Optimization Theory and Methods

Prof. Xiaoe Ruan

Email: wruanxe@xjtu.edu.cn

Tel:13279321898

Fall， 2019

Chap.4 Large-scale unconstrained Optimization Methods

Conjugate Gradient Method

Sparse Quasi-Newton Method*

Finite-memory Quasi-Newton Method*

Nonmemory Quasi-Newton Method*

1. Conjugate Gradient Method(CGM)

Recall1: Conjugate Gradient Method for $\min f(x) = \frac{1}{2} x^{\mathrm{T}} G x - b^{\mathrm{T}} x$

(1.1) Computing α_k **From** $f\left(x^{(k)} + \alpha_k d^{(k)}\right) = \min_{\alpha \geq 0} f\left(x^{(k)} + \alpha d^{(k)}\right)$

yields $$0 = \left.\frac{df\left(x^{(k)} + \alpha d^{(k)}\right)}{d\alpha}\right|_{\alpha=\alpha_k} = \nabla f\left(x^{(k)} + \alpha_k d^{(k)}\right)^{\mathrm{T}} d^{(k)}$$

$$= \left[G\left(x^{(k)} + \alpha_k d^{(k)}\right) - b\right]^{\mathrm{T}} d^{(k)} = -\left(b - Gx^{(k)}\right)^{\mathrm{T}} d^{(k)} + \alpha_k d^{(k)\mathrm{T}} G d^{(k)}$$

where $r^{(k)} = b - Gx^{(k)}$ **residual vector (Negative gradient)**

Then $$\alpha_k = \frac{\left(b - Gx^{(k)}\right)^{\mathrm{T}} d^{(k)}}{d^{(k)\mathrm{T}} G d^{(k)}} = \frac{r^{(k)\mathrm{T}} d^{(k)}}{d^{(k)\mathrm{T}} G d^{(k)}}$$

Conjugate Gradient Method(CGM):

$$\forall x^{(0)},$$

$$r^{(0)} = b - Gx^{(0)} = -\nabla f\left(x^{(0)}\right)$$

$$d^{(0)} = r^{(0)}, \quad \alpha_0 = \frac{r^{(0)\mathrm{T}} d^{(0)}}{d^{(0)\mathrm{T}} G d^{(0)}}$$

$$x^{(1)} = x^{(0)} + \alpha_0 d^{(0)}$$

$$r^{(1)} = b - Gx^{(1)}$$

$$d^{(1)} = r^{(1)} - \frac{r^{(1)\mathrm{T}} G d^{(0)}}{d^{(0)\mathrm{T}} G d^{(0)}} d^{(0)},$$

$$\alpha_1 = \frac{r^{(1)\mathrm{T}} d^{(1)}}{d^{(1)\mathrm{T}} G d^{(1)}}$$

$$x^{(2)} = x^{(1)} + \alpha_1 d^{(1)}$$

$$r^{(2)} = b - Gx^{(2)}$$

$$d^{(k)} = r^{(k)} - \frac{r^{(k)\mathrm{T}} G d^{(0)}}{d^{(0)\mathrm{T}} G d^{(0)}} d^{(0)} - \cdots - \frac{r^{(k)\mathrm{T}} G d^{(k-1)}}{d^{(k-1)\mathrm{T}} G d^{(k-1)}} d^{(k-1)}$$

$$\alpha_k = \frac{r^{(k)\mathrm{T}} d^{(k)}}{d^{(k)\mathrm{T}} G d^{(k)}}$$

$$x^* = x^{(n)} = x^{(n-1)} + \alpha_{n-1} d^{(n-1)}$$

（1.2） Properties of CGM:

Th.1. Let G be a RSPD matrix and $d^{(0)}, d^{(1)}, \cdots, d^{(k)} (k \le n)$ conjugate directions w.r.t G. $r^{(0)}, r^{(1)}, \cdots, r^{(k)} (k \le n-1)$ are nonzero residual directions of CGM.

Then

(1) $\left\langle r^{(k+1)}, d^{(j)} \right\rangle = 0, j = 0, 1, \cdots, k.$

(2) $\left\langle r^{(k)}, d^{(k)} \right\rangle = \left\langle r^{(k)}, r^{(k)} \right\rangle.$

(3) $r^{(0)}, r^{(1)}, \cdots, r^{(k)} (k \le n-1)$ are orthogonal.

(1) Conjugate Gradient Method for PDQ fcn

$$\forall x^{(0)},$$
$$r^{(0)} = b - Gx^{(0)} = -\nabla f\left(x^{(0)}\right)$$

$$d^{(0)} = r^{(0)}, \alpha_0 = \frac{\left\|r^{(0)}\right\|_2^2}{d^{(0)\mathrm{T}} G d^{(0)}}$$

$$x^{(1)} = x^{(0)} + \alpha_0 d^{(0)}$$
$$r^{(1)} = b - Gx^{(1)}$$

$$d^{(1)} = r^{(1)} + \frac{\left\|r^{(1)}\right\|_2^2}{\left\|r^{(0)}\right\|_2^2} d^{(0)},$$
$$\alpha_1 = \frac{\left\|r^{(1)}\right\|_2^2}{d^{(1)\mathrm{T}} G d^{(1)}}$$

$$x^{(2)} = x^{(1)} + \alpha_1 d^{(1)}$$
$$r^{(2)} = b - Gx^{(2)}$$

$$d^{(k)} = r^{(k)} + \frac{\left\|r^{(k)}\right\|_2^2}{\left\|r^{(k-1)}\right\|_2^2} d^{(k-1)}$$
$$\alpha_k = \frac{\left\|r^{(k)}\right\|_2^2}{d^{(k)\mathrm{T}} G d^{(k)}}$$

$$x^* = x^{(n)} = x^{(n-1)} + \alpha_{n-1} d^{(n-1)}$$

Recall2: Huang-class Quasi-Newton Method

For PD quadratic optimization $\min_{x \in R^n} f(x) = \frac{1}{2} x^{\mathrm{T}} G x + b^{\mathrm{T}} x + c$

