$$\begin{aligned} \textit{minimize} \quad & f(x) \\ \textit{s.t.} \quad & x \in E \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function and E is the feasible region.

Let $f : \mathbb{R}^n \to \mathbb{R}$. If $f \in C^1$ then

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_r}(x) \\ \vdots \\ \frac{\partial f}{\partial x_{n-1}}(x) \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

- Derivative of f at a point $x := Df(x) = [\nabla f(x)]^T$.
- sometimes we will use $\nabla f(x) = g(x)$ and $\nabla f(x_k) = g(x_k) = g_k$

Let $f : \mathbb{R}^n \to \mathbb{R}$. If $f \in C^2$ then Hessian of $f(x)$ at point x is $H(x) = \nabla g(x)^T = \nabla\{\nabla^T f(x)\}$. Hence

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}$$

- $H(x)$ is an $n \times n$ square symmetric matrix.

Gradient and Hessian Information

Let $f : \mathbb{R}^n \to \mathbb{R}^m$. If $f \in C^1$ then Jacobian of $f(x)$ at point x is $J(x) = Df(x)$ and is defined as

$$J(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$$

- Here $x = [x_1, x_2, x_3, \cdots, x_n]^T$ and $f(x) = [f_1(x), f_2(x), f_3(x), \cdots, f_m(x)]^T$
- $J(x)$ is an $m \times n$ matrix.

- Let $g : \Omega \subset \mathbb{R}^n \to \mathbb{R}$.
  Let $f : (a, b) \to \Omega$.
  Define $h : (a, b) \to \mathbb{R}$ by $h(t) = g(f(t))$
  Then $Dh(t) = \frac{dh}{dt} = [Dg(f(t))][Df(t)]$, i.e. $=< Dg(f(t)), [Df(t)] >$
  $= \left[\frac{\partial g}{\partial x_1}(f(t)) \quad \frac{\partial g}{\partial x_2}(f(t)) \quad \cdots \quad \frac{\partial g}{\partial x_n}(f(t)) \right] \cdot \left[\frac{\partial f_1}{\partial t} \quad \frac{\partial f_2}{\partial t} \quad \cdots \quad \frac{\partial f_n}{\partial t} \right]^T$
- Let $f, g : \mathbb{R}^n \to \mathbb{R}^m$.
  Define $h : \mathbb{R}^n \to \mathbb{R}$ by $h(x) = [f(x)]^T g(x)$
  Then $Dh(x) = [f(x)]^T Dg(x) + [g(x)]^T Df(x)$. Particularly
  - $D(y^T Ax) = y^T A$
  - $D(x^T Ax) = x^T A + x^T A^T = 2x^T A$(if A is symmetric)
  - $D(y^T x) = y^T$
  - $D(x^T x) = x^T + x^T = 2x^T$

Definition

The level set of a function $f : \mathbb{R}^n \to \mathbb{R}$ *at level* c *is the set* $\{x \in \mathbb{R}^n : f(x) = c\}$.

Particularly

- for $f : \mathbb{R}^2 \to \mathbb{R}$, it is a curve.
- for $f : \mathbb{R}^3 \to \mathbb{R}$, it is a surface.

Let $f : \mathbb{R}^n \to \mathbb{R}$.

- Directional derivative of f at a point x in the direction of d is $< g(x), d >$.

Problem:

$$\begin{aligned} \textit{minimize} \quad & f(x) \\ s.t. \quad & x \in \mathbb{R}^n \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function.

- Directional derivative of f at a point x in the direction of d is $< g(x), d >$.
- The gradient g and its negative $-g$ are the steepest ascent and steepest descent directions.

Input: Function $f(x)$ and initial guess x^0.

Until Convergence Do:

iteration:

$$x^{k+1} = x^k + \alpha_k d(x^k) \quad (1)$$

where x^k is the current estimate of a local minimizer to the problem. $d(x^k)$: is the current search direction in the $\mathbb{R}^n$. In above equation:

$$d(x^k) = -M_k g^k \quad (2)$$

where $g^k = \nabla f(x^k)$ and M_k is $n \times n$ matrix. And, α_k (step size) is the minimizer of a function $\phi : \mathbb{R}^+ \to \mathbb{R}$ defined as

$$\begin{aligned} \phi(\alpha) &= f(x^k + \alpha d(x^k)) \\ &= f(x^k - \alpha M_k g^k) \end{aligned}$$

- $g^k = 0$ (In computation gradient is rarely identically zero).
- $||g^k|| < \varepsilon$ for some pre specified small positive value ε.
- $||f(x^{k+1}) - f(x^k)|| < \varepsilon$ or $||x^{k+1} - x^k|| < \varepsilon$.
- $\frac{||f(x^{k+1})-f(x^k)||}{||f(x^k)||} < \varepsilon$ or $\frac{||x^{k+1}-x^k||}{||x^k||} < \varepsilon$ (relative stoping criteria that is scale-independent).
- $\frac{||f(x^{k+1})-f(x^k)||}{\max\{1,||f(x^k)||\}} < \varepsilon$ or $\frac{||x^{k+1}-x^k||}{\max\{1,||x^k||\}} < \varepsilon$ (to avoid division by a small number).

