Lecture 5: Conjugate Direction Methods, Convergence Analysis, Practicalities

Prof. Krishna R. Pattipati
Dept. of Electrical and Computer Engineering
University of Connecticut
Contact: krishna@engr.uconn.edu (860) 486-2890

ECE 6437
Computational Methods for Optimization

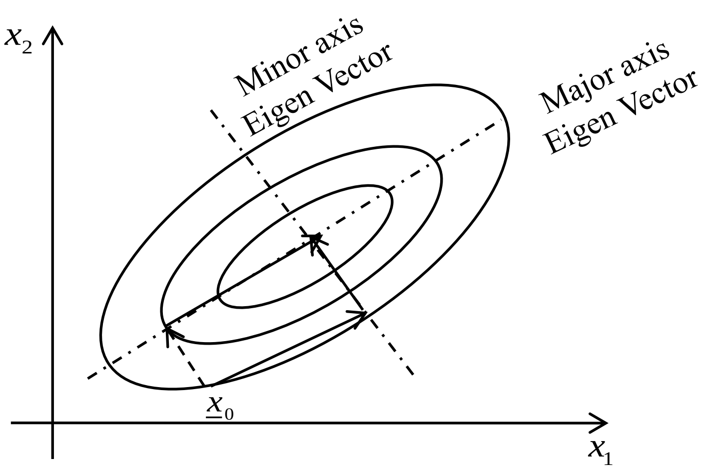
Fall 2009
September 29, 2009

Outline of Lecture 5

- What are Conjugate Directions?
- Conjugate Direction Methods
- Conjugate Gradient Methods
- Convergence Analysis
- Partial Conjugate Gradient Method
- Application to Non-quadratic Problems
- Pre-conditioned Conjugate Gradient (Scaling)

Conjugate Directions - 1

- We will consider quadratic fⁿ : $f(\underline{x})=\frac{1}{2}\underline{x}^T Q\underline{x}-\underline{b}^T\underline{x}+c$ to develop these schemes.
  - If *OK* on a quadratic $f(\underline{x})$, probably *OK* near minimum of general $f(\underline{x})$.
  - Objective is to find "optimal" directions $\{\underline{d}_k\}$ to minimize $f(\underline{x})$ in *as few steps as possible*.
- Recall that for a quadratic function
  - Steepest descent $\Rightarrow \infty$ no. of steps except when $Q=\beta I \Rightarrow$ *SLOW*
  - Newton $\Rightarrow$ 1 step, but needs Hessian $\Rightarrow$ *FAST* but Expensive
  - Conjugate direction & quasi-Newton $\Rightarrow$ *n* steps, $n=dim(\underline{x}) \Rightarrow$ *QUADRATIC TERMINATION*
- Consider the contour curves of $f(\underline{x})$



- What if we pick eigen vectors of Q as search directions (2-dimensional)

$$\underline{d}_0 = \pm\underline{\xi}_0$$
$$\underline{d}_1 = \pm\underline{\xi}_1 \perp \underline{d}_0$$

Then, can obtain $\underline{x}^*$ in two steps

(a) pick $\underline{d}_0=\pm\underline{\xi}_0 \ni \underline{d}_0^T\underline{g}_0<0$, optimize α_0

(b) pick $\underline{d}_1=\pm\underline{\xi}_1 \ni \underline{d}_1^T\underline{g}_1<0$, optimize α_1

NOTE: We can change the order of (a) and (b).

Conjugate Directions - 2

- But the evaluation of eigen vectors is a complex task ≈ O (20 n^3). Can we get a property of $\underline{d}_1$ in terms of $\underline{d}_0$ (Note: $\underline{d}_0 \perp \underline{d}_1$ in the eigen vector case).
- Suppose take $\underline{d}_0 = -\underline{g}_0$ and optimize over α_o and get $\underline{x}_1$. What is the next best direction $\underline{d}_1$?

$$\underline{d}_0^T Q\, \underline{d}_1 = 0 \quad \text{(or) } \underline{d}_0 \text{ is } Q\text{-orthogonal to } \underline{d}_1$$

NOTE: Eigen vectors of Q are Q-orthogonal (since they are ⊥)

- *Def^n^* : A set of vectors $\{\underline{d}_i\}_{i=0}^{k}, \underline{d}_i \in R^n$ are Q-orthogonal (or mutually Q-conjugate) if $\underline{d}_i^T Q \underline{d}_j = \delta_{ij}, \forall\, i, j = 1,2,\ldots,k$
  - Property: If $Q > 0$ and the set $\{\underline{d}_i\}_{i=0}^{n\text{-}1}$ are Q-orthogonal, then $\{\underline{d}_i\}$ are linearly independent
    - *Proof:* If not, $\underline{d}_k = \alpha_1 \underline{d}_1 + \alpha_2 \underline{d}_2 + \ldots\ldots + \alpha_{k-1}\underline{d}_{k-1}$

$$\underline{d}_k^T Q \underline{d}_k = \alpha_1 \underline{d}_k^T Q \underline{d}_1 + \alpha_2 \underline{d}_2^T Q \underline{d}_k + \ldots\ldots + \alpha_{k-1} \underline{d}_k^T Q \underline{d}_{k-1} = 0$$

