# Lecture Notes of Optimization Theory (2024)

Luo Luo
School of Data Science, Fudan University

March 28, 2024

## 1 Review of Linear Algebra

**Woodbury Identity** For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $\mathbf{D} \in \mathbb{R}^{p\times n}$, we have

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}$$

if $\mathbf{A}$ and $\mathbf{A}+\mathbf{BCD}$ are non-singular. For given $\mathbf{A}^{-1}$ and $p \ll n$, achieving $(\mathbf{A}+\mathbf{BCD})^{-1}$ requires

$$\mathcal{O}(n^2+p^3+n^2p) = \mathcal{O}(n^2)$$

flops, which is more efficient than directly computing $(\mathbf{A}+\mathbf{BCD})^{-1}$ that requires $\mathcal{O}(n^3)$.

**Lemma 1.1.** *For $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{m\times n}$, we have*

$$\frac{\partial \mathrm{tr}(\mathbf{A}^\top \mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}.$$

*Proof.* We have

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}, \quad \mathbf{A}^\top = \begin{bmatrix} a_{11} & \dots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \dots & a_{mn} \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{bmatrix},$$

which implies

$$(\mathbf{A}^\top \mathbf{X})_{jj} = \sum_{i=1}^m a_{ij}x_{ij} \quad \text{and} \quad \mathrm{tr}(\mathbf{A}^\top \mathbf{X}) = \sum_{j=1}^m \sum_{i=1}^n a_{ij}x_{ij}.$$

Therefore, we achieve

$$\frac{\partial \mathrm{tr}(\mathbf{A}^\top \mathbf{X})}{\partial x_{ij}} = a_{ij} \quad \text{and} \quad \frac{\partial \mathrm{tr}(\mathbf{A}^\top \mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}.$$

□

**Multivariate Linear Regression** Let

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_N^\top \end{bmatrix} \in \mathbb{R}^{N\times p} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1^\top \\ \vdots \\ \mathbf{b}_N^\top \end{bmatrix} \in \mathbb{R}^{N\times q}$$

We also suppose $\mathbf{A}$ is full rank and $N > p$. For given positive-definite $\mathbf{W} \in \mathbb{R}^{q\times q}$, we consider the loss function $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ as follows

$$f(\mathbf{X}) = \sum_{i=1}^{N} \left\|\mathbf{X}^\top \mathbf{a}_i - \mathbf{b}_i\right\|_{\mathbf{W}}^2 = \mathrm{tr}\left((\mathbf{A}\mathbf{X} - \mathbf{B})\mathbf{W}(\mathbf{X}^\top\mathbf{A}^\top - \mathbf{B}^\top)\right)$$

with respect to $\mathbf{X}$. For $\mathbf{W} = \mathbf{I}$, it is well-known that

$$\begin{aligned} f(\mathbf{X}) =& \mathrm{tr}\left((\mathbf{A}\mathbf{X} - \mathbf{B})(\mathbf{X}^\top\mathbf{A}^\top - \mathbf{B}^\top)\right) \\ =& \mathrm{tr}\left(\mathbf{A}\mathbf{X}\mathbf{X}^\top\mathbf{A}^\top\right) - \mathrm{tr}\left(\mathbf{A}\mathbf{X}\mathbf{B}^\top\right) - \mathrm{tr}\left(\mathbf{B}\mathbf{X}^\top\mathbf{A}^\top\right) + \mathrm{tr}\left(\mathbf{B}\mathbf{B}^\top\right) \\ =& \mathrm{tr}\left(\mathbf{X}^\top\mathbf{A}^\top\mathbf{A}\mathbf{X}\right) - 2\mathrm{tr}\left(\mathbf{X}^\top\mathbf{A}^\top\mathbf{B}\right) + \mathrm{tr}\left(\mathbf{B}\mathbf{B}^\top\right) \end{aligned}$$

and

$$\begin{aligned} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} =& \frac{\partial \mathrm{tr}\left(\mathbf{X}^\top\mathbf{A}^\top\mathbf{A}\mathbf{X}\right) - 2\mathrm{tr}\left((\mathbf{A}^\top\mathbf{B})^\top\mathbf{X}\right) + \mathrm{tr}\left(\mathbf{B}\mathbf{B}^\top\right)}{\partial \mathbf{X}} \\ =& 2(\mathbf{A}^\top\mathbf{A}\mathbf{X} - \mathbf{A}^\top\mathbf{B}). \end{aligned}$$

Setting above gradient be zero leads to

$$\mathbf{X} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{B},$$

which is the solution of

$$\min_{\mathbf{X}\in\mathbb{R}^{p\times q}} f(\mathbf{X}) = \sum_{i=1}^{N} \left\|\mathbf{X}^\top \mathbf{a}_i - \mathbf{b}_i\right\|_2^2$$

**Tricks for Matrix Calculus** Recall the relationship between differential and derivative/gradient as follows

1. For single value input function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\mathrm{d}f(x) = f'(x)\,\mathrm{d}x.$$

2. For vector input function $f : \mathbb{R}^p \to \mathbb{R}$, we have

$$\mathrm{d}f(\mathbf{x}) = \sum_{i=1}^{p} \frac{\partial f(\mathbf{x})}{\partial x_i} \cdot \mathrm{d}x_i = \langle \nabla f(\mathbf{x}), \mathrm{d}\mathbf{x}\rangle,$$

where

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_p} \end{bmatrix} \in \mathbb{R}^p \qquad \text{and} \qquad \mathrm{d}\mathbf{x} = \begin{bmatrix} \mathrm{d}x_1 \\ \vdots \\ \mathrm{d}x_p \end{bmatrix} \in \mathbb{R}^p.$$

3. For scalar variables $x, y \in \mathbb{R}^d$, we have

$$\mathrm{d}(xy) = y\mathrm{d}x + x\mathrm{d}y.$$

For matrix variates $\mathbf{X} \in \mathbb{R}^{p\times q}$ and $\mathbf{Y} \in \mathbb{R}^{q\times r}$, we define

$$\mathrm{d}\mathbf{X} = \begin{bmatrix} \mathrm{d}x_{11} & \dots & \mathrm{d}x_{1q} \\ \vdots & \ddots & \vdots \\ \mathrm{d}x_{p1} & \dots & \mathrm{d}x_{pq} \end{bmatrix} \in \mathbb{R}^{p\times q} \qquad \text{and} \qquad \mathrm{d}\mathbf{Y} = \begin{bmatrix} \mathrm{d}y_{11} & \dots & \mathrm{d}y_{1r} \\ \vdots & \ddots & \vdots \\ \mathrm{d}y_{q1} & \dots & \mathrm{d}y_{qr} \end{bmatrix} \in \mathbb{R}^{q\times r}.$$

It holds that

$$\mathrm{d}(\mathbf{XY}) = (\mathrm{d}\mathbf{X})\mathbf{Y} + \mathbf{X}\mathrm{d}\mathbf{Y}.$$

We can verify above results as follows

$$\begin{aligned}
&(\mathrm{d}(\mathbf{XY}))_{ij} = \mathrm{d}(\mathbf{XY})_{ij} \\
=&\mathrm{d}\sum_{k=1}^{q} x_{ik}y_{kj} = \sum_{k=1}^{q}\mathrm{d}(x_{ik}y_{kj}) \\
=&\sum_{k=1}^{q}\big(x_{ik}\mathrm{d}y_{kj} + (\mathrm{d}x_{ik})y_{kj}\big) \\
=&\big(\mathbf{X}\mathrm{d}\mathbf{Y}\big)_{ij} + \big((\mathrm{d}\mathbf{X})\mathbf{Y}\big)_{ij} \\
=&\big(\mathbf{X}\mathrm{d}\mathbf{Y} + (\mathrm{d}\mathbf{X})\mathbf{Y}\big)_{ij}.
\end{aligned}$$

If $\mathbf{Y} \in \mathbb{R}^{q\times r}$ is constant, we have

$$\mathrm{d}(\mathbf{XY}) = (\mathrm{d}\mathbf{X})\mathbf{Y}.$$

If $\mathbf{X} \in \mathbb{R}^{p\times q}$ is constant, we have

$$\mathrm{d}(\mathbf{XY}) = \mathbf{X}(\mathrm{d}\mathbf{Y}).$$

For $\mathbf{Z} \in \mathbb{R}^{p\times p}$, we have

$$\mathrm{d}\mathrm{tr}(\mathbf{Z}) = \mathrm{d}\left(\sum_{i=1}^{p} z_{ii}\right) = \sum_{i=1}^{p}\mathrm{d}z_{ii} = \mathrm{tr}(\mathrm{d}\mathbf{Z}).$$

4. For matrix input function $f : \mathbb{R}^{p\times q} \to \mathbb{R}$, we have

$$\begin{aligned}
\mathrm{d}f(\mathbf{X}) =& \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\partial f(\mathbf{X})}{\partial x_{ij}}\cdot \mathrm{d}x_{ij} \\
=&\langle \nabla f(\mathbf{X}), \mathrm{d}\mathbf{X}\rangle \\
=&\mathrm{tr}\big(\nabla f(\mathbf{X})^\top \mathrm{d}\mathbf{X}\big),
\end{aligned}$$

where

$$\nabla f(\mathbf{X}) = \begin{bmatrix} \dfrac{\partial f(\mathbf{X})}{\partial x_{11}} & \cdots & \dfrac{\partial f(\mathbf{X})}{\partial x_{1q}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f(\mathbf{X})}{\partial x_{p1}} & \cdots & \dfrac{\partial f(\mathbf{X})}{\partial x_{pq}} \end{bmatrix} \in \mathbb{R}^{p\times q} \qquad \text{and} \qquad \mathrm{d}\mathbf{X} = \begin{bmatrix} \mathrm{d}x_{11} & \cdots & \mathrm{d}x_{1q} \\ \vdots & \ddots & \vdots \\ \mathrm{d}x_{p1} & \cdots & \mathrm{d}x_{pq} \end{bmatrix} \in \mathbb{R}^{p\times q}.$$

This implies if the differential $\mathrm{d}f(\mathbf{X})$ has the form of

$$\mathrm{d}f(\mathbf{X}) = \mathrm{tr}\big(\mathbf{A}^\top \mathrm{d}\mathbf{X}\big),$$

then the gradient of $f(\mathbf{X})$ is $\mathbf{A}$.

**Revisiting Multivariate Linear Regression** We come back to the function

$$f(\mathbf{X}) = \sum_{i=1}^{N}\left\|\mathbf{X}^\top\mathbf{x}_i - \mathbf{y}_i\right\|_{\mathbf{W}}^2 = \mathrm{tr}\left((\mathbf{AX}-\mathbf{B})\mathbf{W}(\mathbf{X}^\top\mathbf{A}^\top - \mathbf{B}^\top)\right),$$

which holds

$$\begin{aligned} f(\mathbf{X}) =& \mathrm{tr}\left((\mathbf{AX}-\mathbf{B})\mathbf{W}(\mathbf{X}^\top\mathbf{A}^\top-\mathbf{B}^\top)\right) \\ =& \mathrm{tr}\left(\mathbf{AXWX}^\top\mathbf{A}^\top\right)-2\mathrm{tr}\left(\mathbf{AXWB}^\top\right)+\mathrm{tr}\left(\mathbf{BWB}^\top\right), \end{aligned} \tag{1}$$

then we write its differential as follows

$$\begin{aligned} \mathrm{d}f(\mathbf{X}) =& \mathrm{dtr}\left(\mathbf{AXWX}^\top\mathbf{A}^\top\right)-2\mathrm{dtr}\left(\mathbf{AXWB}^\top\right)+\mathrm{dtr}\left(\mathbf{BWB}^\top\right) \\ =& \mathrm{tr}\left(\mathrm{d}(\mathbf{AXWX}^\top\mathbf{A}^\top)\right)-2\mathrm{tr}\left(\mathrm{d}(\mathbf{AXWB}^\top)\right). \end{aligned} \tag{2}$$

For the first term, we have

$$\begin{aligned} & \mathrm{d}(\mathbf{AX}\cdot\mathbf{WX}^\top\mathbf{A}^\top) \\ =& \mathrm{d}(\mathbf{AX})\cdot\mathbf{WX}^\top\mathbf{A}^\top+\mathbf{AX}\cdot\mathrm{d}(\mathbf{WX}^\top\mathbf{A}^\top) \\ =& \mathbf{A}(\mathrm{d}\mathbf{X})\mathbf{WX}^\top\mathbf{A}^\top+\mathbf{AXW}(\mathrm{d}\mathbf{X}^\top)\mathbf{A}^\top, \end{aligned}$$

which implies

$$\begin{aligned} & \mathrm{tr}\left(\mathrm{d}(\mathbf{AXWX}^\top\mathbf{A}^\top)\right) \\ =& \mathrm{tr}\left(\mathbf{A}(\mathrm{d}\mathbf{X})\mathbf{WX}^\top\mathbf{A}^\top\right)+\mathrm{tr}\left(\mathbf{AXW}(\mathrm{d}\mathbf{X}^\top)\mathbf{A}^\top\right) \\ =& \mathrm{tr}\left(\mathbf{WX}^\top\mathbf{A}^\top\mathbf{A}\mathrm{d}\mathbf{X}\right)+\mathrm{tr}\left((\mathrm{d}\mathbf{X}^\top)\mathbf{A}^\top\mathbf{AXW}\right) \\ =& \mathrm{tr}\left(\mathbf{WX}^\top\mathbf{A}^\top\mathbf{A}\mathrm{d}\mathbf{X}\right)+\mathrm{tr}\left(\mathbf{WX}^\top\mathbf{A}^\top\mathbf{A}\mathrm{d}\mathbf{X}\right) \\ =& 2\mathrm{tr}\left(\mathbf{WX}^\top\mathbf{A}^\top\mathbf{A}\mathrm{d}\mathbf{X}\right) \end{aligned} \tag{3}$$

For the second term, we have

$$\begin{aligned} & 2\mathrm{tr}\left(\mathrm{d}(\mathbf{AXWB}^\top)\right) \\ =& 2\mathrm{tr}\left(\mathbf{A}(\mathrm{d}\mathbf{X})\mathbf{WB}^\top\right) \\ =& 2\mathrm{tr}\left(\mathbf{WB}^\top\mathbf{A}\mathrm{d}\mathbf{X}\right) \end{aligned} \tag{4}$$

Substituting equations (3) and (4) into (2), we have

$$\begin{aligned} \mathrm{d}f(\mathbf{X}) =& 2\mathrm{tr}\left(\mathbf{WX}^\top\mathbf{A}^\top\mathbf{A}\mathrm{d}\mathbf{X}\right)-2\mathrm{tr}\left(\mathbf{WB}^\top\mathbf{A}\mathrm{d}\mathbf{X}\right) \\ =& \mathrm{tr}\left(2\mathbf{W}(\mathbf{X}^\top\mathbf{A}^\top\mathbf{A}-\mathbf{B}^\top\mathbf{A})\mathrm{d}\mathbf{X}\right), \end{aligned}$$

which means

$$\begin{aligned} \nabla f(\mathbf{X}) =& \left(2\mathbf{W}(\mathbf{X}^\top\mathbf{A}^\top\mathbf{A}-\mathbf{B}^\top\mathbf{A})\right)^\top \\ =& 2\left(\mathbf{X}^\top\mathbf{A}^\top\mathbf{A}-\mathbf{B}^\top\mathbf{A}\right)^\top\mathbf{W} \\ =& 2\left(\mathbf{A}^\top\mathbf{AX}-\mathbf{A}^\top\mathbf{B}\right)\mathbf{W}. \end{aligned}$$

Hence, taking the gradient of $f(\cdot)$ with respect to $\mathbf{X}$ be zero leads to

$$\mathbf{X}=\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{B}.$$

If $\mathbf{A}^\top\mathbf{A}\in\mathbb{R}^{p\times p}$ is singular, the solution contains the term of pseudo-inverse of $\mathbf{A}$. We will give the detailed discussion in later section.

**Example 1.1.** *Let* $\mathbf{A}\in\mathbb{R}^{p\times p}$ *and* $f:\mathbb{R}^{p\times p}\to\mathbb{R}$ *be* $f(\mathbf{X})=\mathrm{tr}(\mathbf{AX}^{-1})$*, then we have*

$$\nabla f(\mathbf{X})=-\mathbf{X}^{-\top}\mathbf{A}^\top\mathbf{X}^{-\top}.$$

*Proof.* It holds

$$\mathbf{0} = \mathrm{d}(\mathbf{A}\mathbf{X}^{-1}\mathbf{X}) = \mathrm{d}(\mathbf{A}\mathbf{X}^{-1}) \cdot \mathbf{X} + \mathbf{A}\mathbf{X}^{-1} \cdot \mathrm{d}\mathbf{X},$$

which means

$$\mathrm{d}(\mathbf{A}\mathbf{X}^{-1}) = -\mathbf{A}\mathbf{X}^{-1} \cdot \mathrm{d}\mathbf{X} \cdot \mathbf{X}^{-1}.$$

Therefore, we have

$$\begin{aligned}\mathrm{tr}(\mathrm{d}(\mathbf{A}\mathbf{X}^{-1})) =& \mathrm{tr}(-\mathbf{A}\mathbf{X}^{-1} \cdot \mathrm{d}\mathbf{X} \cdot \mathbf{X}^{-1}) \\ =& \mathrm{tr}(-\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1} \cdot \mathrm{d}\mathbf{X}),\end{aligned}$$

which implies

$$\nabla f(\mathbf{X}) = (-\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1})^\top = -\mathbf{X}^{-\top}\mathbf{A}^\top\mathbf{X}^{-\top}.$$

□

**In the View of Linear Approximation** For single variable, we have

$$f'(x) = \lim_{\Delta h \to 0} \frac{f(x+h) - f(x)}{h}.$$

We have $g = f'(x)$ if and only if

$$\lim_{h \to 0} \frac{f(x+h) - f(x) - g \cdot h}{h} = 0.$$

That is, we estimate

$$f(x+h) \approx f(x) + f'(x) \cdot h$$

for small $h$. For $\mathbf{X} \in \mathbb{R}^{p \times q}$, we desire

$$f(\mathbf{X} + \mathbf{H}) \approx f(\mathbf{X}) + \langle \nabla f(\mathbf{X}), \mathbf{H} \rangle$$

for small $\mathbf{X}$. We have $\mathbf{G} = \nabla f(\mathbf{X}) \in \mathbb{R}^{p \times q}$ if and only if

$$\lim_{\mathbf{H} \to \mathbf{0}} \frac{f(\mathbf{X} + \mathbf{H}) - f(\mathbf{X}) - \langle \mathbf{G}, \mathbf{H} \rangle}{\|\mathbf{H}\|_F} = 0.$$



**Example 1.2.** *Let $f : \mathbb{S}^p_{++} \to \mathbb{R}$ be $f(\mathbf{X}) = \ln\det(\mathbf{X})$, we have $\nabla f(\mathbf{X}) = \mathbf{X}^{-1}$.*

*Proof.* We have

$$\begin{aligned}f(\mathbf{X} + \mathbf{H}) - f(\mathbf{X}) =& \ln\det(\mathbf{X} + \mathbf{H}) - \ln\det(\mathbf{X}) \\ =& \ln\det(\mathbf{X}^{1/2}(\mathbf{I} + \mathbf{X}^{-1/2}\mathbf{H}\mathbf{X}^{-1/2})\mathbf{X}^{1/2}) - \ln\det(\mathbf{X}) \\ =& \ln\det(\mathbf{X}^{1/2}) + \ln\det(\mathbf{I} + \mathbf{X}^{-1/2}\mathbf{H}\mathbf{X}^{-1/2}) + \ln(\mathbf{X}^{1/2}) - \ln\det(\mathbf{X}) \\ =& \ln\det(\mathbf{I} + \mathbf{X}^{-1/2}\mathbf{H}\mathbf{X}^{-1/2}) \\ =& \ln\prod_{i=1}^{p}(1 + \lambda_i) \\ =& \sum_{i=1}^{p}\ln(1 + \lambda_i)\end{aligned}$$

where $\lambda_i$ is the $i$-th largest eigenvalue of $\mathbf{X}^{-1/2}\mathbf{H}\mathbf{X}^{-1/2}$. We have $\lambda_i \to 0$ for $i = 1, \ldots, p$ when $\mathbf{H} \to \mathbf{0}$. Therefore, it holds

$$\begin{aligned}
0 =& \lim_{\mathbf{H}\to\mathbf{0}} \frac{f(\mathbf{X}+\mathbf{H}) - f(\mathbf{X}) - \sum_{i=1}^{p} \ln(1+\lambda_i)}{\|\mathbf{H}\|_F} \\
=& \lim_{\mathbf{H}\to\mathbf{0}} \frac{f(\mathbf{X}+\mathbf{H}) - f(\mathbf{X}) - \sum_{i=1}^{p} \left(\lambda_i - \frac{\lambda_i^2}{2} + \frac{\lambda_i^3}{3} - \ldots\right)}{\|\mathbf{H}\|_F} \\
=& \lim_{\mathbf{H}\to\mathbf{0}} \frac{f(\mathbf{X}+\mathbf{H}) - f(\mathbf{X}) - \sum_{i=1}^{p} \lambda_i}{\|\mathbf{H}\|_F} \\
=& \lim_{\mathbf{H}\to\mathbf{0}} \frac{f(\mathbf{X}+\mathbf{H}) - f(\mathbf{X}) - \mathrm{tr}(\mathbf{X}^{-1/2}\mathbf{H}\mathbf{X}^{-1/2})}{\|\mathbf{H}\|_F} \\
=& \lim_{\mathbf{H}\to\mathbf{0}} \frac{f(\mathbf{X}+\mathbf{H}) - f(\mathbf{X}) - \mathrm{tr}(\mathbf{X}^{-1}\mathbf{H})}{\|\mathbf{H}\|_F},
\end{aligned}$$

which implies $\nabla f(\mathbf{X}) = \mathbf{X}^{-1}$. The calculation of the high order terms is based on

$$\begin{aligned}
&\left|\frac{\sum_{i=1}^{p}\sum_{k=2}^{\infty}\frac{(-1)^k\lambda_i^k}{k}}{\left\|\mathbf{X}^{-1/2}\mathbf{H}\mathbf{X}^{-1/2}\right\|_F}\right| = \left|\frac{\sum_{i=1}^{p}\sum_{k=2}^{\infty}\frac{(-1)^k\lambda_i^k}{k}}{\sqrt{\sum_{i=1}^{p}\lambda_i^2}}\right| \\
\leq& \frac{\sum_{i=1}^{p}\sum_{k=2}^{\infty}\lambda_i^k}{\sqrt{\sum_{i=1}^{p}\lambda_i^2}} \leq \frac{p\sum_{k=2}^{\infty}\lambda_1^k}{\lambda_1} \\
=& p\lambda_1\sum_{k=0}^{\infty}\lambda_1^k = \frac{p\lambda_1}{1-\lambda_1}
\end{aligned}$$

and

$$\begin{aligned}
&\left\|\mathbf{X}^{-1/2}\mathbf{H}\mathbf{X}^{-1/2}\right\|_F \leq \left\|\mathbf{X}^{-1/2}\right\|_F \left\|\mathbf{H}\right\|_F \left\|\mathbf{X}^{-1/2}\right\|_F \\
\Longrightarrow& \frac{1}{\left\|\mathbf{H}\right\|_F} \leq \frac{\left\|\mathbf{X}^{-1/2}\right\|_F^2}{\left\|\mathbf{X}^{-1/2}\mathbf{H}\mathbf{X}^{-1/2}\right\|_F}.
\end{aligned}$$

□

**The Chain Rule** Let $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ with composite structure as

$$f(\mathbf{W}) = g(\mathbf{C}(\mathbf{W}))$$

such that $g : \mathbb{R}^{m\times n} \to \mathbb{R}$ and $\mathbf{C} : \mathbb{R}^{p\times q} \to \mathbb{R}^{m\times n}$. We can construct the chain rule as follows

$$\begin{aligned}
&\frac{\partial f(\mathbf{W})}{\partial w_{ij}} = \frac{\partial g(\mathbf{C}(\mathbf{W}))}{\partial w_{ij}} \\
=& \sum_{k=1}^{m}\sum_{l=1}^{n} \frac{\partial g(\mathbf{C})}{\partial c_{kl}} \frac{\partial c_{kl}(\mathbf{W})}{\partial w_{ij}} \\
=& \sum_{k=1}^{m}\sum_{l=1}^{n} \left(\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}}\right)_{kl} \left(\frac{\partial \mathbf{C}(\mathbf{W})}{\partial w_{ij}}\right)_{kl} \\
=& \mathrm{tr}\left(\left(\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}}\right)^{\top}\left(\frac{\partial \mathbf{C}(\mathbf{W})}{\partial w_{ij}}\right)\right) \\
=& \frac{\mathrm{tr}\left(\left(\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}}\right)^{\top}\partial \mathbf{C}(\mathbf{W})\right)}{\partial w_{ij}}.
\end{aligned}$$

Hence, we have

$$\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} = \frac{\operatorname{tr}\left(\left(\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}}\right)^\top \partial \mathbf{C}(\mathbf{W})\right)}{\partial \mathbf{W}}.$$

Note that we write $\partial$ before $\mathbf{C}(\mathbf{W})$ (rather than before trace), which means we take derivative on $\mathbf{W}$ by regarding $\partial g(\mathbf{C})/\partial \mathbf{C}$ is fixed.

**Example 1.3.** *We let $\sigma > 0$ be some constant and define $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ as follows*

$$f(\mathbf{W}, \sigma^2) = \ln\det(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}).$$

*We denote*

$$\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} \qquad \textit{and} \qquad g(\mathbf{C}) = \ln\det(\mathbf{C}).$$

*For the term of logarithmic determinant, we have*

$$\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}} = \mathbf{C}^{-1},$$

*and the chain rule implies*

$$\begin{aligned}\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} =& \frac{\operatorname{tr}\left(\left(\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}}\right)^\top \partial(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})\right)}{\partial \mathbf{W}}\\ =& \frac{\operatorname{tr}\big(\mathbf{C}^{-1}\partial(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})\big)}{\partial \mathbf{W}}\\ =& 2\mathbf{C}^{-1}\mathbf{W}\\ =& 2(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{W}.\end{aligned}$$

*The last second equality is because of (for fixed $\mathbf{C}$)*

$$\mathrm{d}(\mathbf{C}^{-1}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})) = \mathbf{C}^{-1}\mathrm{d}(\mathbf{W}\mathbf{W}^\top) = \mathbf{C}^{-1}\big(\mathbf{W}\cdot\mathrm{d}\mathbf{W}^\top + (\mathrm{d}\mathbf{W})\cdot\mathbf{W}^\top\big)$$

*and*

$$\begin{aligned}&\operatorname{tr}\big(\mathrm{d}(\mathbf{C}^{-1}(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}))\big)\\ =&\operatorname{tr}\big(\mathbf{C}^{-1}\big(\mathbf{W}\cdot\mathrm{d}\mathbf{W}^\top + (\mathrm{d}\mathbf{W})\cdot\mathbf{W}^\top\big)\big)\\ =&\operatorname{tr}\big(\mathbf{C}^{-1}\mathbf{W}\cdot\mathrm{d}\mathbf{W}^\top\big) + \operatorname{tr}\big(\mathbf{C}^{-1}\cdot\mathrm{d}\mathbf{W}\cdot\mathbf{W}^\top\big)\\ =&\operatorname{tr}\big(\mathrm{d}\mathbf{W}\cdot\mathbf{W}^\top\mathbf{C}^{-1}\big) + \operatorname{tr}\big(\mathbf{W}^\top\mathbf{C}^{-1}\cdot\mathrm{d}\mathbf{W}\big)\\ =&\operatorname{tr}\big(2\mathbf{W}^\top\mathbf{C}^{-1}\cdot\mathrm{d}\mathbf{W}\big).\end{aligned}$$

**Example 1.4.** *We consider the dataset $\{(\mathbf{a}_i, b_i)\}_{i=1}^n$, where $\mathbf{a}_i \in \mathbb{R}^d$ and $b_i \in \{1, -1\}$. The logistic regression has the objective function*

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \ln(1 + \exp(-b_i\mathbf{a}_i^\top\mathbf{x})).$$

*has gradient*

$$\nabla f(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^n \frac{b_i\mathbf{a}_i}{1 + \exp(b_i\mathbf{a}_i^\top\mathbf{x})}.$$

*Proof.* Let

$$g(z) = \ln(1+\exp(-z))) \qquad \text{and} \qquad f_i(\mathbf{x}) = g(b_i\mathbf{a}_i^\top \mathbf{x}) = \ln(1+\exp(-b_i\mathbf{a}_i^\top \mathbf{x})).$$

We have

$$g'(z) = \frac{-\exp(-z)}{1+\exp(-z)} = -\frac{1}{1+\exp(z)}.$$

We write $z_i = z_i(\mathbf{x}) = b_i\mathbf{a}_i^\top \mathbf{x}$, then

$$f_i(\mathbf{x}) = g(z_i(\mathbf{x})) \qquad \text{and} \qquad \frac{\partial z_i(\mathbf{x})}{\partial \mathbf{x}} = b_i\mathbf{a}_i.$$

Based on the chain rule, we have

$$\begin{aligned}\frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}} =& \frac{\operatorname{tr}\left(\left(\frac{\partial g(z_i)}{\partial z_i}\right)^\top \partial(b_i\mathbf{a}_i^\top \mathbf{x})\right)}{\partial \mathbf{x}} \\ =& \frac{\left(-\frac{1}{1+\exp(z_i)}\right)\partial(b_i\mathbf{a}_i^\top \mathbf{x})}{\partial \mathbf{x}} \\ =& -\frac{b_i\mathbf{a}_i}{1+\exp(b_i\mathbf{a}^\top \mathbf{x})}.\end{aligned}$$

The gradient of $l(\mathbf{x})$ is achieved by taking the average. □

**Example 1.5.** *We consider the network with one hidden layer. We have dataset $\{(\mathbf{a}_i, \mathbf{b}_i)\}_{i=1}^n$, where $\mathbf{a}_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}^q$. The parameters of the model is organized by*

$$\mathbf{W} = \begin{bmatrix}\mathbf{w}_1 & \dots & \mathbf{w}_m\end{bmatrix} \in \mathbb{R}^{d\times m}$$

*For the input $\mathbf{a} \in \mathbb{R}^d$ and the output $\mathbf{b} \in \mathbb{R}^m$, we define $\mathbf{h} : \mathbb{R}^{d\times m} \to \mathbb{R}^m$ and $\mathbf{l} : \mathbb{R}^m \to \mathbb{R}^m$ as*

$$\mathbf{h}(\mathbf{W}) = \begin{bmatrix}h(\mathbf{w}_1^\top \mathbf{a}) \\ \vdots \\ h(\mathbf{w}_m^\top \mathbf{a})\end{bmatrix} \in \mathbb{R}^m \qquad \text{and} \qquad \mathbf{l}(\mathbf{h}) = \begin{bmatrix}l_1(h_1) \\ \vdots \\ l_m(h_m)\end{bmatrix} \in \mathbb{R}^m,$$

*where $\sigma : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are the active function and the loss function, e.g., $h(z) = 1/(1+\exp(-z))$ and $l_i(h_i) = \frac{1}{2}(h_i - b_i)^2$, which leads to the component loss*

$$f(\mathbf{W}) = \frac{1}{2}\left\|\boldsymbol{\sigma}(\mathbf{W}^\top \mathbf{a}) - \mathbf{b}\right\|_2^2,$$

*where $\boldsymbol{\sigma} : \mathbb{R}^m \to \mathbb{R}^m$ is defined as*

$$\boldsymbol{\sigma}(\mathbf{z}) = \begin{bmatrix}\sigma(z_1) \\ \vdots \\ \sigma(z_m)\end{bmatrix} \in \mathbb{R}^m \quad \textit{with} \quad \sigma(z) = \frac{1}{1+\exp(-z)}.$$