Output: $x^* = x^{k+1}$ and $f^* = f(x^*)$

$$x^{k+1} = x^k + \alpha_k d(x^k)$$

where

$$d(x^k) = -M_k g^k$$

and, α_k (step size) is the minimizer of a function $\phi : \mathbb{R}^+ \to \mathbb{R}$ defined as

$$\begin{aligned} \phi(\alpha) &= f(x^k + \alpha d(x^k)) \\ &= f(x^k - \alpha M_k g^k) \end{aligned}$$

Different choice of matrix M_k gives a different name to our algorithm.

- $M_k = I_n$: Steepest Descent Method (SD).
- If M_k is generated by conjugate vectors with respect to an appropriate PD matrix: Conjugate Gradient Method (CG), Fletcher Reeves Method (FR).
- $M_k = [H(x^k)]^{-1} = H_k^{-1}$: Newton Method.
- If M_k is some approximation of Hessian inverse: Quasi-Newton Type Method.

$$x^{k+1} = x^k + \alpha_k d(x^k)$$

where

$$d(x^k) = -M_k g^k$$

and, α_k (step size) is the minimizer of a function $\phi : \mathbb{R}^+ \to \mathbb{R}$ defined as

$$\begin{aligned} \phi(\alpha) &= f(x^k + \alpha d(x^k)) \\ &= f(x^k - \alpha M_k g^k) \end{aligned}$$

Definition

Any iterative method as defined in the main frame-work is said to have descent property if $f(x^{k+1}) < f(x^k), \ \forall k$*, provided* $g^k \neq 0$.

Definition

Any iterative method as defined in the main frame-work is said to have quadratic termination property if the minimum of $f(x) = \frac{1}{2}x^T Ax - b^T x + C$*, A is* $n \times n$ *SPD matrix, is reached in at most n iteration,* $b \in \mathbb{R}^n$*, C is a constant.*

Definition

Any iterative method as defined in the main frame-work is said to be globally convergent if $x^k \to x^*$ (a local minimizer of f) for any initial guess x^0 as $k \to \infty$.

Definition

Any iterative method as defined in the main frame-work is said to have order of convergence p if there exists $0 < a < \infty$ such that

$$\lim_{k\to\infty} \frac{||x^{k+1} - x^*||}{||x^k - x^*||^p} = a$$

If $p = 1$ and $a = 0$ then it is called super linear convergent method.

Algorithm for SD method

$$\textit{minimize} \;\; f(x) \;\; s.t. \;\; x \in \mathbb{R}^n$$

$\surd$ Input $f(x)$, x^0, and ε (for stoping condition)

$\surd$ **Until Convergence Do**: for $k = 0, 1, 2, 3, \cdots$

Step 1 Calculate $g^k := \nabla f(x^k)$

Step 2 Set $d^k = -g^k$

Step 3 Find α_k, the value of α that minimizes $f(x^k + \alpha d^k)$

Step 4 Set $x^{k+1} = x^k + \alpha_k d^k$

$\surd$ EndDo

Theorem

SD is a descent method.

Proof:

$$x^{k+1} = x^k + \alpha_k d^k$$

where $d(x^k) = -g^k$ and, α_k is the minimizer of a function $\phi : \mathbb{R}^+ \to \mathbb{R}$ defined as $\phi(\alpha) = f(x^k + \alpha d^k)$. Since

$$\begin{aligned} \frac{d\phi}{d\alpha}|_{\alpha=0} &= [Df(x^k + \alpha d(x^k))|_{\alpha=0}]d^k \\ &= [g^k]^T d^k = -[d^k]^T d^k = -||d^k||^2 < 0 \end{aligned}$$

So there exists $\overline{\alpha} > 0$ such that $\phi(\alpha) < \phi(0) \ \forall \ 0 < \alpha \leq \overline{\alpha}$. Hence, we have

$$\phi(\alpha_k) = f(x^k + \alpha_k d^k) = f(x^{k+1}) \leq \phi(\alpha) < \phi(0) = f(x^k).$$

Apply Basic frame-work to Quadratic problem

Problem: minimize $f(x) = \frac{1}{2}x^T A x - b^T x$, A is SPD.

Step 1 Calculate $g^k := \nabla f(x^k)$.
Here: $g^k = Ax^k - b = r^k$ (say)

Step 2 Set $d^k = -M_k g^k$. **Here:** $d^k = -M_k r^k$.

Step 3 Find α_k, the value of α that minimizes $\phi(\alpha) := f(x^k + \alpha M_k d^k)$.
Here: $\alpha_k = -\frac{< M_k d^k,\ r^k >}{< M_k d^k,\ A M_k d^k >}$

Step 4 Set $x^{k+1} = x^k + \alpha_k M_k d^k$.
Here: $x^{k+1} = x^k - \frac{< M_k d^k,\ r^k >}{< M_k d^k,\ A M_k d^k >} M_k d^k$

Problem: minimize $f(x) = \frac{1}{2}x^T Ax - b^T x$, A is SPD.

Step 1 Calculate $g^k := \nabla f(x^k)$. **Here:** $g^k = Ax^k - b = r^k$ (say)

Step 2 Set $d^k = -g^k$. **Here:** $d^k = -r^k$.

