Richardson method

$$Q = I$$

$$x^{(k)} = (I - A)x^{(k-1)} + b = x^{(k-1)} + \underbrace{b - Ax^{(k-1)}}_{r^{(k-1)}}$$

converges when $\|I - A\| < 1$ in some induced matrix norm.

Jacobi method

$$Q = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix} = \text{diag}(\text{diag}(A)) = \text{diagonal matrix with same entries as diagonal of } A.$$

$$Q^{-1}A = \begin{bmatrix} 1 & a_{12}/a_{11} & a_{13}/a_{11} & \cdots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 1 & a_{23}/a_{22} & \cdots & a_{2n}/a_{22} \\ \vdots & & & & \\ & & & & a_{n-1,n}/a_{n-1,n-1} \\ a_{n1}/a_{nn} & \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

Thus:

$$\|I - Q^{-1}A\|_\infty = \max_{1 \le i \le n} \sum_{\substack{j=1 \\ j \ne i}}^{n} \frac{|a_{ij}|}{|a_{ii}|}.$$

Theorem If A is diagonally dominant, then the Jacobi method's iterates converge to a solution to $Ax = b$.

Proof.

A diag dominant $\Rightarrow$ $|a_{ii}| \ge \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}|, \quad 1 \le i \le n$

$\Rightarrow \|I - Q^{-1}A\|_\infty < 1$

$\Rightarrow$ Convergence.

Linear algebra reviews

Eigenvalues of A = roots of characteristic polyn. of A

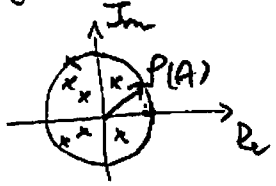
The characteristic poly. of A is defined by:

$$p(\lambda) = \det(\lambda I - A)$$

Spectral radius:

$$\rho(A) = \max\{ |\lambda| \mid \det(\lambda I - A) = 0 \}$$

= radius of smallest circle in $\mathbb{C}$ containing all eigenvalues of A.



Two matrices A and B are similar if there is an invertible matrix X s.t.

$$AX = XB.$$

note: Two similar matrices have same eigenvalues:

$$p_A(\lambda) = \det(\lambda I - A) = \det(X^{-1}(\lambda I - A)X)$$
$$= \det(\lambda I - \underbrace{X^{-1}AX}_{B}) = p_B(\lambda)$$

Here we used the facts:

$$\det(AB) = \det(A)\det(B)$$
$$\det(A^{-1}) = \det(A)^{-1}.$$

Before we continue studying iterative methods for solving linear systems we need two technical results.

Theorem Every matrix is similar to an upper triangular matrix with arbitrarily small off-diagonal elements.

proof: Use Schur factorization

$$A = Q T Q^T, \text{ where } Q^T Q = I$$

this is an eigenvalue revealing factorization as the eigenvalues of A appear on diagonal of T. we will see more about this factorization when we study eigenvalue algorithms.

Let $D = \begin{bmatrix} \varepsilon & & & \\ & \varepsilon^2 & & \\ & & \ddots & \\ & & & \varepsilon^n \end{bmatrix}$

Then

$$(D^{-1} T D)_{ij} = t_{ij}\, \varepsilon^{j-i}$$

But $T = (\nabla)$ thus:

$$t_{ij} = 0 \text{ when } i > j \quad \text{(elements below diag are zero)}$$

Now for elements above diag: $(i < j)$:

$$\left| t_{ij}\, \varepsilon^{\overset{\geq 1}{\overbrace{j-i}}} \right| \leq \varepsilon \left| t_{ij} \right|$$

Thus

$$A = \underbrace{Q D}_{X} \underbrace{D^{-1} T D}_{B} \underbrace{D^{-1} Q^T}_{X^{-1}} . \qquad \text{QED}$$

The second technical result is:

Theorem

$$\rho(A) = \inf_{\|\cdot\|} \|A\|, \text{ where inf is over all induced matrix norms.}$$

This means $\rho(A) = \|A\|_2$ = "smallest" induced norm one can choose to measure A.

proof:

- $\rho(A) \leq \inf_{\|\cdot\|} \|A\|$: Let $\|\cdot\|$ be a vector norm and let λ, x be an eigenvalue/vector of A, then:

$$\|A\|\|x\| \geq \|Ax\| = \|\lambda x\| = |\lambda|\|x\|$$

$$\Rightarrow \quad |\lambda| \leq \|A\| \quad \text{(since eigenvector } x \neq 0\text{)}$$

