Métodos iterativos

Damián Ginestar Peiró

Departamento de Matemática Aplicada

Universidad Politécnica de Valencia

Curso 2022-2023

Índice I

Introducción

Dada una matriz invertible de tamaño $n \times n$ y un vector $b \in \mathbb{R}^n$ la única solución del sistema

$$Ax = b$$

es

$$x = A^{-1}b$$

- Nosotros trabajaremos con matrices vacías (sparse) es decir matrices con un número de elementos no nulos (nnz(A)) del orden

$$nnz(A) = c \cdot n$$

con c independiente de n.

Introducción

- No se puede hacer la inversión de A ya que:
  1. A^{-1} puede dejar de ser vacía, es decir se llena, $\Longrightarrow$ no se puede almacenar.
  2. Cálculo de A^{-1} puede costar $O(n^3)$ operaciones (tiempo de CPU: años).
- Buscaremos métodos aproximados para la resolución del sistema que se basan esencialmente en el producto matriz-vector.

Conceptos básicos

- Un método iterativo obtiene una solución aproximada de $Ax = b$ construyendo una sucesión de vectores:

$$x_1, x_2, \ldots, x_k$$

  desde un vector inicial arbitrario x_0.
- Un método iterativo se dice convergente si

$$\lim_{k \to \infty} x_k = x \ .$$

- El vector error, en cada iteración, se define como

$$\boxed{e_k = x - x_k \ .}$$

Conceptos básicos

- El vector residuo, en cada iteración, se define como

$$\boxed{r_k = b - Ax_k \ .}$$

- Se puede probar que

$$\lim_{k\to\infty} x_k = x \iff \lim_{k\to\infty} \|e_k\| = 0 \iff \lim_{k\to\infty} \|r_k\| = 0$$

Conceptos básicos

- Los métodos directos teóricamente producen la solución exacta; pero en un ordenador dan errores numéricos.
- Un método iterativo nunca da la solución exacta incluso en precisión infinita.
- Hay que establecer un criterio de pararda. Se da a priori una precisión para nuestra solución. Sea TOL el error máximo permitido.

$$\|e_k\| < TOL, \text{ (error absoluto)} \qquad \text{o} \qquad \frac{\|e_k\|}{\|x\|} < TOL \text{ (error relativo)}$$

- Pero x, y e_k no son conocidos y el criterio de parada no es útil.
- Se utiliza el criterio del residuo

$$\|r_k\| < TOL \text{ (absoluto)} \quad \text{o} \quad \frac{\|r_k\|}{\|b\|} < TOL \text{ (relativo)}$$

Conceptos básicos

- La relación entre el error y el residuo es

$$r_k = b - Ax_k = Ax - Ax_k = Ae_k \ .$$

- Usando normas matriciales:

$$\|r_k\| \leq \|A\|\|e_k\| \ (1a); \qquad \|e_k\| \leq \|A^{-1}\|\|r_k\| \ (1b)$$

- Notar además

$$\|x\| \leq \|A^{-1}\|\|b\| \ (2\text{a}); \qquad \|b\| \leq \|A\|\|A^{-1}b\| = \|A\|\|x\| \ (2\text{b})$$

- Combinando (1a) con (2a) y (1b) con (2b) obtenemos

$$\frac{1}{\|A\|\|A^{-1}\|}\frac{\|r_k\|}{\|b\|} \leq \frac{\|e_k\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r_k\|}{\|b\|}$$

- Finalmente, recordando que $\kappa(A) = \|A\|\|A^{-1}\|$:

$$\frac{1}{\kappa(A)}\frac{\|r_k\|}{\|b\|} \leq \frac{\|e_k\|}{\|x\|} \leq \kappa(A)\frac{\|r_k\|}{\|b\|}$$

Conclusión: El test del residuo es fiable si $\kappa(A)$ no es muy grande.

Métodos iterativos estacionarios

Sea A la matriz del sistema $Ax = b$. Podemos considerar la partición (splitting)

$$A = M - N$$

donde $M \neq A$ es una matriz invertible.

Se construye el sistema iterativo

$$\boxed{x_{k+1} = M^{-1}Nx_k + M^{-1}b = \ Hx_k + q, \quad k = 0, 1, \ldots}$$

H es la matriz de iteración y x_0 el vector inicial. Esto es equivalente a

$$\boxed{x_{k+1} = x_k + M^{-1}\left(b - Ax_k\right) = x_k + M^{-1}r_k}$$

Definición

Se dice que un método iterativo es estacionario si la matriz de iteración H es constante en todo el proceso.

Métodos iterativos estacionarios

Sea A tal que $a_{ii} \neq 0$ y consideremos la partición

$$A = L + D + U$$

- L es la parte estrictamente triangular inferior de A,
- D es la parte diagonal de A,
- U es la parte estrictamente triangular superior de A.

1. Método de Jacobi: $M = D$ y $N = -(L + D)$

$$x_{k+1} = -D^{-1}(L + U)x_k + D^{-1}b, \quad k = 0, 1, \ldots$$

2. Método de Gauss-Seidel: $M = D + L$ y $N = -U$

$$x_{k+1} = -(D + L)^{-1}Ux_k + (D + L)^{-1}b, \quad k = 0, 1, \ldots$$

Métodos iterativos estacionarios

- Una iteración de Jacobi es muy *barata*. Sólo hay que hacer multiplicación matriz-vector, además de invertir los elementos diagonales de A. El número de multiplicaciones es del orden nz(A).

$$
\begin{aligned}
x_1^{k+1} &= \frac{1}{a_{11}} \left\{ -a_{12}x_2^k - a_{13}x_3^k - \cdots - a_{1n}x_n^k + b_1 \right\} \\
x_2^{k+1} &= \frac{1}{a_{22}} \left\{ -a_{21}x_1^k - a_{23}x_3^k - \cdots - a_{2n}x_n^k + b_2 \right\} \\
&\vdots \\
x_n^{k+1} &= \frac{1}{a_{nn}} \left\{ -a_{n1}x_1^k - a_{n3}x_3^k - \cdots - a_{n,n-1}x_{n-1}^k + b_n \right\}
\end{aligned}
$$

Métodos iterativos estacionarios

- En el método de Gauss-Seidel las componentes de x_{k+1} que ya conocemos se utilizan en la propia iteración $k+1$.
- Una iteración Gauss-Seidel es barata. Es equivalente a resolver un sistema triangular inferior

$$\boxed{(D+L)x_{k+1} = b - Ux_k.}$$

(Recordar que hay que evitar invertir matrices.)

