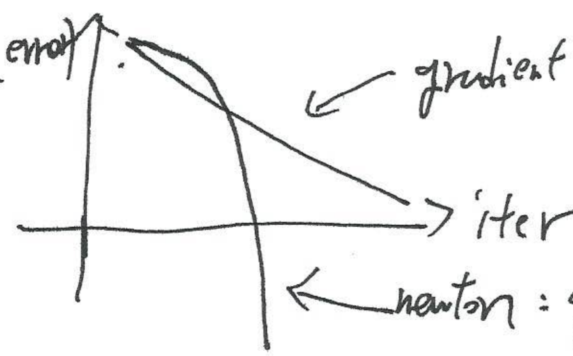
Gradient Descent: $x^{k+1} = x^k - \eta^k \nabla f(x^k) \leftarrow \mathbb{R}^n$

Newton method: $x^{k+1} = x^k - \eta^k \cdot (\underbrace{\nabla^2 f(x^k)}_{\text{Hessian}})^{-1} \cdot \nabla f(x^k)$

Time complexity: GD: usually (time) $\le O(n^2)$ space $O(n)$

Newton: space: $O(n^2)$
time: $O(n^3)$

Convergence:



← gradient descent, linear convergence

← newton: quadratic convergence.

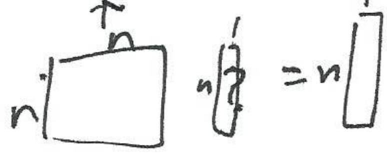
How to scale Newton method to large n?

- "Inexact" Newton method:

compute $p = \nabla^2 f(x)^{-1} \cdot \nabla f(x)$ approximately (inexactly).

$\Updownarrow$

$\nabla^2 f(x) \cdot p = \nabla f(x)$



- *Quasi Newton method: Use B to approximate Hessian matrix.

$p = B^{-1} \cdot \nabla f(x)$

$B \approx \nabla^2 f(x)$

Chapter 5. Conjugate Gradient method (CG)

Goal: Solve $Ax=b$

$A: R^{n\times n}$ positive definite matrix.

$b: R^{n}$

Optimal solution. $x^*=A^{-1}b$

Equivalently, solving the minimization problem:

$$\min_{x} \underbrace{\frac{1}{2}x^TAx - b^Tx}_{f(x)} \Rightarrow Ax^*=b$$

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

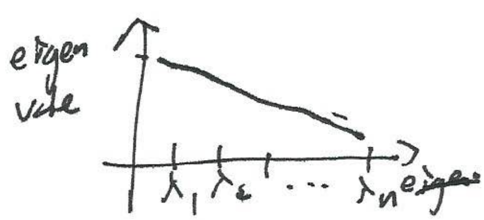
~~$Ax-b = \nabla f(x^*)=0$~~

CG:

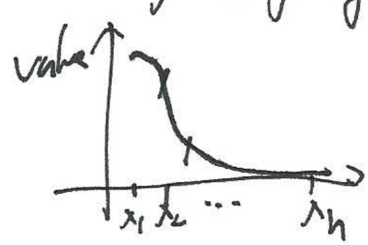
- Only require computing Ap for some p at each iteration.
- Terminate in n iterations.

Usually terminate early (depends on eigen-values)

eigenvalue dist



large eigengap



5.1 Conjugate directions:

Def: A set of vectors (nonzero)

$$\{p_0, p_1, p_2, \cdots, p_{n-1}\}$$

is called "conjugate" with respect to A.

(A symmetric, pd) iff

$$p_i^T A p_j = 0 \quad \forall i, j \le n, \; i \neq j$$

Claim: Given a set of conjugate vectors

$$\{p_0, p_1, \cdots p_{n-1}\} \quad (\text{wrp. } A)$$

we can solve $Ax = b$ by:

Alg I: $x^0 = 0$

For $k = 0, 1, \cdots, n-1$

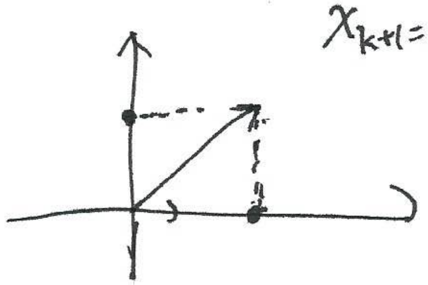
$r^k = Ax^k - b$ ← residual.