Since this is true for any eigenvalue λ we must have:

$$\rho(A) \leq \|A\|$$

Since this is true for any induced matrix norm $\|\cdot\|$, it must be true for the least upper bound as well:

$$\rho(A) \leq \inf_{\|\cdot\|} \|A\| .$$

- $\rho(A) \geqslant \inf_{\|\cdot\|} \|A\|$: $\forall \varepsilon > 0$, there is a non-singular matrix S s.t.

$$S^{-1}AS = D + T \text{, where } \|T\|_\infty < \varepsilon$$

(λ) (∇) (strict)

$$\Rightarrow \quad \|S^{-1}AS\|_\infty = \|D+T\|_\infty \leq \underbrace{\|D\|_\infty}_{=\rho(A)} + \underbrace{\|T\|_\infty}_{<\varepsilon} < \rho(A) + \varepsilon$$

To conclude we notice (this is easily shown) that:

$$\|A\|'_\infty \equiv \|S^{-1}AS\|_\infty \text{ is an induced matrix norm}$$

$$\Rightarrow \inf_{\|\cdot\|} \|A\| \leq \|A\|'_\infty \leq \rho(A)+\varepsilon$$

$$\Rightarrow \inf_{\|\cdot\|} \|A\| \leq \rho(A) \quad \text{since ineq above holds } \forall \varepsilon > 0.$$

QED

We are now ready for the main result of this section. If you notice all convergence results we have given up to now are simply <u>sufficient</u> conditions for convergence. Herein <u>necessary</u> and <u>sufficient</u> condition for convergence.

First let us write the iterative method in a slightly more general way:

$$x^{(k)} = Gx^{(k-1)} + c \qquad (*)$$

The matrix splitting methods (Jacobi etc.) we have seen before fall on this framework:

$$Qx^{(k)} = (Q-A)x^{(k-1)} + b$$

$$x^{(k)} = \underbrace{(I-Q^{-1}A)}_{=G} x^{(k-1)} + \underbrace{Q^{-1}b}_{=c}$$

What is limit of iteration (*)?

$$x = Gx + c \Rightarrow (I-G)x = c$$

$$\boxed{x = (I-G)^{-1}c}$$

Theorem The iterative method

$$x^{(k)} = Gx^{(k-1)} + c$$

converges to $(I-G)^{-1}c$ if and only if $\rho(G) < 1$.

(essentially "sharpest" convergence measure)

proof:

- Assume $\rho(G) < 1$.

Since $\rho(G) = \inf_{\|\cdot\|} \|G\|$, there is an induced matrix norm for which $\|G\| < 1$.

$$x^{(1)} = Gx^{(0)} + c$$
$$x^{(2)} = G^2x^{(0)} + Gc + c$$
$$x^{(3)} = G^3x^{(0)} + G^2c + Gc + c$$
$$\vdots$$
$$x^{(k)} = G^kx^{(0)} + \sum_{j=0}^{k-1} G^jc$$

Note:

$$\|G^kx^{(0)}\| \leq \|G\|^k \|x^{(0)}\| \to 0 \text{ as } k \to \infty$$

and

$$\sum_{j=0}^{\infty} G^jc = (I-G)^{-1}c \quad \text{by Neumann series}$$

Hence $x^{(k)} \to (I-G)^{-1}c$, regardless of initial guess $x^{(0)}$.

- Assume $\rho(G) \geq 1$: Take $u \neq 0$ s.t.

$$Gu = \lambda u \quad \text{and} \quad |\lambda| \geq 1.$$

Then with $x^{(0)} = 0$ and $c = u$:

then: $$z^{(k)} = \sum_{j=0}^{k-1} G^j u = \sum_{j=0}^{k-1} \lambda^j u = \begin{cases} \pm k u & \text{if } \lambda = \pm 1 \\ \dfrac{1-\lambda^k}{1-\lambda} & \text{if } |\lambda| \neq 1 \end{cases}$$

thus $x^{(k)}$ does not converge as $k \to \infty$. QED.