Iterative sequence: $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)} = x^{(k)} - \alpha_k H_k^{\mathrm{T}} \nabla f\left(x^{(k)}\right)$

where, $d^{(k)} = -H_k^{\mathrm{T}} \nabla f\left(x^{(k)}\right)$ **is quasi-Newton direction satisfying**

(1) $d^{(0)}, d^{(1)}, \cdots, d^{(k)}$ **are conjugate w.r.t** G

(2) **General quasi-N condition:** $H_{k+1} y^{(k)} = \rho \delta^{(k)}$, ρ **parameter**

(3) **Update formula**

$$H_{k+1} = H_k + \Delta H_k, \quad \Delta H_k = \delta^{(k)} u^{(k)\mathrm{T}} + H_k y^{(k)} v^{(k)\mathrm{T}}$$

where, $u^{(k)}, v^{(k)}$ **satisfying**

$$u^{(k)} = a_{11} \delta^{(k)} + a_{12} H_k^{\mathrm{T}} y^{(k)}, \quad v^{(k)} = a_{21} \delta^{(k)} + a_{22} H_k^{\mathrm{T}} y^{(k)},$$

$$u^{(k)\mathrm{T}} y^{(k)} = \rho, \qquad v^{(k)\mathrm{T}} y^{(k)} = -1.$$

Th.2 Let $f(x)=\frac{1}{2}x^{\mathrm{T}}Gx+b^{\mathrm{T}}x+c$ where G is SPD.

$x^{(0)}$ and H_0 given.

Huang-class quasi-Newton sequence

$x^{(k+1)}=x^{(k)}-\alpha_k H_k^{\mathrm{T}}\nabla f\left(x^{(k)}\right), k=0,1,\cdots.$ satisfying exact line search

$$f\left(x^{(k)}+\alpha_k d^{(k)}\right)=\min_{\alpha\ge 0} f\left(x^{(k)}+\alpha d^{(k)}\right).$$

Then

Sequence $x^{(0)},x^{(1)},x^{(2)},\cdots,$ does not depend on parameter ρ.

Therefore, the sequence

$x^{(0)},x^{(1)},x^{(2)},\cdots,$ generated by different parameters is identical.

And the searching directions

$$d^{(k+1)}=-\left(I-\sum_{j=0}^{k}\frac{\delta^{(i)}y^{(j)\mathrm{T}}}{y^{(j)\mathrm{T}}\delta^{(i)}}\right)H_0\nabla f\left(x^{(k+1)}\right)$$

are no difference.

Let $H_0 = I$ and $\beta_k = \dfrac{\left\| g^{(k+1)} \right\|_2^2}{\left\| g^{(k)} \right\|_2^2}$. By conjugation and exact line search,

Huang-class Quasi-Newton Method is Conjugate Gradient Method:

$$\forall x^{(0)}, d^{(0)} = -\nabla f\left(x^{(0)}\right)$$

$$\alpha_0 = \frac{\left\| d^{(0)} \right\|_2^2}{d^{(0)\mathrm{T}} G d^{(0)}}$$

$$x^{(1)} = x^{(0)} + \alpha_0 d^{(0)}$$
$$g^{(1)} = G x^{(1)} + b$$

$$d^{(1)} = -g^{(1)} + \beta_0 d^{(0)},$$
$$\alpha_1 = -\frac{g^{(1)\mathrm{T}} d^{(1)}}{d^{(1)\mathrm{T}} G d^{(1)}}, \quad \beta_0 = \frac{\left\| g^{(1)} \right\|_2^2}{\left\| g^{(0)} \right\|_2^2}$$

$$x^{(2)} = x^{(1)} + \alpha_1 d^{(1)}$$
$$g^{(2)} = G x^{(2)} + b$$

$$d^{(k)} = -g^{(k)} + \beta_{k-1} d^{(k-1)}$$
$$\alpha_k = -\frac{g^{(k)\mathrm{T}} d^{(k)}}{d^{(k)\mathrm{T}} G d^{(k)}}, \quad \beta_k = \frac{\left\| g^{(k+1)} \right\|_2^2}{\left\| g^{(k)} \right\|_2^2}$$

$$x^* = x^{(n)} = x^{(n-1)} + \alpha_{n-1} d^{(n-1)}$$

named as Fletcher-Reeves(FR) Conjugate Gradient Method.

If $$\beta_k = \frac{\left(g^{(k+1)} - g^{(k)}\right)^{\mathrm{T}} g^{(k+1)}}{\left\|g^{(k)}\right\|_2^2}$$ **Polak-Ribiere-Polyak(PRP) CGM**

If $$\beta_k = \frac{\left(g^{(k+1)} - g^{(k)}\right)^{\mathrm{T}} g^{(k+1)}}{\left(g^{(k+1)} - g^{(k)}\right)^{\mathrm{T}} d^{(k)}}$$ **--Crowder-Wolfe CGM**

If $$\beta_k = -\frac{g^{(k+1)\mathrm{T}} g^{(k+1)}}{g^{(k)\mathrm{T}} d^{(k)}}$$ **--Dixon CGM**

If exact line search $$f\left(x^{(k)} + \alpha_k d^{(k)}\right) = \min_{\alpha \geq 0} f\left(x^{(k)} + \alpha d^{(k)}\right),$$

The above CGMs are equivalent. Thus properties of CGM hold.

