lower bound $(L/(2\varepsilon))^n$	:	10^{20} calls of the oracle
computational complexity of the oracle	:	at least n arithmetic operations
total complexity	:	10^{21} arithmetic operations
CPU	:	1GHz or 10^9 arithmetic operations per second
total time	:	10^{12} seconds
one year	:	$\leq 3.2 \times 10^7$ seconds
we need	:	≥ 10000 years

- If we change n by $n+1$, the # of calls of the oracle is multiplied by 100.
- If we multiply ε by 2, the arithmetic complexity is reduced by 1000.

We know from Corollary 4.5 that the number of iterations of the uniform grid method is at most $(\lfloor L/(2\varepsilon)\rfloor + 2)^n$. Theorem 4.6 showed that any method which uses only function evaluations requires at least $(\lfloor L/(2\varepsilon)\rfloor)^n$ calls to have a better performance than ε. If for instance we take $\varepsilon = \mathcal{O}(L/n)$, these two bounds coincide up to a constant factor. In this sense, the uniform grid method is an *optimal method for the class of problems* $\mathcal{P}$.

4.3 Optimality Conditions for Smooth Optimization Problems

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function on $\mathbb{R}^n$, $\bar{\boldsymbol{x}} \in \mathbb{R}^n$, and $\boldsymbol{s}$ be a direction in $\mathbb{R}^n$ such that $\|\boldsymbol{s}\|_2 = 1$. Consider the local decrease (or increase) of $f(\boldsymbol{x})$ along $\boldsymbol{s}$:

$$f'(\bar{\boldsymbol{x}}; \boldsymbol{s}) = \lim_{\alpha \to 0} \frac{1}{\alpha} \left[f(\bar{\boldsymbol{x}} + \alpha \boldsymbol{s}) - f(\bar{\boldsymbol{x}}) \right].$$

Since $f(\bar{\boldsymbol{x}} + \alpha\boldsymbol{s}) - f(\bar{\boldsymbol{x}}) = \alpha\langle \boldsymbol{\nabla f}(\bar{\boldsymbol{x}}), \boldsymbol{s}\rangle + o(\|\alpha\boldsymbol{s}\|_2)$, we have $f'(\bar{\boldsymbol{x}}; \boldsymbol{s}) = \langle \boldsymbol{\nabla f}(\bar{\boldsymbol{x}}), \boldsymbol{s}\rangle$.
Using the Cauchy-Schwarz inequality $-\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2 \leq \langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2$,

$$f'(\bar{\boldsymbol{x}}; \boldsymbol{s}) = \langle \boldsymbol{\nabla f}(\bar{\boldsymbol{x}}), \boldsymbol{s}\rangle \geq -\|\boldsymbol{\nabla f}(\bar{\boldsymbol{x}})\|_2.$$

Choosing in particular the direction $\bar{\boldsymbol{s}} = -\boldsymbol{\nabla f}(\bar{\boldsymbol{x}})/\|\boldsymbol{\nabla f}(\bar{\boldsymbol{x}})\|_2$,

$$f'(\bar{x}; \boldsymbol{s}) = -\left\langle \boldsymbol{\nabla f}(\bar{\boldsymbol{x}}), \frac{\boldsymbol{\nabla f}(\bar{\boldsymbol{x}})}{\|\boldsymbol{\nabla f}(\bar{\boldsymbol{x}})\|_2} \right\rangle = -\|\boldsymbol{\nabla f}(\bar{\boldsymbol{x}})\|_2.$$

Thus, the direction $-\boldsymbol{\nabla f}(\bar{\boldsymbol{x}})$ is the direction of the *fastest local decrease of* $f(\boldsymbol{x})$ *at point* $\bar{\boldsymbol{x}}$.

Theorem 4.8 (First-order necessary optimality condition) Let $\boldsymbol{x}^*$ be a local minimum of the differentiable function $f(\boldsymbol{x})$. Then

$$\boldsymbol{\nabla f}(\boldsymbol{x}^*) = \boldsymbol{0}.$$

Proof:
Let $\boldsymbol{x}^*$ be the local minimum of $f(\boldsymbol{x})$. Then, there is $r > 0$ such that for all $\boldsymbol{y}$ with $\|\boldsymbol{y} - \boldsymbol{x}^*\|_2 \leq r$, $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*)$.
Since f is differentiable on $\mathbb{R}^n$,

$$f(\boldsymbol{y}) = f(\boldsymbol{x}^*) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^*\rangle + o(\|\boldsymbol{y} - \boldsymbol{x}^*\|_2) \geq f(\boldsymbol{x}^*).$$

Dividing by $\|\boldsymbol{y} - \boldsymbol{x}^*\|_2$, and taking the limit $\boldsymbol{y} \to \boldsymbol{x}^*$,

$$\langle \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{s}\rangle \geq 0, \quad \forall \boldsymbol{s} \in \mathbb{R}^n, \quad \|\boldsymbol{s}\|_2 = 1.$$

Choosing $\boldsymbol{s} = \boldsymbol{e}_i$ and $\boldsymbol{s} = -\boldsymbol{e}_i$ for $(i = 1, 2, \ldots, n)$, we conclude that $\boldsymbol{\nabla f}(\boldsymbol{x}^*) = 0$. ∎

Remark 4.9 For the first-order sufficient optimality condition, we need convexity for the function $f(\boldsymbol{x})$. See Theorem 5.8.

