Lemma 1.7.3 For any $k, \ell \geq 0,\ k \neq \ell$, we have $\langle f'(\boldsymbol{x}_k), f'(\boldsymbol{x}_\ell)\rangle = 0$.

Proof: Let $k \geq i$, and consider

$$\phi(\boldsymbol{\lambda}) = f\left(\boldsymbol{x}_0 + \sum_{j=1}^{k} \lambda_j f'(\boldsymbol{x}_{j-1})\right).$$

From the previous lemma, there is a $\boldsymbol{\lambda}^*$ such that $\boldsymbol{x}_k = \boldsymbol{x}_0 + \sum_{j=1}^{k} \lambda_j^* f'(\boldsymbol{x}_{j-1})$. Moreover, $\boldsymbol{\lambda}^*$ is the minimum of the function $\phi(\boldsymbol{\lambda})$. Therefore,

$$\frac{\partial \phi}{\partial \lambda_i}(\boldsymbol{\lambda}^*) = \langle f'(\boldsymbol{x}_k), f'(\boldsymbol{x}_{i-1})\rangle = 0.$$

∎

Corollary 1.7.4 The sequence generated by the conjugate gradient method for the convex quadratic function is finite.

Proof: Since the number of orthogonal directions in $\mathbb{R}^n$ cannot exceed n. ∎

Let us define $\boldsymbol{\delta}_i = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i$. It is clear that $\mathcal{L}_k = \mathrm{Lin}\{\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{k-1}\}$.

Lemma 1.7.5 For any $k, \ell \geq 0$, $k \neq \ell$, $\langle \boldsymbol{A}\boldsymbol{\delta}_k, \boldsymbol{\delta}_\ell\rangle = 0$.

Proof: Let $k > \ell$. Then

$$\langle \boldsymbol{A}\boldsymbol{\delta}_k, \boldsymbol{\delta}_\ell\rangle = \langle \boldsymbol{A}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k), \boldsymbol{\delta}_\ell\rangle = \langle f'(\boldsymbol{x}_{k+1}) - f'(\boldsymbol{x}_k), \boldsymbol{x}_{\ell+1} - \boldsymbol{x}_\ell\rangle = 0,$$

due to Lemma 1.7.3. ∎

The vectors $\{\boldsymbol{\delta}_i\}$ are called *conjugate* with respect to matrix $\boldsymbol{A}$.

Now, let us be more precise with the conjugate gradient method. We will define the next iterations as follows:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k f'(\boldsymbol{x}_k) + \sum_{j=0}^{k-1} \lambda_j \boldsymbol{\delta}_j$$

Using the previous properties, we arrive that

$$\lambda_j = 0, \quad (j = 0, 1, \ldots, k-2), \quad \lambda_{k-1} = \frac{h_k \|f'(\boldsymbol{x}_k)\|^2}{\langle f'(\boldsymbol{x}_k) - f'(\boldsymbol{x}_{k-1}), \boldsymbol{\delta}_{k-1}\rangle}. \tag{1.6}$$

Thus

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \boldsymbol{p}_k$$

where

$$\boldsymbol{p}_k = f'(\boldsymbol{x}_k) - \frac{\|f'(\boldsymbol{x}_k)\|^2 \boldsymbol{p}_{k-1}}{\langle f'(\boldsymbol{x}_k) - f'(\boldsymbol{x}_{k-1}), \boldsymbol{p}_{k-1}\rangle}.$$

Finally, we can present the Conjugate Gradient Method

Conjugate Gradient Method	
Step 0:	Let $\boldsymbol{x}_0 \in \mathbb{R}^n$, compute $f(\boldsymbol{x}_0), f'(\boldsymbol{x}_0)$ and set $\boldsymbol{p}_0 := f'(\boldsymbol{x}_0)$, $k := 0$
Step 1:	Find $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k - h_k \boldsymbol{p}_k$ by "approximate line search" on the scalar h_k
Step 2:	Compute $f(\boldsymbol{x}_{k+1})$ and $f'(\boldsymbol{x}_{k+1})$
Step 3:	Compute the coefficient β_{k+1}
Step 4:	Set $p_{k+1} := f'(\boldsymbol{x}_{k+1}) - \beta_{k+1}\boldsymbol{p}_k$, $k := k+1$ and go to Step 1

The most popular choices for the coefficient β_k are:

1. *Hestenes-Stiefel (1952):* $\beta_{k+1} = \frac{\langle f'(\boldsymbol{x}_{k+1}), f'(\boldsymbol{x}_{k+1}) - f'(\boldsymbol{x}_k)\rangle}{\langle f'(\boldsymbol{x}_{k+1}) - f'(\boldsymbol{x}_k), \boldsymbol{p}_k\rangle}$.

