Phương pháp tính MAT1099

Lê Phê Đô
dolp@vnu.edu.vn

07-09-2020

Mục lục

Danh sách hình vẽ

Danh sách bảng

Chương 1

Giải tích sai số

1.1 Giới thiệu về phép tính gần đúng

1.1.1 Một số ví dụ về tính toán khoa học và phương pháp tính

Lý do nghiên cứu Phương pháp tính

- Làm viêc với các số gần đúng
- Giải gần đúng các phương trình và hệ phương trình
- Xấp xỉ hàm số: Phương pháp nội suy, phương pháp xấp xỉ hàm số, chuỗi Taylor hoặc chuỗi Maclaurin
- Số học IEEE

Các nhiệm vụ

- Tìm hiểu và ứng dụng các thuật toán.
- Thể hiện các thuật toán bằng các chương trình.
- Tìm các bài toán thực tiễn.

Trong thực tế chúng ta thường phải xử lý, tính toán với các đại lượng gần đúng như các số đo vật lý, các dữ liệu ban đầu, các số làm tròn...với sai số nào đó, tức là các số gần đúng. Việc ước lượng sai số hợp lý cho phép ta đánh giá được chất lượng của quá trình tính toán, quyết định số chữ số giữ lại trong các phép tính trung gian và trong kết quả. Vì vậy, trước tiên ta cần nghiên cứu về các phép tính gần đúng và sai số.

1.2 Số gần đúng, sai số tuyệt đối và tương đối

1.2.1 Sai số tuyệt đối

Nếu số gần đúng a có giá trị đúng là a_0 thì ta nói a xấp xỉ a_0 hay a là số gần đúng của a_0. Khi đó sai số của a là:

$$E_a = a - a_0 \tag{1.1}$$

Nhưng giá trị này nói chúng ta không biết được mà chỉ ước lượng được cận trên của trị tuyệt đối của nó.

Định nghĩa 1.1: Sai số tuyệt đối (absolute error)

Giá trị ước lượng Δa sao cho

$$|a - a_0| \leq \Delta a \tag{1.2}$$

được gọi là *sai số tuyệt đối* của số gần đúng a.

Sai số tuyệt đối nhỏ nhất có thể biết được gọi là sai số tuyệt đối giới hạn của a. Thông thường ước lượng sai số tuyệt đối giới hạn là khó và nhiều khi không cần thiết nên người ta chỉ cần ước lượng sai số tuyệt đối đủ nhỏ và dùng từ 1 đến 3 chữ số có nghĩa (là số chữ số bắt đầu từ chữ số khác không đầu tiên từ trái sang phải - xem mục 2.1) để biểu diễn sai số tuyệt đối của số gần đúng.

Thay cho 1.2 người ta còn dùng cách biểu diễn sau để chỉ sai số tuyệt đối của a:

$$a_0 = a \pm \Delta a \tag{1.3}$$

Trong thực tế thì sai số E_a không thể biết được nên khi không có sự hiểu lầm người ta còn dùng từ *sai số* để chỉ sai số tuyệt đối E_a.

Ví dụ 1.1. *Căn phòng có chiều dài* $d = 5{,}45\,\text{m}$ *và chiều rộng* $r = 3{,}94\,\text{m}$ *với sai số* $1\,\text{cm}$.

Khi đó ta hiểu là:

$$\Delta d = 0{,}01\,\text{m} \text{ hay } d = (5{,}45 \pm 0{,}01)\,\text{m}$$
$$\Delta r = 0{,}01\,\text{m} \text{ hay } r = (3{,}97 \pm 0{,}01)\,\text{m}$$

Như vậy diện tích của phòng được ước lượng bởi:

$$S = d \cdot r = 5{,}45 \cdot 3{,}94 = 21{,}473\,\text{m}^2$$

với cận trên và cận dưới của S *là:*

$$(5{,}45 - 0{,}01)(3{,}94 - 0{,}01) = 21{,}3792 \leq S \leq (5{,}45 + 0{,}01)(3{,}94 + 0{,}01) = 21{,}567$$

Vậy ta có ước lượng sai số tuyệt đối của S là:

$$|S - S_0| \leq 0{,}094\,\mathrm{m}^2$$

1.2.2 Sai số tương đối

Hai số gần đúng có cùng sai số tuyệt đối sẽ có "mức độ chính xác" khác nhau nếu độ lớn của chúng khác nhau: số bé hơn sẽ có độ chính xác kém hơn. Để biểu diễn độ chính xác này, người ta dùng sai số tương đối.

Định nghĩa 1.2: Sai số tương đối (relative error)

Sai số tương đối của số gần đúng a là tỷ số giữa sai số tuyệt đối và giá trị tuyệt đối của nó, được ký hiệu là δa.

$$\delta a = \frac{\delta a}{|a|} \tag{1.4}$$

Thường sai số tương đối được biểu diễn dưới dạng phần trăm với 2 hoặc 3 chữ số.

Từ 1.4 ta thấy nếu biết δa thì:

$$\Delta a = |a|\,\delta a \tag{1.5}$$

nên ta chỉ cần biết một trong hai loại sai số của nó là được.

Ví dụ 1.2. *Nếu $a = 57$ và $\Delta a = 0{,}5$ thì $\delta a = 0{,}008\,771\,9$ hoặc $0{,}88\,\%$ (gọn hơn là $0{,}9\,\%$).*

1.2.3 Các loại sai số khác

Để hình dung các loại sai số khác ta xét ví dụ sau:

Ví dụ 1.3. *Một vật thể rơi từ độ cao H_0 với vận tốc ban đầu v_0 (được đo nhờ thiết bị nào đó). Tính độ cao $H(t)$ của vật thể sau thời gian t. Bài toán có thể giải như sau:*

Nếu gọi ngoại lực tác động vào vật thể là $F(t)$ (gồm lực hút trọng trường và lực cản), khối lượng vật thể là m thì $H(t)$ là nghiệm của phương trình vi phân cấp hai

$$H''(x) = \frac{-F(t)}{m} \tag{1.6}$$

với điều kiện ban đầu $H(0) = H_0$ và $H'(0) = -v_0$.

Ta chọn một phương pháp gần đúng để giải phương trình này, chẳng hạn nếu giả thiết $\dfrac{F(t)}{m}$ không đổi thì

$$H(t) = H_0 - g\frac{t^2}{2} - v_0 t$$

Qua ví dụ trên ta thấy sai số của kết quả nhận được chịu ảnh hưởng của:

- các số đo H_0, v_0
- cách lập luận để xác định $F(t)$
- phương pháp giải phương trình 1.6
- và các yếu tố ngẫu nhiên khác

Theo các yếu tố ảnh hưởng tới kết quả tính toán ta phân ra các loại sai số sau:

- *Sai số dữ liệu*: Còn gọi là sai số của số liệu ban đầu. Trong thí dụ trên là sai số khi đo H_0 và v_0.
- *Sai số giả thiết*: Sai số này gặp phải khi ta đơn giản hoá bài toán thực tiễn để thiết lập mô hình toán học có thể giải được. Trong thí dụ trên có thể giả thiết ngoại lực chỉ là trọng lực.
- *Sai số phương pháp*: Là sai số của phương pháp giải gần đúng bài toán theo mô hình được lập. Trong thí dụ trên là phương pháp giải phương trình vi phân 1.6.
- *Sai số tính toán*: Là sai số tích luỹ trong quá trình tính toán theo phương pháp được chọn.
- *Sai số làm tròn*: Khi tính toán ta thường phải làm tròn các số nên ảnh hưởng tới kết quả nhiều khi rất đáng kể.
- *Sai số ngẫu nhiên*: Là sai số chịu các quy luật chi phối ngẫu nhiên không tránh được.

Về sau ta quan tâm tới sai số tính toán và sai số phương pháp.

1.3 Sai số tích lũy và các bài toán sai số

1.3.1 Sai số hàm một biến

Cho hàm số $y = f(x)$ và x là số gần đúng của x_0. Ký hiệu Δx và Δy là sai số tuyệt đối tương ứng của đối số và hàm số. Ta sẽ xét các bài toán ước lượng sai số của hàm hoặc của đối số khi biết một trong hai sai số.

Bài toán thuận

Bài toán này ta ước lượng Δy khi biết x và Δx.
Theo công thức số gia hữu hạn ta có:

$$|y - y_0| = \big|f'(c)\big||x - x_0|$$

ở đây y_0 là giá trị đúng của y và c là điểm thuộc miền (x, x_0) nếu $x < x_0$ và thuộc (x_0, x) nếu $x_0 < x$.

Khi Δx bé, x gần x_0 ta có ước lượng:

$$\begin{aligned} \Delta y &\approx \left|f'(x)\right||x - x_0| \\ \text{hay } \Delta y &\leq \left|f'(x)\right| \Delta x \end{aligned} \tag{1.7}$$

Ví dụ 1.4. *Cho $y = \ln x$ ta có ước lượng:*

$$\Delta(\ln x) = \frac{1}{x}\Delta x = \delta x$$

Bài toán ngược

Trong bài toán này, ta biết giá trị gần đúng x, ta cần xác định phải tính x với Δx là bao nhiêu để đảm bảo $\Delta y \leq \Delta$. Với giá trị Δ cho trước, từ công thức 1.7 ta thấy nếu

$$\Delta x \leq \frac{\Delta}{\left|f'(x)\right|} \tag{1.8}$$

thì đủ để $\Delta y \leq \Delta$.

Ví dụ 1.5. *$y = e^x$ với $x \approx 3$ để có $\Delta y \leq 0{,}01$ ta tính x với $\Delta x \leq \dfrac{0{,}01}{e^3}$ là đủ.*

1.3.2 Sai số qua các phép toán số học

Khi tính toán với các số gần đúng thì sai số sẽ tích luỹ qua các phép toán cơ bản. Sau đây ta ước lượng sai số khi cộng trừ, nhân chia các số gần đúng.

Sai số của tổng hoặc hiệu

Mệnh đề. *Sai số tuyệt đối của một tổng hoặc hiệu bằng tổng các sai số tuyệt đối thành phần.*

Chứng minh. Để đơn giản ta xét $u = a \pm b$ với các số a, b có giá trị đúng a_0, b_0 và sai số tuyệt đối Δa, Δb tương ứng. Trường hợp có nhiều số hạng được xét tương tự.

Khi đó ta có:

$$\begin{cases} a_0 - \Delta a \leq a \leq a_0 + \Delta a \\ b_0 - \Delta b \leq b \leq b_0 + \Delta b \end{cases}$$

Do đó ta có:

$$\begin{cases} a_0 + b_0 - (\Delta a + \Delta b) \leq a + b \leq a_0 + b_0 + (\Delta a + \Delta b) \\ a_0 - b_0 - (\Delta a + \Delta b) \leq a - b \leq a_0 - b_0 + (\Delta a + \Delta b) \end{cases}$$

Nên

$$a_0 \pm b_0 - (\Delta a + \Delta b) \le a \pm b \le a_0 \pm b_0 + (\Delta a + \Delta b)$$

đpcm.

Ví dụ 1.6. *Cho* $a = 50{,}5$, $b = 50{,}9$ *với* $\Delta a = \Delta b = 0{,}05$ *và* $u = a - b$.
Ta có $u = 0{,}4$ *với* $\Delta u = 0{,}05 + 0{,}05 = 0{,}1$.
Vậy $\delta u = \dfrac{0{,}1}{0{,}4} = 25\,\%$.

Sai số của tích hoặc thương

Mệnh đề. *Sai số tương đối của tích hoặc thương bằng tổng các sai số tương đối thành phần.*

Chứng minh. Xét

$$u = \frac{x_1 \dots x_m}{y_1 \dots y_n}$$

Ta có thể giả thiết các x_i và y_j đều dương. Khi đó ta có:

$$\ln u = \sum_{i=1}^{m} \ln x_i + \sum_{j=1}^{n} \ln y_j$$

Theo mệnh đề 1.3.2 ở trên, ta suy ra:

$$\delta u = \sum_{i=1}^{m} \delta x_i + \sum_{j=1}^{n} \delta y_j$$

đpcm.

Ví dụ 1.7. *Xét* $S = d \cdot r$ *như ở ví dụ 1.1* $d = 5{,}45$, $r = 3{,}94$, $\Delta d = \Delta r = 0{,}01$.
Ta có:

$$\delta d = 0{,}001\,835$$
$$\delta r = 0{,}002\,538$$
$$\delta S = 0{,}004\,373 \text{ nên } \Delta S = 0{,}094$$

1.3.3 Sai số hàm nhiều biến

Ta xét hàm nhiều biến $u = f(x_1, \dots, x_n)$ với giá trị gần đúng $x_1, \dots, x_n$ và y đã biết.

Bài toán thuận

Trong bài toán này, ta cần ước lượng sai số Δy khi biết $\delta x_i \, \forall i \leq n$.

Tương tự hàm một biến, sử dụng công thức số gia hữu hạn ta có ước lượng:

$$\Delta u = \sum_{i=1}^{n} \left| f_i'(x_1, \ldots, x_n) \right| \Delta x_i \tag{1.9}$$

với f_i' là đạo hàm riêng của u theo biến x_i.