$\Rightarrow Q$ is not PD - a *contradiction*

Orthogonal	$\Rightarrow$	Q-orthogonal
Q-orthogonal	$\not\Rightarrow$	Orthogonal

Conjugate Direction Method (CD) - 1

- Basic idea of CD method
  - Given a collection of mutually Q-conjugate directions $\{\underline{d}_k\}$, k=0,1,2, …, n-1, the conjugate direction method generates the sequence $\{\underline{x}_1, \underline{x}_2, ..., \underline{x}_n\}$ via

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k \quad (\underline{x}_0 \text{ known})$$

where $\underline{x}_0$ is a given vector in R^n & $\alpha_k = \arg\min_{\alpha} f\left(\underline{x}_k + \alpha \underline{d}_k\right)$

- Key property of CD method
  - If $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$ are generated from the conjugate direction method, then

$$\underline{g}_{k+1}\underline{d}_i = 0 \;\; \forall \; i=0,1,2, \ldots, k \qquad (1)$$

- *Proof:* Since $\alpha_k = \arg\min_{\alpha} f\left(\underline{x}_k + \alpha \underline{d}_k\right)$

$$\left.\frac{\partial f(\underline{x}_k + \alpha \underline{d}_k)}{\partial \alpha}\right|_{\alpha=\alpha_k} = 0 \Rightarrow \nabla \underline{f}^T(\underline{x}_{k+1})\underline{d}_k = \underline{g}_{k+1}^T \underline{d}_k = 0 \;\; \forall k$$

- Note that this property is true for *any* generalized gradient method
- To prove (1), we need to prove its validity for $i = 0,1,2,\ldots,k-1$

$$\underline{g}_{k+1}^T \underline{d}_i = \left(Q\underline{x}_{k+1} - \underline{b}\right)^T \underline{d}_i = \left(\underline{x}_{i+1} + \sum_{j=i+1}^{k} \alpha_j \underline{d}_j\right)^T Q\underline{d}_i - \underline{b}^T \underline{d}_i$$

$$= \left(Q\underline{x}_{i+1} - \underline{b}\right)^T \underline{d}_i = \underline{g}_{i+1}^T \underline{d}_i = 0$$

Conjugate Direction Method (CD) -2

- How to create Q- orthogonal (or Q-conjugate) directions ?
  - $\exists$ many choices for $\{\underline{d}_i\}$. For example, $\underline{d}_i = \underline{\xi}_i$ is one possibility.
  - We can create $\{\underline{d}_i\}$ from any linearly independent set of vectors $\{\underline{p}_0, \underline{p}_1, \ldots, \underline{p}_{n-1}\}$. The process is termed GRAM SCHMIDT ORTHOGONALIZATION

$$\underline{d}_0 = \underline{p}_0$$

$$\underline{d}_i = \underline{p}_i + \sum_{j=0}^{i-1} l_{ij} \underline{d}_j$$

- From Q-orthogonality, need $\underline{d}_i^T Q \underline{d}_k = \underline{p}_i^T Q \underline{d}_k + \sum_{j=0}^{i-1} l_{ij} \underline{d}_j^T Q \underline{d}_k = 0\,;\; k = 0,1,2,\ldots,i-1$

$$\Rightarrow \boxed{l_{ik} = \frac{-\underline{p}_i^T Q \underline{d}_k}{\underline{d}_k^T Q \underline{d}_k} \quad \forall \quad \begin{matrix} i=1,2,\ldots,n\text{-}1 \\ k=1,2,\ldots,i\text{-}1 \end{matrix}}$$

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -l_{10} & 1 & 0 & 0 \\ -l_{20} & -l_{21} & 1 & 0 \\ -l_{n-1,0} & -l_{n-1,1} & 0 & 1 \end{bmatrix}$$

$$\boxed{\underline{d}_i = \underline{p}_i - \sum_{j=0}^{i-1} \left(\frac{\underline{p}_i^T Q \underline{d}_j}{\underline{d}_j^T Q \underline{d}_j} \right) \underline{d}_j}$$

or $(\underline{d}_0, \underline{d}_1, \ldots, \underline{d}_{n-1}) L^T = (\underline{p}_0, \underline{p}_1, \ldots, \underline{p}_{n-1})$

$$L^T = \begin{bmatrix} 1 & -l_{10} & -l_{20} & -l_{n-1,0} \\ 0 & 1 & -l_{21} & -l_{n-1,1} \\ 0 & 0 & 1 & . \\ 0 & 0 & & 1 \end{bmatrix}$$

- Indeed $\mathbb{R}\left(\underline{p}_0, \underline{p}_1, \ldots, \underline{p}_{n-1}\right) = \mathbb{R}\left(\underline{d}_0, \underline{d}_1, \ldots, \underline{d}_{n-1}\right)$

Conjugate Gradient Method (CG) - 1

Ultimate goal:

- Generate a set of Q- orthogonal directions $\{\underline{d}_i\}$ without the explicit knowledge of Q. A method that accomplishes this is the Conjugate Gradient (CG) method that takes $\underline{p}_i = -\underline{g}_i$

$$\underline{d}_0 = -\underline{g}_0; \quad \underline{x}_1 = \underline{x}_0 + \alpha_0 \underline{d}_0$$

$$\underline{d}_1 = -\underline{g}_1 + \frac{\underline{g}_1^T Q \underline{d}_0}{\underline{d}_0^T Q \underline{d}_0} \cdot \underline{d}_0$$

