Convergence in $\mathbb{R}^n$

We say a sequence of vectors $v^{(k)} \in \mathbb{R}^n$ converges to $v \in \mathbb{R}^n$ as $k \to \infty$ if

$$\lim_{k\to\infty} \|v^{(k)} - v\| = 0$$

In $\mathbb{R}^n$ the choice of norm does not matter as all <u>norms are equivalent</u>: (this is true for any finite dimensional space)

Two norms $\|\cdot\|$ and $\|\cdot\|$ are equivalent if there are two constants c_1 and $c_2 > 0$ s.t.

$$c_1\|x\| < \|x\| \leq c_2\|x\|.$$

Also $\mathbb{R}^n$ is <u>complete</u> meaning any sequence satisfying the Cauchy criterion converges.

$$\forall \varepsilon > 0 \quad \exists N \text{ s.t. } \forall i,j \geq N \quad \|v^{(i)} - v^{(j)}\| < \varepsilon$$

Neumann Series

Let $A \in \mathbb{R}^{n\times n}$ be such that $\|A\| < 1$, then

i) $I - A$ is invertible

ii) $(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$

<u>Proof</u>: Assume for contradiction that $I - A$ is singular

$\Rightarrow \exists x \in \mathbb{R}^n$ s.t. $(I-A)x = 0$, take w.l.o.g. x s.t. $\|x\| = 1$

then:

$$1 = \|x\| = \|Ax\| \leq \underbrace{\|A\|}_{<1}\|x\| < \|x\| = 1. \quad \text{contradiction!}$$

Now we need to show that

$$\sum_{k=0}^{m} A^k \to (I-A)^{-1} \quad \text{(i.e. partial sums converge)}$$

Notice: (telescoping series)

$$(I-A)\sum_{k=0}^{m} A^k = \sum_{k=0}^{m} A^k - A^{k+1} = A^0 - A^{m+1}$$
$$= I - A^{m+1}$$

Now using properties of induced matrix norm: ($\|AB\| \le \|A\|\|B\|$)

$$\|A^{m+1}\| \le \|A\|^m \to 0$$

Thus:

$$\left\| (I-A)\sum_{k=0}^{m} A^k - I \right\| \le \|A\|^{m+1} \to 0 \text{ as } m \to \infty.$$

QED

Note: This theorem is a very intuitive generalization of what you already know about geometric series.

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots, \text{ when } |x| < 1.$$

Here is the kind of result we can show with Neumann Series:

$$\|(I-A)^{-1}\| \le \sum_{k=0}^{\infty} \|A^k\| \le \sum_{k=0}^{\infty} \|A\|^k = \frac{1}{1-\|A\|}.$$

Note: This theorem is very useful for the theory, however it is seldom used in practice. (There are faster iterative methods for solving $Ax = b$).

Here is an easy but handy generalization of Neumann Series:

Theorem

If A and B are matrices s.t. $\|I - AB\| < 1$ then A, B are invertible and

$$A^{-1} = B \sum_{k=0}^{\infty} (I - AB)^k$$

$$B^{-1} = \left[\sum_{k=0}^{\infty} (I - AB)^k \right] A$$

Proof: By Neumann series:

$I - (I - AB) = AB$ is invertible and:

$$(AB)^{-1} = \sum_{k=0}^{\infty} (I - AB)^k$$

$$\Rightarrow \quad A^{-1} = BB^{-1}A^{-1} = B \sum_{k=0}^{\infty} (I - AB)^k$$

$$B^{-1} = B^{-1}A^{-1}A = \left[\sum_{k=0}^{\infty} (I - AB)^k \right] A$$

Iterative refinement

Let $x^{(0)}$ be an approx. sol to

$$Ax = b.$$

then:

$$x = x^{(0)} + A^{-1}(b - Ax^{(0)})$$

We often call

$r^{(0)} = b - Ax^{(0)}$ = residual vector

$e^{(0)} = A^{-1}r^{(0)}$ = error vector.