Ví dụ 1.8. *Xét $u = a^2 b$ với $a = 2{,}0$, $b = 25{,}0$, $\Delta a = \Delta b = 0{,}1$.*

Ta có:

$$\begin{aligned} u &= 100 \\ \Delta u &= 2ab\Delta a + a^2 \Delta b \\ &= 2 \cdot 2{,}0 \cdot 25{,}0 \cdot 0{,}1 + 2{,}0^2 \cdot 0{,}1 \\ &= 10{,}4 \end{aligned}$$

Bài toán ngược

Bây giờ ta đã biết các số gần đúng x_i, ta phải tính chúng với sai số tuyệt đối như thế nào để có $\Delta y \leq \Delta$; ở đây Δ là số cho trước.

Các phương pháp xử lý bài toán này đều dựa trên công thức 1.9 một cách linh hoạt. Sau đây ta xét hai phương pháp thông dụng.

1. Sai số của đối số như nhau. Ta xét khi:

   $$\Delta x_k = \Delta x \, \forall k \leq n$$

   Từ 1.8 ta có:

   $$\Delta u = \sum_{i=1}^{n} \left| f_i'(x_1, \ldots, x_n) \right| \Delta x$$

   Vậy để cho $\Delta u \leq \Delta$ thì chỉ cần:

   $$\Delta x \leq \frac{\Delta}{\sum_{i=1}^{n} \left| f_i'(x_1, \ldots, x_n) \right|} \tag{1.10}$$

   là đủ.

2. Phân bố đều sai số. Bây giờ ta xét khi:

   $$\left| f_i'(x_1, \ldots, x_n) \right| \Delta x_i = \left| f_k'(x_1, \ldots, x_n) \right| \Delta x_k \, \forall i, k \leq n$$

Khi đó $\forall j \leq n$, từ 1.9 ta có:

$$\Delta u = n\left|f_i'(x_1, \ldots, x_n)\right| \Delta x_j$$

Vậy để cho $\Delta u \leq \Delta$ thì chỉ cần tính:

$$\Delta x_j \leq \frac{\Delta}{n\left|f_i'(x_1, \ldots, x_n)\right|} \forall j = 1, \ldots, n \tag{1.11}$$

Ví dụ 1.9. *Mảnh vườn có cạnh* $d \approx 45{,}0\,\mathrm{m}$ *và* $r \approx 20{,}0\,\mathrm{m}$. *Cần tính* d *và* r *với* Δd, Δr *như thế nào để* $\Delta S \leq 0{,}1\,\mathrm{m}^2$.

Cách 1. *Xét* $\Delta d = \Delta r = \Delta x$, *ta áp dụng 1.10:*

$$\Delta x \leq \frac{0{,}1}{45 + 20} = 0{,}0015\,\mathrm{m}$$

Cách 2. *Khi đo chiều dài thường có sai số lớn hơn chiều rộng nên ta có thể dùng 1.11.*

$$\Delta d \leq \frac{0{,}1}{2{,}20} = 0{,}0025\,\mathrm{m}\,,\ \Delta r \leq \frac{0{,}1}{2{,}45} = 0{,}0010\,\mathrm{m}$$

là đủ để $\Delta S \leq 0{,}1\,\mathrm{m}^2$

1.4 Sai số qui tròn, quan hệ giữa sai số và số chữ số đáng tin

Trong mục này ta xét các số được biểu diễn dưới dạng thập phân. Khi các số là gần đúng, vấn đề đặt ra là nên biểu diễn chúng với bao nhiêu chữ số? Thu gọn chúng như thế nào?

1.4.1 Chữ số có nghĩa

Định nghĩa 1.3: Chữ số có nghĩa

Trong biểu diễn theo cơ số b (trường hợp riêng là biểu diễn thập phân):

- các chữ số kể từ chữ số khác 0 đầu tiên tính từ trái sang phải gọi là *các chữ số có nghĩa*,
- các chữ số 0 bên trái là *không có nghĩa*

Nếu a được viết dưới dạng

$$a = \sum_{k=p}^{n} a_k 10^k \tag{1.12}$$

thì các chữ số 0 bên trái không có ở biểu diễn này, ý nghĩa của các chữ số 0 bên phải liên quan tới cách biểu diễn số gần đúng sẽ xét dưới đây.

Ví dụ 1.10. *Số* $a = 03.4050$ *thì chữ số* 0 *đầu không có nghĩa (người ta có thể điền để tránh viết thêm) còn các chữ số* 3, 4, 0, 5, 0 *là có nghĩa.*

Số $b = 0.034$ *thì các chữ số* 3, 4 *là có nghĩa, hai chữ số* 0 *bên trái không có nghĩa vì nếu biểu diễn theo dạng 1.12 thì các chữ số này không cần đến.*

1.4.2 Chữ số chắc

Định nghĩa 1.4: Chữ số chắc

Xét a có biểu diễn 1.12 với sai số Δa.

- Nếu $\Delta a \leq 0{,}5 \cdot 10^m$ thì a_k là *chữ số chắc (đáng tin)* $\forall k \geq m$ (theo nghĩa hẹp dùng trong tính toán).
- Nếu $0{,}5 \cdot 10^m \leq \Delta a \leq 10^m$ thì a_m là *chắc theo nghĩa rộng.*

Ví dụ 1.11. $a = 21{,}473$ *và* $\Delta a = 0{,}094 = 0{,}94 \cdot 10^{-1}$ *thì:*

- *Các chữ số* 2, 1 *là chắc theo nghĩa hẹp.*
- *Chữ số* 4 *là chắc theo nghĩa rộng.*
- *Các chữ số* 7, 3 *là không đáng tin hay không chắc.*

Khi cho số gần đúng ta có thể cho theo hai cách:

- *Cách 1:* Viết kèm với sai số tuyệt đối.
- *Cách 2:* Chỉ viết các chữ số chắc. Nếu ta có số gần đúng mà không cho sai số thì luôn ngầm hiểu các chữ số có nghĩa là các chữ số chắc. Như vậy các chữ số 0 ở bên phải cho ta biết nó là chữ số chắc.

Trong quá trình tính toán, người ta thường để lại vài chữ số không chắc và trong kết quả thì giữ lại các chữ số chắc theo nghĩa rộng.

1.4.3 Số thu gọn

Khi số a có nhiều chữ số không chắc hoặc có quá nhiều chữ số có nghĩa thì người ta thường thu gọn thành số $\overline{a}$ có ít chữ số có nghĩa hơn. Nếu a có biểu diễn 1.12 và số thu gọn được giữ lại đến a_m $(m > p)$ thì $\overline{a}$ có biểu diễn

$$\overline{a} = \sum_{k=m}^{n} b_k 10^k \tag{1.13}$$

nhờ bỏ đi các chữ số a_k $(k < m)$ theo quy tắc sau:

Quy tắc. *Quy tắc chữ số chẵn: Giả sử $a > 0$ và phần bỏ đi là μ.*

- *Nếu $\mu < 0{,}5 \cdot 10^m$ thì*

$$\overline{a} = \sum_{k=m}^{n} a_k 10^k \tag{1.14}$$

  nghĩa là ta giữ nguyên các chữ số đến hàng m tính từ trái sang phải.

- *Nếu $\mu > 0{,}5 \cdot 10^m$ thì*

$$\overline{a} = \sum_{k=m}^{n} a_k 10^k + 10^m \tag{1.15}$$

- *Nếu $\mu = 0{,}5 \cdot 10^m$, ta xét tiếp:*
  - *Nếu a_m chẵn, làm theo 1.14.*
  - *Nếu a_m lẻ, làm theo 1.15.*

Khi $a < 0$ ta thu gọn giá trị tuyệt đối và giữ nguyên dấu.

Khi thu gọn a thành $\overline{a}$ ta có sai số thu gọn $\Gamma_a \leq 0{,}5 \cdot 10^m$. Để nó ít ảnh hưởng tới sai số tuyệt đối, ta thu gọn số và giữ lại một hoặc hai chữ số không chắc.

Nếu a có biểu diễn 1.12 và a_k chắc với $k \geq m$ thì $\Delta a \leq 10^m$ nên

$$\delta_a = \frac{\Delta a}{|a|} \leq \frac{10^m}{\sum\limits_{k=m}^{n} a_k 10^k} = \frac{1}{\sum\limits_{k=0}^{n} a_{k+m} 10^k}$$

Như vậy sai số tương đối của số gần đúng có thể ước lượng bởi nghịch đảo của số gồm các chữ số chắc của a không có dấu phẩy.

1.4.4 Dấu phẩy động

Chúng ta biết rằng trong biểu diễn thập phân, mọi số thực được biểu diễn bởi một hữu hạn hoặc một dãy vô hạn các chữ số thập phân.

Bây giờ hầu hết các máy tính có hai cách biểu diễn số, được gọi là dấu phẩy tĩnh và dấu phẩy động.

Trong một biểu diễn *dấu phẩy tĩnh* tất cả các số được được đưa ra với một số cố định các số thập phân sau dấu thập phân; ví dụ, số được đưa ra với 3 số thập phân là 62,358, 0,014, 1,000. Trong một văn bản chúng ta sẽ viết, nói, biểu diễn 3 số thập phân là biểu diễn *3D*.

Biểu diễn dấu phẩy tĩnh có ưu thế:

- Thuận tiện trong tính toán hàng ngày,

- và trong tính toán với các số gần đơn vị

Trong hệ thống *dấu phẩy động*, chúng sẽ ta viết, ví dụ,

$$0{,}6247 \cdot 10^3; 0{,}1735 \cdot 10^{-13}; -0{,}2000 \cdot 10^{-1}$$

hay đôi khi ta biểu diễn:

$$6{,}247 \cdot 10^2; 1{,}735 \cdot 10^{-14}; -2{,}000 \cdot 10^{-2}$$

Chúng ta thấy rằng trong biểu diễn này số các chữ số có nghĩa được giữ cố định, trong khi dấu phẩy là "động". Ở đây, một chữ số có nghĩa của một số c là chữ số bất kỳ của c, ngoại trừ chữ số 0 nằm ở bên trái chữ số khác 0 đầu tiên; các chữ số 0 này chỉ để xác định vị trí của dấu phẩy (như vậy, bất kỳ chữ 0 khác đều là chữ số có nghĩa của c).

Ví dụ 1.12.

$$13\,600; 1{,}3600; 0{,}001\,360\,0$$

tất cả đều có 5 chữ số có nghĩa.

Trong văn bản chúng ta nói rút gọn một số đến 5 chữ số có nghĩa là *5S*.

Việc sử dụng số mũ cho phép chúng ta biểu diễn số rất lớn và rất nhỏ. Thật vậy, về mặt lý thuyết số khác 0 bất kỳ a có thể được viết như sau:

$$a = \pm m 10^n \mid 0.1 \leq |m| < 1, n \in \mathbb{Z}$$

Quy tắc. *Để làm tròn số x đến k chữ số sau dấu phẩy, ta cộng vào x lượng $0{,}5 \cdot 10^{-(k+1)}$ và bỏ đi từ chữ số thứ $k+1$ sau dấu phẩy trở đi.*

Số học IEEE

Trong máy tính hiện đại người ta dùng các số nhị phân, ở đây m được giới hạn bởi k chữ số nhị phân (ví dụ, $k = 8$) và n cũng được giới hạn, ta có biểu diễn sau (chỉ biểu diễn được một số hữu hạn số):

$$\overline{a} = \overline{m} \cdot 2^n \mid \overline{m} = 0, d_1 d_2 \dots d_k, d_1 > 0 \tag{1.16}$$

Các số $\overline{a}$ ở đây được gọi là số máy nhị phân k chữ số. Phần sau dấu phẩy m (được gọi là *mantissa*), biểu diễn các chữ số có nghĩa của $\overline{a}$, n được gọi là lũy thừa của $\overline{a}$.

Chương 2

Giải gần đúng phương trình

2.1 Mở đầu

Sự tăng trưởng của dân số thường có thể được mô hình hóa trong khoảng thời gian ngắn bằng cách giả định rằng dân số tăng liên tục theo thời gian tỷ lệ thuận với con số hiện tại vào thời điểm đó. Giả sử $N(t)$ biểu thị số dân tại thời điểm t và λ biểu thị tỷ lệ sinh không đổi của cộng đồng. Khi đó dân số thỏa mãn phương trình vi phân:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \lambda N(t)$$

Nghiệm của phương trình là $N(t) = N_0 e^{\lambda t}$, ở đây N_0 là dân số ban đầu.

Mô hình hàm mũ này chỉ có giá trị khi dân số bị cô lập, không có người nhập cư. Nếu nhập cư được phép ở tốc độ không đổi v thì phương trình vi phân trở thành:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \lambda N(t) + v$$

Nghiệm của nó là:

$$N(t) = N_0 e^{\lambda t} + \frac{v}{\lambda}(e^{\lambda t} - 1)$$

Giả sử ban đầu có $N(0) = 1\,000\,000$ người, và có tới $435\,000$ người nhập cư vào cộng đồng trong năm đầu tiên, vậy $N(1) = 1\,564\,000$ người có mặt vào cuối năm đầu tiên. Để xác định tỷ lệ sinh của cộng đồng dân số này, chúng ta cần tìm λ trong phương trình:

$$1\,564\,000 = 1\,000\,000 e^{\lambda} + \frac{435\,000}{\lambda}(e^{\lambda} - 1)$$

Không thể giải một cách chính xác giá trị λ trong phương trình này, nhưng các phương pháp tính được thảo luận trong chương này có thể được sử dụng để tính gần đúng nghiệm của các phương trình loại này với độ chính xác cao tùy ý.

