Solving equations with Iterative methods

Idea: instead of having a direct method that finds solution to $Ax = b$ in a finite number of steps, use an iterative method to find successive approx that converge to solution.

Here are two simple examples:

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

One strategy is to ~~solve ith eq for i-th unknown~~

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

Jacobi method

We could improve this by using newest value:

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

Gauss-Seidel method

more up to date than Jacobi

actually converges faster!

Note: convergence depends on initial guess

Theory:

$$Ax = b$$

introduce so called Splitting matrix Q:

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

Use to define operation.

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

$x^{(0)}$ given

Not any choice of Q would work, we need:

① $x^{(k)}$ converges to sol of $Ax=b$, and rapidly

② $x^{(k)}$ is cheap to compute.

We shall see that ① follows from $Q^{-1} \approx A^{-1}$

② ——— $Qx=y$ cheap to solve.

Let's try to see what is going on:

If $x^{(k)} \to x_*$ the x_* must satisfy:

$$Qx_* = (Q-A)x_* + b$$

$$\Rightarrow \quad Ax_* = b$$

We assume: A is non-singular

Q ———

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

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

⚠ never written like this in practice! to find $x^{(k)}$ we solve a system.

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

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

$$\vdots$$

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

with appropriate matrix induced norm

Thus: if $\|I - Q^{-1}A\| < 1$ then $x^{(k)} \to x$, sol of $Ax=b$.

Theorem If $\|I - Q^{-1}A\| < 1$ for some induced matrix norm then the seq. $Qx^{(k)} = (Q-A)x^{(k-1)} + b$ converges to sol of $Ax=b$ for any starting vector $x^{(0)}$.

Let $\delta = \| I - Q^{-1}A \| < 1$, One can use as stopping criterion the difference between two consecutive iterates as:

$$\| x^{(k)} - x \| \leq \frac{\delta}{1-\delta} \| x^{(k)} - x^{(k-1)} \|$$

note: can be bad if δ is close to 1.

Richardson method

$$Q = I$$

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

when does this work?

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

Jacobi method or iteration

$$Q = \begin{bmatrix} a_{11} & & \\ & a_{22} & \\ & & \ddots \\ & & & a_{nn} \end{bmatrix}$$ = diagonal matrix with same entries on diagonal as matrix A.

In this case:

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

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

Theorem on convergence of Jacobi method

If A is diagonally dominant, then the sequence of iterates produced by Jacobi method converges to sol of $Ax=b$.

proof:

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

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

$\Rightarrow$ convergence

Algorithm:

for $k = 1, 2, \ldots$

$\quad x = (b - Ax) ./ \text{diag}(A)$

Note: one can avoid divisions if system is preprocessed:

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

$$(D^{-1/2} A D^{-1/2})(D^{1/2} x) = D^{-1/2} b \quad \text{symm}$$

Some preliminaries before we do analysis of general case.

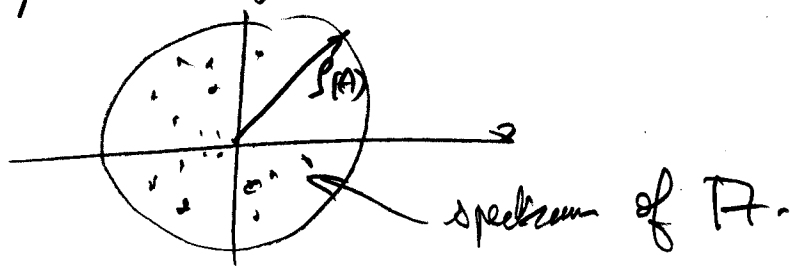
eigenvalues of A = roots of characteristic poly:

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

Spectral radius:

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

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



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

$$AX = XB$$

Theorem on similar upper triangular matrices

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

Use Schur factorization:

$$A = Q T Q^T \qquad Q^T Q = I$$

T eigenvalues of A are on diag.