Métodos iterativos estacionarios

Teorema

Sea A invertible. Un método iterativo estacionario converge, para cualquier vector inicial $x_0 \in \mathbb{R}^n$, a la solución exacta del sistema lineal, si y sólo si,

$$\rho(H) < 1$$

es decir, el mayor valor propio en valor absoluto de la matriz de iteración es menor que uno.

Introduciendo el error $e_k = x_k - x$. Como $Mx = Nx + b$,

$$M\left(x_{k+1} - x\right) = N\left(x_k - x\right)$$
$$e_{k+1} = M^{-1}Ne_k = \left(M^{-1}N\right)^k e_0$$

si $\rho\left(M^{-1}N\right) < 1$, entonces $\lim_{k\to\infty} \left(M^{-1}N\right)^k e_0 = 0$.

Definición

Una matriz $A = [a_{ij}]$ de tamaño $n \times n$ se dice que es estrictamente diagonal dominante por filas si

$$|a_{ii}| > \sum_{j=1,\, j\neq i}^{n} |a_{ij}|, \quad \text{para todo} \quad i = 1, 2, \ldots, n.$$

Métodos iterativos estacionarios

Teorema

Si la matriz A es estrictamente diagonal dominante entonces el método de Jacobi y el de Gauss-Seidel son convergentes.

- Se llama radio de convergencia a $R = -\log_{10}(\rho(H))$. Cuanto más pequeño sea $\rho(H)$ más rápida será la convergencia.

Métodos iterativos estacionarios

- Una generalización del método de Jacobi es el método de sobre-relajación (JOR)

$$x_i^{k+1} = \frac{w}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}} a_{ij} x_j^k \right) + (1-w) x_i^k$$

  donde se ha introducido un parámetro de relajación w.

- Este método es equivalente a la iteración

$$\boxed{x^{k+1} = x^k + w D^{-1} r^k}$$

- Se cumple que si el método de Jacobi converge, entonces el método JOR converge si $0 \leq w \leq 1$.

Métodos iterativos estacionarios

Podemos definir otra descomposición de la matriz A de la forma

$$\boxed{\omega A = (D + \omega L) - (-\omega U + (1 - \omega)D) \ ,}$$

que da lugar al método iterativo conocido como el método SOR (successive over relaxation)

$$(D + \omega L)x^{k+1} = (-\omega U + (1 - \omega)D)x^k + \omega b \ ,$$

Análogamente, se puede definir otro método SOR de la forma

$$(D + \omega U)x^{k+1} = (-\omega L + (1 - \omega)D)x^k + \omega b \ .$$

Un método SOR simétrico, SSOR, viene definido por las ecuaciones

$$\begin{aligned} (D + \omega L)x^{k+1/2} &= (-\omega U + (1 - \omega)D)x^k + \omega b \ , \\ (D + \omega U)x^{k+1} &= (-\omega L + (1 - \omega)D)x^{k+1/2} + \omega b \ . \end{aligned}$$

Métodos iterativos estacionarios

Lema de Kahan

Sea $A \in \mathbb{C}^{n \times n}$ con elementos diagonales no nulos. Entonces el método SOR diverge si $w < 0$ o $w > 2$.

Ostrowski

Si A es simétrica y definida positiva, el método SOR es convergente sí y sólo sí $0 < \omega < 2$

Si A es estrictamente diagonal dominante por filas, el método SOR converge si $0 < \omega \leq 1$.

Métodos de Richardson

Llamaremos métodos de Richardson a métodos de la forma

$$\boxed{x^{k+1} = x^k + \alpha_k P^{-1} r^k}$$

Algoritmo:

1. Cálculo de $z^k = P^{-1} r^k$
2. Cálculo del parámetro de aceleración α_k
3. Cálculo de $x^{k+1} = x^k + \alpha_k z^k$
4. Cálculo del residuo $r^{k+1} = b - Ax^{k+1} = r^k - \alpha_k A z^k$

Si $P = I$ se dice que el método está no precondicionado.
El método de Jacobi se obtiene con $\alpha = 1$ y $P = D$. El método de Gauss-Seidel se obtiene con $\alpha = 1$ y $P = D + L$

Método de Richardson

Consideremos la iteración (Richardson no precondicionado)

$$x^{k+1} = x^k + \alpha \left(b - Ax^k \right)$$

que se puede reescribir como

$$x^{k+1} = (I - \alpha A)\, x^k + \alpha b$$

La matriz de iteración es $H_\alpha = I - \alpha A$.
Si los autovalores de A son λ_i, $i = 1, \ldots, n$

$$\lambda_{\min} \leq \lambda_i \leq \lambda_{\max}$$

los autovalores de H_α, μ_i, satisfacen:

$$1 - \alpha\lambda_{\max} \leq \mu_i \leq 1 - \alpha\lambda_{\min}$$

Métodos de Richardson

- Se puede ver que si $\lambda_{\min} < 0$ y $\lambda_{\max} > 0$ el método diverge.
- Si los autovalores de A son todos positivos, se ha de cumplir

$$1 - \alpha\lambda_{\min} < 1$$
$$1 - \alpha\lambda_{\max} > -1$$

esto es

$$0 < \alpha < \frac{2}{\lambda_{\max}}$$

- El valor de α óptimo es

$$\alpha = \frac{2}{\lambda_{\min} + \lambda_{\max}}$$

Método de direcciones alternadas

Los métodos de direcciones alternadas (ADI) se introdujeron para resolver problemas elípticos

$$\frac{\partial}{\partial x}\left(a(x,y)\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(b(x,y)\frac{\partial u}{\partial y}\right)=f$$

