Lecture 3: QUADRATIC INTERPOLATION, COMBINED GOLDEN SECTION & QUADRATIC INTERPOLATION CONVERGENCE OF GENERALIZED GRADIENT METHOD, STOPPING CRITERIA

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 15, 2009

Outline of Lecture 3

- Quadratic Interpolation
- Combined Golden Section and Quadratic Interpolation
- Convergence of Generalized Gradient Method
- Stopping Criteria
- Some Test Examples

Quadratic Interpolation: Basic Ideas

- **To fit a parabola to the scalar function of α, $g(\alpha) = f(\underline{x}_k + \alpha \underline{d}_k)$, we need three pieces of information, e.g., values of g at three points**
  - Suppose have function values at α_1, α_2 and $\alpha_3 \Rightarrow g(\alpha_1)$, $g(\alpha_2)$ and $g(\alpha_3)$
  - How to get them later. Recall that golden section search also needs it! Suppose

$\alpha_1 < \alpha_2 < \alpha_3$ & $g(\alpha_1) > g(\alpha_2)$ & $g(\alpha_3) > g(\alpha_2) \Rightarrow$ "smaller in the middle"

$\Rightarrow$ a local minimum is bracketed by (α_1, α_3)

such an $(\alpha_1, \alpha_2, \alpha_3) \ni \left. \begin{array}{l} \alpha_1 < \alpha_2 < \alpha_3 \\ g(\alpha_1) > g(\alpha_2) \;\; \& \; g(\alpha_3) > g(\alpha_2) \end{array} \right\}$ is termed a "THREE POINT PATTERN"

  - Since a parabola $a\alpha^2 + b\alpha + c$ is parameterized by (a, b, c), we have

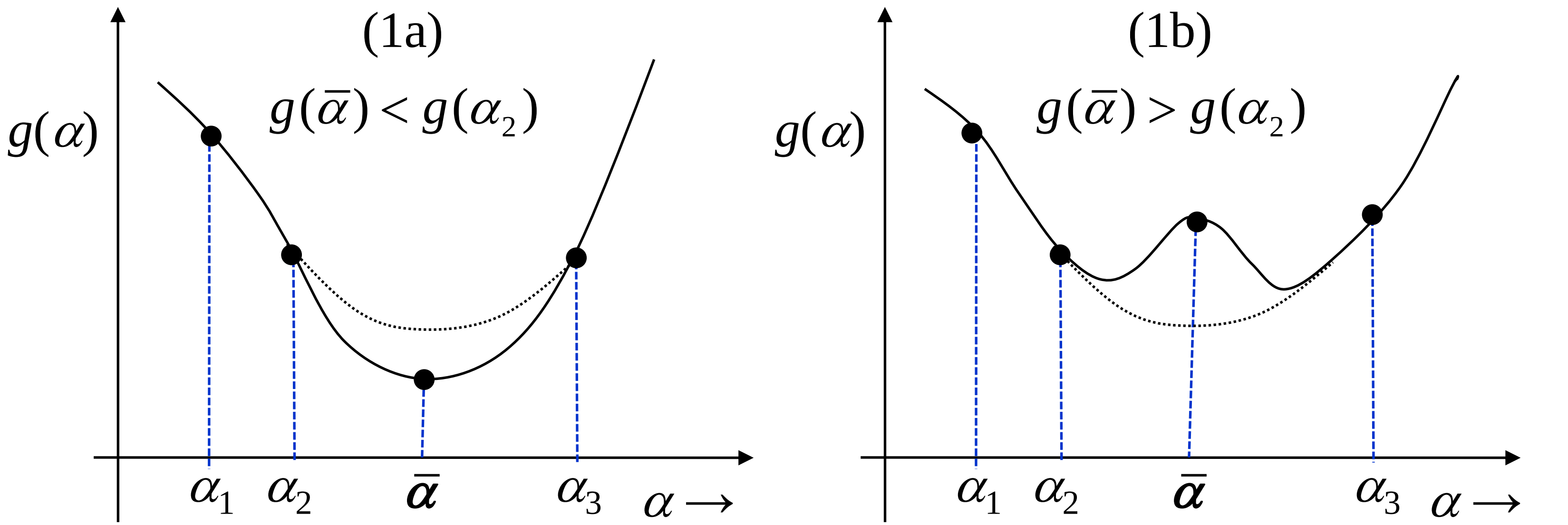
$$\left. \begin{array}{l} g(\alpha_1) = a\alpha_1^2 + b\alpha_1 + c \\ g(\alpha_2) = a\alpha_2^2 + b\alpha_2 + c \\ g(\alpha_3) = a\alpha_3^2 + b\alpha_3 + c \end{array} \right\} \Rightarrow \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{bmatrix} \alpha_1^2 & \alpha_1 & 1 \\ \alpha_2^2 & \alpha_2 & 1 \\ \alpha_3^2 & \alpha_3 & 1 \end{bmatrix}^{-1} \begin{bmatrix} g(\alpha_1) \\ g(\alpha_2) \\ g(\alpha_3) \end{bmatrix}$$

  - Minimum of a parabola is achieved at $\bar{\alpha} = -b/2a$ so that

$$\bar{\alpha} = \frac{1}{2} \frac{g(\alpha_1)(\alpha_3^2 - \alpha_2^2) + g(\alpha_2)(\alpha_1^2 - \alpha_3^2) + g(\alpha_3)(\alpha_2^2 - \alpha_1^2)}{g(\alpha_1)(\alpha_3 - \alpha_2) + g(\alpha_2)(\alpha_1 - \alpha_3) + g(\alpha_3)(\alpha_2 - \alpha_1)}$$

