi_l) - a(e, \varphi_l) = F(\varphi_l)$ สำหรับทุก l ทำให้ $I_h u \in V_h$ เป็นผลเฉลยกาเลอคิน และจากความมีเพียงหนึ่งในทฤษฎีบท 15.3.2 ได้ $u_h = I_h u$ ซึ่งคือข้อ (ก)
3. จากข้อ (ก) และบทตั้งเซอาในนอร์มพลังงาน หรือจากทฤษฎีบท 15.3.4 กับ $u'' = -f$ ได้ $\|u' - u_h'\| = \|u' - (I_h u)'\| \le h\|u''\| = h\|f\|$ ซึ่งคือข้อ (ข)

สรุป: ด้วยภาระแม่น ค่าที่โหนดของเชือกหนึ่งมิติแม่นพอดี และค่าคลาดเคลื่อนในนอร์มพลังงานเป็นอันดับ h

ร่างต้นฉบับ . Preprint . 2026-10-23

เกร็ดเพิ่มเติม (ผู้เรียบเรียง)

ข้อ (ก) ใช้ได้เพราะ $a(v,w)=\int v'w'$ มีสัมประสิทธิ์เป็นหนึ่ง และ φ_l' เป็นค่าคงที่ทีละช่วง เราไม่ได้ตรวจกรณี $b(x)u$ หรือหลายมิติ

ข้อควรระวัง: ที่โหนดแม่น ไม่ได้แปลว่าทุกจุดแม่น

ทฤษฎีบท 15.3.5 ให้ค่าที่โหนดแม่น แต่ระหว่างจุดกริด u_h ต่างจาก u ทฤษฎีบทนี้พิสูจน์เพียงว่าค่าคลาดเคลื่อนในนอร์มพลังงานเป็นอันดับ h จึงอย่าใช้บอกค่าสูงสุดระหว่างจุด

ตัวเลขจากโค้ดและวิธีผลต่างกลาง

ให้ $f(x)=\pi^2\sin(\pi x)$ ผลเฉลยจริงคือ $u=\sin(\pi x)$ ตารางนี้คำนวณด้วยภาระแม่น และค่าคลาดเคลื่อนสูงสุดระหว่างจุดวัดบนจุดตัวอย่าง 4001 จุด

m	h	สูงสุดระหว่างจุด	ค่าคลาดเคลื่อนพลังงาน	ขอบเขต $h\|f\|$
4	$\frac{1}{5}$	4.894×10^{-2}	4.003×10^{-1}	1.396
8	$\frac{1}{9}$	1.519×10^{-2}	2.234×10^{-1}	0.775
16	$\frac{1}{17}$	4.266×10^{-3}	1.184×10^{-1}	0.411
32	$\frac{1}{33}$	1.133×10^{-3}	6.104×10^{-2}	0.211

ที่โหนดทุกจุดค่า FEM แตกต่างจากผลเฉลยจริงประมาณ 10^{-15} ซึ่งเป็นค่าการปัดเศษ วิธีผลต่างกลางให้ค่าคลาดเคลื่อนที่โหนดเท่ากับ 3.19×10^{-2} เมื่อ $m=4$ ความต่างมาจากภาระที่ใช้ ดังทฤษฎีบทต่อไปนี้

ทฤษฎีบท 15.3.6: ภาระแบบกฎเชิงสี่เหลี่ยมคางหมูให้ระบบผลต่างกลาง

ถ้าเลือก $b_l=h\,f(x_l)$ ซึ่งได้จาก**กฎเชิงสี่เหลี่ยมคางหมู** (trapezoidal rule) ระบบ $Kc=b$ เท่ากับ $\frac{1}{h^2}Tc=f(x_l)$ ทุก l ซึ่งคือระบบผลต่างกลางของหัวข้อ 10.1 ทุกประการ

พิสูจน์ (ของเราเอง).

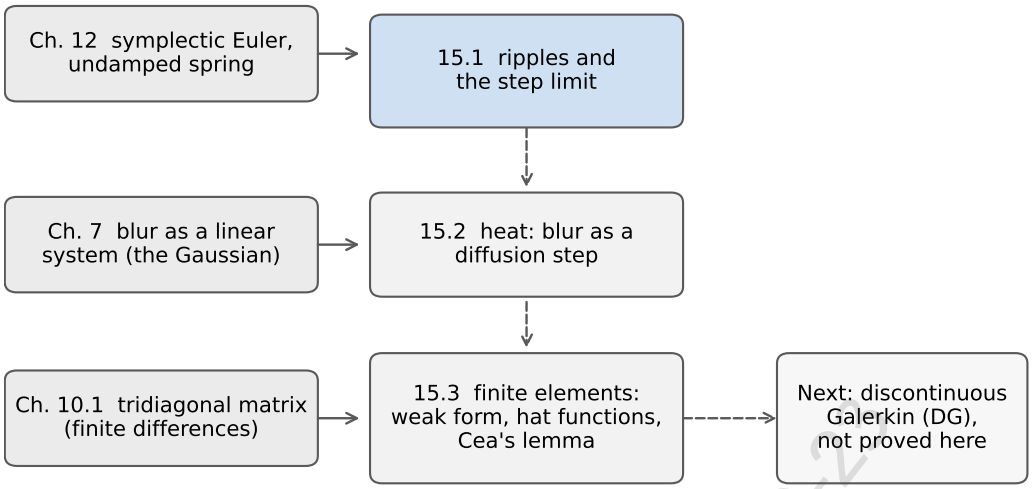
1. φ_l เป็นศูนย์ที่ x_{l-1} และ x_{l+1} และเป็น 1 ที่ x_l กฎเชิงสี่เหลี่ยมคางหมูบนแต่ละช่วงครึ่งหนึ่งให้ $\frac{h}{2}\big(0+f(x_l)\big)$ รวมสองครึ่งได้ $hf(x_l)$
2. ดังนั้น $Kc=b$ เป็น $\frac{1}{h}Tc=hf(x_l)$ ซึ่งเท่ากับ $\frac{1}{h^2}Tc=f(x_l)$

สรุป: ภาระที่แม่นและภาระแบบคางหมูให้สมการต่างกัน แม้เมทริกซ์จะเป็นตัวเดียวกัน

ข้อควรระวัง: ฟังก์ชันรูปหมวกต่อเนื่อง แต่วิธี DG ไม่ต่อเนื่อง

ปริภูมิ V_h ของเราอยู่ใน V จึงใช้บทตั้งเซอาได้ตรง ๆ ส่วนวิธี DG ใช้ฟังก์ชันที่กระโดดที่รอยต่อ ซึ่งไม่อยู่ใน V ต้องใช้การวิเคราะห์อีกแบบที่เรายังไม่ได้พิสูจน์ที่นี่

ฉบับร่าง · Preprint · 2026-10-23



รูปที่ 15.7: แผนผังบทที่ 15 ทั้งบท ลูกศรทึบคือแนวคิดที่หัวข้อหนึ่งใช้ ลูกศรประคือแนวคิดที่ต่อจากหัวข้อก่อนหน้า (สร้างจาก `numanal/ch15.py`)

สรุปใจความของหัวข้อ 15.3: ชิ้นเล็ก ๆ ประกอบกัน

- รูปอ่อนขอแค่ว่าผลคูณกับทุกฟังก์ชันทดสอบเท่ากัน และผลเฉลยแบบฉบับเป็นผลเฉลยแบบอ่อนเพียงหนึ่งเดียว
- ฟังก์ชันรูปหมวกต่อเนื่องและอยู่ใน V ทำให้เมทริกซ์สติฟเนส $K = \frac{1}{h}T$ เป็นแถบสามแนวเฉียง
- บทตั้งเซอา: ผลเฉลยกาเลอคินดีที่สุดในนอร์มพลังงาน ค่าคงที่เท่ากับหนึ่งในปัญหาของเรา
- ด้วยภาระแม่นในหนึ่งมิติ ค่าที่โหนดแม่นพอดี และค่าคลาดเคลื่อนในนอร์มพลังงานไม่เกิน $h\|f\|$

แบบฝึกหัด (ผู้เรียบเรียง)

แบบฝึกหัด 15.3.1 ($\star$, ทำด้วยมือ) ให้ $-u'' = 1$ บน $[0,1]$ กับ $u(0) = u(1) = 0$ ผลเฉลยจริงคือ $u(x) = \frac{1}{2}x(1-x)$ (ตรวจได้โดยหาอนุพันธ์สองครั้ง)

1. เมื่อ $m = 1$ ($h = \frac{1}{2}$) หาค่า K และ $b_1 = \int_0^1 \varphi_1\,dx$ แล้วหา $u_h(\frac{1}{2})$ เทียบกับ $u(\frac{1}{2})$
2. เมื่อ $m = 3$ ($h = \frac{1}{4}$) ตรวจว่า $c = (\frac{3}{32}, \frac{1}{8}, \frac{3}{32})$ เป็นคำตอบของ $Kc = b$ เมื่อ $b_l = \int_0^1 \varphi_l\,dx$

แบบฝึกหัด 15.3.2 ($\star$, ทำด้วยมือ) สำหรับ $m = 4$ ($h = \frac{1}{5}$)

1. เขียนเมทริกซ์ K ขนาด 4×4 โดยใช้ $K_{lk} = \int_0^1 \varphi_k'\varphi_l'\,dx$
2. ค่าลักษณะเฉพาะของ K คือ $\frac{4}{h}\sin^2\frac{j\pi}{2(m+1)}$ สำหรับ $j = 1,\dots,4$ หาค่าที่มากที่สุด และตรวจว่าเท่ากับ $20\sin^2\frac{2\pi}{5}$
3. อธิบายว่าทำไม K เป็นเมทริกซ์สามแนวเฉียง โดยใช้ว่า φ_k และ φ_l มีส่วนที่ซ้อนกันเมื่อ $|k-l| \le 1$ เท่านั้น

ฉบับร่าง · Preprint · 2026-10-21

แบบฝึกหัด 15.3.3 (★★, คำนวณ) ให้ $f(x) = \pi^2 \sin(\pi x)$ บน $[0,1]$ ผลเฉลยจริงคือ $u(x) = \sin(\pi x)$ ทำการคำนวณสำหรับ $m = 4, 8, 16, 32$ ด้วยโค้ดของเราและตอบดังนี้