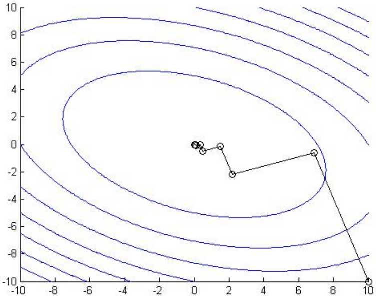
Step 3 Find α_k, the value of α that minimizes $\phi(\alpha) := f(x^k + \alpha d^k)$.
Here: $\alpha_k = \frac{<r^k,\, r^k>}{<r^k,\, Ar^k>}$

Step 4 Set $x^{k+1} = x^k + \alpha_k d^k$. **Here:** $x^{k+1} = x^k - \frac{<r^k,\, r^k>}{<r^k,\, Ar^k>} r^k$

Problem: minimize $x_1^2 + x_1 x_2 + 2x_2^2$.

Here: minimize $f(x) = \frac{1}{2}x^T Ax - b^T x$, $b = 0$ and $A = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$.

$$\begin{pmatrix} Iter & x_1 & x_2 \\ 1 & 10.0000 & -10.0000 \\ 2 & 6.8749 & -0.6247 \\ 3 & 2.1875 & -2.1875 \\ 10 & 0.0157 & -0.0014 \\ 12 & 0.0034 & -0.0003 \\ 15 & 0.0002 & -0.0002 \\ 18 & 0.0000 & -0.0000 \end{pmatrix}$$



Theorem

SD moves in orthogonal directions

Proof:

$$x^{k+1} = x^k + \alpha_k d^k, \quad x^{k+2} = x^{k+1} + \alpha_{k+1} d^{k+1}$$

So,

$$< x^{k+2} - x^{k+1}, x^{k+1} - x^k >= \alpha_k \alpha_{k+1} < d^{k+1}, d^k > \tag{3}$$

Since, α_k is the minimizer of a function $\phi : \mathbb{R}^+ \to \mathbb{R}$ defined as $\phi(\alpha) = f(x^k + \alpha d^k)$. So

$$\begin{aligned} \frac{d\phi}{d\alpha}|_{\alpha=\alpha_k} &= [Df(x^k + \alpha d(x^k))|_{\alpha=\alpha_k}]d^k \\ &= [g^{k+1}]^T d^k = -[d^{k+1}]^T d^k = 0 \end{aligned} \tag{4}$$

Above two equations imply the conclusion.

Theorem

SD is a globally convergent method.

SD convergence rate is linear for quadratic functions

For simplicity, we take

Problem: minimize $f(x) = \frac{1}{2}x^T Ax$, A is SPD. It is known here that $x^* = 0$, $f^* = f(x^*) = 0$. Let max and min eigenvalues of matrix A are B and b respectively.

We know that here SD generates the following sequence for minimizer: $x^{k+1} = x^k - \alpha_k r^k$ where $\alpha_k = \frac{<r^k,\, r^k>}{<r^k,\, Ar^k>}$ and $r^k = \nabla f(x^k) = Ax^k$.

We can show easily that $\frac{f(x^{k+1})}{f(x^k)} = 1 - \beta$ where $\beta = \frac{<r^k, r^k>^2}{<r^k, Ar^k><r^k, A^{-1}r^k>}$.

By Rayleigh inequality:

$\frac{b}{2}||x^{k+1}||^2 \leq f(x^{k+1})$ and $\frac{B}{2}||x^k||^2 \geq f(x^k)$. Hence

$\frac{b}{2}||x^{k+1}||^2 \leq f(x^{k+1}) = (1-\beta)f(x^k) \leq (1-\beta)\frac{B}{2}||x^k||^2$

$\Rightarrow \frac{||x^{k+1}||^2}{||x^k||^2} = \frac{B}{b}(1-\beta) \leq \frac{B}{b}\frac{B-b}{B}$ (By Kantorovich inequality). Hence

$$\frac{||x^{k+1}||}{||x^k||} \leq \sqrt{\frac{B-b}{b}}$$

$$\textit{minimize } f(x) \quad s.t. \ x \in \mathbb{R}^n$$

$\surd$ Input $f(x)$, x^0, and ε (for stoping condition)

$\surd$ **Until Convergence Do**: for $k = 0, 1, 2, 3, \cdots$

Step 1 Calculate $g^k := \nabla f(x^k)$

Step 2 Set $d^k = -[H(x^k)]^{-1} g^k$

Step 3 Set $x^{k+1} = x^k + d^k$

$\surd$ EndDo

Idea:

Let x^k be the current estimator of x^*. Take quadratic approximation q of f near x^k and find its minima instead of f. Minima of q is treated as the next estimator x^{k+1} for the x^*. Thus,

$q(x) = f(x^k) + g(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T H(x^k)(x - x^k)$. This gives:

$Dq(x) = g(x^k)^T + \frac{1}{2}((x - x^k)^T H(x^k) + (x - x^k)^T H(x^k)^T)$

$\quad = g(x^k)^T + (x - x^k)^T H(x^k)$. This implies:

$\nabla q(x) = g(x^k) + H(x^k)(x - x^k) = 0$ gives the step 3.

minimize $f(x) = \frac{1}{2}x^T Ax - b^T x$, A is SPD.