Corollary 4.10 Let $\boldsymbol{x}^*$ be a local minimum of a differentiable function $f(\boldsymbol{x})$ subject to linear equality constraints

$$\boldsymbol{x} \in \mathcal{L} := \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\} \neq \emptyset,$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $m < n$.

Then, there exists a vector of multipliers $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ such that

$$\boldsymbol{\nabla f}(\boldsymbol{x}^*) = \boldsymbol{A}^T\boldsymbol{\lambda}^*.$$

Proof:

Consider the vectors $\boldsymbol{u}_i$ $(i = 1, 2, \dots, k)$ with $k \geq n - m$ which form an orthonormal basis of the null space of $\boldsymbol{A}$. Then, $\boldsymbol{x} \in \mathcal{L}$ can be represented as

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{t}) := \boldsymbol{x}^* + \sum_{i=1}^{k} t_i \boldsymbol{u}_i, \quad \boldsymbol{t} \in \mathbb{R}^k.$$

Moreover, the point $\boldsymbol{t} = \boldsymbol{0}$ is the local minimal solution of the function $\phi(\boldsymbol{t}) = f(\boldsymbol{x}(\boldsymbol{t}))$.

From Theorem 4.8, $\boldsymbol{\nabla}\phi(\boldsymbol{0}) = \boldsymbol{0}$. That is,

$$\frac{\partial \phi}{\partial t_i}(\boldsymbol{0}) = \langle \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{u}_i \rangle = 0, \quad i = 1, 2, \dots, k.$$

Now there is $\boldsymbol{t}^* \in \mathbb{R}^k$ and $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ such that

$$\boldsymbol{\nabla f}(\boldsymbol{x}^*) = \sum_{i=1}^{k} t_i^* \boldsymbol{u}_i + \boldsymbol{A}^T\boldsymbol{\lambda}^*.$$

For each $i = 1, 2, \dots, k$,

$$\langle \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{u}_i \rangle = t_i^* = 0.$$

Therefore, we have the result. ∎

The following type of result is called *theorems of the alternative*, and are closed related to duality theory in optimization.

Corollary 4.11 Given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{c} \in \mathbb{R}^n$, $\eta \in \mathbb{R}$, either

$$\left\{ \begin{array}{c} \langle \boldsymbol{c}, \boldsymbol{x} \rangle < \eta \\ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{array} \right. \text{ has a solution } \boldsymbol{x} \in \mathbb{R}^n, \tag{3}$$

or

$$\left(\begin{array}{c} \left\{ \begin{array}{c} \langle \boldsymbol{b}, \boldsymbol{\lambda} \rangle > 0 \\ \boldsymbol{A}^T\boldsymbol{\lambda} = \boldsymbol{0} \end{array} \right. \\ \text{or} \\ \left\{ \begin{array}{c} \langle \boldsymbol{b}, \boldsymbol{\lambda} \rangle \geq \eta \\ \boldsymbol{A}^T\boldsymbol{\lambda} = \boldsymbol{c} \end{array} \right. \end{array} \right) \text{ has a solution } \boldsymbol{\lambda} \in \mathbb{R}^m, \tag{4}$$

but never both

Proof:

Let us first show that if $\exists \boldsymbol{x} \in \mathbb{R}^n$ satisfying (3), $\nexists \boldsymbol{\lambda} \in \mathbb{R}^m$ satisfying (4). Let us assume by contradiction that $\exists \boldsymbol{\lambda}$. Then $\langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} \rangle = \langle \boldsymbol{\lambda}, \boldsymbol{b} \rangle$ and in the homogeneous case it gives $0 = \langle \boldsymbol{\lambda}, \boldsymbol{b} \rangle > 0$ and in the non-homogeneous case it gives $\eta > \langle \boldsymbol{c}, \boldsymbol{x} \rangle = \langle \boldsymbol{\lambda}, \boldsymbol{b} \rangle \geq \eta$. Both of cases are impossible.

Now, let us assume that $\nexists \boldsymbol{x} \in \mathbb{R}^n$ satisfying (3). If additionally $\nexists \boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, it means that the columns of the matrix $\boldsymbol{A}$ do not spam the vector $\boldsymbol{b}$. Therefore, there is $\boldsymbol{0} \neq \boldsymbol{\lambda} \in \mathbb{R}^m$ which is orthogonal to all of these columns and $\langle \boldsymbol{b}, \boldsymbol{\lambda} \rangle \neq 0$. Selecting the correct sign, we constructed a $\boldsymbol{\lambda}$ which satisfies the homogeneous system of (4). Now, if for all $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ we have

$\langle \boldsymbol{c}, \boldsymbol{x} \rangle \geq \eta$, it means that the minimization of the function $f(\boldsymbol{x}) = \langle \boldsymbol{c}, \boldsymbol{x} \rangle$ subject to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ has an optimal solution $\boldsymbol{x}^*$ with $f(\boldsymbol{x}^*) \geq \eta$ (since $\exists \boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, we can always assume that $m \leq n$ eliminating redundant linear constraints from the system. If $n = m$ and $\boldsymbol{A}$ is nonsingular, take $\boldsymbol{\lambda} = \boldsymbol{A}^{-T}\boldsymbol{c}$. Otherwise, we can eliminate again redundant linear constraint to have $n > m$). From Corollary 4.10, $\exists \boldsymbol{\lambda} \in \mathbb{R}^m$ such that $\boldsymbol{A}^T\boldsymbol{\lambda} = \boldsymbol{c}$, and $\langle \boldsymbol{b}, \boldsymbol{\lambda} \rangle = \langle \boldsymbol{x}^*, \boldsymbol{A}^T\boldsymbol{\lambda} \rangle = \langle \boldsymbol{x}^*, \boldsymbol{c} \rangle \geq \eta$. ∎