1. ทำตารางค่าคลาดเคลื่อนพลังงาน $\|u' - u_h'\|_{L^2}$ และขอบเขต $h\,\|f\|_{L^2}$ บอกว่าขอบเขตเป็นจริงทุกแถวหรือไม่
2. เมื่อ h ลดลงครึ่งหนึ่ง อัตราส่วนของค่าคลาดเคลื่อนพลังงานเข้าใกล้เท่าไร และสอดคล้องกับอันดับของ h เท่าใด
3. ค่าคลาดเคลื่อนสูงสุดระหว่างจุดที่วัดบนจุดตัวอย่าง 4001 จุด ลดลงเป็นกี่เท่าเมื่อ h ลดครึ่งหนึ่ง
4. เมื่อ $m = 4$ ค่าคลาดเคลื่อนที่โหนดของวิธีผลต่างกลางจากหัวข้อ 10.1 คือ 0.0319 ใช้ทฤษฎีบท 15.3.5 และ 15.3.6 อธิบายว่าทำไมวิธีสมาชิกจำกัดที่มีภาระแม่นจึงให้ค่าที่โหนดแม่นกว่า

แบบฝึกหัด 15.3.4 (★★★, พิสูจน์) ให้ $u_h \in V_h$ เป็นผลเฉลยกาเลอคิน และให้ $\|w\|_E^2 = a(w,w)$ เมื่อ $a(v,w) = \int_0^1 v'w'\,dx$ เป็นรูปแบบสมมาตร

1. พิสูจน์ความสัมพันธ์พิทาโกรัส $\|u - v_h\|_E^2 = \|u - u_h\|_E^2 + \|u_h - v_h\|_E^2$ สำหรับทุก $v_h \in V_h$
2. ใช้ข้อ 1 แสดงว่า $\|u - u_h\|_E \le \|u - v_h\|_E$ สำหรับทุก $v_h \in V_h$ (ค่าคงที่ของเซอาเท่ากับหนึ่ง)
3. ให้ $J(v) = \tfrac{1}{2}a(v,v) - F(v)$ แสดงว่า $J(v_h) = J(u_h) + \tfrac{1}{2}\|v_h - u_h\|_E^2$ และสรุปว่า u_h ทำให้ J น้อยที่สุดบน V_h

ตัวช่วยคิด

ในข้อ 1 เขียน $u - v_h = (u - u_h) + (u_h - v_h)$ แล้วกระจายด้วยสมมาตรของ a ใช้ความตั้งฉากแบบกาเลอคินกับ $u_h - v_h \in V_h$ ในข้อ 3 แทน $v_h = u_h + w$ แล้วใช้ $a(u_h, w) = F(w)$

แบบฝึกหัด 15.3.5 (★★, สิ่งที่อาจผิด) ให้ $v(x) = 2x$ เมื่อ $0 \le x < \tfrac{1}{2}$ และ $v(x) = 2x - 1$ เมื่อ $\tfrac{1}{2} < x \le 1$

1. หา $v(0)$ และ $v(1)$ และค่าลิมิตซ้ายกับขวาที่ $x = \tfrac{1}{2}$ อธิบายว่าทำไม $v \notin V$
2. คำนวณ $\int_0^1 (v')^2\,dx$ โดยแยกเป็นสองช่วง ได้ค่าเท่าไร
3. ค่าในข้อ 2 เป็นนอร์มพลังงานของ v หรือไม่ ทฤษฎีบท 15.3.3 ใช้กับ v ได้หรือไม่ และเพราะอะไร

แบบฝึกหัด 15.3.6 (★★★, โครงงานย่อย) ให้แรงกด $f(x) = \exp\!\big(-((x - 0.3)/0.05)^2\big)$ บน $[0,1]$ (แรงกระจุกแคบรอบ $x = 0.3$) ใช้ผลเฉลยอ้างอิงที่ $m = 2047$ ที่โค้ดให้ ซึ่งไม่ใช่ผลเฉลยจริง

1. ใช้โค้ดทำตารางค่าคลาดเคลื่อนสูงสุดระหว่างจุดเทียบกับผลเฉลยอ้างอิงสำหรับ $m = 3, 7, 15, 31, 63, 127$
2. หา m ที่เล็กที่สุดในรายการที่ทำให้ค่าคลาดเคลื่อนต่ำกว่า 10^{-3}
3. ที่โหนดของ m ใดก็ได้ ค่า FEM ต่างจากค่าอ้างอิงเท่าไร และทำไมจึงเป็นเช่นนั้น (อ้างทฤษฎีบท 15.3.5)
4. ระบบของ $m = 15$ มีตัวไม่ทราบกี่ตัว และ K มีสมาชิกที่ไม่เป็นศูนย์กี่ตัว
5. เขียนย่อหน้าสั้น ๆ ว่าถ้าเป็นผ้าที่มีแรงกดแคบเช่นนี้ ควรเลือกชิ้นเล็กอย่างไร โดยอ้างตารางในข้อ 1

สมบรูณ์ Preprint · 2026-10-23

เฉลยแบบฝึกหัด (ผู้เรียบเรียง)

เฉลย 15.3.1.

1. เมื่อ $m = 1$ ได้ $K = \frac{1}{h} \cdot 2 = 4$ และ b_1 เป็นพื้นที่ของสามเหลี่ยมที่ฐานยาว $2h = 1$ และสูง 1 คือ $b_1 = \frac{1}{2}$ ดังนั้น $c_1 = \frac{1/2}{4} = \frac{1}{8}$ และ $u_h(\frac{1}{2}) = \frac{1}{8}$ ส่วนผลเฉลยจริงได้ $u(\frac{1}{2}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$ ค่าตรงกันพอดีที่โหนด (ตรงกับทฤษฎีบท 15.3.5)

2. เมื่อ $m = 3$ แต่ละ $b_l = \int \varphi_l = h = \frac{1}{4}$ และ $K = 4T$ คำนวณ Tc ได้ $(2 \cdot \frac{3}{32} - \frac{1}{8},\ -\frac{3}{32} + 2 \cdot \frac{1}{8} - \frac{3}{32},\ \dots) = (\frac{1}{16}, \frac{1}{16}, \frac{1}{16})$ ดังนั้น $Kc = 4 \cdot \frac{1}{16} = \frac{1}{4} = b_l$ ตรงกันทุกแถว

เฉลย 15.3.2.

1. จาก $K = \frac{1}{h}T$ และ $\frac{1}{h} = 5$ ได้
$$K = \begin{pmatrix} 10 & -5 & 0 & 0 \\ -5 & 10 & -5 & 0 \\ 0 & -5 & 10 & -5 \\ 0 & 0 & -5 & 10 \end{pmatrix}.$$

2. ค่าลักษณะเฉพาะคือ $20\sin^2 \frac{j\pi}{10}$ สำหรับ $j = 1, 2, 3, 4$ ซึ่งได้ $1.909830,\ 6.909830,\ 13.090170,\ 18.090170$ ค่าที่มากที่สุดคือ $20\sin^2 \frac{4\pi}{10} = 20\sin^2 \frac{2\pi}{5} \approx 18.090170$ ตรงกับการคำนวณเชิงตัวเลข

3. ตัวรองรับของ φ_k คือ $[x_{k-1}, x_{k+1}]$ ถ้า $|k - l| \geq 2$ ตัวรองรับของ φ_k และ φ_l ไม่ซ้อนกันเป็นช่วงยาวบวก ผลคูณ $\varphi_k'\varphi_l'$ จึงเป็นศูนย์เกือบทุกจุด และ $K_{lk} = 0$ เมื่อ $|k - l| \geq 2$ ส่วนค่าที่ไม่เป็นศูนย์อยู่บนแนวทแยงและแนวข้างเคียง

เฉลย 15.3.3.

1. ตารางจากโค้ดของเรา ($\|f\|_{L^2} = \pi^2/\sqrt{2} \approx 6.978864$)

m	$\|u' - u_h'\|_{L^2}$	$h\,\|f\|_{L^2}$	ตรวจขอบเขต
4	0.400284	1.395773	จริง
8	0.223393	0.775429	จริง
16	0.118440	0.410521	จริง
32	0.061040	0.211481	จริง

2. อัตราส่วนคือ 1.79184, 1.88613 และ 1.94037 เข้าใกล้ 2 จึงสอดคล้องกับค่าคลาดเคลื่อนเป็นอันดับ h ในนอร์มพลังงาน (ตรงกับขอบเขต $h\|f\|$)

3. ค่าคลาดเคลื่อนสูงสุดบน 4001 จุดคือ 4.89435×10^{-2}, 1.51922×10^{-2}, 4.26582×10^{-3} และ 1.13266×10^{-3} อัตราส่วนคือ 3.22161, 3.56139 และ 3.76620 เข้าใกล้ 4 ซึ่งบอกว่าค่าคลาดเคลื่อนระหว่างจุดลดลงในอันดับ h^2 จากตัวเลขเท่านั้น เรายังไม่ได้พิสูจน์ที่นี่

4. ทฤษฎีบท 15.3.5 บอกว่าภาระแม่น $b_l = \int f\varphi_l$ ทำให้ค่าที่โหนดของ FEM เท่ากับผลเฉลยจริงทุกจุด (ในกรณีนี้ $\approx 10^{-15}$ จากการปัดเศษ) ส่วนวิธีผลต่างกลางใช้ค่า $f(x_l)$ โดยตรงแทนปริพันธ์ ทฤษฎีบท 15.3.6 บอกว่าภาระแบบกฎเชิงสี่เหลี่ยมคางหมู $b_l = hf(x_l)$ ให้ระบบเดียวกับผลต่างกลาง จึงให้ค่าคลาดเคลื่อนที่โหนดเท่ากัน (0.0319 เมื่อ $m = 4$) ความต่างอยู่ที่ภาระเท่านั้น

ฉบับร่าง · Preprint · 2026-10-23

เฉลย 15.3.4.