Algorithm

$x^{(0)}$ given

for $k = 0, 1, \ldots$

$$r^{(k)} = b - Ax^{(k)}$$
$$Ae^{(k)} = r^{(k)} \quad \leftarrow \text{usually inexact solve}$$
$$x^{(k+1)} = x^{(k)} + e^{(k)}$$

Solution can be improved even if the solves are inexact for example:

→ LU factorization is stopped prematurely (= Incomplete LU fact.)

→ $Ae^{(k)} = r^{(k)}$ is solved with an iterative method

→ A is not known exactly.

To be more precise about what we mean by "solving approximatively"

Let $B \equiv$ "approximate" inverse of A

$$\approx A^{-1} \quad \text{(in some sense we shall see)}$$

The iterates are:

$$x^{(0)} = Bb$$
$$x^{(k+1)} = x^{(k)} + B(b - Ax^{(k)}), \quad k \geq 0.$$

Looking at a few iterates we see a pattern emerge.

$$x^{(0)} = Bb$$
$$x^{(1)} = x^{(0)} + B(b - Ax^{(0)})$$
$$= Bb + B(b - ABb)$$
$$= Bb + B(I - AB)b$$
$$x^{(2)} = x^{(1)} + B(b - Ax^{(1)})$$
$$= Bb + B(I-AB)b + B(b - A(Bb + B(I-AB)b))$$
$$= Bb + B(I-AB)b + B((I-AB)b - AB(I-AB)b)$$
$$= Bb + B(I-AB)b + B(I-AB)^2 b$$
$$\vdots$$

$$\boxed{x^{(m)} = B\sum_{k=0}^{m} (I-AB)^k b}$$

Theorem (Iterative refinements)

Iterates from iterative refinement are:

$$x^{(m)} = B\sum_{k=0}^{m} (I-AB)^k b \quad (m \geq 0)$$

and because of the generalized Neuman series:

$$x^{(m)} \to x \text{ as } m \to \infty \text{ when } \|I - AB\| < 1.$$

Proof: Can be done by induction on m.

Solving equations with iterative methods

Idea: Instead of having a direct method which finds solution in finitely many steps, use an iterative method to find succesive approx that converge to sol.

Motivational example

$$\begin{bmatrix} 7 & -6 \\ -8 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

Jacobi method: Solve i-th equation for i-th unknown:

$$\begin{cases} x_1^{(k)} = \frac{3}{7} + \frac{6}{7} x_2^{(k-1)} \\ x_2^{(k)} = \frac{-4}{9} + \frac{8}{9} x_1^{(k-1)} \end{cases}$$

Gauss Seidel method: use improved value of $x_1^{(k)}$ in second equation.

$$\begin{cases} x_1^{(k)} = \frac{3}{7} + \frac{6}{7} x_2^{(k-1)} \\ x_2^{(k)} = \frac{-4}{9} + \frac{8}{9} x_1^{(k)} \end{cases}$$

converges faster!

Note: convergence for these methods depends on initial guess $(x_1^{(0)}, x_2^{(0)})$!

Matrix splitting

Consider the system $Ax = b$.

Introduce a non-singular matrix Q and rewrite system as:

$$Qx = (Q - A)x + b$$

Get iteration:

$$Q x^{(k)} = (Q - A) x^{(k-1)} + b, \quad k \geq 1, \; x^{(0)} \text{ given}$$

Q should have the desirable properties:

① $x^{(k)} \to x$ where x is a sol. of $Ax = b$

② Solving systems with Q is cheap.

If $x^{(k)} \to x$ then the limit x must satisfy:

$$Qx = (Q - A)x + b \quad \Rightarrow \quad Ax = b$$

Assuming A & Q are non-singular:

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

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

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

$$\Rightarrow \|x^{(k)} - x\| \leq \|I - Q^{-1}A\| \, \|x^{(k-1)} - x\|$$

$$\vdots$$

$$\leq \|I - Q^{-1}A\|^k \, \|x^{(0)} - x\|$$

Thus when $\|I - Q^{-1}A\| < 1$ we have $x^{(k)} \to x$, where $Ax = b$.

Theorem If $\|I - Q^{-1}A\| < 1$ for some induced matrix norm then seq.

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

converges to sol. of $Ax = b$ for any starting vector $x^{(0)}$

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:

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

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

Proof.