Until Convergence Do:

Step 1 Calculate $g^k := \nabla f(x^k)$

Step 2 Set $d^k = -[H(x^k)]^{-1} g^k$

Step 3 Set $x^{k+1} = x^k + d^k$

EndDo

First iteration:

Step 1 $g^0 := \nabla f(x^0) = A(x^0) - b$

Step 2
$$\begin{aligned} d^0 &= -A^{-1} g^0 \\ &= -A^{-1}(A(x^0) - b) \\ &= -x^0 + A^{-1} b \end{aligned}$$

Step 3
$$\begin{aligned} x^1 &= x^0 - x^0 + A^{-1} b \\ &= A^{-1} b \end{aligned}$$

$\surd$ $g^1 = A(x^1) - b = 0$ STOP

Newton's Method give exact answer in only one iteration always!

Why?

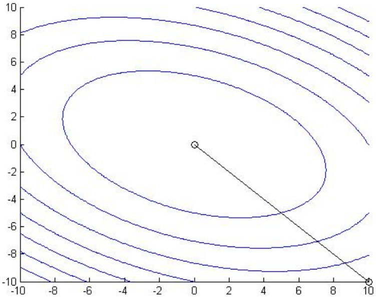
quadratic Approximation of a quadratic function.

Newton's Method satisfies quadratic termination property

Problem: minimize $x_1^2 + x_1 x_2 + 2x_2^2$.

Here: minimize $f(x) = \frac{1}{2}x^T A x - b^T x$, $b = 0$ and $A = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$.

$$\begin{pmatrix} \mathit{Iter} & x_1 & x_2 \\ 0 & 10.0000 & -10.0000 \\ 1 & 0 & 0 \end{pmatrix}$$



(Any nonlinear function: not quadratic)

$$minimize \;\; f(x) \;\; s.t. \;\; x \in \mathbb{R}^n$$

- $\surd$ Input $f(x)$, x^0, and ε (for stoping condition)
- $\surd$ **Until Convergence Do**: for $k = 0, 1, 2, 3, \cdots$
- **Step 1** Calculate $g^k := \nabla f(x^k)$
- **Step 2** Set $d^k = -[H(x^k)]^{-1} g^k$
- **Step 3** Set $x^{k+1} = x^k + d^k$
- $\surd$ EndDo

Problems:

1. $H(x^k)$ must be invertible for each iteration.
2. Above scheme is not descent for any arbitrary x_0 (Why?).

$$\textit{minimize} \;\; f(x) \;\; s.t. \;\; x \in \mathbb{R}^n$$

Modified Newton's Method:

$\surd$ Input $f(x)$, x^0, and ε (for stoping condition)

$\surd$ **Until Convergence Do**: for $k = 0, 1, 2, 3, \cdots$

Step 1 Calculate $g^k := \nabla f(x^k)$

Step 2 Set $d^k = -[H(x^k)]^{-1} g^k$ OR Solve $H(x^k) d^k = -g^k$.

Step 3 Find α_k, the value of α that minimizes $f(x^k + \alpha d^k)$

Step 4 Set $x^{k+1} = x^k + \alpha_k d^k$

$\surd$ EndDo

Theorem

If Hessian is SPD in each iteration, then Modified Newton Method is descent.

Theorem

If Hessian is SPD in each iteration, then Modified Newton Method is descent.

Proof:

$$x^{k+1} = x^k + \alpha_k d^k$$

where $d(x^k) = -[H(x^k)]^{-1} g^k$ and, α_k is the minimizer of a function $\phi : \mathbb{R}^+ \to \mathbb{R}$ defined as $\phi(\alpha) = f(x^k + \alpha d^k)$. Since

$$\begin{aligned} \frac{d\phi}{d\alpha}\Big|_{\alpha=0} &= [Df(x^k + \alpha d(x^k))|_{\alpha=0}] d^k \\ &= [g^k]^T d^k = [d^k]^T g^k = -[d^k]^T H(x^k) d^k < 0 \end{aligned}$$

So there exists $\overline{\alpha} > 0$ such that $\phi(\alpha) < \phi(0) \;\; \forall\; 0 < \alpha \leq \overline{\alpha}$.
Hence, we have

$$\phi(\alpha_k) = f(x^k + \alpha_k d^k) = f(x^{k+1}) \leq \phi(\alpha) < \phi(0) = f(x^k).$$

$$\textit{minimize} \;\; f(x) \;\; s.t. \;\; x \in \mathbb{R}^n$$

Levenberg-Marquardt Method:

$\surd$ Input $f(x)$, x^0, and ε (for stoping condition)

$\surd$ **Until Convergence Do**: for $k = 0, 1, 2, 3, \cdots$

Step 1 Calculate $g^k := \nabla f(x^k)$

Step 2 Solve $[H(x^k) + \mu_k I]d^k = -g^k$ for sufficiently large μ_k.