1. ให้ $e = u - u_h$ และ $w = u_h - v_h \in V_h$ จะได้ $u - v_h = e + w$ เมื่อกระจาย $\|e+w\|_E^2 = a(e,e) + a(e,w) + a(w,e) + a(w,w)$ เพราะ a สมมาตร $a(e,w) = a(w,e)$ และจากความตั้งฉากแบบกาเลอคิน $a(e,w) = 0$ ดังนั้น $\|e+w\|_E^2 = \|e\|_E^2 + \|w\|_E^2$ ซึ่งคือข้อความที่ต้องการ

2. จากข้อ 1 $\|u-u_h\|_E^2 \le \|u-u_h\|_E^2 + \|u_h - v_h\|_E^2 = \|u - v_h\|_E^2$ จึง $\|u-u_h\|_E \le \|u-v_h\|_E$ สำหรับทุก $v_h \in V_h$ ค่าคงที่ M/α ของเซอาในนอร์มนี้เท่ากับหนึ่ง

3. แทน $v_h = u_h + w$ จะได้ $J(v_h) = \frac{1}{2}a(u_h,u_h) + a(u_h,w) + \frac{1}{2}a(w,w) - F(u_h) - F(w)$ ใช้ $a(u_h,w) = F(w)$ ได้ $J(v_h) = \left[\frac{1}{2}a(u_h,u_h) - F(u_h)\right] + \frac{1}{2}\|w\|_E^2 = J(u_h) + \frac{1}{2}\|v_h - u_h\|_E^2$ ค่านี้มีค่าน้อยที่สุดเมื่อ $v_h = u_h$ เท่านั้น ดังนั้น u_h ทำให้ J น้อยที่สุดบน V_h เพียงจุดเดียว

เฉลย 15.3.5.

1. $v(0) = 0$ แต่ $v(1) = 2\cdot 1 - 1 = 1 \neq 0$ ที่ $x = \frac{1}{2}$ ลิมิตซ้ายคือ $2\cdot\frac{1}{2} = 1$ และลิมิตขวาคือ $2\cdot\frac{1}{2} - 1 = 0$ จึงกระโดดที่จุดนี้ V ต้องการฟังก์ชันต่อเนื่องที่ปลายเป็นศูนย์ ดังนั้น $v \notin V$

2. บนแต่ละช่วง $v' = 2$ จึง $\int_0^{1/2} 4\,dx + \int_{1/2}^1 4\,dx = 2 + 2 = 4$

3. ไม่ได้ เพราะนอร์มพลังงานนิยามบน V และ $v \notin V$ ทฤษฎีบท 15.3.3 ต้องการให้ u และ u_h อยู่ใน V และ $V_h \subset V$ เมื่อ $v \notin V$ ค่า 4 นับเฉพาะความชันในแต่ละช่วง และไม่เห็นการกระโดด จึงใช้เป็นขอบเขตความผิดพลาดไม่ได้

เฉลย 15.3.6.

1. จากโค้ด ค่าคลาดเคลื่อนสูงสุดระหว่างจุดคือ 2.74365×10^{-3} ($m=3$), 1.51625×10^{-3} ($m=7$), 4.07404×10^{-4} ($m=15$), 1.19650×10^{-4} ($m=31$), 3.01127×10^{-5} ($m=63$) และ 7.61221×10^{-6} ($m=127$)

2. $m = 15$ ($h = \frac{1}{16}$) เป็นค่าที่เล็กที่สุดในรายการที่ค่าคลาดเคลื่อนต่ำกว่า 10^{-3} ($m=7$ ยังได้ 1.5×10^{-3})

3. ต่างกันราว 10^{-15} (จากการปัดเศษ) เพราะทั้งสองวิธีใช้ภาระแม่น และทฤษฎีบท 15.3.5 บอกว่าค่าที่โหนดของทั้งสองเท่ากับผลเฉลยจริง ผลเฉลยอ้างอิงจึงแม่นที่โหนดของ $m = 3, 7, \ldots$ ด้วย และการเทียบต้องดูระหว่างจุด

4. $m = 15$ มีตัวไม่ทราบ 15 ตัว และ K เป็นแถบสามแนวเฉียงที่มีสมาชิกไม่เป็นศูนย์ $3m - 2 = 43$ ตัวจากทั้งหมด 225 ตัว

5. ตัวอย่างนี้แรงกดกว้างเพียงราว 0.05 ตารางที่หยาบมาก ($m = 3, 7$) ไม่เห็นรูปร่างของแรง เมื่อ $m = 15$ ค่าคลาดเคลื่อนลงใต้ 10^{-3} ดังนั้นควรให้ชิ้นเล็กพอให้เห็นรูปร่างของแรงในบริเวณนั้น และควรดูตารางของตัวเองก่อนเชื่อผลลัพธ์ การทำชิ้นเล็กเฉพาะที่แรงแคบ (ตาราง m ไม่สม่ำเสมอ) เป็นข้อเสนอที่เราไม่ได้ทดสอบในบทนี้

ฉบับร่าง · Preprint · 2026-10-23

สรุปบทที่ 15 ในหน้าเดียว

วิธี	ค่าใช้จ่ายต่อก้าวหรือการแก้	เกณฑ์หรือผลสำคัญ	สิ่งที่ต้องระวัง
ลีปฟ รอก ของ สมการ คลื่น (15.1)	ค่าสองก้าวและเพื่อนบ้าน	$\lambda < \lambda_n^{(2)}$ บนตารางขอบตรึง	ตัวเลข คู แรนต์ λ เป็นอัตราส่วนคนละชนิดกับ μ
วิธี ชัด แจ้ง ของ ความ ร้อน (15.2)	ค่าเก่าและเพื่อนบ้าน	$\mu \le \mu_n \to \frac{1}{4}$	ระเบิดช้าเมื่อใกล้เกณฑ์
วิธีโดย ปริยาย ของ ความ ร้อน (15.2)	แก้ระบบเชิงเส้นทุกก้าว	ไม่ระเบิดทุก μ	ไม่ได้แม่นขึ้นเอง
การ ย้อน การ เบลอ B^{-k} (15.2)	ระบบเชิงเส้น	$\|B^{-k}\|_2 = \sin^{-2k}\frac{\pi}{2(n+1)}$	ไม่ถูกรูปแบบ
FEM รูปหมวกมิติเดียว (15.3)	แก้ระบบแถบสามแนวเฉียงครั้งเดียว	$\|u' - u_h'\| \le h\|f\|$	ค่าที่ โหนด แม่น เมื่อ ภาระ แม่นในมิติเดียว
ผลต่างกลาง (หัวข้อ 10.1)	ระบบ $\frac{1}{h^2}T$	ค่าที่โหนดลดเป็น h^2	ภาระที่ใช้คือ $f(x_l)$

สูตรที่ควรจำ: $\lambda_n^{(2)} = \dfrac{1}{\sqrt{2}\cos(\pi/(2(n+1)))}$ สำหรับลีปฟรอกสองมิติ; $\mu_n = \dfrac{1}{4\cos^2(\pi/(2(n+1)))}$ สำหรับวิธี ชัด แจ้ง; $K = \frac{1}{h}\mathrm{tridiag}(-1,2,-1)$ และ $\|u' - u_h'\| \le h\|f\|$ สำหรับ เชือก หนึ่ง มิติ; และ ค่า ลักษณะ เฉพาะ $\frac{4}{h}\sin^2\frac{j\pi}{2(m+1)}$ ของ K

ก่อนใช้เกณฑ์ใด ๆ ตรวจสามข้อ คือ เกณฑ์นั้นเป็นของมิติและขอบเขตเดียวกันหรือไม่ ตัวเลขคูแรนต์ λ ของหัวข้อ 15.1 กับ μ ของหัวข้อ 15.2 ไม่ปนกัน และค่าที่โหนดของ FEM แม่นเฉพาะเมื่อภาระแม่นในหนึ่งมิติ ตาราง 41×41 ของเราให้ $\lambda_{41}^{(2)} \approx 0.707602$ และ $\mu_{41} \approx 0.250350$ เมื่อเลือกระหว่างวิธีชัดแจ้งกับโดยปริยาย ให้ชั่งก้าวที่ยาวขึ้นกับต้นทุนของการแก้ระบบในแต่ละก้าว

อ่านต่อ. ISE สรุปวิธีสมาชิกจำกัดใน Iserles 2009, §9.1--9.2, pp. 171--192 โดยไม่พิสูจน์ส่วนทฤษฎีเต็ม ทฤษฎีของรูปอ่อนอยู่ใน Ern and Guermond 2004, บทที่ 2 ซึ่งว่าด้วยวิธีกาเลอคิน ปริภูมิประมาณแบบริตซ์–กาเลอคินอยู่ใน Braess 2007, §II.4, pp. 53--59 ส่วนตัวแก้ระบบอยู่ในบทที่ IV และ V ของหนังสือเล่มเดียวกัน

ฉบับร่าง · Preprint · 2026-10-23

บรรณานุกรม

Adenot, Paul and Hongchan Choi (2026). *Web Audio API 1.1*. Version Working Draft, 22 September 2026. URL: https://www.w3.org/TR/2026/WD-webaudio-1.1-20260922/ (visited on 10/22/2026).

Atkinson, Kendall E. (1989). *An Introduction to Numerical Analysis*. 2nd ed. John Wiley & Sons. ISBN: 9780471624899. URL: https://www.wiley.com/en-gb/an-introduction-to-numerical-analysis-2nd-edition-p-9780471624899.

Baraff, David and Andrew Witkin (1998). ``Large Steps in Cloth Simulation". In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. ACM Press, pp. 43–54. DOI: 10.1145/280814.280821.

Birtles, Brian, Dean Jackson, and Matt Rakow (2023). *CSS Easing Functions Level 1*. Version Candidate Recommendation Draft, 13 February 2023. URL: https://www.w3.org/TR/2023/CRD-css-easing-1-20230213/ (visited on 10/22/2026).

Board, Ariane 5 Flight 501 Inquiry (July 19, 1996). *ARIANE 5 Flight 501 Failure: Report by the Inquiry Board*. Inquiry Board report. URL: https://www-users.cse.umn.edu/~arnold/disasters/ariane5rep.html (visited on 10/22/2026).

Braess, Dietrich (2007). *Finite Elements: Theory, Fast Solvers, and Applications in Elasticity Theory*. 3rd ed. Cambridge University Press. ISBN: 9780521705189. URL: https://assets.cambridge.org/97805217/05189/frontmatter/9780521705189_frontmatter.pdf.

Brin, Sergey and Lawrence Page (1998). ``The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Computer Networks and ISDN Systems* 30.1–7, pp. 107–117. DOI: 10.1016/S0169-7552(98)00110-X. URL: https://doi.org/10.1016/S0169-7552(98)00110-X.

Casiez, Géry, Nicolas Roussel, and Daniel Vogel (2012). ``1 € filter: a simple speed-based low-pass filter for noisy input in interactive systems". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp. 2527–2530. DOI: 10.1145/2207676.2208639. URL: https://doi.org/10.1145/2207676.2208639.