Al discretizar el problema se llega a un sistema

$$A_1u+A_2u=b$$

donde A_1 está asociada a la discretizacion de

$$\frac{\partial}{\partial x}\left(a(x,y)\frac{\partial}{\partial x}\right)$$

y A_2 está asociada a la discretización de

$$\frac{\partial}{\partial y}\left(b(x,y)\frac{\partial y}{\partial y}\right)$$

Método de direcciones alternadas

Dado un sistema

$$Ax = b$$

donde $A = A_1 + A_2$, el método ADI resuelve el esquema

$$\begin{aligned}(I + \alpha_1 A_1)\, u^{k+\frac{1}{2}} &= (I - \alpha_1 A_2)\, u^k + \alpha_1 b \\ (I + \alpha_2 A_2)\, u^{k+1} &= (I - \alpha_2 A_1)\, u^{k+\frac{1}{2}} + \alpha_2 b\end{aligned}$$

Este esquema es convergente para $\alpha_1 = \alpha_2 = \alpha > 0$.

Métodos a bloques

Dado un sistema a bloques

$$\begin{pmatrix} A_{11} & \cdots & A_{1q} \\ \vdots & & \vdots \\ A_{q1} & \cdots & A_{qq} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_q \end{pmatrix} = \begin{pmatrix} B_1 \\ \vdots \\ B_q \end{pmatrix}$$

Método de Jacobi a bloques:

for $i = 1, \ldots q$ do

$$X_i^{k+1} = A_{ii}^{-1} \left(B_i - \sum_{\substack{j=1 \\ j \neq i}}^{q} A_{ij} X_j^k \right)$$

end for

Método de Gauss-Seidel a bloques

for $i = 1, \ldots q$ do

$$X_i^{k+1} = A_{ii}^{-1} \left(B_i - \sum_{j=1}^{i-1} A_{ij} X_j^{k+1} - \sum_{j=i+1}^{q} A_{ij} X_j^k \right)$$

end for

Ejercicio: Particularizar estas expresiones para 2×2 bloques

Precondicionadores. Introducción

- Precondicionar un sistema lineal no es otra cosa que (pre)multiplicar el sistema por una matriz nonsingular, denotada por M^{-1},
- Produce el sistema equivalente

$$M^{-1}Ax = M^{-1}b$$

- ¿Qué hay que tener en cuenta para elegir el precondicionador?
  - Condicionar mejor el sistema inicial.
  - El precondicionador M^{-1}, debe ser fácil de invertir, es decir, debe producir un sistema lineal

$$My = c$$

    fácil de resolver.

Precondicionadores. Introducción

- Se pueden definir también precondicionadores por la derecha

$$AM^{-1}y = b \ , \quad y = Mx$$

- y precondicionadores centrados

$$M_L^{-1}AM_R^{-1}y = M_L^{-1}b \ , \quad y = M_R x$$

- Los precondicionadores se pueden dividir en dos categorías: los precondiconadores algebráicos y los funcionales.
  - Los algebráicos son independientes del problema que origina el sistema de ecuaciones y se construyen por procedimientos algebraicos.
  - Los funcionales aprovechan las características especiales de los problemas donde aparecen los sistemas.

Precondicionadores. Introducción

Dado un método iterativo

$$x_{k+1} = Gx_k + f$$

puede verse como una técnica para resolver el sistema

$$(I - G)\, x = f$$

comparando con

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b$$

se tiene que $I - G = M^{-1}N$,
$G = I - M^{-1}N = M^{-1}(M - N) = M^{-1}A$.
Así el método iterativo se puede ver como una técnica para resolver el sistema precondicionado

$$M^{-1}(M - N)\, x = M^{-1}b$$
$$M^{-1}Ax = M^{-1}b$$

Precondicionadores clásicos

- $$M_J = D \ , \quad \text{Jacobi}$$
$$M_{GS} = D - L \ , \quad \text{Gauss} - \text{Seidel}$$
$$M_{SOR} = \frac{1}{\omega}(D - \omega L) \ , \quad \text{SOR}$$

- Factorización incompleta de Cholesky
  $M = \tilde{L}\tilde{L}^T$ donde $\tilde{L}$ es una aproximación del factor triangular obtenido por la factorización de Cholesky. Tenemos que resolver un sistema con la matriz M queremos que L sea lo mas vacía posible. Para ello se permite que $\tilde{L}$ tenga los elementos no cero en las posiciones donde los tiene A, esto es

$$a_{ij} = 0 \Longrightarrow l_{ij} \equiv 0, \qquad IC(0)$$

Precondicionadores clásicos

- LU incompleta

  Se construye $M = \tilde{L}\tilde{U}$ donde $\tilde{L}$ es una matriz vacía triangular inferior que aproxima a L y $\tilde{U}$ es una matriz vacía triangular superior que aproxima a U.

  Fijado un subconjunto $S \subset [1, \ldots, n] \times [1, \ldots, n]$ de posiciones de elementos en la matriz, entonces

$$a_{ij} := \begin{cases} a_{ij} - a_{ik} a_{kk}^{-1} a_{kj} & \text{si } (i,j) \in S \\ 0 & \text{si } (i,j) \notin S \end{cases}$$

  da una factorización incompleta de A que mantiene las propiedades (SPD).

