Conjugate Gradient method

Sunday, March 21, 2021   8:49 AM

Fast algo. for smooth unconstrained optim.

- grad. descent, Nesterov
- Newton (see Boyd + Vandenberghe, self-concordant analysis)
- Nonlinear conjugate gradient
- Quasi-Newton (BFGS, L-BFGS)

• (linear) Conjugate Gradient (these notes follow Nocedal + Wright text and Anne Greenbaum's monograph)

History: Hestenes, Stiefel '50s (nonlinear: Fletcher + Reeves '60s)

CG solves (approximately) $Ax=b$ if $A \succ 0$, A is $n \times n$ (naively: $O(n^3)$ flops)

Motivation / link to optimization:

$$\min_x \underbrace{\tfrac{1}{2}\|\tilde{A}x-\tilde{b}\|^2}_{\varphi(x)}$$

$$\varphi(x) = \tfrac{1}{2} x^T \underbrace{\tilde{A}^T\tilde{A}}_{A} x - \underbrace{\tilde{b}^T\tilde{A}}_{b^T} x + \tfrac{1}{2}\tilde{b}^T\tilde{b}$$

$$= \tfrac{1}{2} x^T A x - b^T x + \text{const.}$$

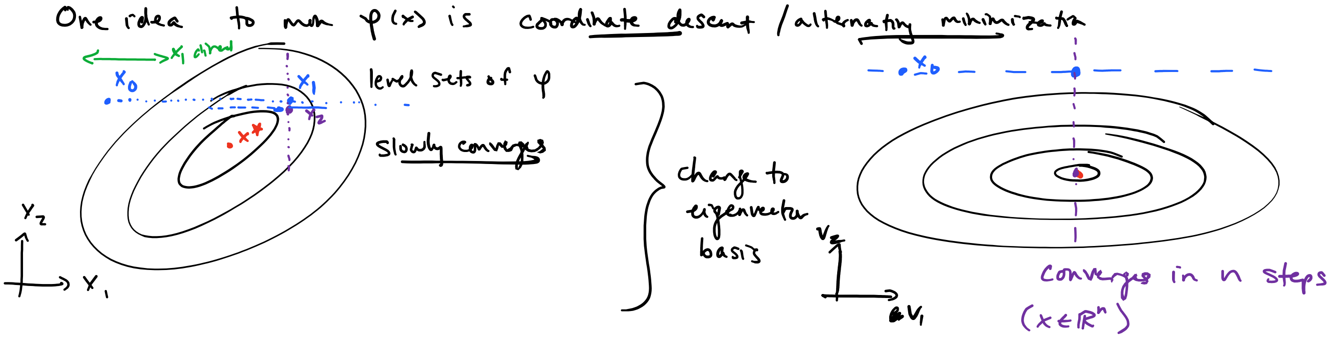
$\succeq 0$ and $A \succ 0$ if $m>n$ + full rank

set $\nabla\varphi(x)=0$ and solve,

$\nabla\varphi(x) = Ax-b$, so find x st $Ax=b$

⚠ If you do want to solve least-squares, don't form $A=\tilde{A}^T\tilde{A}$. Instead use LSQR (CG-variant)

One idea to min $\varphi(x)$ is coordinate descent / alternating minimization



Conjugate Directions

$\{p_i\}$ are conjugate directions if they are A-orthogonal i.e., $\langle p_i, A p_j \rangle = 0$ if $i \neq j$ (Fact: $\langle x, y \rangle_A = x^T A y$ is an inner-product)

Notation: $\underbrace{\langle x | A | y \rangle}_{\langle x, y \rangle_A} = \langle x, Ay \rangle = \langle Ax, y \rangle$

$= x^T A y = (Ax)^T y$ ($A = A^T$ or $A = A^*$)

so, $\langle p_i | A | p_j \rangle = 0$ if $i \neq j$

If we have $\{p_i\}_{i=0}^{n-1}$ A-orthog, it's a basis.

Ex: p_i are eigenvectors: if A is symm, its eigenvectors are orthog.

$\langle p_i | \underbrace{A | p_j}_{\lambda_j p_j} \rangle = \lambda_j \langle p_i | p_j \rangle = 0$ if $i \neq j$

Goal: find $\{p_i\}$ more cheaply than eigenvectors

Conjugate Direction method (abstract): Assume $\{p_i\}_{i=0}^{n-1}$ are conjugate directions.

(*) $x_{k+1} = x_k + \alpha_k p_k$ (like coordinate descent)

α_k solves $\min_\alpha \varphi(x_k + \alpha p_k)$ (exact linesearch)

closed form

$\varphi(x) = \frac{1}{2} \langle x | A | x \rangle - \langle b | x \rangle$

$$\alpha_k = -\frac{\langle A x_k - b \,|\, p_k \rangle}{\langle p_k | A | p_k \rangle} \qquad r_k := A x_k - b$$

Thm: $x_n = x^*$, i.e., the true minimizer.

proof: $\{p_i\}$ is a basis, $x^* - x_0 = \sum_{i=0}^{n-1} \sigma_i p_i$ coeff.

$$p_k^T A (x^* - x_0) = \sum_{i=0}^{n-1} \sigma_i \langle p_k | A | p_i \rangle$$

$$= \sigma_k \langle p_k | A | p_k \rangle$$

$$\text{so } \sigma_k = \underline{\frac{\langle p_k | A | x^* - x_0 \rangle}{\langle p_k | A | p_k \rangle}} \quad \forall k = 0, 1, \dots, n-1$$