Catmull, Edwin and Raphael Rom (1974). ``A Class of Local Interpolating Splines". In: *Computer Aided Geometric Design*. Ed. by Robert E. Barnhill and Richard F. Riesenfeld. Academic Press, pp. 317–326. DOI: 10.1016/B978-0-12-079050-0.50020-5. URL: https://doi.org/10.1016/B978-0-12-079050-0.50020-5.

Catto, Erin (2025). *Box2D: Simulation*. Version 3.2.0. URL: https://box2d.org/documentation/md_simulation.html (visited on 10/22/2026).

Cooley, James W. and John W. Tukey (1965). ``An Algorithm for the Machine Calculation of Complex Fourier Series". In: *Mathematics of Computation* 19.90, pp. 297–301. DOI: 10.2307/2003354. URL: https://doi.org/10.2307/2003354.

draft · preprint · 2026-10-23

Courant, Richard, Kurt Friedrichs, and Hans Lewy (1928). ``Über die partiellen Differenzengleichungen der mathematischen Physik". In: *Mathematische Annalen* 100.1, pp. 32–74. DOI: 10.1007/BF01448839. URL: https://doi.org/10.1007/BF01448839.

Curtiss, C. F. and J. O. Hirschfelder (1952). ``Integration of Stiff Equations". In: *Proceedings of the National Academy of Sciences* 38.3, pp. 235–243. DOI: 10.1073/pnas.38.3.235. URL: https://doi.org/10.1073/pnas.38.3.235.

Ern, Alexandre and Jean-Luc Guermond (2004). *Theory and Practice of Finite Elements*. 1st ed. Springer. ISBN: 9780387205748. DOI: 10.1007/978-1-4757-4355-5. URL: https://link.springer.com/book/10.1007/978-1-4757-4355-5.

Fiedler, Glenn (2004a). *Fix Your Timestep!* URL: https://gafferongames.com/post/fix_your_timestep/ (visited on 10/22/2026).

– (2004b). *Integration Basics*. URL: https://gafferongames.com/post/integration_basics/ (visited on 10/22/2026).

Games, Epic (2026). *Large World Coordinates in Unreal Engine 5*. Version 5.8. URL: https://dev.epicgames.com/documentation/en-us/unreal-engine/large-world-coordinates-in-unreal-engine-5 (visited on 10/22/2026).

Goldberg, David (1991). ``What Every Computer Scientist Should Know About Floating-Point Arithmetic". In: *ACM Computing Surveys* 23.1, pp. 5–48. DOI: 10.1145/103162.103163. URL: https://docs.oracle.com/cd/E19957-01/806-3568/ncg_goldberg.html.

Grossmann, Christian, Hans-Görg Roos, and Martin Stynes (2007). *Numerical Treatment of Partial Differential Equations*. 1st ed. Springer. ISBN: 9783540715825. DOI: 10.1007/978-3-540-71584-9. URL: https://link.springer.com/book/10.1007/978-3-540-71584-9.

Hesthaven, Jan S. and Tim Warburton (2008). *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. 1st ed. Springer. ISBN: 9780387720654. DOI: 10.1007/978-0-387-72067-8. URL: https://link.springer.com/book/10.1007/978-0-387-72067-8.

Hochbruck, Marlis (June 6, 2016). *Numerische Mathematik: Runge-Kutta-Verfahren*.

Iserles, Arieh (2009). *A First Course in the Numerical Analysis of Differential Equations*. 2nd ed. Cambridge University Press. ISBN: 9780521734905. URL: https://www.cambridge.org/core/books/first-course-in-the-numerical-analysis-of-differential-equations/2B4E05F5CFC58CFDC7BBBC6D1150661B.

Jakobsen, Thomas (2001). ``Advanced Character Physics". In: *Proceedings of the Game Developers Conference*. URL: https://edelkas.github.io/n/files/bib/jakobsen.pdf.

Kahan, W. (1965). ``Further remarks on reducing truncation errors". In: *Communications of the ACM* 8.1, p. 40. DOI: 10.1145/363707.363723. URL: https://doi.org/10.1145/363707.363723.

Keeling, Charles D. et al. (1976). ``Atmospheric carbon dioxide variations at Mauna Loa Observatory, Hawaii". In: *Tellus* 28.6, pp. 538–551. DOI: 10.1111/j.2153-3490.1976.tb00701.x. URL: https://doi.org/10.1111/j.2153-3490.1976.tb00701.x.

NOAA Global Monitoring Laboratory (2026). *Trends in Atmospheric Carbon Dioxide: Mauna Loa Monthly Mean Data*. URL: https://gml.noaa.gov/ccgg/trends/data.html (visited on 10/22/2026).

Office, U.S. General Accounting (1992). *Patriot Missile Defense: Software Problem Led to System Failure at Dhahran, Saudi Arabia*. Tech. rep. GAO/IMTEC-92-26. URL: https://www.gao.gov/products/imtec-92-26 (visited on 10/22/2026).

ฉบับร่าง Preprint 2026-10-23

OpenCV contributors (2026). *OpenCV calib3d.hpp: solvePnP documentation*. URL: https://raw.githubusercontent.com/opencv/opencv/4.x/modules/calib3d/include/opencv2/calib3d.hpp (visited on 10/22/2026).

Pérez, Patrick, Michel Gangnet, and Andrew Blake (2003). ``Poisson Image Editing". In: *ACM Transactions on Graphics* 22.3, pp. 313–318. DOI: 10.1145/1201775.882269.

Quinn, Kevin (Nov. 8, 1983). ``Ever Had Problems Rounding Off Figures? This Stock Exchange Has". In: *The Wall Street Journal*, p. 37. URL: https://www5.in.tum.de/~huckle/Vancouv.pdf.

Reed, Nathan (Oct. 21, 2021). *Visualizing Depth Precision*. NVIDIA. URL: https://developer.nvidia.com/blog/visualizing-depth-precision/ (visited on 10/22/2026).

Runge, Carl (1901). ``Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten". In: *Zeitschrift für Mathematik und Physik* 46, pp. 224–243. URL: https://gdz.sub.uni-goettingen.de/id/PPN600494829_0046.

Saad, Youcef and Martin H. Schultz (1986). ``GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems". In: *SIAM Journal on Scientific and Statistical Computing* 7.3, pp. 856–869. DOI: 10.1137/0907058.

Savitzky, Abraham and Marcel J. E. Golay (1964). ``Smoothing and Differentiation of Data by Simplified Least Squares Procedures". In: *Analytical Chemistry* 36.8, pp. 1627–1639. DOI: 10.1021/ac60214a047. URL: https://doi.org/10.1021/ac60214a047.

Software, id (1999). *Quake-III-Arena: code/game/q_math.c*. URL: https://github.com/id-Software/Quake-III-Arena/blob/master/code/game/q_math.c (visited on 10/22/2026).

Stam, Jos (2003). ``Real-Time Fluid Dynamics for Games". In: *Proceedings of the Game Developers Conference*. URL: https://www.dgp.toronto.edu/public_user/stam/reality/Research/pdf/GDC03.pdf.

Verlet, Loup (1967). ``Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules". In: *Physical Review* 159, pp. 98–103. DOI: 10.1103/PhysRev.159.98.

Vorst, Henk A. van der (2003). *Iterative Krylov Methods for Large Linear Systems*. Cambridge University Press. ISBN: 9780521818285. URL: https://www.cambridge.org/core/books/iterative-krylov-methods-for-large-linear-systems/FFB93854B3C47699F045AC396C0A208F.

ฉบับร่าง Preprint 2026-10-23

ดรรชนี (ไทย – English)

A-เสถียร (A-stable), 457

กฎเชิงสี่เหลี่ยมคางหมู (trapezoidal rule), 507
กฎของซิมป์สัน (Simpson's rule), 146, 436
กฎของบูล (Boole's rule), 169, 170
กฎเชิงสี่เหลี่ยมคางหมู (trapezoidal rule), 144, 411
กฎแบบประกอบ (composite rule), 144
กรวยมงจ์ (Monge cone), 488
กระบวนการแกรม–ชมิดท์ (Gram--Schmidt process), 268
กราฟ (graph), 240
กราฟของเมทริกซ์ (graph of a matrix), 239
กลวิธีมิลน์ (Milne device), 173, 445
การกำกัดแบบถ่วงน้ำหนักเต็ม (full weighting), 353
การกำจัดแบบเกาส์ (Gaussian elimination), 217
การค้นหาตามเส้น (line search), 364
การตรวจการชนแบบต่อเนื่อง (continuous collision detection), 416
การตัดทิ้ง (chopping), 32
การตัดทิ้ง (deflation), 311
การทะลุผ่าน (tunnelling), 416
การทับซ้อนความถี่ (aliasing), 124
การทำซ้ำแบบเกาส์–ไซเดล (Gauss--Seidel iteration), 220
การทำอัปวินด์ (upwinding), 487
การทำให้เป็นวิยุตบางส่วน (semi-discretization), 493
การประมาณค่านอกช่วงแบบริชาร์ดสัน (Richardson extrapolation), 167
การประมาณค่าเชิงตรีโกณมิติ (trigonometric interpolation), 122
การประมาณค่าในช่วง (interpolation), 60
การประมาณค่าในช่วงเชิงเส้น (linear interpolation), 62
การประมาณค่าในช่วงเชิงเส้นคู่ (bilinear interpolation), 353
การปรับปรุงคำตอบซ้ำ (iterative refinement), 233
การปรับเรียบก่อน (pre-smoothing), 354
การปรับเรียบหลัง (post-smoothing), 354
การปัดเศษแบบใกล้ที่สุด (rounding to nearest), 15
การผ่อน (easing), 84
การผ่อน (relaxation), 220
การรวมแบบคาฮาน (Kahan summation), 33
การรวมแบบคู่ (pairwise summation), 32
การรวมแบบตรง (sequential summation), 30
การรั่วของสเปกตรัม (spectral leakage), 123
การรีสตาร์ต (restart), 390
การลบแบบหักล้าง (cancellation), 38, 136
การลู่เข้า (convergence), 405
การวนซ้ำอาร์โนลดี (Arnoldi iteration), 317
การวนซ้ำแบบผกผัน (inverse iteration), 299
การสะท้อนแบบเฮาส์โฮลเดอร์ (Householder reflection), 265
การหดตัว (contraction), 183
การหมุนกีฟเวนส์ (Givens rotation), 266
การเก็บแบบแถวบีบอัด (compressed sparse row, CSR), 240
การเร่งแบบไอทเคน (Aitken extrapolation), 206
การเลื่อนค่า (shift), 300
การเลื่อนแบบวิลกินสัน (Wilkinson shift), 310
การแกว่งเท่ากัน (equioscillation), 113
การแก้ไขบนกริดหยาบ (coarse-grid correction), 354
การแปลงภาวะคล้าย (similarity transformation), 308
การแยกค่าเอกฐาน (singular value decomposition), 275