*We have*

$$\sigma'(z) = \frac{\exp(-z)}{(1+\exp(-z))^2} = \frac{\exp(-z)}{1+\exp(-z)} \cdot \frac{1}{1+\exp(-z)} = \sigma(z)(1-\sigma(z)).$$

*Following the chain rule, we let*

$$f(\mathbf{W}) = g(\boldsymbol{\sigma}(\mathbf{W})) \quad \textit{with} \quad g(\boldsymbol{\sigma}) = \frac{1}{2}\|\boldsymbol{\sigma} - \mathbf{b}\|_2^2 \quad \textit{and} \quad \boldsymbol{\sigma} = \begin{bmatrix}\sigma_1 \\ \vdots \\ \sigma_m\end{bmatrix} = \begin{bmatrix}\sigma(\mathbf{w}_1^\top \mathbf{a}) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{a})\end{bmatrix} \in \mathbb{R}^m.$$

*Then we have*

$$\frac{\partial g(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} = \boldsymbol{\sigma} - \mathbf{b} \in \mathbb{R}^m$$

*and*

$$\begin{aligned}
&\frac{\partial f(\mathbf{W})}{\partial \mathbf{w}_k} = \frac{\operatorname{tr}((\boldsymbol{\sigma} - \mathbf{b})^\top \partial \boldsymbol{\sigma}(\mathbf{W}))}{\partial \mathbf{w}_k} \\
=&\frac{\partial \sum_{j=1}^m (\sigma_j - b_j)\sigma(\mathbf{w}_j^\top \mathbf{a})}{\partial \mathbf{w}_k} \\
=&\frac{(\sigma_k - b_k)\sigma(\mathbf{w}_k^\top \mathbf{a})}{\partial \mathbf{w}_k} \\
=&(\sigma(\mathbf{w}_k^\top \mathbf{a}) - b_k)\sigma(\mathbf{w}_k^\top \mathbf{a})(1 - \sigma(\mathbf{w}_k^\top \mathbf{a}))\mathbf{a} \in \mathbb{R}^d.
\end{aligned}$$

*We can write*

$$\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} = \mathbf{a}\big((\boldsymbol{\sigma}(\mathbf{W}^\top \mathbf{a}) - \mathbf{b}) \circ \boldsymbol{\sigma}(\mathbf{W}^\top \mathbf{a}) \circ (\mathbf{1} - \boldsymbol{\sigma}(\mathbf{W}^\top \mathbf{a}))\big)^\top \in \mathbb{R}^{d \times m}.$$

**Example 1.6.** *For logistic regression, we have*

$$f_i(\mathbf{x}) = \ln(1 + \exp(-b_i \mathbf{a}_i^\top \mathbf{x})) \qquad \textit{and} \qquad \frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}} = -\frac{b_i \mathbf{a}_i}{1 + \exp(b_i \mathbf{a}_i^\top \mathbf{x})}.$$

*Let $z_i = b_i \mathbf{a}_i^\top \mathbf{x}$, it holds*

$$\frac{\partial f_i(\mathbf{x})}{\partial x_j} = -\frac{b_i a_{ij}}{1 + \exp(b_i \mathbf{a}_i^\top \mathbf{x})}$$

*and*

$$\begin{aligned}
\frac{\partial f_i(\mathbf{x})}{\partial x_j \partial x_k} =& - b_i a_{ij} \cdot \frac{\partial}{\partial z_i} \frac{1}{1 + \exp(z_i)} \cdot \frac{\partial z_i}{\partial x_k} \\
=& - b_i a_{ij} \cdot \frac{-\exp(z_i)}{(1 + \exp(z_i))^2} \cdot b_i a_{ik} \\
=& \frac{\exp(z_i)}{(1 + \exp(z_i))^2} \cdot a_{ij} a_{ik}.
\end{aligned}$$

*Therefore, we have*

$$\nabla^2 f_i(\mathbf{x}) = \frac{\exp(b_i \mathbf{a}_i^\top \mathbf{x})}{(1 + \exp(b_i \mathbf{a}_i^\top \mathbf{x}))^2} \cdot \mathbf{a}_i \mathbf{a}_i^\top \qquad \textit{and} \qquad \nabla^2 f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{\exp(b_i \mathbf{a}_i^\top \mathbf{x})}{(1 + \exp(b_i \mathbf{a}_i^\top \mathbf{x}))^2} \cdot \mathbf{a}_i \mathbf{a}_i^\top.$$

*For implementation, we prefer to write*

$$\begin{aligned}
\nabla^2 f(\mathbf{x}) =& \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + \exp(-b_i \mathbf{a}_i^\top \mathbf{x})} \cdot \left(1 - \frac{1}{1 + \exp(-b_i \mathbf{a}_i^\top \mathbf{x})}\right) \mathbf{a}_i \mathbf{a}_i^\top \\
=& \frac{1}{n} \sum_{i=1}^n \sigma(-b_i \mathbf{a}_i^\top \mathbf{x})(1 - \sigma(-b_i \mathbf{a}_i^\top \mathbf{x})) \mathbf{a}_i \mathbf{a}_i^\top.
\end{aligned}$$

*Since it holds $\sigma(z) \in (0, 1)$ for any $z \in \mathbb{R}$, the Hessian is positive definite.*

# 2 Introduction and Topology

The examples of different types of sets

- open sets: $\{x : a < x < b\}$, $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\|_2 < 1\}$ and $\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} > \mathbf{0}\}$.
- close sets: $\{x : a \leq x \leq b\}$, $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\|_2 \leq 1\}$ and $\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} \geq \mathbf{0}\}$
- bounded sets: $\{x : a \leq x < b\}$, $\{\mathbf{x} : \|\mathbf{x} - \mathbf{a}\|_2 < 1\}$ and $\{\mathbf{x} : \mathbf{1} > \mathbf{x} \geq \mathbf{0}\}$.

**Example 2.1.** *Let $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\|_2 < 1\}$, then we have $\mathcal{C}^\circ = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\|_2 < 1\}$,*

$$\begin{aligned}\overline{\mathcal{C}} = \mathbb{R}^d \backslash (\mathbb{R}^n \backslash \mathcal{C})^\circ =& \mathbb{R}^n \backslash (\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\|_2 \geq 1\})^\circ \\ =& \mathbb{R}^n \backslash (\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\|_2 > 1\})^\circ \\ =& \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\|_2 \leq 1\}\end{aligned}$$

*and*

$$\begin{aligned}\overline{\mathcal{C}} \backslash \mathcal{C}^\circ =& \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\|_2 \leq 1\} \backslash \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\|_2 < 1\} \\ =& \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\|_2 = 1\}.\end{aligned}$$

**Example 2.2** (PD matrix)**.** *The positive-definite matrices on $\mathbb{R}^{d \times d}$ with spectral norm distance is an open set. That is, the set*

$$\mathbb{S}_{++}^d = \{\mathbf{A} \in \mathbb{R}^{d \times d} : \mathbf{A} \succ \mathbf{0}\}$$

*is open.*

*We need to prove that for any $\mathbf{A} \in \mathbb{S}_{++}^d$, there exists $\delta > 0$ such that*

$$\left\{\mathbf{B} \in \mathbb{R}^{d \times d} : \|\mathbf{A} - \mathbf{B}\|_2 \leq \delta\right\} \subseteq \mathbb{S}_{++}^d.$$

*Let $\mathbf{B} \in \mathbb{R}^{d \times d}$ satisfy $\|\mathbf{A} - \mathbf{B}\|_2 \leq \delta$ for some $\delta > 0$, then for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$|\mathbf{x}^\top (\mathbf{A} - \mathbf{B})\mathbf{x}| \leq \|\mathbf{x}\|_2 \cdot \|(\mathbf{A} - \mathbf{B})\mathbf{x}\|_2 \leq \|\mathbf{x}\|_2 \cdot \|\mathbf{A} - \mathbf{B}\|_2 \cdot \|\mathbf{x}\|_2 \leq \delta \|\mathbf{x}\|_2^2$$

*which implies*

$$-\delta \|\mathbf{x}\|_2^2 \leq \mathbf{x}^\top (\mathbf{A} - \mathbf{B})\mathbf{x} \leq \delta \|\mathbf{x}\|_2^2 .$$

*Hence, we it holds that*

$$\mathbf{x}^\top \mathbf{B}\mathbf{x} \geq \mathbf{x}^\top \mathbf{A}\mathbf{x} - \delta \|\mathbf{x}\|_2^2 \geq (\sigma_{\min}(\mathbf{A}) - \delta) \|\mathbf{x}\|_2^2$$

*Taking $\delta = \sigma_{\min}(\mathbf{A})/2$ guarantees*

$$\mathbf{x}^\top \mathbf{B}\mathbf{x} \geq \frac{\sigma_{\min}(\mathbf{A})}{2} \|\mathbf{x}\|_2^2 > 0$$

*for any non-zero $\mathbf{x} \in \mathbb{R}^d$, which implies $\mathbf{B} \in \mathbb{S}_{++}^d$.*

**Remark 2.1.** *We can show $\mathbb{S}_+^n = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X} \succeq \mathbf{0}\}$ is closed and $\{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{I} \succeq \mathbf{X} \succeq \mathbf{0}\}$ is compact.*

**Example 2.3.** *We can verify the convergence of some sequances as follows:*

- *The sequence* $\{1/k^2\}$ *converges to 0 sublinearly, since we have*

$$\lim_{k\to+\infty} \frac{1/(k+1)^2}{1/k^2} = \lim_{k\to+\infty} \frac{k^2}{(k+1)^2} = 1.$$

- *The sequences* $\{10^{-k}\}$, $\{0.999^k\}$ *converges to 0 linearly, since we have*

$$\lim_{k\to+\infty} \frac{10^{-(k+1)}}{10^{-k}} = 0.1 \quad \textit{and} \quad \lim_{k\to+\infty} \frac{0.999^{(k+1)}}{0.999^k} = 0.999.$$

- *The sequence* $\{0.9^{k(k+1)}\}$ *converges to 0 superlinearly, since we have*

$$\lim_{k\to+\infty} \frac{0.9^{(k+1)(k+2)}}{0.9^{k(k+1)}} = \lim_{k\to+\infty} 0.9^{2(k+1)} = 0$$

- *If* $\{x_k\}$ *holds that* $x_{k+1} = x_k^2$*, the quadratic convergence does not hold for any* $x_0 \in \mathbb{R}$*.*

*Let* $\epsilon > 0$ *be the accuracy and the sequence is generated by some iterative algorithm.*

- *For* $1/k^2 \le \epsilon$*, we require* $k \ge 1/\sqrt{\varepsilon}$*.*
- *For* $10^{-k} = (1-0.9)^k \le \epsilon$*, we require* $k \ge (10/9)\ln(1/\epsilon)$*.*
- *For* $0.999^k = (1-10^{-3})^k \le \epsilon$*, we require* $k \ge 10^3 \ln(1/\epsilon)$*.*
- *For* $0.9^{k(k+1)} \le \epsilon$*, we require* $k(k+1) \ge 10\ln(1/\epsilon)$*, and* $k \ge \sqrt{10\ln(1/\epsilon)}$ *is enough.*
- *For* $x_{k+1} = x_k^2$*, we have*

$$x_1 = x_0^2, \quad x_2 = x_1^2 = x_0^4, \quad x_3 = x_2^2 = x_0^8, \quad \ldots \quad x_k = x_{k-1}^2 = x_0^{2^k}.$$

  *Let* $x_0 = 1 - 10^{-3}$*, then achieving* $x_k \le \epsilon$ *requires*

$$x_0^{2^k} = (1-10^{-3})^{2^k} \le \epsilon \quad \Longleftarrow \quad 2^k \ge 10^3 \ln(1/\epsilon) \quad \Longleftarrow \quad k \ge \frac{\ln(10^3 \ln(1/\epsilon))}{\ln 2}.$$

  *For* $\epsilon = 10^{-18}$*, setting* $k = \lceil 15.339 \rceil = 16$ *can achieve* $x_k \le \epsilon = 10^{-18}$*.*

**Remark 2.2.** *The Bernoulli's inequality says for any* $0 < z < 1$*, we have*

$$\exp(z) = \sum_{k=0}^{+\infty} \frac{z^k}{k!} \le \sum_{k=0}^{+\infty} z^k = \frac{1}{1-z}.$$

*We consider* $x_{t+1} = (1-1/\kappa)x_t$ *for some* $x_0 > 0$*, which leads to*

$$x_t \le \left(1 - \frac{1}{\kappa}\right)^t x_0.$$

*Let* $z = 1/\kappa$ *for some* $\kappa \gg 1$*, then we have (the equality nearly holds)*

$$\exp\left(\frac{1}{\kappa}\right) \le \frac{1}{1-1/\kappa} = \frac{\kappa}{\kappa-1} \implies 1 - \frac{1}{\kappa} = \frac{\kappa-1}{\kappa} \le \exp\left(-\frac{1}{\kappa}\right) \implies x_t = \left(1-\frac{1}{\kappa}\right)^t x_0 \le x_0 \exp\left(-\frac{t}{\kappa}\right).$$

*For* $x_t \le \epsilon$*, it is enough to let*

$$x_0 \exp\left(-\frac{t}{\kappa}\right) \le \epsilon \quad \Longleftarrow \quad \frac{x_0}{\epsilon} \le \exp\left(\frac{t}{\kappa}\right) \quad \Longleftarrow \quad t \ge \kappa \ln\left(\frac{x_0}{\epsilon}\right).$$

**Example 2.4.** *Consider the sequence* $\{x_k\}$ *with*

$$x_k = 2^{-\lceil k/2 \rceil},$$

*which converges to* $x^* = 0$ *linearly. For even* $k$*, we have*

$$\frac{|x_{k+1} - x^*|}{|x_k - x^*|} = \frac{2^{-(k+2)/2}}{2^{-k/2}} = \frac{1}{2}.$$

*For odd* $k$*, we have*

$$\frac{|x_{k+1} - x^*|}{|x_k - x^*|} = \frac{2^{-(k+1)/2}}{2^{-(k+1)/2}} = 1.$$

*Obviously, the sequence* $1, 1/2, 1, 1/2 \ldots$ *does not converge.*

Suppose that the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$. The sequence is said to converge R-linearly to $\mathbf{x}^*$ if there exists a sequence $\{\epsilon_k\}$ such that

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \epsilon_k$$

for all $k$ and $\{\epsilon_k\}$ converges Q-linearly to zero.

**Example 2.5.** *Let*

$$x_k = 2^{-\lceil k/2 \rceil} + 1,$$

*which converges to* $x^* = 1$*. We have*

$$|x_k - x^*| = 2^{-\lceil k/2 \rceil} \leq 2^{-k/2} \triangleq \epsilon_k.$$

*We can verify*

$$\lim_{k \to \infty} \frac{\epsilon_{k+1}}{\epsilon_k} = 2^{-1/2} < 1.$$

*Hence, the sequence* $\{\epsilon_k\}$ *Q-linearly converges to 0, and the sequence* $\{x_k\}$ *R-linearly converges to 1.*



# 3 Convex Analysis

**Theorem 3.1.** *Let* $\mathcal{C}_\theta$ *be convex sets indexed by* $\theta$*, then* $\mathcal{C} = \bigcap_\theta \mathcal{C}_\theta$ *is a convex set.*

*Proof.* Since any $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{C}$ also belongs to $\mathcal{C}_\theta$ for each $\theta$, we have

$$\alpha \mathbf{x} + (1 - \alpha)\mathbf{y} \in \mathcal{C}_\theta \subseteq \mathcal{C}$$

for any $\alpha \in [0, 1]$. Hence, we have proved the set $\mathcal{C}$ is convex. □

**Theorem 3.2.** *The projection* $\mathrm{proj}_{\mathcal{C}}(\mathbf{y})$ *for* $\mathbf{x} \in \mathbb{R}^d$ *on* $\mathcal{C}$ *is uniquely defined for nonempty, closed and convex set* $\mathcal{C} \subseteq \mathbb{R}^d$*.*

*Proof.* If $\mathbf{y} \in \mathcal{C}$, it is clear that $\mathbf{y} = \mathrm{proj}_{\mathcal{C}}(\mathbf{y})$ finish the proof. Now we focus on the case of $\mathbf{y} \notin \mathcal{C}$.
We first consider the existence. We define

$$f(\mathbf{y}, \mathcal{C}) = \inf_{\mathbf{y} \in \mathcal{C}} \|\mathbf{y} - \mathbf{x}\|_2 .$$

This definition of infimum means for any $\epsilon_k > 0$, there exists $\mathbf{w}_k \in \mathcal{C}$ such that

$$f(\mathbf{y}, \mathcal{C}) \leq \|\mathbf{y} - \mathbf{w}_k\|_2 < f(\mathbf{y}, \mathcal{C}) + \epsilon_k.$$

Let $\epsilon_k = 1/k$, then the sequence $\{\mathbf{w}_k\}$ is bounded. Then there exists subsequence $\{\mathbf{w}_{k_j}\}$ which convergence to some point $\mathbf{w} \in \mathbb{R}^d$. Since the set $\mathcal{C}$ is close, we have $\mathbf{w} \in \mathcal{C}$. Taking $k \to +\infty$, we achieve $f(\mathbf{y}, \mathcal{C}) = \|\mathbf{y} - \mathbf{w}\|_2$ and such $\mathbf{w} \in \mathcal{C}$ is just $\text{proj}_{\mathcal{C}}(\mathbf{y})$.

We then consider the uniqueness. We assume there exist $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$ such that

$$\mathbf{x}_1 \neq \mathbf{x}_2 \qquad \text{and} \qquad \|\mathbf{y} - \mathbf{x}_1\|_2^2 = \|\mathbf{y} - \mathbf{x}_2\|_2^2 = f(\mathbf{y}, \mathcal{C}).$$

The assumption $\mathbf{x}_1 \neq \mathbf{x}_2$ implies

$$\begin{aligned}
&\left\|\mathbf{y} - \frac{\mathbf{x}_1 + \mathbf{x}_2}{2}\right\|_2^2 - \frac{1}{2}\|\mathbf{y} - \mathbf{x}_1\|_2^2 - \frac{1}{2}\|\mathbf{y} - \mathbf{x}_2\|_2^2 \\
=& \|\mathbf{y}\|_2^2 - \langle \mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}\rangle + \frac{1}{4}\|\mathbf{x}_1 + \mathbf{x}_2\|_2^2 - \frac{1}{2}\|\mathbf{y}\|_2^2 + \langle \mathbf{y}, \mathbf{x}_1\rangle - \frac{1}{2}\|\mathbf{x}_1\|_2^2 - \frac{1}{2}\|\mathbf{y}\|_2^2 + \langle \mathbf{y}, \mathbf{x}_2\rangle - \frac{1}{2}\|\mathbf{x}_2\|_2^2 \\
=& \frac{1}{4}\|\mathbf{x}_1 + \mathbf{x}_2\|_2^2 - \frac{1}{2}\|\mathbf{x}_1\|_2^2 - \frac{1}{2}\|\mathbf{x}_2\|_2^2 \\
=& -\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2}{4} < 0.
\end{aligned}$$

Arranging above inequality leads to

$$\begin{aligned}
&2\left\|\mathbf{y} - \frac{\mathbf{x}_1 + \mathbf{x}_2}{2}\right\|_2^2 \\
<& \|\mathbf{y} - \mathbf{x}_1\|_2^2 + \|\mathbf{y} - \mathbf{x}_2\|_2^2 \\
=& 2f(\mathbf{y}, \mathcal{C}),
\end{aligned}$$

where the last step use the assumption that $\mathbf{x}_1$ and $\mathbf{x}_2$ both achieve the minimum. It says $(\mathbf{x}_1 + \mathbf{x}_2)/2 \in \mathcal{C}$ is strictly more close to $\mathbf{y}$ than $\mathbf{x}_1$ and $\mathbf{x}_2$, which leads to contradiction. Hence, the projection is unique. □



**Theorem 3.3.** *If $\mathbf{y} \notin \mathcal{C}$ for some close and convex set $\mathcal{C} \subseteq \mathbb{R}^d$, then $\mathbf{z} = \text{proj}_{\mathcal{C}}(\mathbf{y})$ lies on the boundary of $\mathcal{C}$ and the hyperplane*

$$\{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{y} - \mathbf{z}, \mathbf{x} - \mathbf{z}\rangle = 0\}$$

*separates $\mathbf{y}$ and $\mathcal{C}$ in that they lie on different sides, that is*

$$\langle \mathbf{y} - \mathbf{z}, \mathbf{y} - \mathbf{z}\rangle > 0 \quad \textit{and} \quad \langle \mathbf{y} - \mathbf{z}, \mathbf{x} - \mathbf{z}\rangle \leq 0$$

*for any $\mathbf{x} \in \mathcal{C}$. It implies*

$$\|\mathbf{x} - \mathbf{z}\|_2^2 \leq \|\mathbf{x} - \mathbf{y}\|_2^2$$

*for any $\mathbf{x} \in \mathcal{C}$.*

*Proof.* The condition means $\mathbf{y} \neq \mathbf{z}$, then $\langle \mathbf{y} - \mathbf{z}, \mathbf{y} - \mathbf{z}\rangle > 0$.

Given any $\mathbf{x} \in \mathcal{C}$, the definition of $\mathbf{z} = \text{proj}_{\mathcal{C}}(\mathbf{y})$ means $\mathbf{z} \in \mathcal{C}$. Hence, for any $\mathbf{x} \in \mathcal{C}$ and $\alpha \in (0, 1)$, we have

$$\mathbf{w} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{z} \in \mathcal{C}$$

which means

$$\begin{aligned}
\|\mathbf{y} - \mathbf{z}\|_2^2 \leq \|\mathbf{y} - \mathbf{w}\|_2^2 &= \|\mathbf{y} - (\alpha\mathbf{x} + (1 - \alpha)\mathbf{z})\|_2^2 = \|\mathbf{y} - \mathbf{z} - \alpha(\mathbf{x} - \mathbf{z})\|_2^2 \\
&= \|\mathbf{y} - \mathbf{z}\|_2^2 - 2\alpha\langle \mathbf{y} - \mathbf{z}, \mathbf{x} - \mathbf{z}\rangle + \alpha^2\|\mathbf{x} - \mathbf{z}\|_2^2,
\end{aligned}$$

where the inequality is based on $\mathbf{z} = \text{proj}_{\mathcal{C}}(\mathbf{y})$. Therefore, we have

$$2\langle \mathbf{x}-\mathbf{z}, \mathbf{y}-\mathbf{z}\rangle \leq \alpha \|\mathbf{x}-\mathbf{z}\|_2^2$$

By letting $\alpha \to 0$, we obtain the first inequality $\langle \mathbf{y}-\mathbf{z}, \mathbf{x}-\mathbf{z}\rangle < 0$.
We also have

$$\begin{aligned}
&\|\mathbf{x}-\mathbf{z}\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2 \\
=&2\langle \mathbf{x}-\mathbf{z}-(\mathbf{x}-\mathbf{y}), \mathbf{x}-\mathbf{z}+(\mathbf{x}-\mathbf{y})\rangle \\
=&2\langle \mathbf{y}-\mathbf{z}, \mathbf{x}-\mathbf{z}-(\mathbf{y}-\mathbf{x})\rangle \\
=&2\langle \mathbf{y}-\mathbf{z}, \mathbf{x}-\mathbf{z}\rangle - 2\|\mathbf{y}-\mathbf{z}\|_2^2 < 0.
\end{aligned}$$

□

**Theorem 3.4.** *A function $f(\mathbf{x})$ is convex if and only if its epigraph is a convex set.*

*Proof.* **Part I:** Suppose $f : \mathcal{C} \to \mathbb{R}$ is convex. Let $(\mathbf{x}_1, u_1)$ and $(\mathbf{x}_1, u_1)$ in

$$\text{epi}\, f \triangleq \{(\mathbf{x}, u) \in \mathcal{C} \times \mathbb{R} : f(\mathbf{x}) \leq u\}.$$

For any $\alpha \in [0, 1]$, the point

$$\alpha(\mathbf{x}_1, u_1) + (1-\alpha)(\mathbf{x}_2, u_2) = (\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2, \alpha u_1 + (1-\alpha)u_2)$$

satisfies

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2) \leq \alpha u_1 + (1-\alpha)u_2,$$

where the first inequality use the convexity of $f$ and the second one is due to $(\mathbf{x}_1, u_1)$ and $(\mathbf{x}_2, u_2)$ in epi $f$. Hence, the point

$$(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2, \alpha u_1 + (1-\alpha)u_2) = \alpha(\mathbf{x}_1, u_1) + (1-\alpha)(\mathbf{x}_2, u_2)$$

also in epi $f$, which means the epigraph is convex.
**Part II:** Suppose the epigraph

$$\text{epi}\, f \triangleq \{(\mathbf{x}, u) \in \mathcal{C} \times \mathbb{R} : f(\mathbf{x}) \leq u\}$$

is convex. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$, $u_1 = f(\mathbf{x}_1)$ and $u_2 = f(\mathbf{x}_2)$, then we have $(\mathbf{x}_1, u_1), (\mathbf{x}_2, u_2) \in \text{epi}\, f$. The convexity of epigraph means

$$\alpha(\mathbf{x}_1, u_1) + (1-\alpha)(\mathbf{x}_2, u_2) = (\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2, \alpha u_1 + (1-\alpha)u_2) \in \text{epi}\, f,$$

which leads to

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha u_1 + (1-\alpha)u_2 = \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2).$$

This mean function $f$ is convex. □

**Theorem 3.5** (supremum)**.** *If each $f_i : \mathcal{X} \to \mathbb{R}$ is convex for all $\mathbf{y} \in \mathcal{Y}$, then the function*

$$g(\mathbf{x}) = \sup_{i \in \mathcal{I}} f_i(\mathbf{x})$$

*is convex on $\mathcal{X}$, where $\mathcal{I}$ is any indicator set.*

*Proof.* For any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0,1]$, we have

$$\begin{aligned}
&g(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \\
=&\sup_{i\in\mathcal{I}} f_i(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \\
\leq&\sup_{i\in\mathcal{I}} \big(\lambda f_i(\mathbf{x}_1) + (1-\lambda) f_i(\mathbf{x}_2)\big) \\
\leq&\sup_{i\in\mathcal{I}} \lambda f_i(\mathbf{x}_1) + \sup_{i\in\mathcal{I}} (1-\lambda) f_i(\mathbf{x}_2) \\
=&\lambda \sup_{i\in\mathcal{I}} f_i(\mathbf{x}_1) + (1-\lambda) \sup_{i\in\mathcal{I}} f_i(\mathbf{x}_2) \\
=&\lambda g(\mathbf{x}_1) + (1-\lambda) g(\mathbf{x}_2),
\end{aligned}$$

where the first inequality is based on the convexity of $f_i$. □

**Example 3.1.** *We say the function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ is convex-concave if the function $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for any fixed $\mathbf{y} \in \mathbb{R}^y$ and concave in $\mathbf{y}$ for any fixed $\mathbf{x} \in \mathbb{R}^{d_x}$. We define*

$$P(\mathbf{x}) = \sup_{\mathbf{y}\in\mathbb{R}^{d_y}} f(\mathbf{x}, \mathbf{y}).$$

*In the view of Theorem 3.5 by taking $i = \mathbf{y}$ and $\mathcal{I} = \mathbb{R}^{d_y}$, we can conclude $P(\mathbf{x})$ is convex.*

**Theorem 3.6** (partial infimum)**.** *If $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex for all $(\mathbf{x}, \mathbf{y})$ in convex set $\mathcal{X} \times \mathcal{Y}$, then*

$$g(\mathbf{x}) = \inf_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$$

*is convex on $\mathcal{X}$.*

**Remark 3.1.** *There is an incorrect proof. Let $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ such that $g(\mathbf{x}_1) = f(\mathbf{x}_1, \mathbf{y}_1)$ and $g(\mathbf{x}_2) = f(\mathbf{x}_2, \mathbf{y}_2)$. Then we have*

$$\begin{aligned}
&g(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \\
=&\inf_{\mathbf{y}\in\mathcal{Y}} f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \mathbf{y}) \\
\leq&f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda \mathbf{y}_1 + (1-\lambda)\mathbf{y}_2) \\
\leq&\lambda f(\mathbf{x}_1, \mathbf{y}_2) + (1-\lambda) f(\mathbf{x}_2, \mathbf{y}_2) \\
=&\lambda g(\mathbf{x}_1) + (1-\lambda) g(\mathbf{x}_2).
\end{aligned}$$

*This analysis is problematic, since we cannot guarantee the existence of such $\mathbf{y}_1$ and $\mathbf{y}_2$.*

*Proof.* Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0,1]$. For any $\epsilon > 0$, the definition of $g$ means there exist $\mathbf{y}_1$ and $\mathbf{y}_2$ in $\mathcal{Y}$ such that

$$f(\mathbf{x}_1, \mathbf{y}_1) \leq g(\mathbf{x}_1) + \epsilon \qquad \text{and} \qquad f(\mathbf{x}_2, \mathbf{y}_2) \leq g(\mathbf{x}_2) + \epsilon. \tag{5}$$

The convexity of $f$ means

$$\begin{aligned}
&g(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \\
=&\inf_{\mathbf{y}\in\mathcal{Y}} f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \mathbf{y}) \\
\leq&f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda \mathbf{y}_1 + (1-\lambda)\mathbf{y}_2) \\
\leq&\lambda f(\mathbf{x}_1, \mathbf{y}_2) + (1-\lambda) f(\mathbf{x}_2, \mathbf{y}_2) \\
\leq&\lambda (g(\mathbf{x}_1) + \epsilon) + (1-\lambda)(g(\mathbf{x}_2) + \epsilon) \\
=&\lambda g(\mathbf{x}_1) + (1-\lambda) g(\mathbf{x}_2) + \epsilon,
\end{aligned}$$

where the first inequality is based on the definition of infimum and the convexity of $\mathcal{Y}$ that leads to

$$(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda \mathbf{y}_1 + (1-\lambda)\mathbf{y}_2) = \lambda(\mathbf{x}_1, \mathbf{y}_1) + (1-\lambda)(\mathbf{x}_2, \mathbf{y}_2) \in \mathcal{X} \times \mathcal{Y};$$

the second inequality is based on the convexity of $f$; the last inequality is based on inequality (5). Since above result holds for any $\epsilon > 0$, we have

$$g(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda g(\mathbf{x}_1) + (1-\lambda) g(\mathbf{x}_2).$$

□

**Remark 3.2.** *The composition of convex functions may not preserve the convexity. Consider that $g(x) = x^2$ and $h(y) = -y$, then $f(x) = h(g(x)) = -x^2$ is not convex.*

**Remark 3.3.** *Let $h : \mathbb{R}^m \to \mathbb{R}$ be convex define $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ for some $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\mathbf{b} \in \mathbb{R}^m$, then the function $f(\mathbf{x}) = h(g(\mathbf{x}))$ is convex. For any $\mathbf{x}_1, \mathbf{x} \in \mathbb{R}^d$ and $\lambda \in [0,1]$, we have $f(\mathbf{x}) = h(\mathbf{A}\mathbf{x} + \mathbf{b})$ and*

$$\begin{aligned}
&f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \\
=&h(\mathbf{A}(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) + \mathbf{b}) \\
=&h(\lambda(\mathbf{A}\mathbf{x}_1 + \mathbf{b}) + (1-\lambda)(\mathbf{A}\mathbf{x}_2 + \mathbf{b})) \\
\leq&\lambda h(\mathbf{A}\mathbf{x}_1 + \mathbf{b}) + (1-\lambda) h(\mathbf{A}\mathbf{x}_2 + \mathbf{b}) \\
=&\lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2).
\end{aligned}$$

**Example 3.2.** *The function*

$$f(x,y) = \begin{cases} \dfrac{x^2}{y}, & (x,y) \neq (0,0) \\ 0, & (x,y) = (0,0) \end{cases}$$

*with domain $\{(x,y) : x \in \mathbb{R}, y > 0\} \cup \{(0,0)\}$ is not continuous at $(0,0)$. We consider $\epsilon = 1$ and point $(\hat{x}, \hat{y})$ that satisfies $\hat{x}^2 = 2\hat{y}$. Then it always holds that $\hat{x}^2/\hat{y} = 2 > \epsilon$ and $(\hat{x}, \hat{y})$ can be arbitrary close to the point $(0,0)$ by taking $\hat{x} \to 0$ and $\hat{y} \to 0$. However, the minimizer of $f(x,y)$ is $(0,0)$.*

**Theorem 3.7.** *If $\mathbf{x}^*$ is a local solution of the convex problem*

$$\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x}),$$

*then it is also a global solution.*

*Proof.* Assume $\mathbf{x}^*$ is a local solution in $\mathcal{B}_\delta(\mathbf{x}^*)$ for some $\delta > 0$. Given any $\mathbf{x} \in \mathcal{C}$, we consider

$$\hat{\mathbf{x}} = \alpha \mathbf{x} + (1-\alpha)\mathbf{x}^* \in \mathcal{C}.$$

There is a sufficiently small $\alpha > 0$ such that $\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2 \leq \delta$. The local optimality of $\mathbf{x}^*$ implies that

$$f(\mathbf{x}^*) \leq f(\hat{\mathbf{x}}) = f(\alpha \mathbf{x} + (1-\alpha)\mathbf{x}^*) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{x}^*)$$

This implies that $f(\mathbf{x}^*) \leq f(\mathbf{x})$. □

**Theorem 3.8.** *If a function $f$ is differentiable on open set $\mathcal{C}$, then it is convex on $\mathcal{C}$ if and only if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

*hols for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$.*

*Proof.* **Part I:** If $f$ is convex on $\mathcal{C}$, then

$$f(\lambda \mathbf{y} + (1-\lambda)\mathbf{x}) \leq \lambda f(\mathbf{y}) + (1-\lambda) f(\mathbf{x})$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and $\lambda \in [0,1]$. Rewrite the inequality leads to

$$f(\mathbf{x} + \lambda(\mathbf{y}-\mathbf{x})) \leq \lambda(f(\mathbf{y}) - f(\mathbf{x})) + f(\mathbf{x}) \quad \Longrightarrow \quad f(\mathbf{y}) - f(\mathbf{x}) \geq \frac{f(\mathbf{x} + \lambda(\mathbf{y}-\mathbf{x})) - f(\mathbf{x})}{\lambda}.$$

Taking $\lambda \to 0^+$, we achieve $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$.