Step 3 Find α_k, the value of α that minimizes $f(x^k + \alpha d^k)$

Step 4 Set $x^{k+1} = x^k + \alpha_k d^k$

$\surd$ EndDo

$$\textit{minimize} \;\; f(x) \;\; s.t. \;\; x \in \mathbb{R}^n$$

Levenberg-Marquardt Method:

- $\surd$ Input $f(x)$, x^0, and ε (for stoping condition)
- $\surd$ Calculate g^0, $H(x^0)$, and set $\mu = 10^6$ (large number), dif = TRUE.
- $\surd$ **while** $||g^K|| < \varepsilon$

Step 1 if (dif) $\mu = 2\mu$; else $\mu = \frac{\mu}{2}$.

Step 2 Solve $[H(x^k) + \mu_k I]d^k = -g^k$.

Step 3 Set $x^{k+1} = x^k + d^k$

Step 4 Calculate f^{k+1}, g^{k+1}, H^{k+1}

Step 5 if $(f(x^{k+1}) < f(x^k))$ dif = FALSE; else dif = TRUE.

Conjugate Direction

Definition

Let A be any SPD $n \times n$ matrix. Then Vectors $p^0, p^1, p^2, \cdots, p^{n-1}$ are called $A-$conjugate directions iff $< p^i, Ap^j >= 0, \forall i \neq j$.

- This definition generalize the orthogonality.
- Above n directions are LI in $\mathbb{R}^n$.

$$\text{minimize} \;\; f(x) \;\; s.t. \;\; x \in \mathbb{R}^n$$

- $\surd$ Input $f(x)$, x^0, n $A-$conjugate directions $p^0, p^1, p^2, \cdots, p^{n-1}$ and ε (for stoping condition)
- $\surd$ **Until Convergence Do**: for $k = 0, 1, 2, 3, \cdots$

Step 1 Find α_k, the value of α that minimizes $f(x^k + \alpha p^k)$

Step 2 Set $x^{k+1} = x^k + \alpha_k p^k$

- $\surd$ EndDo

Idea:
Instead of negative gradient direction, move in the conjugate directions.

minimize $f(x) = \frac{1}{2}x^TAx - b^Tx$, A is SPD.

$\surd$ Input $f(x)$, x^0, n $A-$conjugate directions $p^0, p^1, p^2, \cdots, p^{n-1}$ and ε (for stoping condition)

$\surd$ **Until Convergence Do**: for $k = 0, 1, 2, 3, \cdots$

Step 1 Find α_k, the value of α that minimizes $f(x^k + \alpha p^k)$.
Here: $\alpha_k = -\frac{<p^k,\ r^k>}{<p^k,\ Ap^k>}$, where $r^k = \nabla f(x^k) = Ax^k - b$.

Step 2 Set $x^{k+1} = x^k + \alpha_k p^k$. **Here:** $x^{k+1} = x^k - \frac{<p^k,\ r^k>}{<p^k,\ Ap^k>}p^k$.

$\surd$ EndDo

Proof:

$\phi(\alpha) = f(x^k + \alpha p^k) = \frac{1}{2}(x^k + \alpha p^k)^T A(x^k + \alpha p^k) - b^T(x^k + \alpha p^k)$. Thus

$$\begin{aligned}\frac{d\phi}{d\alpha} &= \frac{1}{2}[(p^k)^TA(x^k + \alpha p^k) + (x^k + \alpha p^k)^TAp^k] - b^Tp^k \\ &= (p^k)^TAx^k + \alpha(p^k)^TAp^k - (p^k)^Tb.\end{aligned}$$

$\frac{d\phi}{d\alpha} = 0$ gives $\alpha = \alpha_k$.

Basic Conjugate Direction Method satisfies quadratic termination property

minimize $f(x) = \frac{1}{2}x^T Ax - b^T x$, A is SPD.

$\surd$ Input $f(x)$, x^0, n $A-$conjugate directions $p^0, p^1, p^2, \cdots, p^{n-1}$ and ε (for stoping condition)

Step 1 $\alpha_k = -\frac{<p^k,\, r^k>}{<p^k,\, Ap^k>}$, where $r^k = \nabla f(x^k) = Ax^k - b$.

Step 2 Set $x^{k+1} = x^k + \alpha_k p^k$.

Proof: By iterative step:

$$
\begin{aligned}
x^1 &= x^0 + \alpha_0 p^0 \\
x^2 &= x^1 + \alpha_1 p^1 = x^0 + \alpha_0 p^0 + \alpha_1 p^1 \\
x^k &= x^0 + \alpha_0 p^0 + \alpha_1 p^1 + \alpha_2 p^2 + \cdots + \alpha_{k-1} p^{k-1} \qquad (5) \\
x^n &= x^0 + \alpha_0 p^0 + \alpha_1 p^1 + \alpha_2 p^2 + \cdots + \alpha_{n-1} p^{n-1} \qquad (6)
\end{aligned}
$$

We have to show that $x^n = x^*$. Since $x^* \in \mathbb{R}^n$ and $p^0, p^1, p^2, \cdots, p^{n-1}$ is a basis of $x^* \in \mathbb{R}^n$, there exists λ_i's such that

$$x^* - x^0 = \lambda_0 p^0 + \lambda_1 p^1 + \lambda_2 p^2 + \cdots + \lambda_{n-1} p^{n-1} \qquad (7)$$

From (6) and (7), it is clear that to prove $x^n = x^*$, it is sufficient to show that $\lambda_k = \alpha_k$.