Setting up Next Three Point Pattern: Case 1

- **Two cases can occur**
  - **Case 1**: $\bar{\alpha} > \alpha_2$



$$g(\bar{\alpha}) < g(\alpha_2) \Rightarrow \begin{array}{l} \alpha_1 \leftarrow \alpha_2 \\ \alpha_2 \leftarrow \bar{\alpha} \\ \alpha_3 \leftarrow \alpha_3 \end{array}$$

$$g(\bar{\alpha}) > g(\alpha_2) \Rightarrow \begin{array}{l} \alpha_1 \leftarrow \alpha_1 \\ \alpha_2 \leftarrow \alpha_2 \\ \alpha_3 \leftarrow \bar{\alpha} \end{array}$$

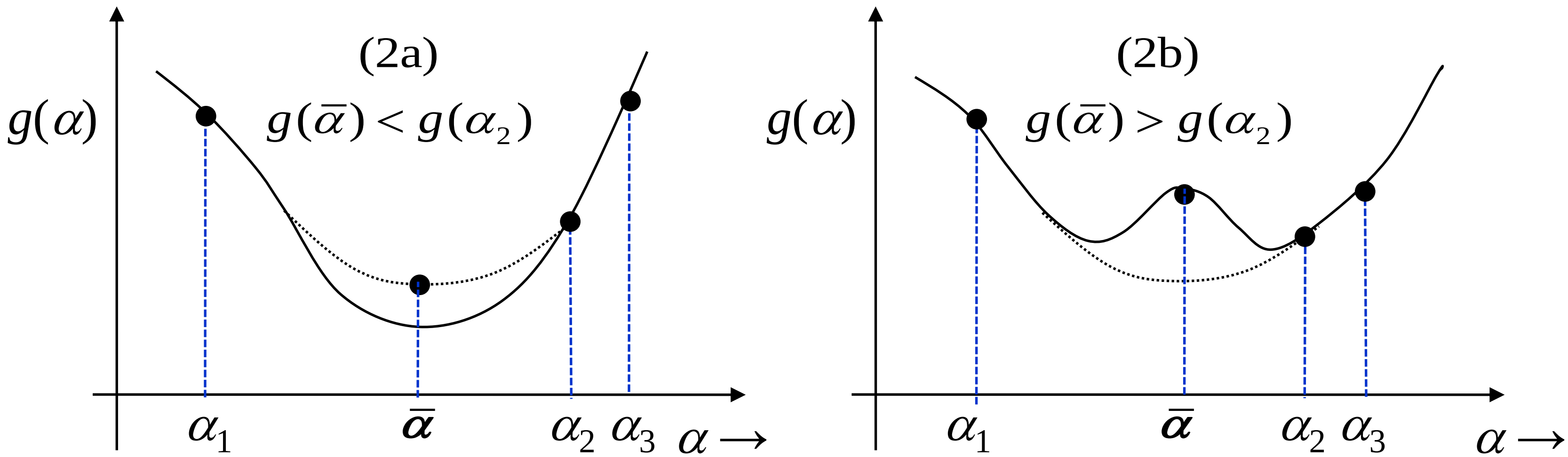
Three point pattern: $\alpha_1 < \alpha_2 < \alpha_3$

$g(\alpha_1) > g(\alpha_2)$ & $g(\alpha_3) > g(\alpha_2)$

Three point pattern conditions are satisfied

Setting up Next Three Point Pattern: Case 2

- **Case 2**: $\bar{\alpha} < \alpha_2$



$$g(\bar{\alpha}) < g(\alpha_2) \Rightarrow \begin{array}{l} \alpha_3 \leftarrow \alpha_2 \\ \alpha_2 \leftarrow \bar{\alpha} \\ \alpha_1 \leftarrow \alpha_1 \end{array}$$

$$g(\bar{\alpha}) > g(\alpha_2) \Rightarrow \begin{array}{l} \alpha_1 \leftarrow \bar{\alpha} \\ \alpha_2 \leftarrow \alpha_2 \\ \alpha_3 \leftarrow \alpha_3 \end{array}$$

- Note that if $g(\bar{\alpha}) \approx g(\alpha_2)$, then a special local search near $\bar{\alpha}$ should be conducted to replace $\bar{\alpha}$ by a point $\bar{\alpha}^*$ with $g(\bar{\alpha}^*) \neq g(\bar{\alpha})$.
- Terminate the computation when the length of the three point pattern is smaller than a certain tolerance $\Rightarrow \alpha_2 \rightarrow \alpha^*$ and $|\alpha_3 - \alpha_1|$ shrinks. Typically require $|\alpha_3 - \alpha_1| \leq \varepsilon|\alpha_2|$, $\varepsilon \approx .01 \sim .0001$

Setting up Initial Three Point Pattern - 1

- **How to pick an initial three point pattern (or equivalently, the initial range)?**
  - Need $\alpha_1 < \alpha_2 < \alpha_3 \ni \begin{matrix} g(\alpha_1) > g(\alpha_2) \\ g(\alpha_3) > g(\alpha_2) \end{matrix}$
  - Procedure Initialize $z_0 = 0$, Δ = increment, $\tau > 1$ increase ratio (e.g., 1.618)