- Note that $\underline{g}_1 - \underline{g}_0 = Q(\underline{x}_1 - \underline{x}_0) = \alpha_0 Q \underline{d}_0$

$$\underline{d}_1 = -\underline{g}_1 + \frac{\underline{g}_1^T(\underline{g}_1 - \underline{g}_0)}{\underline{d}_0^T(\underline{g}_1 - \underline{g}_0)} \cdot \underline{d}_0 \quad \text{(or)} \quad \underline{d}_1 = -\underline{g}_1 + \beta_1 \underline{d}_0$$

By repeating the process for $j = 0,1,2,\ldots,k$, we have

$$\underline{d}_k = -\underline{g}_k + \sum_{j=0}^{k-1} \frac{\underline{g}_k^T Q \underline{d}_j}{\underline{d}_j^T Q \underline{d}_j} \cdot \underline{d}_j = -\underline{g}_k + \sum_{j=0}^{k-1} \frac{\underline{g}_k^T(\underline{g}_{j+1} - \underline{g}_j)}{\underline{d}_j^T(\underline{g}_{j+1} - \underline{g}_j)} \cdot \underline{d}_j$$

- We know: (1) $\mathbb{R}\{\underline{g}_0, \underline{g}_1, \ldots, \underline{g}_{k-1}\} = \mathbb{R}\{\underline{d}_0, \underline{d}_1, \ldots, \underline{d}_{k-1}\} \Rightarrow \underline{d}_j = \sum_{i=0}^{j} \alpha_{ji} \underline{g}_i$

  (2) $\underline{g}_k^T \underline{d}_j = 0; \quad \forall i = 0,1,2,\ldots,k-1$

Conjugate Gradient Method (CG) -2

(1) & (2) $\Rightarrow$ $\underline{g}_k^T \underline{g}_i = 0 \quad \forall i = 0,1,2,\ldots,k-1$

(or) $\underline{d}_k = -\underline{g}_k + \beta_k \underline{d}_{k-1}$

where $\beta_k = \dfrac{\underline{g}_k^T(\underline{g}_k - \underline{g}_{k-1})}{\underline{d}_{k-1}^T(\underline{g}_k - \underline{g}_{k-1})}$ Sorensen-Wolfe (SW) form

- Since $\underline{d}_{k-1}^T(\underline{g}_k - \underline{g}_{k-1}) = (-\underline{g}_{k-1} + \beta_{k-1}\underline{d}_{k-2})^T(\underline{g}_k - \underline{g}_{k-1}) = \underline{g}_{k-1}^T \underline{g}_{k-1}$

$\Rightarrow \beta_k = \dfrac{\underline{g}_k^T(\underline{g}_k - \underline{g}_{k-1})}{\underline{g}_{k-1}^T \underline{g}_{k-1}}$ Polak-Ribiere-Poljak (PRP) form

Since $\underline{g}_k^T \underline{g}_{k-1} = 0 \Rightarrow \beta_k = \dfrac{\underline{g}_k^T \underline{g}_k}{\underline{g}_{k-1}^T \underline{g}_{k-1}}$ **Fletcher-Reeves (FR)form**

- Remarks:
  1) To compute $\underline{d}_k$, need only $\underline{d}_{k-1}$, $\underline{g}_{k-1}$ and $\underline{g}_k$

  $$\underline{d}_k = -\underline{g}_k + \beta_k \underline{d}_{k-1} = \Im(\underline{g}_0, \underline{g}_1, \ldots, \underline{g}_k)$$

  2) Various forms will be useful for NL function minimization. We will discuss some modifications to these later.

Quadratic Termination Property - 1

- Quadratic termination property
  - Note that $Q\underline{x}^*=\underline{b}$
  - At step k
  - Also $$\underline{x}_k = \underline{x}_0 + \sum_{i=0}^{k-1} \alpha_i \underline{d}_i$$
    $$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k$$
  - To find step size α_k, note that $\underline{g}_{k+1}^T \underline{d}_k = 0$
    $$\Rightarrow [Q(\underline{x}_k - \underline{x}^* + \alpha_k \underline{d}_k)]^T \underline{d}_k = 0$$
    (or) $$\alpha_k = \frac{-\underline{g}_k^T \underline{d}_k}{\underline{d}_k^T Q \underline{d}_k}$$
  - Note that
    $$\underline{g}_k = \underline{g}_0 + Q\sum_{i=0}^{k-1} \alpha_i \underline{d}_i$$
    Since $\underline{d}_i^s$ are Q-conjugate, we have
    $$\alpha_k = \frac{-\underline{g}_0^T \underline{d}_k}{\underline{d}_k^T Q \underline{d}_k}$$

Quadratic Termination Property - 2

- $$\underline{x}_n = \underline{x}_0 + \sum_{k=0}^{n-1} \alpha_k \underline{d}_k$$

$$= \underline{x}_0 - \sum_{k=0}^{n-1} \frac{\underline{g}_0^T \underline{d}_k}{\underline{d}_k^T Q \underline{d}_k} \underline{d}_k$$

$$= \underline{x}_0 - \sum_{k=0}^{n-1} \frac{\underline{d}_k^T Q \underline{x}_0}{\underline{d}_k^T Q \underline{d}_k} \underline{d}_k + \sum_{k=0}^{n-1} \frac{\underline{d}_k^T Q \underline{x}^*}{\underline{d}_k^T Q \underline{d}_k} \underline{d}_k$$

$$= \underline{x}_0 - \underline{x}_0 + \underline{x}^* \quad \text{since } \underline{x}_0 = \sum_{k=0}^{n-1} \gamma_k \underline{d}_k \Rightarrow \gamma_k = -\frac{\underline{d}_k^T Q \underline{x}_0}{\underline{d}_k^T Q \underline{d}_k}$$

$$= \underline{x}^*$$

Convergence analysis for $f(\underline{x}) = \frac{1}{2}(\underline{x} - \underline{x}^*)^T Q (\underline{x} - \underline{x}^*)$