Also,

$x_k = x_0 + \alpha_0 p_0 + \alpha_1 p_1 + \dots + \alpha_{k-1} p_{k-1}$ via (*) recursively

$$x_k - x_0 = \sum_{i=0}^{k-1} \alpha_i p_i$$

$$p_k^T A (x_k - x_0) = \sum_{i=0}^{k-1} \alpha_i \langle p_k | A | p_i \rangle = 0$$

$$\text{so } \underline{\langle p_k | A | x_k - x_0 \rangle = 0}$$

$Ax^* = b$

claim $\underline{\underline{r_k}} = \dfrac{\langle p_k | A | \overbrace{x^*}^{=b} - x_k \rangle}{\langle p_k | A | p_k \rangle} = \dfrac{\langle p_k | b - Ax_k \rangle}{\cdots} = \underline{\underline{\alpha_k}}$

$\Rightarrow x_n = x^*$ $\square$

Facts via any CD method,

- $r_{k+1} = r_k + \alpha_k A p_k$
- $\langle r_k, p_i \rangle = 0 \quad i < k$
- x_k minimizes φ over $K(r_0, k-1) = \text{span}\{ r_0, Ar_0, \dots A^{k-1} r_0 \}$ ← Krylov subspace

  i.e., $K(r_0, n-1) = \mathbb{R}^n$
- $\langle r_k, r_i \rangle = 0 \quad i < k$
- $p_k, r_k \in K(r_0, k)$

Conjugate Gradient = C.D. method that cheaply compute p_k

- x_0 anything, $r_0 = Ax_0 - b$, $p_0 = -r_0$
- Iterate

$$p_k = -r_k + \beta_k \cdot p_{k-1}, \quad \beta_k = \frac{\langle r_k | A | p_{k-1} \rangle}{\langle p_{k-1} | A | p_{k-1} \rangle}$$

$$x_{k+1} = x_k + \alpha_k p_k \qquad \text{chosen so } \langle p_k | A | p_{k-1} \rangle = 0$$

$$\alpha_k = \frac{-\langle r_k | p_k \rangle}{\langle p_k | A | p_k \rangle}$$

Magic: $\langle p_k | A | p_i \rangle = 0 \quad \forall i \leq k-1$ ($i = k-1$ is by design)

$i < k$ "are for free"

(see Nocedal + Wright's text)

Cost: one matrix-vector multiply per step

Convergence Result (NW book)

$$\| x_k - x^* \|_A \leq 2 \cdot \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \| x_0 - x^* \|_A$$

$\kappa = \kappa(A)$ is the condition # $= \dfrac{\text{max sing. value}}{\text{min. sing. value}}$

Practical issues exist (eg., stability/roundoff), not discussed here.

- Nonlinear Conjugate Gradient

Linear CG: $\min_x \ \varphi(x) := \frac{1}{2}\langle x|A|x\rangle - \langle b|x\rangle$, $\nabla\varphi(x) = Ax - b = r$

i.e., solve $Ax = b$ ← hence name "linear"

Nonlinear CG: $\min_x \ \varphi(x)$, $\varphi(x)$ isn't quadratic, $\nabla\varphi$ isn't linear

Linear: $x_{k+1} = x_k + \alpha_k p_k$, $\alpha_k = \operatorname{argmin}_\alpha \varphi(x_k + \alpha p_k)$ known in closed form

$p_{k+1} = -\underbrace{r_{k+1}}_{=\nabla\varphi(x_{k+1})} + \beta_{k+1} p_k$, $\beta_{k+1} = \dfrac{\langle r_{k+1}|A|p_k\rangle}{\langle p_k|A|p_k\rangle}$

Nonlinear: $x_{k+1} = x_k + \alpha_k p_k$, $\alpha_k \approx \operatorname{argmin}_\alpha \varphi(x_k + \alpha p_k)$ Not in closed form.

Approximate via linesearch.

⚠ Sensitive! Need accurate linesearch

$p_{k+1} = -\underbrace{\nabla\varphi(x_{k+1})}_{\nabla\varphi_{k+1} \text{ for short}} + \beta_{k+1} p_k$ } Somewhat like Nesterov / momentum methods

β_{k+1} has many choices, all of which reduce to $\dfrac{\langle r_{k+1}|A|p_k\rangle}{\langle p_k|A|p_k\rangle}$ if φ quadratic

Ex: $\beta_{k+1}^{FR} = \dfrac{\|\nabla\varphi_{k+1}\|^2}{\|\nabla\varphi_k\|^2}$

others too (Polyak-Ribière ...)

Comments on non-linear CG

- can be fast, but limited to unconstrained optimization
  - Doesn't play well w/ constraints, even simple $x \geq 0$ ones
  - i.e., if φ is quadratic, CG is essentially optimal among all 1st order methods. So φ "almost quadratic", non-linear CG might do very well
- Finicky w/ α, β terms. Hager + Zhang have many advanced versions, some require many (eg. 20) parameters.
- Global convergence theory isn't great
- Nemirovsky + Yudin proved there exist reasonable cvx functions for which non-linear CG is slower than gradient descent

- IMO, quasi-Newton methods about as fast, and more stable (simpler)

  (ie., if you tuned all non-linear CG param. just right, it might be faster, but parameters are problem dependent!)