$z_1 = \Delta$, $i = 1$

If $g(z_1) < g(z_0)$ then

Do while $g(z_i) < g(z_{i-1})$

$i = i + 1$

$z_i = z_0 + \tau^{i-1}\Delta$

End Do

Three point pattern: $\begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ z_{i-2} & z_{i-1} & z_i \end{pmatrix}$

Else

Do while $g(z_i) > g(z_0)$

$i = i + 1$

$z_i = z_0 + (z_{i-1} - z_0)/\ \tau$

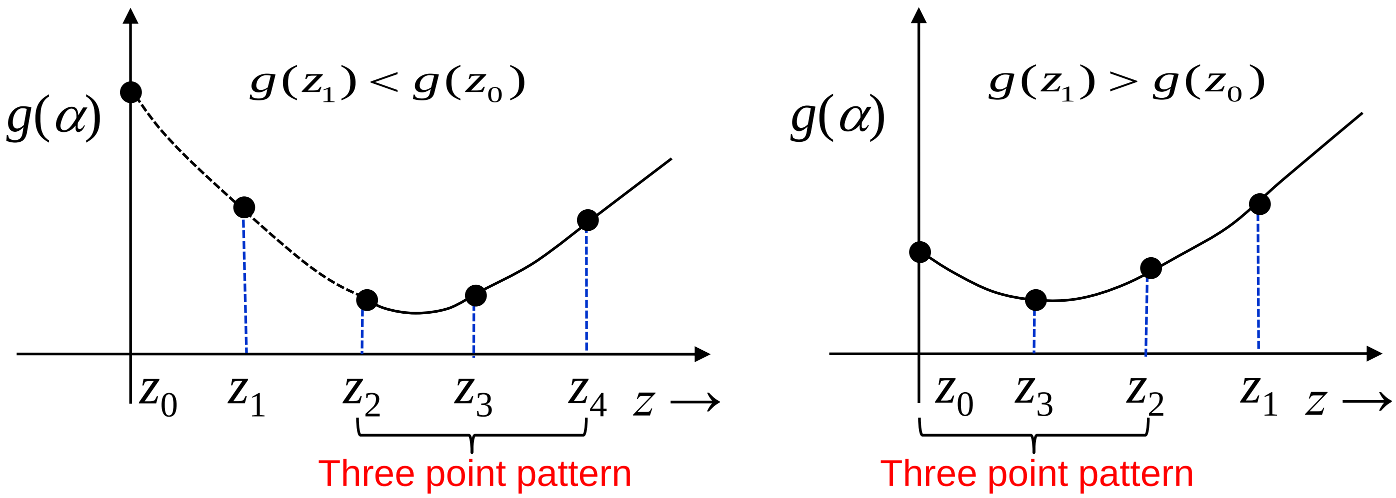
End Do

Three point pattern: $\begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ z_0 & z_i & z_{i-1} \end{pmatrix}$

End If

Setting up Initial Three Point Pattern - 2

- **Illustration**



- Note that if take $\tau = 1.618$, ideal for golden section search.

How to Pick Δ?

- **Picking Δ**: ∃ two methods
  - Pick $\Delta \approx \alpha^*$. Suppose have $g(0)$, $g'(0)$ and $g''(0)$, then we can fit a parabola

$$p(\alpha) = g(0) + g'(0)\alpha + (1/2)g''(0)\alpha^2$$

$$\Rightarrow \Delta \approx \frac{-g'(0)}{g''(0)} = \frac{-\nabla \underline{f}^T(\underline{x}_k)\underline{d}_k}{\underline{d}_k^T \nabla^2 f(\underline{x}_k)\underline{d}_k}$$

  - Suppose have access to $g(0)$ and $g'(0)$ only, then pick Δ to obtain a specified decrease in function value (e.g., 10 ~30%)

$$\Rightarrow g(0) + g'(0)\Delta = \beta g(0) \qquad (\beta = .7 \sim .9)$$

$$\Delta = \frac{(\beta - 1)g(0)}{g'(0)} = \frac{(\beta - 1)|f(\underline{x}_k)|}{\nabla \underline{f}^T(\underline{x}_k)\underline{d}_k}$$

  - $$\Delta \approx \frac{2}{\lambda_{\max} + \lambda_{\min}} \cong \frac{2}{\max_i \frac{\partial^2 f}{\partial x_i^2} + \min_i \frac{\partial^2 f}{\partial x_i^2}}$$

  - $$\Delta = \frac{n}{tr(\nabla^2 f(\underline{x}))}$$

Convergence Analysis

Convergence Analysis of Quadratic Fit

- Define the errors $\bar{e} = \alpha^* - \bar{\alpha},\ e_i = \alpha^* - \alpha_i,\ 1 \le i \le 3$

$$\bar{e} = \alpha^* - \bar{\alpha} = \alpha^* - \frac{1}{2}\frac{g_1(\alpha_3^2 - \alpha_2^2) + g_2(\alpha_1^2 - \alpha_3^2) + g_3(\alpha_2^2 - \alpha_1^2)}{g_1(\alpha_3 - \alpha_2) + g_2(\alpha_1 - \alpha_3) + g_3(\alpha_2 - \alpha_1)}$$

$$= -\frac{1}{2}\frac{[(g_1 - g_2)(\alpha^* - \alpha_3)^2 + (g_2 - g_3)(\alpha^* - \alpha_1)^2 + (g_3 - g_1)(\alpha^* - \alpha_2)^2]}{g_1(\alpha_3 - \alpha_2) + g_2(\alpha_1 - \alpha_3) + g_3(\alpha_2 - \alpha_1)}$$