**Part II:** For any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and $\lambda \in [0,1]$, we let $\mathbf{z} = \lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in \mathcal{C}$. If the first-order condition holds, then we have

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \quad \text{and} \quad f(\mathbf{y}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle.$$

Multiplying the first on by $\lambda$, the second one by $(1-\lambda)$ and adding, we get

$$\begin{aligned} &\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \\ \geq & f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \lambda \mathbf{x} + (1-\lambda)\mathbf{y} - \mathbf{z} \rangle \\ = & f(\mathbf{z}) = f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}). \end{aligned}$$

□

**Remark 3.4.** *In the proof of part one, we use the fact*

$$\lim_{\lambda \to 0} \frac{f(\mathbf{x} + \lambda \mathbf{h}) - f(\mathbf{x})}{\lambda} = \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle,$$

*where* $\mathbf{h} = \mathbf{y} - \mathbf{x} = [h_1, \ldots, h_d]^\top$. *We can verify this result by construct*

$$g(\lambda) = f(\mathbf{x} + \lambda \mathbf{h}),$$

*which means*

$$g'(0) = \lim_{\lambda \to 0} \frac{g(0+\lambda) - g(0)}{\lambda} = \lim_{\lambda \to 0} \frac{f(\mathbf{x} + \lambda \mathbf{h}) - f(\mathbf{x})}{\lambda}.$$

*Let* $\mathbf{y} = \mathbf{y}(\lambda) = \mathbf{x} + \lambda \mathbf{h}$, *then we have* $g(\lambda) = f(\mathbf{y}(\lambda))$ *and the chain rule implies*

$$\begin{aligned} g'(\lambda) = & \frac{\langle \nabla f(\mathbf{y}), \partial \mathbf{y}(\lambda) \rangle}{\partial \lambda} \\ = & \frac{\partial}{\partial \lambda} \sum_{i=1}^{d} \frac{\partial f(\mathbf{y})}{\partial y_i} \cdot (x_i + \lambda h_i) \\ = & \sum_{i=1}^{d} \frac{\partial f(\mathbf{y})}{\partial y_i} \cdot h_i \\ = & \langle \nabla f(\mathbf{y}), \mathbf{h} \rangle \\ = & \langle \nabla f(\mathbf{x} + \lambda \mathbf{h}), \mathbf{h} \rangle. \end{aligned}$$

*Hence, we have* $g'(0) = \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle$.

**Theorem 3.9.** *The subdifferential of* $f(\mathbf{x}) = \|\mathbf{x}\|$ *defined on* $\mathbb{R}^d$ *holds that* $\partial f(\mathbf{0}) = \{\mathbf{g} \in \mathbb{R}^d : \|\mathbf{g}\|_* \leq 1\}$.

*Proof.* The definition of subdifferential means

$$\begin{aligned} \partial f(\mathbf{0}) = & \{\mathbf{g} \in \mathbb{R}^d : \|\mathbf{y}\| \geq \|\mathbf{0}\| + \langle \mathbf{g}, \mathbf{y} - \mathbf{0} \rangle \text{ for all } \mathbf{y} \in \mathbb{R}^d\} \\ = & \{\mathbf{g} \in \mathbb{R}^d : \|\mathbf{y}\| \geq \langle \mathbf{g}, \mathbf{y} \rangle \text{ for all } \mathbf{y} \in \mathbb{R}^d\}. \end{aligned}$$

For any $\mathbf{g}_0 \in \{\mathbf{g} \in \mathbb{R}^d : \|\mathbf{g}\|_* \leq 1\}$ and $\mathbf{y} \in \mathbb{R}^d$, we have

$$\langle \mathbf{g}_0, \mathbf{y} \rangle \leq \|\mathbf{g}_0\|_* \|\mathbf{y}\| = \|\mathbf{y}\|,$$

which implies $\mathbf{g}_0 \in \partial f(\mathbf{0})$.
For any nonzero $\mathbf{g}_0 \in \partial f(\mathbf{0}) = \{\mathbf{g} \in \mathbb{R}^d : \|\mathbf{y}\| \geq \langle \mathbf{g}, \mathbf{y} \rangle \text{ for all } \mathbf{y} \in \mathbb{R}^d\}$, we have

$$\|\mathbf{y}\| \geq \langle \mathbf{g}_0, \mathbf{y} \rangle \qquad \Longleftrightarrow \qquad 0 \geq \langle \mathbf{g}_0, \mathbf{y} \rangle - \|\mathbf{y}\|$$

for any $\mathbf{y} \in \mathbb{R}^d$. Taking supreme on the constraint $\|\mathbf{y}\|_2 = 1$, we have

$$0 \geq \sup_{\|\mathbf{y}\|_2 = 1} (\langle \mathbf{g}_0, \mathbf{y} \rangle - \|\mathbf{y}\|) = \sup_{\|\mathbf{y}\|_2 = 1} (\langle \mathbf{g}_0, \mathbf{y} \rangle - 1) = \|\mathbf{g}_0\|_* - 1$$

that is $\mathbf{g}_0 \in \{\mathbf{g} \in \mathbb{R}^d : \|\mathbf{g}\|_* \leq 1\}$. □

**Remark 3.5.** *Given a norm $\|\cdot\|$ on $\mathbb{R}^d$, its dual norm $\|\cdot\|_*$ on $\mathbb{R}^d$ is defined as follows:*

$$\|\mathbf{u}\|_* = \sup_{\|\mathbf{v}\| = 1} \mathbf{u}^\top \mathbf{v}.$$

*The definition leads to inequality $\mathbf{u}^\top \mathbf{v} \leq \|\mathbf{u}\|_* \|\mathbf{v}\|$ for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ such that $\|\mathbf{v}\|_2 = 1$. For the general vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, we can let $\mathbf{u} = \mathbf{w}/\|\mathbf{w}\|_2$ (the case of $\mathbf{w} = \mathbf{0}$ is trivial), which means*

$$\begin{aligned} \mathbf{u}^\top \mathbf{v} \leq \|\mathbf{u}\|_* \|\mathbf{v}\| \quad &\Longrightarrow \quad \left(\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\right)^\top \mathbf{v} \leq \left\| \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \right\|_* \|\mathbf{v}\| \\ &\Longrightarrow \quad \mathbf{w}^\top \mathbf{v} \leq \|\mathbf{w}\|_* \|\mathbf{v}\|. \end{aligned}$$

*Some norms are commonly used in machine learning:*

1. *$\ell_p$-norm vs. $\ell_q$-norm, where $p, q \in [0, +\infty]$ with $1/p + 1/q = 1$*
2. *$\mathbf{H}$-norm vs. $\mathbf{H}^{-1}$-norm, where $\mathbf{H}$ is positive definite.*

*We consider $f(\mathbf{u}) = \|\mathbf{u}\|_1$ and desire to find its dual norm*

$$\|\mathbf{u}\|_* = \sup_{\|\mathbf{u}\|_1 = 1} \mathbf{u}^\top \mathbf{v}.$$

*We want to maximize $\sum_{i=1}^d u_i v_i$ under the constraint $\sum_{i=1}^d |v_i| = 1$. We have*

$$\sum_{i=1}^d u_i v_i \leq \sum_{i=1}^d |u_i||v_i| \leq \max_{j \in [d]} |u_j| \sum_{i=1}^d |v_i| \leq \max_{j \in [d]} |u_j| = \|\mathbf{u}\|_\infty.$$

*The subdifferential of $f(\cdot) = \|\cdot\|_1$ at $\mathbf{0}$ is*

$$\partial f(\mathbf{0}) = \{\mathbf{g} \in \mathbb{R}^d : \|\mathbf{g}\|_\infty \leq 1\}.$$

*For $d = 1$, we have*

$$\partial f(0) = \{g \in \mathbb{R} : |g| \leq 1\} = [-1, 1].$$

**Theorem 3.10.** *The subdifferential of an indicator function $\mathbb{1}_{\mathcal{C}}(\mathbf{x})$ is*

$$\partial \mathbb{1}_{\mathcal{C}}(\mathbf{x}) = \mathcal{N}_{\mathcal{C}}(\mathbf{x}),$$

*where*

$$\mathcal{N}_{\mathcal{C}}(\mathbf{x}) = \left\{\mathbf{g} \in \mathbb{R}^d : \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq 0 \text{ for all } \mathbf{y} \in \mathcal{C}\right\}$$

*is called the normal cone of $\mathcal{C} \subseteq \mathbb{R}^d$ at $\mathbf{x} \in \mathcal{C}$.*

*Proof.* For any $\mathbf{x} \in \mathcal{C}$, we require $\mathbf{g} \in \mathbb{R}^d$ holds that

$$\mathbb{1}_{\mathcal{C}}(\mathbf{y}) \geq \mathbb{1}_{\mathcal{C}}(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle$$

for any $\mathbf{y} \in \mathbb{R}^d$. We can verify it as follows:

- If $\mathbf{y} \notin \mathcal{C}$, we have $\mathbb{1}_{\mathcal{C}}(\mathbf{y}) = +\infty$ and the condition holds.
- If $\mathbf{y} \in \mathcal{C}$, we have $\mathbb{1}_{\mathcal{C}}(\mathbf{y}) = 0$ and the condition becomes $\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq 0$.

□

**Remark 3.6.** *If $\mathbf{x}$ lies in the interior of $\mathcal{N}_{\mathcal{C}}(\mathbf{x})$, there exists $\delta > 0$ such that $\mathcal{B}_\delta(\mathbf{x}) \subseteq \mathcal{C}$. We can find some $\mathbf{z} \neq \mathbf{0}$ such that $\mathbf{y}_1 = \mathbf{x} + \mathbf{z}$ and $\mathbf{y}_2 = \mathbf{x} - \mathbf{z}$ in $\mathcal{B}_\delta(\mathbf{x}) \subseteq \mathcal{C}$. Then we require subgradient $\mathbf{g}$ holds that*

$$\mathbb{1}_{\mathcal{C}}(\mathbf{y}_1) \geq \mathbb{1}_{\mathcal{C}}(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y}_1 - \mathbf{x} \rangle \Longrightarrow 0 \geq 0 + \langle \mathbf{g}, \mathbf{z} \rangle$$

*and*

$$\mathbb{1}_{\mathcal{C}}(\mathbf{y}_2) \geq \mathbb{1}_{\mathcal{C}}(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y}_2 - \mathbf{x} \rangle \Longrightarrow 0 \geq 0 + \langle \mathbf{g}, -\mathbf{z} \rangle,$$

*which implies $\mathbf{g} = \mathbf{0}$. If $\mathbf{x}$ lies in the boundary of $\mathcal{C}$ and $\mathbf{y} \in \mathcal{C}$, the vector $\mathbf{y} - \mathbf{x}$ and $\mathbf{g}$ should leads to an obtuse angle or an right angle.*

**Theorem 3.11.** *If a convex function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at $\mathbf{x} \in \mathbb{R}^d$, then*

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}.$$

*Proof.* Let $\mathbf{g} \in \partial f(\mathbf{x})$. For any $t > 0$ and $\mathbf{h} \in \mathbb{R}^d$, it holds that

$$f(\mathbf{x} + t\mathbf{h}) \geq f(\mathbf{x}) + \langle \mathbf{g}, t\mathbf{h} \rangle \implies \frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t} \geq \langle \mathbf{g}, \mathbf{h} \rangle.$$

Taking $t \to 0^+$, we have

$$\langle \nabla f(\mathbf{x}), \mathbf{h} \rangle \geq \langle \mathbf{g}, \mathbf{h} \rangle \quad \Longleftrightarrow \quad \langle \nabla f(\mathbf{x}) - \mathbf{g}, \mathbf{h} \rangle \geq 0.$$

The analysis also holds for $-\mathbf{h} \in \mathbb{R}^d$, which leads to

$$\langle \nabla f(\mathbf{x}) - \mathbf{g}, -\mathbf{h} \rangle \geq 0.$$

Hence, we achieve $\mathbf{g} = \nabla f(\mathbf{x})$. □

**Theorem 3.12.** *Let $f_1$ and $f_2$ be proper convex functions on $\mathbb{R}^d$, then*

$$\partial(f_1 + f_2)(\mathbf{x}) \supseteq \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}).$$

*Proof.* Any $\mathbf{g} \in \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x})$ can be written as

$$\mathbf{g} = \mathbf{g}_1 + \mathbf{g}_2,$$

where $\mathbf{g}_1 \in \partial f_1(\mathbf{x})$ and $\mathbf{g}_2 \in \partial f_2(\mathbf{x})$. Then we have

$$f_1(\mathbf{y}) \geq f_1(\mathbf{x}) + \langle \mathbf{g}_1, \mathbf{y} - \mathbf{x} \rangle \qquad \text{and} \qquad f_2(\mathbf{y}) \geq f_2(\mathbf{x}) + \langle \mathbf{g}_2, \mathbf{y} - \mathbf{x} \rangle$$

for any $\mathbf{y} \in \mathbb{R}^d$. Summing over these inequality leads to

$$(f_1 + f_2)(\mathbf{y}) \geq (f_1 + f_2)(\mathbf{x}) + \langle \mathbf{g}_1 + \mathbf{g}_2, \mathbf{y} - \mathbf{x} \rangle,$$

which means $\mathbf{g} = \mathbf{g}_1 + \mathbf{g}_2 \in \partial(f_1 + f_2)(\mathbf{x})$. □

**Relative Interior** The relative interior $\mathrm{ri}(\mathcal{C})$ for convex $\mathcal{C} \subseteq \mathbb{R}^d$ as

$$\begin{aligned}\mathrm{ri}(\mathcal{C}) = \{\mathbf{z} \in \mathcal{C} :\ & \text{for every } \mathbf{x} \in \mathcal{C} \text{ such that} \\ & \text{there exist a } \mu > 1 \text{ such that } (1-\mu)\mathbf{x} + \mu\mathbf{z} \in \mathcal{C}\}.\end{aligned}$$

Let $\mathbf{y} = (1-\mu)\mathbf{x} + \mu\mathbf{z} \in \mathcal{C}$ and $\lambda = 1/\mu \in (0,1)$, then $\mathbf{z} = \lambda\mathbf{y} + (1-\lambda)\mathbf{x}$. The condition means that every line segment in $\mathcal{C}$ having $\mathbf{z}$ as one endpoint can be prolonged beyond $\mathbf{z}$ without leaving $\mathcal{C}$. For example $(0,1)$ is the relative interior of $[0,1]$ in $\mathbb{R}^2$.

**Example 3.3.** *Let $\mathcal{C} = \{(x,y) \in \mathbb{R}^2 : x = 0, y \in [-1,1]\}$, then the point $(0,0)$ is a relative interior point but not a interior point.*

**Example 3.4.** *Consider the functions defined on $\mathbb{R}^2$*

$$f(\mathbf{x}) = \begin{cases} 0, & (x_1+1)^2 + x_2^2 \le 1, \\ +\infty, & \textit{otherwise}, \end{cases} \quad \textit{and} \quad g(\mathbf{x}) = \begin{cases} 0, & (x_1-1)^2 + x_2^2 \le 1, \\ +\infty, & \textit{otherwise}, \end{cases}$$

*then*

$$(f+g)(\mathbf{x}) = \begin{cases} 0, & (x_1,x_2) = (0,0), \\ +\infty, & \textit{otherwise}, \end{cases}$$

*Let $z = (0,0)$, then we have $\partial f(z) = \{(x_1,x_2) : x_1 \ge 0, x_2 = 0\}$ and $g(z) = \{(x_1,x_2) : x_1 \le 0, x_2 = 0\}$, which means*

$$\partial f(z) + \partial g(z) = \{(x_1,x_2) : x_1 \in \mathbb{R}, x_2 = 0\} \subset \partial(f+g)(z) = \mathbb{R}^2.$$

**Theorem 3.13** (Supporting Hyperplane Theorem)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set and $\mathbf{x}_0$ belongs to its boundary. Then, there exists a nonzero vector $\mathbf{w} \in \mathbb{R}^d$ such that*

$$\langle \mathbf{w}, \mathbf{x} - \mathbf{x}_0 \rangle \le \mathbf{0}$$

*for any $\mathbf{x} \in \mathcal{X}$.*

*Proof.* Since $\mathbf{x}_0$ belongs to the boundary of $\mathcal{X}$, for any $\delta_k > 0$, there exists $\mathbf{y}_k \in \mathcal{B}(\mathbf{x}_0, \delta_k)$ and $\mathbf{y}_k \notin \mathcal{X}$. Taking $\delta_k \to 0$, we obtain $\{\mathbf{y}_k\}$ such that $\mathbf{y}_k \to \mathbf{x}_0$. We construct the sequence $\{\mathbf{w}_k\}$ such that

$$\mathbf{w}_k = \frac{\mathbf{y}_k - \mathbf{z}_k}{\|\mathbf{y}_k - \mathbf{z}_k\|_2},$$

where $\mathbf{z}_k = \mathrm{proj}_{\mathcal{X}}(\mathbf{y}_k)$. Noticing that $\{\mathbf{w}_{k_l}\}$ is bounded, therefore, its subsequence $\{\mathbf{w}_{k_l}\}$ converges to some limit point $\mathbf{w} \in \mathbb{R}^d$.

The property of projection (Theorem 3.3) means

$$\langle \mathbf{y}_{k_l} - \mathbf{z}_{k_l}, \mathbf{x} - \mathbf{z}_{k_l} \rangle \le 0 \iff \langle \mathbf{w}_{k_l}, \mathbf{x} - \mathbf{z}_{k_l} \rangle \le 0 \iff \langle \mathbf{w}_{k_l}, \mathbf{x} \rangle \le \langle \mathbf{w}_{k_l}, \mathbf{z}_{k_l} \rangle$$

for any $\mathbf{x} \in \mathcal{X}$. We also have

$$\begin{aligned}\langle \mathbf{w}_{k_l}, \mathbf{z}_{k_l} \rangle =& \langle \mathbf{w}_{k_l}, \mathbf{z}_{k_l} - \mathbf{y}_{k_l} \rangle + \langle \mathbf{w}_{k_l}, \mathbf{y}_{k_l} \rangle \\ =& -\|\mathbf{z}_{k_l} - \mathbf{y}_{k_l}\|_2 + \langle \mathbf{w}_{k_l}, \mathbf{y}_{k_l} \rangle \\ \le& \langle \mathbf{w}_{k_l}, \mathbf{y}_{k_l} \rangle\end{aligned}$$

for all $k_l$. Connecting above inequalities, we have

$$\langle \mathbf{w}_{k_l}, \mathbf{x} \rangle \le \langle \mathbf{w}_{k_l}, \mathbf{y}_{k_l} \rangle$$

for all $\mathbf{x} \in \mathcal{X}$. Since $\mathbf{w}_{k_l} \to \mathbf{w}$ and $\mathbf{y}_{k_l} \to \mathbf{x}_0$, we have

$$\langle \mathbf{w}, \mathbf{x} \rangle \le \langle \mathbf{w}, \mathbf{x}_0 \rangle.$$

□

**Theorem 3.14.** *The convex function has the following properties*

1. *If any* $\mathbf{x} \in \operatorname{dom} f$ *satisfies* $\partial f(\mathbf{x}) \neq \emptyset$*, then* $f$ *is convex.*
2. *If* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex and* $\mathbf{x}$ *belongs to the interior of* $\operatorname{dom} f$*, then* $\partial f(\mathbf{x}) \neq \emptyset$*.*

*Proof.* **Part I:** Let $\mathbf{x}_1, \mathbf{x}_2 \in \operatorname{dom} f$. For any $\alpha \in [0, 1]$, we define

$$\mathbf{z} = \alpha \mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2 \in \operatorname{dom} f.$$

Then there exists $\mathbf{g} \in \partial f(\mathbf{z})$ such that

$$f(\mathbf{x}_1) \geq f(\mathbf{z}) + \langle \mathbf{g}, \mathbf{x}_1 - \mathbf{z} \rangle \qquad \text{and} \qquad f(\mathbf{x}_2) \geq f(\mathbf{z}) + \langle \mathbf{g}, \mathbf{x}_2 - \mathbf{z} \rangle.$$

Taking weighted sum on above inequalities leads to

$$\begin{aligned}
&\alpha f(\mathbf{x}_1) + (1 - \alpha) f(\mathbf{x}_2) \\
\geq &\alpha(f(\mathbf{z}) + \langle \mathbf{g}, \mathbf{x}_1 - \mathbf{z} \rangle) + (1 - \alpha)(f(\mathbf{z}) + \langle \mathbf{g}, \mathbf{x}_2 - \mathbf{z} \rangle) \\
\geq &f(\mathbf{z}) + \langle \mathbf{g}, \alpha(\mathbf{x}_1 - \mathbf{z}) + (1 - \alpha)(\mathbf{x}_2 - \mathbf{z}) \rangle \\
= &f(\mathbf{z}) + \langle \mathbf{g}, \alpha \mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2 - \mathbf{z} \rangle \\
= &f(\alpha \mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2).
\end{aligned}$$

**Part II:** Consider than $(\mathbf{x}, f(\mathbf{x}))$ is on the boundary of $\operatorname{epi} f$. The hyperplane supporting theorem (Theorem 3.13) say there exists $(\mathbf{a}, b)$ with $(\mathbf{a}, b) \neq (\mathbf{0}, 0)$ such that

$$\left\langle \begin{bmatrix} \mathbf{a} \\ b \end{bmatrix}, \begin{bmatrix} \mathbf{y} - \mathbf{x} \\ t - f(\mathbf{x}) \end{bmatrix} \right\rangle \leq 0$$

for any $(\mathbf{y}, t) \in \operatorname{epi} f$. That is

$$\langle \mathbf{a}, \mathbf{y} - \mathbf{x} \rangle + b(t - f(\mathbf{x})) \leq 0.$$

If $\mathbf{a} \neq \mathbf{0}$, we can conclude $b \leq 0$. Otherwise, let $t \to +\infty$ ($t$ can be arbitrary large for fixed $\mathbf{x}, \mathbf{y}$ and $\mathbf{a}$) leads to LHS tends to $+\infty$. Since $\mathbf{x}$ is in the interior of $\operatorname{dom} f$, we can find some $\epsilon > 0$ such that $\mathbf{x} + \epsilon \mathbf{a} \in \operatorname{dom} f$. Then taking $\mathbf{y} = \mathbf{x} + \epsilon \mathbf{a}$ which leads to

$$\epsilon \|\mathbf{a}\|_2^2 + b(t - f(\mathbf{x})) \leq 0.$$

This implies $b \neq 0$. Hence, we can say $b < 0$ and dividing by $b$ obtains

$$\left\langle \frac{\mathbf{a}}{b}, \mathbf{y} - \mathbf{x} \right\rangle + (t - f(\mathbf{x})) \geq 0 \quad \iff \quad t \geq f(\mathbf{x}) + \left\langle -\frac{\mathbf{a}}{b}, \mathbf{y} - \mathbf{x} \right\rangle.$$

Taking $t = f(\mathbf{y})$ means $\mathbf{g} = -\mathbf{a}/b$ is a subgradient at $\mathbf{x}$.

If $\mathbf{a} = \mathbf{0}$, then we have $b \neq 0$. Taking $t \to +\infty$ means $b < 0$, which implies

$$t - f(\mathbf{x}) \geq 0.$$

Hence, we the vector $\mathbf{g} = \mathbf{0}$ is a subgradient at $\mathbf{x}$. □

**Example 3.5.** *Let*

$$f(x) = -\sqrt{x}$$

*defined on* $[0, +\infty)$*. Suppose there exists* $g \in \partial f(0)$*, then we require*

$$f(y) - f(0) = -\sqrt{y} \geq \langle g, y \rangle$$

*for all* $y \geq 0$*. If* $g > 0$*, then* $y = g$ *leads to* $-\sqrt{g} \geq g^2$ *that can not hold. If* $g = 0$*, then for any* $y > 0$*, it should satisfy* $-\sqrt{y} \geq 0$*, which is also can not hold.*

Review:

- If $\mathbf{x}^*$ is a local solution of the convex problem $\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$, then it is also a global solution.
- If a function $f$ is differentiable on open set $\mathcal{C}$, then it is convex on $\mathcal{C}$ if and only if

  $$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

  hols for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$.
- We say a vector $\mathbf{g} \in \mathbb{R}^d$ is a subgradient of a proper convex function $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{x} \in \operatorname{dom} f$ if

  $$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle$$

  holds for any $\mathbf{y} \in \mathbb{R}^d$.
- The set of subgradients at $\mathbf{x} \in \operatorname{dom} f$ is called the subdifferential of $f$ at $\mathbf{x}$, defined as

  $$\partial f(\mathbf{x}) \triangleq \left\{ \mathbf{g} \in \mathbb{R}^d : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \text{ holds for any } \mathbf{y} \in \mathbb{R}^d \right\}.$$

**Theorem 3.15.** *Consider proper closed convex function $f$ and closed convex set $\mathcal{C} \subseteq (\operatorname{dom} f)^\circ$. A point $\mathbf{x}^* \in \mathcal{C}$ is a solution of convex optimization problem*

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$$

*if and only if*

$$\mathbf{0} \in \partial(f(\mathbf{x}^*) + \mathbb{1}_{\mathcal{C}}(\mathbf{x}^*)).$$

*Equivalently, there exists a subgradient $\mathbf{g}^* \in \partial f(\mathbf{x}^*)$, such that any $y \in \mathcal{C} \subseteq \mathbb{R}^d$ satisfies*

$$\langle \mathbf{g}^*, \mathbf{y} - \mathbf{x}^* \rangle \geq 0.$$

*In particular, the point $\mathbf{x}^*$ is the solution of the problem in unconstrained case if*

$$\mathbf{0} \in \partial f(\mathbf{x}^*).$$

*Proof.* **Part I:** The problem can be written as

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \mathbb{1}_{\mathcal{C}}(\mathbf{x}).$$

We show that the first statement is a direct consequence of the definition of subgradient. We have

$$\begin{aligned} &\mathbf{0} \in \partial(f + \mathbb{1}_{\mathcal{C}})(\mathbf{x}^*) \\ \Longleftrightarrow &f(\mathbf{y}) + \mathbb{1}_{\mathcal{C}}(\mathbf{y}) \geq f(\mathbf{x}^*) + \mathbb{1}_{\mathcal{C}}(\mathbf{x}^*) + \langle \mathbf{0}, \mathbf{y} - \mathbf{x}^* \rangle = f(\mathbf{x}^*) + \mathbb{1}_{\mathcal{C}}(\mathbf{x}^*) = f(\mathbf{x}^*) \text{ for any } \mathbf{y} \in \mathbb{R}^d. \end{aligned}$$

**Part II:** The second inequality follows from the subgradient calculus

$$\mathbf{0} = \mathbf{g}^* + \mathbf{g} \in \partial f(\mathbf{x}^*) + \partial \mathbb{1}_{\mathcal{C}}(\mathbf{x}^*) \subseteq \partial(f + \mathbb{1}_{\mathcal{C}})(\mathbf{x}^*),$$

for some $\mathbf{g}^* \in \partial f(\mathbf{x}^*)$ and $\mathbf{g} \in \partial \mathbb{1}_{\mathcal{C}}(\mathbf{x}^*)$. Since vector $\mathbf{g}$ is a subgradient of $\mathbb{1}_{\mathcal{C}}(\mathbf{x}^*)$, we have

$$\mathbf{0} = \mathbb{1}_{\mathcal{C}}(\mathbf{y}) - \mathbb{1}_{\mathcal{C}}(\mathbf{x}^*) \geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x}^* \rangle$$

for all $\mathbf{y} \in \mathcal{C}$. Then we have $\langle \mathbf{g}^*, \mathbf{y} - \mathbf{x}^* \rangle \geq 0$.