If $f(\boldsymbol{x})$ is twice differentiable at $\bar{\boldsymbol{x}} \in \mathbb{R}^n$, then for $\boldsymbol{y} \in \mathbb{R}^n$, we have

$$\boldsymbol{\nabla f}(\boldsymbol{y}) = \boldsymbol{\nabla f}(\bar{\boldsymbol{x}}) + \boldsymbol{\nabla}^2\boldsymbol{f}(\bar{\boldsymbol{x}})(\boldsymbol{y} - \bar{\boldsymbol{x}}) + \boldsymbol{o}(\|\boldsymbol{y} - \bar{\boldsymbol{x}}\|_2),$$

where $\boldsymbol{o}(r)$ is such that $\lim_{r\to 0} \|\boldsymbol{o}(r)\|_2/r = 0$ and $\boldsymbol{o}(0) = 0$.

Theorem 4.12 (Second-order necessary optimality condition) Let $\boldsymbol{x}^*$ be a local minimum of a twice continuously differentiable function $f(\boldsymbol{x})$. Then

$$\boldsymbol{\nabla f}(\boldsymbol{x}^*) = 0, \qquad \boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*) \succeq \boldsymbol{O}.$$

Proof:

Since $\boldsymbol{x}^*$ is a local minimum of $f(\boldsymbol{x})$, $\exists r > 0$ such that for all $\boldsymbol{y} \in \mathbb{R}^n$ which satisfy $\|\boldsymbol{y}-\boldsymbol{x}^*\|_2 \leq r$, $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*)$.

From Theorem 4.8, $\boldsymbol{\nabla f}(\boldsymbol{x}^*) = 0$. Then

$$f(\boldsymbol{y}) = f(\boldsymbol{x}^*) + \frac{1}{2}\langle \boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*)(\boldsymbol{y} - \boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle + o(\|\boldsymbol{y} - \boldsymbol{x}^*\|_2^2) \geq f(\boldsymbol{x}^*).$$

And $\langle \boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*)\boldsymbol{s}, \boldsymbol{s} \rangle \geq 0,\ \forall \boldsymbol{s} \in \mathbb{R}^n$ with $\|\boldsymbol{s}\|_2 = 1$. ∎

Theorem 4.13 (Second-order sufficient optimality condition) Let the function $f(\boldsymbol{x})$ be twice continuously differentiable on $\mathbb{R}^n$, and let $\boldsymbol{x}^*$ satisfy the following conditions:

$$\boldsymbol{\nabla f}(\boldsymbol{x}^*) = 0, \quad \boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*) \succ \boldsymbol{O}.$$

Then, $\boldsymbol{x}^*$ is a strict local minimum of $f(\boldsymbol{x})$.

Proof:

In a small neighborhood of $\boldsymbol{x}^*$, function $f(\boldsymbol{x}^*)$ can be represented as:

$$f(\boldsymbol{y}) = f(\boldsymbol{x}^*) + \frac{1}{2}\langle \boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*)(\boldsymbol{y} - \boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle + o(\|\boldsymbol{y} - \boldsymbol{x}^*\|_2^2).$$

Since $o(r)/r \to 0$, there is a $\bar{r} > 0$ such that for all $r \in [0, \bar{r}]$,

$$|o(r)| \leq \frac{r}{4}\lambda_1(\boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*)),$$

where $\lambda_1(\boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*))$ is the smallest eigenvalue of the symmetric matrix $\boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*)$ which is positive. Then

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*) + \frac{1}{2}\lambda_1(\boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*))\|\boldsymbol{y} - \boldsymbol{x}^*\|_2^2 + o(\|\boldsymbol{y} - \boldsymbol{x}^*\|_2^2).$$

W.L.O.G, considering that $\bar{r} < 1$, $|o(r^2)| \leq r^2\lambda_1(\boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*))/4$ for $r \in [0, \bar{r}]$, finally we arrived at

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*) + \frac{1}{4}\lambda_1(\boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}^*))\|\boldsymbol{y} - \boldsymbol{x}^*\|_2^2 > f(\boldsymbol{x}^*).$$

∎

4.4 Algorithms for Minimizing Smooth Functions

4.4.1 Steepest Descent Method

Consider $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function on its domain.

Steepest Descent Method	
Choose:	$\boldsymbol{x}_0 \in \mathbb{R}^n$
Iterate:	$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \boldsymbol{\nabla f}(\boldsymbol{x}_k), \;\; k = 0, 1, \ldots$

We consider four strategies for the step-size h_k:

1. **Constant Step**

   The sequence $\{h_k\}_{k=0}^{\infty}$ is chosen in *advance*. For example

   $$h_k := h > 0,$$
   $$h_k := \frac{h}{\sqrt{k+1}}.$$

   This is the simplest strategy.

2. **Exact Line Search (Cauchy Step-Size)**

   The sequence $\{h_k\}_{k=0}^{\infty}$ is chosen such that

   $$h_k := \arg\min_{h \geq 0} f(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)).$$

   This choice is only theoretical since even for the one dimensional case, it is very difficult and expensive.