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

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

$\Rightarrow$ Convergence.

Linear algebra reviews

Eigenvalues of A = roots of characteristic polyn. of A

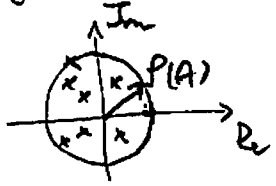
The characteristic poly. of A is defined by:

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

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

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

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.

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

$$(\lambda) \quad (\nabla) \text{ (strict)}$$

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

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

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

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

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

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 sufficient conditions for convergence. Herein necessary and sufficient 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 (*)?

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

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

Theorem The iterative method

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

converges to $(I-G)^{-1}c$ ← essentially "sharpest" convergence measure

if and only if $\rho(G) < 1$.

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:

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

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 $\rho(G) \ge 1$: Take $u \neq 0$ s.t.

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

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. $\|x\|_\infty = 1$.

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

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

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:

$(1-\omega) \times [\; D x^{(k)} = D x^{(k-1)} \;]$ 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:

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

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

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

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

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 $x^{(0)}, x^{(1)}, \dots, x^{(k)}$.

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

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

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

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

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

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.

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

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}.

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

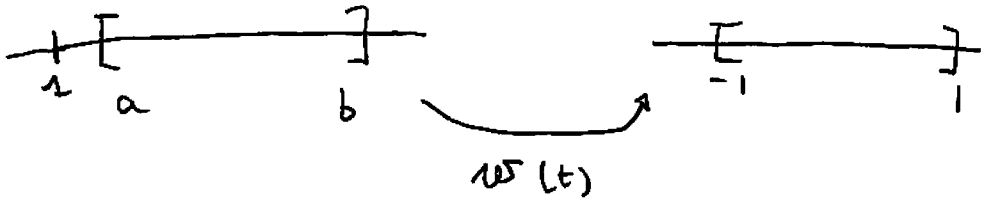
why? Because

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

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)

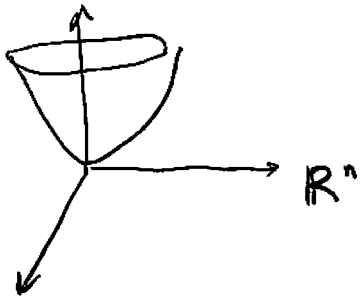
Also we have the following property.

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:

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

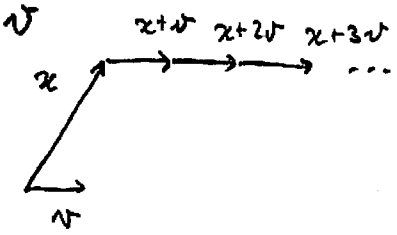
$$\nabla^2 q(x) = A = \text{S.P.D.}$$

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 $v \neq 0$.

This gives idea on how to find solution to $Ax = b$.

$$x^{(k+1)} = x^{(k)} + t_k v^{(k)}$$

t_k: step size; $v^{(k)}$: "search direction"

If $v^{(k)}$ is given then our calculation reveals that the best possible stepsize we can take (i.e. the one that reduces the most value of objective function $q(x)$) is:

$$t_k = \frac{\langle v^{(k)}, b - Ax^{(k)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} = \frac{\langle v^{(k)}, r^{(k)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle}$$

where $r^{(k)} = b - Ax^{(k)}$ = residual at step k.

One example of such method is:

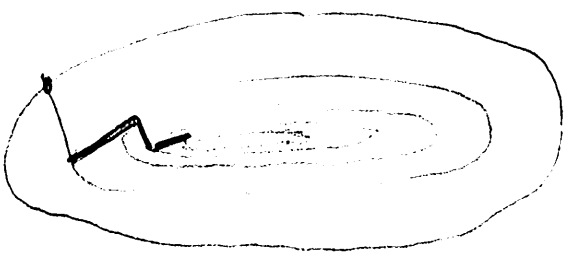
Steepest descent where $v^{(k)} = r^{(k)} = -\nabla q(x^{(k)})$

Algorithm:

for $k = 1, 2, \ldots$

$$v = b - Ax$$
$$t = \frac{(v, v)}{(v, Av)}$$
$$x = x + tv$$

This is not a very efficient method. It has slow, typically "zig-zagging" convergence to solution:



Level sets of $q(x)$ are ellipses when $x \in \mathbb{R}^2$.