ฉบับร่าง · Preprint · 2026-05-23

การแยกตัวประกอบ LU (LU factorization), 218, 329
การแยกตัวประกอบ QR (QR factorization), 267
การแยกตัวประกอบโชเลสกี (Cholesky decomposition), 381
ก้าวย่อย (sub-step), 453
ก้าวย่อย (substep), 449
ก้าวแรกแบบเทย์เลอร์ (Taylor start), 484
ขนาดขั้น (step size), 134
ขอบเขตการพึ่งพา (domain of dependence), 487
ขั้นตอนวิธี QR (QR method), 309
ขั้นตอนวิธีคัธฮิลล์–แมกคีแบบกลับลำดับ (reverse Cuthill--McKee), 244
ควอดราเจอร์ (quadrature), 143, 434
ควอดราเจอร์แบบอะแดปทีฟ (adaptive quadrature), 167
ควอดราเจอร์แบบเกาส์ (Gaussian quadrature), 156
ความกว้างแถบ (bandwidth), 241
ความคลาดเคลื่อนยินยอม (tolerance), 170, 443
ความชันของชั้น (stage slope), 435
ความถี่ไนควิสต์ (Nyquist frequency), 124
ความยาวส่วนโค้ง (arc length), 158
ความสัมพันธ์เวียนเกิดสามพจน์ (three-term recurrence), 110
ความเยื้องศูนย์กลาง (eccentricity), 438
ความเอนเอียง (bias), 32
คอนดิชันนัมเบอร์ (condition number), 40, 50, 226, 228, 334
คำตอบกำลังสองน้อยสุด (least squares solution), 263
คีย์เฟรม (keyframe), 60
คู่วิธีรุงเงอ-คุททาแบบฝังตัว (embedded Runge--Kutta pair), 445
ค่าคงตัวการหดตัว (contraction constant), 183
ค่าคลาดเคลื่อนการประมาณค่า (interpolation error), 72
ค่าคลาดเคลื่อนการปัดเศษ (round-off error), 29
ค่าคลาดเคลื่อนตัดปลาย (truncation error), 135
ค่าคลาดเคลื่อนตัดปลายเฉพาะที่ (local truncation error), 404, 436
ค่าคลาดเคลื่อนย้อนกลับ (backward error), 51
ค่าคลาดเคลื่อนวงกว้าง (global error), 404
ค่าคลาดเคลื่อนสัมบูรณ์ (absolute error), 28, 220
ค่าคลาดเคลื่อนสัมพัทธ์ (relative error), 28, 346
ค่าคลาดเคลื่อนไปข้างหน้า (forward error), 51, 136
ค่าตกค้าง (residual), 99, 206, 221, 263, 346
ค่าริตซ์ (Ritz value), 319
ค่าลักษณะเฉพาะ (eigenvalue), 162, 227, 286
ค่าเดาตั้งต้น (initial guess), 179
ค่าเอกฐาน (singular value), 275
ค่าไม่ใช่ตัวเลข (NaN), 20
จานเกอร์ชโกริน (Gerschgorin disc), 288
จาโคบีแบบถ่วง (weighted Jacobi), 352
จำนวนการดำเนินการ (operation count), 219
จำนวนเต็มที่เก็บได้ตรง ๆ (exact integers), 17
จุดตรึง (fixed point), 183, 253, 411
จุดนำทาง (waypoint), 85
จุดยอด (vertex), 240
จุดไกลดาวที่สุด (apoapsis), 438
ชักตัวอย่าง (sampling), 416
ช่วงล้อม (bracket), 181
ช่องความถี่ (frequency bin), 118
ซี-ไฟต์ติง (z-fighting), 49
ตัวขยายความคลาดเคลื่อนจากเรขาคณิต (dilution of precision), 268
ตัวคูณหมุน (twiddle factor), 125
ตัวดำเนินการลาปลาซแบบห้าจุด (five-point Laplace operator), 484
ตัวประกอบการหน่วง (damping factor), 296
ตัวประมาณค่าคลาดเคลื่อน (error estimate), 444
ตัวประมาณปาเดอ (Padé approximant), 474
ตัวปรับสภาพเงื่อนไข (preconditioner), 382
ตัวปรับเรียบ (smoother), 352
ตัวผกผันเทียม (pseudo-inverse), 277
ตัวสะสมเวลา (accumulator), 448
ตัวหลัก (pivot), 217
ตัวเลขคูแรนต์ (Courant number), 484, 493
ตารางบุชเชอร์ (Butcher tableau), 435
ต้นไม้ (tree), 243
ถูกรูปแบบ (well posed), 498
ทฤษฎีบทจุดตรึงของบานาค (Banach fixed-point theorem), 411
ทฤษฎีบทออสตรอฟสกี–ไรช์ (Ostrowski--Reich

ฉบับร่าง · Preprint · 2026-10-23

theorem), 346
ที่มีภาวะรากซ้ำ (multiplicity), 196
นอร์ม L_2 (L2 norm), 109
นอร์มของเวกเตอร์ (vector norm), 221
นอร์มค่าสูงสุด (maximum norm), 109
นอร์มพลังงาน (energy norm), 342, 363, 505
นอร์มสอง (2-norm), 221
นอร์มสเปกตรัม (spectral norm), 276
นอร์มหนึ่ง (1-norm), 221
นอร์มอนันต์ (infinity norm), 221
นอร์มเมทริกซ์ที่ถูกเหนี่ยวนำ (induced matrix norm), 221
นิวตันแบบหน่วง (damped Newton's method), 204
น้ำหนักของสูตร (quadrature weight), 144
น้ำหนักของหน้าต่าง (window weights), 103
บวกแน่นอน (positive definite), 242
บัฟเฟอร์ความลึก (depth buffer), 49
บิต (bit), 4
ปรากฏการณ์รุงเง (Runge phenomenon), 76
ปริภูมิย่อยไครลอฟ (Krylov subspace), 317, 367
ปริภูมิโซโบเลฟ (Sobolev space), 503
ปริภูมิไฟไนต์เอลิเมนต์แบบคอนฟอร์มมิง (conforming finite element space), 503
ปัญหากำลังสองน้อยสุด (least squares problem), 99, 263
ปัญหาค่าเริ่มต้น (initial value problem), 403
ปัญหาที่ไม่ถูกรูปแบบ (ill-posed problem), 497
ปัญหาสภาวะเลว (ill-conditioned problem), 50
ปัญหาเคปเลอร์ (Kepler problem), 438
ปัดเข้าหาเลขคู่ (ties to even), 15
ผลการแปลงฟูเรียร์แบบเร็ว (fast Fourier transform), 125
ผลการแปลงฟูเรียร์แบบไม่ต่อเนื่อง (discrete Fourier transform), 120
ผลคูณจุด (dot product), 38
ผลคูณภายใน (inner product), 110
ผลคูณโครเนกเกอร์ (Kronecker product), 334
ผลต่างหาร (divided difference), 65
ผลต่างอันตะ (finite difference), 332
ผลบวกย่อย (partial sum), 29
ผลบวกโครเนกเกอร์ (Kronecker sum), 396
ผลหารเชิงผลต่าง (difference quotient), 197
ผลหารเชิงผลต่างกลาง (central difference quotient), 134
ผลหารเชิงผลต่างไปข้างหน้า (forward difference quotient), 134
ผลหารเรย์ลี (Rayleigh quotient), 289
ผลเฉลยแบบอ่อน (weak solution), 502
พลังงานต่อหน่วยมวล (energy per unit mass), 439
พหุนามตรีโกณมิติ (trigonometric polynomial), 122
พหุนามประมาณค่าในช่วง (interpolation polynomial), 62
พหุนามลากรานจ์ (Lagrange polynomial), 64
พหุนามเชบีเชฟ (Chebyshev polynomial), 75, 373
พหุนามเชบีเชฟ (Chebyshev polynomials), 110
พหุนามเชิงตั้งฉาก (orthogonal polynomials), 110, 160
พหุนามเลอฌ็องดร์ (Legendre polynomials), 110
พหุนามโหนด (node polynomial), 73
พิกเซล (pixel), 226
ฟังก์ชันกำลังสอง (quadratic function), 363
ฟังก์ชันถ่วงน้ำหนัก (weight function), 110
ฟังก์ชันทดสอบ (test function), 502
ฟังก์ชันรูปหมวก (hat function), 503
ฟังก์ชันฮาร์มอนิก (harmonic function), 369
ฟังก์ชันเสถียรภาพ (stability function), 456
ฟิลอิน (fill-in), 239, 242, 335
มาตรฐาน IEEE 754 (IEEE 754), 5
มินิแมกซ์ (minimax), 113
มุมเซนิท (zenith angle), 48
รอบแบบ V (V-cycle), 354
ระดับความแม่นยำ (degree of precision), 147
ระบบที่มีสภาวะเลว (ill-conditioned system), 229
ระบบสมการเกิน (overdetermined system), 263
ระบบสมการเชิงเส้น (linear system), 217
รัศมีสเปกตรัม (spectral radius), 253, 341
ราก (root), 181
รากปฐมฐานของหนึ่งอันดับ N (primitive root of unity), 119
รีจิสเตอร์ (register), 9
รูปแบบลากรานจ์ (Lagrange form), 64