- Consider algorithms of the form

$$\underline{x}_1 = \underline{x}_0 + \gamma_{00}\, \underline{g}_0$$

$$\underline{x}_2 = \underline{x}_1 + \alpha_1 \underline{d}_1 = \underline{x}_0 + \gamma_{10}\, \underline{g}_0 + \gamma_{11} \underline{g}_1$$

$$\underline{x}_{k+1} = \underline{x}_0 + \gamma_{k0}\, \underline{g}_0 + \gamma_{k1}\, \underline{g}_1 + \dots + \gamma_{kk}\, \underline{g}_k$$

or $$\underline{x}_{k+1} = \underline{x}_0 + \sum_{l=0}^{k} \gamma_{kl}\, \underline{g}_l$$

Convergence Analysis -1

- We can express g_i in terms of $\underline{x}_i - \underline{x}^*$ as follows

$$\underline{g}_i = Q(\underline{x}_i - \underline{x}^*) = Q(\underline{x}_i - \underline{x}_{i-1}) + Q(\underline{x}_{i-1} - \underline{x}^*)$$

or $\boxed{\underline{g}_i = \underline{g}_{i-1} + \alpha_{i-1} Q \underline{d}_{i-1}}$, $\underline{d}_0 = -\underline{g}_0$

$$\Rightarrow \underline{g}_1 = \underline{g}_0 - \alpha_{i-1} Q \underline{g}_0 = Q(\underline{x}_0 - \underline{x}^*) - \alpha_{i-1} Q^2 (\underline{x}_0 - \underline{x}^*)$$

- Similarly $\underline{g}_i = \beta_{i1} Q(\underline{x}_0 - \underline{x}^*) + \beta_{i2} Q^2 (\underline{x}_0 - \underline{x}^*) + \ldots + \beta_{i,i+1} Q^{i+1} (\underline{x}_0 - \underline{x}^*)$

- So,

$$\underline{x}_{k+1} - \underline{x}^* = \underline{x}_0 - \underline{x}^* + \sum_{l=0}^{k} \gamma_{kl} \sum_{j=1}^{l+1} \beta_{lj} Q^j (\underline{x}_0 - \underline{x}^*)$$

$$= (\underline{x}_0 - \underline{x}^*) + \sum_{j=1}^{k+1} \left(\sum_{l=j-1}^{k} \gamma_{kl} \beta_{lj} \right) Q^j (\underline{x}_0 - \underline{x}^*)$$

$$= (\underline{x}_0 - \underline{x}^*) + \sum_{j=1}^{k+1} \xi_{kj} Q^j (\underline{x}_0 - \underline{x}^*)$$

Final result $(\underline{x}_{k+1} - \underline{x}^*) = \left[I + Q P_k(Q) \right] (\underline{x}_0 - \underline{x}^*)$ Matrix polynomial of degree k

where $P_k(Q) = \xi_{k1} + \xi_{k2} Q + \xi_{k3} Q^2 + \ldots. + \xi_{k(k+1)} Q^k$

Convergence Analysis - 2

- Among all algorithms of the form $\underline{x}_k = \underline{x}_0 + \sum_{i=0}^{k} \gamma_{ki}\, \underline{g}_i$, the conjugate gradient method is optimal in the sense that $\forall$ k, it minimizes $f(\underline{x}_{k+1})$ over all sets of coefficients $\gamma_{k0}, \gamma_{k1}, \ldots, \gamma_{kk}$

- Let us consider the function value at $\underline{x}_{k+1}$

$$f(\underline{x}_{k+1}) = \frac{1}{2}(\underline{x}_{k+1} - \underline{x}^*)^T Q(\underline{x}_{k+1} - \underline{x}^*)$$

$$= \min_{P_k} \frac{1}{2}(\underline{x}_0 - \underline{x}^*)^T Q\,[I + Q\,P_k(Q)]^2 (\underline{x}_0 - \underline{x}^*)$$

- To simplify, let $\underline{\zeta}_1, \underline{\zeta}_2, \ldots, \underline{\zeta}_n$ be eigen vectors and $\lambda_1, \lambda_2, \ldots, \lambda_n$ be eigen values of Q. Note that $\underline{\zeta}_i$ are orthonormal

$$\text{Let } (\underline{x}_0 - \underline{x}^*) = \sum_{i=1}^{n} \gamma_i \underline{\varsigma}_i \Rightarrow Q(\underline{x}_0 - \underline{x}^*) = \sum_{i=1}^{n} \lambda_i \gamma_i \underline{\varsigma}_i$$

$$\text{Then} \quad f(\underline{x}_0) = \frac{1}{2}\left(\sum_{i=1}^{n} \lambda_i \gamma_i \underline{\varsigma}_i\right)^T \left(\sum_{j=1}^{n} \gamma_j \underline{\varsigma}_j\right) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i \gamma_i^2$$

$$f(\underline{x}_{k+1}) = \min_{P_k} \frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \gamma_i \gamma_j \underline{\varsigma}_i^T \; Q[I + QP_k(Q)]^2 \underline{\varsigma}_j$$

$$\leq \frac{1}{2}\sum_{i=1}^{n} \lambda_i \gamma_i^2 [1 + \lambda_i P_k(\lambda_i)]^2$$