Advantages: Simple algorithmic structure, small computing load, no construction of searching directions and so on.

Thus they fit for large-scale optimization.

(1.2) CGM for non-quadratic fcn

Th.3 **Let** $f\left(x\right)\in C^{1}$ **be lower bounded.** **FR-CGM:**

$x^{(0)}$ **arbitrary initial point,** $d^{(0)}=-g^{(0)}=-\nabla f\left(x^{(0)}\right),$

$$f\left(x^{(0)}+\alpha_0 d^{(0)}\right)=\min_{\alpha\ge 0} f\left(x^{(0)}+\alpha d^{(0)}\right),\ x^{(k+1)}=x^{(k)}+\alpha_k d^{(k)},$$

$$\beta_k=\frac{\left\|g^{(k+1)}\right\|_2^2}{\left\|g^{(k)}\right\|_2^2},\qquad d^{(k+1)}=-g^{(k+1)}+\beta_k d^{(k)},$$

$$f\left(x^{(k)}+\alpha_k d^{(k)}\right)=\min_{\alpha\ge 0} f\left(x^{(k)}+\alpha d^{(k)}\right),\quad k=0,1,\cdots$$

If level set $L\left(x^{(0)}\right)=\left\{x \mid f\left(x\right)\le f\left(x^{(0)}\right)\right\}$ **is bounded,** **then**

sequence $\left\{x^{(k)}\right\}$ **exists at least an accumulation point,**

which is a stationary point.

Proof: (1) If there exists k_0 s.t. $g^{(k_0)} = \nabla f\left(x^{(k_0)}\right) = 0.$ True

(2) Assume that for all k, $g^{(k)} = \nabla f\left(x^{(k)}\right) \neq 0.$

Because $f\left(x^{(k-1)} + \alpha_{k-1} d^{(k-1)}\right) = \min_{\alpha \geq 0} f\left(x^{(k-1)} + \alpha d^{(k-1)}\right),$

the equality $\nabla f\left(x^{(k-1)} + \alpha_{k-1} d^{(k-1)}\right)^{\mathrm{T}} d^{(k-1)} = g^{(k)\mathrm{T}} d^{(k-1)} = 0$ holds.

From $d^{(k)} = -g^{(k)} + \beta_{k-1} d^{(k-1)}$ achieves $g^{(k)\mathrm{T}} d^{(k)} = -g^{(k)\mathrm{T}} g^{(k)} < 0$

i.e. $d^{(k)}$ is descent direction. From $L\left(x^{(0)}\right) = \left\{x \mid f(x) \leq f\left(x^{(0)}\right)\right\}$ bounded results $\left\{f\left(x^{(k)}\right)\right\}$ is monotone descending but bounded. Thus $\lim_{k\to\infty} f\left(x^{(k)}\right)$ exists and $\left\{x^{(k)}\right\} \subset L\left(x^{(0)}\right)$. Then there exists convergent subsequence $\left\{x^{(k_i)}\right\}$. Let $\lim_{k_i\to\infty} x^{(k_i)} = x^*$. Then

$f\left(x^*\right) = f\left(\lim_{k_i\to\infty} x^{(k_i)}\right) = \lim_{k_i\to\infty} f\left(x^{(k_i)}\right) = \min_{k_i} f\left(x^{(k_i)}\right).$ Therefore $g\left(x^*\right) = \nabla f\left(x^*\right) = 0.$

Th.4 Let $f(x)\in C^1$ be lower bounded. **FR-CGM:**

$x^{(0)}$ arbitrary initial point, $d^{(0)}=-g^{(0)}=-\nabla f\left(x^{(0)}\right)$,

$$x^{(k+1)}=x^{(k)}+\alpha_k d^{(k)},\quad \beta_k=\left\|g^{(k+1)}\right\|^2\left\|g^{(k)}\right\|^{-2},$$

$$d^{(k+1)}=-g^{(k+1)}+\beta_k d^{(k)},\qquad k=0,1,\cdots$$

where $0<\alpha_k$ satisfies Wolfe criterion.

Given $\mu\in\left(0,\frac{1}{2}\right)$ and $\sigma\in\left(\mu,\frac{1}{2}\right)$ satisfying

$$f\left(x^{(k)}+\alpha_k d^{(k)}\right)\le f\left(x^{(k)}\right)+\alpha_k\mu\nabla f\left(x^{(k)}\right)^{\mathrm{T}}d^{(k)}$$ and

$$\left|g^{(k+1)\mathrm{T}}d^{(k)}\right|=\left|\nabla f\left(x^{(k)}+\alpha_k d^{(k)}\right)^{\mathrm{T}}d^{(k)}\right|\le-\sigma\nabla f\left(x^{(k)}\right)^{\mathrm{T}}d^{(k)}=-\sigma g^{(k)\mathrm{T}}d^{(k)}.$$

Then $$-\sum_{j=0}^{k}\sigma^j\le\frac{g^{(k)\mathrm{T}}d^{(k)}}{\left\|g^{(k)}\right\|^2}\le-2+\sum_{j=0}^{k}\sigma^j<0.$$

i.e. $d^{(k)}$ is descent direction.