2. *Fletcher-Reeves (1964):* $\beta_{k+1} = \frac{\|f'(\boldsymbol{x}_{k+1})\|^2}{\|f'(\boldsymbol{x}_k)\|^2}$.

3. *Polak-Ribière:* $\beta_{k+1} = \frac{\langle f'(\boldsymbol{x}_{k+1}), f'(\boldsymbol{x}_{k+1}) - f'(\boldsymbol{x}_k)\rangle}{\|f'(\boldsymbol{x}_k)\|^2}$.

4. *Polak-Ribière plus:* $\beta_{k+1} = \max\left\{0, \frac{\langle f'(\boldsymbol{x}_{k+1}), f'(\boldsymbol{x}_{k+1}) - f'(\boldsymbol{x}_k)\rangle}{\|f'(\boldsymbol{x}_k)\|^2}\right\}$.

5. *Dai-Yuan (1999):* $\beta_{k+1} = \frac{\|f'(\boldsymbol{x}_{k+1})\|^2}{\langle f'(\boldsymbol{x}_{k+1}) - f'(\boldsymbol{x}_k), \boldsymbol{p}_k\rangle}$.

1.8 Quasi-Newton methods

The basic idea of the quasi-Newton methods is to approximate the Hessian matrix (or its inverse) which we need to compute in the Newton method. There are of course infinitely many ways to do so, but we choose the ones which satisfy the *secant equation*:

$$\boldsymbol{H}_{k+1}\boldsymbol{y}_k = \boldsymbol{s}_k$$

where $\boldsymbol{y}_k = f'(\boldsymbol{x}_{k+1}) - f'(\boldsymbol{x}_k)$, $\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$.

The general scheme of the quasi-Newton method is as follows.

Quasi-Newton Method	
Step 0:	Let $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\boldsymbol{H}_0 := \boldsymbol{I}$, $k := 0$. Compute $f(\boldsymbol{x}_0), f'(\boldsymbol{x}_0)$
Step 1:	Set $\boldsymbol{p}_k := \boldsymbol{H}_k f'(\boldsymbol{x}_k)$
Step 2:	Find $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k - h_k \boldsymbol{p}_k$ by "approximate line search" on the scalar h_k
Step 3:	Compute $f(\boldsymbol{x}_{k+1})$ and $f'(\boldsymbol{x}_{k+1})$
Step 4:	Compute $\boldsymbol{H}_{k+1}$ from $\boldsymbol{H}_k$, $k := k+1$ and go to Step 1

The most popular updates for $\boldsymbol{H}_{k+1}$ are:

1. *BFGS (Broyden-Fletcher-Goldfarb-Shanno)*

$$\boldsymbol{H}_{k+1} := \left(\boldsymbol{I} - \frac{\boldsymbol{s}_k(\boldsymbol{y}_k)^T}{(\boldsymbol{s}_k)^T\boldsymbol{y}_k}\right)\boldsymbol{H}_k\left(\boldsymbol{I} - \frac{\boldsymbol{y}_k(\boldsymbol{s}_k)^T}{(\boldsymbol{s}_k)^T\boldsymbol{y}_k}\right) + \frac{\boldsymbol{s}_k(\boldsymbol{s}_k)^T}{(\boldsymbol{s}_k)^T\boldsymbol{y}_k}$$

2. *DFP (Davidon-Fletcher-Powell)*

$$\boldsymbol{H}_{k+1} := \boldsymbol{H}_k + \frac{\boldsymbol{s}_k(\boldsymbol{s}_k)^T}{(\boldsymbol{y}_k)^T\boldsymbol{s}_k} - \frac{\boldsymbol{H}_k\boldsymbol{y}_k(\boldsymbol{y}_k)^T\boldsymbol{H}_k}{(\boldsymbol{y}_k)^T\boldsymbol{H}_k\boldsymbol{y}_k}$$

3. *Symmetric-Rank-One*

$$\boldsymbol{H}_{k+1} := \boldsymbol{H}_k + \frac{(\boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k)(\boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k)^T}{(\boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k)^T\boldsymbol{y}_k}$$

In the same way for the conjugate gradient method, we can show that the quasi-Newton method converges in finite number of iterations for a strictly convex quadratic function. Moreover, under some strict convexity conditions at the neighborhood of the local minimum, it is possible to show that its iterates converge super-linearly [NOCEDAL2006].

1.9 Exercises

1. In view of Theorem 1.3.4, find a twice continuously differentiable function on $\mathbb{R}^n$ which satisfies $f'(\boldsymbol{x}^*) = 0, \quad f''(\boldsymbol{x}^*) \succeq \boldsymbol{O}$, but $\boldsymbol{x}^*$ is not a local minimum of $f(\boldsymbol{x})$.