- As $k \to \infty$, $\bar{e}$ must be a polynomial function of e_1, e_2, e_3. Must be second order since quadratic fit. $\bar{e} \to 0$ if any two of e_1, e_2, $e_3 \to 0$. Must be symmetric

$$\Rightarrow \bar{e} \approx M(e_1 e_2 + e_2 e_3 + e_3 e_1)$$

- As $k \to \infty$

$$\text{Recall } e_{k+1} = \beta e_k^r \Rightarrow \ln M e_{k+1} = \ln \underbrace{\frac{\beta}{M^{r-1}}}_{\beta'} + r \ln M e_k$$

$$e_{k+2} = M e_k e_{k-1} \Rightarrow M e_{k+2} = (M e_k)(M e_{k-1})$$

$$\ln M e_{k+2} = \ln M e_k + \ln M e_{k-1}$$

$$y_{k+2} = y_k + y_{k-1}$$

characteristic Eq^n : $z^3 - z - 1 = 0 \Rightarrow r = 1.33$

$$y_{k+1} \approx 1.33\, y_k$$
$$\Rightarrow \ln M e_{k+1} = 1.33 \ln M e_k$$

Super linear convergence

Hybrid Golden Section & Quadratic Interpolation

Combining Golden Section Search and Quadratic Fit

- Set up the three point pattern using $\tau = 1.618$, $z_i = z_0 + \tau^{i-1}\Delta$

$$z_i = z_0 + \frac{z_{i-1} - z_0}{\tau}, \qquad i \geq 2$$

- Use golden section search to reduce the interval $[l_1, r_1]$ to

$$(l_N, r_N) \ni (r_N - l_N) = \frac{r_1 - l_1}{PG}, \quad PG = 40$$

- Use the quadratic search procedure to reduce the interval by a factor of PQ (PQ = 100-1,000)

Combined Armijo & Quadratic Interpolation

- **Combining Armijo Step Size Rule and Quadratic Fit**

Given $\sigma \in (0, \frac{1}{2})$, and s

$$l = 0.1$$

$$k = 0$$

$$\alpha_k = s$$

Do while $f(\underline{x}_k + \alpha_k \underline{d}_k) > f(\underline{x}_k) + \sigma \alpha_k \nabla \underline{f}^T(\underline{x}_k)\underline{d}_k$

$$\gamma_k = \frac{-\alpha_k^2 \nabla \underline{f}^T(\underline{x}_k)\underline{d}_k}{2[f(\underline{x}_k + \alpha_k \underline{d}_k) - \alpha_k \nabla \underline{f}^T(\underline{x}_k)\underline{d}_k - f(\underline{x}_k)]}$$

$$\alpha_k = \begin{cases} \alpha_k / \beta, & \text{if } f(\underline{x}_k + \gamma_k \underline{d}_k) \geq f(\underline{x}_k + \alpha_k \underline{d}_k) \\ \gamma_k, & \text{if } f(\underline{x}_k + \gamma_k \underline{d}_k) < f(\underline{x}_k + \alpha_k \underline{d}_k) \ \& \ \gamma_k \geq l\alpha_k \\ l\alpha_k, & \text{otherwise} \end{cases}$$

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

$$k \leftarrow k + 1$$

End do

Convergence of Generalized Gradient Method - 1

Convergence Analysis of the generalized Gradient Method

$$H_k = \begin{cases} I, & SD \\ Diag(\frac{1}{d^2 f / dx_i^2}), & \text{Diagonal Scaling} \\ [\nabla^2 f(\underline{x})]^{-1}, & \text{Hessian} \end{cases}$$

- Consider the generalized gradient method

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k = \underline{x}_k - \alpha_k H_k \nabla \underline{f}(\underline{x}_k) = \underline{x}_k - \alpha_k H_k \underline{g}_k$$

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

- Does the algorithm converge? Yes, as long as $\nabla \underline{f}^T(\underline{x}_k)\underline{d}_k < 0 \ \ \forall \nabla \underline{f}(\underline{x}_k) \ni \left\|\nabla \underline{f}(\underline{x}_k)\right\| \neq 0$ and $\left\|\underline{d}_k\right\| < \infty$, α_k from Armijo, Goldstein, Armijo-quadratic or Golden section and quadratic
- How fast does it converge to a local minimum?…speed or rate of convergence

Convergence of Generalized Gradient Method - 2

- Let us consider a quadratic object function. Why quadratic?

Recall most functions can be approximated by a quadratic function near minimum

$$f(\underline{x}) = f(\underline{x}^*) + \underbrace{\frac{1}{2}(\underline{x}-\underline{x}^*)^{\mathrm{T}}\nabla^2 f(\underline{x}^*)(\underline{x}-\underline{x}^*)}_{\text{a quadratic surface}}$$

$f(\underline{x}^*)$ ↓ constant

- Conisider the qudratic fuction $f(\underline{x}) = \frac{1}{2}(\underline{x}-\underline{x}^*)^{\mathrm{T}}Q(\underline{x}-\underline{x}^*),\ Q>0$

min at $\underline{x}=\underline{x}^*$ and $f(\underline{x}^*)=0$

$$\nabla f(\underline{x}_k) = Q(\underline{x}_k-\underline{x}^*) = \underline{g}_k$$

$$\underline{d}_k = -H_k\underline{g}_k$$

Optimal $\alpha=\alpha_k = -\dfrac{\underline{g}_k^T\underline{d}_k}{\underline{d}_k^TQ\underline{d}_k} = \dfrac{\underline{g}_k^TH_k\underline{g}_k}{\underline{g}_k^TH_kQ_kH_k\underline{g}_k}$