**Part III:** In unconstrained case, we have $\mathcal{C} = \mathbb{R}^d$ and $\mathbb{1}_{\mathcal{C}}(\mathbf{y}) = 0$ for all $\mathbf{y} \in \mathbb{R}^d$, which means

$$\mathbf{0} \in \partial f(\mathbf{x}^*).$$

□

**Example 3.6.** *Some example for optimal condition:*

1. *Let $f(x) = |x|$, then $0 \in \partial f(0) = [-1, 1]$ and $x = 0$ is the minimizer.*
2. *Let*

$$f(x) = \begin{cases} x, & x \in (0, 1], \\ 0.5x, & x \in [-1, 0], \\ -x - 1.5, & x \in (\infty, -1), \end{cases}$$

   *then $\partial f(0) = [0.5, 1]$ and $\partial f(-1) = [-1, 0.5]$. Hence $x = -1$ is the minimizer.*

3. *If we consider minimizing above $f(x)$ on $\mathcal{C} = [0, +\infty]$, then $1 \in \partial f(x^*)$ for $x^* = 0$ and we have*

$$1 \cdot (y - x^*) = y \geq 0$$

   *for any $\mathbf{y} \in \mathcal{C}$. Hence, the point $x^* = 0$ is the solution of the problem.*

**Theorem 3.16.** *If there exists some*

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x})$$

*for strongly convex function $f : \mathcal{C} \to \mathbb{R}$, then it is the unique minimizer.*

*Proof.* Suppose the point $\mathbf{y} \in \mathcal{C}$ is another minimizer such that $\mathbf{y} \neq \mathbf{x}^*$ and $f(\mathbf{x}^*) = f(\mathbf{y})$, then we have

$$\begin{aligned} & f(\alpha \mathbf{x}^* + (1 - \alpha)\mathbf{y}) \\ \leq & \alpha f(\mathbf{x}^*) + (1 - \alpha) f(\mathbf{y}) - \frac{\mu \alpha (1 - \alpha)}{2} \|\mathbf{x}^* - \mathbf{y}\|_2^2 \\ = & f(\mathbf{x}^*) - \frac{\mu \alpha (1 - \alpha)}{2} \|\mathbf{x}^* - \mathbf{y}\|_2^2 \end{aligned}$$

holds for any $\alpha \in [0, 1]$. For any $\alpha \in (0, 1)$, the point $\mathbf{z} = \alpha \mathbf{x}^* + (1 - \alpha)\mathbf{y}$ holds $f(\mathbf{z}) < f(\mathbf{x}^*)$, which leads to contradiction. □

**Remark 3.7.** *For any approximate solution $\hat{\mathbf{x}}$ satisfying $f(\mathbf{x}) \leq f(\mathbf{x}^*) + \epsilon$ for any $\mathbf{x}$, we have*

$$\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 \leq 2\epsilon/\mu.$$

*Let $\mathbf{g} \in \partial f(\mathbf{x}^*)$, then we have*

$$\begin{aligned} f(\mathbf{x}) \geq & f(\mathbf{x}^*) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x}^* \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2 \\ \geq & f(\mathbf{x}^*) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2 \\ \geq & f(\mathbf{x}) - \epsilon + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2 . \end{aligned}$$

**Remark 3.8.** *However, the strong convexity alone cannot guarantee the existence of a minimizer. Consider the function*

$$f(x) = \begin{cases} x^2, & \text{if } x > 0, \\ 1, & \text{if } x = 0. \end{cases}$$

*We can verify the strong convexity based on finding $\mu > 0$ for*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y) - \frac{\mu \alpha (1 - \alpha)}{2} \|x - y\|_2^2 ,$$

*where $x, y \in [0, +\infty)$ and $\alpha \in [0, 1]$. If $x, y \in (0, +\infty)$ or $x = y = 0$, it obviously holds for $\mu = 2$. If $x > 0$ and $y = 0$, the condition can be written as*

$$(\alpha x)^2 \leq \alpha x^2 + (1 - \alpha) - \frac{\mu\alpha(1 - \alpha)x^2}{2}.$$

*Taking $\mu = 2$, it can be written as*

$$\alpha^2 x^2 \leq \alpha x^2 + (1 - \alpha) - \alpha(1 - \alpha)x^2 \quad \iff \quad 0 \leq 1 - \alpha.$$

*Hence, this function is 2-strongly convex but it has no minimizer.*

**Remark 3.9.** *Besides the strong convexity, the existence of minimizer also require the function $f : \mathcal{C} \to \mathbb{R}$ is lower semi-continuous, i.e., for any $\mathbf{x}_0 \in \mathcal{C}$ and $y \in \mathbb{R}$ with $y < f(\mathbf{x}_0)$, there exists $\delta > 0$ such that $y < f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{B}_\delta(\mathbf{x}_0) \cap \mathcal{C}$.*

**Remark 3.10.** *Lower semi-continuouity alone cannot leads to the existence of minimizer, such as the function $f(x) = \exp(x)$.*

**Theorem 3.17.** *A convex function $f$ is $G$-Lipschitz continuous on $(\operatorname{dom} f)^\circ$ if and only if*

$$\|\mathbf{g}\|_2 \leq G$$

*for all $\mathbf{g} \in \partial f(\mathbf{x})$ and $\mathbf{x} \in (\operatorname{dom} f)^\circ$.*

*Proof.* **Part I:** Suppose the subgradient is bounded. There exists $\mathbf{g}_1 \in \partial f(\mathbf{x}_1)$ and $\mathbf{g}_2 \in \partial f(\mathbf{x}_2)$, we have

$$f(\mathbf{x}_2) - f(\mathbf{x}_1) \leq \langle \mathbf{g}_2, \mathbf{x}_2 - \mathbf{x}_1 \rangle \leq \|\mathbf{g}_2\|_2 \|\mathbf{x}_2 - \mathbf{x}_1\|_2 \leq G \|\mathbf{x}_2 - \mathbf{x}_1\|_2$$

and

$$f(\mathbf{x}_1) - f(\mathbf{x}_2) \leq \langle \mathbf{g}_1, \mathbf{x}_1 - \mathbf{x}_2 \rangle \leq \|\mathbf{g}_1\|_2 \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq G \|\mathbf{x}_1 - \mathbf{x}_2\|_2 ,$$

which means $|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq G \|\mathbf{x}_1 - \mathbf{x}_2\|_2$.

**Part II:** Suppose $f(\cdot)$ is $G$-Lipschitz continuous. For any $\mathbf{x} \in (\operatorname{dom} f)^\circ$ and $\mathbf{g} \in \partial f(\mathbf{x})$, we have

$$G \|\mathbf{y} - \mathbf{x}\|_2 \geq f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle$$

for all $\mathbf{y}$. Let $\mathbf{y} = \mathbf{x} + \epsilon\mathbf{g}$ for sufficient small $\epsilon > 0$ such that $\mathbf{y}$ in the interior of the domain, then we have

$$G \|\epsilon\mathbf{g}\|_2 \geq \langle \mathbf{g}, \epsilon\mathbf{g} \rangle,$$

that is $\|\mathbf{g}\|_2 \leq G$. □

**Theorem 3.18.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is $L$-smooth (possibly nonconvex), then it holds*

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 .$$

*for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

*Proof.* For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we define

$$g(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$$

on $t \in [0, 1]$. It holds that

$$g(0) = f(\mathbf{x}), \qquad g(1) = f(\mathbf{y}) \qquad \text{and} \qquad g'(t) = \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle.$$

Then we have

$$\begin{aligned}
&|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \\
=&|g(1) - g(0) - g'(0)| \\
=&\left| \int_0^1 g'(t)\,\mathrm{d}t - \int_0^1 g'(0)\,\mathrm{d}t \right| \\
\leq& \int_0^1 |g'(t) - g'(0)|\,\mathrm{d}t \\
=& \int_0^1 |\langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle|\,\mathrm{d}t \\
\leq& \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|_2 \|\mathbf{y} - \mathbf{x}\|_2\,\mathrm{d}t \\
\leq& \int_0^1 Lt \|\mathbf{y} - \mathbf{x}\|_2^2\,\mathrm{d}t = \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.
\end{aligned}$$

□

**Theorem 3.19.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and $L$-smooth, then we have*

1. $0 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$
2. $f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2^2 \leq f(\mathbf{y})$
3. $\frac{1}{L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2^2 \leq \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$

*for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

*Proof.* **Part I:** Apply Theorem 3.8 and 3.19.
**Part II:** Define the function

$$\phi(\mathbf{x}) = f(\mathbf{x}) - \langle \nabla f(\mathbf{x}_0), \mathbf{x} \rangle,$$

which is convex and $L$-smooth, i.e., we have

$$\begin{aligned}
&\phi(\mathbf{y}) \geq \phi(\mathbf{x}) + \langle \nabla \phi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\
\Longleftrightarrow& f(\mathbf{y}) - \langle \nabla f(\mathbf{x}_0), \mathbf{y} \rangle \geq f(\mathbf{x}) - \langle \nabla f(\mathbf{x}_0), \mathbf{x} \rangle + \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_0), \mathbf{y} - \mathbf{x} \rangle \\
\Longleftrightarrow& f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla \phi(\mathbf{y}) - \nabla \phi(\mathbf{x})\|_2 &= \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}_0) - (\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_0))\|_2 \\
&= \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2.
\end{aligned}$$

We can verify $\mathbf{y}^* = \mathbf{x}_0$ is a minimizer of $\phi(\cdot)$, then

$$\begin{aligned}
\phi(\mathbf{x}_0) =& \min_{\mathbf{y} \in \mathbb{R}} \phi(\mathbf{y}) \leq \min_{\mathbf{y} \in \mathbb{R}} \left( \phi(\mathbf{x}) + \langle \nabla \phi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \right) \\
=& \min_{\mathbf{y} \in \mathbb{R}} \left( \phi(\mathbf{x}) + \frac{L}{2} \left\| \mathbf{y} - \mathbf{x} + \frac{1}{L}\nabla \phi(\mathbf{x}) \right\|_2^2 - \frac{1}{2L}\|\nabla \phi(\mathbf{x})\|_2^2 \right) \\
=& \phi(\mathbf{x}) - \frac{1}{2L}\|\nabla \phi(\mathbf{x})\|_2^2.
\end{aligned}$$

We can verify $\nabla\phi(\mathbf{x}) = \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_0)$, which implies

$$f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{x}_0\rangle \leq f(\mathbf{x}) - \langle \nabla f(\mathbf{x}_0), \mathbf{x}\rangle - \frac{1}{2L}\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_0)\right\|_2^2.$$

Since $\mathbf{x}_0$ and $\mathbf{x}$ are arbitrary, we finish the proof.

**Part III:** Summing over the second inequality by changing the role of $\mathbf{x}$ and $\mathbf{y}$, we obtain

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{x}\rangle + \frac{1}{L}\left\|\nabla f(\mathbf{y}) - \nabla(\mathbf{x})\right\|_2^2 \leq 0.$$

Arranging above inequality achieve the desired result. □

**Remark 3.11.** *Under convex assumption, the L-smoothness and these three condition are equivalent. In above proof, we have shown L-smooth $\Longrightarrow$ point 1 $\Longrightarrow$ point 2 $\Longrightarrow$ point 3. We can also show the last result can lead to $\Longrightarrow$ L-smooth. Combining Cauchy–Schwarz inequality, we obtain*

$$\frac{1}{L}\left\|\nabla f(\mathbf{y}) - \nabla(\mathbf{x})\right\|_2^2 \leq \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \leq \left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\|_2 \left\|\mathbf{x} - \mathbf{y}\right\|_2,$$

*which implies $\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\|_2 \leq L\left\|\mathbf{x} - \mathbf{y}\right\|_2$.*

**Theorem 3.20** (second-order condition)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. Suppose that the Hessian $\nabla^2 f(\cdot)$ is continuous in an open neighborhood of $\mathbf{x}^* \in \mathbb{R}^d$.*

1. *If $\mathbf{x}^*$ is a local minimizer of $f(\cdot)$, then it holds that*
$$\nabla f(\mathbf{x}^*) = \mathbf{0} \qquad \textit{and} \qquad \nabla^2 f(\mathbf{x}^*) \succeq \mathbf{0}.$$
2. *If it holds that*
$$\nabla f(\mathbf{x}^*) = \mathbf{0} \qquad \textit{and} \qquad \nabla^2 f(\mathbf{x}^*) \succ \mathbf{0},$$
*then the point $\mathbf{x}^*$ is a strict local minimizer of $f(\cdot)$.*

*Proof.* **Part I:** Suppose $\nabla f(\mathbf{x}^*) \neq 0$. We define

$$\mathbf{p} = -\nabla f(\mathbf{x}^*),$$

which means $\langle \mathbf{p}, \nabla f(\mathbf{x}^*)\rangle < 0$. The continuity of $\nabla f(\cdot)$ means there exists some $T > 0$ such that

$$\langle \mathbf{p}, \nabla f(\mathbf{x}^* + t\mathbf{p})\rangle < 0$$

for any $t \in (0, T)$. For any $\hat{t} \in (0, T)$, Taylor's theorem means there exist some $t \in (0, \hat{t}) \subseteq (0, T]$ such that

$$f(\mathbf{x}^* + \hat{t}\mathbf{p}) = f(\mathbf{x}^*) + \langle t\mathbf{p}, \nabla f(\mathbf{x}^* + t\mathbf{p})\rangle < f(\mathbf{x}^*),$$

which leads to contradiction. Hence, we conclude $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Suppose the Hessian $\nabla f(\mathbf{x}^*)$ is not positive semi-definite. Then we can find some vector $\mathbf{p} \in \mathbb{R}^d$ such that $\langle \nabla^2 f(\mathbf{x}^*)\mathbf{p}, \mathbf{p}\rangle < 0$. The continuity of Hessian means there exist some $T > 0$ such that for any $t \in [0, T]$ holds that

$$\langle \nabla^2 f(\mathbf{x}^* + t\mathbf{p})\mathbf{p}, \mathbf{p}\rangle < 0$$

Doing Taylor expansion around $\mathbf{x}^*$, we have for all $\hat{t} \in (0, T)$, there exist some $t \in (0, \hat{t}) \subseteq (0, T]$ such that

$$f(\mathbf{x}^* + \hat{t}\mathbf{p}) = f(\mathbf{x}^*) + \langle \hat{t}\mathbf{p}, \nabla f(\mathbf{x}^*)\rangle + \frac{1}{2}\mathbf{p}^\top \nabla^2(\mathbf{x}^* + t\mathbf{p})\mathbf{p} < f(\mathbf{x}^*),$$

which leads to contradiction. Hence, we conclude $\nabla f(\mathbf{x}^*)$ is positive semi-definite.

**Part II:** The continuity of Hessian means the positive definiteness of Hessian still hold in $\mathcal{B}(\mathbf{x}^*, \delta)$ for some $\delta > 0$. For any $\mathbf{p} \in \mathbb{R}^d$ with $\left\|\mathbf{p}\right\|_2 < \delta$, then we have

$$f(\mathbf{x}^* + \mathbf{p}) = f(\mathbf{x}^*) + \langle \mathbf{p}, \nabla f(\mathbf{x}^*)\rangle + \frac{1}{2}\mathbf{p}^\top \nabla^2(\mathbf{x}^* + t\mathbf{p})\mathbf{p} > f(\mathbf{x}^*)$$

for some $t \in (0, 1)$. Hence, the point $\mathbf{x}^*$ is a strict local minimizer. □

**Remark 3.12.** *We cannot state "if and only if". Consider the function $f(x) = x^3$ at $x = 0$.*

**Remark 3.13.** *We can also define third-order necessary condition for $\mathbf{x}$ as follows*

1. $\nabla f(\mathbf{x}) = \mathbf{0}$,
2. $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$,
3. *Any $\mathbf{u} \in \mathbb{R}^d$ satisfies $\mathbf{u}^\top \nabla^2 f(\mathbf{x})\mathbf{u} = 0$ holds that $D^3 f(\mathbf{x})[\mathbf{u}, \mathbf{u}, \mathbf{u}] = 0$,*

*where we denote*

$$D^3 f(\mathbf{x})(\mathbf{u}, \mathbf{u}, \mathbf{u}) = \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} \frac{\partial^3 f(\mathbf{x})}{\partial u_i \partial u_j \partial u_k} \cdot u_i u_j u_k.$$

**Theorem 3.21** (Smoothness and Convexity)**.** *Let $f(\cdot)$ be a twice differentiable function defined on $\mathbb{R}^d$*

1. *It is $L$-smooth if and only if*

$$-L\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$$

   *for all $\mathbf{x} \in \mathbb{R}^d$.*

2. *It is convex if and only if*

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$$

   *for all $\mathbf{x} \in \mathbb{R}^d$.*

3. *It is $\mu$-strongly-convex if and only if*

$$\nabla^2 f(\mathbf{x}) \succeq \mu\mathbf{I}$$

   *for all $\mathbf{x} \in \mathbb{R}^d$.*



*Proof.* **Part I:** Suppose $f$ is $L$-smooth. For any $\mathbf{x}, \mathbf{p} \in \mathbb{R}^d$, we have

$$\nabla^2 f(\mathbf{x})\mathbf{p} = \lim_{t \to 0} \frac{\nabla f(\mathbf{x} + t\mathbf{p}) - \nabla f(\mathbf{x})}{t}.$$

Taking the $\ell_2$-norm on both sides, we obtain

$$\begin{aligned} \left\|\nabla^2 f(\mathbf{x})\mathbf{p}\right\|_2 &\leq \lim_{t \to 0} \left\| \frac{\nabla f(\mathbf{x} + t\mathbf{p}) - \nabla f(\mathbf{x})}{t} \right\|_2 \\ &\leq \lim_{t \to 0} \left\| \frac{t\mathbf{p}}{t} \right\|_2 = \|\mathbf{p}\|_2 , \end{aligned}$$

which means $\left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L$.

Suppose any $\mathbf{x} \in \mathbb{R}^d$ holds that $\left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L$. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we construct $\mathbf{v} : \mathbb{R} \to \mathbb{R}^d$ as follows

$$\mathbf{v}(t) = \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})),$$

which means

$$\mathbf{v}'(t) = \nabla^2 f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y}).$$

Then we have

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 = \|\mathbf{v}(1) - \mathbf{v}(0)\|_2 = \left\| \int_0^1 \mathbf{v}'(t)\, \mathrm{d}t \right\|_2$$

$$
\begin{aligned}
&\leq \left\| \int_0^1 \nabla^2 f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y}) \, \mathrm{d}t \right\|_2 \\
&\leq \int_0^1 \left\| \nabla^2 f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) \right\|_2 \|\mathbf{x} - \mathbf{y}\|_2 \, \mathrm{d}t \\
&\leq L \|\mathbf{x} - \mathbf{y}\|_2 .
\end{aligned}
$$

**Part II:** Suppose $f$ is convex. We construct $g : \mathbb{R}^d \to \mathbb{R}$ as follows

$$
g(\mathbf{y}) = f(\mathbf{y}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.
$$

Then for any $\mathbf{y} \in \mathbb{R}^d$, we have

$$
\nabla g(\mathbf{y}) = \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) \quad \text{and} \quad \nabla g^2(\mathbf{y}) = \nabla^2 f(\mathbf{y}).
$$

Therefore, the point $\mathbf{x}$ is a minimizer of $\mathbf{g}(\cdot)$ since we can verify $\mathbf{g}(\cdot)$ is convex and $\nabla g(\mathbf{x}) = 0$. The second-order optimal condition (Theorem 3.20) means $\nabla^2 f(\mathbf{y}) = \nabla^2 g(\mathbf{x}) \succeq \mathbf{0}$.

Suppose we have the Hessian is positive semi-definite on $\mathbb{R}^d$. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, Taylor's theorem implies there exist some $t \in [0, 1]$ such that

$$
f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle,
$$

which just is the first-order condition of convex function. Then we achieve the convexity.

**Part II:** Recall that the strongly convexity of $f(\mathbf{x})$ means $f(\mathbf{x}) - \frac{\mu}{2}$ is convex. Using above result, we have

$$
\nabla^2 \left( f(\mathbf{x}) - \frac{\mu}{2} \|\mathbf{x}\|_2^2 \right) = \nabla^2 f(\mathbf{x}) - \mu \mathbf{I} \succeq \mathbf{0} \Longleftrightarrow \nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}.
$$

□

**Example 3.7.** *For unconstrained quadratic problem*

$$
\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{2} \mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbf{b}^\top \mathbf{x},
$$

*where* $\mathbf{A} \in \mathbb{R}^{d \times d}$ *is positive-definite and* $\mathbf{b} \in \mathbb{R}^d$*. We can check its property by*

$$
\nabla^2 f(\mathbf{x}) = \mathbf{A}.
$$

**Example 3.8.** *For regularized generalized linear model*

$$
\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n} \sum_{i=1}^n \phi_i(\mathbf{a}^\top \mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x}\|_2^2 .
$$

*where* $\phi_i : \mathbb{R}^d \to \mathbb{R}$ *is smooth and twice differentiable. We have*

$$
\nabla f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \phi_i'(\mathbf{a}_i^\top \mathbf{x}) \mathbf{a} + \lambda \mathbf{x} \qquad \textit{and} \qquad \nabla^2 f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \phi_i''(\mathbf{a}_i^\top \mathbf{x}) \mathbf{a}_i \mathbf{a}_i^\top + \lambda \mathbf{I}.
$$

*For logistic loss* $\phi(z) = \ln(1 + \exp(-z))$*, we have*

$$
\phi'(z) = \frac{-1}{1 + \exp(-z)} \qquad \textit{and} \qquad \phi''(z) = \frac{\exp(-z)}{(1 + \exp(-z))^2} > 0.
$$

*We can verify*

$$
\lim_{z \to +\infty} \phi''(z) = 0, \quad \lim_{z \to -\infty} \phi''(z) = 0 \quad \textit{and} \quad 0 < \phi''(z) \leq \frac{1}{4},
$$

*then*

$$\lambda \mathbf{I} \prec \nabla^2 f(\mathbf{x}) \preceq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{4}\mathbf{a}_i\mathbf{a}_i^\top + \lambda \mathbf{I} \preceq \frac{1}{4n}\mathbf{A}^\top\mathbf{A} + \lambda \mathbf{I} \preceq \frac{\left\|\mathbf{A}^\top\mathbf{A}\right\|_2}{4n} + \lambda \mathbf{I} \quad \textit{and} \quad \mathbf{A} = \begin{bmatrix}\mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_n^\top\end{bmatrix} \in \mathbb{R}^{n\times d}.$$

*If $\lambda = 0$, the function is strictly convex, rather than strongly convex. Note that we have $\phi(z) \to 0$ when $z \to 0$.*

**Example 3.9.** *Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$, the solution of minimization problem*

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \triangleq \frac{1}{2}\left\|\mathbf{A}\mathbf{x} - \mathbf{b}\right\|_2^2.$$

*is $\hat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^n$.*

*Proof.* We can verify

$$\nabla f(\mathbf{x}) = \mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}) = \mathbf{A}^\top\mathbf{A} \succeq \mathbf{0},$$

which means $f(\mathbf{x})$ is convex. Hence, we only needs to solve the linear system

$$\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0}.$$

If $\mathbf{A}^\top\mathbf{A}$ is full rank, we have

$$\mathbf{x}^* = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}.$$

Otherwise, we let $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top$ be the condense SVD, where $r$ is the rank of $\mathbf{A}$. We denote the solution of $\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0}$ be

$$\mathcal{X} = \left\{\mathbf{x} : \mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0}\right\}$$

and denote $\mathcal{X}_1 = \left\{\mathbf{x} : \mathbf{x} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y},\ \ \mathbf{y} \in \mathbb{R}^n\right\}$. We can verify that $\mathbf{x}^* = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y} \in \mathcal{X}_1$ satisfies $\mathbf{x}^* \in \mathcal{X}$ as follows

$$\begin{aligned}
&\mathbf{A}^\top\mathbf{A}\mathbf{x}^* - \mathbf{A}^\top\mathbf{b} \\
=&\mathbf{A}^\top\mathbf{A}\left(\mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y}\right) - \mathbf{A}^\top\mathbf{b} \\
=&\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\dagger - \mathbf{I})\mathbf{b} + \mathbf{A}^\top\mathbf{A}\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{y} \\
=&\mathbf{V}_r\mathbf{\Sigma}_r\mathbf{U}_r^\top(\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top\mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^\top - \mathbf{I})\mathbf{b} + \mathbf{V}_r\mathbf{\Sigma}_r\mathbf{U}_r^\top\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top\left(\mathbf{I} - \mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^\top\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top\right)\mathbf{y} \\
=&\mathbf{V}_r\mathbf{\Sigma}_r\mathbf{U}_r^\top(\mathbf{U}_r\mathbf{U}_r^\top - \mathbf{I})\mathbf{b} + \mathbf{V}_r\mathbf{\Sigma}_r^2\mathbf{V}_r^\top\left(\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top\right)\mathbf{y} \\
=&\mathbf{V}_r\mathbf{\Sigma}_r(\mathbf{U}_r^\top - \mathbf{U}_r^\top)\mathbf{b} + \mathbf{V}_r\mathbf{\Sigma}_r^2\left(\mathbf{V}_r^\top - \mathbf{V}_r^\top\right)\mathbf{y} = \mathbf{0}.
\end{aligned}$$

Hence, we have $\mathcal{X}_1 \subseteq \mathcal{X}$.

For any $\mathbf{x} \in \mathcal{X}$, we have

$$\begin{aligned}
&\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0} \\
\Longleftrightarrow&\mathbf{V}_r\mathbf{\Sigma}_r^2\mathbf{V}_r^\top\mathbf{x} - \mathbf{V}_r\mathbf{\Sigma}_r\mathbf{U}_r^\top\mathbf{b} = \mathbf{0} \\
\Longleftrightarrow&\mathbf{\Sigma}_r^2\mathbf{V}_r^\top\mathbf{x} - \mathbf{\Sigma}_r\mathbf{U}_r^\top\mathbf{b} = \mathbf{0} \\
\Longleftrightarrow&\mathbf{V}_r^\top\mathbf{x} = \mathbf{\Sigma}_r^{-1}\mathbf{U}_r^\top\mathbf{b} \\
\Longleftrightarrow&\mathbf{V}_r\mathbf{V}_r^\top\mathbf{x} = \mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^\top\mathbf{b} \\
\Longleftrightarrow&\mathbf{x} - (\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top)\mathbf{x} = \mathbf{A}^\dagger\mathbf{b} \\
\Longleftrightarrow&\mathbf{x} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top)\mathbf{x} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}.
\end{aligned}$$

Then we have $\mathbf{x} \in \left\{\mathbf{x} : \mathbf{x} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top)\mathbf{x}\right\} \subseteq \mathcal{X}_1$, which means $\mathcal{X} \subseteq \mathcal{X}_1$. Hence, we conclude $\mathcal{X} = \mathcal{X}_1$. □

# 4 Gradient Descent Methods

**Theorem 4.1.** *For the minimization problem*

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \tag{6}$$

*with $L$-smooth function $f:\mathbb{R}^d\to\mathbb{R}$ and optimal solution $\mathbf{x}^*$, we generate $\mathbf{x}_t$ by gradient descent method*

$$\mathbf{x}_{t+1}=\mathbf{x}_t-\eta_t\nabla f(\mathbf{x}_t).$$

*for $\eta_t=\eta\le 1/L$. Then we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t)\le f(\hat{\mathbf{x}})+\frac{\|\mathbf{x}_0-\hat{\mathbf{x}}\|_2^2}{2\eta T}$$

*for any $\hat{\mathbf{x}}\in\mathbb{R}^d$.*

*Proof.* Theorem 3.19 means

$$\begin{aligned} f(\mathbf{x}_{t+1}) \le& f(\mathbf{x}_t)+\langle\nabla f(\mathbf{x}_t),\mathbf{x}_{t+1}-\mathbf{x}_t\rangle+\frac{L}{2}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|_2^2\\ \le& f(\mathbf{x}_t)-\left(\eta-\frac{L\eta^2}{2}\right)\|\nabla f(\mathbf{x}_t)\|_2^2\\ \le& f(\mathbf{x}_t)-\frac{\eta}{2}\|\nabla f(\mathbf{x}_t)\|_2^2 \end{aligned} \tag{7}$$

Now we obtain

$$\begin{aligned} &\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\|_2^2\\ =&\|\mathbf{x}_t-\eta\nabla f(\mathbf{x}_t)-\hat{\mathbf{x}}\|_2^2\\ =&\|\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2-2\eta\langle\nabla f(\mathbf{x}_t),\mathbf{x}_t-\hat{\mathbf{x}}\rangle+\eta^2\|\nabla f(\mathbf{x}_t)\|_2^2\\ \le&\|\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2+2\eta(f(\hat{\mathbf{x}})-f(\mathbf{x}_t))+\eta^2\|\nabla f(\mathbf{x}_t)\|_2^2\\ \le&\|\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2+2\eta(f(\hat{\mathbf{x}})-f(\mathbf{x}_t))+2\eta(f(\mathbf{x}_t)-f(\mathbf{x}_{t+1}))\\ \le&\|\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2+2\eta(f(\hat{\mathbf{x}})-f(\mathbf{x}_{t+1})), \end{aligned}$$

where the first inequality uses the convexity of $f$ such that $f(\hat{\mathbf{x}})\ge f(\mathbf{x}_t)+\langle\hat{\mathbf{x}}-\mathbf{x}_t,\nabla f(\mathbf{x}_t)\rangle$ and the second inequality uses (7). Taking the average over above result with $t=0,\dots,T-1$, we finish the proof. □

**Remark 4.1.** *Additionally suppose $f(\cdot)$ has a minimizer $\mathbf{x}^*$ and let $\bar{\mathbf{x}}=\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}_t$, then we need*

$$T=\left\lceil\frac{L\|\mathbf{x}_0-\mathbf{x}^*\|_2^2}{2}\cdot\frac{1}{2\epsilon}\right\rceil$$

*to guarantee $f(\bar{\mathbf{x}})-f(\mathbf{x}^*)\le\epsilon$.*

**Theorem 4.2.** *Under the setting of Theorem 4.1, we suppose $\mathbf{x}^*$ is a minimizer of the problem, then*

$$f(\mathbf{x}_T)-f(\mathbf{x}^*)\le\frac{2L\|\mathbf{x}_0-\mathbf{x}^*\|_2^2}{t+4}.$$

*Proof.* We have

$$\begin{aligned} &\|\mathbf{x}_{t+1}-\mathbf{x}^*\|_2^2\\ =&\|\mathbf{x}_t-\eta\nabla f(\mathbf{x}_t)-\mathbf{x}^*\|_2^2 \end{aligned}$$

$$
\begin{aligned}
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle + \eta^2 \|\nabla f(\mathbf{x}_t)\|_2^2 \\
&\le \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \frac{2\eta}{L} \|\nabla f(\mathbf{x}_t)\|_2^2 + \eta^2 \|\nabla f(\mathbf{x}_t)\|_2^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \left(\frac{2\eta}{L} - \eta^2\right) \|\nabla f(\mathbf{x}_t)\|_2^2 ,
\end{aligned}
$$

where the inequality is based on the last statement of Theorem 3.19 that leads to

$$
\frac{1}{L} \|\nabla f(\mathbf{x}) - \nabla(\mathbf{x}^*)\|_2^2 \le \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle.
$$

Therefore, we have $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \le \|\mathbf{x}_t - \mathbf{x}^*\|_2 \le \cdots \le \|\mathbf{x}_0 - \mathbf{x}^*\|_2$ and the convexity means

$$
f(\mathbf{x}_t) - f(\mathbf{x}^*) \le \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle \le \|\nabla f(\mathbf{x}_t)\|_2 \|\mathbf{x}_t - \mathbf{x}^*\|_2 \le \|\nabla f(\mathbf{x}_t)\|_2 \|\mathbf{x}_0 - \mathbf{x}^*\|_2 .
$$

Then the inequality (7) in the proof of previous theorem means

$$
\begin{aligned}
& f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \\
\le & f(\mathbf{x}_t) - f(\mathbf{x}^*) - \frac{\eta}{2} \|\nabla f(\mathbf{x}_t)\|_2^2 \\
\le & f(\mathbf{x}_t) - f(\mathbf{x}^*) - \frac{\eta (f(\mathbf{x}_t) - f(\mathbf{x}^*))^2}{2 \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}.
\end{aligned}
$$

Consequently, we obtain

$$
\frac{1}{f(\mathbf{x}_t) - f(\mathbf{x}^*)} \le \frac{1}{f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)} - \frac{\eta (f(\mathbf{x}_t) - f(\mathbf{x}^*))}{2 \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 (f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*))} \le \frac{1}{f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)} - \frac{\eta}{2 \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2},
$$

where the last step use the non-decreasing of $f(\mathbf{x}_t) - f(\mathbf{x}^*)$. Summing over above over $t = 0, \ldots, T-1$ means

$$
\frac{1}{f(\mathbf{x}_0) - f(\mathbf{x}^*)} \le \frac{1}{f(\mathbf{x}_T) - f(\mathbf{x}^*)} - \frac{t}{2L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2},
$$

that is

$$
f(\mathbf{x}_T) - f(\mathbf{x}^*) \le \frac{2L(f(\mathbf{x}_0) - f(\mathbf{x}^*)) \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{t(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + 2L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2} = \frac{2L(f(\mathbf{x}_0) - f(\mathbf{x}^*)) \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{t(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + 2L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}.
$$