Precondicionadores clásicos

- Si se hace una factorización LU con el mismo patrón de ceros que la matriz A se obtiene el precondicionador ILU(0). ILU(m), si se permite que el mismo patrón de dispersidad que A^{m+1}.
- La factorización incompleta puede fallar incluso si la matriz inicial admite factorización.
- El fallo ocurre cuando $a_{kk} = 0$. Sin embargo, en la práctica es raro que hayan fallos.

Precondicionadores polinomiales

- Estos precondicionadores son de la forma

$$M^{-1} = p(A)$$

- Un caso particular son los precondicionadores de Neuman. Se supone que la matriz A se escribe

$$A = D - C = \left(I - CD^{-1}\right) D$$

con lo que

$$A^{-1} = D^{-1}\left(I - CD^{-1}\right)^{-1} = D^{-1}\left(I + CD^{-1} + \left(CD^{-1}\right)^2 + \cdots\right)$$

Se obtienen los precondicionadores de Neuman truncando la serie. Este método funciona si $\rho(CD^{-1}) < 1$.

Precondicionadores polinomiales

Otra estrategia es partir de la ecuación escalar

$$x^{-1} - a = 0$$

si se utiliza el método de Newton, se tiene

$$x_{i+1} = x_i + \left(x_i^{-1} - a\right) x_i^2 = 2x_i - ax_i^2$$

En forma matricial

$$P^{-1} - A = 0$$

y

$$P_{i+1} = 2P_i - P_i A P_i$$

el producto

$$P_{i+1} r = \begin{cases} u = P_i r \\ v = Au \\ w = 2u - P_i v \end{cases}$$

Método de descenso rápido

- Resolver $Ax = b$, con A simétrica y definida positiva (SPD).
- Definimos la función cuadrática $\phi : \mathbb{R}^n \to \mathbb{R}$

$$\phi(y) = \frac{1}{2}(y-x)^T A(y-x) = \frac{1}{2}e^T Ae \ .$$

- Se tiene $\phi(y) \geq 0 \ \forall y \neq 0$ (definición de matriz SPD).
- Error $e = y - x$.

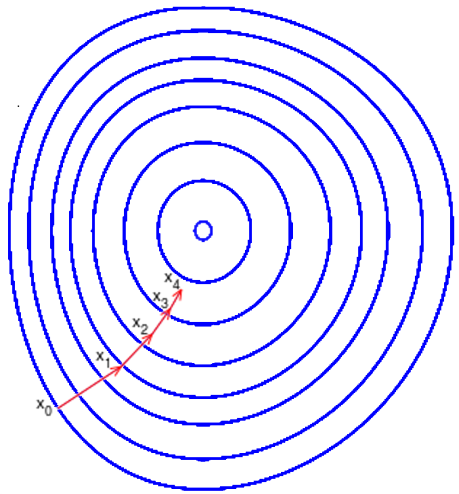
Teorema

La solución del sistema $Ax = b$ es el mínimo de la función $\phi(y)$.

Método de descenso rápido

$$\phi(y) = \frac{1}{2}(y-x)^T A(y-x) = \frac{1}{2}e^T Ae$$

- $\phi(y_k) = \text{constant}$ representa un hiperelipsoide en un espacio de dimensión n.
- El centro geométrico es la solución x del sistema lineal (mínimo).
- Construir una sucesión $\{y_k\}$ tal que $\lim_{k\to\infty} y_k = x$.



$$y_{k+1} = y_k + \alpha_k p_k$$

- Hace falta determinar la dirección p_k y α.

Método de descenso rápido

- El método de descenso rápido construye una sucesión que va hacia el centro del hiperelipsoide en la dirección del gradiente.
- El gradiente de ϕ en el punto y_k es

$$\nabla\phi(y_k) = \frac{1}{2}\nabla e_k^T A e_k = \nabla\left(\frac{1}{2}y_k^T A y_k - y_k^T b + \frac{1}{2}x^T A x\right) = Ay_k - b = -r_k$$

- Como la dirección del vector gradiente es hacia fuera, la dirección buscada coincide con el residuo r_k en la aproximación actual.
- En consecuencia la nueva aproximación es

$$y_{k+1} = y_k + \alpha_k r_k$$

donde α_k es una constante a determinar. ¿Cómo?
Minimizando $\phi(y)$ en la dirección buscada r_k.

- Es decir, nuestro α_k es la solución

$$\boxed{\alpha_k = \text{argmin}_\beta \phi(y_k + \beta r_k)}$$

Método de descenso rápido

Desarrollando la función $\phi(y_k + \beta r_k)$ se tiene un polinomio de segundo grado en la variable β.

$$\begin{aligned}
\phi(y_k + \beta r_k) &= (y_k + \beta r_k - x)^T A(y_k + \beta r_k - x) \\
&= (y_k + \beta r_k - x)^T (Ay_k + \beta Ar_k - b) \\
&= (y_k + \beta r_k - x)^T (\beta Ar_k - r_k) \\
&= (\beta r_k - e_k)^T (\beta Ar_k - r_k) \\
&= \beta^2 r_k^T Ar_k - \beta \left(r_k^T r_k + e_k^T Ar_k \right) + x^T r_k \\
&= \beta^2 r_k^T Ar_k - 2\beta r_k^T r_k + \text{const.}
\end{aligned}$$

Método de descenso rápido

Como $r_k^T A r_k > 0$ el mínimo de ϕ se alcanza cuando

$$\alpha_k \equiv \beta = \frac{r_k^T r_k}{r_k^T A r_k}$$

Otra forma: Resolver

$$\boxed{\frac{\partial \phi}{\partial \beta} = 0\,.}$$

Método de descenso rápido

- La $k+1$ iteración se puede representar como

$$\begin{aligned} r_k &= b - Ax_k \\ \alpha_k &= \frac{r_k^T r_k}{r_k^T A r_k} \\ y_{k+1} &= y_k + \alpha_k r_k \end{aligned}$$

  Notar que el coste computacional es principalmente dos productos matriz-vector.