Proof: **By Induction.**

(1) If $k=0,\quad d^{(0)}=-g^{(0)}=-\nabla f\left(x^{(0)}\right),\quad \sigma^0=1.$ **Then**

$$-\sum_{j=0}^{0}\sigma^j=\frac{g^{(0)\mathrm{T}}d^{(0)}}{\left\|g^{(0)}\right\|^2}=-1=-2+\sum_{j=0}^{0}\sigma^j.$$ **True**

(2) Assume that for all $k>0$, **if** $g^{(k)}=\nabla f\left(x^{(k)}\right)\neq 0.$ **True**

i.e.
$$-\sum_{j=0}^{k}\sigma^j\le\frac{g^{(k)\mathrm{T}}d^{(k)}}{\left\|g^{(k)}\right\|^2}\le-2+\sum_{j=0}^{k}\sigma^j<0.$$

Then for $k+1,\ d^{(k+1)}=-g^{(k+1)}+\beta_k d^{(k)}=-g^{(k+1)}+\frac{\left\|g^{(k+1)}\right\|^2}{\left\|g^{(k)}\right\|^2}d^{(k)},$

$$\frac{g^{(k+1)\mathrm{T}}d^{(k+1)}}{\left\|g^{(k+1)}\right\|^2}=-1+\frac{g^{(k+1)\mathrm{T}}d^{(k)}}{\left\|g^{(k)}\right\|^2}\ge-1+\frac{\sigma g^{(k)\mathrm{T}}d^{(k)}}{\left\|g^{(k)}\right\|^2}\ge-1-\sigma\sum_{j=0}^{k}\sigma^j=-\sum_{j=0}^{k+1}\sigma^j$$

and $-1+\frac{g^{(k+1)\mathrm{T}}d^{(k)}}{\left\|g^{(k)}\right\|^2}\le-1-\frac{\sigma g^{(k)\mathrm{T}}d^{(k)}}{\left\|g^{(k)}\right\|^2}\le-1+\sigma\sum_{j=0}^{k}\sigma^j\le-2+\sum_{j=0}^{k+1}\sigma^j.$ **True**

Th.5 Let $f(x)\in C^2$. Level set $L\left(x^{(0)}\right)$ is bounded.

FR-CGM: $x^{(0)}$ arbitrary initial point. $d^{(0)}=-g^{(0)}=-\nabla f\left(x^{(0)}\right)$,

$$x^{(k+1)}=x^{(k)}+\alpha_k d^{(k)},\quad \beta_k=\left\|g^{(k+1)}\right\|^2\left\|g^{(k)}\right\|^{-2},$$

$$d^{(k+1)}=-g^{(k+1)}+\beta_k d^{(k)},\ k=0,1,\cdots$$

where $0<\alpha_k$ satisfies Wolfe criterion.

Given $\mu\in\left(0,\frac{1}{2}\right)$ and $\sigma\in\left(\mu,\frac{1}{2}\right)$ satisfying

$$f\left(x^{(k)}+\alpha_k d^{(k)}\right)\le f\left(x^{(k)}\right)+\alpha_k\mu\nabla f\left(x^{(k)}\right)^{\mathrm{T}}d^{(k)}$$

and $$\left|g^{(k+1)\mathrm{T}}d^{(k)}\right|\le-\sigma g^{(k)\mathrm{T}}d^{(k)}.$$

Then $$\liminf_{k\to\infty}\left\|g^{(k)}\right\|=0.$$

No convexity requirement for objective fcn.

Proof: **First, prove** $\left\|d^{(k)}\right\|^2 \le \frac{1-\sigma}{1+\sigma}\left\|g^{(k)}\right\|^4 \sum_{j=0}^{k}\left\|g^{(j)}\right\|^{-2}.$

From Wolfe criterion and Th.4 gives rise to

$$\left|g^{(k)\mathrm{T}}d^{(k-1)}\right| \le -\sigma g^{(k-1)\mathrm{T}}d^{(k-1)} \le \left\|g^{(k-1)}\right\|^2 \sigma\sum_{j=0}^{k-1}\sigma^j \le \left\|g^{(k-1)}\right\|^2 \frac{\sigma}{1-\sigma}.$$

$$\beta_{k-1} = \frac{\left\|g^{(k)}\right\|^2}{\left\|g^{(k-1)}\right\|^2}$$

Then $\left\|d^{(k)}\right\|^2 = \left(-g^{(k)}+\beta_{k-1}d^{(k-1)}\right)^{\mathrm{T}}\left(-g^{(k)}+\beta_{k-1}d^{(k-1)}\right)$

$$= \left\|g^{(k)}\right\|^2 - 2\beta_{k-1}g^{(k)\mathrm{T}}d^{(k-1)} + \beta_{k-1}^2\left\|d^{(k-1)}\right\|^2$$

$$\le \left\|g^{(k)}\right\|^2 + \frac{2\sigma}{1-\sigma}\left\|g^{(k)}\right\|^2 + \beta_{k-1}^2\left\|d^{(k-1)}\right\|^2 = \frac{1+\sigma}{1-\sigma}\left\|g^{(k)}\right\|^2 + \beta_{k-1}^2\left\|d^{(k-1)}\right\|^2$$

$$\le \frac{1+\sigma}{1-\sigma}\left\|g^{(k)}\right\|^2 + \beta_{k-1}^2\left(\frac{1+\sigma}{1-\sigma}\left\|g^{(k-1)}\right\|^2 + \beta_{k-2}^2\left\|d^{(k-2)}\right\|^2\right)$$

$$\le \frac{1+\sigma}{1-\sigma}\left\|g^{(k)}\right\|^2 + \left(\frac{1+\sigma}{1-\sigma}\frac{\left\|g^{(k)}\right\|^4}{\left\|g^{(k-1)}\right\|^2} + \frac{1+\sigma}{1-\sigma}\frac{\left\|g^{(k)}\right\|^4}{\left\|g^{(k-2)}\right\|^2} + \cdots + \frac{1+\sigma}{1-\sigma}\frac{\left\|g^{(k)}\right\|^4}{\left\|g^{(0)}\right\|^2}\right)$$

$$= \frac{1-\sigma}{1+\sigma}\left\|g^{(k)}\right\|^4 \sum_{j=0}^{k}\left\|g^{(j)}\right\|^{-2}.$$

