Numerical Methods I: Iterative solvers for $A\boldsymbol{x} = \boldsymbol{b}$

Georg Stadler
Courant Institute, NYU
stadler@cims.nyu.edu

December 7, 2017

Iterative solution of (symmetric) linear systems

Target problems: very large ($n = 10^5, 10^6, \ldots$), A is usually sparse and has specific properties.
To solve

$$A\boldsymbol{x} = \boldsymbol{b}$$

we construct a sequence

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$$

of iterates that converges as fast as possible to the solution $\boldsymbol{x}$, where $\boldsymbol{x}_{k+1}$ can be computed from $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ with as little cost as possible (e.g., one matrix-vector multiplication).

Iterative solution of (symmetric) linear systems

Let Q be invertible, then

$Q \in \mathbb{R}^{n \times n}$, invertible

$$\begin{aligned} A\boldsymbol{x} = \boldsymbol{b} &\Leftrightarrow Q^{-1}(\boldsymbol{b} - A\boldsymbol{x}) = 0 \\ &\Leftrightarrow (I - Q^{-1}A)\boldsymbol{x} + Q^{-1}\boldsymbol{b} = \boldsymbol{x} \\ &\Leftrightarrow G\boldsymbol{x} + \boldsymbol{c} = \boldsymbol{x} \end{aligned}$$

Fixed point method: $\boxed{x_{k+1} = G x_k + c,}$

$k = 0, 1, 2, \ldots$

$x_0 \in \mathbb{R}^n$ initialization

Iterative solution of (symmetric) linear systems

Theorem: The fixed point method $\boldsymbol{x}_{k+1} = G\boldsymbol{x}_k + \boldsymbol{c}$ with an invertible G converges for each starting point $\boldsymbol{x}_o$ if and only if

$$\rho(G) < 1,$$

where $\rho(G)$ is the largest eigenvalue of G (i.e., the spectral radius).

Proof: G spd. $\exists\ Q$ orthogonal s.t. $\boxed{x^* = A^{-1}b}$

$$QGQ^T = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

Since $|\lambda_i| \le \rho(G) < 1 \Longrightarrow \Lambda^k \longrightarrow 0$ as $k \to \infty$

$$\Longrightarrow (Q\Lambda Q^T)^k = G^k = Q\Lambda^k Q^T \longrightarrow 0 \text{ as } k \to \infty$$

Since $\rho(G) \le \|G\|$ for any matrix norm induced by vector norm.

$$x_k - x^* = Gx_{k-1} + c - (Gx^* + c) = G(x_{k-1} - x^*) = \ldots G^k(x_0 - x^*)$$

$$\longrightarrow \|x_k - x^*\| \le \|G^k\| \, \|x_0 - x^*\| \longrightarrow 0. \text{ if } \|G\| < 1.$$

Iterative solution of (symmetric) linear systems

Choices for Q:

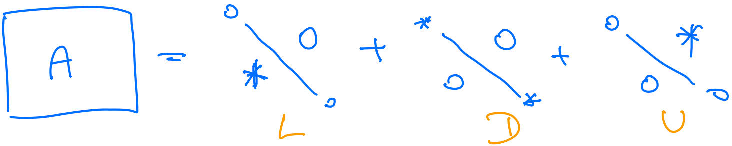
- Choose $Q = I \ldots$ Richardson method

$$G = I - Q^{-1}A = I - A, \qquad x_{k+1} = x_k - Ax_k + b$$

$$A \text{ spd}, \quad \rho(G) = \rho(I - A) = \max\{|1-\lambda_{min}(A)|, |1-\lambda_{max}(A)|\}$$

$$\Longrightarrow \rho(G) < 1 \quad \text{if} \quad \lambda_{max}(A) < 2$$

For more choices, consider $A = L + D + U$, where D is diagonal, L and U are lower and upper triangular with zero diagonal.



Iterative solution of (symmetric) linear systems

$A = L + D + U$

- Choose $Q = D$ ... Jacobi method

$G = I - Q^{-1}A = I - D^{-1}A = I - D^{-1}(L+D+U) =$

$= -D^{-1}(L+U)$

$$\boxed{x_{k+1} = -D^{-1}(L+U)x_k + D^{-1}b}$$

Theorem: The Jacobi method converges for any starting point $\boldsymbol{x}_o$ to the solution of $A\boldsymbol{x} = \boldsymbol{b}$ if A is strictly diagonal dominant, i.e.,

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \qquad \text{for } i = 1, \dots, n.$$

Proof: $\rho(G) = \rho(-D^{-1}(L+R)) \leq \|D^{-1}(L+R)\|_\infty =$

$= \max_i \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1$ since A is strictly diagonal dom.