2. Prove Lemma 1.4.5.

3. Give a geometric interpretation of the following step-size strategies:

   Let $0 < c_1 < c_2 < 1$,

   - Wolfe condition

$$f(\boldsymbol{x}_k - hf'(\boldsymbol{x}_k)) \leq f(\boldsymbol{x}_k) - c_1h\|f'(\boldsymbol{x}_k)\|^2,$$
$$\langle f'(\boldsymbol{x}_k - hf'(\boldsymbol{x}_k)), f'(\boldsymbol{x}_k)\rangle \leq c_2\|f'(\boldsymbol{x}_k)\|^2.$$

   - Strong Wolfe condition

$$f(\boldsymbol{x}_k - hf'(\boldsymbol{x}_k)) \leq f(\boldsymbol{x}_k) - c_1h\|f'(\boldsymbol{x}_k)\|^2,$$
$$|\langle f'(\boldsymbol{x}_k - hf'(\boldsymbol{x}_k)), f'(\boldsymbol{x}_k)\rangle| \leq c_2\|f'(\boldsymbol{x}_k)\|^2.$$

4. Consider a sequence $\{\beta_k\}_{k=0}^{\infty}$ which converges to zero.

   The sequence is said to converge *Q-linearly* if there exists a scalar $\rho \in (0, 1)$ such that

$$\frac{\beta_{k+1}}{\beta_k} \leq \rho,$$

   for all k sufficiently large. *Q-superlinear* convergence occurs when we have

$$\lim_{k\to\infty} \frac{\beta_{k+1}}{\beta_k} = 0,$$

   while the convergence is *Q-quadratic* if there is a constant C such that

$$\frac{\beta_{k+1}}{\beta_k^2} \leq C$$

for all k sufficiently large. *Q-superquadratic* convergence is indicated by

$$\lim_{k\to\infty} \frac{\beta_{k+1}}{\beta_k^2} = 0.$$

(a) Show that the following implications are valid: Q-superquadratic $\Rightarrow$ Q-quadratic $\Rightarrow$ Q-superlinear $\Rightarrow$ Q-linear.

(b) Give examples of sequences which do not imply the opposite directions in the three cases above.

A zero converging sequence $\{\beta_k\}_{k=0}^{\infty}$ is said to converge *R-linearly* if it is dominated by a Q-linearly converging sequence. That is, if there is a Q-linearly converging sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ such that $0 \leq \beta_k \leq \hat{\beta}_k$.

(c) Give a sequence which is R-linearly converging but not Q-linearly converging.

5. Let $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$ such that $\boldsymbol{Q}$ is symmetric, and indefinite. Apply the gradient method with constant step. Show that if the starting point $\boldsymbol{x}_0$ belongs to the space spanned by the negative eigenvectors, the sequence generated by the gradient method diverges.

6. In light of Theorem 1.6.3, show that under Assumption 1.6.2, if we want to obtain $\|\boldsymbol{x}_k - \boldsymbol{x}^*\| < \varepsilon$, we need an order of $\ln(\ln \varepsilon^{-1})$ iterations for the Newton method.

7. In the Section 1.7, show that $\mathcal{L}_k = \{\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{k-1}\}$.

8. In the same section, arrive at the expression (1.6) for a strictly convex quadratic function.

9. Show that the secant equation is valid for BFGS, DFP and symmetric-rank-one formulae.

10. Given $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ and a non-singular matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, if $1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{u} \neq 0$, then the

$$(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{\boldsymbol{M}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{M}^{-1}}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{u}}. \quad \text{(Sherman-Morrison formula)}$$

Apply this formula to compute the inverses $\boldsymbol{B}_{k+1}$ of $\boldsymbol{H}_{k+1}$ for BFGS, DFP and symmetric-rank-one formulae.

11. Apply the quasi-Newton method with BFGS, DFP, and Symmetric-Rank-One updates for the strictly convex function $f(\boldsymbol{x}) = \alpha + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle$ with $\boldsymbol{A} \succ \boldsymbol{O}$.

Chapter 2

Smooth Convex Optimization

2.1 Smooth convex functions

Definition 2.1.1 A continuously differentiable function $f(\boldsymbol{x})$ is called *convex* on $\mathbb{R}^n$ (notation $\mathcal{F}^1(\mathbb{R}^n)$) if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle f'(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

if $-f(\boldsymbol{x})$ is convex, $f(\boldsymbol{x})$ is called *concave*.