Second, prove $\liminf_{k\to\infty} \left\| g^{(k)} \right\| = 0.$ $\quad 0 < \sigma < \frac{1}{2}$

Contrarily if $\liminf_{k\to\infty} \left\| g^{(k)} \right\| \neq 0$ **i.e.** $\exists \varepsilon > 0,$ **s.t.** $\left\| g^{(k)} \right\| \geq \varepsilon > 0.$

Due to boundedness of level set $L\left(x^{(0)}\right)$, **there exist** $c_1 > 0$ **and**

$M > 0$ **s.t.** $\left\| d^{(k)} \right\|^2 \leq c_1 (k+1),\ k = 0,1,\cdots,$ **and** $\left\| \nabla^2 f(x) \right\| \leq M.$

Thus
$$\cos\theta_k = \frac{-g^{(k)\mathrm{T}} d^{(k)}}{\left\| g^{(k)} \right\| \left\| d^{(k)} \right\|} \geq \left(2 - \sum_{j=0}^{k} \sigma^j \right) \frac{\left\| g^{(k)} \right\|}{\left\| d^{(k)} \right\|} \geq \frac{1-2\sigma}{1-\sigma} \frac{\left\| g^{(k)} \right\|}{\left\| d^{(k)} \right\|}.$$

Then
$$\sum_{k=0}^{\infty} (\cos\theta_k)^2 \geq \sum_{k=0}^{\infty} \left(\frac{1-2\sigma}{1-\sigma} \right)^2 \frac{\varepsilon^2}{c_1 (k+1)} \geq \left(\frac{1-2\sigma}{1-\sigma} \right)^2 \frac{\varepsilon^2}{c_1} \sum_{k=0}^{\infty} \frac{1}{(k+1)}$$
diverges.

On the other hand
$$\nabla f\left(x^{(k)} + \alpha_k d^{(k)} \right) = \nabla f\left(x^{(k)} \right) + \alpha_k \nabla^2 f\left(\xi^{(k)} \right) d^{(k)}.$$

Therefore
$$\boxed{\sigma g^{(k)\mathrm{T}} d^{(k)} \leq} \ g^{(k+1)\mathrm{T}} d^{(k)} \leq g^{(k)\mathrm{T}} d^{(k)} + \alpha_k M \left\| d^{(k)} \right\|^2.$$

Then $\alpha_k \geq -\dfrac{1-\sigma}{M}\dfrac{g^{(k)\mathrm{T}}d^{(k)}}{\left\|d^{(k)}\right\|^2}$. **Substituted into Wolfe criterion delivers**

$$f\left(x^{(k)}+\alpha_k d^{(k)}\right) \leq f\left(x^{(k)}\right)+\alpha_k \mu \nabla f\left(x^{(k)}\right)^{\mathrm{T}} d^{(k)},$$

$$\leq f\left(x^{(k)}\right)-\frac{\mu(1-\sigma)}{M}\left(\frac{g^{(k)\mathrm{T}}d^{(k)}}{\left\|d^{(k)}\right\|}\right)^2$$

$$= f\left(x^{(k)}\right)-\frac{\mu(1-\sigma)}{M}\left\|g^{(k)}\right\|^2\left(\cos\theta_k\right)^2.$$

Therefore

$$\left(\cos\theta_k\right)^2 \leq \frac{f\left(x^{(k)}\right)-f\left(x^{(k+1)}\right)}{\mu(1-\sigma)}\frac{M}{\left\|g^{(k)}\right\|^2} \leq \frac{M}{\varepsilon^2\mu(1-\sigma)}\left(f\left(x^{(k)}\right)-f\left(x^{(k+1)}\right)\right)$$

i.e.

$$\sum_{k=0}^{\infty}\left(\cos\theta_k\right)^2 \leq C\sum_{k=0}^{\infty}\left(f\left(x^{(k)}\right)-f\left(x^{(k+1)}\right)\right) \leq C\left(f\left(x^{(0)}\right)-f\left(x^{(\infty)}\right)\right) \leq C_0.$$

$x^{(0)}$ **arbitrary initial point.** $d^{(0)} = -g^{(0)} = -\nabla f\left(x^{(0)}\right),$

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)},\ d^{(k+1)} = -g^{(k+1)} + \beta_k d^{(k)},\ k = 0,1,\cdots$$

PRP-CGM

$$\beta_k = \frac{\left(g^{(k+1)} - g^{(k)}\right)^{\mathrm{T}} g^{(k+1)}}{\left\|g^{(k)}\right\|^2}$$

FR-CGM

$$\beta_k = \frac{\left\|g^{(k+1)}\right\|^2}{\left\|g^{(k)}\right\|^2}$$

If $f\left(x^{(k)} + \alpha_k d^{(k)}\right) = \min_{\alpha \geq 0} f\left(x^{(k)} + \alpha d^{(k)}\right),$

then $\nabla f\left(x^{(k)} + \alpha_k d^{(k)}\right)^{\mathrm{T}} d^{(k)} = \left[\nabla f\left(x^{(k)}\right) + \nabla^2 f\left(\xi^{(k)}\right)\alpha_k d^{(k)}\right]^{\mathrm{T}} d^{(k)} = 0.$

Powell discovered: **For some** k, **if** $d^{(k)}$ **is almost orthogonal to** $g^{(k)}$

Then $\alpha_k \approx 0.$ **Thus** $x^{(k+1)} \approx x^{(k)},\quad g^{(k+1)} \approx g^{(k)}.$

PRP-CGM: $\beta_k \approx 0,\quad d^{(k+1)} \approx -g^{(k+1)}$ **Restart!**

Besides， Powell raised an example conveying that

for PRP-CGM applying to general nonconvex fcn, the result

$\liminf_{k\to\infty} \left\| g^{(k)} \right\| = 0$ **of above Theorems possibly does not hold.**

If the objective fcn is consistently convex, PRP-CGM is globally convergent.