3. **Goldstein-Armijo Rule**

   Find a sequence $\{h_k\}_{k=0}^{\infty}$ such that

   $$\begin{aligned} \alpha\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle &\leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}), \\ \beta\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle &\geq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}), \end{aligned}$$

   where $0 < \alpha < \beta < 1$ are fixed parameters.

   Since $f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k - h_k \boldsymbol{\nabla f}(\boldsymbol{x}_k))$,

   $$f(\boldsymbol{x}_k) - \beta h_k \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \leq f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \alpha h_k \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2.$$

   The acceptable steps exist unless $f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k))$ is not bounded from below.

4. **Barzilai-Borwein Step-Size**[1]

   Let us define $\boldsymbol{s}_{k-1} := \boldsymbol{x}_k - \boldsymbol{x}_{k-1}$ and $\boldsymbol{y}_{k-1} := \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}_{k-1})$. Then, we can define the Barzilai-Borwein (BB) step sizes $\{h_k^1\}_{k=1}^{\infty}$ and $\{h_k^2\}_{k=1}^{\infty}$:

   $$h_k^1 := \frac{\|\boldsymbol{s}_{k-1}\|_2^2}{\langle \boldsymbol{s}_{k-1}, \boldsymbol{y}_{k-1}\rangle},$$
   $$h_k^2 := \frac{\langle \boldsymbol{s}_{k-1}, \boldsymbol{y}_{k-1}\rangle}{\|\boldsymbol{y}_{k-1}\|_2^2}.$$

   The first step-size is the one which minimizes the following secant condition $\|\frac{1}{h}\boldsymbol{s}_{k-1} - \boldsymbol{y}_{k-1}\|_2^2$ while the second one minimizes $\|\boldsymbol{s}_{k-1} - h\boldsymbol{y}_{k-1}\|_2^2$.

[1] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA Journal of Numerical Analysis*, **8** (1988), pp. 141–148.

Now, consider the problem

$$\boxed{\begin{array}{c}\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})\\[2ex] \text{where } f\in\mathcal{C}_L^{1,1}(\mathbb{R}^n), \text{ and } f(\boldsymbol{x}) \text{ is bounded from below.}\end{array}}$$

Let us evaluate the result of one step of the steepest descent method.
Consider $\boldsymbol{y}=\boldsymbol{x}-h\boldsymbol{\nabla f}(\boldsymbol{x})$. From Lemma 3.6,

$$\begin{aligned} f(\boldsymbol{y}) &\leq f(\boldsymbol{x})+\langle\boldsymbol{\nabla f}(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle+\frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2\\ &= f(\boldsymbol{x})-h\|\boldsymbol{\nabla f}(\boldsymbol{x})\|_2^2+\frac{h^2L}{2}\|\boldsymbol{\nabla f}(\boldsymbol{x})\|_2^2\\ &= f(\boldsymbol{x})-h\left(1-\frac{h}{2}L\right)\|\boldsymbol{\nabla f}(\boldsymbol{x})\|_2^2. \end{aligned} \tag{5}$$

Thus, one step of the steepest descent method decreases the value of the objective function at least the following value for $h^*=1/L$.

$$f(\boldsymbol{y})\leq f(\boldsymbol{x})-\frac{1}{2L}\|\boldsymbol{\nabla f}(\boldsymbol{x})\|_2^2.$$

Now, for the Goldstein-Armijo Rule, since $\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-h_k\boldsymbol{\nabla f}(\boldsymbol{x}_k)$, we have:

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})\leq\beta h_k\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2,$$

and from (5)

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})\geq h_k\left(1-\frac{h_k}{2}L\right)\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2.$$

Therefore, $h_k\geq 2(1-\beta)/L$.

Also, substituting in

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})\geq\alpha h_k\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2\geq\frac{2}{L}\alpha(1-\beta)\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2.$$

Thus, in the three step-size strategies excepting the BB step size considered here, we can say that

$$\boxed{f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})\geq\frac{\omega}{L}\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2}$$

for some positive constant ω.

Summing up the above inequality we have:

$$\frac{\omega}{L}\sum_{k=0}^{N}\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2\leq f(\boldsymbol{x}_0)-f(\boldsymbol{x}_{N+1})\leq f(\boldsymbol{x}_0)-f^*$$

where f^* is the optimal value of the problem.

As a simple consequence we have

$$\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2\to 0 \quad\text{as}\quad k\to\infty.$$

Finally,

$$g_N^*:=\min_{0\leq k\leq N}\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2\leq\frac{1}{\sqrt{N+1}}\left[\frac{L}{\omega}(f(\boldsymbol{x}_0)-f^*)\right]^{1/2}. \tag{6}$$

Remark 4.14 $g_N^* \to 0$, but we cannot say anything about the rate of convergence of the sequence $\{f(\boldsymbol{x}_k)\}$ or $\{\boldsymbol{x}_k\}$.

Example 4.15 Consider the function $f(x,y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$. $(0,-1)^T$ and $(0,1)^T$ are local minimal solutions, but $(0,0)^T$ is a stationary point.

If we start the steepest descent method from $(1,0)^T$, we will only converge to the stationary point.