Corollary The matrix splitting iteration:

$$Q x^{(k)} = (Q - A) x^{(k-1)} + b$$

converges to solution of $Ax = b$

if and only if

$$\rho(I - Q^{-1}A) < 1.$$

Gauss Seidel method:

Q = lower triangular part of A (including diag)
= easy to solve systems (forward substitution)

Theorem If A is diag. dominant then Gauss-Seidel method converges to sol. of $Ax = b$, for any starting vector $x^{(0)}$.

proof: we need to show $\rho(I - Q^{-1}A) < 1$

Let λ be an eigenvalue of $I - Q^{-1}A$ and x a corresp. eigenvector chosen s.t. $\|x\|_\infty = 1$.

$$(I - Q^{-1}A)x = \lambda x \iff \overbrace{(Q - A)}^{= -\text{strictly upper triangular part}} x = \lambda Q x$$

$$\iff -\sum_{j=i+1}^{n} a_{ij} x_j = \lambda \sum_{j=1}^{i} a_{ij} x_j$$

for $i = 1, \dots, n$.

$$\Rightarrow \lambda a_{ii} x_i = -\lambda \sum_{j=1}^{i-1} a_{ij} x_j - \sum_{j=i+1}^{n} a_{ij} x_j \quad (i=1,\dots,n)$$

Let i be s.t. $|x_i| = 1$ (i.e. element realizing $\max_i |x_i| = 1$)

$$|\lambda|\,|a_{ii}| \le |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

$$\Rightarrow |\lambda| \le \frac{\sum_{j=i+1}^{n} |a_{ij}|}{|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|} < 1 \quad \text{by diagonal dominance.}$$

$\Rightarrow \rho(I - Q^{-1}A) < 1 \Rightarrow$ convergence. QED.

let $A = D - E - F$

(↘) (Δ) strict (∇) strict

Successive over relaxation (SOR):

$$Q = \frac{1}{\omega}(D - \omega E), \text{ where } \omega > 0.$$

Iteration becomes:

$$\frac{1}{\omega}(D - \omega E)\, x^{(k)} = \left(\frac{1}{\omega}(D - \omega E) - D + E + F\right) x^{(k)} + b$$

$$= \frac{1}{\omega}\left((1-\omega)D + \omega F\right) x^{(k)} + b$$

$$(D - \omega E)\, x^{(k)} = \left((1-\omega)D + \omega F\right) x^{(k)} + \omega b$$

Complexity is similar to Gauss-Seidel, but because of extra parameter ω one can get better convergence or even convergence in some systems where G-S does not converge.

Note: When $\omega = 1$ SOR $\equiv$ Gauss Seidel.

Here is another way of looking at SOR:

$(1-\omega) \times [\quad D x^{(k)} = D x^{(k-1)} \quad]$ identity.

$\omega \times [(D - E) x^{(k)} = F x^{(k-1)} + b]$ G-S.

$$(D - \omega E) x^{(k)} = ((1-\omega) D + \omega F) x^{(k)} + \omega b$$

Depending on matrix $A = D - E - F$, there are ways of choosing ω optimally.

Extrapolation methods:

$\gamma \times [\quad x^{(k)} = G x^{(k-1)} + c \quad]$

$(1-\gamma) \times [\quad x^{(k)} = x^{(k-1)} \quad]$

$$x^{(k)} = \gamma (G x^{(k-1)} + c) + (1-\gamma) x^{(k-1)}$$

$$= G_\gamma x^{(k-1)} + \gamma c$$

where $G_\gamma = \gamma G + (1-\gamma) I$

= interp between G and I.

If iteration converges then:

$$x = \gamma (Gx + c) + (1-\gamma) x$$

$$\Rightarrow \quad x = Gx + c$$

So setting $G = (I - Q^{-1}A)$ and $c = Q^{-1}b$ we can solve $Ax = b$.

Question How do we choose parameter γ in

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

s.t. $\rho(G_\gamma)$ = minimum

[the smaller $\rho(G_\gamma)$ is the faster convergence rate]

If we know

$$a \le \lambda(G) \le b$$

why?

then

$\lambda(G_\gamma)$ is between $\gamma a + (1-\gamma)$ and $\gamma b + (1-\gamma)$.

$$\Rightarrow \quad \rho(G_\gamma) = \max |\lambda(G_\gamma)| \quad \text{check.}$$

$$= \max |\gamma \lambda(G) + 1 - \gamma|$$

$$\le \max_{a \le \lambda \le b} |\gamma \lambda + 1 - \gamma| .$$

This problem has explicit solution:

Theorem If $\lambda(G) \in [a,b]$, and $1 \notin [a,b]$, then the best choice for γ is:

$$\gamma_* = \frac{2}{2-a-b} \quad \text{and} \quad \rho(G_\gamma) \le 1 - |\gamma| d$$

where d = distance from 1 to $[a,b]$.