Th.6 **Let** $f(x) \in C^2$. **Level set** $L\left(x^{(0)}\right)$ **is bounded.**

PRP-CGM: $x^{(0)}$ **arbitrarily given.** $d^{(0)} = -g^{(0)} = -\nabla f\left(x^{(0)}\right)$,

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}, \qquad d^{(k+1)} = -g^{(k+1)} + \beta_k d^{(k)},$$

where $f\left(x^{(k)} + \alpha_k d^{(k)}\right) = \min_{\alpha \ge 0} f\left(x^{(k)} + \alpha d^{(k)}\right)$, $\beta_k = \dfrac{\left(g^{(k+1)} - g^{(k)}\right)^{\mathrm{T}} g^{(k+1)}}{\left\| g^{(k)} \right\|^2}$.

If $\exists m > 0$, **s.t.** $m\|u\|^2 \le u^{\mathrm{T}} \nabla^2 f(x) u, \forall x \in L\left(x^{(0)}\right), \forall u \in R^n$.

Then $\lim_{k\to\infty} x^{(k)} = x^*$, $f\left(x^*\right) = \min_{x \in R^n} f(x)$.

Recall3:

Th3.1 **Let** $\nabla f(x)$ **be consistently continuous on level set**

$$L\left(x^{(0)}\right)=\left\{x \mid f(x)\le f\left(x^{(0)}\right)\right\}.\quad 0\le\theta_k\le\frac{\pi}{2}-\bar{\mu},(\bar{\mu}>0),k=1,2,\cdots$$

$$f\left(x^{(k)}+\alpha_k d^{(k)}\right)=\min_{\alpha\ge 0} f\left(x^{(k)}+\alpha d^{(k)}\right).$$

Iterative sequence $x^{(k+1)}=x^{(k)}+\alpha_k d^{(k)},k=0,1,\cdots.$

Then **(1) There exists** k **s.t.** $\nabla f\left(x^{(k)}\right)=0.$

or **(2)** $f\left(x^{(k)}\right)\to-\infty(k\to\infty).$

or **(3)** $\nabla f\left(x^{(k)}\right)\to 0(k\to\infty).$

If $f(x)$ **is continuous, then level set** $L\left(x^{(0)}\right)$ **is closed. From consistent continuity of continuous fcn on bounded closed set, proof of Th6 is equivalent to the following**

$0\le\theta_k\le\frac{\pi}{2}-\bar{\mu},(\bar{\mu}>0),k=1,2,\cdots$ **or** $\cos\theta_k\ge\rho>0,k=1,2,\cdots$

Proof: **Because of** $f\left(x^{(k-1)}+\alpha_{k-1}d^{(k-1)}\right)=\min_{\alpha\geq 0} f\left(x^{(k-1)}+\alpha d^{(k-1)}\right)$,

then $\nabla f\left(x^{(k-1)}+\alpha_{k-1}d^{(k-1)}\right)^{\mathrm{T}} d^{(k-1)} = g^{(k)\mathrm{T}} d^{(k-1)} = 0.$

In addition $d^{(k)}=-g^{(k)}+\beta_{k-1}d^{(k-1)}$. **Thus**

$$\cos\theta_k = \frac{-g^{(k)\mathrm{T}}d^{(k)}}{\left\|g^{(k)}\right\|\left\|d^{(k)}\right\|} = \frac{-g^{(k)\mathrm{T}}\left(-g^{(k)}+\beta_{k-1}d^{(k-1)}\right)}{\left\|g^{(k)}\right\|\left\|d^{(k)}\right\|} = \frac{\left\|g^{(k)}\right\|^2}{\left\|g^{(k)}\right\|\left\|d^{(k)}\right\|}$$

i.e.

$$\cos^2\theta_k = \frac{\left\|g^{(k)}\right\|^2}{\left\|d^{(k)}\right\|^2} = \frac{\left\|g^{(k)}\right\|^2}{\left\|-g^{(k)}+\beta_{k-1}d^{(k-1)}\right\|^2} = \frac{\left\|g^{(k)}\right\|^2}{\left\|g^{(k)}\right\|^2+\left\|\beta_{k-1}d^{(k-1)}\right\|^2}.$$

Evaluation of β_{k-1} **as following.**

From $\nabla f\left(x^{(k-1)}+\alpha_{k-1}d^{(k-1)}\right)=\nabla f\left(x^{(k-1)}\right)+\alpha_{k-1}\nabla^2 f\left(\xi^{(k-1)}\right)d^{(k-1)}$

and $f\left(x^{(k-1)}+\alpha_{k-1}d^{(k-1)}\right)=\min\limits_{\alpha\geq 0} f\left(x^{(k-1)}+\alpha d^{(k-1)}\right)$ **reduces**

$$\nabla f\left(x^{(k-1)}+\alpha_{k-1}d^{(k-1)}\right)^{\mathrm{T}} d^{(k-1)}=\left(\nabla f\left(x^{(k-1)}\right)+\alpha_{k-1}\nabla^2 f\left(\xi^{(k-1)}\right)d^{(k-1)}\right)^{\mathrm{T}} d^{(k-1)}=0.$$