Convergence of Generalized Gradient Method - 3

Let $\underline{y}_k = H_k^{1/2} \underline{g}_k$; $H_k^{1/2}$ symmetric $\underline{y}_k \to \begin{cases} \underline{g}_k & SD \\ [\nabla^2 f(\underline{x}_k)]^{-1/2} \underline{g}_k & \text{Newton} \end{cases}$

$L_k = H_k^{1/2} Q H_k^{1/2}$ $L_k \to \begin{cases} Q & SD \\ I & \text{Newton} \end{cases}$

$$\Rightarrow \alpha_k = \frac{\underline{y}_k^T \underline{y}_k}{\underline{y}_k^T L_k \underline{y}_k} \le \frac{1}{\lambda_{\min}(L_k)}$$

$$\underline{g}_k = Q(\underline{x}_k - \underline{x}^*)$$
$$\underline{g}_k^T H_k \underline{g}_k = \underline{y}_k^T \underline{y}_k$$
$$\underline{g}_k^T H_k Q \underline{H_k \underline{g}_k} = \underline{y}_k^T L_k \underline{y}_k$$

Note: $f(\underline{x}_k) = (1/2) \underline{y}_k^T L_k^{-1} \underline{y}_k$

$$f(\underline{x}_{k+1}) = (1/2)[\underline{x}_k - \alpha_k H_k \underline{g}_k - \underline{x}^*]^T Q [\underline{x}_k - \alpha_k H_k \underline{g}_k - \underline{x}^*]$$
$$= f(\underline{x}_k) - \alpha_k (\underline{x}_k - \underline{x}^*)^T Q H_k \underline{g}_k + (1/2)\alpha_k^2 \underline{g}_k^T H_k Q H_k \underline{g}_k$$
$$= f(\underline{x}_k) - \alpha_k \underline{g}_k^T H_k \underline{g}_k + (1/2)\alpha_k^2 \underline{g}_k^T H_k Q H_k \underline{g}_k$$
$$= f(\underline{x}_k) - \left(\frac{\underline{y}_k^T \underline{y}_k}{\underline{y}_k^T L_k \underline{y}_k} \right) \left[\underline{y}_k^T \underline{y}_k - (1/2) \frac{\underline{y}_k^T \underline{y}_k}{\underline{y}_k^T L_k \underline{y}_k} \underline{y}_k^T L_k \underline{y}_k \right]$$
$$= f(\underline{x}_k) - (1/2) \frac{(\underline{y}_k^T \underline{y}_k)^2}{(\underline{y}_k^T L_k \underline{y}_k)}$$

Convergence of Generalized Gradient Method - 4

Find results:

$$f(\underline{x}_{k+1}) = \overbrace{\left[1 - \frac{(\underline{y}_k^T \underline{y}_k)^2}{(\underline{y}_k^T L_k \underline{y}_k)[\underline{y}_k^T L_k^{-1} \underline{y}_k]}\right]}^{\beta_k} f(\underline{x}_k)$$

Special cases:

$$f(\underline{x}_{k+1}) = \left[1 - \frac{(\underline{g}_k^T \underline{g}_k)^2}{(\underline{g}_k^T Q \underline{g}_k)[\underline{g}_k^T Q^{-1} \underline{g}_k]}\right] f(\underline{x}_k) \cdots \text{SD}$$

$$f(\underline{x}_{k+1}) = 0 \qquad \text{Newton}$$

At first glance appears to be linearly convergent. Let us explore further

1) Rayleigh inequality:

$$\lambda_{\min}(L_k) \le \frac{\underline{y}^T L_k \underline{y}}{\underline{y}^T \underline{y}} \le \lambda_{\max}(L_k), \qquad L_k = L_k^T$$

$$\text{(or)} \quad \frac{1}{\lambda_{\max}(L_k)} \le \frac{\underline{y}^T \underline{y}}{\underline{y}^T L_k \underline{y}} \le \frac{1}{\lambda_{\min}(L_k)}$$

Convergence of Generalized Gradient Method - 5

Similarly

$$\frac{1}{\lambda_{\max}(L_k^{-1})} = \lambda_{\min}(L_k) \le \frac{\underline{y}^T\underline{y}}{\underline{y}^T L_k^{-1}\underline{y}} \le \frac{1}{\lambda_{\min}(L_k^{-1})} = \lambda_{\max}(L_k)$$

Use lower bound $\Rightarrow \beta_k \le [1-\frac{\lambda_{\min}(L_k)}{\lambda_{\max}(L_k)}]=(1-\frac{1}{\kappa(L_k)})$

$$\kappa(L_k) = \text{condition number of } L_k = \sqrt{\lambda_{\max}(L_k L_k^T)/\lambda_{\min}(L_k L_k^T)} = \frac{\lambda_{\max}(L_k)}{\lambda_{\min}(L_k)}$$

$\kappa(L_k)$ large $\Rightarrow$ BAD NEWS since $\beta_k \approx 1$

would like $\kappa(L_k) \approx 1 \Rightarrow \beta_k \approx 0 \Rightarrow$ approach Newton's method