We focus now on the following problem class:

Model:	1. $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ 2. $f \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$ 3. $f(\boldsymbol{x})$ is bounded from below
Oracle:	Only function and gradient values are available
Approximate solution:	Find $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ such that $f(\bar{\boldsymbol{x}}) \le f(\boldsymbol{x}_0)$ and $\|\boldsymbol{\nabla f}(\bar{\boldsymbol{x}})\|_2 < \epsilon$

From (6), we have

$$g_N^* < \varepsilon \quad \text{if} \quad N+1 > \frac{L}{\omega\varepsilon^2}(f(\boldsymbol{x}_0) - f^*).$$

Remark 4.16 This is much better than the result of Theorem 4.6, since *it does not depend on* n.

Finally, consider the following problem under Assumption 4.17.

$$\boxed{\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})}$$

Assumption 4.17

1. $f \in \mathcal{C}_M^{2,2}(\mathbb{R}^n)$;
2. There is a local minimum $\boldsymbol{x}^*$ of the function $f(\boldsymbol{x})$;
3. We know some bound $0 < \ell \le L < \infty$ for the Hessian at $\boldsymbol{x}^*$:
$$\ell \boldsymbol{I} \preceq \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^*) \preceq L\boldsymbol{I};$$
4. Our starting point $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}^*$.

Theorem 4.18 Let $f(\boldsymbol{x})$ satisfy our assumptions above and let the starting point $\boldsymbol{x}_0$ be close enough to a local minimum:

$$r_0 = \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 < \bar{r} := \frac{2\ell}{M}.$$

Then, the steepest descent method with step-size $h^* = 2/(L+\ell)$ converges as follows:

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 \le \frac{\bar{r} r_0}{\bar{r} - r_0}\left(1 - \frac{2\ell}{L+3\ell}\right)^k.$$

This rate of convergence is called (R-)*linear.*

Proof:
In the steepest descent method, the iterates are $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \boldsymbol{\nabla f}(\boldsymbol{x}_k)$.
Since $\boldsymbol{\nabla f}(\boldsymbol{x}^*) = 0$,

$$\boldsymbol{\nabla} \boldsymbol{f}(\boldsymbol{x}_k) = \boldsymbol{\nabla} \boldsymbol{f}(\boldsymbol{x}_k) - \boldsymbol{\nabla} \boldsymbol{f}(\boldsymbol{x}^*) = \int_0^1 \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_k - \boldsymbol{x}^*))(\boldsymbol{x}_k - \boldsymbol{x}^*) d\tau = \boldsymbol{G}_k(\boldsymbol{x}_k - \boldsymbol{x}^*),$$

and therefore,

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}^* = \boldsymbol{x}_k - \boldsymbol{x}^* - h_k \boldsymbol{G}_k(\boldsymbol{x}_k - \boldsymbol{x}^*) = (\boldsymbol{I} - h_k \boldsymbol{G}_k)(\boldsymbol{x}_k - \boldsymbol{x}^*).$$

Let $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$. From Lemma 3.8,

$$\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^*) - \tau M r_k \boldsymbol{I} \preceq \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_k - \boldsymbol{x}^*)) \preceq \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^*) + \tau M r_k \boldsymbol{I}.$$

Integrating all parts from 0 to 1 and using our hypothesis,

$$(\ell - \frac{r_k}{2} M)\boldsymbol{I} \preceq \boldsymbol{G}_k \preceq (L + \frac{r_k}{2} M)\boldsymbol{I}.$$

Therefore,

$$\left(1 - h_k(L + \frac{r_k}{2} M)\right) \boldsymbol{I} \preceq \boldsymbol{I} - h_k \boldsymbol{G}_k \preceq \left(1 - h_k(\ell - \frac{r_k}{2} M)\right) \boldsymbol{I}.$$

We arrive at

$$\|\boldsymbol{I} - h_k \boldsymbol{G}_k\|_2 \leq \max\{|a_k(h_k)|, |b_k(h_k)|\}$$

where $a_k(h) = 1 - h(\ell - \frac{r_k}{2} M)$ and $b_k(h) = h(L + \frac{r_k}{2} M) - 1$.

Notice that $a_k(0) = 1$ and $b_k(0) = -1$.

Now, let us use our hypothesis that $r_0 < \bar{r}$.

When $a_k(h) = b_k(h)$, we have $1 - h(\ell - \frac{r_k}{2} M) = h(L + \frac{r_k}{2} M) - 1$, and therefore

$$h_k^* = \frac{2}{L + \ell}.$$

(Surprisingly, it does not depend neither on M nor r_k). Finally,

$$r_{k+1} = \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2 \leq \left(1 - \frac{2}{L + \ell}\left(\ell - \frac{r_k}{2} M\right)\right) \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2.$$

That is,

$$r_{k+1} \leq \left(\frac{L - \ell}{L + \ell} + \frac{r_k M}{L + \ell}\right) r_k.$$

and $r_{k+1} < r_k < \bar{r}$.