Since RHS in above is increasing in $f(\mathbf{x}_0) - f(\mathbf{x}^*)$ and

$$
f(\mathbf{x}_0) - f(\mathbf{x}^*) \le \langle \nabla f(\mathbf{x}^*), \mathbf{x}_0 - \mathbf{x}^*\rangle + \frac{L}{2} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 = \frac{L}{2} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 ,
$$

we have

$$
f(\mathbf{x}_T) - f(\mathbf{x}^*) \le \frac{2L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{t + 4}.
$$

□

**Remark 4.2** (nonconvex case)**.** *Noticing that inequality (7) holds even if the function is nonconvex, then we have*

$$
\frac{\eta}{2T} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|_2^2 \le \frac{f(\mathbf{x}_0) - f(\mathbf{x}_T)}{T}.
$$

*Let* $\hat{\mathbf{x}}$ *be uniformly sampled from* $\{\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}\}$*, we have*

$$
\mathbb{E} \|\nabla f(\hat{\mathbf{x}})\|_2^2 \le \frac{2(f(\mathbf{x}_0) - f(\mathbf{x}_T))}{\eta T} \le \frac{2L(f(\mathbf{x}_0) - f^*)}{T},
$$

*where we suppose*

$$f^* = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty.$$

*Hence, taking* $T \geq 2L(f(\mathbf{x}_0) - f^*)\epsilon^{-2}$ *leads to an* $\epsilon$*-stationary point in expectation.*

**Theorem 4.3.** *Under the setting of Theorem 4.1, we additionally suppose the objective is* $\mu$*-strongly-convex, then*

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \left(1 - \frac{\mu}{L}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^*))$$

*Proof.* The strong convexity means

$$\begin{aligned} f(\mathbf{x}^*) \geq & f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|_2 \\ = & f(\mathbf{x}) + \frac{\mu}{2} \left\| \mathbf{x} - \mathbf{x}^* - \frac{1}{\mu}\nabla f(\mathbf{x}) \right\|_2 - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2 \\ \geq & f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2 . \end{aligned}$$

for any $\mathbf{x} \in \mathbb{R}^d$. Using the result of (7), we have

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\eta}{2} \|\nabla f(\mathbf{x}_t)\|_2^2 \leq f(\mathbf{x}_t) - \mu\eta(f(\mathbf{x}_t) - f(\mathbf{x}^*)),$$

that is

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq \left(1 - \frac{\mu}{L}\right) (f(\mathbf{x}_t) - f(\mathbf{x}^*)).$$

Then we obtain $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq (1 - \mu/L)^T (f(\mathbf{x}_0) - f(\mathbf{x}^*))$. □

**Remark 4.3.** *We can find* $\mathbf{x}_T$ *such that* $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \epsilon$ *within* $\kappa \ln((f(\mathbf{x}_T) - f(\mathbf{x}^*))/\epsilon)$ *first-order oracle complexity, where* $\kappa \triangleq L/\mu$ *is the condition number. The Bernoulli's inequality says for any* $0 < z < 1$*, we have*

$$\exp(z) = \sum_{k=0}^{+\infty} \frac{z^k}{k!} \leq \sum_{k=0}^{+\infty} z^k = \frac{1}{1-z}.$$

*Let* $z = 1/\kappa$*, we achieve*

$$\begin{aligned} & \exp\left(\frac{1}{\kappa}\right) \leq \frac{\kappa}{\kappa - 1} \Longrightarrow \left(1 - \frac{1}{\kappa}\right)^T \leq \exp\left(-\frac{T}{\kappa}\right) \leq \epsilon \\ \Longrightarrow & \left(1 - \frac{1}{\kappa}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^*)) \leq \exp\left(-\frac{T}{\kappa}\right) (f(\mathbf{x}_0) - f(\mathbf{x}^*)) \leq \epsilon, \end{aligned}$$

*where the last inequality requires*

$$\frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{\epsilon} \leq \exp\left(\frac{T}{\kappa}\right) \Longleftarrow \kappa \ln\left(\frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{\epsilon}\right) \leq T.$$

**Example 4.1.** *Define* $f : \mathbb{R}^d \to \mathbb{R}$ *as*

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x},$$

*where* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *is nonzero positive semi-definite matrix (but not positive definite). Since matrix* $\mathbf{A}$ *is not full rank, there exists* $\mathbf{x}^* \in \mathbb{R}^d$ *such that* $\mathbf{A}\mathbf{x}^* = \mathbf{b}$*. Then we have*

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} - \left(\frac{1}{2}\mathbf{x}^{*\top} \mathbf{A}\mathbf{x}^* - \mathbf{b}^\top \mathbf{x}^*\right)$$

$$
\begin{aligned}
=&\frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x}-{\mathbf{x}^*}^\top\mathbf{A}\mathbf{x}-\left(\frac{1}{2}{\mathbf{x}^*}^\top\mathbf{A}\mathbf{x}^*-{\mathbf{x}^*}^\top\mathbf{A}\mathbf{x}^*\right)\\
=&\frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x}-{\mathbf{x}^*}^\top\mathbf{A}\mathbf{x}+\frac{1}{2}{\mathbf{x}^*}^\top\mathbf{A}\mathbf{x}^*\\
=&\frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^\top\mathbf{A}(\mathbf{x}-\mathbf{x}^*)
\end{aligned}
$$

*and*

$$
\|\nabla f(\mathbf{x})\|_2^2=\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2=\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_2^2=(\mathbf{x}-\mathbf{x}^*)^\top\mathbf{A}^2(\mathbf{x}-\mathbf{x}^*).
$$

Taking $\mu=\lambda_k(\mathbf{A})$, where $\lambda_k(\mathbf{A})$ is the smallest nonzero eigenvalue of $\mathbf{A}$. Then it holds that $\mu\mathbf{A}\preceq\mathbf{A}^2$ and

$$
f(\mathbf{x})-f(\mathbf{x}^*)\le\frac{1}{2\mu}\|\nabla f(\mathbf{x})\|_2^2.
$$

**Theorem 4.4.** *Let $g:\mathbb{R}^m\to\mathbb{R}$ be smooth and $\mu$-strongly convex and $\mathbf{A}\in\mathbb{R}^{m\times d}$ with $\operatorname{rank}(\mathbf{A})=m$. Define the function $f:\mathbb{R}^d\to\mathbb{R}$ as $f(\mathbf{x})=g(\mathbf{A}\mathbf{x})$, then it satisfies PL condition with parameter $\mu/\left\|\mathbf{A}\mathbf{A}^\top\right\|_2$.*

*Proof.* We can verify

$$
\nabla f(\mathbf{x})=\mathbf{A}^\top\nabla g(\mathbf{A}\mathbf{x}).
$$

For any $\hat{\mathbf{x}}$, we have

$$
\begin{aligned}
&f(\mathbf{x})-f^*\\
=&g(\mathbf{A}\mathbf{x})-f^*\\
\le&g(\mathbf{A}\mathbf{x})-g^*\\
\le&\frac{1}{2\mu}\|\nabla g(\mathbf{A}\mathbf{x})\|_2^2\\
\le&\frac{1}{2\mu_f}\|\nabla f(\mathbf{x})\|_2^2\\
=&\frac{1}{2\mu_f}(\nabla g(\mathbf{A}\mathbf{x}))^\top\mathbf{A}\mathbf{A}^\top\nabla g(\mathbf{A}\mathbf{x})
\end{aligned}
$$

where the first inequality is due to $\mathbf{A}\mathbf{x}\subseteq\mathbb{R}^m$ and the last inequality requires

$$
\frac{1}{\mu}\mathbf{I}\preceq\frac{1}{\mu_f}\mathbf{A}\mathbf{A}^\top\quad\Longleftarrow\quad\mu_f\mathbf{I}\preceq\mu\mathbf{A}\mathbf{A}^\top\quad\Longleftarrow\quad\mu_f=\mu\lambda_m(\mathbf{A}\mathbf{A}^\top).
$$

□

**Example 4.2.** *Nonconvex function may also hold PL condition, such as $f(x)=x^2+3(\sin x)^2$.*

**Remark 4.4.** *The simple condition such that*

$$
f(\mathbf{x}_t+\alpha_t\mathbf{p}_t)<f(\mathbf{x}_t)
$$

*is not sufficient. Consider the problem*

$$
\min_{x\in\mathbb{R}^d}f(x)\triangleq x^2.
$$

*We set $x_0=1$, $p_t=-\operatorname{sign}(x)$ and $\alpha_t=1/3^{t+1}$, then*

$$
x_t=1-\left(\frac{1}{3}+\frac{1}{3^2}+\cdots+\frac{1}{3^t}\right)=\frac{1}{2}\left(1+\frac{1}{3^t}\right)
$$

*convergence to $1/2$. Additionally the Armijo condition is also not enough. Since the condition*

$$
f(\mathbf{x}+\alpha\mathbf{p})\le f(\mathbf{x})+c_1\alpha\langle\nabla f(\mathbf{x}),\mathbf{p}\rangle\quad\Longrightarrow\quad(x-\alpha)^2=1-2\alpha x+\alpha^2\le1-2c_1\alpha x
$$

*always holds for sufficient small $\alpha>0$ and $c_1\in(0,1)$.*

**Theorem 4.5.** *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable. Let $\mathbf{p}_t$ be a descent direction at $\mathbf{x}_t$ and assume $\phi(\alpha) = f(\mathbf{x}_t + \alpha\mathbf{p}_t)$ is bounded below on $\alpha \in (0, +\infty)$. Then there exist intervals of step lengths satisfying the Wolfe condition*

$$f(\mathbf{x}_t + \alpha_t\mathbf{p}_t) \leq f(\mathbf{x}_t) + c_1\alpha_t\langle\nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle, \tag{8}$$

$$\langle\nabla f(\mathbf{x}_t + \alpha_t\mathbf{p}_t), \mathbf{p}_t\rangle \geq c_2\langle\nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle \tag{9}$$

*with $0 < c_1 < c_2 < 1$.*

*Proof.* Consider that

$$\phi'(\alpha) = \langle\nabla f(\mathbf{x}_t + \alpha\mathbf{p}_t), \mathbf{p}_t\rangle.$$

Since $\phi(\alpha)$ is bounded below on $\alpha \in (0, +\infty)$ and the decent directions $\mathbf{p}_t$ means $\phi'(0) = \langle\nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle < 0$, the line

$$l(\alpha) = f(\mathbf{x}_k) + \alpha c_1\langle\nabla f(\mathbf{x}_l), \mathbf{p}_k\rangle$$

must intersect $\phi(\alpha)$ at least once. Let $\alpha' > 0$ be the smallest intersecting value of $\alpha$, that is

$$f(\mathbf{x}_t + \alpha'\mathbf{p}_t) = f(\mathbf{x}_t) + \alpha' c_1\langle\nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle. \tag{10}$$

Then condition (8) clearly holds for all $\alpha < \alpha'$.

By the mean value theorem, there exists $\alpha'' \in (0, \alpha')$ such that

$$\begin{aligned}
&\phi(0) = \phi(\alpha') + \phi(\alpha'')(0 - \alpha') \\
\iff &\phi(\alpha') - \phi(0) = \phi(\alpha'')\alpha' \\
\iff &f(\mathbf{x}_t + \alpha'\mathbf{p}_t) - f(\mathbf{x}_t) = \alpha'\langle\nabla f(\mathbf{x}_t + \alpha''\mathbf{p}_t), \mathbf{p}_t\rangle.
\end{aligned} \tag{11}$$

By combining (10) and (11), we obtain

$$\langle\nabla f(\mathbf{x}_t + \alpha''\mathbf{p}_t), \mathbf{p}_t\rangle = c_1\langle\nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle > c_2\langle\nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle,$$

where we use the condition $0 < c_1 < c_2$ and $\mathbf{p}_t$ is a descent direction. □

**Theorem 4.6.** *Consider any iteration of the form*

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t\mathbf{p}_t,$$

*where $\mathbf{p}_t$ is a descent direction such that*

$$\langle\mathbf{p}_t, \nabla f(\mathbf{x}_k)\rangle < 0.$$

*and $\alpha_k$ satisfies the Wolfe conditions (8)-(9). Suppose that continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and lower bounded on $\mathbb{R}^d$ and continuously differentiable. Then*

$$\sum_{t=0}^{+\infty}(\cos\theta_t)^2 \left\|\nabla f(\mathbf{x}_t)\right\|_2^2 < +\infty, \qquad \textit{where } \cos\theta_t = \frac{-\langle\nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle}{\left\|\nabla f(\mathbf{x}_t)\right\|_2 \left\|\mathbf{p}_t\right\|_2}.$$

*Proof.* From the iteration $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t\mathbf{p}_t$ and condition $\langle\nabla f(\mathbf{x}_t + \alpha_t\mathbf{p}_t), \mathbf{p}_t\rangle \geq c_2\langle\nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle$ we have

$$\langle\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle \geq (c_2 - 1)\langle\nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle.$$

The smoothness of $f$ means

$$\langle\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle \leq \left\|\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)\right\|_2 \left\|\mathbf{p}_t\right\|_2 \leq L\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2 \left\|\mathbf{p}_t\right\|_2 \leq \alpha_t L\left\|\mathbf{p}_t\right\|_2^2.$$

Combining above relations, we obtain

$$\alpha_t \geq \frac{(c_2-1)\langle \nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle}{L\left\|\mathbf{p}_t\right\|_2^2}.$$

By substituting this inequality into $f(\mathbf{x}_t+\alpha_t\mathbf{p}_t) \leq f(\mathbf{x}_t) + c_1\alpha_t\langle \nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle$, we obtain

$$f(\mathbf{x}_{t+1}) = (\mathbf{x}_t+\alpha_t\mathbf{p}_t) \leq f(\mathbf{x}_t) - \frac{c_1(1-c_2)(\langle \nabla f(\mathbf{x}_t), \mathbf{p}_t\rangle)^2}{L\left\|\mathbf{p}_t\right\|_2^2} = f(\mathbf{x}_t) - \frac{c(\cos\theta_t)^2\left\|\nabla f(\mathbf{x}_t)\right\|_2^2}{L},$$

where $c = c_1(1-c_2)$. Summing over above inequality with $t = 1, \ldots, k$ leads to

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_0) - \frac{c}{L}\sum_{t=0}^{k}(\cos\theta_t)^2\left\|\nabla f(\mathbf{x}_t)\right\|_2^2.$$

Since $f$ is lower bounded, we have

$$\sum_{t=0}^{k}(\cos\theta_t)^2\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \leq \frac{L}{c}\left(f(\mathbf{x}_0) - f(\mathbf{x}_{t+1})\right) < +\infty.$$

Taking $t \to +\infty$ finishes the proof. □

**Remark 4.5.** *This result implies*

$$\lim_{t\to+\infty}(\cos\theta_t)^2\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 = 0.$$

*If the search directions ensures are never too close to orthogonality with the gradient, that is*

$$\cos\theta_t \geq \delta > 0$$

*for all $t$, then* $\lim_{t\to+\infty}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 = 0$.

**Barzilai–Borwein Step Size** Taylor expansion says

$$f(\mathbf{x}_t+\mathbf{v}) = f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{v}\rangle + \frac{1}{2}\left\langle \mathbf{v}, \int_0^1 \nabla^2 f(\mathbf{x}_t+\tau\mathbf{v})\mathbf{v}\,\mathrm{d}\tau\right\rangle$$

for some $\tau \in [0,1]$. Minimizing RHS with approximation

$$\int_0^1 \nabla^2 f(\mathbf{x}_t+\tau\mathbf{v})\,\mathrm{d}\tau \approx \nabla^2 f(\mathbf{x}_t)$$

leads to $\mathbf{v} = -(\nabla^2 f(\mathbf{x}_t))^{-1}\nabla f(\mathbf{x}_t)$ and Newton's method

$$\mathbf{x}_{t+1} = \mathbf{x}_t - (\nabla^2 f(\mathbf{x}_t))^{-1}\nabla f(\mathbf{x}_t)$$

The Hessian holds the scent condition

$$\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) = \int_0^1 \nabla^2 f(\mathbf{x}_t+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))(\mathbf{x}_{t+1}-\mathbf{x}_t)\,\mathrm{d}\tau.$$

We consider the following approximation

$$\int_0^1 \nabla^2 f(\mathbf{x}_t+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))\,\mathrm{d}\tau \approx \frac{1}{\alpha}\mathbf{I} \Longrightarrow \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) = \alpha^{-1}(\mathbf{x}_{t+1}-\mathbf{x}_t).$$

for scent condition, which implies

$$\min_{\alpha>0}\left\|\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) - \alpha^{-1}(\mathbf{x}_{t+1}-\mathbf{x}_t)\right\|_2^2 \quad \text{or} \quad \min_{\alpha>0}\left\|\alpha(\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)) - (\mathbf{x}_{t+1}-\mathbf{x}_t)\right\|_2^2,$$

which leads to

$$\alpha^{\mathrm{BB1}} = \frac{\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2}{\langle \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1}-\mathbf{x}_t\rangle} \qquad \text{and} \qquad \alpha^{\mathrm{BB2}} = \frac{\langle \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1}-\mathbf{x}_t\rangle}{\left\|\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)\right\|_2^2}.$$

In practice, we use the BB step size obtain from the previous iteration.

# 5 Acceleration

**Theorem 5.1.** *Consider the quadratic problem*

$$\min_{\mathbf{x}\in\mathbb{R}^d} Q(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}, \tag{12}$$

*where* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *is positive definite and* $\mathbf{b} \in \mathbb{R}^d$*. The gradient descent method*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla Q(\mathbf{x}_t)$$

*with* $\eta \in (0, 2/L)$ *holds that*

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2 \le \rho^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2$$

*with* $\rho = \max\{1-\eta\mu, |1-\eta L|\} < 1$*, where* $L = \lambda_1(\mathbf{A})$ *and* $\mu = \lambda_d(\mathbf{A})$*.*

*Proof.* We can verify $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$ and

$$\nabla Q(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{A}(\mathbf{x} - \mathbf{x}^*),$$

then

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \\
=&\|\mathbf{x}_t - \eta\nabla Q(\mathbf{x}_t) - \mathbf{x}^*\|_2 \\
=&\|\mathbf{x}_t - \eta\mathbf{A}(\mathbf{x} - \mathbf{x}^*) - \mathbf{x}^*\|_2 \\
=&\|(\mathbf{I} - \eta\mathbf{A})(\mathbf{x} - \mathbf{x}^*)\|_2 \\
\le&\max\{|1-\eta\mu|, |1-\eta L|\}\|(\mathbf{x} - \mathbf{x}^*)\|_2 .
\end{aligned}$$

□

**Remark 5.1.** *Letting* $1 - \eta\mu = \eta L - 1$ *leads to* $\eta = 2/(L+\mu)$ *and* $\rho = (L-\mu)/(L+\mu) \approx 1 - 2/\kappa$*. Recall that for general strongly convex function, we set* $\eta = 1/L$ *and the decay coefficient is* $1 - 1/\kappa$*.*

**Theorem 5.2.** *Solving problem (12) in above theorem by Polyak's heavy ball method*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\nabla Q(\mathbf{x}_t) + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}),$$

*where* $\eta > 0$ *and* $\beta \in (0,1)$ *such that* $\beta \ge \max\{(1-\sqrt{\eta L})^2, (1-\sqrt{\eta\mu})^2\}$*. Then we have*

$$\begin{bmatrix} \mathbf{x}_{t+1} - \mathbf{x}^* \\ \mathbf{x}_t - \mathbf{x}^* \end{bmatrix} = \mathbf{M} \begin{bmatrix} \mathbf{x}_t - \mathbf{x}^* \\ \mathbf{x}_{t-1} - \mathbf{x}^* \end{bmatrix}.$$

*all* $t \ge 0$ *and some* $\mathbf{M}$ *with spectral radius of* $\beta$*.*

*Proof.* We have

$$\begin{aligned}
&\mathbf{x}_{t+1} - \mathbf{x}^* \\
=&\mathbf{x}_t - \eta\nabla Q(\mathbf{x}_t) + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}) - \mathbf{x}^* \\
=&\mathbf{x}_t - \eta\mathbf{A}(\mathbf{x}_t - \mathbf{x}^*) + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}) - \mathbf{x}^* \\
=&(\mathbf{I} - \eta\mathbf{A})(\mathbf{x}_t - \mathbf{x}^*) + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}) \\
=&(\mathbf{I} - \eta\mathbf{A})(\mathbf{x}_t - \mathbf{x}^*) + \beta(\mathbf{x}_t - \mathbf{x}^*) - \beta(\mathbf{x}_{t-1} - \mathbf{x}^*) \\
=&((1+\beta)\mathbf{I} - \eta\mathbf{A})(\mathbf{x}_t - \mathbf{x}^*) - \beta(\mathbf{x}_{t-1} - \mathbf{x}^*).
\end{aligned}$$

We present above result in matrix form as follows

$$\begin{bmatrix} \mathbf{x}_{t+1} - \mathbf{x}^* \\ \mathbf{x}_t - \mathbf{x}^* \end{bmatrix} = \begin{bmatrix} (1+\beta)\mathbf{I} - \eta\mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_t - \mathbf{x}^* \\ \mathbf{x}_{t-1} - \mathbf{x}^* \end{bmatrix}.$$

Then we study the eigenvalues of

$$\mathbf{M} = \begin{bmatrix} (1+\beta)\mathbf{I} - \eta\mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}.$$

Let $\mathbf{A}$ has eigenvalue decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ and define the orthogonal matrix

$$\mathbf{V} = \begin{bmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{bmatrix}.$$

Then we have

$$\begin{aligned}
&\mathbf{V}^\top\mathbf{M}\mathbf{V} = \begin{bmatrix} \mathbf{U}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{U}^\top \end{bmatrix} \begin{bmatrix} (1+\beta)\mathbf{I} - \eta\mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{bmatrix} \\
=& \begin{bmatrix} (1+\beta)\mathbf{U}^\top - \eta\mathbf{U}^\top\mathbf{A} & -\beta\mathbf{U}^\top \\ \mathbf{U}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{bmatrix} \\
=& \begin{bmatrix} (1+\beta)\mathbf{U}^\top\mathbf{U} - \eta\mathbf{U}^\top\mathbf{A}\mathbf{U} & -\beta\mathbf{U}^\top\mathbf{U} \\ \mathbf{U}^\top\mathbf{U} & \mathbf{0} \end{bmatrix} \\
=& \begin{bmatrix} (1+\beta)\mathbf{I} - \eta\mathbf{\Lambda} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}.
\end{aligned}$$

Recall that determinant will not be changed by multiply orthogonal matrix and change two rows (or column) only changes its sign. So, we can rearrange $\mathbf{V}^\top\mathbf{M}\mathbf{V}$ into block diagonal matrix and consider the block component

$$\mathbf{M}_2(\lambda_k) = \begin{bmatrix} 1+\beta-\eta\lambda_k & -\beta \\ \mathbf{I} & 0 \end{bmatrix},$$

where $\lambda_k$ is the $k$-th largest (absolute value) eigenvalue of $\mathbf{A}$. The eigenvalues of $\mathbf{M}_2(\lambda_k)$ are

$$\gamma_{k,1} = \frac{1}{2}\left(1+\beta-\eta\lambda_k + \sqrt{(1+\beta-\eta\lambda_k)^2 - 4\beta}\right) \text{ and } \gamma_{k,2} = \frac{1}{2}\left(1+\beta-\eta\lambda_k - \sqrt{(1+\beta-\eta\lambda_k)^2 - 4\beta}\right).$$

Since $\lambda_k \in [\mu, L]$, the condition on $\beta$ means $\beta \geq \left(1-\sqrt{\eta\lambda_k}\right)^2$, which implies

$$\begin{aligned}
&(1+\beta-\eta\lambda_k)^2 - 4\beta \\
\leq& (1+\beta-\eta\lambda_k - 2\sqrt{\beta})(1+\beta-\eta\lambda_k + 2\sqrt{\beta}) \\
\leq& \left((1-\sqrt{\beta}\,)^2 - \eta\lambda_k\right)\left((1+\sqrt{\beta}\,)^2 - \eta\lambda_k\right) \\
\leq& \left(1-\sqrt{\beta}-\sqrt{\eta\lambda_k}\,\right)\left(1-\sqrt{\beta}+\sqrt{\eta\lambda_k}\,\right)\left(1+\sqrt{\beta}-\sqrt{\eta\lambda_k}\,\right)\left(1+\sqrt{\beta}+\sqrt{\eta\lambda_k}\,\right) \\
\leq& \left((1-\sqrt{\eta\lambda_k})^2 - \beta\right)\left(1-\sqrt{\beta}+\sqrt{\eta\lambda_k}\,\right)\left(1+\sqrt{\beta}+\sqrt{\eta\lambda_k}\,\right) \leq 0,
\end{aligned}$$

where the last step is based on $\beta < 1$. Hence, we have

$$|\gamma_{k,1}| = |\gamma_{k,2}| = \frac{1}{2}\sqrt{(1+\beta-\eta\lambda_k)^2 + 4\beta - (1+\beta-\eta\lambda_k)^2} = \sqrt{\beta}.$$

□

**Remark 5.2.** *Let $\rho(\mathbf{A})$ be spectral radius of $\mathbf{A}$, then we have*

$$\lim_{k\to+\infty} \left\|\mathbf{A}^k\right\|_2^{1/k} = \rho(\mathbf{A}).$$

*For any $\epsilon > 0$, we define*

$$\mathbf{A}_+ = \frac{1}{\rho(\mathbf{A})+\epsilon}\mathbf{A} \qquad \textit{and} \qquad \mathbf{A}_- = \frac{1}{\rho(\mathbf{A})-\epsilon}\mathbf{A}.$$

*Then*

$$\rho(\mathbf{A}_+) = \frac{\rho(\mathbf{A})}{\rho(\mathbf{A}) + \epsilon} < 1 \qquad \textit{and} \qquad \rho(\mathbf{A}_-) = \frac{\rho(\mathbf{A})}{\rho(\mathbf{A}) - \epsilon} > 1,$$

*which means*

$$\lim_{k\to\infty} \mathbf{A}_+^k = \mathbf{0}.$$

*Hence, there exists some $N^+$ such that for all $k \geq N^+$, we have $\left\|\mathbf{A}_+^k\right\|_2 < 1$. Then we obtain*

$$\left\|\mathbf{A}^k\right\|_2 = \left\|(\rho(\mathbf{A}) + \epsilon)^k \mathbf{A}_+^k\right\|_2 = (\rho(\mathbf{A}) + \epsilon)^k \left\|\mathbf{A}_+^k\right\|_2 < (\rho(\mathbf{A}) + \epsilon)^k. \tag{13}$$

*Similarly, $\rho(\mathbf{A}_-) > 1$ means $\mathbf{A}_-^k$ is unbounded. Hence, there exists some $N^-$ such that for all $k \geq N^-$, we have $\left\|\mathbf{A}_-^k\right\|_2 > 1$. Then we obtain*

$$\left\|\mathbf{A}^k\right\|_2 = \left\|(\rho(\mathbf{A}) - \epsilon)^k \mathbf{A}_-^k\right\|_2 = (\rho(\mathbf{A}) - \epsilon)^k \left\|\mathbf{A}_-^k\right\|_2 > (\rho(\mathbf{A}) - \epsilon)^k.$$

*Combing above results, we have*

$$\lim_{k\to+\infty} \left\|\mathbf{A}^k\right\|_2^{1/k} = \rho(\mathbf{A}).$$

**Remark 5.3.** *For heavy ball method, we are interested in the bound (13). We define*

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{x}_{t+1} - \mathbf{x}^* \\ \mathbf{x}_t - \mathbf{x}^* \end{bmatrix}.$$

*Then for any $\epsilon > 0$, there exist $N^+ \in \mathbb{N}$ such that for all $t > N^+$, we have*

$$\|\mathbf{z}_t\|_2 = \left\|\mathbf{M}^t \mathbf{z}_0\right\|_2 \leq \left\|\mathbf{M}^t\right\|_2 \|\mathbf{z}_0\|_2 < (\rho(\mathbf{M}) + \epsilon)^t \|\mathbf{z}_0\|_2 .$$

*Let*

$$\eta = \left(\frac{2}{\sqrt{L} + \sqrt{\mu}}\right)^2,$$

*then we have*

$$\beta = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \approx 1 - \frac{2}{\sqrt{\kappa}}$$

*when $\kappa \gg 1$. The first-order oracle complexity to obtain $\|\mathbf{z}_t\|_2 \leq \epsilon$ is $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$.*

**Remark 5.4.** *Although the heavy ball method was stated for general nonlinear optimization by Polyak, only asymptotic convergence was proved.*

**Remark 5.5** (Strong Convexity to Non-Strong Convexity)**.** *We apply AGD to optimize*

$$\min_{\mathbf{x}\in\mathbb{R}^d} \hat{f}(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\delta}{2}\|\mathbf{x}\|_2^2 .$$

*Let the solution of above problem be $\hat{\mathbf{x}}^*$, then we have*

$$\begin{aligned} & f(\mathbf{x}_t) - \left(f(\mathbf{x}^*) + \frac{\delta}{2}\|\mathbf{x}^*\|_2^2\right) \\ \leq & f(\mathbf{x}_t) - \left(f(\hat{\mathbf{x}}^*) + \frac{\delta}{2}\|\hat{\mathbf{x}}^*\|_2^2\right) \end{aligned}$$

$$
\begin{aligned}
&\leq f(\mathbf{x}_t) + \frac{\delta}{2}\|\mathbf{x}_t\|_2^2 - \left(f(\hat{\mathbf{x}}^*) + \frac{\delta}{2}\|\hat{\mathbf{x}}^*\|_2^2\right) \\
&= \hat{f}(\mathbf{x}_t) - \hat{f}(\hat{\mathbf{x}}^*) \\
&\leq \left(1 - \sqrt{\frac{\delta}{L+\delta}}\right)^t \left(\hat{f}(\mathbf{x}_0) - \hat{f}(\hat{\mathbf{x}}^*) + \frac{\delta}{2}\|\mathbf{x}_0 - \hat{\mathbf{x}}^*\|_2^2\right),
\end{aligned}
$$

*which implies*

$$
f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \left(1 - \sqrt{\frac{\delta}{L+\delta}}\right)^t \left(\hat{f}(\mathbf{x}_0) - \hat{f}(\hat{\mathbf{x}}^*) + \frac{\delta}{2}\|\mathbf{x}_0 - \hat{\mathbf{x}}^*\|_2^2\right) + \frac{\delta}{2}\|\mathbf{x}^*\|_2^2.
$$

*Hence, setting $\delta = \mathcal{O}(\epsilon)$ and $t = \mathcal{O}\big(\sqrt{L/\epsilon}\log(1/\epsilon)\big)$ can find an $\epsilon$ suboptimal solution.*

**Definition 5.1.** *A pair of sequences $\{\phi_t : \mathbb{R}^d \to \mathbb{R}\}_{t=0}^{+\infty}$ and $\{\lambda_t \geq 0\}_{t=0}^{+\infty}$ is called an estimate sequence of function $f(\cdot)$ if*

$$
\lim_{t \to +\infty} \lambda_t = 0
$$

*and for any $\mathbf{x} \in \mathbb{R}^d$ and all $t \geq 0$ we have*

$$
\phi_t(\mathbf{x}) \leq (1 - \lambda_t) f(\mathbf{x}) + \lambda_t \phi_0(\mathbf{x}).
$$

**Lemma 5.1.** *We follow the notation of Definition 5.1. If we have*

$$
f(\mathbf{x}_{t+1}) \leq \phi_t^* \triangleq \min_{\mathbf{x} \in \mathbb{R}^d} \phi_t(\mathbf{x}),
$$

*for all $t \geq 0$, then $f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq \lambda_t(\phi_0(\mathbf{x}^*) - f(\mathbf{x}^*))$, where $\mathbf{x}^*$ is the minimizer of $f(\cdot)$.*