Note: $r^{(k)}$ is $\perp$ to level sets of $q(x)$.

Conjugate Gradient method (Hestenes & Stiefel, 1952)

$x_{k+1} = x_k + \alpha_k p_k$ where α_k, p_k is determined s.t. x_{k+1} solves

(1) $\min \frac{1}{2} x^T A x - b^T x$

$x \in x_0 + \text{span}\{p_0, \dots, p_{k-1}, r_k\}$

$x = \hat{x} + x_0$

$\Leftrightarrow$ (2) $\min \frac{1}{2} \hat{x}^T A \hat{x} - \hat{x}^T r_0$

$\hat{x} \in \text{span}\{p_0, \dots, p_{k-1}, r_k\}$

Recall: $\text{span}\{u_1, u_2, \dots u_n\} = \left\{ \sum_{i=1}^{n} \alpha_i u_i \;\middle|\; \alpha_i \in \mathbb{R} \text{ for } i = 1, \dots, n \right\}$

= set of all possible linear combinations of vectors $u_1, \dots u_n$

= linear span of family $\{u_1, \dots u_n\}$.

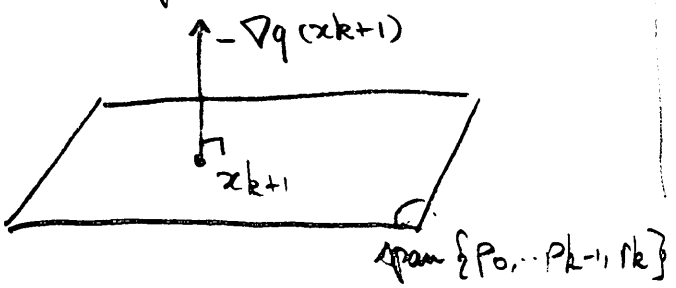
It can be shown that $\hat{x}_{k+1}$ solves (2) iff:

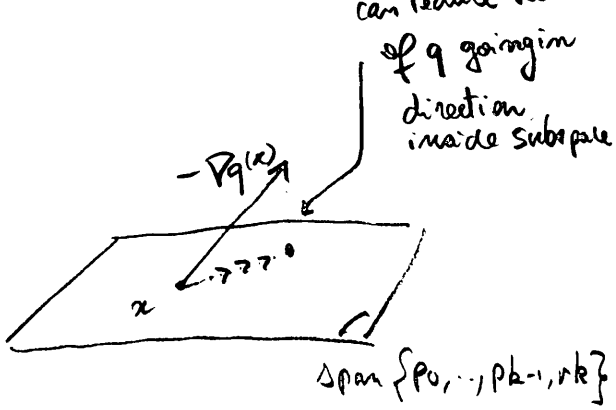
$$(A\hat{x}_{k+1} - r_0)^T v = 0 \quad \forall v \in \text{span}\{p_0, \dots, p_{k-1}, r_k\}$$

$\Leftrightarrow$ $(A x_{k+1} - b)^T v = 0$ " " "

$= -\nabla q(x_{k+1})$

Intuitively:





OPTIMAL: no direction of descent in span $\{\dots\}$ **SUB-OPTIMAL**

note: v is called a descent direction for q if

$$v^T \nabla q(x) < 0.$$

Looking at Taylor expansion around x:

$$q(x+tv) = q(x) + \underbrace{t\, \nabla q(x)^T v}_{<0 \text{ when } t>0 \text{ and } \nabla q(x)^T v < 0.} + o(t)$$

$\Rightarrow$ means there is a step length $t>0$ for which $q(x+tv) < q(x)$.