Now, let us analyze the rate of convergence. Multiplying the above inequality by $M/(L + \ell)$,

$$\frac{M r_{k+1}}{L + \ell} \leq \frac{M(L - \ell)}{(L + \ell)^2} r_k + \frac{M^2 r_k^2}{(L + \ell)^2}.$$

Calling $\alpha_k = \frac{M r_k}{L + \ell}$ and $q = \frac{2\ell}{L + \ell}$, we have

$$\alpha_{k+1} \leq (1 - q)\alpha_k + \alpha_k^2 = \alpha_k(1 + \alpha_k - q) = \frac{\alpha_k(1 - (\alpha_k - q)^2)}{1 - (\alpha_k - q)}. \tag{7}$$

Now, since $r_k < \frac{2\ell}{M}$, $\alpha_k - q = \frac{M r_k}{L + \ell} - \frac{2\ell}{L + \ell} < 0$, and $1 + (\alpha_k - q) = \frac{L - \ell}{L + \ell} + \frac{M r_k}{L + \ell} > 0$. Therefore, $-1 < \alpha_k - q < 0$, and (7) becomes $\leq \frac{\alpha_k}{1 + q - \alpha_k}$.

$$\frac{1}{\alpha_{k+1}} \geq \frac{1 + q}{\alpha_k} - 1.$$

$$\frac{q}{\alpha_{k+1}} - 1 \geq \frac{q(1 + q)}{\alpha_k} - q - 1 = (1 + q)\left(\frac{q}{\alpha_k} - 1\right).$$

and then,

$$\frac{q}{\alpha_k} - 1 \geq (1 + q)^k \left(\frac{q}{\alpha_0} - 1\right) = (1 + q)^k \left(\frac{2\ell}{L + \ell} \frac{L + \ell}{M r_0} - 1\right) = (1 + q)^k \left(\frac{\bar{r}}{r_0} - 1\right),$$

$$\frac{\alpha_k}{q} \leq \left(\frac{1}{1 + q}\right)^k \left(\frac{r_0}{\bar{r} - r_0}\right).$$

Finally, we arrive at

$$r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 \leq \frac{\bar{r} r_0}{\bar{r} - r_0}\left(1 - \frac{2\ell}{L + 3\ell}\right)^k.$$

∎

4.4.2 The Newton Method

Example 4.19 Let us apply the Newton method to find the root of the following function

$$g(x) = \frac{x}{\sqrt{1 + x^2}}.$$

Clearly $x^* = 0$.

The Newton method will give:

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)} = x_k - x_k(1 + x_k^2) = -x_k^3.$$

Therefore, the method converges if $|x_0| < 1$, it oscillates if $|x_0| = 1$, and finally, diverges if $|x_0| > 1$.

Assumption 4.20

1. $f \in \mathcal{C}_M^{2,2}(\mathbb{R}^n)$;
2. There is a local minimum $\boldsymbol{x}^*$ of the function $f(\boldsymbol{x})$;
3. The Hessian is positive definite at $\boldsymbol{x}^*$:

$$\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^*) \succeq \ell \boldsymbol{I}, \quad \ell > 0;$$

4. Our starting point $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}^*$.

Theorem 4.21 Let the function $f(\boldsymbol{x})$ satisfy the above assumptions. Suppose that the initial starting point $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}^*$:

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 < \bar{r} := \frac{2\ell}{3M}.$$

Then $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 < \bar{r}$ for all k of the Newton method and it converges (Q-)quadratically:

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2 \leq \frac{M\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2}{2(\ell - M\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2)}.$$

Proof:

Let $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$. From Lemma 3.8 and the assumption, we have for $k = 0$,

$$\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0) \succeq \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^*) - M r_0 \boldsymbol{I} \succeq (\ell - M r_0)\boldsymbol{I}. \tag{8}$$

Since $r_0 < \bar{r} = \frac{2\ell}{3M} < \frac{\ell}{M}$, we have $\ell - M r_0 > 0$ and therefore, $\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)$ is invertible.

Consider the Newton method for $k = 0$, $\boldsymbol{x}_1 = \boldsymbol{x}_0 - [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1} \boldsymbol{\nabla} \boldsymbol{f}(\boldsymbol{x}_0)$.

Then

$$\begin{aligned}
\boldsymbol{x}_1 - \boldsymbol{x}^* &= \boldsymbol{x}_0 - \boldsymbol{x}^* - [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1} \boldsymbol{\nabla} \boldsymbol{f}(\boldsymbol{x}_0) \\
&= \boldsymbol{x}_0 - \boldsymbol{x}^* - [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1} \int_0^1 \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_0 - \boldsymbol{x}^*))(\boldsymbol{x}_0 - \boldsymbol{x}^*) d\tau \\
&= [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1} \boldsymbol{G}_0 (\boldsymbol{x}_0 - \boldsymbol{x}^*)
\end{aligned}$$

where $\boldsymbol{G}_0 = \int_0^1 [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_0 - \boldsymbol{x}^*))] d\tau$.

Then

$$\begin{aligned}
\|\boldsymbol{G}_0\|_2 &= \left\| \int_0^1 [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_0 - \boldsymbol{x}^*))] d\tau \right\|_2 \\
&\leq \int_0^1 \|\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_0 - \boldsymbol{x}^*))\|_2 d\tau \\
&\leq \int_0^1 M|1-\tau| r_0 d\tau = \frac{r_0}{2} M.
\end{aligned}$$

From (8),

$$\|[\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1}\|_2 \leq (\ell - M r_0)^{-1}.$$

Then

$$r_1 \leq \frac{M r_0^2}{2(\ell - M r_0)}.$$

Since $r_0 < \bar{r} = \frac{2\ell}{3M}$, we have $M r_0 < 2(\ell - M r_0)$ and therefore $\frac{M r_0}{2(\ell - M r_0)} < 1$. Then $r_1 < r_0$.