ฉบับร่าง · Preprint 2026-10-23

รูปแบบเชิงเส้นคู่ (bilinear form), 502
ลำดับชั้นของเมทริกซ์ (rank), 275
ลิพชิทซ์ (Lipschitz), 183
ลู่เข้าแบบเชิงเส้น (linear convergence), 186
วิธีกาเลอคิน (Galerkin method), 503
วิธีกาเลอคินแบบไม่ต่อเนื่อง (discontinuous Galerkin method), 489
วิธีกำลัง (power method), 296
วิธีกึ่งนิวตัน (quasi-Newton method), 253
วิธีค่าตกค้างน้อยสุดนัยทั่วไป (generalized minimal residual method (GMRES)), 361
วิธีชัดแจ้ง (explicit), 493
วิธีทำซ้ำเชบีเชฟ (Chebyshev iterative method), 356
วิธีทำซ้ำเชิงเส้นขั้นเดียวแบบคงที่ (linear one-step stationary scheme), 341
วิธีทีตา (theta method), 411
วิธีนิวตัน (Newton's method), 192, 251
วิธีนิวตัน-ราฟสัน (Newton–Raphson method), 464
วิธีนิวตัน-ราฟสันแบบดัดแปลง (modified Newton--Raphson method), 253
วิธีนิวตัน–โคตส์ (Newton–Cotes rule), 150
วิธีผสม (hybrid method), 207
วิธีรุงเงอ-คุททา (Runge--Kutta method), 435
วิธีรุงเงอ-คุททาแบบโดยปริยาย (implicit Runge–Kutta (IRK)), 472
วิธีลงโทษ (penalty method), 467
วิธีลานซอส (Lanczos method), 318
วิธีลีปฟรอก (leapfrog scheme), 484
วิธีสมาชิกจำกัด (finite element method), 501
วิธีสองกริด (two-grid method), 354
วิธีหลายกริด (multigrid method), 351
วิธีหลายกริดเต็มรูป (full multigrid), 356
วิธีหลายขั้น (multistep method), 465
วิธีหลายขั้นเชิงเส้น (linear multistep method), 422
วิธีอดัมส์-บาชฟอร์ท (Adams--Bashforth method), 422
วิธีออยเลอร์ (Euler method), 401, 404
วิธีออยเลอร์ซิมเพลกติก (symplectic Euler method), 413
วิธีออยเลอร์ถอยหลัง (backward Euler method), 464, 496
วิธีเกรเดียนต์สังยุค (conjugate gradients), 361
วิธีเกรเดียนต์สังยุคที่ปรับสภาพเงื่อนไข (preconditioned conjugate gradients), 382
วิธีเกาส์–เลอฌ็องดร์ (Gauss–Legendre method), 473
วิธีเซแคนต์ (secant method), 197
วิธีเวอร์เลต (Verlet integration), 422
วิธีเส้น (method of lines), 493
วิธีแบบคอนฟอร์มมิง (conforming method), 503
วิธีแบบโดยปริยาย (implicit method), 410, 464, 496
วิธีแบ่งครึ่ง (bisection method), 159, 181
วิธีแลกเปลี่ยนจุด (exchange method), 114
สกัดด์ (Scud), 9
สมการการพา (advection equation), 487
สมการกาเลอคิน (Galerkin equations), 503
สมการคลื่น (wave equation), 484
สมการความร้อน (heat equation), 492
สมการตัวตึง (stiff equation), 457
สมการปรกติ (normal equations), 100, 264
สมการปัวซง (Poisson equation), 331
สมการเชิงอนุพันธ์สามัญ (ordinary differential equation (ODE)), 403
สมการเชิงอนุพันธ์สามัญ (ordinary differential equation), 493
สังยุค (complex conjugate), 121
สังยุคเทียบ A (conjugate with respect to A), 366
สัญญาณรบกวน (noise), 226
สูตร BDF (backward differentiation formula (BDF)), 464, 465
สูตรห้าจุด (five-point formula), 332
สูตรออยเลอร์-แมคลอริน (Euler--Maclaurin formula), 151
สเปกตรัม (spectrum), 286
สไปลน์ลูกบาศก์ (cubic spline), 84
ส่วนยกกำลัง (exponent), 4
ส่วนเศษ (fraction field), 5
หน่วยปัด (unit roundoff), 15
หน้าต่างเบลแมน (Blackman window), 118
หน้าต่างเลื่อน (sliding window), 102
หุ่นแร็กดอลล์ (ragdoll), 453
อนันต์ (infinity), 19

ฉบับร่าง · Preprint · 2026-10-23

อนุกรมเชบีเชฟ (Chebyshev series), 113
อนุพันธ์ (derivative), 38, 132
อัตราการลู่เข้า (rate of convergence), 186
อัตราส่วนความตึง (stiffness ratio), 458
อันดับของการลู่เข้า (order of convergence), 186
อันดับของวิธี (order of a method), 404, 436
อันเดอร์โฟลว์ (underflow), 19
อัลกอริทึมโกลับ–เวลช์ (Golub--Welsch algorithm), 162
อาร์กโคไซน์ (arccos), 38
อีควอไลเซอร์ (equaliser), 118
เกาส์–นิวตัน (Gauss--Newton), 268
เคอร์เนลเปอาโน (Peano kernel), 147
เงื่อนไข CFL (Courant–Friedrichs–Lewy condition), 487
เงื่อนไขขอบแบบดีรีเคล (Dirichlet boundary condition), 483
เงื่อนไขราก (root condition), 424
เงื่อนไขลิพชิทซ์ (Lipschitz condition), 403
เชิงตั้งฉาก (orthogonal), 110
เด่นทแยงแบบลดทอนไม่ได้ (irreducibly diagonally dominant), 343
เบลอ (blur), 226
เป็นคาบ (periodic), 119
เป็นอิสระเชิงเส้น (linearly independent), 101
เมทริกซ์การกำกัด (restriction matrix), 353
เมทริกซ์การขยาย (prolongation matrix), 353
เมทริกซ์การทำซ้ำ (iteration matrix), 341
เมทริกซ์การเรียงสับเปลี่ยน (permutation matrix), 232
เมทริกซ์ของสูตรห้าจุด (five-point matrix), 332
เมทริกซ์จาโคบี (Jacobi matrix), 162
เมทริกซ์จาโคเบียน (Jacobian matrix), 250, 268
เมทริกซ์ที่เวกเตอร์ลักษณะเฉพาะไม่ครบ (defective matrix), 291
เมทริกซ์ปรกติ (normal matrix), 487
เมทริกซ์มากเลขศูนย์ (sparse matrix), 215, 329
เมทริกซ์ลำดับชั้นหนึ่ง (rank-one matrix), 275
เมทริกซ์สติฟเนส (stiffness matrix), 458, 504
เมทริกซ์สมมาตรเสมือน (skew-symmetric matrix), 396
เมทริกซ์สามแนวเฉียง (tridiagonal matrix), 162, 217, 286, 333
เมทริกซ์ฮิลเบิร์ต (Hilbert matrix), 109
เมทริกซ์เชิงตั้งฉาก (orthogonal matrix), 265, 458, 485
เมทริกซ์เด่นทแยงอย่างเข้ม (strictly diagonally dominant matrix), 343
เมทริกซ์เลื่อนวงกลม (cyclic shift matrix), 393
เมทริกซ์เอ็ม (M-matrix), 384
เมทริกซ์เฮสเซนเบิร์ก (Hessenberg matrix), 308
เมทริกซ์แถบ (banded matrix), 241
เมทริกซ์แวนเดอร์มอนด์ (Vandermonde matrix), 63
เมทริกซ์ไม่เอกฐาน (nonsingular matrix), 217
เรย์ (ray), 47
เรียงสับเปลี่ยน (permutation), 244
เลขคลื่น (wavenumber), 352
เลขความแม่นคู่ (double precision), 6
เลขความแม่นเดี่ยว (single precision), 2, 132
เลขความแม่นเดี่ยวปกติ (normal float32), 7
เลขจุดคงที่ (fixed-point number), 4
เลขจุดลอยตัว (floating-point number), 2
เลขซับนอร์มอล (subnormal number), 19
เลขฐานสอง (binary number), 4
เลขนัยสำคัญ (significant digits), 29
เลขปกติ (normal number), 7
เลขเปเคลต์ (Péclet number), 390
เลือกตัวหลักตามหลัก (partial pivoting), 232
เวกเตอร์ภาระ (load vector), 504
เวกเตอร์ริตซ์ (Ritz vector), 319
เวกเตอร์ลักษณะเฉพาะ (eigenvector), 286
เวกเตอร์หนึ่งหน่วย (unit vector), 38
เส้นเชื่อม (edge), 240
เส้นโค้งเบซิเยร์ (Bézier curve), 84
เอปซิลอนของเครื่อง (machine epsilon), 15
แคทมัลล์–รอม (Catmull--Rom), 85
แบบกำลังสอง (quadratic convergence), 195
แมนทิสซา (significand), 4
แอร์มีต (Hermite), 84
แอ่งดึงดูด (basin of attraction), 203
โคเออร์ซีฟ (coercive), 505
โชเลสกี (Cholesky factorization), 242
โดเมนเสถียรภาพเชิงเส้น (linear stability domain), 456
โรมเบิร์ก (Romberg integration), 167

ฉบับร่าง · Preprint · 2026-10-23

โหนด (node), 62, 144
โหนดเชบีเชฟ (Chebyshev nodes), 75
โอเวอร์โฟลว์ (overflow), 19
ไบอัส (bias), 6
ไฟไนต์เอลิเมนต์ (finite element), 501

ฉบับร่าง · Preprint · 2026-10-23

ดรรชนี (English – ไทย)