$A = L + D + U$

- Choose $Q = D + L$ ... Gauss-Seidel method

$G = I - Q^{-1}A = I - (D+L)^{-1}A = I - (D+L)^{-1}(L+D+U)$

$= -(D+L)^{-1}U$

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

Theorem: The Gauss-Seidel method converges for any starting point $\boldsymbol{x}_o$ if A is spd.

Iterative solution of (symmetric) linear systems

Relaxation methods: Use linear combination between new and previous iterate:

$G_\omega = \omega G + (1-\omega) I$

$$\boldsymbol{x}_{k+1} = \omega(G\boldsymbol{x}_k + c) + (1-\omega)\boldsymbol{x}_k = G_\omega \boldsymbol{x}_k + \omega c,$$

where $\omega \in [0,1]$ is a damping/relaxation parameter (sometimes, $\omega > 1$ is used, leading to overrelaxation). Target is to choose ω such that $\rho(G_\omega)$ is as small as possible.

Def: A fixed point method $\boldsymbol{x}_{k+1} = G\boldsymbol{x}_k + \boldsymbol{c}$ with $G = G(A)$ is called *symmetrizable* if for any spd matrix A, $I - G$ is similar to an spd matrix. That is, $\exists\ W \in \mathbb{R}^{n\times n}$ invertible such that

$W(I-G)W^{-1}$ is spd.

Examples: Richardson $\quad G = I - A$, so $I - G =$

$= A$ is spd

$[W = I]$

Jacobi: $\quad G = I - D^{-1}A$, $W := D^{-\frac{1}{2}}$

$D^{-\frac{1}{2}}(I-G)D^{\frac{1}{2}} = D^{-\frac{1}{2}}(\cancel{I} - \cancel{I} + D^{-1}A)D^{\frac{1}{2}} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ spd

if A spd.

Iterative solution of (symmetric) linear systems

Let the fixed point method be symmetrizable, and A an spd matrix. Then all eigenvalues of G are real and less than 1.

Proof: method is symmetrizable $\Longrightarrow$ $I-G$ is similar to an spd matrix $\Longrightarrow$ eigenvalues of $I-G$ are real and positive $\Longrightarrow$ eigenvalues of G are real and <1.

Iterative solution of (symmetric) linear systems

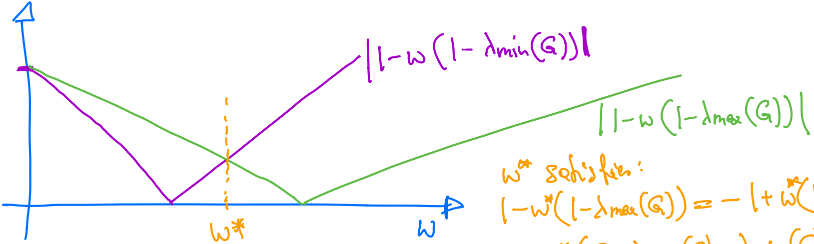
Finding the optimal damping parameter:

symmetrizable method, $\lambda_{min} \le \lambda_{max} < 1$ extreme eigenvalues of G.

Eigenvalues of $G_\omega = \omega G + (1-\omega) I$:

$$\lambda_i(G_\omega) = \omega \lambda_i(G) + (1-\omega) = 1 - \omega(1-\lambda_i(G)) < 1$$

$$\rho(G_\omega) = \max\{|1-\omega(1-\lambda_{min}(G))|, |1-\omega(1-\lambda_{max}(G))|\}$$



ω^* satisfies:

$$1-\omega^*(1-\lambda_{max}(G)) = -1 + \omega^*(1-\lambda_{min}(G)).$$

$$\Longrightarrow 2 = \omega^*(2-\lambda_{max}(G)-\lambda_{min}(G))$$

Iterative solution of (symmetric) linear systems

Krylov methods:

Idea: Build a basis for the Krylov subspace $\{\boldsymbol{r}_0, A\boldsymbol{r}_0, A^2\boldsymbol{r}_0 \ldots\}$ and reduce residual optimally in that space.

- spd matrices: Conjugate gradient (CG) method
- symmetric matrices: Minimal residual method (MINRES)
- general matrices: Generalized residual method (GMRES), BiCG, BiCGSTAB

Iterative solution of (symmetric) linear systems

Krylov methods:

Idea: Build a basis for the Krylov subspace $\{\boldsymbol{r}_0, A\boldsymbol{r}_0, A^2\boldsymbol{r}_0 \ldots\}$ and reduce residual optimally in that space.