*Proof.* We have

$$
\begin{aligned}
&f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \\
\leq &\phi_t^* - f(\mathbf{x}^*) \\
\leq &\min_{\mathbf{x} \in \mathbb{R}^d} ((1 - \lambda_t) f(\mathbf{x}) + \lambda_t \phi_0(\mathbf{x})) - f(\mathbf{x}^*) \\
\leq &(1 - \lambda_t) f(\mathbf{x}^*) + \lambda_t \phi_0(\mathbf{x}^*) - f(\mathbf{x}^*) \\
\leq &\lambda_t(\phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)).
\end{aligned}
$$

□

**Remark 5.6.** *The convergence rate of sequence $\{\mathbf{x}_t\}$ can be formed by the convergence rate of $\{\lambda_t\}$.*

**Lemma 5.2.** *For $L$-smooth and $\mu$-strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$, we define $\mathbf{x}_{t+1} = \mathbf{y}_t - \eta \nabla f(\mathbf{y}_t)$ and*

$$
\psi_t(\mathbf{x}; \mathbf{y}_t) = f(\mathbf{x}_{t+1}) - \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{y}_t\|_2^2 + \frac{1}{\eta_t}\langle \mathbf{y}_t - \mathbf{x}_{t+1}, \mathbf{x} - \mathbf{x}_{t+1}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_t\|_2^2
$$

*for $\eta \leq 1/L$. Then it holds*

$$
\psi(\mathbf{x}; \mathbf{y}_t) \leq f(\mathbf{x}).
$$

*Proof.* We have

$$
\begin{aligned}
f(\mathbf{x}) \geq &f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x} - \mathbf{y}_t\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_t\|_2^2 \\
= &f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_{t+1} - \mathbf{y}_t\rangle + \langle \nabla f(\mathbf{y}_t), \mathbf{x} - \mathbf{x}_{t+1}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_t\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
\geq & f(\mathbf{x}_{t+1}) - \frac{1}{\eta_t}\|\mathbf{x}_{t+1} - \mathbf{y}_t\|_2^2 + \langle \nabla f(\mathbf{y}_t), \mathbf{x} - \mathbf{x}_{t+1}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_t\|_2^2 \\
= & f(\mathbf{x}_{t+1}) - \frac{1}{\eta_t}\|\mathbf{x}_{t+1} - \mathbf{y}_t\|_2^2 + \frac{1}{\eta_t}\langle \mathbf{y}_t - \mathbf{x}_{t+1}, \mathbf{x} - \mathbf{x}_{t+1}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_t\|_2^2 \\
= & \psi_t(\mathbf{x}; \mathbf{y}_t),
\end{aligned}
$$

where the inequality are based on the strong convexity and smoothness of $f$. □

**Remark 5.7.** *We can define an estimate sequence* $\{(\phi_t, \lambda_t)\}_{t=0}^{+\infty}$ *recursively as*

$$
\phi_t(\mathbf{x}) = (1-\theta_t)\phi_{t-1}(\mathbf{x}) + \theta_t \psi_t(\mathbf{x}; \mathbf{y}_t) \quad \textit{and} \quad \lambda_t = (1-\theta_t)\lambda_{t-1} \tag{14}
$$

*with*

$$
\phi_0(\mathbf{x}) = f(\mathbf{x}_0) + \frac{\gamma_0}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \qquad \textit{and} \qquad \lambda_0 = 1
$$

*for some* $\theta_t \in (0,1)$ *such that* $\lim_{t\to+\infty} \prod_{s=0}^{t}(1-\theta_s) = 0$. *We prove the property of* $\phi_t$ *by induction*

1. *For* $t = 0$, *we can verify* $\phi_0(\mathbf{x}) = (1-\lambda_0)f(\mathbf{x}) + \lambda_0\phi_0(\mathbf{x})$ *since* $\lambda_0 = 1$.
2. *Suppose we have* $\phi_{t-1}(\mathbf{x}) \leq (1-\lambda_{t-1})f(\mathbf{x}) + \lambda_{t-1}\phi_0(\mathbf{x})$ *for any* $\mathbf{x} \in \mathbb{R}^d$, *then*

$$
\begin{aligned}
\phi_t(\mathbf{x}) = & (1-\theta_t)\phi_{t-1}(\mathbf{x}) + \theta_t\psi_t(\mathbf{x}; \mathbf{y}_t) \\
\leq & (1-\theta_t)((1-\lambda_{t-1})f(\mathbf{x}) + \lambda_{t-1}\phi_0(\mathbf{x})) + \theta_t f(\mathbf{x}) \\
= & (1-\theta_t)(1-\lambda_{t-1})f(\mathbf{x}) + (1-\theta_t)\lambda_{t-1}\phi_0(\mathbf{x}) + \theta_t f(\mathbf{x}) \\
= & ((1-\theta_t)(1-\lambda_{t-1}) + \theta_t)f(\mathbf{x}) + (1-\theta_t)\lambda_{t-1}\phi_0(\mathbf{x}) \\
= & (1-\lambda_t)f(\mathbf{x}) + \lambda_t\phi_0(\mathbf{x}),
\end{aligned}
$$

*where the inequality is based on induction hypnosis and Lemma 5.2.*

**Remark 5.8.** *Then we only needs to find* $\theta_t$ *that guarantees* $f(\mathbf{x}_{t+1}) \leq \min_{\mathbf{x}\in\mathbb{R}^d}\phi_t(\mathbf{x})$. *We define*

$$
\mathbf{v}_t = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \phi_t(\mathbf{x}).
$$

*Note that the definition of* $\phi_t(\cdot)$ *means it must can be written as*

$$
\phi_t(\mathbf{x}) = \phi_t(\mathbf{v}_t) + \frac{\gamma_t}{2}\|\mathbf{x} - \mathbf{v}_t\|_2^2,
$$

*where* $\gamma_t = (1-\theta_t)\gamma_{t-1} + \theta_t\mu$. *Combining relation (14) with above expression (for* $t-1$*), we have*

$$
\begin{aligned}
\phi_t(\mathbf{x}) = & (1-\theta_t)\phi_{t-1}(\mathbf{x}) + \theta_t\psi_t(\mathbf{x}; \mathbf{y}_t) \\
= & (1-\theta_t)\left(\phi_{t-1}(\mathbf{v}_{t-1}) + \frac{\gamma_{t-1}}{2}\|\mathbf{x} - \mathbf{v}_{t-1}\|_2^2\right) + \theta_t\psi_t(\mathbf{x}; \mathbf{y}_t).
\end{aligned}
$$

*Since the minimizer of* $\phi_t(\cdot)$ *is* $\mathbf{v}_t$, *the first-order condition means*

$$
\begin{aligned}
\mathbf{0} = & \nabla\phi_t(\mathbf{v}_t) = (1-\theta_t)\gamma_{t-1}(\mathbf{v}_t - \mathbf{v}_{t-1}) + \theta_t\nabla\psi_t(\mathbf{v}_t; \mathbf{y}_t) \\
= & (1-\theta_t)\gamma_{t-1}(\mathbf{v}_t - \mathbf{v}_{t-1}) + \theta_t\nabla\psi_t(\mathbf{v}_t; \mathbf{y}_t) \\
= & (1-\theta_t)\gamma_{t-1}(\mathbf{v}_t - \mathbf{v}_{t-1}) + \frac{\theta_t}{\eta_t}(\mathbf{y}_t - \mathbf{x}_{t+1}) + \theta_t\mu(\mathbf{v}_t - \mathbf{y}_t) \\
\overset{(18)}{=} & \left(\frac{\theta_t^2}{\eta_t} - \theta_t\mu\right)(\mathbf{v}_t - \mathbf{v}_{t-1}) + \frac{\theta_t}{\eta_t}(\mathbf{y}_t - \mathbf{x}_{t+1}) + \theta_t\mu(\mathbf{v}_t - \mathbf{y}_t) \\
= & \frac{\theta_t^2}{\eta_t}\mathbf{v}_t - \left(\frac{\theta_t^2}{\eta_t} - \theta_t\mu\right)\mathbf{v}_{t-1} + \frac{\theta_t}{\eta_t}(\mathbf{y}_t - \mathbf{x}_{t+1}) - \theta_t\mu\mathbf{y}_t
\end{aligned} \tag{15}
$$

*which implies*

$$\begin{aligned}\mathbf{v}_t =& \frac{(\theta_t - \mu\eta_t)\,\mathbf{v}_{t-1} - (1-\mu\eta_t)\mathbf{y}_t + \mathbf{x}_{t+1}}{\theta_t} \\ =& \frac{(\theta_t - \mu\eta_t)\,\mathbf{v}_{t-1} - (1-\mu\eta_t)(\mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1})) + \mathbf{x}_{t+1}}{\theta_t}.\end{aligned} \tag{16}$$

*Now we use induction to show*

$$\mathbf{v}_t = \frac{1}{\theta_t}(\mathbf{x}_{t+1} - (1-\theta_t)\mathbf{x}_t).$$

*Suppose* $\mathbf{v}_{t-1} = \theta_{t-1}^{-1}(\mathbf{x}_t - (1-\theta_{t-1})\mathbf{x}_{t-1})$ *holds, then (16) and induction hypothesis implies*

$$\begin{aligned}\mathbf{v}_t =& \frac{(\theta_t - \mu\eta_t)\,\mathbf{v}_{t-1} - (1-\mu\eta_t)(\mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1})) + \mathbf{x}_{t+1}}{\theta_t} \\ =& \frac{\theta_{t-1}^{-1}\,(\theta_t - \mu\eta_t)\,(\mathbf{x}_t - (1-\theta_{t-1})\mathbf{x}_{t-1}) - (1-\mu\eta_t)(\mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1})) + \mathbf{x}_{t+1}}{\theta_t} \\ =& \frac{\theta_{t-1}^{-1}\,(\theta_t - \mu\eta_t)\,\mathbf{x}_t - (1-\mu\eta_t)(1+\beta_t)\mathbf{x}_t + \mathbf{x}_{t+1}}{\theta_t} \\ \overset{(20)}{=}& \frac{1}{\theta_t}(\mathbf{x}_{t+1} - (1-\theta_t)\mathbf{x}_t).\end{aligned} \tag{17}$$

*Above calculation have used the parameters setting*

$$\frac{\theta_t^2}{\eta_t} = \theta_t\mu + (1-\theta_t)\gamma_{t-1}, \tag{18}$$

$$\gamma_t = (1-\theta_t)\gamma_{t-1} + \theta_t\mu, \tag{19}$$

$$\beta_t = \frac{(\theta_t - \mu\eta_t)(1-\theta_{t-1})}{\theta_{t-1}(1-\mu\eta_t)}. \tag{20}$$

*From the third line of (15) and (19), we have*

$$\mathbf{x}_{t+1} - \mathbf{y}_t = \eta_t\mu(\mathbf{v}_t - \mathbf{y}_t) + a_t(\mathbf{v}_t - \mathbf{v}_{t-1}),$$

*where* $a_t = (\theta_t^{-1} - 1)\eta_t\gamma_{t-1}$. *The third line of (15) also implies*

$$-\|\mathbf{x}_{t+1} - \mathbf{y}_t\|_2^2 = -(\eta_t\mu + a_t)\left(\eta_t\mu\,\|\mathbf{v}_t - \mathbf{y}_t\|_2^2 - a_t\,\|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2^2\right) + \eta_t\mu a_t\,\|\mathbf{v}_{t-1} - \mathbf{y}_t\|_2^2. \tag{21}$$

*Finally, we prove the* $f(\mathbf{x}_t) \le \phi_t(\mathbf{v}_t)$ *by induction. We suppose* $f(\mathbf{x}_t) \le \phi_{t-1}(\mathbf{v}_{t-1})$, *then*

$$\begin{aligned}&\phi_t(\mathbf{v}_t) \\ =&(1-\theta_t)\left(\phi_{t-1}(\mathbf{v}_{t-1}) + \frac{\gamma_{t-1}}{2}\,\|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2^2\right) + \theta_t\psi_t(\mathbf{v}_t;\mathbf{y}_t) \\ \ge&(1-\theta_t)\left(f(\mathbf{x}_t) + \frac{\gamma_{t-1}}{2}\,\|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2^2\right) + \theta_t\psi_t(\mathbf{v}_t;\mathbf{y}_t) \\ \ge&(1-\theta_t)\left(\psi_t(\mathbf{x}_t;\mathbf{y}_t) + \frac{\gamma_{t-1}}{2}\,\|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2^2\right) + \theta_t\psi_t(\mathbf{v}_t;\mathbf{y}_t) \\ =&(1-\theta_t)\left(f(\mathbf{x}_{t+1}) - \frac{1}{2\eta_t}\,\|\mathbf{x}_{t+1} - \mathbf{y}_t\|_2^2 + \frac{1}{\eta_t}\langle\mathbf{y}_t - \mathbf{x}_{t+1}, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle + \frac{\mu}{2}\,\|\mathbf{x}_t - \mathbf{y}_t\|_2^2\right) \\ &+ \frac{(1-\theta_t)\gamma_{t-1}}{2}\,\|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2^2 \\ &+ \theta_t\left(f(\mathbf{x}_{t+1}) - \frac{1}{2\eta_t}\,\|\mathbf{x}_{t+1} - \mathbf{y}_t\|_2^2 + \frac{1}{\eta_t}\langle\mathbf{y}_t - \mathbf{x}_{t+1}, \mathbf{v}_t - \mathbf{x}_{t+1}\rangle + \frac{\mu}{2}\,\|\mathbf{v}_t - \mathbf{y}_t\|_2^2\right)\end{aligned}$$

$$
\begin{aligned}
&\overset{(17)}{=} f(\mathbf{x}_{t+1}) - \frac{1}{2\eta_t}\|\mathbf{y}_t - \mathbf{x}_{t+1}\|_2^2 + \frac{(1-\theta_t)\mu}{2}\|\mathbf{x}_t - \mathbf{y}_t\|_2^2 + \frac{\theta_t\lambda}{2}\|\mathbf{v}_t - \mathbf{y}_t\|_2^2 + \frac{(1-\theta_t)\gamma_{t-1}}{2}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2^2 \\
&\overset{(21)}{=} f(\mathbf{x}_{t+1}) + \frac{(1-\theta_t)\mu}{2}\|\mathbf{x}_t - \mathbf{y}_t\|_2^2 + \frac{(\theta_t^{-1}-1)\eta_t\gamma_{t-1}\mu}{2}\|\mathbf{v}_{t-1} - \mathbf{y}_t\|_2^2 \\
&\geq f(\mathbf{x}_{t+1}).
\end{aligned}
$$

*Hence, Lemma 5.1 implies the convergence.*

Recall the the general framework of Nesterov's acceleration

$$
\begin{cases}
\dfrac{\theta_t^2}{\eta_t} = \theta_t\mu + (1-\theta_t)\gamma_{t-1} \\
\gamma_t = (1-\theta_t)\gamma_{t-1} + \theta_t\mu \\
\beta_t = \dfrac{(\theta_t - \mu\eta_t)(1-\theta_{t-1})}{\theta_{t-1}(1-\mu\eta_t)} \\
\mathbf{y}_t = \mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1}) \\
\mathbf{x}_{t+1} = \mathbf{y}_t - \eta_t\nabla f(\mathbf{y}_t)
\end{cases}
$$

We have

$$
f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \lambda_t\left(f(\mathbf{x}_0) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right), \quad \text{where} \quad \lambda_t = \prod_{s=0}^{t}(1-\theta_s).
$$

1. For strongly convex case, we set $\eta_t = 1/L$, $\theta_t = \sqrt{\mu\eta}$ and $\gamma_0 = \mu$, then

$$
\gamma_t = \mu, \qquad \beta = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \quad \text{and} \qquad \lambda_t = \left(1 - \sqrt{\frac{\mu}{L}}\right)^t.
$$

2. For non-strongly convex case, we set $\mu = 0$ and $\gamma_0 = 1/\eta$. Then we have

$$
\theta_t^2 = (1-\theta_t)\gamma_{t-1}\eta_t \quad \text{and} \quad \gamma_t = (1-\theta_t)\gamma_{t-1},
$$

which implies $\theta_t^2 = \gamma_t\eta$. Thus $\theta_{t-1}^2 = \gamma_{t-1}\eta$ and

$$
\theta_t^2 = (1-\theta_t)\gamma_{t-1}\eta_t = (1-\theta_t)\theta_{t-1}^2.
$$

The setting $\theta_0 = 1$ leads to $\theta_t \leq 2/(t+2)$. We verify this result by induction.

(a) For $t = 0$, we have $\theta_0 = 1 = 2/(0+2)$.

(b) Suppose we have $\theta_{t-1} \leq 2/(t+1)$. The recursion means

$$
\theta_t^2 = (1-\theta_t)\theta_{t-1}^2 \leq \frac{4(1-\theta_t)}{(t+1)^2}.
$$

Thus, we have

$$
\theta_t \leq \frac{-4 + \sqrt{16 + 16(t+1)^2}}{2(t+1)^2} \leq \frac{2}{t+2}
$$

The last step can be checked by setting $x = t+1$ and

$$
\begin{aligned}
&\frac{-1+\sqrt{1+x^2}}{x^2} \leq \frac{1}{x+1} \iff -(x+1) + (x+1)\sqrt{1+x^2} \leq x^2 \\
\iff &(x+1)\sqrt{1+x^2} \leq x^2 + x + 1 \iff (x+1)^2(1+x^2) \leq (x^2+x+1)^2 \\
\iff &x^4 + 2x^3 + 2x^2 + 2x + 1 \leq x^4 + 2x^3 + 4x^2 + 2x + 1.
\end{aligned}
$$

Then we have

$$\lambda_t = (1-\theta_t)\lambda_{t-1} = \frac{\theta_t^2}{\theta_{t-1}^2}\lambda_{t-1}$$

which means

$$\lambda_t \le \theta_t^2 \le \frac{4}{(t+2)^2}.$$

Hence, we have

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \le \lambda_t \left(f(\mathbf{x}_0) - f(\mathbf{x}^*)\right) \le \frac{4(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{(t+2)^2} \le \frac{2L\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{(t+2)^2}$$

# 6 Lower Complexity Bound

**Assumption 6.1.** *An iterative method $\mathcal{M}$ generates a sequence of test points $\{\mathbf{x}_t\}$ such that*

$$\mathbf{x}_t \in \mathbf{x}_0 + \operatorname{span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_{k-1})\}.$$

**Remark 6.1.** *For AGD, we have*

$$\begin{aligned}
&\textit{(view as } \mathbf{x}_1\textit{)} \quad \mathbf{x}_1 = \mathbf{x}_0 - \eta_t \nabla f(\mathbf{x}_0),\\
&\textit{(view as } \mathbf{x}_2\textit{)} \quad \mathbf{y}_1 = \mathbf{x}_1 + \beta_t(\mathbf{x}_1 - \mathbf{x}_0) = -\beta_t \mathbf{x}_0 + (1+\beta_t)(\mathbf{x}_0 - \eta_t \nabla f(\mathbf{x}_0)),\\
&\textit{(view as } \mathbf{x}_2\textit{)} \quad \mathbf{x}_2 = \mathbf{y}_1 - \eta_t \nabla f(\mathbf{y}_1).
\end{aligned}$$

Consider the following functions

$$f_t(\mathbf{x}) = \frac{L}{4}\left(\frac{1}{2}\left(x_1^2 + \sum_{i=1}^{t-1}(x_i - x_{i+1})^2 + x_t^2\right) - x_1\right),$$

for $k = 1, \ldots, d$, where $\mathbf{x} = [x_1, \ldots, x_d]^\top$.

We can verify $\nabla^2 f(\mathbf{x}) = \frac{L}{4}\mathbf{A}_t$ with

$$\mathbf{A}_t = \begin{bmatrix}
2 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0\\
-1 & 2 & -1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0\\
0 & -1 & 2 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & 0 & \cdots & 2 & -1 & 0 & 0 & \cdots & 0\\
0 & 0 & 0 & \cdots & -1 & 2 & -1 & 0 & \cdots & 0\\
0 & 0 & 0 & \cdots & \cdots & -1 & 2 & 0 & \cdots & 0\\
0 & 0 & 0 & \cdots & \cdots & 0 & 0 & 0 & \cdots & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & 0 & \cdots & \cdots & 0 & 0 & 0 & \cdots & 0.
\end{bmatrix}$$

The quadratic function holds that

$$\langle \mathbf{s}, \nabla^2 f(\mathbf{x})\mathbf{s}\rangle = \frac{L}{4}\left(s_1^2 + \sum_{i=1}^{t-1}(s_i - s_{i+1})^2 + s_t^2\right) \ge \mathbf{0}$$

for all $\mathbf{x} \in \mathbb{R}^d$, where the first step is because of any quadratic function $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$ holds that

$$\mathbf{s}^\top \nabla^2 g(\mathbf{x})\mathbf{s} = \mathbf{s}^\top \mathbf{A}\mathbf{s} = 2g(\mathbf{s}) + 2\mathbf{b}^\top \mathbf{x}.$$

We also have

$$\langle \mathbf{s}, \nabla^2 f(\mathbf{x})\mathbf{s}\rangle \le \frac{L}{4}\left(s_1^2 + \sum_{i=1}^{t-1}(2s_i^2 + 2s_{i+1}^2) + s_t^2\right) \le L\left\|\mathbf{s}\right\|_2^2.$$

Hence, the function $f$ is convex and $L$-smooth.

The equation $\nabla f_t(\mathbf{x}) = \mathbf{0}$ is equivalent to $\nabla f_t(\mathbf{x}) = \mathbf{0}$, that is

$$\mathbf{A}_t\mathbf{x} - \mathbf{e}_1 = \mathbf{0} \Longleftrightarrow \begin{cases} 2x_1 - x_2 = 1 \\ -x_1 + 2x_2 - x_3 = 0 \\ \cdots \\ -x_{t-2} + 2x_{t-1} - x_t = 0 \\ -x_{t-1} + 2x_t = 0 \end{cases}$$

$$\Longleftrightarrow \begin{cases} x_1 = tx_t \\ x_2 = (t-1)x_t \\ \cdots \\ x_{t-2} = 3x_t \\ x_{t-1} = 2x_t \end{cases} \Longleftrightarrow \begin{cases} 1 = (t+1)x_t \\ x_1 = tx_t \\ x_2 = (k-1)x_t \\ \cdots \\ x_{t-2} = 3x_t \\ x_{t-1} = 2x_t \end{cases} \Longleftrightarrow x_i = \begin{cases} 1 - \dfrac{i}{t+1}, & i = 1, \ldots, t, \\ 0, & i = t+1 \ldots, d. \end{cases}$$

Then the optimal function value is

$$f_t^* = f(\mathbf{x}_t^*) = \frac{L}{4}\left(\frac{1}{2}\langle \mathbf{A}_t\mathbf{x}_t^*, \mathbf{x}_t^*\rangle - \langle \mathbf{e}_1, \mathbf{x}_t^*\rangle\right) = \frac{L}{4}\left(\frac{1}{2}\langle \mathbf{e}_1, \mathbf{x}_t^*\rangle - \langle \mathbf{e}_1, \mathbf{x}_t^*\rangle\right) = -\frac{L}{8}\left(1 - \frac{1}{t+1}\right).$$

We also note that

$$\begin{aligned} \left\|\mathbf{x}_t^*\right\|_2^2 &= \sum_{i=1}^{t}\left(1 - \frac{i}{t+1}\right)^2 = k - \frac{2}{t+1}\sum_{i=1}^{t} i + \frac{1}{(t+1)^2}\sum_{i=1}^{t} i^2 \\ &\le t - \frac{2t(t+1)}{2(t+1)} + \frac{t(t+1)(2t+1)}{6(t+1)^2} \le \frac{(t+1)^3}{3(t+1)^2} = \frac{t+1}{3}. \end{aligned} \tag{22}$$

**Lemma 6.1.** *Let $\mathbf{x}_0 = \mathbf{0}$. Then for any sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ satisfying the condition*

$$\mathbf{x}_t \in \mathcal{L}_t = \operatorname{span}\{\nabla f_t(\mathbf{x}_0), \ldots, \nabla f_t(\mathbf{x}_{t-1})\},$$

*we have $\mathcal{L}_t \subseteq \mathbb{R}^{t,d}$.*

**Lemma 6.2.** *For all $\mathbf{x} \in \mathbb{R}^{t,d}$, we have $f_t(\mathbf{x}) = f_p(\mathbf{x})$ for $p = t, t+1, \ldots, d$.*

*Proof.* We consider $p = t+1$ and $\mathbf{x} \in \mathbb{R}^{t,d}$. Let $\mathbf{x} = [x_1, \ldots, x_t, 0]^\top \in \mathbb{R}^{t+1}$. Then we have

$$\mathbf{x}^\top \mathbf{A}_{t+1}\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_t & 0 \end{bmatrix} \left( \mathbf{A}_t + \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix} \right) \begin{bmatrix} x_1 \\ \vdots \\ x_t \\ 0 \end{bmatrix}$$

and

$$\mathbf{x}^\top \mathbf{A}_{t+1}\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_t & 0 \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_t \\ 0 \end{bmatrix} = \begin{bmatrix} x_1 & \cdots & x_t & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -x_t \end{bmatrix} = 0.$$

□

**Corollary 6.1.** *For any $\{\mathbf{x}_t\}_{t=1}^p$ with $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{x}_t \in \mathcal{L}_t$, we have $\mathbf{x}_t \in \mathbb{R}^{t,d}$ and $f_p(\mathbf{x}_t) = f_t(\mathbf{x}_t) \geq f_t^*$ for any $p = t, t+1, \ldots, d$.*

**Theorem 6.1.** *For any $t$ such that $t \in [1, (d-1)/2]$ and any $\mathbf{x}_0 \in \mathbb{R}^d$, there exists an $L$-smooth and convex function $f$ such that for any first-order algorithm $\mathcal{M}$ with*

$$\mathbf{x}_t \in \mathbf{x}_0 + \operatorname{span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_{t-1})\},$$

*we have*

$$f(\mathbf{x}_t) - f^* \geq \frac{3L \left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{8(t+1)^2},$$

*where $\mathbf{x}^*$ is the minimizer of $f$ and $f^* = f(\mathbf{x}^*)$.*

*Proof.* Apply algorithm $\mathcal{M}$ to minimize

$$f(\mathbf{x}) \triangleq f_{2t+1}(\mathbf{x}),$$

which starts from $\mathbf{x}_0$ generate $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$. We suppose $\mathbf{x}_0 = \mathbf{0}$, otherwise we just need to consider minimize $f(\mathbf{x} + \mathbf{x}_0)$ with initial point $\mathbf{0}$.

Using Corollary 6.1 with $p = 2t+1$, we have

$$f_{2t+1}(\mathbf{x}_t) \geq f_t^* = -\frac{L}{8}\left(1 - \frac{i}{t+1}\right).$$

On the other hand, we have

$$f^* = f_{2t+1}^*(\mathbf{x}_t) = -\frac{L}{8}\left(1 - \frac{i}{2t+2}\right).$$

Combining inequality (22), we have

$$\frac{f(\mathbf{x}_t) - f^*}{\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2} \geq \frac{-\frac{L}{8}\left(1 - \frac{1}{t+1}\right) + \frac{L}{8}\left(1 - \frac{1}{2t+2}\right)}{\frac{2t+2}{3}} = \frac{3L}{8(t+1)^2}$$

□

**Remark 6.2.** *For any $\epsilon > 0$, there exists function $f : \mathbb{R}^d \to \mathbb{R}$ with $d = \Theta\left(\sqrt{L/\epsilon}\right)$ such that finding $\mathbf{x}$ with $f(\mathbf{x}) - f^* \leq \epsilon$ requires at least $\Omega\left(\sqrt{L/\epsilon}\right)$ iterations of first-order methods.*

Now we consider the lower complexity bound for minimizing strongly-convex function. We introduce

$$\begin{aligned} f(\mathbf{x}) =& \frac{L-\mu}{4}\left(\frac{1}{2}\left(x_1^2 + \sum_{i=1}^{d-1}(x_i - x_{i+1})^2 + \left(1 - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)x_d^2\right) - x_1\right) + \frac{\mu}{2}\left\|\mathbf{x}\right\|_2^2 \\ =& \frac{L-\mu}{4}\left(\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{e}_1^\top \mathbf{x}\right) + \frac{\mu}{2}\left\|\mathbf{x}\right\|_2 \end{aligned}$$

for $t = 1, \ldots, d$, where $\mathbf{x} = [x_1, \ldots, x_d]^\top$ and

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & \cdots & -1 & 2 - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \end{bmatrix}.$$

We show some properties of above function:

1. For any $\mathbf{s} \in \mathbb{R}^d$, we have

$$\begin{aligned}\langle \mathbf{s}, \nabla^2 f(\mathbf{x})\mathbf{s}\rangle =& \frac{L-\mu}{4}\left(s_1^2 + \sum_{i=1}^{d-1}(s_i - s_{i+1})^2 + \left(1 - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)s_d^2\right) + \mu\|\mathbf{s}\|_2^2 \\ \leq& \frac{L-\mu}{4}\left(s_1^2 + \sum_{i=1}^{d-1}(2s_i^2 + 2s_{i+1}^2) + \left(1 - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)s_d^2\right) + \mu\|\mathbf{s}\|_2^2 \\ \leq& (L-\mu)\|\mathbf{s}\|_2^2 + \mu\|\mathbf{s}\|_2^2 = L\|\mathbf{x}\|_2^2\end{aligned}$$

   and $\langle \mathbf{s}, \nabla^2 f(\mathbf{x})\mathbf{s}\rangle \geq \mu\|\mathbf{s}\|_2^2$. Hence, the function is $L$-smooth and $\mu$-strongly convex.