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

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

But $T = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$ then:

$$t_{ij} = 0 \text{ if } j < i$$

now if $j > i$ (elements above diag):

$$|t_{ij}\, \varepsilon^{(j-i)}| \leq \varepsilon |t_{ij}|$$

($j - i \geq 1$)

then:

$$A = \underbrace{Q\,D}_{X}\ \underbrace{D^{-1}TD}_{B}\ \underbrace{D^{-1}Q^T}_{X^{-1}}$$

↓ sought after

Theorem on spectral radius :

$$\rho(A) = \inf_{\|\cdot\|} \|A\|$$ in which infimum is over all induced matrix norms.

This shows $\rho(A) = \|A\|_2$ is "smallest" induced norm for A.

proof:

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

Then:

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

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

$$\Rightarrow \rho(A) \leq \inf_{\|\cdot\|} \|A\| \quad \text{(Greatest lower bound)}$$

For reverse inequality:

$\forall \varepsilon > 0$, there is a non-singular matrix S s.t.

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

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

$$\leq \rho(A) + \varepsilon$$

this is a subordinate norm $\|A\|'_\infty$

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

$\leq \rho(A)$ since ε is arbitrary.

Necessary and Sufficient conditions for convergence of an iterative method.

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

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

proof: if $\rho(G) < 1$ then we have induced norm for which

$$\|G\| < 1$$

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

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

Moreover Neumann Series gives:

$$\sum_{j=0}^{\infty} G^j c = (I-G)^{-1} c$$

Hence $x^{(k)} \to (I-G)^{-1}c$ as $k \to \infty$.

Now we show converse. Suppose $\rho(G) \geq 1$ and take u s.t.

$$Gu = \lambda u \text{ and } |\lambda| \geq 1. \quad (u \neq 0)$$

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

$$x^{(k)} = \sum_{j=0}^{k-1} G^j u = \sum_{j=0}^{k-1} \lambda^j u$$

$$= \begin{cases} ku & \text{if } \lambda = 1 \\ \dfrac{1-\lambda^k}{1-\lambda} u & \text{if } \lambda \neq 1 \end{cases}$$

which doesn't converge as $k \to \infty$.

Corollary w/ splitting matrix

The iteration:

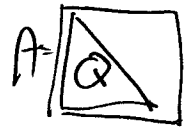
$$Qx^{(k)} = (Q-A)x^{(k-1)} + b \text{ converges to } Ax = b$$

for any starting point $x^{(0)}$ iff $\rho(I - Q^{-1}A) < 1$

Theorem Gauss Seidel convergence

If A is diag dominant then G.S. method converges for any starting $x^{(0)}$.

proof: Q in Gauss Seidel is lower triang part of A (including diag)



Need to show $\boxed{\rho(I - Q^{-1}A) < 1}$

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

$$(I - Q^{-1}A)x = \lambda x \iff (Q - A)x = \lambda Q x$$

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

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

Let i be s.t. $|x_i| = 1$:

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

$$|\lambda| \le \frac{\sum_{j=i+1}^{n} |a_{ij}|}{|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|} < 1 \qquad \text{(by diag dom)}$$

A note on parallelism:

Assume we can make operations in parallel w/o parallelization.

What method is more parallelizable?

Jacobi: each eq can be solved independently

GS: x_i requires comp of $x_1, x_2, \dots, x_{i-1}$

Here is a summary of different methods we have seen:

Richardson $\quad Q = I$

$G = I - A$

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

$A = D - E - F$

(\) $\triangle$ $\triangledown$

Strict.

Jacobi $\quad Q = D$

$G = D^{-1}(E+F)$

$Dx^{(k)} = (E+F)x^{(k-1)} + b$

Gauss Seidel $\quad Q = D - E$

$G = (D-E)^{-1}F$

$(D-E)x^{(k)} = Fx^{(k-1)} + b$

SOR Successive over relaxation (lin comb of Jac & GS).

$Q = \omega^{-1}(D - \omega E)$

$G = (D-\omega E)^{-1}(\omega U + (1-\omega)D)$

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

One can choose param ω optimally

Extrapolation: (in general)

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

$$x^{(k)} = \gamma\left(Gx^{(k-1)} + c\right) + (1-\gamma)x^{(k-1)}$$

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

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

$\gamma = 1$ get old iter., $\gamma = 0$ get identity.