2.2 Phương pháp chia đôi

Giả sử f là hàm số xác định và liên tục trên khoảng $[a, b]$, với $f(a)$ và $f(b)$ trái dấu. *Định lý giá trị trung gian* nói rằng tồn tại một số $p \in (a, b)$ với $f(p) = 0$.

Định lí 2.1: Định lý giá trị trung gian (Intermediate Value Theorem)

Nếu f liên tục trên $[a, b]$ và K nằm giữa $f(a)$ và $f(b)$, tồn tại $c \in (a, b)$ sao cho $f(c) = K$.

Cụ thể hơn, do $f(a)$ và $f(b)$ trái dấu, do đó 0 nằm giữa $f(a)$ và $f(b)$, do đó tồn tại nghiệm $p \in (a, b)$.

Kết quả trên là một trường hợp đặc biệt ($f(a)$, $f(b)$ trái dấu, $K = 0$) của định lý giá trị trung gian, còn được gọi là *định lý Bolzano*.

Mặc dù có thể tồn tại nhiều hơn một nghiệm trong khoảng (a, b), nhưng để thuận lợi, chúng ta giả thiết chỉ có duy nhất một nghiệm trong khoảng này. Khi đó, ta có thể dùng phương pháp sau:

Phương pháp 2.1: Phương pháp chia đôi (Bisection method)

Phương pháp này cho phép tìm nghiệm p của $f(p) = 0$ trong khoảng $[a, b]$, với $f(a)$ và $f(b)$ trái dấu.
Để bắt đầu, ta đặt $a_1 = a$ và $b_1 = b$, và đặt p_1 là điểm giữa của $[a, b]$; nghĩa là:

$$p_1 = a_1 + \frac{b_1 - a_1}{2} = \frac{a_1 + b_1}{2}$$

- Nếu $f(p_1) = 0$ thì $p = p_1$.
- Nếu $f(p_1) \neq 0$ thì $f(p_1)$ cùng dấu với $f(a_1)$ hoặc $f(b_1)$.
  - Nếu $f(p_1)$ cùng dấu với $f(a_1)$ thì $p \in [p_1, b_1]$. Đặt $a_2 = p_1$, $b_2 = b_1$.
  - Nếu $f(p_1)$ cùng dấu với $f(b_1)$ thì $p \in [a_1, p_1]$. Đặt $a_2 = a_1$, $b_2 = p_1$.

  sau đó làm tiếp phương pháp trên với khoảng $[a_2, b_2]$.

Các cách dừng khác (còn gọi là *tiêu chí dừng*) có thể được áp dụng trong phương pháp trên hoặc trong bất kỳ các kỹ thuật lặp lại trong chương này. Ví dụ, chúng ta có thể chọn một dung sai $\varepsilon > 0$ và tạo dãy $p_1, ..., p_N$ cho đến khi đáp ứng một trong các điều kiện sau:

$$|p_N - p_{N-1}| < \varepsilon, \tag{2.1}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \varepsilon,\ p_N \neq 0 \text{ hoặc} \tag{2.2}$$

$$\big|f(p_N)\big| < \varepsilon \tag{2.3}$$

Không may, khó khăn có thể phát sinh với bất kỳ tiêu chí dừng nào. Ví dụ, có các chuỗi $\{p_n\}_{n=1}^{\infty}$ mà hiệu $p_n - p_{n-1}$ hội tụ về 0 trong khi dãy đó lại phân kỳ. Cũng có thể có $f(p_n)$ gần bằng 0 trong khi p_n khác đáng kể so với p. Nếu không có kiến thức bổ sung về f hoặc p, bất đẳng thức 2.2 là tiêu chuẩn dừng tốt nhất để áp dụng vì nó sát nhất với sai số tương đối.

Khi dùng máy tính để tính xấp xỉ, nên thiết lập một giới hạn trên về số lần lặp lại. Điều này giúp tránh vòng lặp vô hạn, một tình huống có thể phát sinh khi chuỗi $\{p_N\}_{n=0}^{\infty}$ phân kỳ (và cả khi chương trình sai).

Ví dụ 2.1. *Chứng minh rằng $f(x) = x^3 + 4x^2 - 10 = 0$ có nghiệm trong khoảng $[1, 2]$, và dùng phương pháp chia đôi để xác định nghiệm đúng đến 10^{-4}.*

Vì $f(1) = -5$ và $f(2) = 14$, $f(x) = 0$ chắc chắn có nghiệm trong khoảng $[1, 2]$.

Ta có bảng sau:

n	a_n	b_n	p_n	$f(p_n)$
1	1,0	2,0	1,5	2,375
2	1,0	1,5	1,25	−1,796 87
3	1,25	1,5	1,375	0,162 11
4	1,25	1,375	1,3125	−0,848 39
5	1,3125	1,375	1,343 75	−0,350 98
6	1,343 75	1,375	1,359 375	−0,096 41
7	1,359 375	1,375	1,367 187 5	0,032 36
8	1,359 375	1,367 187 5	1,363 281 25	−0,032 15
9	1,363 281 25	1,367 187 5	1,365 234 375	0,000 072
10	1,363 281 25	1,365 234 375	1,364 257 813	−0,016 05
11	1,364 257 813	1,365 234 375	1,364 746 094	−0,007 99
12	1,364 746 094	1,365 234 375	1,364 990 234	−0,003 96
13	1,364 990 234	1,365 234 375	1,365 112 305	−0,001 94

Sau 13 lần lặp, $p_{13} = 1{,}365\,112\,305$ xấp xỉ nghiệm p với sai số:

$$|p - p_{13}| < |b_{14} - a_{14}| = |1{,}365\,234\,375 - 1{,}365\,112\,305| = 0{,}000\,122\,070$$

Do $|a_{14}| < |p|$ (khoảng đang xét dương), ta có:

$$\frac{|p - p_{13}|}{|p|} < \frac{|b_{14} - a_{14}|}{|a_{14}|} \leq 9 \times 10^{-5}$$

Cần chú ý rằng, p_9 thực sự gần với p hơn kết quả cuối cùng p_{13}, tuy nhiên

khi thực hiện thuật toán ta không thể biết đều này. Hơn nữa, $|f(p_9)| < |f(p_{13})|$ *cũng không liên quan đến việc* p_9 *sát với* p *hơn.*

Phương pháp chia đôi có hai điểm yếu lớn:

- Cần số vòng lặp N lớn
- Vô tình bỏ qua các xấp xỉ tốt

Dù vậy, phương pháp này lại có một ưu điểm lớn là đảm bảo dãy $\{p_N\}_{n=0}^{\infty}$ hội tụ đến một nghiệm. Do ưu điểm này, phương pháp chia đôi thường được dùng để tìm điểm bắt đầu cho các phương pháp khác hiệu quả hơn mà sẽ được giới thiệu sau.

Định lí 2.2

Cho hàm $f \in [a, b]$ và $f(a)\dot{f}(b) < 0$. Phương pháp chia đôi tạo ra một chuỗi $\{p_n\}_{n=1}^{\infty}$ xấp xỉ nghiệm p của f với sai số như sau:

$$|p_n - p| \leq \frac{b-a}{2^n},\ n \geq 1$$

Chứng minh. Với mọi $n \geq 1$, ta có:

$$b_n - a_n = \frac{1}{2^{n-1}}(b-a) \text{ và } p \in (a_n, b_n)$$

Do

$$p_n = \frac{1}{2}(a_n + b_n)$$

ta suy ra được

$$\begin{aligned} & \frac{1}{2}(a_n + b_n) - b_n \leq p_n - p \leq \frac{1}{2}(a_n + b_n) - a_n \\ \iff & \frac{1}{2}(a_n - b_n) \leq p_n - p \leq \frac{1}{2}(b_n - a_n) \\ \iff & |p_n - p| \leq \frac{1}{2}(b_n - a_n) = \frac{b-a}{2^n} \end{aligned}$$

đpcm.

2.3 Phương pháp điểm bất động

2.3.1 Điểm bất động và bài toán tìm nghiệm

Điểm bất động (*fixed point*) của một hàm là số mà tại đó giá trị của hàm số bằng đúng giá trị của đối số.

Định nghĩa 2.1: S

p được gọi là điểm bất động của hàm số g nếu $g(p) = p$.

Trong phần này, chúng ta xét việc đưa bài toán tìm nghiệm về bài toán tìm điểm bất động và tìm sự liên hệ giữa chúng.

Các bài toán tìm nghiệm và các bài toán tìm điểm cố định là các lớp tương đương theo nghĩa sau đây:

- Từ bài toán tìm nghiệm của phương trình $f(p) = 0$, ta có thể xác định hàm g với điểm bất động tại p theo một số cách, ví dụ,

$$g(x) = x - 3f(x)$$

  vì khi thay p vào, $g(p) = p - 3f(p) = p$.

- Ngược lại, nếu hàm g có một điểm bất động tại p, thì hàm f xác định bởi

$$f(x) = x - g(x)$$

  có nghiệm tại p.

Mặc dù các bài toán ta muốn giải quyết là dạng tìm nghiệm, nhưng dạng điểm bất động dễ thực hiện hơn và có một số lựa chọn điểm bất động dẫn tới kỹ thuật tìm nghiệm rất hiệu quả. Trước hết ta cần đi đến dạng bài toán mới này một cách thoải mái, và đưa ra quyết định khi nào hàm số có điểm bất động và điểm bất động được xấp xỉ với độ chính xác bao nhiêu.

Các điểm bất động xuất hiện trong nhiều lĩnh vực toán học khác nhau, và là công cụ chính của các nhà kinh tế dùng để chứng minh các kết quả liên quan đến tính cân bằng. Mặc dù ý tưởng đằng sau kỹ thuật là cũ, nhưng thuật ngữ được sử dụng lần đầu bởi nhà toán học Hà Lan L. E. J. Brouwer (1882 - 1962) trong đầu những năm 1900.

2.3.2 Điều kiện tồn tại của điểm bất động

Ví dụ 2.2. *Hãy xác định điểm bất động của hàm* $g(x) = x^2 - 2$.

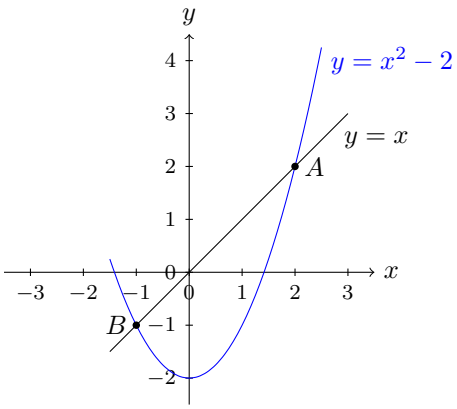
Điểm bất động p *của* g *có tính chất:*

$$p = g(p) \iff p = p^2 - 2$$

Suy ra

$$p^2 - p - 2 = (p + 1)(p - 2) = 0$$

Điểm bất động xảy ra đúng khi khi đồ thị của hàm số $y = g(x)$ *cắt đồ thị hàm số* $y = x$, *vì vậy* g *có 2 điểm bất động là* -1 *và* 2. *Điều này được minh họa bởi hình 2.1:*



Hình 2.1: Điểm bất động của $y = x^2 - 2$.

Định lý sau cho điều kiện đủ để hàm số *có ít nhất một* và *có duy nhất một* điểm bất động.

Định lí 2.3

1. Nếu $g \in C[a,b]$, và $g(x) \in [a,b] \, \forall x \in [a,b]$, khi đó g có ít nhất một điểm bất động trên $[a,b]$.
2. Hơn nữa, nếu $g'(x)$ tồn tại trên (a,b) và $\left|g'(x)\right| < 1 \, \forall x \in [a,b]$, khi đó, tồn tại duy nhất một điểm bất động trên $[a,b]$.

Trước khi chứng minh định lí trên, ta cần biết *định lí giá trị trung bình*.

Định lí 2.4: Định lí giá trị trung bình (Mean Value Theorem)

Nếu f liên tục trên $[a,b]$ và khả vi trên (a,b), tồn tại một điểm $c \in (a,b)$ sao cho tiếp tuyến tại c song song với cát tuyến qua hai điểm mút $(a, f(a))$ và $(b, f(b))$, hay nói cách khác:

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

Chứng minh Định lí 2.3.

1. Nếu $g(a) = a$ hoặc $g(b) = b$, g có điểm bất động a hoặc b. Nếu không, $g(a) > a$ và đồng thời $g(b) < b$; ta sẽ xét trường hợp này.

   Hàm $h(x) = g(x) - x$ liên tục trên $[a,b]$ với:

   $$h(a) - a > 0 \text{ và } h(b) - b > 0$$

Định lý giá trị trung gian khẳng định rằng tồn tại $p \in (a, b)$ sao cho $h(p) = 0$. Điểm p này là điểm bất động của g vì:

$$0 = h(p) = g(p) - p \iff g(p) = p$$

2. Giả sử g có hai điểm bất động p, q trên $[a, b]$. Không mất tính tổng quát, giả sử $p < q$. Theo định lí giá trị trung bình, tồn tại $\xi \in (p, q)$ sao cho:

$$g'(\xi) = \frac{g(p) - g(q)}{p - q}$$

Ta có:

$$|p - q| = \big|g(p) - g(q)\big| = \big|g'(\xi)\big||p - q| < |p - q| \text{ (vô lí)}$$

Giả thuyết g có hai điểm bất động trên $[a, b]$ sai. Vậy với điều kiện ban đầu, chỉ có duy nhất một điểm bất động trên $[a, b]$.