2. The optimal solution should satisfies

$$\left(\mathbf{A} + \frac{4}{\kappa-1}\mathbf{I}\right)\mathbf{x} = \mathbf{e}_1,$$

   which leads to

$$\begin{cases} \frac{2(\kappa+1)}{\kappa-1}x_1 - x_2 = 1 \\ -x_1 + \frac{2(\kappa+1)}{\kappa-1}x_2 - x_3 = 0 \\ \cdots\cdots \\ -x_{d-2} + \frac{2(\kappa+1)}{\kappa-1}x_{d-1} - x_d = 0 \\ -x_{d-1} + \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}x_d = 0 \end{cases} \Longrightarrow \begin{cases} \frac{2(\kappa+1)}{\kappa-1}x_1 - x_2 = 1 \\ x_i = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{i-1} x_1 \end{cases} \Longrightarrow x_i = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i = q^i.$$

Let $d = 2t$. Combining above results with zero-chain property, we have

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \geq \sum_{i=t+1}^{d}\|x_i^*\|_2^2 = \sum_{i=t+1}^{d} q^{2i} = \sum_{i=t+1}^{2t} q^{2i}.$$

On the other hand, we have

$$\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 = \sum_{i=1}^{d} q^{2i} = \sum_{i=1}^{2t} q^{2i}.$$

Finally, we achieve

$$\begin{aligned}\frac{f(\mathbf{x}_t) - f^*}{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2} \geq& \frac{\mu\|\mathbf{x}_t - \mathbf{x}^*\|_2^2}{2\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2} = \frac{\mu\sum_{i=t+1}^{2t} q^{2i}}{2\sum_{i=1}^{2t} q^{2i}} \\ =& \frac{\mu q^{2t}\sum_{i=1}^{t} q^{2i}}{2(1+q^{2t})\sum_{i=1}^{t} q^{2i}} = \frac{\mu q^{2t}}{2(1+q^{2t})} \\ =& \frac{\mu(\sqrt{\kappa}-1)^{2t}}{2((\sqrt{\kappa}+1)^{2t} + (\sqrt{\kappa}-1)^{2t})} \geq \frac{\mu}{4}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t}\end{aligned}$$

**Remark 6.3.** *For any $\epsilon > 0$, there exists function $f : \mathbb{R}^d \to \mathbb{R}$ with $d = \Theta\big(\sqrt{\kappa}\log(1/\epsilon)\big)$ such that finding $\mathbf{x}$ with $f(\mathbf{x}) - f^* \leq \epsilon$ requires at least $\Omega\big(\sqrt{\kappa}\log(1/\epsilon)\big)$ iterations of first-order methods.*

# 7 Nonsmooth Convex Optimization

**Example 7.1.** *Consider the function*

$$f(x) = |x|.$$

*The optimal solution is $x = 0$. For any constant learning rate $\eta_t = \eta$, if we take $x_0 \neq \eta/2$, then*

$$\mathbf{x}_1 = -\frac{\eta}{2}, \quad \mathbf{x}_2 = \frac{\eta}{2}, \quad \mathbf{x}_3 = -\frac{\eta}{2} \dots$$

*Therefore the algorithm does not converge with a constant step size.*

**Example 7.2.** *We consider the SVM formulation*

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \max\{1 - b_i \mathbf{a}_i^\top \mathbf{x}, 0\} + \frac{\lambda}{2}\|\mathbf{x}\|_2^2$$

*which is nonsmooth. The function $f$ is not Lipschitz globally over $\mathbb{R}^d$. However, assume that we start with $\mathbf{x}_0 = \mathbf{0}$ and consider the region matters for optimization:*

$$\mathcal{C} \triangleq \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \leq f(\mathbf{0})\} = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \leq 1\} \subseteq \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq \sqrt{\frac{2}{\lambda}}\right\}$$

*Then the function is Lipschitz in $\mathcal{C}$.*

**Theorem 7.1.** *We assume the convex function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies*

$$\max_{\mathbf{g}\in\partial f(\mathbf{x})} \{\|\mathbf{g}\|_2\} \leq G$$

*on convex and closed domain $\mathcal{C}$. Then for all $\hat{\mathbf{x}} \in \mathcal{C}$, the iteration*

$$\begin{cases} \tilde{\mathbf{x}}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}_t, \\ \mathbf{x}_{t+1} = \mathrm{proj}_{\mathcal{C}}\left(\tilde{\mathbf{x}}_{t+1}\right) \end{cases}$$

*for $t = 0, 1 \dots$ with $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ and*

$$\lim_{t\to+\infty} \eta_t = 0$$

*satisfies*

$$\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t f(\mathbf{x}_t) \leq f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \sum_{t=0}^{T-1} G^2\eta_t^2}{2\sum_{t=0}^{T-1}\eta_t}.$$

*Proof.* Given $\hat{\mathbf{x}} \in \mathcal{C}$, we have

$$\begin{aligned}
&\|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \\
=&\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_t + \mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\
=&\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|_2^2 + 2\langle\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_t, \mathbf{x}_t - \hat{\mathbf{x}}\rangle + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\
=&\eta_t^2\|\mathbf{g}_t\|_2^2 - 2\eta_t\langle\mathbf{g}_t, \mathbf{x}_t - \hat{\mathbf{x}}\rangle + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\
\leq&\eta_t^2 G^2 - 2\eta_t\langle\mathbf{g}_t, \mathbf{x}_t - \hat{\mathbf{x}}\rangle + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\
\leq&\eta_t^2 G^2 + 2\eta_t(f(\hat{\mathbf{x}}) - f(\mathbf{x}_t)) + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2,
\end{aligned}$$

where the first inequality is based on the bounded subgradient assumption and the second one use the definition of subgradient. Using Theorem 3.3, we obtain

$$\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 \le \|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \le \eta_t^2 G^2 + 2\eta_t (f(\hat{\mathbf{x}}) - f(\mathbf{x}_t)) + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 .$$

We sum above inequality over $t = 0, \ldots, T-1$ and obtain

$$0 \le \|\mathbf{x}_T - \hat{\mathbf{x}}\|_2^2 \le \sum_{t=0}^{T-1} \eta_t^2 G^2 + 2 \sum_{t=0}^{T-1} \eta_t (f(\hat{\mathbf{x}}) - f(\mathbf{x}_t)) + \|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 ,$$

which implies the desired result. □

**Remark 7.1.** *We consider two types of stepsizes to understand the convergence rate:*

1. *Taking $\eta_t = \eta_0/\sqrt{T}$ leads to*

$$\frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) \le f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \eta_0^2 G^2}{2\eta_0 \sqrt{T}}.$$

   *Hence, the complexity to find $\epsilon$-suboptimal solution requires $\mathcal{O}(1/\varepsilon^2)$ subgradient oracle complexity.*

2. *If we do not know $T$ a prior, we can take $\eta_t = \eta_0/(\sqrt{t+1} + \sqrt{t}\,)$, which leads to*

$$\sum_{t=0}^{T-1} \eta_t = \sum_{t=0}^{T-1} \frac{\eta_0}{\sqrt{t+1} + \sqrt{t}} = \eta_0 \sum_{t=0}^{T-1} \left( \sqrt{t+1} - \sqrt{t} \right) = \eta_0 \sqrt{T}.$$

   *and*

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta_t^2 &= \sum_{t=0}^{T-1} \frac{\eta_0^2}{\left(\sqrt{t+1} + \sqrt{t}\right)^2} \\
&\le \sum_{t=0}^{T-1} \frac{\eta_0^2}{2t+1} = \eta_0^2 + \eta_0^2 \sum_{t=1}^{T-1} \frac{1}{2t+1} \\
&\le \eta_0^2 + \eta_0^2 \int_0^{T-1} \frac{1}{2x+1} \, \mathrm{d}x \\
&= \eta_0^2 + \frac{\eta_0^2}{2} \ln(2x+1) \Big|_0^{T-1} \\
&= \eta_0^2 + \frac{\eta_0^2}{2} \ln(2T-1).
\end{aligned}$$

   *Combining above results and Theorem 7.1, we have*

$$\sum_{t=0}^{T-1} \frac{f(\mathbf{x}_t)}{\sqrt{T(t+1)} + \sqrt{Tt}} \le f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \eta_0^2 (\ln(2T-1) + 2) G^2/2}{2\eta_0 \sqrt{T}}.$$

**Theorem 7.2.** *Under the settings of Theorem 7.1, we suppose $f$ is $\mu$-strongly convex and set*

$$\eta_t = \frac{2}{\mu(t+1)}.$$

*Then*

$$\sum_{t=0}^{T-1} \frac{t}{T(T-1)} f(\mathbf{x}_t) \le f(\hat{\mathbf{x}}) + \frac{2G^2}{\mu(T-1)}.$$

*Proof.* We have

$$\begin{aligned}
&\langle \mathbf{g}_t, \mathbf{x}_t - \hat{\mathbf{x}} \rangle \\
=&\frac{1}{\eta_t} \langle \mathbf{x}_t - \tilde{\mathbf{x}}_{t+1}, \mathbf{x}_t - \hat{\mathbf{x}} \rangle \\
=&\frac{1}{2\eta_t} \left( \|\mathbf{x}_t - \tilde{\mathbf{x}}_{t+1}\|_2^2 + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \right) \\
=&\frac{1}{2\eta_t} \left( \eta_t^2 \|\mathbf{g}_t\|_2^2 + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \right) \\
\leq&\frac{1}{2\eta_t} \left( \eta_t^2 G^2 + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \right) \\
=&\frac{1}{2\eta_t} \left( \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \right) + \frac{\eta_t G^2}{2} \\
\leq&\frac{1}{2\eta_t} \left( \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 \right) + \frac{\eta_t G^2}{2}
\end{aligned}$$

where the last step is based on Theorem 3.3. Combining with the strong convexity, we obtain

$$\begin{aligned}
&f(\mathbf{x}_t) - f(\hat{\mathbf{x}}) \\
\leq&\langle \mathbf{g}_t, \mathbf{x}_t - \hat{\mathbf{x}} \rangle - \frac{\mu}{2} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\
\leq&\frac{1}{2\eta_t} \left( \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 \right) + \frac{\eta_t G^2}{2} - \frac{\mu}{2} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\
=&\frac{\mu(t+1)}{4} \left( \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 \right) + \frac{G^2}{\mu(t+1)} - \frac{\mu}{2} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\
=&\frac{\mu(t-1)}{4} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \frac{\mu(t+1)}{4} \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 + \frac{G^2}{\mu(t+1)},
\end{aligned}$$

which implies

$$\begin{aligned}
t(f(\mathbf{x}_t) - f(\hat{\mathbf{x}})) \leq&\frac{\mu(t-1)t}{4} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \frac{\mu t(t+1)}{4} \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 + \frac{G^2 t}{\mu(t+1)} \\
\leq&\frac{\mu(t-1)t}{4} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \frac{\mu t(t+1)}{4} \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 + \frac{G^2}{\mu}.
\end{aligned}$$

We sum over above inequality over $t = 0, \ldots, T-1$ and obtain

$$\sum_{t=0}^{T-1} t(f(\mathbf{x}_t) - f(\hat{\mathbf{x}})) \leq -\frac{\mu(T-1)T}{4} \|\mathbf{x}_T - \hat{\mathbf{x}}\|_2^2 + \frac{TG^2}{\mu} \leq \frac{TG^2}{\mu}$$

Hence, we have

$$\sum_{t=0}^{T-1} \frac{t}{T(T-1)} f(\mathbf{x}_t) \leq f(\hat{\mathbf{x}}) + \frac{2G^2}{\mu(T-1)}.$$

□

(Optimality of Subgradient Methods)

**Example 7.3.** *Consider the composite convex optimization problem*

$$\min_{\mathbf{x} \in \mathbb{R}^d} \phi(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}).$$

1. *Let $\mathcal{C}$ be a convex set and take $g(\mathbf{x}) = \mathbb{1}_{\mathcal{C}}(\mathbf{x})$. Then the problem is equivalent to*

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}).$$

2. *Let*

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \quad \textit{and} \quad g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1,$$

*then we obtain the Lasso problem.*

3. *Let $h(x) = \lambda|x|$. Consider the proximal operator*

$$\mathrm{prox}_h(x) = \arg\min_{z\in\mathbb{R}} \left(\frac{1}{2}(z-x)^2 + \lambda|x|\right).$$

*Given $x \in \mathbb{R}$, we have $z - x + \lambda\partial|z| = 0$, then*

   (a) *For $z > 0$, we have $z - x + \lambda = 0$, which means $z = x - \lambda > 0$. Hence, $x > \lambda$.*
   (b) *For $z < 0$, we have $z - x - \lambda = 0$, which means $z = x + \lambda < 0$. Hence, $x < -\lambda$.*
   (c) *For $z = 0$, we have $z - x - \lambda\partial|x| = 0$, which means $z \in [x-\lambda, x+\lambda]$. Hence, $x \in [-\lambda - z, \lambda - z]$.*

*In summary, we have*

$$\arg\min_{\mathbf{z}\in\mathbb{R}} \left(\frac{1}{2}(z-x)^2 + \lambda|x|\right) = \mathrm{sign}(x)\max\{|x|-\lambda, 0\}.$$

**Theorem 7.3.** *If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and G-Lipschitz continuous, then*

$$\tilde{f}(\mathbf{x}) = \min_{\mathbf{z}\in\mathbb{R}^d} \left(f(\mathbf{z}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|_2^2\right).$$

*is an $(L, G^2/(2L))$-smooth approximation of $f(\mathbf{x})$.*

*Proof.* We can write

$$\tilde{f}(\mathbf{x}) = \min_{\mathbf{z}\in\mathbb{R}^d} f(\mathbf{z}) + \frac{1}{2\gamma}\|\mathbf{z}-\mathbf{x}\|_2^2,$$

where $\gamma = 1/L$. We define

$$\mathrm{prox}_{\gamma g}(\mathbf{x}) = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \gamma f(\mathbf{z}) + \frac{1}{2}\|\mathbf{z}-\mathbf{x}\|_2^2.$$

1. The convexity can be proved by showing

$$f(\mathbf{z}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|_2^2$$

   is jointly convex of $\mathbf{x}$ and $\mathbf{z}$.

2. Now we prove $\tilde{f}$ is smooth and

$$\nabla\tilde{f}(\mathbf{x}) = \frac{\mathbf{x} - \mathrm{prox}_{\gamma g}(\mathbf{x})}{\gamma}. \tag{23}$$

   For any $\mathbf{x} \in \mathbb{R}^d$, the equation (23) is equivalent to

$$\lim_{\mathbf{y}\to\mathbf{x}} \frac{\tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) - \langle \mathbf{y}-\mathbf{x}, \frac{1}{\gamma}(\mathbf{x} - \mathrm{prox}_{\gamma g}(\mathbf{x}))\rangle}{\|\mathbf{y}-\mathbf{x}\|_2} = 0.$$

Let $\mathbf{u} = \text{prox}_{\gamma g}(\mathbf{x})$ and $\mathbf{v} = \text{prox}_{\gamma g}(\mathbf{y})$. The optimal condition means

$$\frac{1}{\gamma}(\mathbf{x} - \mathbf{u}) \in \partial f(\mathbf{u}) \qquad \text{and} \qquad \frac{1}{\gamma}(\mathbf{y} - \mathbf{v}) \in \partial f(\mathbf{v}). \tag{24}$$

Then

$$\begin{aligned}
&\tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) \\
=&f(\mathbf{v}) + \frac{1}{2\gamma}\|\mathbf{v} - \mathbf{y}\|_2^2 - \left(f(\mathbf{u}) + \frac{1}{2\gamma}\|\mathbf{u} - \mathbf{x}\|_2^2\right) \\
=&\frac{1}{2\gamma}\left(2\gamma(f(\mathbf{v}) - f(\mathbf{u})) + \|\mathbf{v} - \mathbf{y}\|_2^2 - \|\mathbf{u} - \mathbf{x}\|_2^2\right) \\
\geq&\frac{1}{2\gamma}\left(2\langle \mathbf{x} - \mathbf{u}, \mathbf{v} - \mathbf{u}\rangle + \|\mathbf{v} - \mathbf{y}\|_2^2 - \|\mathbf{u} - \mathbf{x}\|_2^2\right) \\
=&\frac{1}{2\gamma}\left(\|\mathbf{v} - \mathbf{y} - (\mathbf{u} - \mathbf{x})\|_2^2 + 2\langle \mathbf{y} - \mathbf{x}, \mathbf{x} - \mathbf{u}\rangle\right) \\
=&\frac{1}{\gamma}\langle \mathbf{y} - \mathbf{x}, \mathbf{x} - \mathbf{u}\rangle,
\end{aligned}$$

where the first inequality use (24). Swapping the roles of $\mathbf{x}$ and $\mathbf{y}$ leads to

$$\tilde{f}(\mathbf{x}) - \tilde{f}(\mathbf{y}) \geq \frac{1}{\gamma}\langle \mathbf{x} - \mathbf{y}, \mathbf{y} - \mathbf{v}\rangle \qquad \Longleftrightarrow \qquad \tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) \leq \frac{1}{\gamma}\langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{v}\rangle.$$

Combing above results, we have

$$\begin{aligned}
0 \leq&\tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) - \frac{1}{\gamma}\langle \mathbf{y} - \mathbf{x}, \mathbf{x} - \mathbf{u}\rangle \\
\leq&\frac{1}{\gamma}\left(\langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{v}\rangle - \langle \mathbf{y} - \mathbf{x}, \mathbf{x} - \mathbf{u}\rangle\right) \\
=&\frac{1}{\gamma}\left(\langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{v}\rangle - \langle \mathbf{y} - \mathbf{x}, \mathbf{x} - \mathbf{u}\rangle\right) \\
=&\frac{1}{\gamma}\left(\|\mathbf{y} - \mathbf{x}\|_2^2 - \langle \mathbf{y} - \mathbf{x}, \mathbf{v} - \mathbf{u}\rangle\right) \\
\leq&\frac{1}{\gamma}\|\mathbf{y} - \mathbf{x}\|_2^2,
\end{aligned}$$

where the last step is because of the result (24) leads to

$$\begin{cases} f(\mathbf{u}) \geq f(\mathbf{v}) + \frac{1}{\gamma}\langle \mathbf{y} - \mathbf{v}, \mathbf{u} - \mathbf{v}\rangle \\ f(\mathbf{v}) \geq f(\mathbf{u}) + \frac{1}{\gamma}\langle \mathbf{x} - \mathbf{u}, \mathbf{v} - \mathbf{u}\rangle \end{cases} \Longrightarrow \langle \mathbf{v} - \mathbf{u}, \mathbf{y} - \mathbf{x}\rangle \geq \|\mathbf{u} - \mathbf{v}\|_2^2 \geq 0.$$

This implies (23) and

$$0 \leq \tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) - \langle \mathbf{y} - \mathbf{x}, \tilde{\nabla} f(\mathbf{x})\rangle \leq L\|\mathbf{y} - \mathbf{x}\|_2^2,$$

3. For given $\mathbf{x} \in \mathbb{R}^d$, let

$$\mathbf{z}^* = \arg\min_{\mathbf{z} \in \mathbb{R}^d}\left(f(\mathbf{z}) + \frac{L}{2}\|\mathbf{z} - \mathbf{x}\|_2^2\right).$$

Hence, we have

$$\tilde{f}(\mathbf{x}) = f(\mathbf{z}^*) + \frac{L}{2}\|\mathbf{z}^* - \mathbf{x}\|_2^2$$

$$
\begin{aligned}
&\geq f(\mathbf{x}) - G\left\|\mathbf{x}-\mathbf{z}^*\right\|_2 + \frac{L}{2}\left\|\mathbf{z}^*-\mathbf{x}\right\|_2^2 \\
&\geq f(\mathbf{x}) + \frac{L}{2}\left(\left\|\mathbf{x}-\mathbf{z}^*\right\|_2 - \frac{G}{L}\right)^2 - \frac{G^2}{2L} \\
&\geq f(\mathbf{x}) - \frac{G^2}{2L}.
\end{aligned}
$$

□

**Example 7.4.** *Let $f(x) = |x|$ and*

$$\tilde{f}(x) = \min_{z\in\mathbb{R}} f(z) + \frac{1}{2\epsilon}(z-x)^2.$$

*Then we want to find $z$ such that*

$$\frac{1}{\epsilon}(x-z) \in \partial|z|.$$

1. *For $z > 0$, we have*

$$\frac{1}{\epsilon}(x-z) = 1 \quad \Longleftrightarrow \quad z = x - \epsilon > 0,$$

   *where $x > \epsilon$.*

2. *For $z < 0$, we have*

$$\frac{1}{\epsilon}(x-z) = -1 \quad \Longleftrightarrow \quad z = x + \epsilon < 0,$$

   *where $x < -\epsilon$.*

3. *For $z = 0$, we have*

$$\frac{1}{\epsilon}(x-z) \in [-1,1] \quad \Longleftrightarrow \quad x \in [-\epsilon, \epsilon].$$

*Then we obtain*

$$\tilde{f}(x) = \begin{cases} |x| - \dfrac{\epsilon}{2}, & |x| \geq \epsilon, \\ \dfrac{x^2}{2\epsilon}, & \textit{otherwise.} \end{cases}$$



**Proximal Gradient Method** For composite problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \phi(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}),$$

we can minimize RHS of

$$\phi(\mathbf{y}) = f(\mathbf{y}) + g(\mathbf{y}) \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y}-\mathbf{x}_t\rangle + \frac{L}{2}\left\|\mathbf{y}-\mathbf{x}_t\right\|_2^2 + g(\mathbf{y}),$$

which is

$$\arg\min_{\mathbf{y}\in\mathbb{R}^d} f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\left\|\mathbf{y}-\mathbf{x}_t\right\|_2^2 + g(\mathbf{y})$$

$$
\begin{aligned}
&= \arg\min_{\mathbf{y}\in\mathbb{R}^d} \frac{1}{L}\langle \nabla f(\mathbf{x}_t), \mathbf{y}-\mathbf{x}_t\rangle + \frac{1}{2}\|\mathbf{y}-\mathbf{x}_t\|_2^2 + \frac{1}{L}g(\mathbf{y}) \\
&= \arg\min_{\mathbf{y}\in\mathbb{R}^d} \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}_t-\frac{1}{L}\nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{1}{L}g(\mathbf{y}) \\
&= \mathrm{prox}_{\eta g}(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t))
\end{aligned}
$$

with $\eta = 1/L$.

**Gradient Mapping** The proximal gradient iteration can be written as

$$
\begin{aligned}
\mathbf{x}_{t+1} =& \mathrm{prox}_{\eta g}(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)) \\
=& \mathbf{x}_t - \eta\cdot\frac{\mathbf{x}_t - \mathrm{prox}_{\eta g}(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)))}{\eta} \\
=& \mathbf{x}_t - \eta\mathcal{G}_{\eta g, f}(\mathbf{x}_t).
\end{aligned}
$$

If $g(\mathbf{x}) = 0$, then $\mathcal{G}_{\eta g,f}(\mathbf{x}) = \nabla f(\mathbf{x})$.

**Lemma 7.1.** *We consider the composite convex problem*

$$
\min_{\mathbf{x}\in\mathbb{R}^d} \phi(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}),
$$

*where $f:\mathbb{R}^d\to\mathbb{R}$ is $L$-smooth and convex and $g:\mathbb{R}^d\to\mathbb{R}$ is convex but possibly nonsmooth. Let $\mathbf{x}^+ = \mathrm{prox}_{\eta g}(\mathbf{x}-\eta\nabla f(\mathbf{x}))$. Then we has the following results:*

1. *The point $\mathbf{x}^*$ is an optimal solution if and only if $\mathcal{G}_{\eta g,f}(\mathbf{x}^*) = \mathbf{0}$.*
2. *Suppose $g$ is $\mu_g$-strongly convex and $\eta < 2/(L-\mu)$, then*

$$
\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2 \le \frac{2/\eta}{2-\eta(L-\mu_g)}(\phi(\mathbf{x}) - \phi(\mathbf{x}^+)).
$$

3. *Suppose $\phi$ is $\mu_\phi$-strongly convex and $\eta \ge 1/L$, then*

$$
\phi(\mathbf{x}^+) \le \phi(\mathbf{x}^*) + \frac{1}{2\mu_\phi}\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2.
$$



*Proof.* **Part 1:** The definition of subgradient means $\mathbf{x}^*$ is an optimal solution if and only if there exists $\boldsymbol{\xi}^* \in \partial g(\mathbf{x}^*)$ such that

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

That is, at $\mathbf{z} = \mathbf{x}^*$, we have

$$
\mathbf{z} - (\mathbf{x}^* - \eta\nabla f(\mathbf{x}^*)) + \eta\boldsymbol{\xi}^* = \mathbf{0},
$$

which is equivalent to $\mathbf{z} = \mathbf{x}^*$ is the optimal solution of

$$
\min_{\mathbf{z}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{z} - (\mathbf{x}^* - \eta\nabla f(\mathbf{x}^*))\|_2^2 + \eta g(\mathbf{z}).
$$

Hence, we have $\mathbf{x}^* = \mathrm{prox}_{\eta g}(\mathbf{x}^* - \eta\nabla f(\mathbf{x}^*))$, which implies desired results.

**Part 2:** Let

$$
Q(\mathbf{z}) = f(\mathbf{x}) + \langle\nabla f(\mathbf{x}), \mathbf{z}-\mathbf{x}\rangle + \frac{1}{2\eta}\|\mathbf{z}-\mathbf{x}\|_2^2 + g(\mathbf{z}),
$$

then $\mathbf{x}^+$ is the solution of $\min_{\mathbf{z}\in\mathbb{R}^d} Q(\mathbf{z})$ and $Q$ is $(\eta^{-1}+\mu_g)$-strongly convex. This implies

$$Q(\mathbf{x}) - Q(\mathbf{x}^+) \geq \langle \mathbf{x}-\mathbf{x}^+, \partial Q(\mathbf{x}^+)\rangle + \frac{\eta^{-1}+\mu_g}{2}\left\|\mathbf{x}-\mathbf{x}^+\right\|_2^2 = \frac{\eta^{-1}+\mu_g}{2}\left\|\mathbf{x}-\mathbf{x}^+\right\|_2^2. \tag{25}$$

From the smoothness of $f$, we have

$$\begin{aligned}
&\phi(\mathbf{x}^+) = f(\mathbf{x}^+) + g(\mathbf{x}^+)\\
\leq& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x}\rangle + \frac{L}{2}\left\|\mathbf{x}^+-\mathbf{x}\right\|_2^2 + g(\mathbf{x}^+)\\
\leq& Q(\mathbf{x}^+) + \frac{L-\eta^{-1}}{2}\left\|\mathbf{x}^+-\mathbf{x}\right\|_2^2\\
\overset{(25)}{\leq}& Q(\mathbf{x}) + \frac{L-\mu_g-2\eta^{-1}}{2}\left\|\mathbf{x}^+-\mathbf{x}\right\|_2^2\\
=& \phi(\mathbf{x}) + \frac{(L-\mu_g)\eta^2-2\eta}{2}\left\|\mathcal{G}_{\eta g,f}(\mathbf{x})\right\|_2^2,
\end{aligned}$$

which implies the desired result.

**Part 3:** The optimality of $\mathbf{x}^+$ in the view of minimizing $Q(\mathbf{z})$ means there exists $\boldsymbol{\xi}^+ \in \partial g(\mathbf{x}^+)$ such that for all $\hat{\mathbf{x}} \in \mathcal{C}$, we have

$$\langle \nabla f(\mathbf{x}) + \eta^{-1}(\mathbf{x}^+ - \mathbf{x}) + \boldsymbol{\xi}^+, \hat{\mathbf{x}} - \mathbf{x}^+\rangle \geq 0.$$

This implies

$$\begin{aligned}
&\phi(\hat{\mathbf{x}}) - \phi(\mathbf{x}^+) - \frac{\mu_\phi}{2}\left\|\mathbf{x}^+-\hat{\mathbf{x}}\right\|_2^2\\
\geq& \langle \nabla f(\mathbf{x}^+) + \boldsymbol{\xi}^+, \hat{\mathbf{x}} - \mathbf{x}^+\rangle\\
=& \langle \nabla f(\mathbf{x}^+) - \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}^+\rangle + \langle \nabla f(\mathbf{x}^+) + \boldsymbol{\xi}^+, \hat{\mathbf{x}} - \mathbf{x}^+\rangle\\
\geq& \langle \nabla f(\mathbf{x}^+) - \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}^+\rangle + \eta^{-1}\langle \mathbf{x}-\mathbf{x}^+, \hat{\mathbf{x}} - \mathbf{x}^+\rangle\\
\geq& \langle \nabla \tilde{f}(\mathbf{x}^+) - \nabla \tilde{f}(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}^+\rangle\\
\geq& -\left\|\nabla \tilde{f}(\mathbf{x}^+) - \nabla \tilde{f}(\mathbf{x})\right\|_2 \left\|\hat{\mathbf{x}}-\mathbf{x}^+\right\|_2\\
\geq& -\eta^{-1}\left\|\mathbf{x}^+-\mathbf{x}\right\|_2 \left\|\hat{\mathbf{x}}-\mathbf{x}^+\right\|_2,
\end{aligned}$$

where $\tilde{f}(\mathbf{z}) = f(\mathbf{z}) - \frac{1}{2\eta}\left\|\mathbf{z}\right\|_2^2$ and we can show it is $\eta^{-1}$ smooth because the smoothness of $f$ means

$$\begin{aligned}
&0 \leq f(\mathbf{u}) - f(\mathbf{v}) - \langle \nabla f(\mathbf{v}), \mathbf{u}-\mathbf{v}\rangle \leq \frac{L}{2}\left\|\mathbf{u}-\mathbf{v}\right\|_2^2\\
\iff& -\frac{\eta^{-1}}{2}\left\|\mathbf{u}-\mathbf{v}\right\|_2^2 \leq f(\mathbf{u}) - \frac{\eta^{-1}}{2}\left\|\mathbf{u}\right\|_2^2 - \left(f(\mathbf{v}) - \frac{\eta^{-1}}{2}\left\|\mathbf{v}\right\|_2^2\right) - \left\langle \nabla f(\mathbf{v}) - \eta^{-1}\mathbf{v}, \mathbf{u}-\mathbf{v}\right\rangle \leq \frac{L-\eta^{-1}}{2}\left\|\mathbf{u}-\mathbf{v}\right\|_2^2\\
\iff& -\frac{\eta^{-1}}{2}\left\|\mathbf{u}-\mathbf{v}\right\|_2^2 \leq \tilde{f}(\mathbf{u}) - \tilde{f}(\mathbf{v}) - \langle \nabla \tilde{f}(\mathbf{v}), \mathbf{u}-\mathbf{v}\rangle \leq \frac{L-\eta^{-1}}{2}\left\|\mathbf{u}-\mathbf{v}\right\|_2^2.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\phi(\hat{\mathbf{x}}) - \phi(\mathbf{x}^+) \geq& \frac{\mu_\phi}{2}\left\|\mathbf{x}^+-\hat{\mathbf{x}}\right\|_2^2 - \eta^{-1}\left\|\mathbf{x}^+-\mathbf{x}\right\|_2\left\|\hat{\mathbf{x}}-\mathbf{x}^+\right\|_2\\
\geq& \inf_{\mathbf{z}\in\mathbb{R}^d}\left(\frac{\mu_\phi}{2}\left\|\mathbf{x}^+-\mathbf{z}\right\|_2^2 - \eta^{-1}\left\|\mathbf{x}^+-\mathbf{x}\right\|_2\left\|\mathbf{z}-\mathbf{x}^+\right\|_2\right)\\
=& -\frac{1}{2\mu_\phi\eta^2}\left\|\mathbf{x}^+-\mathbf{x}\right\|_2^2 = -\frac{1}{2\mu_\phi}\left\|\mathbf{x}^+-\mathbf{x}\right\|_2^2
\end{aligned}$$

□

**Remark 7.2.** *These results corresponds to property of $\nabla f(\mathbf{x})$ in convex optimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.*

1. *The optimal condition is* $\nabla f(\mathbf{x}^*) = \mathbf{0}$.
2. *Let* $\eta = 1/L$, *we have* $f(\mathbf{x}^+) \leq f(\mathbf{x}) - \frac{1}{2L}\|\nabla f(\mathbf{x})\|_2^2$.
3. *For strongly convex* $f$, *we have* $f(\mathbf{x}) \leq f(\mathbf{x}^*) + \frac{1}{2\mu}\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2$.