One can see now that the same argument is valid for all k's. ∎

- Comparing this result with the rate of convergence of the steepest descent, we see that the Newton method is much faster.
- Surprisingly, the region of *quadratic convergence* of the Newton method is almost the same as the region of the *linear convergence* of the gradient method.

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 < \frac{2\ell}{M} \quad \text{(steepest descent method)} \quad \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 < \frac{2\ell}{3M} \quad \text{(Newton method)}$$

- This justifies a standard recommendation to use the steepest descent method only at the initial stage of the minimization process in order to get close to a local minimum and then perform the Newton method to refine.

4.4.3 The Conjugate Gradient Methods

The conjugate gradient methods were initially proposed for minimizing convex quadratic functions. Consider the problem

$$\boxed{\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})}$$

with $f(\boldsymbol{x}) = \alpha + \langle \boldsymbol{a}, \boldsymbol{x} \rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle$, $\boldsymbol{A}^T = \boldsymbol{A} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{A} \succ \boldsymbol{O}$. Since its minimal solution is $\boldsymbol{x}^* = -\boldsymbol{A}^{-1}\boldsymbol{a}$, we can rewrite $f(\boldsymbol{x})$ as:

$$\begin{aligned}
f(\boldsymbol{x}) &= \alpha - \langle \boldsymbol{A}\boldsymbol{x}^*, \boldsymbol{x} \rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle \\
&= \alpha - \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}^*, \boldsymbol{x}^* \rangle + \frac{1}{2}\langle \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle.
\end{aligned}$$

Thus, $f(\boldsymbol{x}^*) = \alpha - \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}^*, \boldsymbol{x}^* \rangle$ and $\boldsymbol{\nabla} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*)$.

Definition 4.22 Given a starting point $\boldsymbol{x}_0$, the linear *Krylov subspaces* are defined as

$$\mathcal{L}_k := \text{span}\{\boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}^*), \ldots, \boldsymbol{A}^k(\boldsymbol{x}_0 - \boldsymbol{x}^*)\}, \quad k \geq 1,$$

where $\text{span}\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p\}$ is the linear subspace of $\mathbb{R}^n$ spanned by the vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p \in \mathbb{R}^n$.

We claim temporarily that the sequence of points generated by a *conjugate gradient method* is defined as follows:

$$\boxed{\boldsymbol{x}_k := \arg\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \boldsymbol{x}_0 + \mathcal{L}_k\}, \;\; k \geq 1.}$$

Lemma 4.23 For any $k \geq 1$, $\mathcal{L}_k = \mathrm{span}\{\boldsymbol{\nabla f}(\boldsymbol{x}_0), \dots, \boldsymbol{\nabla f}(\boldsymbol{x}_{k-1})\}$.

Proof:

Let us prove by induction hypothesis.

For $k = 1$, the statement is true since $\boldsymbol{\nabla f}(\boldsymbol{x}_0) = \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}^*)$.

Suppose the claim is true for some $k \geq 1$. Then from the definition of the conjugate gradient method,

$$\boldsymbol{x}_k = \boldsymbol{x}_0 + \sum_{i=1}^{k} \lambda_i \boldsymbol{A}^i(\boldsymbol{x}_0 - \boldsymbol{x}^*)$$

with some $\lambda_i \in \mathbb{R}, \quad i = 1, \dots, k$. Therefore,

$$\boldsymbol{\nabla f}(\boldsymbol{x}_k) = \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}^*) + \sum_{i=1}^{k} \lambda_i \boldsymbol{A}^{i+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) = \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}^*) + \sum_{i=1}^{k-1} \lambda_i \boldsymbol{A}^{i+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) + \lambda_k \boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*).$$

The first two terms of the last expression belongs to $\mathcal{L}_k$ from the definition. And then,

$$\mathrm{span}\{\mathcal{L}_k, \boldsymbol{\nabla f}(\boldsymbol{x}_k)\} \subseteq \mathrm{span}\{\mathcal{L}_k, \boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*)\} = \mathcal{L}_{k+1}.$$

There are two ways to show that the equality holds.

Assume that $\boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) \in \mathcal{L}_k$. Then it is obvious and $\mathcal{L}_k = \mathcal{L}_{k+1}$. If $\boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) \notin \mathcal{L}_k$, the equality holds unless $\lambda_k = 0$. However, this possibility implies that $\boldsymbol{x}_k \in \mathcal{L}_{k-1}$, $\boldsymbol{x}_{k-1} = \boldsymbol{x}_k$ and therefore, $\mathcal{L}_{k-1} = \mathcal{L}_k = \mathcal{L}_{k+1}$ again.

An alternative way is to use contradiction. If the equality does not hold, $\boldsymbol{\nabla f}(\boldsymbol{x}_k) \in \mathcal{L}_k$ implies $\boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) \in \mathcal{L}_k$, which again implies the equality, or $\lambda_k = 0$, which implies that $\boldsymbol{x}_k = \boldsymbol{x}_{k-1}$ (algorithm terminated). ∎

Lemma 4.24 For any $k, \ell \geq 0, \; k \neq \ell$, we have $\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{\nabla f}(\boldsymbol{x}_\ell)\rangle = 0$.

Proof:

Let $k \geq i$, and consider

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

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

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

∎

Corollary 4.25 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 $\mathcal{L}_k = \mathrm{span}\{\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_{k-1}\}$ (Exercise 10).