By premultiplying (7) with $(p^k)^T A$ we have

$$\lambda_k = \frac{< p^k, A(x^* - x^0) >}{< p^k, Ap^k >} \tag{8}$$

Now, $x^* - x^0 = (x^* - x^k) + (x^k - x^0)$ implies

$$\begin{aligned} < p^k, A(x^* - x^0) > &= < p^k, A(x^* - x^k) > + < p^k, A(x^k - x^0) > \\ &= < p^k, A(x^* - x^k) > \\ &\quad (\because \text{second term is zero by (5)}) \\ &= < p^k, b - Ax^k > = - < p^k, r^k > \end{aligned} \tag{9}$$

From (8) and (9), it is clear that $\lambda_k = \alpha_k$

One more important property of the Basic Conjugate Direction Method for quadratic problem

minimize $f(x) = \frac{1}{2}x^T Ax - b^T x$, A is SPD.

$\surd$ Input $f(x)$, x^0, n $A-$conjugate directions $p^0, p^1, p^2, \cdots, p^{n-1}$ and ε (for stoping condition)

Step 1 $\alpha_k = -\frac{<p^k,\, r^k>}{<p^k,\, Ap^k>}$, where $r^k = \nabla f(x^k) = Ax^k - b$.

Step 2 Set $x^{k+1} = x^k + \alpha_k p^k$.

Theorem : New residual $\perp$ all old conjugate directions

For any $k \in \{0, 1, 2, \cdots, n-1\}$, $< r^{k+1}, p^i >= 0 \; \forall i = 0, 1, 2, \cdots, k$.

Proof: For $k = 0$, see

$$\begin{aligned} < r^1, p^0 > &= < Ax^1 - b, p^0 >=< Ax^0 + \alpha_0 Ap^0 - b, p^0 > \\ &= < Ax^0 - b + \alpha_0 Ap^0, p^0 >=< r^0 + \alpha_0 Ap^0, p^0 > \\ &= < r^0, p^0 > + \alpha_0 < Ap^0, p^0 >= 0. \end{aligned}$$

Last equality is due to the definition of α_0.

One more important property of the Basic Conjugate Direction Method for quadratic problem

Now Assume that for some $k-1$ and $i = 0, 1, 2, \cdots, k-1$ statement is correct, i.e., $< r^k, p^i >= 0 \, \forall i = 0, 1, 2, \cdots, k-1$. Then for k and $i = 0, 1, 2, \cdots, k-1$, we have

$$
\begin{aligned}
< r^{k+1}, p^i > &= < r^k + \alpha_k A p^k, p^i > \\
&\quad \text{(by the iteration used in the algo: } r^{k+1} = r^k + \alpha_k A p^k) \\
&= < r^k, p^i > + \alpha_k < A p^k, p^i >= 0 \\
&\quad \text{(first term is zero by the assumption of induction)} \\
&\quad \text{(and second is zero by the conjugacy of } p^i \, vectors.)
\end{aligned}
$$

Now its remains to show that $< r^{k+1}, p^k >= 0$. To see it

$$
\begin{aligned}
< r^{k+1}, p^k > &= < r^k + \alpha_k A p^k, p^k > \\
&= < r^k, p^k > + \alpha_k < A p^k, p^k >= 0.
\end{aligned}
$$

Last equality is due to the definition of α_k.

Problem in consideration: minimize $f(x) = \frac{1}{2}x^T Ax - b^T x$, A is SPD.
Iteration step in Basic Conjugate Direction Method:
$x^{k+1} = x^k - \alpha_k p^k$, where $\alpha_k = \frac{<p^k,\, r^k>}{<p^k,\, Ap^k>}$ and $r^k = Ax^k - b$.
Construction of p^i, $i = 0, 1, 2, \cdots, n-1$:

$$p^0 = -r^0 \tag{10}$$
$$p^k = -r^k + \beta_{k-1} p^{k-1},\ k = 1, 2, \cdots, n-1 \tag{11}$$

where

$$\text{where } \beta_{k-1} = \frac{< r^k,\, Ap^{k-1} >}{< p^{k-1},\, Ap^{k-1} >},\ k = 1, 2, \cdots, n-1 \tag{12}$$

This choice of β_{k-1} in equation (11) gives the following facts:

Gradient vectors r^k, $k = 0, 1, 2, \cdots, n-1$ are mutually orthogonal

For any $k \in \{0, 1, 2, \cdots, n-1\}$, $< r^{k+1},\, r^i >= 0\ \forall i = 0, 1, 2, \cdots, k$.

p^k, $k = 0, 1, 2, \cdots, n-1$ are mutually A conjugate directions

For any $k \in \{0, 1, 2, \cdots, n-1\}$, $< p^{k+1},\, Ap^i >= 0\ \forall i = 0, 1, 2, \cdots, k$.

Theorem: Gradient vectors r^k, $k = 0, 1, 2, \cdots, n-1$ are mutually orthogonal

Proof: for arbitrary i, using equation (11) we have

$$\begin{aligned} < r^{k+1}, r^i > &= < r^{k+1}, -p^i + \beta_{i-1} p^{i-1} > \\ &= - < r^{k+1}, p^i > + \beta_{i-1} < r^{k+1}, p^{i-1} > \\ &= 0, \end{aligned} \tag{13}$$

by the previous result that, for any $k \in \{0, 1, 2, \cdots, n-1\}$, $< r^{k+1}, p^i >= 0 \; \forall i = 0, 1, 2, \cdots, k$.