1-norm (นอร์มหนึ่ง), 221
2-norm (นอร์มสอง), 221

A-stable (A-เสถียร), 457
absolute error (ค่าคลาดเคลื่อนสัมบูรณ์), 28, 220
accumulator (ตัวสะสมเวลา), 448
Adams--Bashforth method (วิธีอดัมส์-บาชฟอร์ท), 422
adaptive quadrature (ควอดราเจอร์แบบอะแดปทีฟ), 167
advection equation (สมการการพา), 487
Aitken extrapolation (การเร่งแบบไอทเคน), 206
aliasing (การทับซ้อนความถี่), 124
apoapsis (จุดไกลดาวที่สุด), 438
arc length (ความยาวส่วนโค้ง), 158
arccos (อาร์กโคไซน์), 38
Arnoldi iteration (การวนซ้ำอาร์โนลดี), 317

backward differentiation formula (BDF) (สูตร BDF), 464, 465
backward error (ค่าคลาดเคลื่อนย้อนกลับ), 51
backward Euler method (วิธีออยเลอร์ถอยหลัง), 464, 496
Banach fixed-point theorem (ทฤษฎีบทจุดตรึงของบานาค), 411
banded matrix (เมทริกซ์แถบ), 241
bandwidth (ความกว้างแถบ), 241
basin of attraction (แอ่งดึงดูด), 203
bias (ความเอนเอียง), 32
bias (ไบอัส), 6
bilinear form (รูปแบบเชิงเส้นคู่), 502
bilinear interpolation (การประมาณค่าในช่วงเชิงเส้นคู่), 353
binary number (เลขฐานสอง), 4
bisection method (วิธีแบ่งครึ่ง), 159, 181
bit (บิต), 4
Blackman window (หน้าต่างเบลแมน), 118
blur (เบลอ), 226
Boole's rule (กฎของบูล), 169, 170
bracket (ช่วงล้อม), 181
Butcher tableau (ตารางบุชเชอร์), 435
Bézier curve (เส้นโค้งเบซิเยร์), 84

cancellation (การลบแบบหักล้าง), 38, 136
Catmull--Rom (แคทมัลล์–รอม), 85
central difference quotient (ผลหารเชิงผลต่างกลาง), 134
Chebyshev iterative method (วิธีทำซ้ำเชบีเชฟ), 356
Chebyshev nodes (โหนดเชบีเชฟ), 75
Chebyshev polynomial (พหุนามเชบีเชฟ), 75, 373
Chebyshev polynomials (พหุนามเชบีเชฟ), 110
Chebyshev series (อนุกรมเชบีเชฟ), 113
Cholesky decomposition (การแยกตัวประกอบโชเลสกี), 381
Cholesky factorization (โชเลสกี), 242
chopping (การตัดทิ้ง), 32
coarse-grid correction (การแก้ไขบนกริดหยาบ), 354
coercive (โคเออร์ซีฟ), 505
complex conjugate (สังยุค), 121
composite rule (กฎแบบประกอบ), 144
compressed sparse row, CSR (การเก็บแบบแถวบีบอัด), 240
condition number (คอนดิชันนัมเบอร์), 40, 50, 226, 228, 334

ฉบับร่าง • Preprint • 2026-10-23

conforming finite element space (ปริภูมิไฟไนต์เอลิเมนต์แบบคอนฟอร์มมิง), 503
conforming method (วิธีแบบคอนฟอร์มมิง), 503
conjugate gradients (วิธีเกรเดียนต์สังยุค), 361
conjugate with respect to A (สังยุคเทียบ A), 366
continuous collision detection (การตรวจการชนแบบต่อเนื่อง), 416
contraction (การหดตัว), 183
contraction constant (ค่าคงตัวการหดตัว), 183
convergence (การลู่เข้า), 405
Courant number (ตัวเลขคูแรนต์), 484, 493
Courant–Friedrichs–Lewy condition (เงื่อนไข CFL), 487
cubic spline (สไปลน์ลูกบาศก์), 84
cyclic shift matrix (เมทริกซ์เลื่อนวงกลม), 393

damped Newton's method (นิวตันแบบหน่วง), 204
damping factor (ตัวประกอบการหน่วง), 296
defective matrix (เมทริกซ์ที่เวกเตอร์ลักษณะเฉพาะไม่ครบ), 291
deflation (การตัดทิ้ง), 311
degree of precision (ระดับความแม่นยำ), 147
depth buffer (บัฟเฟอร์ความลึก), 49
derivative (อนุพันธ์), 38, 132
difference quotient (ผลหารเชิงผลต่าง), 197
dilution of precision (ตัวขยายความคลาดเคลื่อนจากเรขาคณิต), 268
Dirichlet boundary condition (เงื่อนไขขอบแบบดีรีเคล), 483
discontinuous Galerkin method (วิธีกาเลอคินแบบไม่ต่อเนื่อง), 489
discrete Fourier transform (ผลการแปลงฟูเรียร์แบบไม่ต่อเนื่อง), 120
divided difference (ผลต่างหาร), 65
domain of dependence (ขอบเขตการพึ่งพา), 487
dot product (ผลคูณจุด), 38
double precision (เลขความแม่นคู่), 6

easing (การผ่อน), 84
eccentricity (ความเยื้องศูนย์กลาง), 438
edge (เส้นเชื่อม), 240
eigenvalue (ค่าลักษณะเฉพาะ), 162, 227, 286
eigenvector (เวกเตอร์ลักษณะเฉพาะ), 286
embedded Runge--Kutta pair (คู่วิธีรุงเงอ-คุททาแบบฝังตัว), 445
energy norm (นอร์มพลังงาน), 342, 363, 505
energy per unit mass (พลังงานต่อหน่วยมวล), 439
equaliser (อีควอไลเซอร์), 118
equioscillation (การแกว่งเท่ากัน), 113
error (ค่าคลาดเคลื่อนสัมบูรณ์), 221
error estimate (ตัวประมาณค่าคลาดเคลื่อน), 444
Euler method (วิธีออยเลอร์), 401, 404
Euler--Maclaurin formula (สูตรออยเลอร์-แมคลอริน), 151
exact integers (จำนวนเต็มที่เก็บได้ตรง ๆ), 17
exchange method (วิธีแลกเปลี่ยนจุด), 114
explicit (วิธีชัดแจ้ง), 493
exponent (ส่วนยกกำลัง), 4

fast Fourier transform (ผลการแปลงฟูเรียร์แบบเร็ว), 125
fill-in (ฟิลอิน), 239, 242, 335
finite difference (ผลต่างอันตะ), 332
finite element (ไฟไนต์เอลิเมนต์), 501
finite element method (วิธีสมาชิกจำกัด), 501
five-point formula (สูตรห้าจุด), 332
five-point Laplace operator (ตัวดำเนินการลาปลาซแบบห้าจุด), 484
five-point matrix (เมทริกซ์ของสูตรห้าจุด), 332
fixed point (จุดตรึง), 183, 253, 411
fixed-point number (เลขจุดคงที่), 4
floating-point number (เลขจุดลอยตัว), 2
forward difference quotient (ผลหารเชิงผลต่างไปข้างหน้า), 134
forward error (ค่าคลาดเคลื่อนไปข้างหน้า), 51, 136
fraction field (ส่วนเศษ), 5
frequency bin (ช่องความถี่), 118
full multigrid (วิธีหลายกริดเต็มรูป), 356
full weighting (การกำกัดแบบถ่วงน้ำหนักเต็ม), 353

Galerkin equations (สมการกาเลอคิน), 503
Galerkin method (วิธีกาเลอคิน), 503

ฉบับร่าง Preprint 2026-03-23

Gauss--Newton (เกาส์–นิวตัน), 268
Gauss--Seidel iteration (การทำซ้ำแบบเกาส์–ไซเดล), 220
Gaussian elimination (การกำจัดแบบเกาส์), 217
Gaussian quadrature (ควอดราเจอร์แบบเกาส์), 156
Gauss–Legendre method (วิธีเกาส์–เลอฌ็องดร์), 473
generalized minimal residual method (GMRES) (วิธีค่าตกค้างน้อยสุดนัยทั่วไป), 361
Gerschgorin disc (จานเกอร์ชโกริน), 288
Givens rotation (การหมุนกีฟเวนส์), 266
global error (ค่าคลาดเคลื่อนวงกว้าง), 404
Golub--Welsch algorithm (อัลกอริทึมโกลับ–เวลช์), 162
Gram--Schmidt process (กระบวนการแกรม–ชมิดท์), 268
graph (กราฟ), 240
graph of a matrix (กราฟของเมทริกซ์), 239

harmonic function (ฟังก์ชันฮาร์มอนิก), 369
hat function (ฟังก์ชันรูปหมวก), 503
heat equation (สมการความร้อน), 492
Hermite (แอร์มีต), 84
Hessenberg matrix (เมทริกซ์เฮสเซนเบิร์ก), 308
Hilbert matrix (เมทริกซ์ฮิลเบิร์ต), 109
Householder reflection (การสะท้อนแบบเฮาส์โฮลเดอร์), 265
hybrid method (วิธีผสม), 207

IEEE 754 (มาตรฐาน IEEE 754), 5
ill-conditioned problem (ปัญหาสภาวะเลว), 50
ill-conditioned system (ระบบที่มีสภาวะเลว), 229
ill-posed problem (ปัญหาที่ไม่ถูกรูปแบบ), 497
implicit method (วิธีแบบโดยปริยาย), 410, 464, 496
implicit Runge–Kutta (IRK) (วิธีรุงเงอ-คุททาแบบโดยปริยาย), 472
induced matrix norm (นอร์มเมทริกซ์ที่ถูกเหนี่ยวนำ), 221
infinity (อนันต์), 19
infinity norm (นอร์มอนันต์), 221
initial guess (ค่าเดาตั้งต้น), 179
initial value problem (ปัญหาค่าเริ่มต้น), 403
inner product (ผลคูณภายใน), 110
interpolation (การประมาณค่าในช่วง), 60
interpolation error (ค่าคลาดเคลื่อนการประมาณค่า), 72
interpolation polynomial (พหุนามประมาณค่าในช่วง), 62
inverse iteration (การวนซ้ำแบบผกผัน), 299
irreducibly diagonally dominant (เด่นทแยงแบบลดทอนไม่ได้), 343
iteration matrix (เมทริกซ์การทำซ้ำ), 341
iterative refinement (การปรับปรุงคำตอบซ้ำ), 233

Jacobi matrix (เมทริกซ์จาโคบี), 162
Jacobian matrix (เมทริกซ์จาโคเบียน), 250, 268

Kahan summation (การรวมแบบคาฮาน), 33
Kepler problem (ปัญหาเคปเลอร์), 438
keyframe (คีย์เฟรม), 60
Kronecker product (ผลคูณโครเนกเกอร์), 334
Kronecker sum (ผลบวกโครเนกเกอร์), 396
Krylov subspace (ปริภูมิย่อยไครลอฟ), 317, 367

L2 norm (นอร์ม L_2), 109
Lagrange form (รูปแบบลากรานจ์), 64
Lagrange polynomial (พหุนามลากรานจ์), 64
Lanczos method (วิธีลานซอส), 318
leapfrog scheme (วิธีลีปฟรอก), 484
least squares problem (ปัญหากำลังสองน้อยสุด), 99, 263
least squares solution (คำตอบกำลังสองน้อยสุด), 263
Legendre polynomials (พหุนามเลอฌ็องดร์), 110
line search (การค้นหาตามเส้น), 364
linear convergence (ลู่เข้าแบบเชิงเส้น), 186
linear interpolation (การประมาณค่าในช่วงเชิงเส้น), 62
linear multistep method (วิธีหลายขั้นเชิงเส้น), 422
linear one-step stationary scheme (วิธีทำซ้ำเชิงเส้นขั้นเดียวแบบคงที่), 341
linear stability domain (โดเมนเสถียรภาพเชิงเส้น), 456
linear system (ระบบสมการเชิงเส้น), 217
linearly independent (เป็นอิสระเชิงเส้น), 101
Lipschitz (ลิพชิทซ์), 183
Lipschitz condition (เงื่อนไขลิพชิทซ์), 403
load vector (เวกเตอร์ภาระ), 504