- De $y_{k+1} = y_k + \alpha_k r_k$ se sigue que

$$r_{k+1} = b - Ax_{k+1} = b - Ax_k - A\alpha_k r_k = r_k - \alpha_k A r_k,$$

- Los residuos consecutivos r_{k+1}, r_k son ortogonales (demostración: Ejercicio).
- El error e_{k+1} es A ortogonal a la dirección r_k. (demostración: Ejercicio).

Método de descenso rápido

Algoritmo: Descenso rápido

Input: $y_0, A, b, k_{\max}$, tol

- $r_0 = b - Ay_0, \;\; k = 0$
- while $\|r_k\| >$ tol $\|b\|$ and $k < k_{\max}$ do
  1. $z = Ar_k$
  2. $\alpha_k = \dfrac{r_k^T r_k}{z^T r_k}$
  3. $y_{k+1} = y_k + \alpha_k r_k$
  4. $r_{k+1} = r_k - \alpha_k z$
  5. $k = k + 1$
- end while

Método de descenso rápido

Lema

Sea A simétrica definida positiva y sean $0 < \lambda_n \leq \cdots \leq \lambda_2 \leq \lambda_1$ sus valores propios. Si $P(t)$ es un polinomio real, entonces

$$||P(A)x||_A \leq \max_{1 \leq j \leq n} |P(\lambda_j)| \cdot ||x||_A, \quad x \in \mathbb{R}^n$$

donde $||x||_A = \sqrt{x^T A x}$.

Teorema

Sean las mismas condiciones que en el lema anterior. La sucesión $\{y_k\}$ del método de descenso rápido satisface

$$||y_k - x||_A \leq \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \right)^k ||y_0 - x||_A$$

donde x es la solución exacta del sistema.

Método de descenso rápido

Teorema

$$\sqrt{\phi(y_k)} = \sqrt{e_k^T A e_k} = \|e_k\|_A \leq \mu^k \|e_0\|_A, \quad \text{donde} \quad \mu = \frac{\kappa(A) - 1}{\kappa(A) + 1}$$

- Cuando los sistemas vienen de discretizar ecuaciones EDPs, $\kappa(A)$ puede ser muy grande.

Método de descenso rápido

- Se estima el número de iteraciones para ganar p digitos en la aproximación de la solución:

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq 10^{-p} \text{ resolviendo } \left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^k \leq 10^{-p}$$

- Tomando logaritmos y usando la aproximación de primer orden de Taylor se tiene $\log\left(\frac{\kappa(A)-1}{\kappa(A)+1}\right) \approx \frac{-2}{\kappa(A)+1}$.

  Con lo que:

$$\boxed{k \approx \frac{\log(10)}{2} p\,(\kappa(A)+1)}$$

Método del gradiente conjugado

- Es una mejora del Descenso rápido. La sucesión de recurrencia es similar

$$y_{k+1} = y_k + \alpha_k p_k$$

- Las direcciones se construyen como

$$\begin{aligned} p_0 &= r_0 \\ p_k &= r_k + \beta_k p_{k-1}, \qquad k > 0 \end{aligned}$$

- Se exige que las direcciones sean A conjugadas

$$\boxed{p_{k-1}^T A p_k = 0 \ ,}$$

  es decir, p_k y p_{k-1} son A-ortogonales.
- Por tanto, se debe cumplir:

$$\boxed{\beta_k = -\frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}}$$

Método del gradiente conjugado

- Como en el método de descenso más rápido, la elección de α_k se obtiene minimizando $\phi(y_{k+1}) = \phi(y_k + \alpha_k p_k)$ dando la expresión

$$\boxed{\alpha_k = \frac{r_k^T p_k}{p_k^T A p_k}}$$

- Residuos consecutivos como en el método de descenso más rápido satisfacen la relación de recurrencia

$$\boxed{r_{k+1} = r_k + \alpha_k A p_k}$$

Método del gradiente conjugado

Teorema

Las sucesiones de vectores $\{r_i\}$ y $\{p_i\}$ satisfacen las siguientes relaciones:

(i) $p_i^T r_j = 0, \quad 0 \leq i < j \leq k,$
(ii) $r_i^T r_j = 0, \quad i \neq j, \quad 0 \leq i, j \leq k,$
(iii) $p_i^T A p_j = 0, \quad i \neq j, \quad 0 \leq i, j \leq k,$
(iv) $\text{env}\{r_0, r_1, \ldots, r_k\} = \text{env}\{p_0, p_1, \ldots, p_k\}$

Corolario

El método del gradiente conjugado obtiene la solución del sistema de n ecuaciones como máximo en n iteraciones.

Método del gradiente conjugado

Otras relaciones útiles:

1. $p_k^T r_k = r_k^T r_k$. Ya que de $e_k^T A p_j = 0$ se sigue $r_k^T p_j = 0$ y, por tanto,

$$p_k^T r_k = (r_k + \beta_{k-1} p_{k-1})^T r_k = r_k^T r_k$$

2. $r_k^T A p_k = p_k^T A p_k$.
3. Combinando 1 y 2, se obtiene una definición alternativa de α_k:

$$\boxed{\alpha_k = \frac{r_k^T p_k}{p_k^T A p_k} = \frac{r_k^T r_k}{r_k^T A p_k}}$$

4. Formulación alternativa de β_k. Como $p_k^T A p_k = p_k^T \frac{1}{\alpha_k}(r_k - r_{k+1}) = \frac{1}{\alpha_k} r_k^T r_k$

$$r_{k+1}^T A p_k = r_{k+1}^T \frac{1}{\alpha_k}(r_k - r_{k+1}) = -\frac{1}{\alpha_k} r_{k+1}^T r_{k+1}$$