We shall not prove this result.

However be aware one can think of more complicated functionals of G and ask the question of which one gives faster convergence

Here is one example that has ties to Chebyshev polynomials which you should have seen in Math 5610, when one is allowed to move the interpolation nodes in order to reduce the Lagrange interpolation error.

Chebyshev acceleration idea

We start with iterative method:

$$x^{(k)} = G x^{(k-1)} + c$$

and ask whether a linear combination of previous iterates

$$u^{(k)} = \sum_{i=0}^{k} a_i^{(k)} x^{(i)}$$

with coeff s.t. $\sum_{i=0}^{k} a_i^{(k)} = 1$ do a better approx than the $x^{(0)}, x^{(1)}, \dots, x^{(k)}$.

$$u^{(k)} = \sum_{i=0}^{k} a_i^{(k)} x^{(i)}$$

$$\ominus \qquad x = \sum_{i=0}^{k} a_i^{(k)} x$$

$$u^{(k)} - x = \sum_{i=0}^{k} a_i^{(k)} (x^{(i)} - x)$$

$$= \sum_{i=0}^{k} a_i^{(k)} G^i (x^{(0)} - x)$$

$$= P(G)(x^{(0)} - x) \quad \Rightarrow \|u^{(k)} - x\| \le \|P(G)\| \, \|x^{(0)} - x\|$$

where P is the polynomial:

$$p(z) = \sum_{i=0}^{k} a_i^{(k)} z^i$$

note: $P(A)$ means exactly what you would have expected.

$$P(A) = \sum_{i=0}^{k} a_i^{(k)} A^i \text{, with } A^0 \equiv I.$$

Now it is not hard to show that

$$\lambda(A) \equiv \text{eigenvalue of } A \Rightarrow \quad p(\lambda(A)) \equiv \text{eigenvalue of } p(A)$$

Thus if we want best approximation we must have:

$$\rho(p(G)) = \max_{1 \le i \le n} |p(\lambda_i(G))|$$

$$\le \max_{z \in S} |p(z)|$$

where polynomial p is s.t.

$$p(1) = 1 \quad \left(\text{since this is equiv. to } \sum_{i=0}^{k} a_i^{(k)} = 1\right)$$

$\Rightarrow$ p is a <u>monic</u> polynomial

Moral of story so that we have reduced the problem of finding the best approx in terms of previous iterates to the problem of finding a polynomial with smallest maximum over some set S.

When $S = [-1, 1]$ the solution can be given in terms of <u>Chebyshev polynomials</u>:

$T_k(z) \equiv$ unique polynomial minimizing $\max_{z \in [-1,1]} |T_k(z)|$ with leading coefficient being 2^{k-1}.

$$\equiv \cos(k \cos^{-1} z)$$

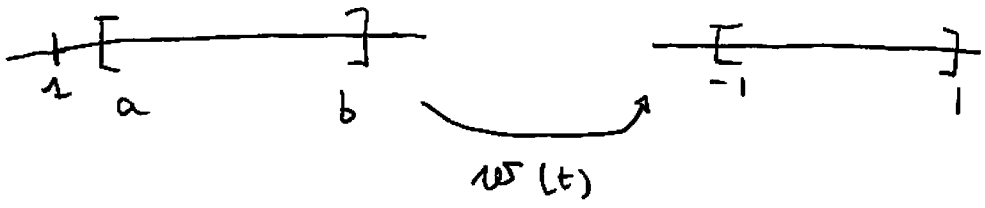
why? Because

$$T_k(z) = \min_{\substack{p \in \Pi_k \\ \text{Largest coeff} \\ = 2^{k-1}}} \max_{z \in [-1,1]} |p(z)|$$

it follows that for $\beta \in \mathbb{R} \setminus (-1,1)$: (to avoid dividing by zero below)
(all roots of T_k are in $(-1,1)$)

$$\frac{T_k(z)}{T_k(\beta)} = \min_{\substack{p \in \Pi_k \\ p(\beta)=1}} \max_{z \in [-1,1]} |p(z)|$$

If $S = [a,b]$ = interval where eigenvalues of G lie,
and $[a,b] \not\ni 1$, then we can find transformation s.t.:



where

$$w(t) = (-1)\frac{t-b}{a-b} + (1)\frac{a-t}{a-b} \qquad \text{(Lagrange interp)}$$

$$= \frac{a+b-2t}{a-b}$$

then:

$$\underbrace{\frac{T_k(w(t))}{T_k(w(1))}}_{L_k} = \min_{\substack{p \in \Pi_k \\ p(1)=1}} \max_{z \in [a,b]} |p(z)|$$

L_k can be computed on the fly with a three term recurrence.
So we only need to remember 2 past iterates.