ฉบับร่าง · Preprint 2026-10-26

local truncation error (ค่าคลาดเคลื่อนตัดปลายเฉพาะที่), 404, 436
LU factorization (การแยกตัวประกอบ LU), 218, 329

M-matrix (เมทริกซ์เอ็ม), 384
machine epsilon (เอปซิลอนของเครื่อง), 15
maximum norm (นอร์มค่าสูงสุด), 109
method of lines (วิธีเส้น), 493
Milne device (กลวิธีมิลน์), 173, 445
minimax (มินิแมกซ์), 113
modified Newton--Raphson method (วิธีนิวตัน-ราฟสันแบบดัดแปลง), 253
Monge cone (กรวยมงจ์), 488
multigrid method (วิธีหลายกริด), 351
multiplicity (ที่มีภาวะรากซ้ำ), 196
multistep method (วิธีหลายขั้น), 465

NaN (ค่าไม่ใช่ตัวเลข), 20
Newton's method (วิธีนิวตัน), 192, 251
Newton–Cotes rule (วิธีนิวตัน–โคตส์), 150
Newton–Raphson method (วิธีนิวตัน-ราฟสัน), 464
node (โหนด), 62, 144
node polynomial (พหุนามโหนด), 73
noise (สัญญาณรบกวน), 226
nonsingular matrix (เมทริกซ์ไม่เอกฐาน), 217
normal equations (สมการปรกติ), 100, 264
normal float32 (เลขความแม่นเดี่ยวปกติ), 7
normal matrix (เมทริกซ์ปรกติ), 487
normal number (เลขปกติ), 7
Nyquist frequency (ความถี่ไนควิสต์), 124

operation count (จำนวนการดำเนินการ), 219
order of a method (อันดับของวิธี), 404, 436
order of convergence (อันดับของการลู่เข้า), 186
ordinary differential equation (ODE) (สมการเชิงอนุพันธ์สามัญ), 403
ordinary differential equation (สมการเชิงอนุพันธ์สามัญ), 493
orthogonal (เชิงตั้งฉาก), 110
orthogonal matrix (เมทริกซ์เชิงตั้งฉาก), 265, 458, 485
orthogonal polynomials (พหุนามเชิงตั้งฉาก), 110, 160
Ostrowski--Reich theorem (ทฤษฎีบทออสตรอฟสกี–ไรช์), 346
overdetermined system (ระบบสมการเกิน), 263
overflow (โอเวอร์โฟลว์), 19

Padé approximant (ตัวประมาณปาเดอ), 474
pairwise summation (การรวมแบบคู่), 32
partial pivoting (เลือกตัวหลักตามหลัก), 232
partial sum (ผลบวกย่อย), 29
Peano kernel (เคอร์เนลเปอาโน), 147
penalty method (วิธีลงโทษ), 467
periodic (เป็นคาบ), 119
permutation (เรียงสับเปลี่ยน), 244
permutation matrix (เมทริกซ์การเรียงสับเปลี่ยน), 232
pivot (ตัวหลัก), 217
pixel (พิกเซล), 226
Poisson equation (สมการปัวซง), 331
positive definite (บวกแน่นอน), 242
post-smoothing (การปรับเรียบหลัง), 354
power method (วิธีกำลัง), 296
pre-smoothing (การปรับเรียบก่อน), 354
preconditioned conjugate gradients (วิธีเกรเดียนต์สังยุคที่ปรับสภาพเงื่อนไข), 382
preconditioner (ตัวปรับสภาพเงื่อนไข), 382
primitive root of unity (รากปฐมฐานของหนึ่งอันดับ N), 119
prolongation matrix (เมทริกซ์การขยาย), 353
pseudo-inverse (ตัวผกผันเทียม), 277
Péclet number (เลขเปเคลต์), 390

QR factorization (การแยกตัวประกอบ QR), 267
QR method (ขั้นตอนวิธี QR), 309
quadratic convergence (แบบกำลังสอง), 195
quadratic function (ฟังก์ชันกำลังสอง), 363
quadrature (ควอดราเจอร์), 143, 434
quadrature weight (น้ำหนักของสูตร), 144
quasi-Newton method (วิธีกึ่งนิวตัน), 253

ragdoll (หุ่นแร็กดอลล์), 453
rank (ลำดับชั้นของเมทริกซ์), 275
rank-one matrix (เมทริกซ์ลำดับชั้นหนึ่ง), 275

ฉบับร่าง · Preprint · 2026-10-23

rate of convergence (อัตราการลู่เข้า), 186
ray (เรย์), 47
Rayleigh quotient (ผลหารเรย์ลี), 289
register (รีจิสเตอร์), 9
relative error (ค่าคลาดเคลื่อนสัมพัทธ์), 28, 346
relaxation (การผ่อน), 220
residual (ค่าตกค้าง), 99, 206, 221, 263, 346
restart (การรีสตาร์ต), 390
restriction matrix (เมทริกซ์การกำกัด), 353
reverse Cuthill--McKee (ขั้นตอนวิธีคัธฮิลล์–แมกคีแบบกลับลำดับ), 244
Richardson extrapolation (การประมาณค่านอกช่วงแบบริชาร์ดสัน), 167
Ritz value (ค่าริตซ์), 319
Ritz vector (เวกเตอร์ริตซ์), 319
Romberg integration (โรมเบิร์ก), 167
root (ราก), 181
root condition (เงื่อนไขราก), 424
round-off error (ค่าคลาดเคลื่อนการปัดเศษ), 29
rounding to nearest (การปัดเศษแบบใกล้ที่สุด), 15
Runge phenomenon (ปรากฏการณ์รุงเง), 76
Runge--Kutta method (วิธีรุงเงอ-คุททา), 435

sampling (ชักตัวอย่าง), 416
Scud (สกัดด์), 9
secant method (วิธีเซแคนต์), 197
semi-discretization (การทำให้เป็นวิยุตบางส่วน), 493
sequential summation (การรวมแบบตรง), 30
shift (การเลื่อนค่า), 300
significand (แมนทิสซา), 4
significant digits (เลขนัยสำคัญ), 29
similarity transformation (การแปลงภาวะคล้าย), 308
Simpson's rule (กฎของซิมป์สัน), 146, 436
single precision (เลขความแม่นเดี่ยว), 2, 132
singular value (ค่าเอกฐาน), 275
singular value decomposition (การแยกค่าเอกฐาน), 275
skew-symmetric matrix (เมทริกซ์สมมาตรเสมือน), 396
sliding window (หน้าต่างเลื่อน), 102
smoother (ตัวปรับเรียบ), 352
Sobolev space (ปริภูมิโซโบเลฟ), 503
sparse matrix (เมทริกซ์มากเลขศูนย์), 215, 329
spectral leakage (การรั่วของสเปกตรัม), 123
spectral norm (นอร์มสเปกตรัม), 276
spectral radius (รัศมีสเปกตรัม), 253, 341
spectrum (สเปกตรัม), 286
stability function (ฟังก์ชันเสถียรภาพ), 456
stage slope (ความชันของชั้น), 435
step size (ขนาดขั้น), 134
stiff equation (สมการตัวตึง), 457
stiffness matrix (เมทริกซ์สติฟเนส), 458, 504
stiffness ratio (อัตราส่วนความตึง), 458
strictly diagonally dominant matrix (เมทริกซ์เด่นทแยงอย่างเข้ม), 343
sub-step (ก้าวย่อย), 453
subnormal number (เลขซับนอร์มอล), 19
substep (ก้าวย่อย), 449
symplectic Euler method (วิธีออยเลอร์ซิมเพลกติก), 413

Taylor start (ก้าวแรกแบบเทย์เลอร์), 484
test function (ฟังก์ชันทดสอบ), 502
theta method (วิธีทีตา), 411
three-term recurrence (ความสัมพันธ์เวียนเกิดสามพจน์), 110
ties to even (ปัดเข้าหาเลขคู่), 15
tolerance (ความคลาดเคลื่อนยินยอม), 170, 443
trapezoidal rule (กฎเชิงสี่เหลี่ยมคางหมู), 507
trapezoidal rule (กฎเชิงสี่เหลี่ยมคางหมู), 144, 411
tree (ต้นไม้), 243
tridiagonal matrix (เมทริกซ์สามแนวเฉียง), 162, 217, 286, 333
trigonometric interpolation (การประมาณค่าเชิงตรีโกณมิติ), 122
trigonometric polynomial (พหุนามตรีโกณมิติ), 122
truncation error (ค่าคลาดเคลื่อนตัดปลาย), 135
tunnelling (การทะลุผ่าน), 416
twiddle factor (ตัวคูณหมุน), 125
two-grid method (วิธีสองกริด), 354

underflow (อันเดอร์โฟลว์), 19

ฉบับร่าง · Preprint · 2026-10-23

unit roundoff (หน่วยปัด), 15
unit vector (เวกเตอร์หนึ่งหน่วย), 38
upwinding (การทำอัปวินด์), 487

V-cycle (รอบแบบ V), 354
Vandermonde matrix (เมทริกซ์แวนเดอร์มอนด์), 63
vector norm (นอร์มของเวกเตอร์), 221
Verlet integration (วิธีเวอร์เลต), 422
vertex (จุดยอด), 240

wave equation (สมการคลื่น), 484
wavenumber (เลขคลื่น), 352
waypoint (จุดนำทาง), 85
weak solution (ผลเฉลยแบบอ่อน), 502
weight function (ฟังก์ชันถ่วงน้ำหนัก), 110
weighted Jacobi (จาโคบีแบบถ่วง), 352
well posed (ถูกรูปแบบ), 498
Wilkinson shift (การเลื่อนแบบวิลกินสัน), 310
window weights (น้ำหนักของหน้าต่าง), 103

z-fighting (ซี-ไฟต์ติง), 49
zenith angle (มุมเซนิท), 48

ฉบับร่าง · Preprint · 2026-10-23