Лекція 9. Ітераційні та прямі методи

Системи лінійних алгебраїчних рівнянь можна розв‘язувати за допомогою як прямих, так і ітераційних методів.

Метод розв‘язання СЛАУ відносять до класу **прямих, або точних**, якщо за умови відсутності округлень він дає точний розв‘язок задачі після скінченного числа арифметичних і логічних операцій. До цих методів належать метод Гаусса і його модифікації, метод відбиття (Хаусхолдера), метод обертань (Гівенса), метод квадратного кореня і метод Холесського.

Ітераційні методи розв‘язання СЛАУ - це методи наближеного розв‘язування, що базуються на послідовному наближенні до розв‘язку шляхом багатократного застосування деякої обчислювальної процедури, при цьому вихідними даними для кожної наступної процедури є результати застосування попередніх процедур. Наслідком такого ітераційного процесу є послідовність, яка при виконанні деяких умов збігається до розв‘язку задачі.

Для систем середньої вимірності часто більш привабливими є прямі методи. Ітераційні методи застосовуються головним чином для розв‘язування складних задач великої вимірності, для яких внаслідок обмежень, що накладаються на об‘єм робочої пам‘яті і число арифметичних операцій, використання прямих методів виявляється достатньо важким або значно менше ефективним. Наприклад, ітераційні методи, як правило, застосовуються для розв‘язання задач з трьома просторовими змінними, задач, що включають системи нелінійних рівнянь, задач, що виникають при дискретизації систем рівнянь в частинних похідних, а також для

розв‘язування нестаціонарних задач з більш ніж однією просторовою змінною.

Формулювання і застосування ітераційних методів потребують спеціальних знань і деякого досвіду. Ефективному використанню ітераційних методів заважають три обставини:

1) невідомо, який метод належить застосовувати і яким чином його можна реалізувати;
2) невідомо, як обирати ітераційні параметри, що необхідні для роботи конкретних методів, наприклад, коефіцієнт релаксації для методу послідовної верхньої релаксації.
3) неясний вибір моменту закінчення ітераційного процесу.

Внаслідок великого різноманіття задач та ітераційних процедур повністю позбавитися цих невизначеностей неможливо.

Вибір ефективного ітераційного методу розв‘язування конкретної задачі залежить від її характерних властивостей та від архітектури обчислювальної машини, на якій буде розв‘язуватися задача.

З огляду на це, жодних загальних правил вибору найкращого методу розв‘язування не існує. Проте, знання порівняльних характеристик ряду ітераційних процедур загального вигляду може суттєво спростити проблему. Отже, підхід до вибору методу має грунтуватися на аналізі теоретичних та обчислювальних принципів загальних ітераційних методів. Ці принципи потім можна реально використовувати при виборі ефективного ітераційного методу.

Розглянемо клас ітераційних методів, призначених для розв‘язування лінійних систем

$$Ax = b, \qquad (1)$$

де A - задана дійсна невироджена матриця розміру $n \times n$, а b - заданий вектор-стовпець, що складається з n дійсних компонент.

Всі методи, що будуть нами розглянуті, є **лінійними стаціонарними методами першого порядка**, які можуть бути записані в такому вигляді:

$$x^{(k+1)} = Gx^{(k)} + F, \ k = 0,1,2,... \qquad (2)$$

де G - дійсна матриця переходу даного методу розміру $n \times n$, а F - відповідний відомий вектор. Такий метод є методом першого порядка, оскільки наближення $x^{(k+1)}$ залежить явно лише від $x^{(k)}$, але не залежить явно від $x^{(k-1)}, x^{(k-2)}, ..., x^{(0)}$. Метод є лінійним, оскільки ані матриця G, ані вектор F не залежать від $x^{(k)}$. Цей метод є стаціонарним, оскільки ані G, ані F не залежать від номера ітерації n. В подальшому до основних ітераційних методів ми будемо відносити будь-який ітераційний метод виду (2). До найбільш відомих основних ітераційних методів належать: метод Річардсона (RF), метод Якобі (J), метод Гаусса-Зейделя (GZ), метод послідовної верхньої релаксації (SOR) і симетричної послідовної верхньої релаксаії (SSOR).

Всюди будемо припускати, що

$$G = I - Q^{-1}A, \ F = Q^{-1}b \qquad (3)$$

для деякої невиродженої матриці Q. Така матриця Q називається **матрицею розщеплення.**

Із припущення (3) і того факту, що матриця A є невиродженою, витікає, що $\bar{x}$ є розв‘язком суміжної системи

$$(I - G)x = F \qquad (4)$$

тоді і тільки тоді, коли $\bar{x}$ є також розв‘язком системи (1), тобто

$$\bar{x} = A^{-1}b . \qquad (5)$$

Ітераційний метод (2), суміжна система (4) для якого має єдиний розв‘язок $\bar{x}$, що співпадає із розв‘язком (1), називається **цілком узгодженим**.