Theorem 2.1.2 If $f \in \mathcal{F}^1(\mathbb{R}^n)$ and $f'(\boldsymbol{x}^*) = 0$, then $\boldsymbol{x}^*$ is the *global minimum* of $f(\boldsymbol{x})$ on $\mathbb{R}^n$.

Proof: Left for exercise. ∎

Lemma 2.1.3 If $f \in \mathcal{F}^1(\mathbb{R}^m)$, $\boldsymbol{b} \in \mathbb{R}^m$, and $\boldsymbol{A} : \mathbb{R}^n \to \mathbb{R}^m$, then

$$\phi(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) \in \mathcal{F}^1(\mathbb{R}^n).$$

Proof: Left for exercise. ∎

Example 2.1.4 The following functions are differentiable and convex:

1. $f(x) = e^x$
2. $f(x) = |x|^p, \quad p > 1$
3. $f(x) = \frac{x^2}{1+|x|}$
4. $f(x) = |x| - \ln(1 + |x|)$
5. $f(\boldsymbol{x}) = \sum_{i=1}^m e^{\alpha_i + \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle}$
6. $f(\boldsymbol{x}) = \sum_{i=1}^m |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i|^p, \quad p > 1$

Theorem 2.1.5 Let f be a continuously differentiable function. The following conditions are equivalent:

1. $f \in \mathcal{F}^1(\mathbb{R}^n)$.

2. $f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \ \forall \alpha \in [0,1]$.

3. $\langle f'(\boldsymbol{x}) - f'(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Proof: Left for exercise. ∎

Theorem 2.1.6 Let f be a twice continuously differentiable function. Then $f \in \mathcal{F}^2(\mathbb{R}^n)$ if and only if

$$f''(\boldsymbol{x}) \succeq \boldsymbol{O}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Proof: Let $f \in \mathcal{F}^2(\mathbb{R}^n)$, and denote $\boldsymbol{x}_\tau = \boldsymbol{x} + \tau \boldsymbol{s}, \ \tau > 0$. Then, from the previous result

$$\begin{aligned} 0 &\leq \frac{1}{\tau^2} \langle f'(\boldsymbol{x}_\tau) - f'(\boldsymbol{x}), \boldsymbol{x}_\tau - \boldsymbol{x} \rangle = \frac{1}{\tau} \langle f'(\boldsymbol{x}_\tau) - f'(\boldsymbol{x}), \boldsymbol{s} \rangle \\ &= \frac{1}{\tau} \int_0^\tau \langle f''(\boldsymbol{x} + \lambda \boldsymbol{s}) \boldsymbol{s}, \boldsymbol{s} \rangle d\lambda \\ &= \frac{F(\tau) - F(0)}{\tau} \end{aligned}$$

where $F(\tau) = \int_0^\tau \langle f''(\boldsymbol{x} + \lambda \boldsymbol{s}) \boldsymbol{s}, \boldsymbol{s} \rangle d\lambda$. Therefore, tending τ to 0, we get $0 \leq F'(0) = \langle f''(\boldsymbol{x}) \boldsymbol{s}, \boldsymbol{s} \rangle$, and we have the result.

Conversely, $\forall \boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned} f(\boldsymbol{y}) &= f(\boldsymbol{x}) + \langle f'(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \int_0^1 \int_0^\tau \langle f''(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle d\lambda d\tau \\ &\geq f(\boldsymbol{x}) + \langle f'(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle. \end{aligned}$$

∎

Corollary 2.1.7 Let f be a two times continuously differentiable function. $f \in \mathcal{F}_L^{2,1}(\mathbb{R}^n)$ if and only if $\boldsymbol{O} \preceq f''(\boldsymbol{x}) \preceq L\boldsymbol{I}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n$.

Theorem 2.1.8 Let f be a continuously differentiable function in $\mathbb{R}^n$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and $\alpha \in [0,1]$. Then the following conditions are equivalent:

1. $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$.

2. $0 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle f'(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$.

3. $f(\boldsymbol{x}) + \langle f'(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2L} \|f'(\boldsymbol{x}) - f'(\boldsymbol{y})\|^2 \leq f(\boldsymbol{y})$.

4. $0 \leq \frac{1}{L} \|f'(\boldsymbol{x}) - f'(\boldsymbol{y})\|^2 \leq \langle f'(\boldsymbol{x}) - f'(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$.

5. $0 \leq \langle f'(\boldsymbol{x}) - f'(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq L \|\boldsymbol{x} - \boldsymbol{y}\|^2$.

6. $f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) + \frac{\alpha(1-\alpha)}{2L} \|f'(\boldsymbol{x}) - f'(\boldsymbol{y})\|^2 \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})$.

7. $0 \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) - f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha(1-\alpha) \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$.