Convergence Analysis - 3

- Since $\lambda_i > 0$

$$f(\underline{x}_{k+1}) \le \max_{\lambda_i} [1+\lambda_i P_k(\lambda_i)]^2 \, f(\underline{x}_0)$$

- NOTE: P_k is any polynomial of degree k. Examine implications
    1. I am still free to pick $P_k(\lambda)$.
    2. $1+\lambda P_k(\lambda)$ is a $(k+1)^{st}$ order polynomial in λ.
    3. If $k = n$-1, then pick n coefficients of $P_{n\text{-}1} \ni 1+\lambda P_{n\text{-}1}(\lambda)$ has its roots at $\lambda_i(Q) \Rightarrow f(\underline{x}_n)=0$ as we know it should since CG method exhibits quadratic termination.
    4. Suppose Q has only $r<n$ distinct eigen values λ_i. Then note that we can select $P_{r\text{-}1} \ni 1+\lambda P_{r\text{-}1}(\lambda)$ has roots at $\lambda_1, \lambda_2, \ldots, \lambda_r$ (distinct) $\Rightarrow$ get termination in r steps with CG method.
- Now, we are in a position to ascertain the convergence rate of CG method.
- Pick $P_k(\lambda) \ni 1+\lambda P_k(\lambda)$ most closely approximates 0 over the interval $\lambda_{min}(Q)$ and $\lambda_{max}(Q)$. Such a polynomial is the following Chebyshev polynomial

$$1+\lambda P_k(\lambda) = \frac{T_{k+1}\left(\dfrac{\lambda_{\min}+\lambda_{\max}-2\lambda}{\lambda_{\max}-\lambda_{\min}}\right)}{T_{k+1}\left(\dfrac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)}$$

where $T_k(\lambda) = \cos(k \cos^{-1} \lambda)$

Partial Conjugate Gradient Method (PCG)

- Can show (Problem 8.10 of Luenberger) that

$$f(\underline{x}_{k+1}) \le 4\left(\frac{\sqrt{k}-1}{\sqrt{k}+1}\right)^2 f(\underline{x}_k) \qquad k=\frac{\lambda_{max}}{\lambda_{min}}$$

$\Rightarrow$ Linear convergence, but convergence ratio $\beta_{CG} \approx 2\sqrt{\beta_{SD}}$

❑ Partial Conjugate Gradient (PCG) method

- Suppose go n-steps and compute $\underline{g}_n = -\nabla f(\underline{x}_n) \neq 0$. We missed the minimum due to inaccurate α_k, $\underline{d}_k$, $\underline{g}_k$, etc. Then, what?
- Best to start over $\underline{x}_n \rightarrow \underline{x}_o$

  $\underline{g}_n \rightarrow \underline{g}_0$

  That is, apply CG method for n-steps $\underline{x}_{k+1}=\underline{x}_k+\alpha_k\underline{d}_k$, $k=0,1,2,\ldots,n-1$ & redo with a gradient step. But, we could restart after $m<n$ steps also

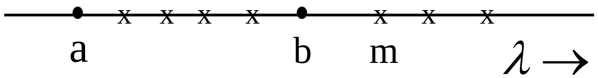
  $\Rightarrow$ PCG. Are there advantages to this approach?
- Recall that SD $\equiv$ PCG with $m=0$. Can we reduce the effective condition number k_e by restarting at $m>1$? For quadratic, we can restart with $m=n \Rightarrow k_e=1$.

Convergence Analysis of PCG - 1

- Theorem:
  - Suppose $Q>0$ and has $(n-m)$ eigen values in the interval $[a, b]$, $a>0$ and the remaining m eigen values are $>b$, then for every $\underline{x}_0$ the vector $\underline{x}_{m+1}$ generated after $(m+1)$ steps of the CG method satisfies.

$$f(\underline{x}_{m+1}) \le \left(\frac{k_e - 1}{k_e + 1}\right)^2 f(\underline{x}_0) \; ; k_e = \frac{b}{a}$$

- Proof:



Let $\lambda_1, \lambda_2, \ldots, \lambda_m > b$. Can select any P_m we like

We know,

$$f(\underline{x}_{m+1}) \le \max_{\lambda_i} [1 + \lambda_i P_m(\lambda_i)]^2 f(\underline{x}_0)$$

Define $P_m(\lambda)$ by

$$1 + \lambda P_m(\lambda) = \frac{2}{(a+b)\lambda_1 \ldots \lambda_m}\left(\frac{a+b}{2} - \lambda\right)(\lambda_1 - \lambda)(\lambda_2 - \lambda)\ldots(\lambda_m - \lambda)$$

Since $1 + \lambda_i P_m(\lambda_i) = 0 \;\forall\; i = 1, 2, \ldots, m$ we have

$$f(\underline{x}_{m+1}) \le \max_{a \le \lambda \le b} [1 + \lambda P_m(\lambda)]^2 f(\underline{x}_0)$$