Lemma 4.26 For any $k, \ell \geq 0$, $k \neq \ell$, $\langle \boldsymbol{A}\boldsymbol{\delta}_k, \boldsymbol{\delta}_\ell \rangle = 0$.

Proof:
Left for exercise. ∎

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:

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

Using the previous properties, we arrive that (see Exercise 11)

$$\lambda_j = 0, \quad (j = 0, 1, \ldots, k-2), \quad \lambda_{k-1} = \frac{h_k \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2}{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}_{k-1}), \boldsymbol{\delta}_{k-1} \rangle}. \tag{9}$$

Thus

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

where

$$\boldsymbol{p}_k = \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \frac{\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \boldsymbol{p}_{k-1}}{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}_{k-1}), \boldsymbol{p}_{k-1} \rangle}.$$

Finally, we can present the Conjugate Gradient Method

Conjugate Gradient Method	
Step 0:	Let $\boldsymbol{x}_0 \in \mathbb{R}^n$, compute $f(\boldsymbol{x}_0)$, $\boldsymbol{\nabla f}(\boldsymbol{x}_0)$ and set $\boldsymbol{p}_0 := \boldsymbol{\nabla f}(\boldsymbol{x}_0)$, $k := 0$
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 $\boldsymbol{\nabla f}(\boldsymbol{x}_{k+1})$
Step 3:	Compute the coefficient β_{k+1}
Step 4:	Set $p_{k+1} := \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \beta_{k+1} \boldsymbol{p}_k$, $k := k+1$ and go to Step 1

The most popular choices for the coefficient β_k are:

1. *Hestenes-Stiefel (1952):* $\beta_{k+1} = \frac{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}), \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k) \rangle}{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{p}_k \rangle}$.
2. *Fletcher-Reeves (1964):* $\beta_{k+1} = \frac{\|\boldsymbol{\nabla f}(\boldsymbol{x}_{k+1})\|_2^2}{\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2}$.
3. *Polak-Ribière (1969):* $\beta_{k+1} = \frac{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}), \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k) \rangle}{\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2}$.
4. *Polak-Ribière plus:* $\beta_{k+1} = \max\left\{0, \frac{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}), \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k) \rangle}{\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2}\right\}$.
5. *Dai-Yuan (1999):* $\beta_{k+1} = \frac{\|\boldsymbol{\nabla f}(\boldsymbol{x}_{k+1})\|_2^2}{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{p}_k \rangle}$.

Among them, Hestenes-Stiefel and Polak-Ribière are empirically preferred.

4.4.4 Quasi-Newton Methods

The basic idea of 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 $\boldsymbol{y}_k = \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k)$, $\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$.

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 $f(\boldsymbol{x}_0), \boldsymbol{\nabla f}(\boldsymbol{x}_0)$
Step 1:	Set $\boldsymbol{p}_k := \boldsymbol{H}_k \boldsymbol{\nabla f}(\boldsymbol{x}_k)$
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 $\boldsymbol{\nabla f}(\boldsymbol{x}_{k+1})$
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)*

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

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

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

3. *Symmetric-Rank-One*

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

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 [Nocedal].

4.5 Exercises

1. Let $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ continuously differentiable functions and $\boldsymbol{h} \in \mathbb{R}^m$. Define the following optimization problem.

$$\begin{cases} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & g(\boldsymbol{x}) = \boldsymbol{h} \\ & \boldsymbol{x} \in \mathbb{R}^n \end{cases}$$

   Write the Karush-Kuhn-Tucker (KKT) conditions corresponding to the above problem.

2. In view of Theorem 4.13, find a twice continuously differentiable function on $\mathbb{R}^n$ which satisfies $\boldsymbol{\nabla f}(\boldsymbol{x}^*) = 0, \quad \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^*) \succeq \boldsymbol{O}$, but $\boldsymbol{x}^*$ is not a local minimum of $f(\boldsymbol{x})$.

3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous differentiable and convex function. If $\boldsymbol{x}^* \in \mathbb{R}^n$ is such that $\boldsymbol{\nabla f}(\boldsymbol{x}^*) = \boldsymbol{0}$, then show that $\boldsymbol{x}^*$ is a global minimum for $f(\boldsymbol{x})$.

4. Determine the Cauchy step-size $h_k \in \mathbb{R}$ for the following strictly convex quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x} + \gamma$, where $\boldsymbol{Q}$ is a $n \times n$ real positive definite matrix.

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

   Let $0 < c_1 < c_2 < 1$,

   - Wolfe condition

   $$\begin{aligned} f(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)) &\le f(\boldsymbol{x}_k) - c_1 h \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2, \\ \langle \boldsymbol{\nabla f}(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)), \boldsymbol{\nabla f}(\boldsymbol{x}_k)\rangle &\le c_2 \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2. \end{aligned}$$

   - Strong Wolfe condition

   $$\begin{aligned} f(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)) &\le f(\boldsymbol{x}_k) - c_1 h \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2, \\ |\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)), \boldsymbol{\nabla f}(\boldsymbol{x}_k)\rangle| &\le c_2 \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2. \end{aligned}$$

6. 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

   $$\left| \frac{\beta_{k+1}}{\beta_k} \right| \le \rho,$$

   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

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

   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 $0 \le |\beta_k| \le \hat{\beta}_k$.

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

7. Let $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$ such that $\boldsymbol{Q}$ is symmetric and indefinite. Apply the steepest descent 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 steepest descent method diverges.

8. Prove Lemma 4.26.