- spd matrices: Conjugate gradient (CG) method
- symmetric matrices: Minimal residual method (MINRES)
- general matrices: Generalized residual method (GMRES), BiCG, BiCGSTAB

Properties:

Do not require eigenvalue estimates; require usually one matrix-vector multiplication per iteration; convergence depends on eigenvalue structure of matrix (clustering of eigenvalues aids convergence). Availability of a good preconditioner is often important. Some methods require storage of iteration vectors.

The conjugate gradient method (CG, pcg in Matlab)

Solve $Ax=b$, A spd, $x, b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ unique solution

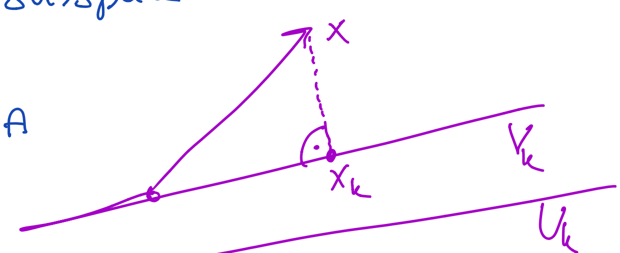
Solving $Ax = b \iff \min_{x \in \mathbb{R}^n} \frac{1}{2} x^T A x - b^T x$

A-weighted norm

$$\|y\|_A = \sqrt{(y, Ay)} \quad \text{is a norm}$$

$x = A^{-1} b$, Search approximations $x_k \in \mathbb{R}^n$ of x in affine subspace $V_k = x_0 + U_k$, $U_k \subset \mathbb{R}^n$ subspace

$$x_k = \arg\min_{y \in V_k} \|y - x\|_A$$



$$\Updownarrow$$

$$(x - x_k, u)_A := (A(x - x_k), u) = 0 \quad \text{for all } u \in U_k$$

Why choose A-weighted norm?

$r_k = b - Ax_k$ residual

$$(x - x_k, u)_A = (A(x - x_k), u) = (r_k, u)$$

does not include x, which is not known.

i.e: residual $\perp$ u in Euclidean inner product

$\Leftrightarrow$ error $x - x_k \perp$ to u in A-weighted inner product.

Let $p_1, \dots, p_k$ A-orthogonal basis in U_k, i.e.

$$(p_i, p_j)_A = \delta_{ij}$$

A-orthogonal also called A-conjugate

$$x_k = \underbrace{P_k}_{\text{orthog. proj on } V_k} x = x_0 + \sum_{j=1}^{k} \frac{(p_j, x - x_0)_A}{(p_j, p_j)_A} p_j =$$

$$= x_0 + \sum_{j=1}^{k} \frac{(p_j, \overbrace{A(x - x_0)}^{r_0})}{(p_j, p_j)_A} p_j$$

$$= x_0 + \sum_{j=1}^{k} \underbrace{\frac{(p_j, r_0)}{(p_j, p_j)_A}}_{\alpha_k} p_j \leftarrow \text{indep. of } x$$

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

$$\boxed{r_k = A(x - x_k) = A(x - x_{k-1} - \alpha p_k) = r_{k-1} - A \alpha_k p_k}$$

Specific ~~...~~ choices: $V_k = x_0 + U_k$

$$U_k = \text{span}\{r_0, A r_0, A^2 r_0, \dots A^{k-1} r_0\}$$

$\boxed{\text{Krylov Spaces}}$

$$\rightarrow \boxed{p_{k+1} = r_k - \underbrace{\frac{(r_k, p_k)}{(p_k, p_k)_A}}_{\beta_{k+1}} p_k}$$

<u>We have:</u> ∘ Method to choose V_k, and compute A-orthog. basis

∘ method to compute x_k that does not need x

∘ recurrence for r_k

<u>Algorithm:</u> A spd, x_0 starting value

$$p_1 = r_0 = b - Ax_0$$

for $k = 1, 2, \ldots$

$$\alpha_k = \frac{(r_{k-1}, r_{k-1})}{(p_k, Ap_k)}$$

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

if converged stop

$$r_k := r_{k-1} - \alpha_k A p_k$$

$$\beta_{k+1} := \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})}$$

$$p_{k+1} := r_k + \beta_k p_k$$

end

Cost:

1 application of A

2 inner products

∘ Converges after at most n iterations because

$$x_k = \arg\min_{y \in V_k} \|x - y\|_A, \quad V_k = \mathbb{R}^n \text{ after } n \text{ iterations}$$

- Converges much faster depending on the eigenvalues of A, in particular if eigenvalues of A are clustered, i.e.