Якщо $\{x^{(k)}\}$ є послідовністю ітераційних наближень, що визначаються за допомогою (2), то із властивості цілковитої (повної) узгодженості витікає, що 1) якщо $x^{(k)} = \bar{x}$ для деякого k, то $x^{(k+1)} = x^{(k+1)} = \ldots = \bar{x}$ і 2) якщо послідовність $\{x^{(k)}\}$ збігається до деякого вектора $\hat{x}$, то $\hat{x} = \bar{x}$.

Завжди будемо припускати, що основний ітераційний метод (2) є цілком узгодженим, оскільки наявність цієї властивості є істотною для будь-якого розумного методу.

Іншою властивістю основних ітераційних методів, яку ми не завжди будемо заздалегідь припускати, є їхня збіжність. Кажуть, що метод (2) **збігається**, якщо для будь-якого початкового наближення $x^{(0)}$ послідовність $x^{(1)}, x^{(2)}, \ldots$, що визначається за допомогою (2), збігається до $\bar{x}$.

Необхідна і достатня умова збіжності має вигляд

$$S(G) < 1 \qquad (6)$$

Для вимірювання швидкості збіжності лінійного стаціонарного ітераційного методу (2) визначимо **вектор похибки** $\varepsilon^{(k)}$:

$$\varepsilon^{(k)} = x^{(k)} - \bar{x} . \qquad (7)$$

Використовуючи (2) і той факт, що $\bar{x}$ також задовольняє суміжну систему (2), отримуємо, що

$$\varepsilon^{(k)} = G\varepsilon^{(k-1)} = G^k\varepsilon^{(0)}. \quad (8)$$

Доведення.

$$x^{(k)} = Gx^{(k-1)} + F = Gx^{(k-1)} + (I-G)\bar{x} = G\left(x^{(k-1)} - \bar{x}\right) + \bar{x} \Rightarrow$$

$$\Rightarrow \qquad x^{(k)} - \bar{x} = G\left(x^{(k-1)} - \bar{x}\right) \qquad \Rightarrow$$

$$\varepsilon^{(k)} = G\varepsilon^{(k-1)} = \ldots = G^k\varepsilon^{(0)}.$$

Таким чином, для будь-якої векторної норми α, $\alpha = 2,\infty$ і відповідної матричної норми α, $\alpha = 2,\infty$ внаслідок властивості

$$\|Av\|_\alpha \le \|A\|_\alpha \|v\|_\alpha$$

маємо

$$\left\|\varepsilon^{(k)}\right\|_\alpha \le \left\|G^k\right\|_\alpha \left\|\varepsilon^{(0)}\right\|_\alpha \quad (9)$$

Отже, величина $\left\|G^k\right\|_\alpha$ визначає, в скільки разів було зменшено норму похибки після k ітерацій.

Визначимо **середню швидкість збіжності** методу (2) в такий спосіб:

$$R_k(G) \equiv -\frac{1}{k}\ln\left\|G^k\right\|_\alpha. \quad (10)$$

Можна показати, що якщо $S(G) < 1$, то

$$\lim_{k\to\infty}\left(\left\|G^k\right\|_\alpha\right)^{\frac{1}{k}} = S(G). \quad (11)$$

Отже, ми прийшли до визначення **асимптотичної швидкості збіжності**

$$R_\infty(G) \equiv \lim_{k\to\infty} R_k(G) = -\ln S(G). \quad (12)$$

Зауважимо, що в той час, як $R_k(G)$ залежить від конкретного виду норми α, величина $R_\infty(G)$ не залежить

від α. Часто $R_\infty(G)$ називають просто **швидкістю збіжності**.

Якщо $S(G) < 1$, то грубе наближення для кількості ітерацій, що потрібні для зменшення норми вектора початкової похибки в ζ^{-1} разів, виначається за формулою

$$k \approx -\frac{\ln \zeta}{R_\infty(G)} \tag{13}$$

Для більшості методів прискорення наявність збіжності основного методу (2) не є необхідною. Як правило, виявляється достатньо, щоб метод був **симетризованим.**

Означення 1. *Ітераційний метод (2) є симетризованим, якщо для деякої невиродженої матриці W матриця $W(I-G)W^{-1}$ виявляється симетричною і додатно визначеною.*

Така матриця W називається **матрицею симетризації**.

Ітераційний метод, що не є симетризованим, називається **несиметризованим**.