If iteration converges:

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

$$x = Gx + c \quad (\text{since } \gamma \neq 0)$$

If $G = I - Q^{-1}A$ and $c = Q^{-1}b$ $\rightsquigarrow$ we solve $Ax = b$.

Fact from linear algebra: If λ is an eigenvalue of A then $p(\lambda)$ is an eigenvalue of $p(A)$

Question: how do we choose parameter γ in $G_\gamma = \gamma G + (1-\gamma) I$?

If we know:

$$a \leq \underbrace{\lambda(G)}_{\text{eigenvalue of } G} \leq b$$

then

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

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

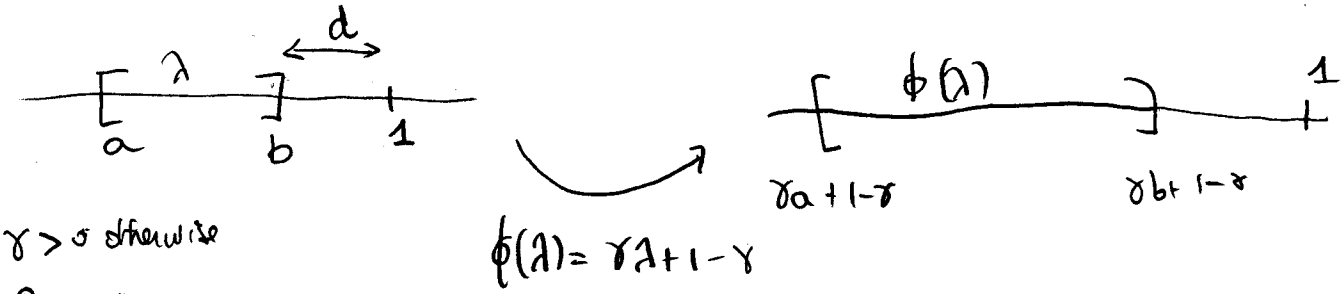
$$\leq \max_{a \leq \lambda \leq b} |\gamma \lambda + 1 - \gamma|$$

Theorem: If all we know about G is that its eigenvalues lie in interval $[a,b]$, and if $1 \notin [a,b]$, then best choice for γ is:

$$\gamma_* = \frac{2}{2-a-b}$$

and $\rho(G_\gamma) \leq 1 - |\gamma| d$ where d = distance from 1 to $[a,b]$.

proof: Since $1 \notin [a,b]$ either $a > 1$ or $b < 1$. ← we do this case:



$\phi(\lambda) = \gamma \lambda + 1 - \gamma$

note: $\gamma > 0$ otherwise $\rho(G_\gamma) > 1$

$\gamma b + 1 - \gamma = 1 + \gamma(b-1) = 1 - \gamma d$

$\gamma a + 1 - \gamma = \gamma(a+b-2) + 1 + \gamma(1-b) =$

$\gamma d + \gamma(a+b-2) + 1 \leq \phi(\lambda) \leq 1 - \gamma d$

take $\gamma_* = \frac{-2}{a+b-2}$ so that

$\gamma d - 1 \leq \phi(\lambda) \leq 1 - \gamma d$

$|\phi(\lambda)| < 1 - \gamma d$

$\rho(G_\gamma) \leq 1 - \gamma d$

Why is this best choice?

$\gamma > \gamma_* \rightarrow \gamma a + 1 - \gamma \leq \gamma_* d - 1 \rightarrow \rho(G_\gamma)$ increases

$\gamma < \gamma_* \rightarrow 1 - \gamma d > 1 - \gamma_* d \rightarrow$ ———— .

Notation: $m(G) =$ smallest eigenvalue of G } assuming G has only real eigenvalues.
$M(G) =$ largest ———— }

What is spectral radius of optimal extrapolated Richardson's?