2) Kantorovich inequality:

$$\frac{(\underline{y}^T\underline{y})^2}{(\underline{y}^T L_k \underline{y})(\underline{y}^T L_k^{-1}\underline{y})} \ge \frac{4\lambda_{\min}(L_k)\lambda_{\max}(L_k)}{[\lambda_{\min}(L_k)+\lambda_{\max}(L_k)]^2} = \frac{4\kappa(L_k)}{[\kappa(L_k)+1]^2}$$

$$L_k \text{ symmetric} \Rightarrow \exists \text{ an orthogonal matrix } T \ni T^T L T = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} = \Lambda$$

Convergence of Generalized Gradient Method - 6

Assume $\lambda_{\min} = \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n = \lambda_{\max}$

$$\text{Let } \underline{z} = T^T \underline{y} \Rightarrow \underline{y}^T L_k \underline{y} = \underline{z}^T \Lambda \underline{z} = \sum_{i=1}^{n} \lambda_i z_i^2$$

$$\underline{y}^T L_k^{-1} \underline{y} = \underline{z}^T \Lambda^{-1} \underline{z} = \sum_{i=1}^{n} z_i^2 / \lambda_i$$

$$\Rightarrow \frac{(\underline{y}^T \underline{y})^2}{(\underline{y}^T L_k \underline{y})(\underline{y}^T L_k^{-1} \underline{y})} = \frac{(\underline{z}^T \underline{z})^2}{(\sum_{i=1}^{n} \lambda_i z_i^2)(\sum_{i=1}^{n} z_i^2 / \lambda_i)}$$

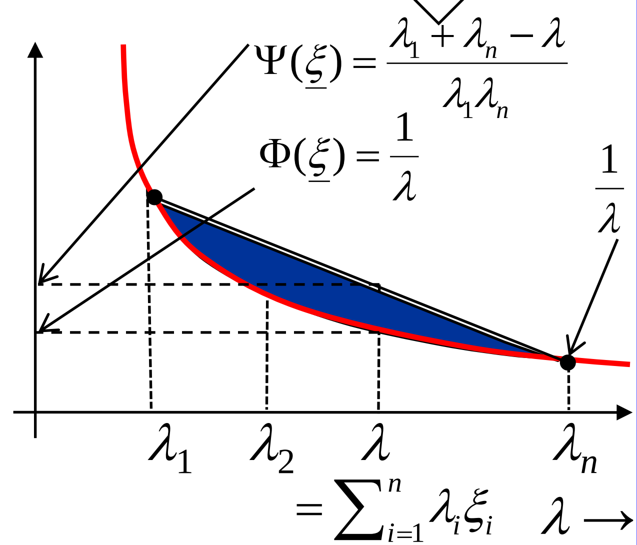
$$\text{Let } \xi_i = \frac{z_i^2}{\underline{z}^T \underline{z}} \Rightarrow \xi_i \ge 0 \ \& \quad \underline{\xi}^T \underline{e} = 1$$

$$\Rightarrow \frac{(\underline{y}^T \underline{y})^2}{(\underline{y}^T L_k \underline{y})(\underline{y}^T L_k^{-1} \underline{y})} = \frac{1/(\sum_{i=1}^{n} \lambda_i \xi_i)}{(\sum_{i=1}^{n} \xi_i / \lambda_i)} = \frac{\Phi(\underline{\xi})}{\Psi(\underline{\xi})}$$

$\sum_{i=1}^{n} \lambda_i \xi_i = \lambda$ is a point on the line segment (λ_1, λ_n)

$\Psi(\underline{\xi}) = \sum_{i=1}^{n} \xi_i / \lambda_i$ is a convex combimation of $1/\lambda_i$

$$\frac{1}{\lambda_1} + \frac{1/\lambda_n - 1/\lambda_1}{(\lambda_n - \lambda_1)}(\lambda - \lambda_1)$$
$$= \frac{\lambda_n - \lambda_1 + (\lambda_1/\lambda_n - 1)(\lambda - \lambda_1)}{\lambda_1(\lambda_n - \lambda_1)}$$
$$= \frac{\lambda_n(\lambda_n - \lambda_1) + (\lambda_1 - \lambda_n)(\lambda - \lambda_1)}{\lambda_1 \lambda_n (\lambda_n - \lambda_1)}$$
$$= \frac{\lambda_1 + \lambda_n - \lambda}{\lambda_1 \lambda_n}$$



Convergence of Generalized Gradient Method - 7

$$\frac{\Phi(\underline{\xi})}{\Psi(\underline{\xi})} \geq \min_{\lambda_1 \leq \lambda \leq \lambda_n} \frac{1/\lambda}{1/\lambda_1 + \frac{(1/\lambda_n - 1/\lambda_1)}{(\lambda_n - \lambda_1)}(\lambda - \lambda_1)} = \min_{\lambda_1 \leq \lambda \leq \lambda_n} \frac{1/\lambda}{\frac{\lambda_1 + \lambda_n - \lambda}{\lambda_1 \lambda_n}}$$

Optimum at $\frac{\lambda_1 + \lambda_n}{2} = \lambda^*$

$$\Rightarrow \frac{\Phi(\underline{\xi})}{\Psi(\underline{\xi})} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2} = \frac{4\lambda_{\min}(L_k)\lambda_{\max}(L_k)}{[\lambda_{\min}(L_k) + \lambda_{\max}(L_k)]^2} = \frac{4\kappa(L_k)}{[\kappa(L_k)+1]^2}$$

$$\Rightarrow f(\underline{x}_{k+1}) \leq \left(1 - \frac{4\kappa(L_k)}{[\kappa(L_k)+1]^2}\right) f(\underline{x}_k) = \frac{(\kappa(L_k)-1)^2}{[\kappa(L_k)+1]^2} f(\underline{x}_k) = \left(\frac{\kappa(L_k)-1}{\kappa(L_k)+1}\right)^2 f(\underline{x}_k)$$