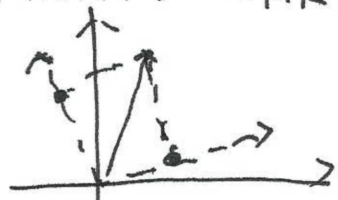
$\alpha_k = -(r^k)^T p_k \,/\, p_k^T A p_k$ ← project residual to p_k

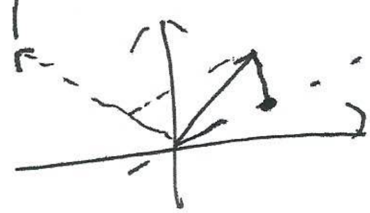
$x_{k+1} = x_k + \alpha_k p_k$ ← update x

$\downarrow$

$$x_{k+1} = \alpha_0 p_0 + \alpha_1 p_1 + \cdots + \alpha_k p_k + x_0$$







Thm: The x^* computed by Alg I is the solution of $Ax=b$.

pf: $x^* - x^0 = \sigma_0 p_0 + \sigma_1 p_1 + \cdots + \sigma_{n-1} p_{n-1}$

for some $\sigma_0, \sigma_1, \cdots, \sigma_{n-1}$

$$\cancel{p_k^T(x^*} \; p_k^T A(x^* - x^0) = \sigma_0 \underbrace{p_k^T A p_0}_{=0} + \sigma_1 \underbrace{p_k^T A p_1}_{=0} + \cdots + \sigma_{n-1} p_k^T A p_{n-1}$$

$$= \sigma_k p_k^T A p_k \quad \checkmark$$

$$\boxed{\Rightarrow \sigma_k = p_k^T A(x^* - x^0) / p_k^T A p_k}$$

At step k, $x_k = x_0 + \alpha_0 p_0 + \alpha_1 p_1 + \cdots + \alpha_{k-1} p_{k-1}$

so $p_k^T A(x_k - x_0) = 0$

$$\Rightarrow p_k^T A(x^* - x_0) = p_k^T A(x^* - x_k + x_k - x_0)$$

$$= p_k^T A \cdot (x^* - x_k)$$

$$\cancel{= p_k^T A (b - A x_k)}$$

$$= p_k^T (\underbrace{b - A x_k}_{-r_k})$$

$\Rightarrow \sigma_k = \alpha_k$

- Also,
$$\begin{aligned} r_{k+1} &= Ax_{k+1} - b \\ &= A(x_k + \alpha_k p_k) - b \\ &= Ax_k - b + \alpha_k A p_k = \underbrace{r_k + \alpha_k A p_k} \end{aligned}$$

Thm: If $\{x_k\}$ is generated by Alg I, then

$$r_k^T p_i = 0 \quad \forall i = 0, 1, \cdots, k-1$$

pf: By induction. if this is true for $1, \cdots k$,

for $k+1$,

$$\begin{aligned} \forall i < k, \quad r_{k+1}^T p_i &= (r_k + \alpha_k A p_k)^T p_i \\ &= \underset{0}{\underset{\parallel}{r_k^T p_i}} + \alpha_k \underset{0}{\underset{\parallel}{p_k^T A p_i}} \\ &= 0 \end{aligned}$$

$$r_{k+1}^T p_k = r_k^T p_k + \alpha_k p_k^T A p_k = r_k^T p_k - r_k^T p_k = 0$$

#

5.2 Conjugate Gradient Method.

($Ax=b$ or $\min_x \underbrace{\frac{1}{2}x^TAx - b^Tx}_{f(x)}$)

$\nabla f(x) = Ax-b$
$= r$

Alg: CG method

Given x_0

set $r_0 = Ax_0 - b$, $p_0 = -r_0$

For $k=0,1,\cdots,n-1$

$\alpha_k = -r_k^Tp_k / p_k^TAp_k$

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

$r_{k+1} = r_k + \alpha_k A p_k$ ← check residual ?= 0

$$\boxed{\begin{aligned} \beta_{k+1} &= (r_{k+1}^T A p_k) / p_k^T A p_k \\ p_{k+1} &= -r_{k+1} + \beta_{k+1} p_k \end{aligned}}$$ ← find new direction p_{k+1}

end

Want to get p_{k+1} st $p_k^TAp_{k+1} = 0$

assume $p_{k+1} = -\underbrace{r_{k+1}}_{\substack{\uparrow \\ \text{negative gradient.}}} + \beta_{k+1} p_k$.