Теорема. *Якщо ітераційний метод (2) є симетризованим, то*

1) власні значення матриці G є дійсними;

2) *алгебраїчно найбільше власне число $\lambda_{\max}$ матриці G менше одиниці;*

3) *множина власних значень матриці G містить базис відповідного векторного простору.*

Численні ітераційні методи є симетризованими. Наприклад, основний метод (2) є симетризованим, якщо матриця A і матриця розщеплення Q в (3) виявляються

симетричними і додатно визначеними. В цьому випадку матриці $A^{\frac{1}{2}}$ і $Q^{\frac{1}{2}}$ є матрицями симетризації.

Для симетричної додатно визначеної матриці A квадратним коренем $A^{\frac{1}{2}}$ називається матриця V: $V^2 = A$. Оскільки симетрична матриця A завжди може бути записаною у вигляді

$$A = PDP^T,$$

де $D = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ - діагональна матриця, діагональні елементи якої дорівнюють власним числам A, а P - ортогональна матриця, стовпчики якої є власними векторами матриці A, то

$$A^{\frac{1}{2}} = PD^{\frac{1}{2}}P^T,$$

де $D^{\frac{1}{2}} = diag\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_n}\right)$ - діагональна матриця, діагональні елементи якої дорівнюють додатним квадратним кореням із власних чисел матриці A.

Крім того, матрицею симетризації є також будь-яка матриця W, така що

$$Q = W^T W.$$

Зауважимо, що із властивості симетризованості не обов‘язково випливає властивість збіжності. Якщо ітераційний метод (2) є симетризованим, то власні числа матриці G виявляються меншими за одиницю, але не обов‘язково менше одиниці за модулем.

Отже, умова збіжності не завжди виконується. Проте завжди існує так званий **екстрапольований метод**, заснований на (2), який є збіжним, коли основний метод є симетризованим.

Екстрапольований метод визначається так:

$$x^{(k+1)} = \gamma\left(Gx^{(k)} + F\right) + \left(1-\gamma\right)x^{(k)} = G_{[\gamma]}x^{(k)} + \gamma F\,, \qquad (14)$$

де

$$G_{[\gamma]} \equiv \gamma G + \left(1-\gamma\right)I \qquad (15)$$

Тут γ **-** параметр, який часто називають "коефіцієнтом екстраполяції". Якщо ітераційний метод є симетризованим, то оптимальне значення $\bar{\gamma}$ для параметра γ в сенсі мінімізації $S\left(G_{[\gamma]}\right)$ визначається за формулою

$$\bar{\gamma} = \frac{2}{2 - \lambda_{\max}(G) - \lambda_{\min}(G)}\,, \qquad (16)$$

де $\lambda_{\min}(G)$ і $\lambda_{\max}(G)$ є найменшим і найбільшим власним числом матриці G, відповідно.

Легко показати, що

$$S\left(G_{[\gamma]}\right) = \frac{\lambda_{\max} - \lambda_{\min}}{2 - \lambda_{\max} - \lambda_{\min}} = \frac{1}{\dfrac{2}{\lambda_{\max} - \lambda_{\min}} - \dfrac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}}\,.$$

Оскільки $\lambda_{\min} + \lambda_{\max} < 2$, внаслідок того факту, що обидва вони менші одиниці, маємо

$$\frac{2}{\lambda_{\max} - \lambda_{\min}} - \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} > 0\,.$$

Покажемо, що

$$\frac{2}{\lambda_{\max} - \lambda_{\min}} - \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} > 1\,.$$

Це еквівалентно тому, що

$$\frac{2}{\lambda_{\max} - \lambda_{\min}} > 1 + \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} = \frac{2\lambda_{\max}}{\lambda_{\max} - \lambda_{\min}}\,,$$

а це, в свою чергу, рівносильно умові

$$\lambda_{\max} < 1.$$

З цього випливає, що оптимальний екстрапольований метод, що визначається співвідношенням

$$x^{(k+1)} = G_{[\bar{\gamma}]} x^{(k)} + \bar{\gamma} F, \qquad (17)$$

є збіжним.

Порівняння прямих та ітераційних методів

Як правило, великі розріджені системи пов‘язані з деякою різницевою сіткою. Нехай $\frac{1}{m}$ - величина кроку сітки. Тоді у тривимірному випадку типовою є стрічкова матриця із шириною стрічки $\approx m^2 = n^{\frac{2}{3}} \Rightarrow n = m^3$. Оцінювана кількість операцій із числами з плаваючою комою (флопів) при застосуванні методу Гаусса в типовому випадку дорівнює $O\left(nm^4\right) \approx n^{2\frac{1}{3}}$. Для двовимірної ситуації ширина стрічки дорівнює $\approx n^{\frac{1}{2}}$, так що число флопів для прямого методу змінюється як n^2. Якщо потрібно розв‘язувати велику кількість систем з різними правими частинами, треба лише один раз привести матрицю до трикутного вигляду, отже кількість операцій при розв‘язанні кожної системи змінюється як $n^{\frac{5}{3}}$ для тривимірної задачі та $n^{\frac{3}{2}}$ для 2-вимірної задачі.