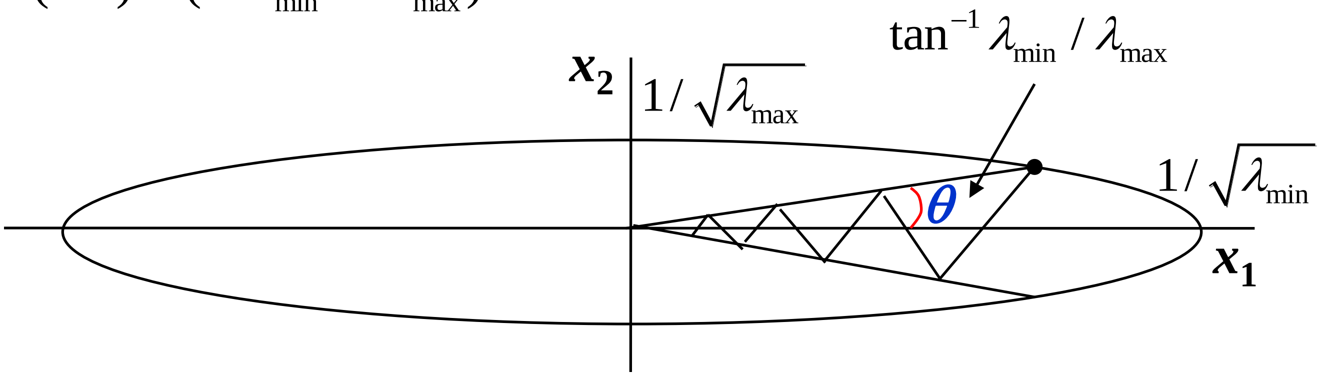
- Convergence ratio $\beta = \lim_{k \to \infty} \left(\frac{\lambda_{\max}(L_k) - \lambda_{\min}(L_k)}{\lambda_{\max}(L_k) + \lambda_{\min}(L_k)} \right)^2$
- $\kappa(L_k) = 1 \Rightarrow \lambda_{\max}(H_k^{1/2} Q H_k^{1/2}) \approx \lambda_{\min}(H_k^{1/2} Q H_k^{1/2})$ (or) $H_k = Q^{-1}$

  $\beta = 0 \Rightarrow$ super linear convergence
- When $\kappa(L_k) \gg 1$, (e.g., steepest descent with $\lambda_{\max}(Q)/\lambda_{\min}(Q) \gg 1$), convergence can be very slow for certain $\underline{x}_0$.

Zigzagging Behavior of Steepest Descent - 1

- **Example:** $f(\underline{x}) = \frac{1}{2}x_1^2 + \frac{9}{2}x_2^2 \quad \text{min at } (0,0)$

$$\underline{x}_0 = (9 \;\; 1) \propto (1/\lambda_{\min} \;\; 1/\lambda_{\max})$$



$$\underline{g}_k = Q\underline{x}_k = \begin{pmatrix} x_1 \\ 9x_2 \end{pmatrix}; \quad \underline{g}_k^T Q \underline{g}_k = x_1^2 + 729x_2^2; \qquad \underline{g}_k^T \underline{g}_k = x_1^2 + 81x_2^2;$$

$$\underline{g}_k^T Q^{-1} \underline{g}_k = x_1^2 + 9x_2^2 \Rightarrow \beta_k = [1 - \frac{(x_1^2 + 81x_2^2)^2}{(x_1^2 + 9x_2^2)(x_1^2 + 729x_2^2)}]_k = \left(\frac{576x_1^2 x_2^2}{(x_1^2 + 9x_2^2)(x_1^2 + 729x_2^2)} \right)_k$$

$$\underline{x}_{k+1} = \underline{x}_k - \frac{\underline{g}_k^T \underline{g}_k}{\underline{g}_k^T Q \underline{g}_k} \underline{g}_k = \begin{pmatrix} 648x_1x_2^2 \\ -8x_1^2x_2 \end{pmatrix}_k \frac{1}{(x_1^2 + 729x_2^2)_k}$$

Zigzagging Behavior of Steepest Descent - 2

$$\underline{x}_1 = \begin{pmatrix} (648)9 \\ -648 \end{pmatrix} \frac{1}{810} = \begin{pmatrix} 9 \\ -1 \end{pmatrix} .8;\ \underline{x}_2 = (.8)^2 \begin{pmatrix} 9 \\ 1 \end{pmatrix} \Rightarrow \underline{x}_k \begin{bmatrix} 9 \\ (-1)^k \end{bmatrix} .8^k = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^k \begin{pmatrix} \frac{9}{\lambda_{\min}} \\ \frac{(-1)^k 9}{\lambda_{\max}} \end{pmatrix}$$

In general,

$$f(\underline{x}) = \sum_{i=1}^{n} \lambda_i x_i^2; \qquad \underline{x}_0 = \left(\frac{1}{\lambda_{\min}}\ 0\ 0 \cdots 0\ \frac{1}{\lambda_{\max}} \right)^T$$

$$\underline{x}_k = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^k \left(\frac{1}{\lambda_{\min}}\ 0\ 0 \cdots 0\ (-1)^k \frac{1}{\lambda_{\max}} \right)^T$$