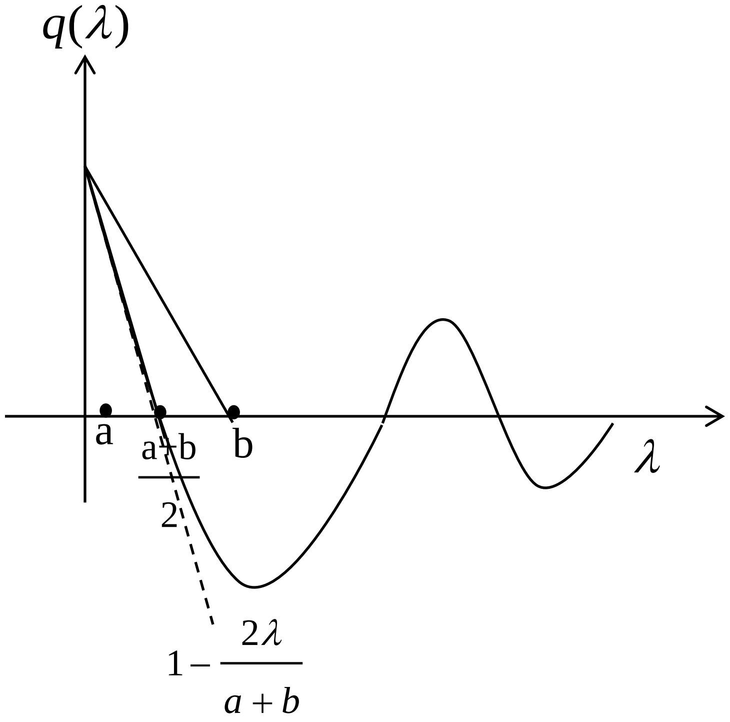
Convergence Analysis of PCG - 2

$$f(\underline{x}_{m+1}) \le \max_{a\le\lambda\le b} [1+\lambda P_m(\lambda)]^2 f(\underline{x}_0)$$

$$\le \max_{a\le\lambda\le b} \frac{\left(\lambda - \frac{a+b}{2}\right)^2}{\left(\frac{a+b}{2}\right)^2} f(\underline{x}_0)$$

$$= \left(\frac{b-a}{b+a}\right)^2 f(\underline{x}_0)$$

$$= \left(\frac{k_e - 1}{k_e + 1}\right)^2 f(\underline{x}_0)$$



- Application to non-quadratic problems

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k$$

$$\alpha_k = \arg\min_{\alpha} f(\underline{x}_k + \alpha \underline{d}_k)$$

$$\underline{d}_k = -\underline{g}_k + \beta_k \underline{d}_{k-1}$$

Application to Non Quadratic Problems - 1

$$\beta_k = \begin{cases} \dfrac{\underline{g}_k^T \underline{g}_k}{\underline{g}_{k-1}^T \underline{g}_{k-1}} & \text{Fletcher-Reeves (F-R)} \\ \max\left(\dfrac{\underline{g}_k^T (\underline{g}_k - \underline{g}_{k-1})}{\underline{g}_{k-1}^T \underline{g}_{k-1}}, 0 \right) & \text{Polak-Ribiare-Polyak (PRP)} \\ \dfrac{\underline{g}_k^T (\underline{g}_k - \underline{g}_{k-1})}{(\underline{g}_k - \underline{g}_{k-1})^T \underline{d}_{k-1}} & \text{Sorensen-Wolfe (S-W)} \end{cases}$$

$$\underline{g}_k^T \underline{d}_k = -\underline{g}_k^T \underline{g}_k + \beta_k \underline{g}_k^T \underline{d}_{k-1} < 0 \quad \text{descent direction}$$

$$\Downarrow$$

$= 0$ for quadratic case & exact line search

- ❑ Computational experience suggests that FR is worse than PRP or SW (why?)
  - conjugacy of $\{\underline{d}_k\}$ is lost due to inexact line searches
  - a situation occurs wherein $\underline{d}_k^T \underline{g}_k \approx 0 \Rightarrow \cos\theta_k = \dfrac{-\underline{g}_k^T \underline{d}_k}{\|\underline{g}_k\| \|\underline{d}_k\|} \approx 0$

$$\theta_k \approx \pi/2 \qquad \cos\theta_k = \frac{-\underline{g}_k^T \underline{d}_k}{\|\underline{g}_k\| \|\underline{d}_k\|} \Rightarrow |\underline{g}_k^T \underline{d}_k| \text{ small}$$

Application to Non Quadratic Problems -2

small $\underline{d}_k^T \underline{g}_k \Rightarrow \underline{x}_{k+1} \approx \underline{x}_k \;\Rightarrow \underline{g}_{k+1} \approx \underline{g}_k$...... recall $\alpha_k = \dfrac{-\underline{g}_k^T \underline{d}_k}{\underline{d}_k^T Q \underline{d}_k}$

$\Rightarrow \beta_{k+1} \approx 1$ for FR

$\beta_{k+1} \approx 0$ for PRP

$\Rightarrow \|\underline{d}_{k+1}\| \approx \|\underline{d}_k\|$ for FR

$\|\underline{d}_{k+1}\| \approx \|\underline{g}_{k+1}\|$ for PRP gets out of jamming