đpcm.

2.3.3 Phương pháp điểm bất động

Xét chuỗi sau:

$$\{p_n\}_{n=0}^{\infty} \mid p_n = g(p_{n-1})\, \forall n \geq 1$$

Giả sử chuỗi này hội tụ tới p, và g liên tục, thì:

$$p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} g(p_{n-1}) = g(\lim_{n\to\infty} p_{n-1}) = g(p)$$

Khi này p chính là điểm bất động của g. Đây chính là tiền đề cho *phương pháp điểm bất động*.

Cần chú ý rằng phương pháp này chỉ đúng khi chuỗi $\{p_n\}_{n=0}^{\infty}$ hội tụ về p.

Phương pháp 2.2: Phương pháp điểm bất động (fixed-point method)

Phương pháp này cho phép tìm điểm bất động p của g, khi biết một điểm bắt đầu p_0.
Đặt $p = g(p_0)$.

- Nếu $|p - p_0|$ đủ nhỏ, thì ta có p cần tìm.
- Nếu $|p - p_0|$ chưa đủ nhỏ, ta đặt $p_0 = p$ rồi làm tiếp phương pháp trên.

Cũng như với phương pháp chia đôi, có thể dùng nhiều điều kiện dừng khác nhau. Ví dụ trên sử dụng điều kiện $|p - p_0|$ nhỏ hơn một mốc ϵ nào đó thì dừng lại.

Ta cần nhắc lại rằng điểm quan trọng nhất của phương pháp trên là giả sử $\{p_n\}_{n=0}^{\infty}$ hội tụ về p, tức ta phải chọn hàm g một cách phù hợp, chứ không áp dụng được cho mọi hàm g. Ví dụ sau cho ta thấy sự quan trọng của hàm này.

Ví dụ 2.3. *Thử tìm và biện luận cho cách tìm nghiệm của phương trình* $x^3 + 4x^2 - 10 = 0$ *trong* $[1, 2]$ *bằng phương pháp điểm bất động.*

Ta có một số lựa chọn về hàm g, được chọn ngẫu nhiên:

a) $x = g_1(x) = x - x^3 - 4x^2 + 10$

b) $x = g_2(x) = \left(\frac{10}{x} - 4x\right)^{0.5}$

c) $x = \frac{1}{2}(10 - x^3)^{0.5}$

d) $x = g_4(x) = \left(\frac{10}{x+4}\right)^{0.5}$

e) $x = g_5(x) = x - \frac{x^3+4x^2-10}{3x^2+8x}$

Với $p_0 = 1{,}5$*, ta có bảng sau:*

n	*a)*	*b)*	*c)*	*d)*	*e)*
0	1,5	1,5	1,5	1,5	1,5
1	−0,875	0,8165	1,286 953 768	1,348 399 725	1,373 333 333
2	6,732	2,9969	1,402 540 804	1,367 376 372	1,365 262 015
3	−469,7	$\sqrt{-8{,}65}$	1,345 458 374	1,364 957 015	1,365 230 014
4	$1{,}03 \times 10^8$		1,375 170 253	1,365 264 748	1,365 230 013
5			1,360 094 193	1,365 225 594	
6			1,367 846 968	1,365 230 576	
7			1,363 887 004	1,365 229 942	
8			1,365 916 734	1,365 230 022	
9			1,364 878 217	1,365 230 012	
10			1,365 410 062	1,365 230 014	
15			1,365 223 680	1,365 230 013	
20			1,365 230 236		
25			1,365 230 006		
30			1,365 230 013		

Với nghiệm thực 1,365 230 013*, ta thấy lựa chọn c), d), e) có tiềm năng nhất. Phương pháp chia đôi cần 27 lần lặp để đạt được kết quả này, tuy nhiên d) chỉ cần 15 lần, còn e) thậm chí chỉ cần 4. Ngược lại, a) thì phân kì còn b) thậm chí không xác định (căn của số âm).*

2.3.4 Tìm g phù hợp

Tiếp tục nhắc lại điểm quan trọng nhất để làm được phương pháp điểm bất động là chọn được g phù hợp. Định lí sau và hệ quả của nó cho ta một số gợi ý về việc chọn những hàm phù hợp, hay quan trọng hơn, loại bỏ những hàm không phù hợp.

Định lí 2.5: Định lí điểm bất động

Cho hàm g liên tục trên $[a,b]$ sao cho $g(x) \in [a,b]\,\forall x \in [a,b]$. Giả sử thêm rằng g khả vi trên (a,b) và

$$\left|g'(x)\right| < 1\,\forall x \in (a,b)$$

Thì với mọi $p_0 \in [a,b]$, chuỗi

$$p_n = g(p_{n-1})\,\forall n \geq 1$$

hội tụ về p là điểm bất động duy nhất của g trên $[a,b]$.

Chứng minh. Dựa vào 2.3, có một điểm bất động duy nhất p trong khoảng $[a,b]$.

Do $g(x) \in [a,b]\,\forall x \in [a,b]$, ta chắc chắn dãy $\{p_n\}_{n=0}^{\infty}$ tồn tại.

Theo điều kiện $\left|g'(x)\right| < 1\,\forall x \in (a,b)$, tồn tại $0 < k < 1$ thỏa mãn:

$$\left|g'(x)\right| \leq k\,\forall x \in (a,b)$$

Kết hợp điều trên với định lí giá trị trung bình, ta có:

$$|p_n - p| = \left|g(p_{n-1}) - g(p)\right| = \left|g'(\xi_n)\right||p_{n-1} - p| \leq k|p_{n-1} - p|$$

với $\xi_n \in (a,b)$. Quy nạp kết quả này ta có:

$$\begin{aligned} &|p_n - p| \leq k^n |p_0 - p| \\ \iff &\lim_{n\to\infty} |p_n - p| \leq \lim_{n\to\infty} k^n |p_0 - p| = 0 \end{aligned}$$

Vậy ta thấy $\{p_n\}_{n=0}^{\infty}$ hội tụ về p. đpcm.

Ta tiếp tục xem xét một số hệ quả hữu dụng của định lí trên.

Hệ quả 2.1: Hệ quả của định lí điểm bất động

Nếu g thỏa mãn các điều kiện trong định lí điểm bất động, ta có cận trên của sai số tuyệt đối khi ước lượng p bằng p_n:

$$|p_n - p| \leq k^n \max\{p_0 - a, b - p_0\} \tag{2.4}$$

và

$$|p_n - p| \leq \frac{k^n}{1-k}|p_1 - p_0| \ \forall n \geq 1 \tag{2.5}$$

với k như đã định nghĩa trong chứng minh của định lí điểm bất động.

Chứng minh. Ta có:

$$|p_n - p| \leq k^n |p_0 - p|$$

Vì $p \in [a, b]$ nên ta suy ra được bất đẳng thức 2.4:

$$|p_n - p| \leq k^n |p_0 - p| \leq k^n \max\{p_0 - a, b - p_0\}$$

Xét khi $n \geq 1$, bằng quy nạp và định lí giá trị trung bình, ta có:

$$|p_{n+1} - p_n| = \big|g(p_n) - g(p_{n-1})\big| \leq k^n |p_1 - p_0|$$

Do đó với $m > n \geq 1$:

$$\begin{aligned}
|p_m - p_n| &= |p_m - p_{m-1} + p_{m-1} - \ldots - p_{n+1} + p_{n+1} - p_n| \\
&\leq |p_m - p_{m-1}| + \ldots + |p_{n+1} - p_n| \\
&\leq k^{m-1}|p_1 - p_0| + \ldots + k^n |p_1 - p_0| \\
&= |p_1 - p_0| \sum_{i=n}^{m-1} k^i \\
&= |p_1 - p_0| \frac{k^m - k^n}{k-1}
\end{aligned}$$

Lấy giới hạn hai vế với $m \to \infty$, ta có được bất đẳng thức 2.5:

$$\begin{aligned}
\lim_{m\to\infty} |p_m - p_n| &= |p_1 - p_0| \lim_{m\to\infty} \frac{k^m - k^n}{k-1} \\
\iff |p_n - p| &= \frac{k^n}{1-k}|p_1 - p_0|
\end{aligned}$$

đpcm.

Qua các kết quả trên, ta rút ra hai quy tắc chọn hàm g:

- $\left|g'(x)\right| < 1 \, \forall x \in (a, b)$
- đạo hàm của g càng nhỏ càng tốt

2.4 Phương pháp Newton & các phương pháp liên quan

Phương pháp Newton, hay *Newton-Raphson*, là một trong những phương pháp mạnh nhất và phổ biến để tìm nghiệm phương trình. Trong phần này, ta sẽ sử dụng khai triển Taylor để biện luận về phương pháp Newton, và hơn nữa là về cận cho sai số của phương pháp này.

2.4.1 Phương pháp Newton

Giả sử $f \in C^2[a, b]$ (f khả vi liên tục đến cấp 2). Xét $p_0 \in [a, b]$ là một xấp xỉ của nghiệm p sao cho $f'(p_0) \neq 0$ và $|p - p_0|$ "nhỏ".

Khai triển Taylor quanh p_0 và tại $x = p$, ta có:

$$f(p) = f(p_0) + (p - p_0)f'(p_0) + \frac{(p - p_0)^2}{2} f''(\xi(p))$$

Với $\xi(p)$ nằm giữa p và p_0.

Do giả thiết $|p - p_0|$ nhỏ, $(p - p_0)^2$ thậm chí còn nhỏ hơn. Phương trình trên trở thành:

$$\begin{aligned} 0 &\approx f(p_0) + (p - p_0)f'(p_0) \\ \Rightarrow p &\approx p_0 - \frac{f(p_0)}{f'(p_0)} \equiv p_1 \end{aligned}$$

Phương trình trên là tiền đề cho phương pháp Newton. Phương pháp Newton bắt đầu từ một p_0 cho trước, và tạo dãy $\{p_n\}_{n=0}^{\infty}$ theo quy tắc:

$$p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n)} \tag{2.6}$$

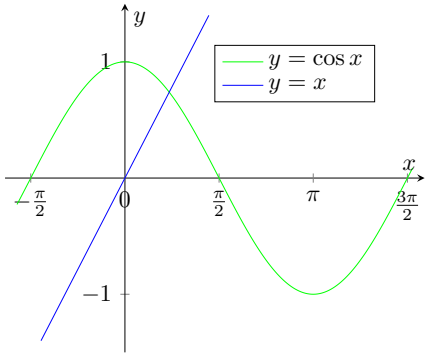
Phương pháp 2.3: Phương pháp Newton (Newton's Method)

Phương pháp này cho phép tìm nghiệm p của f, khi biết một điểm bắt đầu p_0.
Đặt $p = p_0 - \frac{f(p_0)}{f'(p_0)}$.

- Nếu $|p - p_0|$ đủ nhỏ, thì ta có p cần tìm.
- Nếu $|p - p_0|$ chưa đủ nhỏ, ta đặt $p_0 = p$ rồi làm tiếp phương pháp trên.

Tương tự phương pháp chia đôi, ta có thể tùy chọn nhiều điều kiện dừng khác nhau.

Ví dụ 2.4. *Xấp xỉ nghiệm của* $f(x) = \cos x - x = 0$ *bằng phương pháp Newton.*
Xét đồ thị của hàm $y = \cos x$ *và* $y = x$*:*



Hình 2.2: Đồ thị của $y = \cos x$ và $y = x$.

Dựa theo đồ thị, ta biết phương trình có nghiệm trong khoảng $[0, \frac{\pi}{2}]$*, nên ta xét điểm khởi đầu* $p_0 = \frac{\pi}{4}$*.*

Ta có $f'(x) = -sin(x) - 1$*. Áp dụng phương pháp Newton, ta có công thức cho chuỗi xấp xỉ sau:*

$$\begin{cases} p_0 = \frac{\pi}{4} \\ p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n)} \, \forall n \geq 0 \end{cases}$$

Ta có bảng lặp, đạt kết quả tốt ngay khi $n = 3$*:*

n	p_n
0	$\frac{\pi}{4}$
1	0,739 536 134
2	0,739 085 178
3	0,739 085 133
4	0,739 085 133

2.4.2 Khả năng hội tụ của phương pháp Newton

Với phương pháp điểm bất động, việc chọn hàm g là mấu chốt để đạt được kết quả tốt. Với phương pháp Newton, điểm khởi đầu p_0 lại đóng vai trò tối quan trọng. Biện luận trong 2.4.1 chỉ đúng khi $|p - p_0|$ nhỏ, tức ta có xấp xỉ khởi đầu tốt (tuy nhiên vẫn có những ngoại lệ, không cần xấp xỉ tốt vẫn có thể hội tụ tới nghiệm).

Sau đây ta sẽ xem xét kĩ hơn khả năng hội tụ của phương pháp Newton để thấy được sự nhạy cảm của phương pháp này với điểm khởi đầu.

Định lí 2.6

Cho $f \in C^2[a,b]$. Nếu $f = 0$ có nghiệm $p \in (a,b)$ sao cho $f'(p) \neq 0$, thì tồn tại $\delta > 0$ sao cho phương pháp Newton hội tụ với bất kì xấp xỉ ban đầu $p_0 \in [p-\delta, p+\delta]$.