Besides $d^{(k-1)}=-g^{(k-1)}+\beta_{k-2}d^{(k-2)}$. **Let** $\bar{G}_{k-1}=\nabla^2 f\left(\xi^{(k-1)}\right)^{\mathrm{T}}$. **Then**

$$\alpha_{k-1}=\frac{-g^{(k-1)\mathrm{T}}d^{(k-1)}}{d^{(k-1)\mathrm{T}}\bar{G}_{k-1}d^{(k-1)}}=\frac{-g^{(k-1)\mathrm{T}}\left(-g^{(k-1)}+\beta_{k-2}d^{(k-2)}\right)}{d^{(k-1)\mathrm{T}}\bar{G}_{k-1}d^{(k-1)}}=\frac{\left\|g^{(k-1)}\right\|^2}{d^{(k-1)\mathrm{T}}\bar{G}_{k-1}d^{(k-1)}}$$

and

$$\beta_{k-1}=\frac{\left(g^{(k)}-g^{(k-1)}\right)^{\mathrm{T}}g^{(k)}}{\left\|g^{(k-1)}\right\|^2}=\frac{\alpha_{k-1}d^{(k-1)\mathrm{T}}\bar{G}_{k-1}g^{(k)}}{\left\|g^{(k-1)}\right\|^2}=\frac{d^{(k-1)\mathrm{T}}\bar{G}_{k-1}g^{(k)}}{d^{(k-1)\mathrm{T}}\bar{G}_{k-1}d^{(k-1)}}.$$

Let $M=\max\limits_{x\in L\left(x^{(0)}\right)}\left\|G^{\mathrm{T}}(x)\right\|.$

Then $$\left|\beta_{k-1}\right|=\frac{\left\|d^{(k-1)\mathrm{T}}\bar{G}_{k-1}g^{(k)}\right\|}{\left\|d^{(k-1)\mathrm{T}}\bar{G}_{k-1}d^{(k-1)}\right\|}\leq\frac{M\left\|d^{(k-1)}\right\|\left\|g^{(k)}\right\|}{m\left\|d^{(k-1)}\right\|^2}=\frac{M\left\|g^{(k)}\right\|}{m\left\|d^{(k-1)}\right\|}$$

Substituting $\left|\beta_{k-1}\right| \le \dfrac{M\left\|g^{(k)}\right\|}{m\left\|d^{(k-1)}\right\|}$ **into** $\cos^2\theta_k = \dfrac{\left\|g^{(k)}\right\|^2}{\left\|g^{(k)}\right\|^2 + \left\|\beta_{k-1} d^{(k-1)}\right\|^2}$

makes $\cos^2\theta_k \ge \dfrac{\left\|g^{(k)}\right\|^2}{\left\|g^{(k)}\right\|^2 + \dfrac{M^2\left\|g^{(k)}\right\|^2}{m^2\left\|d^{(k-1)}\right\|^2}\left\|d^{(k-1)}\right\|^2} = \dfrac{m^2}{m^2+M^2} = \rho^2 > 0$

From Th3.1 the proof of Th.6 is complete.

Next is convergence rate comparison of PRP-CGM with that of SDM for positive definite quadratic objective function.

Given PDQ fcn $f(x)=\frac{1}{2}x^{\mathrm{T}}Gx.$ **Starting from** $x^{(k)}$

$$g^{(k)}=Gx^{(k)}$$

Steepest descent iterative sequence:

$$x_{\mathrm{SD}}^{(k+1)}=x^{(k)}+\alpha_k^{\mathrm{SD}}d_{\mathrm{SD}}^{(k)},\qquad d_{\mathrm{SD}}^{(k)}=-g^{(k)},$$

$$x^{(k)}=G^{-1}g^{(k)}$$

$$f\left(x^{(k)}+\alpha_k^{\mathrm{SD}}d_{\mathrm{SD}}^{(k)}\right)=\min_{\alpha\geq 0} f\left(x^{(k)}+\alpha d_{\mathrm{SD}}^{(k)}\right).$$

i.e.

$$\alpha_k^{\mathrm{SD}}=-\frac{g^{(k)\mathrm{T}}d_{\mathrm{SD}}^{(k)}}{d_{\mathrm{SD}}^{(k)\mathrm{T}}Gd_{\mathrm{SD}}^{(k)}}=\frac{\left\|g^{(k)}\right\|^2}{g^{(k)\mathrm{T}}Gg^{(k)}}.$$

Then

$$f\left(x^{(k)}+\alpha_k^{\mathrm{SD}}d_{\mathrm{SD}}^{(k)}\right)=\frac{1}{2}\left(x^{(k)}-\frac{\left\|g^{(k)}\right\|^2}{g^{(k)\mathrm{T}}Gg^{(k)}}g^{(k)}\right)^{\mathrm{T}}G\left(x^{(k)}-\frac{\left\|g^{(k)}\right\|^2}{g^{(k)\mathrm{T}}Gg^{(k)}}g^{(k)}\right)$$

$$=\frac{1}{2}\left[x^{(k)\mathrm{T}}Gx^{(k)}-2\frac{\left\|g^{(k)}\right\|^2}{g^{(k)\mathrm{T}}Gg^{(k)}}x^{(k)\mathrm{T}}Gg^{(k)}+\frac{\left\|g^{(k)}\right\|^4}{g^{(k)\mathrm{T}}Gg^{(k)}}\right]=\frac{1}{2}g^{(k)\mathrm{T}}G^{-1}g^{(k)}-\frac{1}{2}\frac{\left\|g^{(k)}\right\|^4}{g^{(k)\mathrm{T}}Gg^{(k)}}$$