Now look at previous step $\hat{x}_k$, which must solve:

$$\min \tfrac{1}{2}\hat{x}^T A \hat{x} - \hat{x}^T r_0 \quad \Leftrightarrow \quad v^T(b - Ax_k) = 0, \quad \forall v \in \text{span}\{p_0, \dots p_{k-1}\}$$

$$\hat{x} \in \text{span}\{p_0, \cdots, p_{k-1}\}$$

$$\Leftrightarrow \boxed{p_j^T (b - Ax_k) = 0 \quad \text{for } j = 0, \dots, k-1.}$$

We also have:

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

$$0 = (Ax_{k+1} - b)^T p_j = (Ax_k - b)^T p_j + \alpha_k p_k^T A p_j \qquad (Ax_k - b)^T p_j = 0$$

for $j = 0, \dots, k-1$

$$\Rightarrow \boxed{p_k \text{ is } A\text{-orthogonal to previous } k \text{ directions}}$$

A-orthogonal means orthogonality w.r.t. inner product:

$$\langle u, v \rangle_A = \langle u, Av \rangle = u^T A v.$$

It is not hard to check $\langle u, v \rangle_A$ is indeed an inner product when A is s.p.d.

So we can obtain p_k be A-orthogonalizing family of vectors $\{r_0, r_1, \ldots, r_{k-1}, r_k\}$.

Use Gram-Schmidt orthogonalization.

$$p_0 = r_0$$
$$p_k = r_k - \sum_{j=0}^{k-1} \underbrace{\frac{r_k^T A p_j}{p_j^T A p_j} p_j}$$

$= \perp$ proj w.r.t. $\langle \cdot, \cdot \rangle_A$ inner prod of r_k along p_j.

Now take the optimal step size along this direction:

$$\alpha_k = \frac{p_k^T (b - Ax_k)}{p_k^T A p_k} = \frac{p_k^T r_k}{p_k^T A p_k}$$

Gram Schmidt + this choice of step size is essentially a preliminary version of CG.

However in this version:

- cost of iteration grows linearly with iteration number.
- storage needed " " " " " .

Fortunately the Gram-Schmidt orthogonalization takes a simpler form if we look closely at subspaces we use.

The subspace that CG uses to look for new direction is:

$$\text{span}\{p_0, p_1, \dots, p_{k-1}, r_k\} = \text{span}\{p_0, \dots, p_{k-1}, p_k\}$$
$$= \text{span}\{r_0, \dots, r_{k-1}, r_k\}$$
$$= \mathcal{K}_{k+1}(A, r_0)$$
$$= \underline{\text{Krylov subspace}}$$

why?

$$r_0 = b - Ax_0 \in \mathcal{K}_1(A, r_0)$$

If $r_{k-1} = b - Ax_{k-1} \in \mathcal{K}_k(A, r_0)$ then

$$r_k = b - Ax_k = b - A(x_{k-1} + \alpha p_{k-1})$$
$$= \underbrace{b - Ax_{k-1}}_{= r_{k-1} \in \mathcal{K}_k} - \alpha \underbrace{A \underbrace{p_{k-1}}_{\in \mathcal{K}_k}}_{\in \mathcal{K}_{k+1}}$$

$$\Rightarrow \boxed{r_k \in \mathcal{K}_{k+1}}$$

Conjugate Gradient forms part of a large family of iterative solvers called <u>Krylov Subspace methods</u>.

<u>Why did we introduce these subspaces?</u>

Recall optimality conditions, now written with Krylov subspaces.

$$r_{k+1}^T v = 0, \quad \forall v \in \mathcal{K}_{k+1}(A, r_0)$$
$$r_k^T v = 0, \quad \forall v \in \mathcal{K}_k(A, r_0)$$

Take $p_i \in \mathcal{K}_{k-1}(A, r_0) \Rightarrow Ap_i \in \mathcal{K}_k(A, r_0)$

$$\Rightarrow r_k^T A p_i = 0, \quad \text{for } i = 0, \dots, k-2$$

This orthogonality greatly simplifies Gram Schmidt and reduces sum to only one term:

$$p_k = r_k - \underbrace{\frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}}_{\beta_k} p_{k-1}$$

- Storage and cost of iteration remains the same regardless of k.
- One needs roughly only 5 vectors of length n.
- One requires only knowledge of action of A on a vector.
  $\Rightarrow$ no need to know entries of A as in direct methods.
  A could be even specified via a "black-box" code.