Chứng minh. Xét công thức chuỗi của phương pháp Newton.

$$\begin{cases} p_{n+1} = g(p_n) \\ g(x) = x - \frac{f(x)}{f'(x)} \end{cases}$$

Dễ thấy công thức của phương pháp Newton có dạng giống như phương pháp điểm bất định (kết quả của g ở vòng lặp này là đầu vào cho g ở vòng lặp kế tiếp; p là điểm bất động của g). Vậy hướng chứng minh tiềm năng là sử dụng định lí điểm bất động, chứng minh hàm g với các điều kiện đã cho là một hàm thỏa mãn các điều kiện trong định lí điểm bất động, từ đó đảm bảo được sự hội tụ của $\{p_n\}$.

Ta chia chứng minh thành hai phần: tồn tại $\delta > 0$ mà:

1. g khả vi liên tục trên $I = [p-\delta, p+\delta]$ và $\left|g'(x)\right| < 1 \, \forall x \in I$, và
2. $g(x) \in I \, \forall x \in I$,

Ta chứng minh điều 1 thông qua bổ đề sau:

Bổ đề 2.1. *Cho $f \in C[a,b]$. Xét $p \in (a,b)$ sao cho $f(p) \neq 0$.*

(a) Nếu $f(p) \neq 0$ thì tồn tại $\delta > 0$ sao cho:

$$f(x) \neq 0 \, \forall x \in [p-\delta, p+\delta] \subseteq [a,b]$$

(b) Nếu $f(p) = 0$ và cho $K > 0$ thì tồn tại $\delta > 0$ sao cho:

$$\left|f(x)\right| \leq K \, \forall x \in [p-\delta, p+\delta] \subseteq [a,b]$$

Chứng minh Bổ đề 2.1.
Do $f \in C[a,b]$, mà $p \in (a,b)$, theo định nghĩa tính liên tục, ta có:

$$\lim_{x \to p} f(x) = f(p)$$

Phân tích phương trình trên, theo định nghĩa giới hạn, ta biết rằng với mọi $E > 0$, tồn tại Δ sao cho:

$$\left|f(x) - f(p)\right| < E \, \forall x \in (p-\Delta, p+\Delta) \cap (a,b)$$

Với mỗi E, ta lại chọn được $\delta < \Delta$ sao cho:

$$\big|f(x) - f(p)\big| < E \,\forall x \in [p-\delta, p+\delta] \subseteq [a,b]$$

(a) Xét khi $E = \varepsilon < \big|f(p)\big|$, ta có:

$$f(p) - \varepsilon < f(x) < f(p) + \varepsilon$$

Nếu $f(p) < 0$, $f(p) + \varepsilon < 0$, dẫn đến $f(x) < 0 \,\forall x \in [p-\delta, p+\delta]$.

Nếu $f(p) > 0$, $f(p) - \varepsilon > 0$, dẫn đến $f(x) > 0 \,\forall x \in [p-\delta, p+\delta]$.

Vậy luôn chọn được δ thỏa mãn yêu cầu.

(b) Xét khi $E = K$. Do $f(p) = 0$, ta có ngay kết quả:

$$\big|f(x)\big| < K \,\forall x \in [p-\delta, p+\delta] \subseteq [a,b]$$

đpcm.

Ta chứng minh điều 1 như sau:

Vì $f \in C^2[a,b]$, nên $f' \in C^1[a,b]$. Áp dụng phần a của bổ đề 2.1 với f', ta thấy rằng tồn tại $\delta_1 > 0$ sao cho:

$$f'(x) \neq 0 \,\forall x \in [p-\delta_1, p+\delta_1] \subseteq [a,b]$$

Vì $f'(x) \neq 0$, nên g xác định trên $[p-\delta_1, p+\delta_1]$. Nói cách khác, $g \in C[a,b]$.

Lấy đạo hàm g, ta có:

$$g'(x) = 1 - \frac{f'(x)f'(x) - f''(x)f(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2}$$

Vì $f \in C^2[a,b]$, nên $f'' \in C[a,b]$, nên $g'(x)$ xác định trên $[p-\delta_1, p+\delta_1]$. Nói cách khác, $g \in C^1[a,b]$.

Xét $g'(p)$. Theo giả thuyết, $f(p) = 0$, nên $g'(p) = 0$. Áp dụng phần b của bổ đề 2.1 cho g' và cho $0 < k < 1$ bất kì, ta thấy rằng tồn tại $\delta_2 > 0$ sao cho:

$$\big|g'(x)\big| \leq k \,\forall x \in [p-\delta_2, p+\delta_2] \subseteq [a,b]$$

Chọn $0 < \delta < \min\{\delta_1, \delta_2\}$. Vậy ta chứng minh được điều 1, rằng tồn tại $\delta > 0$ sao cho:

$$g \in C^1 I \text{ và } \big|g'(x)\big| \leq k < 1 \,\forall x \in I \text{ với } I = [p-\delta, p+\delta]$$

Ta chứng minh điều 2 như sau:

Theo định lí giá trị trung bình, với mọi $x \in I$, tồn tại ξ nằm giữa x và p sao cho:

$$\big|g(x) - g(p)\big| = \big|g'(\xi)\big||x-p|$$

Theo công thức chuỗi của phương pháp Newton, p là điểm bất động của g, tức $p = g(p)$. Thế vào phương trình trên, ta có:

$$\big|g(x) - p\big| = \big|g(x) - g(p)\big| = \big|g'(\xi)\big||x - p| \leq k|x - p| < |x - p|$$

Do $x \in I$ nên $|x - p| < \delta$, do đó:

$$\big|g(x) - p\big| < |x - p| < \delta \iff g(x) \in I$$

Vậy ta chứng minh được điều 2.

Tổng hợp lại, đến đây, ta chứng minh được tồn tại $\delta > 0$ sao cho:

$$\begin{cases} g \in C^1 I = [p - \delta, p + \delta] \text{ và } \big|g(x)\big| < 1 \, \forall x \in I \\ g(x) \in I \, \forall x \in I \end{cases}$$

Áp dụng định lí điểm bất động, ta thấy chuỗi $\{p_n\}_{n=0}^{\infty}$ cho bởi công thức của phương pháp Newton hội tụ về điểm bất động p của g, cũng là nghiệm của f, với p_0 khởi đầu bất kì thuộc $[p - \delta, p + \delta]$. đpcm.

Kết quả trên có ý nghĩa quan trọng về lí thuyết do đảm bảo sự hội tụ của phương pháp Newton khi biết δ "phù hợp", tuy nhiên trong bài tập lại ít dùng do không nói đến cách tìm δ.

Theo kinh nghiệm, phương pháp Newton có một đặc điểm hữu dụng là nó hội tụ nhanh với p_0 tốt, đồng thời phân kì nhanh với p_0 xấu, cho phép loại bỏ các điểm khởi đầu kém.

2.4.3 Phương pháp dây cung

Một điểm yếu quan trọng của phương pháp Newton là cần phải tính đạo hàm của f, và có nhiều trường hợp đạo hàm rất khó tính hay ước lượng.

Để tránh điểm yếu này, *phương pháp dây cung* (*secant method*) có một chút thay đổi trong công thức chuỗi.

Theo định nghĩa:

$$f'(p_n) = \lim_{x \to p_n} \frac{f(x) - f(p_n)}{x - p_n}$$

Nếu p_{n-1} gần với p_n, ta có:

$$f'(p_n) \approx \frac{f(p_{n-1}) - f(p_n)}{p_{n-1} - p_n} = \frac{f(p_n) - f(p_{n-1})}{p_n - p_{n-1}}$$

(Công thức trên là một xấp xỉ *sai phân hữu hạn* (*finite difference*)). Thay vào công thức 2.6, ta có:

$$p_{n+1} = p_n - \frac{f(p_n)}{\frac{f(p_n) - f(p_{n-1})}{p_n - p_{n-1}}} = p_n - \frac{f(p_n)(p_n - p_{n-1})}{f(p_n) - f(p_{n-1})} \tag{2.7}$$

Đây chính là công thức được dùng trong phương pháp dây cung.

Phương pháp 2.4: Phương pháp dây cung (secant method)

Phương pháp này cho phép tìm nghiệm p của $f(p) = 0$ khi biết *hai* điểm khởi đầu p_1 và p_2.
Đặt p_3 là hoành độ của giao điểm của đường thẳng nối $(p_1, f(p_1))$ và $(p_2, f(p_2))$ với Ox. Tiếp tục thực hiện phương pháp trên với p_2 và p_3 và các số sau đó, chuỗi sẽ hội tụ đến p.

Chú ý rằng, công thức 2.7 có thể được rút ngắn thành:

$$p_{n+1} = \frac{f(p_n)p_{n-1} - f(p_{n-1})p_n}{f(p_n) - f(p_{n-1})} \tag{2.8}$$

Công thức trên cũng chính là công thức nhận được khi sử dụng cách tính thông thường (tìm phương trình đường thẳng nối hai điểm $(p_1, f(p_1))$ và $(p_2, f(p_2))$ rồi tìm giao điểm với Ox). So sánh với công thức (2.7), ta thấy:

- Công thức (2.8) có tỉ số và mẫu số đều là những số rất nhỏ (do số trừ và số bị trừ ở hai hiệu đều xấp xỉ nhau), dẫn đến sai lệch khi tính toán.

- Công thức (2.7) nhân hiệu ở tử số với $f(p_n)$ *trước khi* chia, do đó giảm được sai lệch khi tính toán.

Do vậy, trong sử dụng thực tế, công thức 2.7 được sử dụng nhiều hơn.

2.4.4 Phương pháp điểm sai

Phương pháp điểm sai (*false position method*, hay *Regula Falsi*) sinh ra các xấp xỉ theo cách gần giống với phương pháp dây cung, tuy nhiên phương pháp này có một khác biệt quan trọng: nó đảm bảo nghiệm luôn nằm trong khoảng đang tìm, giống phương pháp chia đôi, và thu hẹp khoảng này sau mỗi lần lặp. Tính chất này gọi là *bracketing*, và phương pháp Newton và dây cung không có tính chất này.

Phương pháp 2.5: Phương pháp điểm sai (Regula Falsi)

Phương pháp này cho phép tìm nghiệm p của $f(p) = 0$ khi biết *hai* điểm khởi đầu p_0 và p_1 sao cho $f(p_0)$ và $f(p_1)$ trái dấu.
Đặt p_2 là hoành độ của giao điểm của đường thẳng nối $(p_0, f(p_0))$ và $(p_1, f(p_1))$ với Ox.
Cách tính p_3 như sau:

- Nếu $f(p_1$ và $f(p_2)$ trái dấu, thì p_3 là hoành độ của giao điểm của đường thẳng nối $(p_2, f(p_2))$ và $(p_1, f(p_1))$ với Ox.
- Nếu không, thì p_3 là hoành độ của giao điểm của đường thẳng nối $(p_2, f(p_2))$ và $(p_0, f(p_0))$ với Ox; đồng thời, tráo đổi giá trị của p_0 và p_1.

Tương tự với p_3, tính được p_4, p_5,...

Ta thấy yêu cầu $f(p_0)$ và $f(p_1)$ trái dấu là hiển nhiên do phương pháp này có tính "bracketing", nếu không có nghiệm trong khoảng giữa p_0 và p_1 thì phương pháp sẽ không hội tụ.

Ta cũng thấy yêu cầu tráo giá trị của p_0 và p_1 để đảm bảo nghiệm nằm giữa p_2 và p_3. Ngoài ra, do việc tráo giá trị, khi thuật toán kết thúc, p chỉ đảm bảo kẹp giữa 2 giá trị cuối p_N và p_{N-1}.

Chương 3

Giải hệ phương trình

Chúng ta có 2 loại hệ phương trình:

- Hệ phương trình tuyến tính
- Hệ phương trình phi tuyến

Chúng ta đã biết các phương pháp giải trực tiếp và gián tiếp hệ phương trình tuyến tính:

- Các phương pháp giải trực tiếp:
  - Phương pháp Cramer
  - Phương pháp thế
  - Phương pháp sử dụng ma trận nghịch đảo
  - Phương pháp khử Gauss, Gauss-Jordan
- Các phương pháp giải gián tiếp
  - Phương pháp lặp đơn
  - Phương pháp lặp Seidel
- Các phương pháp tìm trị riêng và véc tơ riêng của ma trận

3.1 Đặt bài toán và phương pháp giải

Hệ phương trình tuyến tính n phương trình, n ẩn x_1, x_2, ..., x_n là tập n phương trình E_1, E_2, ..., E_n dạng

$$\begin{cases} E_1 : a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ E_2 : a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \ldots \\ E_n : a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{cases}$$

với các hệ số a_{jk} và b_j đã biết. Hệ được gọi là thuần nhất nếu các b_j bằng 0, trong trường hợp ngược lại được gọi là hệ không thuần nhất.

Dùng cách biểu diễn ma trận, ta có thể viết hệ (3.1) dưới dạng:

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{3.1}$$

Ở đây, ma trận hệ số $A = [a_{jk}]$ là ma trận vuông cấp n, x và b là các véc tơ cột:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Ma trận $\tilde{\boldsymbol{A}}$ sau được gọi là ma trận mở rộng (augmented matrix) của hệ (3.1):

$$\tilde{\boldsymbol{A}} = [\boldsymbol{A}, \boldsymbol{b}] = \left(\begin{array}{cccc:c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{array}\right)$$

Nghiệm của (3.1) là bộ số x_1, x_2, ..., x_n thỏa mãn tất cả phương trình của hệ.