so $p_k^TAp_{k+1} = -p_k^TAr_{k+1} + \beta_{k+1} p_k^TAp_k = 0$

$\beta_{k+1} = p_k^TAr_{k+1} / p_k^TAp_k$

Why $p_i^TAp_{k+1} = 0 \quad \forall i=0,1,\cdots,k-1$

Thm: The iterates $\{x_k\}$ generated by CG has the following property:

$$\text{span}\{r_0, r_1 \cdots r_k\} = \text{span}\{r_0, Ar_0, A^2r_0, \cdots A^k r_0\}$$
$$= \text{span}\{p_0, p_1, p_2, \cdots, p_k\}$$

kylov space

pf: By induction, assume this is true up to k.

For $k+1$

① span $\{r_0, r_1, \cdots, r_k; r_{k+1}\} = \text{span}\{r_0, Ar_0, \cdots, A^{k+1}r_0\}$

$$r_{k+1} = r_k + \alpha_k A p_k$$
$$A p_k \in \text{span}\{Ar_0, A^2r_0, \cdots, A^{k+1}r_0\}$$
$$r_{k+1} \in \text{span}\{r_0, Ar_0, \cdots, A^{k+1}r_0\}$$
$$A^{k+1}r_0 = A(A^k r_0) \in \text{span}\{Ap_0, Ap_1, \cdots, Ap_k\}$$

$$\alpha_k A p_k = r_{k+1} - r_k$$
$$\Rightarrow A p_k = r_{k+1}/\alpha_k - r_k/\alpha_k \in \text{span}\{r_0, \cdots r_{k+1}\}$$

$$A^{k+1} r_0 \in \text{span}\{r_0, \cdots, r_{k+1}\}$$
$$\Rightarrow \{r_0, r_1, \cdots r_{k+1}\} = \{r_0, Ar_0, \cdots, A^{k+1}r_0\}$$

② $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$.

$$\text{span}\{p_0, \cdots, p_{k+1}\} = \text{span}\{r_0, \cdots, r_k, r_{k+1}\}$$

Thm: For CG algorithm, $p_i^T A p_j = 0 \quad \forall i \neq j$

pf: We already know $p_k^T A p_{k+1} = 0$.

Now we want to show $p_i^T A p_{k+1} = 0 \quad \forall i = 0, \cdots, k-1$

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

$$= -p_i^T A r_{k+1} + \beta_{k+1} \underbrace{p_i^T A p_k}_{=0}$$

$$= -p_i^T A r_{k+1}$$

$$p_i \in \operatorname{span}\{r_0, Ar_0, A^2 r_0, \cdots, A^i r_0\}$$

$$A p_i \in \operatorname{span}\{Ar_0, A^2 r_0, \cdots, A^{i+1} r_0\}$$

$$\in \operatorname{span}\{r_0, r_1, \cdots, r_{i+1}\}$$

$$= \operatorname{span}\{p_0, p_1, \cdots, p_{i+1}\}$$

$$p_i^T r_{k+1} = 0 \quad \forall i \leq k$$

$$\Rightarrow p_i^T A p_{k+1} = 0 \quad \#$$

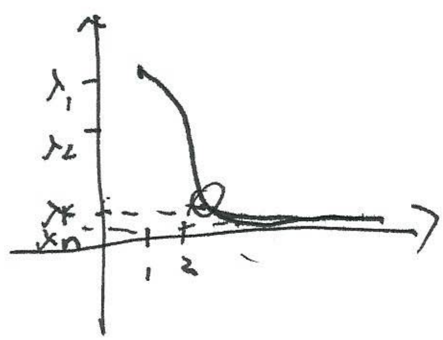
5.3 Rate of Convergence.

Thm: If A has eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n$$

Then $\|x_{k+1} - x^*\|_A^2 \leq \left(\frac{\lambda_k - \lambda_n}{\lambda_k + \lambda_n}\right)^2 \|x_0 - x^*\|_A^2$

$$\|v\|_A^2 = v^T A v.$$



5.4 Combine CG with Newton