Por tanto,

$$\boxed{\beta_k = -\frac{r_{k+1}^T p_k}{p_k^T A p_k} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}}$$

Método del gradiente conjugado

Algoritmo: Gradiente conjugado

Input: $y_0, A, b, k_{\max}$, tol

- $r_0 = p_0 = b - Ax_0, \;\; k = 0$
- while $\|r_k\| >$ tol $\|b\|$ and $k < k_{\max}$ do
  1. $z = Ap_k$
  2. $\alpha_k = \dfrac{p_k^T r_k}{z^T p_k}$
  3. $y_{k+1} = y_k + \alpha_k p_k$
  4. $r_{k+1} = r_k - \alpha_k z$
  5. $\beta_k = \dfrac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$
  6. $p_{k+1} = r_{k+1} + \beta_k p_k$
  7. $k = k + 1$
- end while

Método del gradiente conjugado

Ejercicio

Aplicar el algoritmo del gradiente conjugado para el problema

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

usando como aproximación inicial $x_0 = (0, 0)^T$.

Método del gradiente conjugado

Solución:

- $x_0 = (0,0)^T$.
- $p_0 = r_0 = b = (1,0)^T$.
- $\alpha_0 = \frac{r_0^T r_0}{p_0^T A p_0} = \frac{1}{2}$, $x_1 = x_0 + \alpha_0 p_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}$
- $r_1 = r_0 - \alpha_0 A p_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}$, $r_1^T r_0 = 0$
- $\beta_0 = \frac{r_1^T r_1}{r_0^T r_0} = \frac{1}{4}$, $p_1 = r_1 + \beta_0 p_0 = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \end{pmatrix}$
- $\alpha_1 = \frac{r_1^T r_1}{p_1^T A P_1} = \frac{2}{3}$

  $x_2 = x_1 + \alpha_1 p_1 = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} + \frac{2}{3}\begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ \frac{1}{3} \end{pmatrix}$
- $r_2 = 0$ solución exacta

Método del gradiente conjugado

Teorema

Sea $A \in \mathbb{R}^{n \times n}$ simétrica y definida positiva. Sea x la solución exacta del sistema $Ax = b$. Entonces la sucesión de vectores del Gradiente Conjugado $\{y_k\}$ cumple

$$||x - y_k||_A \leq 2 \left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^k ||x - y_0||_A$$

donde $\kappa_2(A) = ||A||_2 ||A^{-1}||_2$.

Métodos de Krylov

Si consideramos el método de Richardson no precondicionado para el residuo, se tiene la sucesión

$$r^k = \prod_{j=0}^{k-1} (I - \alpha_j A)\, r^0$$

así $r^k = p_k(A) r^0$, donde $p_k(A)$ es un polinomio de grado k.
Definiendo subespacio de Krylov de orden m

$$K_m(A; v) = \text{span}\left\{v, Av, \ldots, A^{m-1} v\right\}$$

se tiene que $r^k \in K_{k+1}(A; r^0)$.

Métodos de Krylov

Si consideramos

$$x^k = x^0 + \sum_{j=0}^{k-1} \alpha_j r^j$$

x^k pertenece al espacio

$$W_k = \left\{ v = x^0 + y; y \in K_k(A, r^0) \right\}$$

Se pueden plantear métodos

$$\boxed{x^k = x^0 + q_{k-1}(A) r^0}$$

donde el polinomio $q_{k-1}(A)$ se selecciona de forma que x^k sea la mejor aproximación de x en W_k.

Métodos de Krylov

Por ejemplo, como ya hemos visto, el método de descenso más rápido se basa en iteraciones de la forma

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k r_k \\ r_{k+1} &= b - Ax_{k+1} = r_k - \alpha_k A r_k \ . \end{aligned}$$

Para este método $x_k \in W_k$

Métodos de Krylov

Los métodos de proyección de un paso se basan en una combinación óptima de los dos últimos vectores base del subespacio de Krylov.
¿Es posible contruir una combinación lineal óptima de todos los vectores base del subespacio de Krylov?

La respuesta en dos pasos:

1. Primero vemos cómo construir una base para $K_k(A; r_0)$;
2. Después veremos cómo construir una aproximación óptima como una combinación lineal de los vectores base. Primero para matrices simétricas.

Métodos de Krylov

La base más simple para el subespacio de Krylov $K_k(A; r_0)$ es la base: $r_0, Ar_0, A^2r_0, \cdots A^{k-1}r_0$.

Pero esta base está mal condicionada ya que $A^{k-1}r_0$ apuntará cada vez más en la dirección del autovector dominante de A.

Una base estable y ortogonal se puede construir usando el método de Arnoldi.

Métodos de Krylov

Se elige un vector inicial q_1 con $\|q_1\|_2 = 1$.

for $k = 1, \cdots$ **do** (iteración)

$v = Aq_k$

for $i = 1, ..., k$ (ortogonalización)

$h_{i,k} = v^T q_i$

$v = v - h_{i,k} q_i$

end for

$h_{k+1,k} = \|v\|_2$

if $h_{k+1,k} = 0$ **stop** (subespacio invariante)

$q_{k+1} = v / h_{k+1,k}$ (nuevo vector)

end for

El método de Arnoldi se puede resumir en:

$$H_k = \begin{pmatrix} h_{1,1} & \cdots & \cdots & h_{1,k} \\ h_{2,1} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ O & & h_{k,k-1} & h_{k,k} \end{pmatrix}$$

y $Q_k = [q_1 \ q_2 \cdots q_k]$ entonces

$$AQ_k = Q_k H_k + h_{k+1,k} q_{k+1} e_k^T$$

donde e_k es el $k-$ésimo vector de la base canónica de $\mathbb{R}^k$.