Véc tơ nghiệm của (3.1) là $\boldsymbol{x}$ mà các thành phần của nó lập thành một nghiệm của (3.1).

Hệ phương trình (3.1) có thể giải được bằng

- *phương pháp trực tiếp* (phương pháp khử Gauss, ...), hoặc
- *phương pháp gián tiếp* hay *phương pháp lặp*

3.2 Phương pháp khử Gauss

Chúng ta đã biết phương pháp sử dụng định thức để giải hệ phương trình (3.1), đó là *phương pháp Cramer*. Ở đây ta xét *phương pháp khử Gauss*, cũng là một phương pháp trực tiếp, để giải hệ phương trình tuyến tính.

3.2.1 Phương pháp khử Gauss

Phương pháp 3.1: Phương pháp khử Gauss

Phương pháp này khử liên tiếp các ẩn để từng bước đưa ma trận mở rộng $\tilde{\boldsymbol{A}}$ về dạng tam giác trên (upper triangular matrix, hay dạng bậc thang). Phương pháp gồm hai phần thực hiện lần lượt như sau:

1. *Quá trình thuận (forward elimination)*: Dùng các phép biến đổi hàng sơ cấp để đưa $\tilde{\boldsymbol{A}}$ về dạng tam giác trên:
   (a) Khử x_1 từ $E_{\geq 2}$ bằng cách:
   $$E_j \coloneqq E_j - \frac{a_{j1}}{a_{11}} E_1 \, \forall j \in [2, n] \tag{3.2}$$
   a_{11} gọi là *phần tử trục xoay (pivot)*, E_1 gọi là *phương trình chính (pivot equation)*.
   (b) Với $i = 2, \ldots, n-1$, khử x_i từ $E_{>i}$ bằng cách tương tự như trên, cuối cùng thu được dạng tam giác trên.
2. *Quá trình ngược (back substitution)*: Giải $\boldsymbol{x}$ từ cuối lên:
   (a) Giải x_n từ E_n, giải tiếp được x_{n-1} do đã biết x_n.
   (b) Tương tự giải được đến x_1, cuối cùng thu được nghiệm $\boldsymbol{x}$.

(*"back"* trong cụm *"back substitution"* nghĩa là các ẩn được giải từ cuối lên, do trước đó đã đưa được $\boldsymbol{A}$ về dạng tam giác trên. Tương tự, nếu $\boldsymbol{A}$ ở dạng tam giác dưới, việc thế từ trên xuống sẽ gọi là *"forward substitution"*. Dù hai dạng này tương đương, trong khử Gauss ta chỉ nói đến back substitution).

Ta kí hiệu ma trận $\boldsymbol{A}$ khởi đầu là $\boldsymbol{A}^{(1)}$. Sau khi kết thúc bước khử x_i, ma trận $\boldsymbol{A}^{(i)}$ sẽ biến đổi thành $\boldsymbol{A}^{(i+1)}$. Bước cuối cùng là bước khử x_{n-1} (vì x_n không cần khử), và ma trận sau cùng là $\boldsymbol{A}^{(n)}$.

Xem xét kĩ hơn, ta thấy (3.2) có một chi tiết nhạy cảm là phép chia. Nếu $a_{11} = 0$, phương pháp không thể thực hiện được, cho dù thực tế hệ phương trình có thể có nghiệm. Một cách xử lí trường hợp này là *đổi vị trí hàng*, như trong hai ví dụ sau:

Ví dụ 3.1. *Hãy giải hệ phương trình:*

$$8x_2 + 2x_3 = -7 \tag{E_1}$$
$$3x_1 + 5x_2 + 2x_3 = 8 \tag{E_2}$$
$$6x_1 + 2x_2 + 8x_3 = 26 \tag{E_3}$$

Chúng ta xoay trục từ E_1 , nhưng do E_1 không có ẩn x_1, trong khi đó hệ số của x_1 trong phương trình E_3 là lớn nhất. Vì vậy ta đổi chỗ E_1 và E_3 cho nhau.

Tới đây ta có ma trận mở rộng như sau:

$$\tilde{\boldsymbol{A}} = \left(\begin{array}{ccc:c} 6 & 2 & 8 & 26 \\ 3 & 5 & 2 & 8 \\ & 8 & 2 & -7 \end{array}\right)$$

Khử x_1 được:

$$\tilde{\boldsymbol{A}}^{(1)} = \left(\begin{array}{ccc:c} 6 & 2 & 8 & 26 \\ & 4 & -2 & -5 \\ & 8 & 2 & -7 \end{array}\right)$$

Khử x_2 được:

$$\tilde{\boldsymbol{A}}^{(2)} = \left(\begin{array}{ccc:c} 6 & 2 & 8 & 26 \\ & 4 & -2 & -5 \\ & & 6 & 3 \end{array}\right)$$

Vậy ta giải được $x_3 = 0{,}5$, $x_2 = -1$, $x_1 = 4$.

Ví dụ 3.2. *Hãy giải hệ phương trình:*

$$\begin{cases} x_1 - x_2 + 2x_3 - x_4 = -8 & (E_1) \\ 2x_1 - 2x_2 + 3x_3 - 3x_4 = -20 & (E_2) \\ x_1 + x_2 + x_3 = -2 & (E_3) \\ x_1 - x_2 + 4x_3 + 3x_4 = 4 & (E_4) \end{cases}$$

Ta có ma trận mở rộng như sau:

$$\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{A}}^{(1)} = \left(\begin{array}{cccc:c} 1 & -1 & 2 & -1 & -8 \\ 2 & -2 & 3 & -3 & -20 \\ 1 & 1 & 1 & 0 & -2 \\ 1 & -1 & 4 & 3 & 4 \end{array}\right)$$

Khử x_1 bằng chuỗi biến đổi

$$E_2 \coloneqq E_2 - 2E_1;\ E_3 \coloneqq E_3 - E_1;\ E_4 \coloneqq E_4 - E_1$$

cho kết quả:

$$\tilde{\boldsymbol{A}}^{(2)} = \left(\begin{array}{cccc:c} 1 & -1 & 2 & -1 & -8 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right)$$

Điểm quay $a_{22}^{(2)} = 0$, do đó cần phải đổi hàng. Ta chọn $a_{32} \neq 0$, do đó đổi chỗ E_2 và E_3:

$$\tilde{\boldsymbol{A}}^{(2)} = \left(\begin{array}{cccc:c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right)$$

Do x_2 đã được khử khỏi E_3 và E_4, ta có:

$$\tilde{\boldsymbol{A}}^{(3)} = \tilde{\boldsymbol{A}}^{(2)} = \left(\begin{array}{cccc:c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right)$$

Khử x_3 bằng

$$E_4 \coloneqq E_4 - (-2)E_3$$

cho kết quả:

$$\tilde{\boldsymbol{A}}^{(4)} = \left(\begin{array}{cccc:c} 1 & 1 & 2 & 1 & -8 \\ 0 & 2 & 1 & 1 & 6 \\ 0 & 0 & 1 & 1 & -4 \\ 0 & 0 & 0 & 2 & 4 \end{array}\right)$$

Vậy ta giải được $x_4 = 2$, $x_3 = 2$, $x_2 = 3$, $x_1 = -7$.

Phương pháp khử Gauss còn có một số biến thể:

- Phương pháp Gauss-Jordan: đưa $\boldsymbol{A}$ về dạng đường chéo thay vì dạng tam giác trên.
- Phương pháp Doolittle, phương pháp Crout, phương pháp Cholesky: đều dựa trên phân tích LU (LU factorization), sẽ giới thiệu ở phần sau.

3.2.2 Độ phức tạp

Ta phân tích độ phức tạp của phương pháp khử Gauss:

- Có tổng cộng $n - 1$ bước khử.
- Ở bước k, ta khử x_k trong các phương trình $E_{>k}$, tổng cộng là $n - k$ phương trình.
- Trong mỗi phương trình:
  - Có 1 phép chia: ví dụ $\frac{a_{i1}}{a_1 1}$ trong (3.2)
  - Có $n - k + 1$ phép nhân: ví dụ $\frac{a_{i1}}{a_1 1}E_1$ trong (3.2)
  - Có $n - k + 1$ phép trừ: ví dụ chính (3.2)

Do đó, tổng số phép tính của phương pháp này là:

$$\begin{aligned} C(n) &= \sum_{k=1}^{n-1}(n-k) + 2\sum_{k=1}^{n-1}(n-k)(n-k+1) \\ &= \sum_{s=1}^{n-1} s + 2\sum_{k=1}^{n-1} s(s+1) \,(\text{with } s = n-k) \\ &= \frac{1}{2}n(n-1) + \frac{2}{3}n(n^2-1) \\ &\approx \frac{2}{3}n^3 \end{aligned}$$

Ta nói rằng phương pháp khử Gauss có độ phức tạp $\mathcal{O}(n^3)$.

3.3 Phân tích LU và Ma trận nghịch đảo

Chúng ta tiếp tục thảo luận các phương pháp số giải hệ phương trình tuyến tính n phương trình, n ẩn. Trong phần này, ta xem xét ba phương pháp cải tiến từ phương pháp khử Gauss, cho phép tìm nghiệm nhanh chóng hơn, gồm phương pháp Doolittle, phương pháp Cholevsky, và phương pháp Crout. Cả ba phương pháp đều dựa trên phân tích LU.

Điểm chung của các phương pháp này là đều cố gắng đưa ma trận về tích của các ma trận tam giác trên (Upper triangular matrix) và ma trận tam giác dưới (Lower triangular matrix). Hai dạng ma trận này rất hữu ích vì cho phép tìm $\boldsymbol{x}$ với độ phức tạp $\mathcal{O}(n^2)$; nếu xét theo khía cạnh này, phương pháp khử Gauss khác ở điểm là *trực tiếp* biến đổi về dạng tam giác. Tất nhiên, phần lớn tính toán lại chuyển về việc phân tích ra dạng ma trận đặc biệt này, và với một số dạng ma trận đặc biệt, độ phức tạp có thể thấp hơn $\mathcal{O}(n^3)$.

Ba phương pháp được trình bày trong phần này đều gắn chặt với một kiểu phân tích ra ma trận tam giác. Vì rất dễ để giải nghiệm từ dạng tam giác, tên phương pháp vừa chỉ phương pháp phân tích ra dạng tam giác tương ứng, vừa chỉ phương pháp giải hệ tuyến tính từ dạng tam giác đã phân tích.

3.3.1 Phân tích LU & phương pháp Doolittle

Định nghĩa 3.1: Phân tích LU (LU factorization)

Phân tích LU là việc phân tích ma trận vuông $\boldsymbol{A}$ thành tích của hai ma trận

$$\boldsymbol{A} = \boldsymbol{LU}$$

trong đó $\boldsymbol{L}$ là ma trận tam giác dưới và $\boldsymbol{U}$ là ma trận tam giác trên.

Một số lưu ý về phân tích LU:

- Không phải mọi ma trận vuông đều có phân tích LU. Tuy nhiên, ta thừa nhận một kết quả quan trọng là mọi ma trận khả nghịch $\boldsymbol{A}_0$ đều có thể sắp xếp lại các hàng để thu được một ma trận $\boldsymbol{A}$ có phân tích LU.

- Có thể có nhiều cách phân tích LU.

Sau đây ta tìm hiểu phương pháp Doolittle để phân tích LU. Ta sẽ xem xét kết quả của phương pháp phân tích LU này với bài toán giải hệ phương trình tuyến tính trước, sau đó mới đi sâu vào công thức toán học của nó.

Phương pháp Doolittle: Dùng dạng LU để giải hệ tuyến tính

Định lí 3.1

Nếu hệ tuyến tính $\boldsymbol{Ax} = \boldsymbol{b}$ có thể giải được bằng khử Gauss mà không cần đổi vị trí hàng, thì $\boldsymbol{A}$ có thể phân tích thành tích của ma trận tam giác trên $\boldsymbol{L}$ và tam giác dưới $\boldsymbol{U}$ (tức $\boldsymbol{A} = \boldsymbol{LU}$) với dạng sau:

$$\boldsymbol{U} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \end{pmatrix}, \text{ và } \boldsymbol{L} = \begin{pmatrix} 1 & & & \\ m_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ m_{n1} & \dots & m_{n,n-1} & 1 \end{pmatrix}$$

trong đó

- a_{ji}^{k} là hệ số của x_i trong phương trình E_j tại bước khử thứ k
- $m_{ji} = \dfrac{a_{ji}^{(i)}}{a_{ii}^{(i)}}$ là hệ số của E_j trong, ví dụ, công thức (3.2).

$\boldsymbol{U}$ cũng chính là ma trận $\boldsymbol{A}$ thu được sau khi khử Gauss.

Định lí trên cũng chính là công thức cho phương pháp Doolittle.

Chú ý rằng $\boldsymbol{U}$ là $\boldsymbol{A}$ sau khi khử Gauss, tức ma trận $\tilde{\boldsymbol{A}}$ trong phần 3.2 nhưng bỏ đi cột cuối.

Cần nhắc lại rằng, đến đây, ta vẫn cần giả sử hệ có thể giải được bằng khử Gauss mà *không* cần đổi vị trí hàng. Mở rộng phương pháp Doolittle cho trường hợp cần đổi vị trí hàng không khó. Trước hết, ta cần tìm cách biểu diễn việc tráo đổi vị trí hàng.

Định nghĩa 3.2: Ma trận hoán vị (permutation matrix)

Ma trận hoán vị $\boldsymbol{P}$ là ma trận có được bằng cách sắp xếp lại các hàng của I tùy ý.