In the classical formulation of CG α_k and β_k take different forms, which can be obtained by applying optimality conditions:

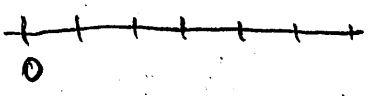
$$\alpha_k = \frac{\|r_k\|^2}{p_k^T A p_k}$$

$$\beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}$$

Convergence of CG:

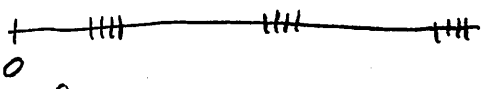
- If A is s.p.d. CG converges in at most n steps, where n = dimension of A.
- In general convergence of CG depends on how many "clusters" of eigenvalues does A have.

Spectrum of A (all eigenvalues)



Slow convergence

(S1)



fast convergence in roughly as many iterations as there are clusters (here $\sim$ iter)

(S2)

Preconditioning

To speed up convergence of CG (i.e. going from (S1) to (S2)) one solves system:

$$M^{-1}Ax = M^{-1}b, \text{ where } M \text{ is invertible}$$

$$\Leftrightarrow \quad Ax = b$$

M = (left) preconditioner

= approximation of A that is easy to invert.

One can also think of right preconditioners:

$$A\underbrace{M^{-1}y}_{=x} = b \qquad \Leftrightarrow Ax = b$$

For symmetric systems, we don't want to loose symmetry, so letting

$$M = CC^T, \ M \text{ invertible:}$$

$$Ax = b \iff \underbrace{C^{-1}AC^{-T}}\, C^T x = C^{-1}b$$

s.p.d. if A s.p.d.

Note: there is no single way of choosing a preconditioner. A preconditioner that works for a particular problem may not work as well for another.

Here are some preconditioning techniques that work (see Trefethen and Bau for a broader overview)

- incomplete LU or Cholesky factorization: drop elements that are below a threshold to get sparse LU (L) factors that give cheap systems to solve
- Jacobi: use diagonal or block-diagonal of A
- Discretization based: use solution to a coarse grid problem, a constant coeff problem, a periodic problem or solve problem in alternating directions (we will see this more closely when we get to ADI methods)

The final algorithm is as follows:

CONJUGATE GRADIENT (w/ preconditioned)

x_0 = given

$r_0 = b - Ax_0$

$p_0 = r_0$

for $n = 1, 2, 3, \ldots$

$$\alpha_n = \frac{r_{n-1}^T r_{n-1}}{p_{n-1}^T A p_{n-1}}$$

$$x_n = x_{n-1} + \alpha_n p_{n-1}$$

$$r_n = r_{n-1} - \alpha_n A p_{n-1}$$ ← check convergence here by looking at $\|r_n\|$

$$\beta_n = \frac{r_n^T r_n}{r_{n-1}^T r_{n-1}}$$

$$p_n = r_n + \beta_n p_{n-1}$$

and the preconditioned version is:

PRECONDITIONED CONJUGATE GRADIENT

x_0 = given

M = preconditioner = given

$r_0 = b - A x_0$

Solve $M y_0 = r_0$ for y_0

$p_0 = y_0$

for $n = 1, 2, 3, \cdots$

$$\alpha_n = \frac{r_{n-1}^T y_{n-1}}{p_{n-1}^T A p_{n-1}}$$

$$x_n = x_{n-1} + \alpha_n p_{n-1}$$

$$r_n = r_{n-1} - \alpha_n A p_{n-1}$$ ← check convergence here.

Solve $M y_n = r_n$ for r_n. should be y_n

$$\beta_n = \frac{y_n^T r_n}{y_{n-1}^T r_{n-1}}$$

$$p_n = y_n + \beta_n p_{n-1}$$

note: This is equivalent to applying CG to system

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

where $M = C C^T$ = preconditioner.

we will not see equivalence but notice:

prec. residual

$$y_n = M^{-1} r_n \Rightarrow r_n^T y_n = r_n^T M^{-1} r_n = (C^{-1} r_n)^T (C^{-1} r_n)$$