CONJUGATE GRADIENT

Objective: Solve $Ax = b$ when $A \in \mathbb{R}^{n \times n}$, symm. pos. def.

We shall use the notation $\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$

- inner product.

Recall properties of an inner product (these hold in general not only in $\mathbb{R}^n$)

inner product $\equiv$ positive symmetric bilinear form

i) $\langle x, y \rangle = \langle y, x \rangle$ (symm)

ii) $\langle x, x \rangle \geq 0$ for all $x \in \mathbb{R}^n$ and $\langle x, x \rangle = 0 \iff x = 0$ } (pos.)

iii) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ (bilin)

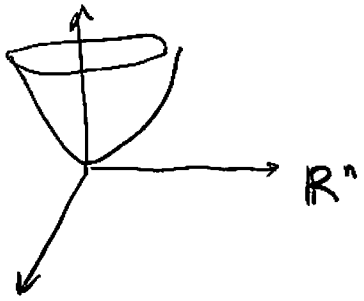
Also we have the following property.

iv)
$$\langle x, Ay \rangle = x^T (Ay) = (A^T x)^T y = \langle A^T x, y \rangle$$

Conjugate gradient exploits the following equivalence between linear system and an optimization problem:

(*) | Find x s.t. $Ax = b$ | $\iff$ | $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T A x - x^T b = q(x)$

Here $q(x) \equiv$ quadratic function. When $x \in \mathbb{R}^2$:



Finding solution of $Ax = b$ $\Leftrightarrow$ finding bottom of bowl.

note: the fact that we have an (upward) bowl and not a hyperboloid is due to A being S.P.D.
($x^T Ax \geq 0 \quad \forall x \in \mathbb{R}^n$, so we cannot have branches going to $-\infty$)

If you know some vector calculus and first and second order sufficient conditions for having a minimizer then showing (*) ... is simply:

$$\nabla q(x) = Ax - b$$

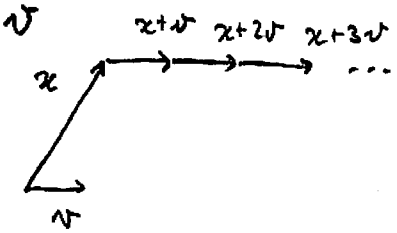
$$\nabla^2 q(x) = A = \text{S.P.D.}$$

Think of $f: \mathbb{R} \to \mathbb{R}$; $f \in C^2$
$f'(x_*) = 0$ and $f(x_*) > 0$
$\Rightarrow x_*$ is a minimizer of f

However here is a direct proof of (*)

proof of (*):

Look at ray $x + tv$



$$q(x+tv) = \tfrac{1}{2}(x+tv)^T A (x+tv) - (x+tv)^T b$$
$$= q(x) + t\, v^T A x - t\, v^T b + \tfrac{1}{2} t^2 v^T A v$$

$$\frac{d}{dt} q(x+tv) = v^T(Ax-b) + t\, v^T A v$$

q is minimum when $t = t_* = -\dfrac{v^T(Ax-b)}{v^T A v}$.

and:

$$q(x+t_* v) = q(x) + t_* \left(v^T(Ax-b) + \tfrac{1}{2} t_* v^T A v \right)$$
$$= q(x) + t_* \left(v^T(Ax-b) - \tfrac{1}{2} v^T(Ax-b) \right)$$
$$= q(x) - \frac{(v^T(Ax-b))^2}{2\, v^T A v}$$

What does this mean?

If x is minimizer of $q(x)$, then if we move from x in <u>any</u> direction v, then we should not be able to get a smaller value for q.

$\Rightarrow$ $v^T(Ax-b) = 0$ for all directions v

$\Rightarrow$ $Ax - b = 0$

<u>Conversely</u>: If $Ax = b$ then

$$q(x+tv) = q(x) + \tfrac{1}{2} t^2 v^T A v > q(x)$$

when $v \neq 0$.