Nhân P vào bên trái A sẽ tráo các hàng của A theo đúng cách tráo các hàng của I để tạo ra P. Nói cách khác, P là tích các ma trận của *biến đổi sơ cấp tráo hàng (row swapping elementary operation)*.

Ta chọn $\boldsymbol{P}$ sao cho $\boldsymbol{PA}$ có thể giải bằng khử Gauss mà không cần tráo vị trí hàng. $\boldsymbol{P}$ được xây dựng đơn giản bằng cách áp dụng cách tráo hàng của Gauss cho ma trận $\boldsymbol{I}$.

Ví dụ 3.3. *Tìm ma trận hoán vị cho ma trận trong ví dụ 3.1, sao cho sau khi ma trận đó với ma trận gốc, nhận được một ma trận có thể dùng phương pháp khử Gauss.*

Trong ví dụ trên, hệ có thể dùng phương pháp khử Gauss sau khi đổi chỗ E_1 và E_3. Vậy một ma trận hoán vị phù hợp là:

$$\boldsymbol{P} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Sau khi tìm được $\boldsymbol{P}$, $\boldsymbol{PA}$ có thể áp dụng phương pháp khử Gauss mà không cần tráo vị trí hàng. Theo định lí 3.1, $\boldsymbol{PA}$ có phân tích LU:

$$\begin{aligned} &\boldsymbol{PA} = \boldsymbol{LU} \\ \iff &\boldsymbol{A} = \boldsymbol{P}^{-1}\boldsymbol{LU} = (\boldsymbol{P}^t\boldsymbol{L})\boldsymbol{U} \,(\text{do } \boldsymbol{P}^{-1} = \boldsymbol{P}^t) \end{aligned}$$

Chứng minh phương pháp Doolittle

Phần này chứng minh kĩ hơn về tính đúng đắn của phương pháp Doolittle, và có thể bỏ qua.

Chứng minh định lí 3.1. Ta xem xét công thức (3.2). Xét trường hợp $j = 2$:

$$E_2 \coloneqq E_2 - \frac{a_{21}}{a_{11}}E_1 = E_2 - m_{21}E_1$$

Sử dụng phép biến đổi sơ cấp cộng một hàng với α lần một hàng khác (row addition elementary operation), ta có ma trận biến đổi sau:

$$\boldsymbol{M}_2^{(1)} = \begin{pmatrix} 1 & & & \\ -m_{21} & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

Tương tự, cùng trong bước khử đầu tiên (khử x_1) này, ta có dạng tổng quát hơn của $\boldsymbol{M}_j^{(1)}$:

$$\boldsymbol{M}_j^{(1)} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ -m_{j1} & & \ddots & \\ & & & 1 \end{pmatrix}$$

Sau khi kết thúc bước khử x_1, $\boldsymbol{A}$ được biến đổi thành

$$\boldsymbol{M}_2^{(1)}\boldsymbol{M}_3^{(1)}\dots\boldsymbol{M}_n^{(1)}\boldsymbol{A} = \boldsymbol{M}^{(1)}\boldsymbol{A} = \boldsymbol{M}^{(1)}\boldsymbol{A}^{(1)} = \boldsymbol{A}^{(2)}$$

và hơn nữa

$$\boldsymbol{A}^{(2)}\boldsymbol{x} = \boldsymbol{M}^{(1)}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{M}^{(1)}\boldsymbol{b} = \boldsymbol{b}^{(2)}$$

Dễ dàng chứng minh được $\boldsymbol{M}^{(1)}$ ở trên, gọi là ma trận biến đổi Gauss thứ nhất, có dạng sau:

$$\boldsymbol{M}^{(1)} = \begin{pmatrix} 1 & & & \\ -m_{21} & 1 & & \\ \vdots & & \ddots & \\ -m_{n1} & & & 1 \\ & & & \end{pmatrix}$$

Tương tự, ta chứng minh được $\boldsymbol{M}^k$ (ma trận biến đổi Gauss thứ k) có dạng sau:

$$\boldsymbol{M}^{(k)} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -m_{k+1,k} & \ddots & & \\ & & \vdots & & \ddots & \\ & & -m_{n,k} & & & 1 \\ & & & & & \end{pmatrix}$$

Cuối cùng, sau khi kết thúc bước khử x_n:

$$\boldsymbol{M}^{(1)}\boldsymbol{M}^{(2)}\dots\boldsymbol{M}^{(n)}\boldsymbol{A} = \boldsymbol{M}\boldsymbol{A} = \boldsymbol{A}^{(n)}$$

trong đó $\boldsymbol{A}^{(n)}$ chính là ma trận $\boldsymbol{A}$ sau khi khử Gauss:

$$\boldsymbol{A}^{(n)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \\ & & & \end{pmatrix}$$

Tới đây, chúng ta đã biến đổi được phương trình ban đầu (3.1) sang dạng sau:

$$\boldsymbol{A}^{(n)}\boldsymbol{x} = \boldsymbol{M}^{(n-1)}\boldsymbol{A}^{(n-1)}\boldsymbol{x} = \boldsymbol{M}^{(n-1)}\boldsymbol{b}^{(n-1)} = \boldsymbol{b}^{(n)} \tag{3.3}$$

trong đó $\boldsymbol{A}^{(n)}$ là một ma trận tam giác trên.

Giờ đây, nếu coi $\boldsymbol{A}^{(n)}$ là thành phần $\boldsymbol{U}$ cần tìm, ta cần nhân vào trước hai vế của (3.3) một thành phần $\boldsymbol{L}$ nào đó để đưa (3.3) về lại (3.1).

Thành phần $\boldsymbol{U}$ đã thấy chỉ đơn giản là $\boldsymbol{A}$ qua một chuỗi các biến đổi sơ cấp cộng một hàng với α lần một hàng khác. Do đó $\boldsymbol{L}$ chỉ cần là một ma trận có thể đảo ngược chuỗi biến đổi này.

Xét biến đổi $\boldsymbol{M}_2^{(1)}$, biến đổi

$$\boldsymbol{L}_2^{(1)} = \begin{pmatrix} 1 & & & \\ m_{21} & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

sẽ đảo ngược được $\boldsymbol{M}_2^{(1)}$. Lý do cũng rất đơn giản:

- $\boldsymbol{M}_2^{(1)}$ lấy E_2 trừ đi m_{21} lần E_1, thì
- $\boldsymbol{L}_2^{(1)}$ lấy E_2 cộng với m_{21} lần E_1

Tương tự, ta dễ dàng thấy

$$\boldsymbol{L}^{(k)} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & m_{k+1,k} & \ddots & & \\ & & \vdots & & \ddots & \\ & & m_{n,k} & & & 1 \end{pmatrix}$$

sẽ đảo ngược được $\boldsymbol{M}^{(k)}$.

Nhân $\boldsymbol{L}^{(k)}$ vào trước hai vế của (3.3) theo thứ tự k tăng dần từ 1 đến $n-1$, ta có:

$$\begin{aligned} & \boldsymbol{L}^{(1)}\boldsymbol{L}^{(2)}\dots\boldsymbol{L}^{(n-1)}\boldsymbol{A}^{(n)}\boldsymbol{x} = \boldsymbol{L}^{(1)}\boldsymbol{L}^{(2)}\dots\boldsymbol{L}^{(n-1)}\boldsymbol{b}^{(n)} \\ \iff\ & \boldsymbol{L}^{(1)}\boldsymbol{L}^{(2)}\dots\boldsymbol{L}^{(n-1)}\boldsymbol{M}^{(n-1)}\dots\boldsymbol{M}^{(2)}\boldsymbol{M}^{(1)}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{L}^{(1)}\boldsymbol{L}^{(2)}\dots\boldsymbol{L}^{(n-1)}\boldsymbol{M}^{(n-1)}\dots\boldsymbol{M}^{(2)}\boldsymbol{M}^{(1)}\boldsymbol{b} \\ & \iff \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

Đến đây, ta nhận được phương trình (3.1) ban đầu. Vậy tích các $\boldsymbol{L}^{(k)}$ theo thứ tự trên là một giá trị $\boldsymbol{L}$ phù hợp. Không khó để chứng minh rằng:

$$\boldsymbol{L} = \boldsymbol{L}^{(1)}\boldsymbol{L}^{(2)}\dots\boldsymbol{L}^{(n-1)} = \begin{pmatrix} 1 & & & & & \\ m_{21} & \ddots & & & & \\ \vdots & \ddots & 1 & & & \\ \vdots & \vdots & m_{k+1,k} & \ddots & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ m_{n1} & \dots & m_{n,k} & \dots & m_{n,n-1} & 1 \end{pmatrix}$$

Vậy ta đã xây dựng được phân tích LU của $\boldsymbol{A}$, với $\boldsymbol{L}$ và $\boldsymbol{U}$ có dạng như yêu cầu.

đpcm.

3.3.2 Phân tích LL' & phương pháp Cholevsky

Thuật toán Cholesky phân tích một ma trận xác định dương ra dạng LL'. Để tìm hiểu phân tích này, ta cần biết về ma trận xác định dương.

Định nghĩa 3.3: Ma trận xác định dương (positive definitive matrix)

Ma trận $\boldsymbol{A}$ $n \times n$ gọi là xác định dương nếu:

- $\boldsymbol{A}$ đối xứng, và
- $\boldsymbol{x}^t \boldsymbol{A}\boldsymbol{x} > 0 \forall \boldsymbol{x} \neq 0$

$\boldsymbol{L}$ có dạng như sau (viết rõ lại để tiện biểu diễn về sau):

$$\boldsymbol{L} = \begin{pmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & \dots & \dots & l_{nn} \\ & & & \end{pmatrix}$$

Ta thừa nhận định lí sau:

Định lí 3.2

Ma trận $\boldsymbol{A}$ là xác định dương khi và chỉ khi nó phân tích được ra dạng $\boldsymbol{L}\boldsymbol{L}^t$, trong đó $\boldsymbol{L}$ là ma trận tam giác dưới với đường chéo chính khác 0.

Do có số ẩn không lớn (với ma trận $n \times n$ cần tìm tổng cộng $\frac{n(n+1)}{2}$ ẩn), đồng thời ma trận tích có dạng phù hợp, nên phương pháp này có thể giải bằng tay các ẩn theo cách thế thông thường với n nhỏ. Nếu không tiện tính tay, ta có phương pháp chính xác sau:

Phương pháp 3.2: Phương pháp Cholevsky

Phương pháp Cholevsky phân tích ma trận xác định dương $\boldsymbol{A}$ thành dạng LL'.
Lần lượt tính phần tử khác 0 ở cột 1, 2, ... Bước thứ i sẽ tính các phần tử thuộc cột i như sau:

- Tính l_{ii}:

$$l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \right)^{0,5}$$

- Tính các phần tử còn lại, nếu có:

$$l_{ji} = \frac{1}{l_{ii}} \left(a_{ji} - \sum_{k=i+1}^{n} l_{jk} l_{ik} \right) \mid j \in [i+1, n]$$

3.3.3 Phân tích LU cho ma trận dải & thuật toán Crout

Ta quay trở lại với phân tích LU, tuy nhiên sử dụng thuật toán khác, nhanh vượt trội so với Doolittle, cho một loại ma trận đặc biệt, ma trận dải.

Định nghĩa 3.4: Ma trận dải (band matrix)

Ma trận $n \times n$ gọi là ma trận dải nếu có $1 < p, q < n$ sao cho $a_{ij} = 0$ khi $j - i \geq p$ hoặc $i - j \geq q$.

Nói cách khác, hai chỉ số p, q chỉ số đường chéo mà các phần tử trên đó không nhất thiết bằng 0:

- p đường chéo trên đường chéo chính, gồm cả đường chéo chính
- q đường chéo dưới đường chéo chính, gồm cả đường chéo chính

Ta lại xét tiếp một trường hợp đặc biệt: ma trận dải với $p = q = 2$, gọi là *ma trận ba đường chéo (tridiagonal matrix)*.
Nếu $\boldsymbol{A}$ trong (3.1) có dạng ba đường chéo:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & & & \\ a_{21} & a_{22} & a_{23} & & \\ & a_{32} & a_{33} & \ddots & \\ & & \ddots & \ddots & a_{n-1,n} \\ & & & a_{n,n-1} & a_{nn} \end{pmatrix}$$

thì có thể phân tích $\boldsymbol{A}$ ra dạng LU như sau:

$$\boldsymbol{L} = \begin{pmatrix} l_{11} & & & & \\ l_{21} & l_{22} & & & \\ & l_{32} & l_{33} & & \\ & & \ddots & \ddots & \\ & & & l_{n,n-1} & l_{nn} \end{pmatrix}, \text{ và } \boldsymbol{U} = \begin{pmatrix} 1 & u_{12} & & & \\ & 1 & u_{23} & & \\ & & \ddots & \ddots & \\ & & & \ddots & u_{n-1,n} \\ & & & & 1 \end{pmatrix}$$

Do có số ẩn không lớn (với ma trận $n \times n$ cần tìm tổng cộng $3n - 2$ ẩn), đồng thời ma trận tích có dạng phù hợp, nên phương pháp này có thể giải bằng tay các ẩn theo cách thế thông thường với n nhỏ. Nếu không tiện tính tay, ta giới thiệu qua phương pháp sau:

Phương pháp 3.3: Phương pháp Crout cho ma trận ba đường chéo

Phương pháp Crout dùng để phân tích ma trận ra dạng LU. Dạng LU của phương pháp Crout có khác biệt so phương pháp Doolittle:

- Phương pháp Crout tạo ra ma trận $\boldsymbol{U}$ có đường chéo chính bằng 1.
- Phương pháp Doolittle tạo ra ma trận $\boldsymbol{L}$ có đường chéo chính bằng 1.