$m(G) = 1 - M(A)$
$M(G) = 1 - m(A)$

acceleration possible when $M(A) < 0$ (all negative)
$m(A) > 0$ (all pos.)

and $\gamma_* = \frac{-2}{1 - M(A) + 1 - m(A) - 2} = \frac{2}{m(A) + M(A)}$

the spectral radius is:

$$\rho(G_\gamma) \leq 1 - \frac{2}{m(A)+M(A)}\left(1 - (1 - m(A))\right) = \frac{M(A) - m(A)}{M(A) + m(A)}$$

(assuming $1 - m(A) < 1$. or $m(A) > 0$.

Jacobi $\quad \gamma_i = \dfrac{2}{m(D^{-1}A) + M(D^{-1}A)}$

$$\rho(G_J) = \frac{M(D^{-1}A) - m(D^{-1}A)}{M(D^{-1}A) + m(D^{-1}A)}$$

CONJUGATE GRADIENT METHOD

Objective: solve system $Ax = b$ when

$A \in \mathbb{R}^{n \times n}$, symm pos def.

we shall use notation: $\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$ for inner product.

Recall properties (positive, symmetric bilinear form)

$\langle x, y \rangle = \langle y, x \rangle$

$\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Leftrightarrow x = 0$

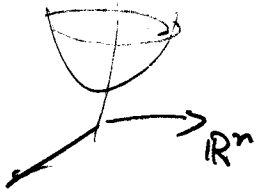
$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$

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

First thing to notice is that:

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

$q(x)$ = quadratic function

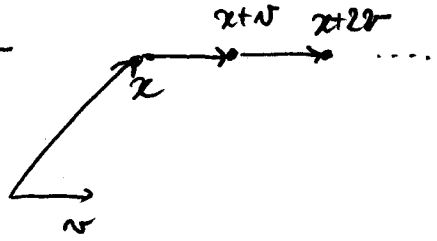


easy way of seeing equivalence: $\nabla q(x) = Ax - b$

(if you know 2nd order suff cond for existence of a minimizer) $\nabla^2 q(x) = A$ s.p.d.

Proof of equivalence (*):

Look at ray $x + tv$



$$q(x+tv) = \frac{1}{2}(x+tv)^T A(x+tv) - (x+tv)^T b$$

$$= q(x) + t v^T Ax - t v^T b + \frac{1}{2} t^2 v^T Av$$

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

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

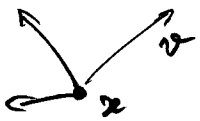
And:

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

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

$$= q(x) - \frac{(v^T(Ax-b))^2}{v^T Av}$$

Interpretation:



we look from x at different directions.

If x is a min of $q(x)$ then if we go along direction v, we should not be able to reduce function value any more.

$\Rightarrow$ $v^T(Ax-b) = 0 \quad \forall v$

$\Rightarrow$ $Ax = b$.

If $Ax = b$ then: $q(x+tv) = q(x) + \frac{1}{2} t^2 v^T Av > q(x)$ if $v \neq 0$

This gives an idea on how to find sol to $Ax = b$:

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

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

We see that optimal step size we can take is:

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

One example is <u>Steepest Descent</u> where $v^{(k)} = r^{(k)} = b - Ax^{(k)}$ = residual.

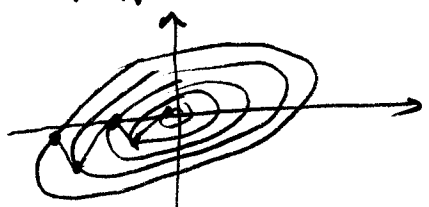
for $k = 1, 2, \ldots$

$r = b - Ax$

$t = \dfrac{(r, r)}{(r, Ar)}$

$x = x + tr$

$\rightsquigarrow$ not very efficient method.



typically "zig zags" to solution

<u>Conjugate gradient method</u>

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

(1) $$\min \tfrac{1}{2} x^T A x - b^T x$$
$$x \in x_0 + \operatorname{span}\{p_0, \ldots, p_{k-1}, r_k\}$$

where $r_k = -\nabla q(x_k) = b - Ax_k$

(1) $\Leftrightarrow$ (2) $\min \frac{1}{2} \hat{x}^T A \hat{x} - \hat{x}^T r_0$ $\qquad x = \hat{x} + x_0$