- Powell suggests using $\max(0, \beta_k)$ for PRP

- When should you restart or reset

  - Restart with SD after n steps or if $\left|\underline{g}_k^T \underline{g}_{k-1}\right| > \gamma\, \underline{g}_k^T \underline{g}_k \;;\; 0 < \gamma < 1,\ \gamma = 0.2$
  - If conjugate, we will have $\underline{g}_k^T \underline{g}_{k-1} \approx 0$
  - very useful methods for large n

CG for Quadratic Functions

- Conjugate gradient for minimizing quadratic functions $f(\underline{x}) = \frac{1}{2}\underline{x}^T Q \underline{x} - \underline{b}^T \underline{x} + c$
  - Given $\underline{x}_0$

$$g_0 = Q\underline{x}_0 - b$$
$$d_0 = -g_0$$
$$k=0$$

Do while $\underline{d}_k \neq \underline{0}$

$$\alpha_k = \frac{\underline{g}_k^T \underline{g}_k}{\underline{d}_k^T Q \underline{d}_k}$$

for NL fn.'s do line search

$$\alpha_k = \arg\min_{\alpha} f(\underline{x}_k + \alpha \underline{d}_k)$$

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k$$

$$\underline{g}_{k+1} = \underline{g}_k + \alpha_k Q \underline{d}_k$$

for NL fn.'s evaluate gradient

$$\beta_{k+1} = \begin{cases} \dfrac{\underline{g}_{k+1}^T \underline{g}_{k+1}}{\underline{g}_k^T \underline{g}_k} & \text{FR} \\ \dfrac{\underline{g}_{k+1}^T (\underline{g}_{k+1} - \underline{g}_k)}{\underline{g}_k^T \underline{g}_k} & \text{PRP} \\ \dfrac{\underline{g}_{k+1}^T (\underline{g}_{k+1} - \underline{g}_k)}{\underline{d}_k^T (\underline{g}_{k+1} - \underline{g}_k)} & \text{S-W} \end{cases}$$

$$\underline{d}_{k+1} = -\underline{g}_{k+1} + \beta_{k+1} \underline{d}_k$$
$$k = k+1$$

end Do

Solves $Q\underline{x} = \underline{b}$

Good for sparse problems

Pre-conditioning to Improve Conditioning - 1

- Pre-conditioned CG
  - Change of variable $\underline{x} = S\underline{y}$

$$h(\underline{y}) = f(S\underline{y}) = \frac{1}{2}\underline{y}^T S^T Q S \underline{y} - \underline{b}^T S \underline{y}$$

$$\underline{y}_{k+1} = \underline{y}_k + \alpha_k \underline{\tilde{d}}_k \qquad \alpha_k = \min_{\alpha} h(\underline{y}_k + \alpha_k \underline{\tilde{d}}_k)$$

$$\tilde{d}_k = -\nabla \underline{h}(\underline{y}_k) + \beta_k \underline{\tilde{d}}_{k-1} \; ; \quad \tilde{d}_0 = -\nabla h(\underline{y}_0)$$

$$\beta_k = \frac{-\nabla \underline{h}^T(\underline{y}_k)\nabla \underline{h}(\underline{y}_k)}{\nabla \underline{h}^T(\underline{y}_{k-1})\nabla \underline{h}(\underline{y}_{k-1})}$$

  - But can do all computations in $\underline{x}$ domain:

$$\nabla h(\underline{y}_k) = S^T Q S \underline{y}_k - S^T \underline{b} = S^T \nabla f(\underline{x}_k) = S^T \underline{g}_k$$

$$\tilde{d}_0 = -S^T \nabla f(\underline{x}_k) = -S^T \underline{g}_k$$

So,

$$\underline{y}_{k+1} = \underline{y}_k + \alpha_k \underline{\tilde{d}}_k$$

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k \qquad \underline{d}_k = S\underline{\tilde{d}}_k$$

$$d_{k+1} = SS^T \underline{g}_k + \beta_{k+1} \underline{d}_k$$

Pre-conditioning to Improve Conditioning - 2

$$\beta_{k+1} = \frac{\underline{g}_{k+1}{}^T SS^T \underline{g}_{k+1}}{\underline{g}_k{}^T SS^T \underline{g}_k}$$

$$\alpha_k = \arg\min f(\underline{x}_k + \alpha_k \underline{d}_k)$$

- Note that SS^T always occurs as an entity
- S via incomplete cholesky of $Q \Rightarrow S = (L^{-1})^T$

  incomplete $\Rightarrow$ if $q_{ij} = 0$, set $l_{ij} = 0 \Rightarrow$ skip or maintain sparsity of L

Pre-conditioned CG

- Given $\underline{x}_0$, pre-conditioner $M = (S^{-1})^T S^{-1} = LL^T$

$$\underline{g}_0 = Q\underline{x}_0 - \underline{b}$$

$$\text{Solve } M\ \underline{z}_0 = \underline{g}_0 \Rightarrow \underline{z}_0 = SS^T \underline{g}_0$$

$$\underline{d}_0 = -\underline{g}_0 \qquad k \leftarrow 0$$

$$\text{while } \|\underline{g}_k\| \neq 0$$

$$\alpha_k = \frac{\underline{g}_k^T \underline{g}_k}{\underline{d}_k^T Q \underline{d}_k}$$