Proof: p^1 and p^0 are A conjugate by the definition of β_0. Now, we show that p^2 is A conjugate to p^1 and p^0. It is clear that p^2 and p^1 are A conjugate by the definition of β_1.

$$
\begin{aligned}
< p^2, Ap^0 > &= < -r^2 + \beta_1 p^1, Ap^0 >= - < r^2, Ap^0 > + \beta_1 < p^1, Ap^0 > \\
&= - < r^2, Ap^0 > (\because \text{second term } < p^1, Ap^0 >= 0) \\
&= - < r^2, \frac{r^1 - r^0}{\alpha_0} >= \frac{1}{\alpha_0}[< r^2, r^0 > - < r^2, r^1 >] \\
& \quad \because \text{iterative step } x^{k+1} = x^k + \alpha_k p^k \Rightarrow Ap^k = \frac{r^{k+1} - r^k}{\alpha_k} \\
&= 0 \text{ by the previous result.}
\end{aligned}
$$

Now Assume that for some $k-1$ and $i = 0, 1, 2, \cdots, k-1$ statement is correct, i.e., $< p^k, p^i >= 0 \, \forall i = 0, 1, 2, \cdots, k-1$. Then for k and $i = 0, 1, 2, \cdots, k-1$, we have

$$\begin{aligned}
< p^{k+1}, Ap^i > &= < -r^{k+1} + \beta_k p^k, Ap^i > \\
&= - < r^{k+1}, Ap^i > + \beta_k < p^k, Ap^i > \\
&\quad \text{(second term is zero by the assumption of induction)} \\
&= - < r^{k+1}, \frac{r^{i+1} - r^i}{\alpha_i} > \\
&= \frac{1}{\alpha_i}[< r^{k+1}, r^i > - < r^{k+1}, r^{i+1} >] \\
&\quad \because \text{iterative step } x^{k+1} = x^k + \alpha_k p^k \Rightarrow Ap^k = \frac{r^{k+1} - r^k}{\alpha_k} \\
&= 0 \text{ by the previous result.}
\end{aligned}$$

Conjugate Gradient Method for the quadratic Problem

minimize $f(x) = \frac{1}{2}x^T Ax - b^T x$, A is SPD.

Step 1 Set $k = 0$, and select the initial point x^0.

Step 2 $g^0 = \nabla f(x^0) = Ax^0 - b = r^0$. If $g^0 = 0$ STOP else $p^0 = -g^0$.

Step 3 $\alpha_k = -\frac{<p^k,\, g^k>}{<p^k,\, Ap^k>}$.

Step 4 $x^{k+1} = x^k + \alpha_k p^k$

Step 5 $g^{k+1} = \nabla f(x^{k+1}) = Ax^{k+1} - b = r^{k+1}$. If $g^{k+1} = 0$ STOP.

Step 6 $\beta_k = \frac{<g^{k+1}, Ap^k>}{<p^k, Ap^k>}$

Step 7 $p^{k+1} = -g^{k+1} + \beta_k p^k$

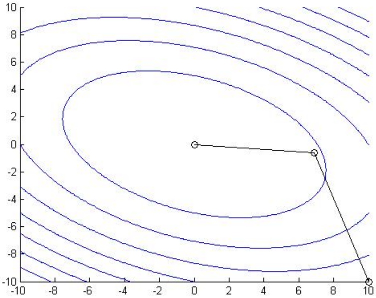
Step 8 Set $k = k + 1$ and go to Step 3.

Since CG method is a conjugate direction method so minimizes the given problem in at most n steps, where n is the order of the matrix A.

Problem: minimize $x_1^2 + x_1 x_2 + 2x_2^2$.

Here: minimize $f(x) = \frac{1}{2}x^T A x - b^T x$, $b = 0$ and $A = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$.

$$\begin{pmatrix} \mathit{Iter} & x_1 & x_2 \\ 0 & 10.0000 & -10.0000 \\ 1 & 6.8750 & -0.6250 \\ 2 & 0 & 0 \end{pmatrix}$$



minimize $f(x) : x \in \mathbb{R}^n$.

Step 1 Set $k = 0$, and select the initial point x^0.

Step 2 $g^0 = \nabla f(x^0) = Ax^0 - b = r^0$. If $g^0 = 0$ STOP else $p^0 = -g^0$.

Step 3 $\alpha_k = -\frac{<p^k,\, g^k>}{<p^k,\, Ap^k>}$.

Step 4 $x^{k+1} = x^k + \alpha_k p^k$

Step 5 $g^{k+1} = \nabla f(x^{k+1}) = Ax^{k+1} - b = r^{k+1}$. If $g^{k+1} = 0$ STOP.

Step 6 $\beta_k = \frac{<g^{k+1}, Ap^k>}{<p^k, Ap^k>}$

Step 7 $p^{k+1} = -g^{k+1} + \beta_k p^k$

Step 8 Set $k = k + 1$ and go to Step 3.

minimize $f(x) : x \in \mathbb{R}^n$.