Zigzagging or hemstitching behavior typical of steepest descent

Key to improved convergence: Make $\kappa(H_k^{1/2} Q H_k^{1/2}) \approx 1$; In general $\kappa(H_k^{1/2} \nabla^2 f(\underline{x}_k) H_k^{1/2}) \approx 1$ $\Rightarrow H_k = [\nabla^2 f(\underline{x}_k)]^{-1}$

Diagonal scaling: $H_k = Diag([d^2 f / dx_i^2]^{-1})$

Also, $s = 1$ will generally work with Armijo step size rule

Stopping Criteria - 1

- **Gradient Related**

know $\|\nabla f\| = 0$ at $\underline{x}^*$, check $\|\nabla \underline{f}^T \nabla \underline{f}\|_2 \leq \varepsilon$

Problem: $\|\nabla f(\underline{x})\|$ strongly depends on the scaling of both f and $\underline{x}$

If $f \in (10^{-7}, 10^{-5})$ $\forall \underline{x}$, then condition may be satisfied for all $\underline{x}$

If $f \in (10^{5}, 10^{7})$ $\forall \underline{x}$, then condition may never be satisfied.

Alternative 1: $\|\nabla \underline{f}^T(\underline{x}_k) \nabla^2 \underline{f}^{-1}(\underline{x}_k) \nabla \underline{f}(\underline{x}_k)\| \leq \varepsilon$ good, but needs Hessian.

Alternative 2: Relative gradient of f at $\underline{x}_k$: $\dfrac{\Delta f / f}{\Delta x / x}$ "BODE SENSITIVITY"

$$\text{component } i : \frac{(\partial f / \partial x_i)|x_i|}{|f|} \Rightarrow \max_i \frac{(\partial f / \partial x_i)|x_i|}{|f|} \leq \varepsilon$$

$$\text{what if } f \text{ or } x_i = 0 \Rightarrow \max_i (\partial f / \partial x_i) \frac{\max\{|x_i|, \text{typical } x_i\}}{\max\{|f|, \text{typical } f\}} \leq \varepsilon$$

Stopping Criteria - 2

- **Variable Related Test**

$$\frac{\|\underline{x}_{k+1}-\underline{x}_k\|_\infty}{\max\{\|\underline{x}_k\|,\ \max_i\ \text{typical}\ x_i\}} \le \varepsilon_k \approx 10^{-6} \sim 10^{-7}$$

- **Put a limit on max number of iterations**

- **Put a limit on maximum step length**

$$\alpha_k \le \frac{1}{\lambda_{\min}(\nabla^2 f(\underline{x}_k))}$$

$$\alpha_{\max} \approx 1{,}000\|\underline{x}_0\|_\infty$$

- **Function related**

$$\| f_{k+1} - f_k \| \le \varepsilon \| f_k \|$$

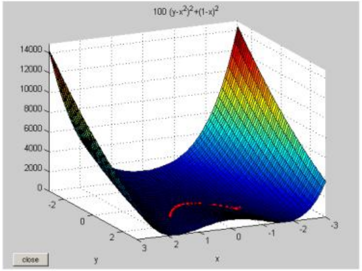
Some Test Examples

Some Test Examples:

1. $f(\underline{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1^2)^2$ Rosenbrock's Banana Function

   opt. (1,1)

   start at (−1.2,1)



2. Gear train inertia problem

$$f(\underline{x}) = [12 + x_1^2 + \frac{1 + x_2^2}{x_1^2} + \frac{x_1^{\ 2} x_2^{\ 2} + 100}{(x_1 x_2)^4}]\frac{1}{10}; \qquad \underline{x}_0 = (.5, 5)$$

3. Wood's function

$$f(\underline{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3^2)^2 + (1 - x_3)^2$$
$$+ 10 \cdot 1[(x_2 - 1)^2 + (x_4 - 1)^2] + 19.8(x_2 - 1)(x_4 - 1)$$

4. Himmelblau function

$$f(\underline{x}) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

5. Also see pp. 47, 53, 79 of Bertsekas' book

6. Lot of test functions on the web

References for Test Examples

- **More Test Example From**
  - R. Sargent and D.J. Sebastian, “Numerical experience with algorithms for unconstrained minimization”, in *Numerical methods for nonlinear optimization*, F.A. Lootsma, Academic, 1971
  - Crowder, H., R.S. Dembo and J.M. Mulvey, “On repenting computational experiments with Math software”, *ACM trans on Math software*, vol. 5, no. 2, 198-203, 1979
  - Carpenter, W.C. and E.A. Smith, “Computational efficiency in structural optimization”, *Eng. Optimization*, vol. 1, no. 3, 169-188, 1975
  - Miele, A. and S. Gonzalez, “On the comparative evaluation of algorithms for math programming problems”, in *NLP III*, Mangasarian, Meyer and Robinson, eds, Academic, 1978, 337-359
  - Shanno, D.F. and K.H. Phua, “Numerical comparison of several variable metric algorithms”, *JOTA*, vol. 25 no. 4, 507-518, 1978
  - Find some more references on the web!!!

Summary

- Quadratic Interpolation
  - Super-linear convergence
- Combined Golden Section and Quadratic Interpolation
- Combined Armijo Rule and Quadratic Interpolation
- Convergence of Generalized Gradient Method
  - Larger the condition number, slower is the convergence
  - Scale the gradient so that condition number is close to 1 to improve convergence
- Stopping Criteria