3.4 Các phương pháp lặp

Phương pháp khử Gauss và các biến thể của nó trong hai phần cuối cùng là các *phương pháp trực tiếp* để giải hệ phương trình tuyến tính; đây là những phương pháp đưa ra nghiệm sau một số tính toán được xác định trước. Ngược lại, trong trường hợp *giải gián tiếp* hoặc *phương pháp lặp (iterative method)* chúng ta bắt đầu từ một giá trị xấp xỉ nghiệm đúng và, nếu thành công, sẽ có được xấp xỉ tốt hơn và tốt hơn từ một quá trình tính toán lặp đi lặp lại. Trong các phương pháp này, số phép tính phụ thuộc vào độ chính xác cần thiết.

Chúng ta áp dụng các phương pháp lặp nếu tốc độ hội tụ đủ nhanh (nếu ma trận có các phần tử nằm trên đường chéo chính lớn hơn các phần tử nằm ngoài, như ta sẽ thấy), hoặc với *ma trận thưa (sparse matrix)*, tức ma trận có phần lớn các phần tử là 0.

3.4.1 Phương pháp lặp Gauss - Seidel

Đây là một phương pháp lặp quan trọng, được nghiên cứu và sử dụng nhiều. Trước khi đi vào công thức chi tiết, ta xem xét phương pháp qua ví dụ sau:

Ví dụ 3.4. *Xét hệ phương trình*

$$\begin{cases} x_1 - 0{,}25x_2 - 0{,}25x_3 = 50 \\ -0{,}25x_1 + x_2 - 0{,}25x_4 = 50 \\ -0{,}25x_1 + x_3 - 0{,}25x_4 = 25 \\ -0{,}25x_2 - 0{,}25x_3 + x_4 = 25 \end{cases}$$

Viết lại hệ theo cách sau:

$$\begin{cases} x_1 = 0{,}25x_2 + 0{,}25x_3 + 50 \\ x_2 = 0{,}25x_1 + 0{,}25x_4 + 50 \\ x_3 = 0{,}25x_1 + 0{,}25x_4 + 25 \\ x_4 = 0{,}25x_2 + 0{,}25x_3 + 25 \end{cases} \tag{1}$$

Ta bắt đầu bằng một xấp xỉ (cho dù có thể khác xa nghiệm cần tìm), ví dụ $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = x_4^{(0)} = 100$ *và tính* $\boldsymbol{x}^{(1)}$ *từ các giá trị này, theo công thức* (1) *như sau:*

Giá trị cũ (chưa có giá trị mới) ↓

$$\begin{aligned} x_1^{(1)} &= 0{,}25x_2^{(0)} + 0{,}25x_3^{(0)} + 50 = 100 \\ x_2^{(1)} &= 0{,}25x_1^{(1)} + 0{,}25x_4^{(0)} + 50 = 100 \\ x_3^{(1)} &= 0{,}25x_1^{(1)} + 0{,}25x_4^{(0)} + 25 = 75 \\ x_4^{(1)} &= 0{,}25x_2^{(1)} + 0{,}25x_3^{(1)} + 25 = 68{,}75 \end{aligned}$$

↑ *Giá trị mới*

Trong tính toán trên, $\boldsymbol{x}^{(1)}$ *được tính theo công thức* (1)*, nhưng với các giá trị* x *mới nhất có được ở thời điểm tính toán. Tiếp tục tính toán, ta có bảng sau:*

Rõ ràng, nghiệm hội tụ rất nhanh.

Phương pháp Gauss-Seidel

Ta đã thấy Gauss-Seidel, cũng giống như phương pháp Newton hay phương pháp điểm bất động, là một phương pháp lặp (iterative method), tức tính một hàm lặp đi lặp lại, kết quả lần lặp trước là đầu vào của lần lặp sau. Trong phần này, ta sẽ đưa ra công thức lặp Gauss-Seidel theo một cách trực quan từ ví dụ 3.4.

Ví dụ 3.4 thực hiện được dễ dàng nhất khi hệ số của x_i trong phương trình thứ i là 1, tức đường chéo chính của $\boldsymbol{A}$ chỉ chứa 1. Để tiện trong việc đưa ra công thức, ta giả sử rằng $a_{jj} = 1 \, \forall j \in [1, n]$, do mọi hệ phương trình có thể đưa được về dạng này, thông qua các biến đổi hàng sơ cấp.

Ta tách $\boldsymbol{A}$ như sau:

$$\boldsymbol{A} = \boldsymbol{I} + \boldsymbol{L} + \boldsymbol{U}$$

trong đó $\boldsymbol{I}$ là ma trận đơn vị, $\boldsymbol{L}$ là ma trận tam giác dưới chặt (strictly), $\boldsymbol{U}$ là ma trận tam giác trên chặt.

Thay vào (3.1), ta có:

$$\begin{aligned} &\boldsymbol{Ax} = (\boldsymbol{I} + \boldsymbol{L} + \boldsymbol{U})\boldsymbol{x} = \boldsymbol{b} \\ \iff &\boldsymbol{x} = \boldsymbol{b} - \boldsymbol{Lx} - \boldsymbol{Ux} \end{aligned}$$

Theo ví dụ 3.4, ta thấy $\boldsymbol{U}$ chứa các giá trị x cũ, còn $\boldsymbol{L}$ chứa các giá trị x mới. Từ đó, ta có công thức lặp tổng quát:

$$\boldsymbol{x}^{(m+1)} = \boldsymbol{b} - \boldsymbol{L}\boldsymbol{x}^{(m+1)} - \boldsymbol{U}\boldsymbol{x}^{(m)} \tag{3.4}$$

trong đó $\boldsymbol{x}^{(m)}$ là véc tơ nghiệm xấp xỉ thứ m.

Phương pháp 3.4: Phương pháp Gauss-Seidel

Phương pháp này xấp xỉ nghiệm $\boldsymbol{x}$ của hệ $\boldsymbol{Ax} = \boldsymbol{b}$ khi

- biết xấp xỉ bắt đầu $\boldsymbol{x}^{(0)}$, và
- $\boldsymbol{A}$ không chứa 0 trên đường chéo chính

Gọi các phương trình trong hệ là E_i. Ta đưa hết x_i sang riêng vế trái của E_i sao cho hệ số của x_i là 1. Nói cách khác, ta tính x_i qua các x_j, $j \neq i$. Phương pháp Gauss-Seidel thực hiện như sau:

1. Trong mỗi lần lặp, tính $\boldsymbol{x}$ như sau:
   - lần lượt tính x_i theo thứ tự i tăng dần như công thức đã tách vế trái ở trên, và
   - dùng giá trị x_j với mới nhất có thể khi tính x_i
2. Dừng lại khi đạt điều kiện dừng, nếu không lặp lại bước trên

Chú ý rằng, phương pháp trên chỉ cần điều kiện đường chéo chính của $\boldsymbol{A}$ không chứa 0. Điều kiện đường chéo chính chứa toàn 1 được bỏ qua vì trong bước đưa x_i sang vế trái của E_i, hệ số của x_i đã được chuyển thành 1.

Điều kiện hội tụ

Ta viết lại (3.4) như sau:

$$\begin{aligned} &\boldsymbol{x}^{(m+1)} = \boldsymbol{b} - \boldsymbol{L}\boldsymbol{x}^{(m+1)} - \boldsymbol{U}\boldsymbol{x}^{(m)} \\ \iff &(\boldsymbol{I} + \boldsymbol{L})\boldsymbol{x}^{(m+1)} = \boldsymbol{b} - \boldsymbol{U}\boldsymbol{x}^{(m)} \\ \iff &\boldsymbol{x}^{(m+1)} = \boldsymbol{C}\boldsymbol{x}^{(m)} + (\boldsymbol{I} + \boldsymbol{L})^{-1}\boldsymbol{b} \text{ với } \boldsymbol{C} = -(\boldsymbol{I} + \boldsymbol{L})^{-1}\boldsymbol{U} \end{aligned}$$

Chú ý rằng, nếu muốn tính $\boldsymbol{C}$ theo công thức trên, $\boldsymbol{A}$ phải ở dạng đường chéo chính chứa toàn 1.

Trước hết ta nhớ lại định nghĩa về *véc tơ riêng* và *giá trị riêng* của ma trận:

Định nghĩa 3.5: Véc tơ riêng (eigenvector) & giá trị riêng (eigenvalue)

Với mỗi ma trận $\boldsymbol{A}$, véc tơ $\boldsymbol{v} \neq \vec{0}$ và vô hướng λ được gọi lần lượt là véc tơ riêng và giá trị riêng ứng với véc tơ riêng đó nếu:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

Điều kiện cần và đủ để λ là giá trị riêng là:

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$$

Quay trở lại phương pháp Gauss-Seidel, ta thừa nhận phương pháp này hội tụ khi và chỉ khi tất cả trị riêng của $\boldsymbol{C}$ có giá trị tuyệt đối nhỏ hơn 1. Cụ thể hơn, ta xét *bán kính phổ* của $\boldsymbol{C}$.

Định nghĩa 3.6: Bán kính phổ (spectral radius)

Bán kính phổ của ma trận $\boldsymbol{C}$ là giá trị tuyệt đối lớn nhất của các giá trị riêng của $\boldsymbol{C}$.

Phương pháp Gauss-Seidel, với bán kính phổ r của $\boldsymbol{C}$, sẽ hội tụ:

- khi và chỉ khi $r < 1$
- nhanh hơn khi r nhỏ

Như vậy, ta đã biết điều kiện *cần và đủ* để phương pháp Gauss-Seidel hội tụ. Giờ ta sẽ xem xét một số điều kiện *đủ* cho sự hội tụ này.

Phương pháp Gauss-Seidel sẽ hội tụ khi

$$\|\boldsymbol{C}\| < 1$$

$\|\boldsymbol{C}\|$ là *chuẩn (norm)* của $\boldsymbol{C}$. Ta có một số chuẩn hay gặp như sau:

- Chuẩn Frobenius, tức căn bậc hai của tổng của bình phương mọi phần tử:

$$\|\boldsymbol{C}\| = \sqrt{\sum_{j=1}^{n}\sum_{k=1}^{n} c_{jk}^2}$$

- Chuẩn tổng cột, tức giá trị lớn nhất trong tổng các trị tuyệt đối của phần tử một cột:

$$\|\boldsymbol{C}\| = \max_k \sum_{j=1}^{n} |c_{jk}|$$

- Chuẩn tổng hàng, tức giá trị lớn nhất trong tổng các trị tuyệt đối của phần tử một hàng:

$$\|\boldsymbol{C}\| = \max_j \sum_{k=1}^{n} \left|c_{jk}\right|$$

Ba chuẩn này không tương đương với nhau. Hoàn toàn có thể có trường hợp một chuẩn thỏa mãn điều kiện trị tuyệt đối nhỏ hơn 1, đủ để kết luận phương pháp hội tụ, nhưng dùng chuẩn khác thì lại không thể kết luận về sự hội tụ.

3.4.2 Phương pháp Jacobi

Phương pháp Gauss-Seidel là một phương pháp *hiệu chỉnh liên tiếp (successive correction)* vì *trong* mỗi bước lặp, ta cập nhật giá trị một thành phần x_j mỗi khi thành phần đó có xấp xỉ mới. *Phương pháp Jacobi* sẽ giới thiệu dưới đây lại thuộc loại *hiệu chỉnh đồng thời*, tức xấp xỉ mới của x_j không được dùng cho đến khi toàn bộ $\boldsymbol{x}$ trong một bước được tính xong.

Khác biệt về việc sử dụng giá trị xấp xỉ mới nói ở trên là điểm khác biệt chính yếu của hai phương pháp này. Do đó, ta có thể viết công thức lặp tổng quát cho phương pháp Jacobi dựa trên (3.4) như sau:

$$\begin{aligned}
\boldsymbol{x}^{(m+1)} &= \boldsymbol{b} - \boldsymbol{L}\boldsymbol{x}^{(m)} - \boldsymbol{U}\boldsymbol{x}^{(m)} \\
&= \boldsymbol{b} - (\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x}^{(m)} \\
&= \boldsymbol{b} - (\boldsymbol{A} - \boldsymbol{I})\boldsymbol{x}^{(m)} \\
&= \boldsymbol{b} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}^{(m)}
\end{aligned} \tag{3.5}$$

Tiếp tục chú ý, $\boldsymbol{A}$ ở đây vẫn phải ở dạng có đường chéo chính toàn 1.

Phương pháp Jacobi sẽ hội tụ với mọi $\boldsymbol{x}^{(0)}$ khi và chỉ khi bán kính phổ của $\boldsymbol{I} - \boldsymbol{A}$ nhỏ hơn 1.

Phương pháp Jacobi gần đây được chú ý nhiều vì cho phép tính toán song song. Bản chất cập nhật lập tức của phương pháp Gauss-Seidel khiến việc song song hóa rất khó khăn, còn việc cập nhật sau mỗi lần lặp khiến phương pháp Jacobi có thể được song song hóa vô cùng dễ dàng.