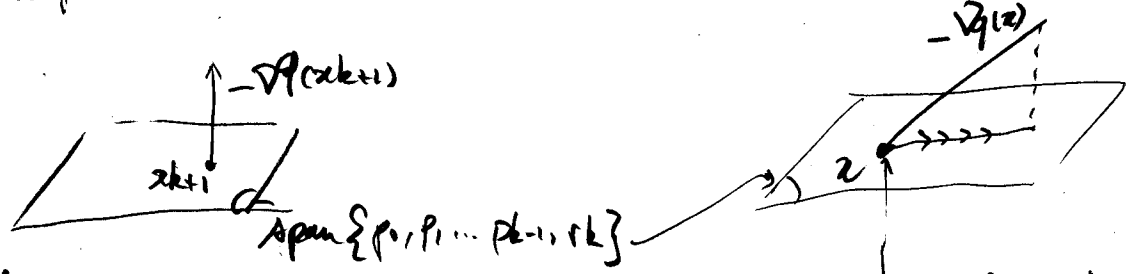
$\hat{x} \in \operatorname{span}\{p_0, \ldots, p_{k-1}, r_k\}$

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

$$(A\hat{x}_{k+1} - r_0)^T v = 0 \quad \forall v \in \text{span}\{p_0, p_1, \dots p_{k-1}, r_k\}$$

$$\underbrace{(Ax_{k+1} - b)}_{-\nabla q(x_{k+1})}{}^T v = 0$$

intuitively:



optimal since there is no descent direction in span{...}

non optimal one can find descent direc (and better pt)

Now in previous step $\hat{x}_k$ solves

$$\min \frac{1}{2}\hat{x}^T A\hat{x} - \hat{x}^T r_0 \quad \Rightarrow \quad p_j^T(b - Ax_k) = 0 \quad j = 0, \dots, k-1$$

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

We also have:

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

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

$\Rightarrow$ p_k is A- orthogonal to previous k directions.

note: $\langle u, v \rangle_A = u^T A v$ is an inner product when A is s.p.d.

We can get p_k using Gram-Schmidt orthogonalization:

$$p_0 = r_0$$
$$p_k = r_k - \sum_{j=0}^{k-1} \frac{r_k^T A p_j}{p_j^T A p_j} p_j \qquad (65)$$

Now that we know direction, we know step size too:

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

$\rightsquigarrow$ preliminary version where # ops grows linearly with iter # - storage

Let us look again at subspace we used:

$$\begin{aligned} \text{span}\{p_0, p_1, \dots p_{k-1}, r_k\} &= \text{span}\{p_0, \dots p_{k-1}, p_k\} \\ &= \text{span}\{r_0, r_1, \dots r_k\} \\ &= \mathcal{K}_{k+1}(A, r_0) \\ &= \text{span}\{r_0, Ar_0, A^2 r_0, \dots, A^k r_0\} \\ &= \text{Krylov subspace} \end{aligned}$$

why?
$$r_0 = b - A x_0 \in \mathcal{K}_1(A, r_0)$$
$$r_1 = b - A x_1 = b - A(x_0 + \alpha(b - A x_0)) \in \mathcal{K}_2(A, r_0)$$
etc...

why is this useful?

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

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

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

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

(GS) reduces to:

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

$= \beta_k$

storage: x_k, r_k, p_k, Ap_k
→ requires only knowledge of matvec.

classical formulation of CG:

$$\alpha_k = \frac{\|r_k\|^2}{p_k^T A p_k} \qquad \beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}$$

Convergence results: convergence depends on how many "clusters" of eigenvalues of A.

(S1) ||||||||||||| slow

(S2) ——— fast (~ 3 iter)

preconditioning: $Ax = b \iff MAx = Mb$

where M = invertible matrix
= preconditioner
= cheap approx. inv to A.
that puts (S1) into (S2).

pre. examples: incomplete LU (drop elements that are too small)
Jacobi, Gauss-Seidel or a few of these iterations
discretization based (solve x- dir then y- dir)