Si A es simétrica de acuerdo a la relación de Arnoldi

$$Q_k^T A Q_k = H_k \ .$$

Como A es simétrica

$$H_k^T = Q_k^T A^T Q_k = Q_k^T A Q_k = H_k \ .$$

Así H_k es simétrica y Hessenberg superior luego H_k es tridiagonal.

Así

$$H_k = \begin{pmatrix} h_{1,1} & h_{2,1} & & O \\ h_{2,1} & \ddots & \ddots & \\ & \ddots & \ddots & h_{k,k-1} \\ O & & h_{k,k-1} & h_{k,k} \end{pmatrix}.$$

Con $\alpha_k = h_{k,k}$ y $\beta_k = h_{k-1,k}$ el método de Arnoldi se simplifica en el método de Lanczos.

Con el método de Lanczos es posible calcular un nuevo vector base ortogonal utilizando sólo los dos vectores base previos.

Métodos de Krylov

El método de Arnoldi y el método de Lanczos se propusieron originalmente como métodos iterativos para calcular autovalores de la matriz A:

$$Q_k^T A Q_k = H_k$$

es 'casi' una transformación de similaridad. Los autovalores de H_k se llaman los valores de Ritz de A.

Métodos de Krylov

El método de Lanczos nos proporciona un método económico para calcular vectores base ortogonales para el subespacio de Krylov $K_k(A; r_0)$.
Nuestra aproximación se escribe

$$x_k = x_0 + Q_k y_k$$

donde y_k se determina de forma que:

- o bien se minimiza

$$f(x_k) = \|x_k - x\|_A^2 = (x_k - x)^T A(x_k - x)$$

  respecto de la norma inducida por A (simétrica y definida positiva)
- o bien se minimiza la norma del residuo

$$g(x_k) = \|A(x_k - x)\|_2^2 = r_k^T r_k.$$

Métodos de Krylov

Se considera la minimización del error en la A-norma:

$$x_k = x_0 + Q_k y_k \Rightarrow f(x_k) = (x_0 + Q_k y_k - x)^T A(x_0 + Q_k y_k - x) \ .$$

Se calcula la derivada respecto de y_k, $\frac{\partial f(x_k)}{\partial y_k} = 0$.
Con lo que

$$Q_k^T A Q_k y_k = Q_k^T r_0$$

y con $Q_k^T A Q_k = T_k, r_0 = \|r_0\| q_1$ se tiene

$$T_k y_k = \|r_0\| e_1$$

con e_1 el primer vector de la base canónica.

Métodos de Krylov

Es fácil ver que los residuos son orgonales a los vectores base

$$r_k = r_0 - AQ_k y_k \Rightarrow Q_k^T r_k = Q_k^T r_0 - Q_k^T A Q_k y_k = 0$$

Esta condición es equivalente a minimizar $f(x_k)$ cuando A es SPD.

En este caso se obtiene el método del gradiente conjugado.
El gradiente conjugado minimiza la A-norma del error.

Método del residuo conjugado

Otra forma de construir una aproximación óptima x_k es minimizar el residuo

$$g(x_k) = \|A(x_k - x)\|_2^2 = r_k^T r_k$$

sobre todos los $x_k \in \{x_0 \cup K_k(A; b)\}$.
Definiendo

$$\underline{T}_k = \begin{pmatrix} \alpha_1 & \beta_2 & & & 0 \\ \beta_2 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \ddots & \beta_k \\ & & & \beta_k & \alpha_k \\ & & & & \beta_{k+1} \end{pmatrix}$$

El método de Lanczos se escribe

$$AQ_k = Q_{k+1}\underline{T}_k$$

Método del residuo conjugado

El problema es encontrar $x_k = x_0 + Q_k y_k$ tal que $\|r_k\|$ es mínimo.

$$r_k = b - Ax_k = r_0 - AQ_k y_k = \|r_0\| q_1 - AQ_k y_k$$

así hay que minimizar

$$\begin{aligned} \|r_k\| &= \|\|r_0\| q_1 - AQ_k y_k\| \\ &= \|\|r_0\| Q_{k+1} e_1 - Q_{k+1} \underline{T}_k y_k\| \\ &= \|\|r_0\| e_1 - \underline{T}_k y_k\| \end{aligned}$$

Método del residuo conjugado

Resolviendo el sistema sobredeterminado $\underline{T}_k y_k = \|r_0\| e_1$ se tienen las iteraciones

$$x_k = x_0 + Q_k y_k$$

que minimizan el residuo.

El algoritmo resultante se llama MINRES (o residuo conjugado)

Método del residuo conjugado

MINRES

$r_0 = b - Ax_0;$ $\quad p_0 = r_0$ initializacion

for $k = 0, 1, \cdots,$ **do**

$$\alpha_k = \frac{r_k^T A r_k}{(Ap_k)^T Ap_k}$$

$$x_{k+1} = x_k + \alpha_k p_k$$

$$r_{k+1} = r_k - \alpha_k A p_k$$ actualizacion residuo

$$\beta_k = \frac{r_{k+1}^T A r_{k+1}}{r_k^T A r_k}$$

$$p_{k+1} = r_{k+1} + \beta_k p_k$$ actualizacion direccion

$$Ap_{k+1} = Ar_{k+1} + \beta_k Ap_k$$

end for

Método FOM

- En el caso de A simétrica el método de Lanczos es muy eficiente, los vectores base nuevos se pueden calcular con una recurrencia de tres términos.
- Esto permite construir métodos muy eficientes que combinan recurrencias cortas y una condición de optimalidad para el error.
- Veremos ahora cómo se pueden construir métodos para el caso no simétrico.
- Estos métodos usan el método de Arnoldi.

Método FOM

Recordemos el Método de Arnoldi.