2. Для симетричної додатно визначеної системи зменшення похибки за одну ітерацію методу спряжених

градієнті приблизно дорівнює $\frac{\sqrt{k}-1}{\sqrt{k}+1}$, де $k = \|A\|_2 \|A^{-1}\|_2$. При дискретизації диференціальних рівнянь в частинних похідних другого порядку на сітці з кроком $\frac{1}{m}$, як правило, $k \approx m^2$. Отже, для 3-вимірної задачі ми отримуємо $k \approx n^{\frac{2}{3}}$, а для 2-вимірних задач: $k \approx n$. Для зменшення похибки до рівня ε необхідно, щоб

$$\left(\frac{1-\frac{1}{\sqrt{k}}}{1+\frac{1}{\sqrt{k}}}\right)^j \approx \left(1-\frac{2}{\sqrt{k}}\right)^j \approx e^{-\frac{2j}{\sqrt{k}}} < \varepsilon,$$

де j - номер ітерації.

Для тривимірної задачі

$$j \approx -\frac{\ln \varepsilon}{2}\sqrt{k} \approx -\frac{\ln \varepsilon}{2} n^{\frac{1}{3}},$$

а для 2-вимірної -

$$j \approx -\frac{\ln \varepsilon}{2} n^{\frac{1}{2}}.$$

Якщо ми припустимо, що кількість флопів на одну ітерацію приблизно дорівнює *fn* (де f — кількість ненульових елементів в рядку матриці), то кількість флопів,

яка потрібна для зменшення похибки до значення, меншого ε, дорівнює

1) для тривимірної задачі

$$f_\varepsilon \approx -fn^{\frac{4}{3}} \ln \varepsilon;$$

2) для двовимірної задачі

$$f_\varepsilon \approx -fn^{\frac{3}{2}}.$$

Отже, маємо висновок:

якщо ми повинні розв‘язувати одну систему, то при великому n, або маленькому f, або помірному ε ітераційні методи можуть бути більш привабливими;

якщо ми маємо розв‘язувати подібні системи із різними правими частинами і якщо ми припускаємо, що їх кількість настільки велика, що кількість операцій для декомпозиції A відносно мала, то для 2-вимірного випадку прямі методи можуть бути більш ефективними. В той же час, для 3-вимірного випадку це є сумнівним, оскільки кількість флопів для прямого розв‘язання СЛАУ є величиною порядку $\approx n^{\frac{5}{3}}$, а для ітераційного розв‘язання - $\approx n^{\frac{4}{3}}$

Приклад. Horst Simon оцінив час, який потрібен для розв‘язання системи з 5×10^9 невідомими найбільш ефективним і економічним прямим методів, відомим на даний час, як 520040 років при обчисленнях із швидкістю 1 TFLOP/сек. З іншого боку, при використанні передобумовленого методу спряжених градієнтів з тією ж швидкістю обчислень оцінюваний час дорівнює 575 сек.

Крім того, вимоги до розмірів пам‘яті при використанні ітераційних методів, як правило, менше. Часто це є аргументом для використання ітераційних методів в 2-вимірних задачах, коли кількість флопів для обох класів методів є приблизно однаковими.

Зауваження.

При відповідному передобумовленні ми можемо отримати $\sqrt{k} \approx n^{\frac{1}{6}}$, і тоді кількість флопів може стати рівною $-fn^{\frac{7}{6}} \ln \varepsilon$. Для спектральних задач найшвидшими можуть бути спеціальні методи, наприклад, багатосіткові. Вимоги до пам‘яті роблять ітераційні методи більш привабливими. Для матриць, які не є симетричними додатно визначеними, ситуація можу бути більш проблематичною: часто важко знайти відповідний і придатний ітераційний метод або передобумовлення. Проте, проекційні методи, такі як GMRES, Bi-CG, CGS і Bi-CGSTAB часто мають кількість флопів, яка близька до CG.

Ітераційні методи можуть бути привабливими навіть для щільних матриць. Якщо матриця є симетричною, додатно визначеною і число обумовленості дорівнює $n^{2-2\varepsilon}$, то кількість флопів на ітерацію дорівнює $\approx n^2$, а кількість ітераційних кроків $\approx n^{1-\varepsilon}$. Отже, загальна кількість флопів грубо оцінюється як $n^{3-\varepsilon}$ і це є асимптотично менше, ніж в методі Холесського, де $f_\varepsilon \approx n^3$.

Література

1. Хейгеман Л., Янг Д. Прикладные итерационные методы. — М.: Мир, 1986. 446 с.