PRP-CGM iterative sequence:

$$\beta_{k-1} = \frac{\left(g^{(k)} - g^{(k-1)}\right)^{\mathrm{T}} g^{(k)}}{\left\|g^{(k-1)}\right\|^2}$$

$$x_{\mathrm{PRP}}^{(k+1)} = x^{(k)} + \alpha_k^{\mathrm{PRP}} d_{\mathrm{PRP}}^{(k)},$$

$$d_{\mathrm{PRP}}^{(k)} = -g^{(k)} + \beta_{k-1} d_{\mathrm{PRP}}^{(k-1)}, \quad f\left(x^{(k)} + \alpha_k^{\mathrm{PRP}} d_{\mathrm{PRP}}^{(k)}\right) = \min_{\alpha \ge 0} f\left(x^{(k)} + \alpha d_{\mathrm{PRP}}^{(k)}\right).$$

Then $$\alpha_k^{\mathrm{PRP}} = -\frac{g^{(k)\mathrm{T}} d_{\mathrm{PRP}}^{(k)}}{d_{\mathrm{PRP}}^{(k)\mathrm{T}} G d_{\mathrm{PRP}}^{(k)}} = \frac{-g^{(k)\mathrm{T}}\left(-g^{(k)} + \beta_{k-1} d_{\mathrm{PRP}}^{(k-1)}\right)}{d_{\mathrm{PRP}}^{(k)\mathrm{T}} G d_{\mathrm{PRP}}^{(k)}} = \frac{\left\|g^{(k)}\right\|^2}{d_{\mathrm{PRP}}^{(k)\mathrm{T}} G d_{\mathrm{PRP}}^{(k)}}$$

Thus $$f\left(x^{(k)} + \alpha_k^{\mathrm{PRP}} d_{\mathrm{PRP}}^{(k)}\right) = \frac{1}{2} g^{(k)\mathrm{T}} G^{-1} g^{(k)} - \frac{1}{2} \frac{\left\|g^{(k)}\right\|^4}{d_{\mathrm{PRP}}^{(k)\mathrm{T}} G d_{\mathrm{PRP}}^{(k)}}$$

Further $$d_{\mathrm{PRP}}^{(k)} = -g^{(k)} + \beta_{k-1} d_{\mathrm{PRP}}^{(k-1)},$$ **i.e.** $$g^{(k)} = -d_{\mathrm{PRP}}^{(k)} + \beta_{k-1} d_{\mathrm{PRP}}^{(k-1)},$$

Therefore $$g^{(k)\mathrm{T}} G g^{(k)} = d_{\mathrm{PRP}}^{(k)\mathrm{T}} G d_{\mathrm{PRP}}^{(k)} + \beta_{k-1}^2 d_{\mathrm{PRP}}^{(k-1)\mathrm{T}} G d_{\mathrm{PRP}}^{(k-1)} \ge d_{\mathrm{PRP}}^{(k)\mathrm{T}} G d_{\mathrm{PRP}}^{(k)}$$

Then $$f\left(x^{(k)} + \alpha_k^{\mathrm{PRP}} d_{\mathrm{PRP}}^{(k)}\right) \le f\left(x^{(k)} + \alpha_k^{\mathrm{SD}} d_{\mathrm{SD}}^{(k)}\right)$$

This means that PRP-CGM converges faster than SDM.

(1.3) Restarted CGM

Advantages of PRP-CGM:

A1. At least linearly convergent rate.

A2. The objective fcn decreasing faster than SDM.

A3. Quadratic termination.

Improvement:

As for a non-quadratic convex fcn, it approximates a positive definite quadratic fcn near the minimizer, it is efficient to restart along negative gradient.

The improved method is called Restarted CGM.

Then the search directions of proceeding n iterations are approaching to conjugate directions. Thus, the n-th iterative point approaches the minimizer and thus converges faster.

May multi-n-iteration restart the searching until reach the minimizer.

Th.7 Let $f(x) \in C^2$.

n-iteration-restarted PRP-CGM sequence: $\left\{x^{(k)}\right\}$.

If $\lim\limits_{k\to\infty} x^{(k)} = x^*$, $\nabla f\left(x^*\right) = 0$ and $\nabla^2 f\left(x^*\right)$ is positive definite.

Then $\lim\limits_{i\to\infty} \dfrac{\left\|x^{((i+1)n+k_0)} - x^*\right\|}{\left\|x^{(in+k_0)} - x^*\right\|} = 0.$ n-iteration superlinear convergent rate.

Th.8 Let $\nabla^2 f(x)$ be Lipchitzs continuous.

n-iteration-restarted PRP-CGM sequence: $\left\{x^{(k)}\right\}$.

If $\lim\limits_{k\to\infty} x^{(k)} = x^*$, $\nabla f\left(x^*\right) = 0$ and $\nabla^2 f\left(x^*\right)$ is positive definite.

Then $\exists M > 0$ s.t. $\left\|x^{((i+1)n+k_0)} - x^*\right\| \le M \left\|x^{(in+k_0)} - x^*\right\|^2$, $i = 1,2,\cdots$

--- n-iteration-quadratic convergent rate.

It possible r-iteration-restarted PR-CGM, and so on.

THANK YOU FOR ATTENDING