Se elige un vector inicial q_1 con $\|q_1\|_2 = 1$.

for $k = 1, \cdots$ **do** (iteración)

$v = Aq_k$

for $i = 1, ..., k$ (ortogonalización)

$h_{i,k} = v^T q_i$

$v = v - h_{i,k} q_i$

end for

$h_{k+1,k} = \|v\|_2$

if $h_{k+1,k} = 0$ **stop** (subespacio invariante)

$q_{k+1} = v/h_{k+1,k}$ (nuevo vector base)

end for

Método FOM

En forma compacta

$$H_k = \begin{pmatrix} h_{1,1} & \cdots & \cdots & h_{1,k} \\ h_{2,1} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ O & & h_{k,k-1} & h_{k,k} \end{pmatrix}$$

y $Q_k = [q_1 \ q_2 \cdots q_k]$ entonces

$$AQ_k = Q_k H_k + h_{k+1,k} q_{k+1} e_k^T$$

Método FOM

Definiendo

$$\underline{H}_k = \begin{pmatrix} h_{1,1} & \cdots & \cdots & h_{1,k} \\ h_{2,1} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & h_{k,k-1} & h_{k,k} \\ O & & & h_{k+1,k} \end{pmatrix}$$

el método de Arnoldi se escribe

$$\boxed{AQ_k = Q_{k+1}\underline{H}_k}$$

Método FOM

El método de Anoldi nos da una base ortogonal para el subespacio de Krylov $K_k(A; r_0)$.
Se tiene la aproximación

$$x_k = x_0 + Q_k y_k$$

y_k es tal que:

- minimiza el error

$$f(x_k) = \|x_k - x\|_A^2 = (x_k - x)^T A(x_k - x)$$

  respecto de la A-norma,
- o bien minimiza el residuo,

$$g(x_k) = \|A(x_k - x)\|_2^2 = r_k^T r_k.$$

Método FOM

- Si A no es SPD, la A-norma no está bien definida.
- Se impone que los residuos sean ortogonales a Q_k. Se tiene

$$Q_k^T(r_0 - AQ_k y_k) = 0 \Rightarrow \|r_0\|e_1 - H_k y_k = 0$$

Resolviendo el sistema pequeño

$$y_k = \|r_0\|H_k^{-1}e_1, \quad x_k = x_0 + Q_k y_k$$

Este método se llama FOM (Full orthogonalization method).

Método FOM

El FOM es equivalente al GC si A es SPD. Pero tiene inconvenientes:

- FOM no trata de forma eficiente la memoria: Q_k se tiene que guardar completa. En cada iteración, un nuevo vector base se tiene que calcular y guardar. La ortogonalización va siendo cada vez más cara con k.
- FOM no tiene una propiedad de optimalidad.
- El método es finito, pero esto sólo tiene interés teórico ya que se tendrían que calcular n vectores base y guardarlos.
- FOM no es robusto ya que H_k podría ser singular.

Método FOM

- El método FOM para obtener x_k es parte de una familia de técnicas para extraer una solución aproximada a partir de un espacio de búsqueda Q_k haciendo que el residuo sea ortogonal a un espacio test W_k.
- Esto se formula como:

  Sea $x_k = x_0 + Q_k y_k$. Encontrar y_k tal que

  $$W_k^T (r_0 - A Q_k y_k) = 0$$

- Estas son las condiciones de Petrov-Galerkin.
  Si $W_k = Q_k$ se llaman condiciones de Galerkin.

Método GMRES

Veamos un segundo método para obtener un mínimo del residuo.

Minimizar

$$g(x_k) = \|A(x_k - x)\|_2^2 = r_k^T r_k$$

es un problema bien definido incluso si A es no simétrica.

Método GMRES

El problema es encontrar $x_k = x_0 + Q_k y_k$ tal que $\|r_k\|$ es mínimo.

$$r_k = b - Ax_k = r_0 - AQ_k y_k = \|r_0\| q_1 - AQ_k y_k$$

así

$$\begin{aligned} \|r_k\| &= \|\|r_0\| q_1 - AQ_k y_k\| \\ &= \|\|r_0\| Q_{k+1} e_1 - Q_{k+1} \underline{H}_k y_k\| \\ &= \|\|r_0\| e_1 - \underline{H}_k y_k\| \end{aligned}$$

Método GMRES

Resolviendo el problema sobredeterminado

$$\underline{H}_k y_k = \|r_0\| e_1$$

se tienen las iteraciones

$$x_k = x_0 + Q_k y_k$$

que minimizan el residuo.

El algoritmo resultante se llama GMRES.
GMRES, es uno de los métodos más populares para resolver sistemas no simétricos

Método GMRES

El método GMRES es equivalente al MINRES si A es simétrica.
Otras características son:

- GMRES no tiene limitada la memoria: Q_k se ha de guardar completamente. En cada iteración se ha de calcular y guardar un nuevo vector de la base. La ortogonalización de un nuevo vector se hace cada vez más cara al aumentar k.
- GMRES minimiza la norma del residuo.
- El método es finito, pero esto sólo tiene interés teórico. Para llegar a este punto, se han de calcular n vectores de la base y se han de guardar.
- GMRES es robusto: $\underline{H}_k y_k = \|r_0\| e_1$ tiene siempre una solución de mínimos cuadrados.

Método GMRES

- Los métodos de Arnoldi minimzan el residuo, calculando un nuevo vector ortogonal de la base en cada iteración. El coste de hacer esto se hace prohibitivo si se hacen muchas iteraciones.
- Se suele fijar un número máximo de m pasos consecutivos. Si el método no ha convergido se hace $x^0 = x^m$ y se reinicia el método. (se obtiene GMRES(m), por ejemplo)
- Hay otro tipo de métodos que se basan en el método de bi-Lanczos. Usan relaciones de recurrencia cortas, pero las iteraciones no son óptimas. (Bi-Lanczos, Bi-CG, Bi-CGSTAB)