For NL functions, do line search

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k$$

$$\underline{g}_{k+1} = \underline{g}_k + \alpha_k Q \underline{d}_k$$

For NL functions, evaluate gradient

$$\text{Solve } M\underline{z}_{k+1} = \underline{g}_{k+1}$$

$$\beta_{k+1} = \frac{\underline{g}_{k+1}^T \underline{z}_{k+1}}{\underline{g}_k^T \underline{z}_k}$$

FR used here. PRP and SW are possible

$$\underline{d}_{k+1} = -\underline{g}_{k+1} + \underline{\beta}_{k+1} \underline{d}_k$$

$$k \leftarrow k+1$$

end DO

CG in Trust Region Method -1

- Use of CG in trust region method
  - Recall that in trust region method, we want $\underline{d}_k \ni$

$$\min_{\underline{d}_k} f_k + \underline{g}_k^T \underline{d}_k + \frac{1}{2} \underline{d}_k^T F_k \underline{d}_k \qquad f_k = f(\underline{x}_k)$$

$$\text{s.t. } \|\underline{d}_k\| \le \delta$$

remove subscript k for simplicity.

See Nocedal's Book

  - In this problem $\underline{x} \sim \underline{d}$

    so, we use $\underline{p}$ for directions

  - Given $\varepsilon > 0$

    Set $\underline{d}_0 = \underline{0}$, $\underline{p}_o = -\underline{g}$

    If $\|\underline{g}\| < \varepsilon$

    return $\underline{d} = \underline{d}_0 \Rightarrow$ stop

    for $k = 0, 1, 2, \ldots$

    If $\underline{p}_k^T F \underline{p}_k \le 0$

    find τ such that $\underline{d} = \underline{d}_k + \tau \underline{p}_k$ minimizes

$$f + \underline{g}^T \underline{d} + \frac{1}{2} \underline{d}^T F \underline{d} \text{ and } \|\underline{d}\| = \delta$$

CG in Trust Region Method -2

$$f + \underline{g}^T(\underline{d}_k + \tau \underline{p}_k) + \frac{1}{2}(\underline{d}_k + \tau \underline{p}_k)^T F(\underline{d}_k + \tau \underline{p}_k)$$

$\Rightarrow \tau$ is the solution of

$$\underline{d}_k^T \underline{d}_k + 2\tau \underline{d}_k^T \underline{p}_k + \tau^2 \underline{p}_k^T \underline{p}_k = \delta^2$$

return $\underline{d}_k + \tau \underline{p}_k = \underline{d}$

else

set $\alpha_k = \dfrac{\underline{g}_k^T \underline{g}_k}{\underline{p}_k^T F \underline{p}_k}$

CG in Trust Region Method -3

Set $\underline{d}_{k+1} = \underline{d}_k + \alpha_k \underline{p}_k$

If $\|\underline{d}_{k+1}\| \geq \delta$

Find τ such that $\|\underline{d}_k + \tau \underline{p}_k\| = \delta$

return $\underline{d} = \underline{d}_k + \tau \underline{p}_k$

end if

Set $\underline{g}_{k+1} = \underline{g}_k + \alpha_k F \underline{p}_k$

If $\|\underline{g}_{k+1}\| < \varepsilon \|\underline{g}_0\|$

return $\underline{d} = \underline{d}_{k+1}$

end if

$$\text{set } \beta_{k+1} = \begin{cases} \dfrac{\underline{g}_{k+1}^T \underline{g}_{k+1}}{\underline{g}_k^T \underline{g}_k} & \text{FR} \\[2ex] \dfrac{\underline{g}_{k+1}^T (\underline{g}_{k+1} - \underline{g}_k)}{\underline{g}_k^T \underline{g}_k} & \text{PRP} \\[2ex] \dfrac{\underline{g}_{k+1}^T (\underline{g}_{k+1} - \underline{g}_k)}{\underline{\alpha}_k^T (\underline{g}_{k+1} - \underline{g}_k)} & \text{S-W} \end{cases}$$

$\underline{p}_{k+1} = -\underline{g}_{k+1} + \beta_{k+1} \underline{d}_k$

$k = k + 1$

end DO

References

1. D. F. Shanno, “ Conjugate gradient methods with inexact searches”, Mathematics of OR, Vol. 3, No. 3, August 1978, pp.244-256
2. M. J. D. Powell, “Restart procedures for conjugate gradient method”, Math. Programming, Vol. 12, 1977, pp.241-254.
3. M. J.D. Powell, “Some convergence properties of the conjugate gradient method”, Math. Programming, Vol. 11, pp. 42-49, 1976.
4. M.Al-Baali, “Descent property and global convergence of the Fletcher-Reeves method with inexact line searches”, IMA J. Numerical Analysis, Vol.5, 1985, pp. 121-124.
5. M.J.D. Powell, “ Nonconvex minimization calculations and the conjugate gradient method”, in Numerical analysis proceedings, Dundee 1983, D.F. Griffiths, ed., springer-verlag, Berlin 1984.
6. M.J.D. Powell, “ Convergence properties of Algorithms for non linear optimization”, SIAM Review, Vol. 28, No. 4, pp. 487-500, Dec 1986.

Summary

- What are Conjugate Directions?
- Conjugate Direction Methods
- Conjugate Gradient Methods
- Convergence Analysis
- Partial Conjugate Gradient Method
- Application to Non-quadratic problems
- Pre-conditioned Conjugate Gradient (Scaling)