Step 1 Set $k = 0$, and select the initial point x^0.

Step 2 $g^0 = \nabla f(x^0) = Ax^0 - b = r^0$. If $g^0 = 0$ STOP else $p^0 = -g^0$.

Step 3 $\alpha_k = -\frac{<p^k,\, g^k>}{<p^k,\, Ap^k>}$.

Step 4 $x^{k+1} = x^k + \alpha_k p^k$

Step 5 $g^{k+1} = \nabla f(x^{k+1}) = Ax^{k+1} - b = r^{k+1}$. If $g^{k+1} = 0$ STOP.

Step 6 $\beta_k = \frac{<g^{k+1}, Ap^k>}{<p^k, Ap^k>}$.

Step 7 $p^{k+1} = -g^{k+1} + \beta_k p^k$

Step 8 Set $k = k + 1$ and go to Step 3.

minimize $f(x) : x \in \mathbb{R}^n$.

Step 1 Set $k = 0$, and select the initial point x^0.

Step 2 $g^0 = \nabla f(x^0)$. If $g^0 = 0$ STOP else $p^0 = -g^0$.

Step 3 $\alpha_k = -\frac{<p^k,\, g^k>}{<p^k,\, Ap^k>}$.

Step 4 $x^{k+1} = x^k + \alpha_k p^k$

Step 5 $g^{k+1} = \nabla f(x^{k+1}) = Ax^{k+1} - b = r^{k+1}$. If $g^{k+1} = 0$ STOP.

Step 6 $\beta_k = \frac{<g^{k+1}, Ap^k>}{<p^k, Ap^k>}$.

Step 7 $p^{k+1} = -g^{k+1} + \beta_k p^k$

Step 8 Set $k = k + 1$ and go to Step 3.

minimize $f(x) : x \in \mathbb{R}^n$.

Step 1 Set $k = 0$, and select the initial point x^0.

Step 2 $g^0 = \nabla f(x^0)$. If $g^0 = 0$ STOP else $p^0 = -g^0$.

Step 3 Find α_k, the value of α that minimizes $f(x^k + \alpha p^k)$.

Step 4 $x^{k+1} = x^k + \alpha_k p^k$

Step 5 $g^{k+1} = \nabla f(x^{k+1}) = Ax^{k+1} - b = r^{k+1}$. If $g^{k+1} = 0$ STOP.

Step 6 $\beta_k = \frac{<g^{k+1}, Ap^k>}{<p^k, Ap^k>}$.

Step 7 $p^{k+1} = -g^{k+1} + \beta_k p^k$

Step 8 Set $k = k + 1$ and go to Step 3.

minimize $f(x) : x \in \mathbb{R}^n$.

Step 1 Set $k = 0$, and select the initial point x^0.

Step 2 $g^0 = \nabla f(x^0)$. If $g^0 = 0$ STOP else $p^0 = -g^0$.

Step 3 Find α_k, the value of α that minimizes $f(x^k + \alpha p^k)$.

Step 4 $x^{k+1} = x^k + \alpha_k p^k$

Step 5 $g^{k+1} = \nabla f(x^{k+1})$. If $g^{k+1} = 0$ STOP.

Step 6 $\beta_k = \frac{<g^{k+1}, Ap^k>}{<p^k, Ap^k>}$.

Step 7 $p^{k+1} = -g^{k+1} + \beta_k p^k$

Step 8 Set $k = k + 1$ and go to Step 3.

Recall: $\beta_k = \frac{<g^{k+1}, Ap^k>}{<p^k, Ap^k>}$
Actually replacement of A in the above formula is $H(x^k)$. But fortunately, algebraic manipulation in the above formula is possible with the knowledge of the function value $f(x^k)$ and gradient value g^k.

Hestenes-Stiefel (SF) Modification:

It is clear that iterative step that $x^{k+1} = x^k + \alpha_k p^k \Rightarrow Ap^k = \frac{g^{k+1} - g^k}{\alpha_k}$
Use this value in the formula of β_k in place of Ap^k, thus:

$$\beta_k = \frac{< g^{k+1}, g^{k+1} - g^k >}{< p^k, g^{k+1} - g^k >}$$

Polak-Ribi$\grave{e}$re (PR) Modification:

In the SF formula of β_k, see denominator:
$< p^k, g^{k+1} - g^k > = - < p^k, g^k >$ (because new gradient is orthogonal to all old conjugate directions). Thus
$< p^k, g^{k+1} - g^k > = - < -g^k + \beta_{k-1} p^{k-1}, g^k > = < g^k, g^k >$ gives:

$$\beta_k = \frac{< g^{k+1}, g^{k+1} - g^k >}{< g^k, g^k >}$$

Fletcher - Reeves (FR) Modification:
Use the fact that all new gradients are orthogonal to the old ones in the numerator of the PR modification to get:

$$\beta_k = \frac{< g^{k+1}, g^{k+1} >}{< g^k, g^k >}$$

Important Remark: Stopping criterion.
For nonquadratic problems, the algorithm will not usually converge in n steps, and as the algorithm, the conjugate direction will tend to deteriorate. Thus a common practice is to reinitialize the direction vector to the negative of gradient after every few iteration (e.g. n).