**Convergence Analysis of Proximal Gradient Method** We set $\eta = 1/L$. There are several results for different cases.

1. For strongly-convex case, we have

$$\|\mathcal{G}_{\eta g,f}(\mathbf{x}_t)\|_2^2 \leq 2L(\phi(\mathbf{x}_t) - \phi(\mathbf{x}_{t+1}))$$

and

$$\phi(\mathbf{x}_{t+1}) \leq \phi(\mathbf{x}^*) + \frac{1}{2\mu_\phi}\|\mathcal{G}_{\eta g,f}(\mathbf{x}_t)\|_2^2.$$

Thus, we obtain

$$\phi(\mathbf{x}_{t+1}) \leq \phi(\mathbf{x}^*) + \frac{1}{2\mu_\phi}\|\mathcal{G}_{\eta g,f}(\mathbf{x}_t)\|_2^2 \leq \phi(\mathbf{x}^*) + \frac{L}{\mu_\phi}(\phi(\mathbf{x}_t) - \phi(\mathbf{x}_{t+1})),$$

that is

$$\phi(\mathbf{x}_{t+1}) - \phi(\mathbf{x}^*) \leq \left(1 - \frac{\mu_\phi}{L + \mu_\phi}\right)(\phi(\mathbf{x}_t) - \phi(\mathbf{x}^*)).$$

2. For convex case, we first note that $\mathbf{x}^+ = \text{prox}_{\eta g}(\mathbf{x} - \eta\nabla f(\mathbf{x}))$ means

$$\mathbf{x}^+ - (\mathbf{x} - \eta\nabla f(\mathbf{x})) + \eta\boldsymbol{\xi}^+ = \mathbf{0} \iff \mathcal{G}_{\eta g,f}(\mathbf{x}) = \frac{\mathbf{x} - \mathbf{x}^+}{\eta} = \nabla f(\mathbf{x}) + \boldsymbol{\xi}^+.$$

Then for any $\mathbf{z} \in \mathbb{R}^d$, we have

$$\begin{aligned}
&\phi(\mathbf{x}^+) = \phi(\mathbf{x} - \eta\mathcal{G}_{\eta g,f}(\mathbf{x})) \\
=&f(\mathbf{x} - \eta\mathcal{G}_{\eta g,f}(\mathbf{x})) + g(\mathbf{x} - \eta\mathcal{G}_{\eta g,f}(\mathbf{x})) \\
\leq&f(\mathbf{x}) - \eta\langle\nabla f(\mathbf{x}), \mathcal{G}_{\eta g,f}(\mathbf{x})\rangle + \frac{L\eta^2}{2}\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2 + g(\mathbf{x} - \eta\mathcal{G}_{\eta g,f}(\mathbf{x})) \\
\leq&f(\mathbf{z}) + \langle\nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z}\rangle - \eta\langle\nabla f(\mathbf{x}), \mathcal{G}_{\eta g,f}(\mathbf{x})\rangle + \frac{L\eta^2}{2}\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2 + g(\mathbf{z}) - \langle\boldsymbol{\xi}^+, \mathbf{z} - (\mathbf{x} - \eta\mathcal{G}_{\eta g,f}(\mathbf{x}))\rangle \\
=&\phi(\mathbf{z}) + \langle\nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} - \eta\mathcal{G}_{\eta g,f}(\mathbf{x})\rangle + \frac{L\eta^2}{2}\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2 - \langle\boldsymbol{\xi}^+, \mathbf{z} - (\mathbf{x} - \eta\mathcal{G}_{\eta g,f}(\mathbf{x}))\rangle \\
=&\phi(\mathbf{z}) + \langle\nabla f(\mathbf{x}) + \boldsymbol{\xi}^+, \mathbf{x} - \mathbf{z} - \eta\mathcal{G}_{\eta g,f}(\mathbf{x})\rangle + \frac{L\eta^2}{2}\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2 \\
=&\phi(\mathbf{z}) + \langle\mathcal{G}_{\eta g,f}(\mathbf{x}), \mathbf{x} - \mathbf{z} - \eta\mathcal{G}_{\eta g,f}(\mathbf{x})\rangle + \frac{L\eta^2}{2}\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2 \\
=&\phi(\mathbf{z}) + \langle\mathcal{G}_{\eta g,f}(\mathbf{x}), \mathbf{x} - \mathbf{z}\rangle - \eta\left(1 - \frac{L\eta}{2}\right)\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2,
\end{aligned} \tag{26}$$

where the first inequality uses smoothness of $f$; the second inequality uses the convexity of $f$ and $g$. Applying (26) with $\mathbf{x}^+ = \mathbf{x}_{t+1}$, $\mathbf{x} = \mathbf{x}_t$, $\mathbf{z} = \mathbf{x}^*$ and $\eta = 1/L$, we achieve

$$\phi(\mathbf{x}_{t+1}) \leq \phi(\mathbf{x}^*) + \langle\mathcal{G}_{\eta g,f}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle - \eta\left(1 - \frac{L\eta}{2}\right)\|\mathcal{G}_{\eta g,f}(\mathbf{x}_t)\|_2^2$$

$$
\begin{aligned}
=&\phi(\mathbf{x}^*) + \langle \mathcal{G}_{\eta g,f}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle - \frac{1}{2L} \left\| \mathcal{G}_{\eta g,f}(\mathbf{x}_t) \right\|_2^2 \\
=&\phi(\mathbf{x}^*) + \frac{L}{2} \left( \left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \left\| \mathbf{x}_t - \frac{1}{L} \mathcal{G}_{\eta g,f}(\mathbf{x}_t) - \mathbf{x}^* \right\|_2^2 \right) \\
=&\phi(\mathbf{x}^*) + \frac{L}{2} \left( \left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \left\| \mathbf{x}_{t+1} - \mathbf{x}^* \right\|_2^2 \right)
\end{aligned}
$$

Summing over above inequality with $t = 0, \ldots, T-1$, we obtain

$$
\begin{aligned}
\phi(\mathbf{x}_T) \leq& \frac{1}{T} \sum_{t=1}^{T} \phi(\mathbf{x}_t) \\
\leq& \phi(\mathbf{x}^*) + \frac{L}{2T} \left( \left\| \mathbf{x}_0 - \mathbf{x}^* \right\|_2^2 - \left\| \mathbf{x}_T - \mathbf{x}^* \right\|_2^2 \right) \\
\leq& \phi(\mathbf{x}^*) + \frac{L}{2T} \left\| \mathbf{x}_0 - \mathbf{x}^* \right\|_2^2,
\end{aligned}
$$

where the first inequality is because of the second statement of Lemma 7.1 that says $\phi(\mathbf{x}_t)$ is non-decreasing.

3. If we only suppose $g$ is convex but allow $f$ be nonconvex, the second statement of Lemma 7.1 still holds with $\mu_g = 0$. Let $\eta = 1/L$, then it implies

$$
\left\| \mathcal{G}_{\eta g,f}(\mathbf{x}_t) \right\|_2^2 \leq 2L(\phi(\mathbf{x}_t) - \phi(\mathbf{x}_{t+1})).
$$

Summing over above inequality with $t = 0, \ldots, T-1$, we obtain

$$
\begin{aligned}
\mathbb{E} \left\| \mathcal{G}_{\eta g,f}(\hat{\mathbf{x}}) \right\|_2^2 =& \frac{1}{T} \sum_{t=0}^{T-1} \left\| \mathcal{G}_{\eta g,f}(\mathbf{x}_t) \right\|_2^2 \\
\leq& \frac{2L(\phi(\mathbf{x}_0) - \phi(\mathbf{x}_T))}{T} \\
\leq& \frac{2L(\phi(\mathbf{x}_0) - \phi^*)}{T},
\end{aligned}
$$

where $\hat{\mathbf{x}}$ is uniformly sampled from $\{\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}\}$.



# 8 Newton's Method

**Theorem 8.1.** *Suppose the twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ has $L_2$-Lipschitz continuous Hessian and local minimizer $\mathbf{x}^*$ with $\nabla^2 f(\mathbf{x}^*) \succeq \mu \mathbf{I}$, then the Newton's method*

$$
\mathbf{x}_{t+1} = \mathbf{x}_t - (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t)
$$

*with $\left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2 \leq \mu/(2L_2)$ holds that*

$$
\left\| \mathbf{x}_{t+1} - \mathbf{x}^* \right\|_2 \leq \frac{L_2}{\mu} \left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2.
$$

*Proof.* For any $\mathbf{x} \in \mathbb{R}^d$, we have

$$
\left\| \nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}^*) \right\|_2 \leq L_2 \left\| \mathbf{x} - \mathbf{x}^* \right\|_2
$$

which means

$$
|\lambda_i(\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}^*))| \leq L_2 \left\| \mathbf{x} - \mathbf{x}^* \right\|_2
$$

$$
\begin{aligned}
&\iff -L_2 \left\|\mathbf{x}-\mathbf{x}^*\right\|_2 \leq \lambda_i(\nabla^2 f(\mathbf{x})-\nabla^2 f(\mathbf{x}^*)) \leq L_2 \left\|\mathbf{x}-\mathbf{x}^*\right\|_2 \\
&\iff -L_2 \left\|\mathbf{x}-\mathbf{x}^*\right\|_2 \mathbf{I} \preceq \nabla^2 f(\mathbf{x})-\nabla^2 f(\mathbf{x}^*) \preceq L_2 \left\|\mathbf{x}-\mathbf{x}^*\right\|_2 \mathbf{I} \\
&\iff \nabla^2 f(\mathbf{x}^*)-L_2 \left\|\mathbf{x}-\mathbf{x}^*\right\|_2 \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L_2 \left\|\mathbf{x}-\mathbf{x}^*\right\|_2 \mathbf{I}+\nabla^2 f(\mathbf{x}^*) \\
&\Longrightarrow \nabla^2 f(\mathbf{x}) \succeq (\mu-L_2 \left\|\mathbf{x}-\mathbf{x}^*\right\|_2)\mathbf{I}.
\end{aligned}
$$

Hence, we have Taylor's expansion means

$$
\begin{aligned}
&\mathbf{x}_{t+1}-\mathbf{x}^* \\
=&\mathbf{x}_t-(\nabla^2 f(\mathbf{x}_t))^{-1}\nabla f(\mathbf{x}_t)-\mathbf{x}^* \\
=&\mathbf{x}_t-\mathbf{x}^*-(\nabla^2 f(\mathbf{x}_t))^{-1}\int_0^1 \nabla^2 f(\mathbf{x}^*+\tau(\mathbf{x}_t-\mathbf{x}^*))(\mathbf{x}_t-\mathbf{x}^*)\,\mathrm{d}\tau \\
=&(\nabla^2 f(\mathbf{x}_t))^{-1}\left(\nabla^2 f(\mathbf{x}_t)-\int_0^1 \nabla^2 f(\mathbf{x}^*+\tau(\mathbf{x}_t-\mathbf{x}^*))(\mathbf{x}_t-\mathbf{x}^*)\,\mathrm{d}\tau\right).
\end{aligned}
$$

Suppose that $\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2 \leq \mu/(2L_2)$, then we obtain

$$
\nabla^2 f(\mathbf{x}_t) \succeq (\mu-L_2 \left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2)\mathbf{I} \succeq \frac{\mu}{2}\mathbf{I}
$$

and

$$
\begin{aligned}
&\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2 \\
=&\left\|(\nabla^2 f(\mathbf{x}_t))^{-1}\left(\int_0^1 \left(\nabla^2 f(\mathbf{x}_t)-\nabla^2 f(\mathbf{x}^*+\tau(\mathbf{x}_t-\mathbf{x}^*))\right)(\mathbf{x}_t-\mathbf{x}^*)\,\mathrm{d}\tau\right)\right\|_2 \\
\leq&\left\|(\nabla^2 f(\mathbf{x}_t))^{-1}\right\|_2 \int_0^1 \left\|\nabla^2 f(\mathbf{x}_t)-\nabla^2 f(\mathbf{x}^*+\tau(\mathbf{x}_t-\mathbf{x}^*))\right\|_2 \left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2 \,\mathrm{d}\tau \\
\leq&\frac{2}{\mu}\int_0^1 L_2(1-\tau)\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2 \left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2 \,\mathrm{d}\tau \\
=&\frac{L_2}{\mu}\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2 \leq \frac{1}{2}\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2.
\end{aligned}
$$

Hence, the quadratic convergence holds if $\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2 \leq \mu/(2L_2)$. □

**Remark 8.1.** *The quadratic convergence means*

$$
\frac{L_2}{\mu}\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2 \leq \left(\frac{L_2}{\mu}\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2\right)^2 \quad\Longrightarrow\quad \frac{L_2}{\mu}\left\|\mathbf{x}_T-\mathbf{x}^*\right\|_2 \leq \left(\frac{L_2}{\mu}\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2\right)^{2^T}.
$$

*In the local region, Newton's method requires $T=\mathcal{O}(\ln\ln(1/\epsilon))$ iterations to achieve $\left\|\mathbf{x}_T-\mathbf{x}^*\right\|_2$. Even for $\epsilon=10^{-20}$, we have $\ln\ln(1/\epsilon)<4$.*

**Projected/Proximal Newton Methods** We consider

$$
\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}),
$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is strongly-convex function with Lipschitz continuous Hessian and $\mathcal{C}\subseteq\mathbb{R}^d$ is a convex set. Directly following projected gradient descent leads to

$$
\begin{cases}
\tilde{\mathbf{x}}_{t+1}=\mathbf{x}_t-(\nabla f(\mathbf{x}_t))^{-1}\nabla f(\mathbf{x}_t), \\
\mathbf{x}_{t+1}=\mathrm{proj}_{\mathcal{C}}(\tilde{\mathbf{x}}_{t+1}),
\end{cases}
\tag{27}
$$

which is not reasonable because of Newton's methods do not depends on Euclidean norm. The correct update should be

$$\begin{aligned}\mathbf{x}_{t+1} =& \arg\min_{\mathbf{x}\in\mathcal{C}} \left( f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{1}{2} \langle \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle \right) \\ =& \arg\min_{\mathbf{x}\in\mathcal{C}} \frac{1}{2} \left\| \mathbf{x} - \left( \mathbf{x}_t - (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t) \right) \right\|^2_{\nabla^2 f(\mathbf{x}_t)},\end{aligned}$$

which is the projection with respect to $\nabla^2 f(\mathbf{x}_t)$-norm. The proximal Newton methods is similar.

**Lemma 8.1.** *If $f : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-strongly convex and has $L_2$-Lipschitz continuous Hessian, then it is $M$-strongly self-concordant with $M = L_2/\mu^{3/2}$.*

*Proof.* The Lipschitz continuity of Hessian means

$$\left\| \nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y}) \right\|_2 \leq L_2 \left\| \mathbf{x} - \mathbf{y} \right\|_2^2$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, which means

$$\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y}) \preceq L_2 \left\| \mathbf{x} - \mathbf{y} \right\|_2 \mathbf{I} \preceq L_2 \sqrt{\left\langle \mathbf{x} - \mathbf{y}, \frac{1}{\mu} \nabla^2 f(\mathbf{z})(\mathbf{x} - \mathbf{y}) \right\rangle} \frac{\nabla^2 f(\mathbf{w})}{\mu} = \frac{L_2}{\mu^{3/2}} \left\| \mathbf{x} - \mathbf{y} \right\|_{\mathbf{z}} \nabla^2 f(\mathbf{w}),$$

for any $\mathbf{w}, \mathbf{z} \in \mathbb{R}^d$, where $\|\cdot\|_{\mathbf{z}}$ is the weighted norm with respect to $\nabla^2 f(\mathbf{z})$. □

**Affine Invariance** Consider function $\phi(\mathbf{x}) = f(\mathbf{A}\mathbf{y})$, where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is non-singular and $f : \mathbb{R}^d \to \mathbb{R}$. Let $\{\mathbf{x}_t\}$ be sequence, generated by

$$\mathbf{x}_{t+1} = \mathbf{x}_t - (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t).$$

Let $\{\mathbf{y}_t\}$ be sequence, generated by

$$\mathbf{y}_{t+1} = \mathbf{y}_t - (\nabla^2 \phi(\mathbf{y}_t))^{-1} \nabla \phi(\mathbf{y}_t).$$

Let $\mathbf{y}_t = \mathbf{A}^{-1}\mathbf{x}_t$, then

$$\begin{aligned}\mathbf{y}_{t+1} =& \mathbf{y}_t - (\nabla^2 \phi(\mathbf{y}_t))^{-1} \nabla \phi(\mathbf{y}_t) \\ =& \mathbf{y}_t - (\mathbf{A}^\top \nabla^2 f(\mathbf{A}\mathbf{y}_t)\mathbf{A})^{-1} \mathbf{A}^\top \nabla f(\mathbf{A}\mathbf{y}_t) \\ =& \mathbf{A}^{-1}\mathbf{x}_t - \mathbf{A}^{-1} (\nabla^2 f(\mathbf{x}_t))^{-1} \mathbf{A}^{-\top} \mathbf{A}^\top \nabla f(\mathbf{x}_t) \\ =& \mathbf{A}^{-1} \left( \mathbf{x}_t - (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t) \right) \\ =& \mathbf{A}^{-1} \mathbf{x}_{t+1}.\end{aligned}$$

If we run GD

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t) \quad \text{and} \quad \mathbf{y}_{t+1} = \mathbf{y}_t - \eta \nabla \phi(\mathbf{y}_t).$$

then

$$\begin{aligned}\mathbf{y}_{t+1} =& \mathbf{y}_t - \eta \nabla \phi(\mathbf{y}_t) \\ =& \mathbf{y}_t - \eta \mathbf{A}^\top \nabla f(\mathbf{A}\mathbf{y}_t) \\ =& \mathbf{A}^{-1} \left( \mathbf{A}\mathbf{y}_t - \eta \mathbf{A}\mathbf{A}^\top \nabla f(\mathbf{A}\mathbf{y}_t) \right) \\ =& \mathbf{A}^{-1} \left( \mathbf{x}_t - \eta \mathbf{A}\mathbf{A}^\top \nabla f(\mathbf{x}_t) \right) \neq \mathbf{A}^{-1} \mathbf{x}_{t+1}.\end{aligned}$$

The coefficient $L_2/\mu$ is not affine invariant, while the parameter of self-concordant is affine invariant.

**Lemma 8.2.** *If $f : \mathbb{R}^d \to \mathbb{R}$ is $M$-strongly self-concordant and $\phi(\mathbf{x}) = f(\mathbf{A}\mathbf{x})$ for some non-singular $\mathbf{A} \in \mathbb{R}^{d\times d}$, then $\phi$ is $M$-strongly self-concordant.*

*Proof.* For any $\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z} \in \mathbb{R}^d$, we let $\hat{\mathbf{z}} = \mathbf{A}\mathbf{z}$ and $\hat{\mathbf{w}} = \mathbf{A}\mathbf{w}$ . Then

$$\begin{aligned}
&\nabla^2\phi(\mathbf{x}) - \nabla^2\phi(\mathbf{y})\\
=&\mathbf{A}^\top \left(\nabla^2 f(\mathbf{A}\mathbf{x}) - \nabla^2 f(\mathbf{A}\mathbf{y})\right)\mathbf{A}\\
\preceq&\mathbf{A}^\top \left(M \left\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\right\|_{\nabla^2 f(\hat{\mathbf{z}})} \nabla^2 f(\hat{\mathbf{w}})\right)\mathbf{A}\\
=&M(\mathbf{x}-\mathbf{y})^\top \mathbf{A}^\top \nabla^2 f(\hat{\mathbf{z}})\mathbf{A}(\mathbf{x}-\mathbf{y})\mathbf{A}^\top \nabla^2 f(\hat{\mathbf{w}})\mathbf{A}\\
=&M(\mathbf{x}-\mathbf{y})^\top \mathbf{A}^\top \nabla^2 f(\mathbf{A}\mathbf{z})\mathbf{A}(\mathbf{x}-\mathbf{y})\mathbf{A}^\top \nabla^2 f(\mathbf{A}\mathbf{w})\mathbf{A}\\
=&M(\mathbf{x}-\mathbf{y})^\top \nabla^2\phi(\mathbf{z})(\mathbf{x}-\mathbf{y})\nabla^2\phi(\mathbf{w})\\
=&M\left\|\mathbf{x}-\mathbf{y}\right\|_{\nabla^2\phi(\mathbf{z})} \nabla^2\phi(\mathbf{w})
\end{aligned}$$

□

**Lemma 8.3.** *If $f : \mathbb{R}^d \to \mathbb{R}$ is $M$-strongly self-concordant, then we have*

$$|D^3 f(\mathbf{x})[\mathbf{h},\mathbf{h},\mathbf{h}]| \le M \left\|\mathbf{h}\right\|_{\mathbf{x}}^3$$

*for any $\mathbf{h} \in \mathbb{R}^d$, where $\|\cdot\|_{\mathbf{x}}$ is the weighted norm with respect to $\nabla^2 f(\mathbf{x})$.*

*Proof.* Recall the $M$-strongly self-concordant condition means we have

$$\nabla^2 f(\mathbf{y}) - \nabla^2 f(\mathbf{x}) \preceq M\left\|\mathbf{y}-\mathbf{x}\right\|_{\mathbf{z}} \nabla^2 f(\mathbf{w}),$$

for any $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w} \in \mathbb{R}^d$. For any $\mathbf{h} \in \mathbb{R}^d$, we have

$$\langle(\nabla^2 f(\mathbf{y}) - \nabla^2 f(\mathbf{x}))\mathbf{h},\mathbf{h}\rangle \le M\left\|\mathbf{y}-\mathbf{x}\right\|_{\mathbf{z}} \langle\nabla^2 f(\mathbf{w})\mathbf{h},\mathbf{h}\rangle.$$

Let $\mathbf{y} = \mathbf{x} + t\mathbf{h}$ and $\mathbf{w} = \mathbf{z} = \mathbf{x}$ for $t > 0$, then we achieve

$$\langle(\nabla^2 f(\mathbf{x}+t\mathbf{h}) - \nabla^2 f(\mathbf{x}))\mathbf{h},\mathbf{h}\rangle \le Mt\left\|\mathbf{h}\right\|_{\mathbf{x}} \langle\nabla^2 f(\mathbf{x})\mathbf{h},\mathbf{h}\rangle = Mt\left\|\mathbf{h}\right\|_{\mathbf{x}}^3,$$

that is

$$\frac{\langle(\nabla^2 f(\mathbf{x}+t\mathbf{h}) - \nabla^2 f(\mathbf{x}))\mathbf{h},\mathbf{h}\rangle}{t} \le M\left\|\mathbf{h}\right\|_{\mathbf{x}}^3.$$

We obtain the desired result by taking $t \to 0^+$. □

**Remark 8.2.** *We say $f : \mathbb{R}^d \to \mathbb{R}$ is $M_f$-self-concordant if*

$$|D^3 f(\mathbf{x})[\mathbf{h},\mathbf{h},\mathbf{h}]| \le 2M_f \left\|\mathbf{h}\right\|_{\mathbf{x}}^3$$

*for any $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$. This property is also affine invariant.*

**Remark 8.3.** *Theorem 2.1.1 of "Arkadii Nemirovskii, Yurii Nesterov. Interior-Point Polynomial Algorithms in Convex Programming, SIAM 1994" says the condition*

$$|D^3 f(\mathbf{x})[\mathbf{h},\mathbf{h},\mathbf{h}]| \le 2M_f \left\|\mathbf{h}\right\|_{\mathbf{x}}^3$$

*for any $\mathbf{h} \in \mathbb{R}^d$ is equivalent to*

$$|D^3 f(\mathbf{x})[\mathbf{h}_1,\mathbf{h}_2,\mathbf{h}_3]| \le 2M_f \prod_{i=1}^3 \left\|\mathbf{h}_i\right\|_{\mathbf{x}}^3 \tag{28}$$

*for any $\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3 \in \mathbb{R}^d$.*

**Global Convergence Analysis** We consider minimizing $M$-strongly self-concordant function with parameter $M = 2$. Then damped Newton method can be written as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{1 + \lambda_f(\mathbf{x}_t)} \big(\nabla^2 f(\mathbf{x}_t)\big)^{-1} \nabla f(\mathbf{x}_t).$$

We first consider one-dimensional case. For $M = 2$, Lemma 8.3 means

$$|f'''(t)| \le 2(f''(t))^{3/2} \quad \Longrightarrow \quad \left|\frac{\mathrm{d}(f''(t))^{-1/2}}{\mathrm{d}t}\right| = \frac{1}{2}\big|f'''(t)(f''(t))^{-3/2}\big| \le 1.$$

for any $t \in \mathbb{R}$. Taking integral over $t = 0$ to $v$ ($v > 0$), we achieve

$$|f''(v)^{-1/2} - f''(0)^{-1/2}| \le v \quad \Longrightarrow \quad -v \le f''(v)^{-1/2} - f''(0)^{-1/2} \le v.$$

Suppose that $1 - vf''(0)^{1/2} > 0$, then

$$\frac{f''(0)}{(1 + vf''(0)^{1/2})^2} \le f''(v) \le \frac{f''(0)}{(1 - vf''(0)^{1/2})^2}. \tag{29}$$

Considering that

$$\int \frac{f''(0)}{(1 - vf''(0)^{1/2})^2}\,\mathrm{d}v = \frac{f''(0)^{1/2}}{1 - vf''(0)^{1/2}}$$

and

$$\int \frac{f''(0)}{(1 + vf''(0)^{1/2})^2}\,\mathrm{d}v = -\frac{f''(0)^{1/2}}{1 + vf''(0)^{1/2}}.$$

Hence, taking integral over $v = 0$ to $u$ ($u > 0$) leads to

$$-\frac{f''(0)^{1/2}}{1 + uf''(0)^{1/2}} + f''(0)^{1/2} \le f'(u) - f'(0) \le \frac{f''(0)^{1/2}}{1 - uf''(0)^{1/2}} - f''(0)^{1/2}.$$

Considering that

$$\int \left(-\frac{f''(0)^{1/2}}{1 + uf''(0)^{1/2}} + f''(0)^{1/2}\right)\mathrm{d}u = -f''(0)^{1/2}\ln(1 + f''(0)^{1/2}u) + f''(0)^{1/2}u = \rho(-f''(0)^{1/2}u)$$

and

$$\int \left(-\frac{f''(0)^{1/2}}{1 + uf''(0)^{1/2}} + f''(0)^{1/2}\right)\mathrm{d}u = -f''(0)^{1/2}\ln(1 - f''(0)^{1/2}x) - f''(0)^{1/2}x = \rho(f''(0)^{1/2}u),$$

where $\rho(z) = -\ln(1 - z) - z$. Taking integral over $u = 0$ to $x$ leads to

$$\rho(-f''(0)^{1/2}x) \le f(x) - f(0) - f'(0)x = \rho(f''(0)^{1/2}x), \tag{30}$$

since $\rho(0) = 0$.

**Lemma 8.4** (Homework)**.** *If $f : \mathbb{R}^d \to \mathbb{R}$ is self-concordant, then $a(t) = f(\mathbf{x} + t\Delta)$ is a one-dimensional self-concordant function for given $\Delta \in \mathbb{R}^d$.*

**Remark 8.4.** *We have $\rho(z) \sim z^2/2$ when $z \to 0$.*

**Lemma 8.5.** *For 1-strongly self-concordant function $f(\mathbf{x})$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that*

$$\delta = \sqrt{(\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x})} < 1,$$

*then*

$$\rho(-\delta) \le f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \le \rho(\delta), \tag{31}$$

$$(1-\delta)^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq \frac{1}{(1-\delta)^2}\nabla^2 f(\mathbf{x}), \tag{32}$$

$$\left\|\nabla f(\mathbf{x})^{-1/2}\big(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})\big)\right\|_2 \le \frac{\delta^2}{1-\delta}. \tag{33}$$

*Proof.* We prove these results as follows.

1. Let $a(t) = f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$, then
$$a'(t) = \langle \mathbf{y}-\mathbf{x}, \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))\rangle \quad \text{and} \quad a''(t) = \langle \mathbf{y}-\mathbf{x}, \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x})\rangle.$$
Then $a''(0) = \delta^2$ and using (30) (with $f(\cdot) = a(\cdot)$ and $x = 1$) leads to
$$\begin{aligned}&\rho(-a''(0)^{1/2}) \le a(1) - a(0) - a'(0) = \rho(a''(0)^{1/2})\\ \Longleftrightarrow &\rho(-\delta) \le f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle = \rho(\delta).\end{aligned}$$

2. Let $b(t) = \mathbf{h}^\top \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))\mathbf{h}$. We have $b'(t) = D^3 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))[\mathbf{h},\mathbf{h},\mathbf{y}-\mathbf{x}]$ and
$$\begin{aligned}|b'(t)| \le& 2\mathbf{h}^\top \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))\mathbf{h}\sqrt{(\mathbf{y}-\mathbf{x})^\top \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x})}\\ =& 2b(t)a''(t)^{1/2} \le \frac{2b(t)\delta}{1-t\delta},\end{aligned}$$
where the first inequality use (28); the second inequality use (29) with $a(\cdot) = f(\cdot)$, $v = t$ and $a''(0) = \delta^2$. This implies for $\delta t \in [0,1)$, we have
$$\frac{\mathrm{d}(1-\delta t)^2 b(t)}{\mathrm{d}t} = -2\delta(1-\delta t)b(t) + (1-\delta t)^2 b'(t) \le 0,$$
which implies $(1-\delta t)^2 b(t)$ is non-increasing in $t$. Hence we have
$$(1-\delta t)^2 b(t) \le b(0) \quad \Longrightarrow \quad b(t) \le \frac{b(0)}{(1-\delta t)^2} = \frac{\mathbf{h}^\top \nabla^2 f(\mathbf{x})\mathbf{h}}{(1-\delta t)^2} \tag{34}$$
for any $\mathbf{h} \in \mathbb{R}^d$, which means RHS of the result. The LHS can be proved similarly.

3. Let $g(t) = \langle \mathbf{h}, \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))\rangle$, for which, $g'(t) = \langle \mathbf{h}, \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x})\rangle$, and
$$g''(t) = D^3 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))[\mathbf{h},\mathbf{y}-\mathbf{x},\mathbf{y}-\mathbf{x}].$$
Inequality (28) means
$$|g''(t)| \le 2\langle \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle\sqrt{\langle \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))\mathbf{h},\mathbf{h}\rangle} = 2a''(t)b(t)^{1/2},$$
where the inequality use equation (28). Combing above inequality with inequality (29) (with $a(\cdot) = f(\cdot)$, $v = t$ and $a''(0) = \delta^2$) and (34), we achieve
$$|g''(t)| \le \frac{2\delta^2}{(1-\delta t)^2}\sqrt{\frac{\mathbf{h}^\top \nabla^2 f(\mathbf{x})\mathbf{h}}{(1-\delta t)^2}} = \frac{2\delta^2}{(1-\delta t)^3}(\mathbf{h}^\top \nabla^2 f(\mathbf{x})\mathbf{h})^{1/2}.$$
We can then integrate above one twice, using $g(0) = \langle \mathbf{h}, \nabla f(\mathbf{x})\rangle$ and $g'(0) = \langle \mathbf{h}, \nabla f(\mathbf{x})(\mathbf{y}-\mathbf{x})\rangle$, to get
$$\begin{aligned}&\langle \mathbf{h}, \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})\rangle \le (\mathbf{h}^\top \nabla^2 f(\mathbf{x})\mathbf{h})^{1/2}\frac{\delta^2}{1-\delta}\\ \Longleftrightarrow &\frac{\langle \mathbf{h}, \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})\rangle}{(\mathbf{h}^\top \nabla^2 f(\mathbf{x})\mathbf{h})^{1/2}} \le \frac{\delta^2}{1-\delta}.\end{aligned}$$
Taking
$$\mathbf{h} = \big(\nabla^2 f(\mathbf{x})\big)^{-1}\big(\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})\big),$$
we prove the desired result.

**Then we show the main result.** Consider the iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{1+\lambda_f(\mathbf{x}_t)}\left(\nabla^2 f(\mathbf{x}_t)\right)^{-1}\nabla f(\mathbf{x}_t),$$

where

$$\lambda_f(\mathbf{x}_t) = \sqrt{\left\langle \nabla f(\mathbf{x}_t), \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1}\nabla f(\mathbf{x}_t)\right\rangle}.$$

Using inequality (31) with $\mathbf{y} = \mathbf{x}_{t+1}$, $\mathbf{x} = \mathbf{x}_t$ and

$$\delta^2 = (\mathbf{x}_{t+1}-\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1}-\mathbf{x}_t) = \frac{\nabla f(\mathbf{x}_t)^\top (\nabla^2 f(\mathbf{x}_t))^{-1}\nabla f(\mathbf{x}_t)}{(1+\lambda_f(\mathbf{x}_t))^2} = \frac{(\lambda_f(\mathbf{x}_t))^2}{(1+\lambda_f(\mathbf{x}_t))^2} < 1.$$

Then we have

$$\begin{aligned}
&f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)\\
\leq&\langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1}-\mathbf{x}_t\rangle + \rho(\delta)\\
=&-\frac{\nabla f(\mathbf{x}_t)^\top (\nabla^2 f(\mathbf{x}_t))^{-1}\nabla f(\mathbf{x}_t)}{1+\lambda_f(\mathbf{x}_t)} + \rho\left(\frac{\lambda_f(\mathbf{x}_t)}{1+\lambda_f(\mathbf{x}_t)}\right)\\
=&-\frac{(\lambda_f(\mathbf{x}_t))^2}{1+\lambda_f(\mathbf{x}_t)} - \ln\left(1-\frac{\lambda_f(\mathbf{x}_t)}{1+\lambda_f(\mathbf{x}_t)}\right) - \frac{\lambda_f(\mathbf{x}_t)}{1+\lambda_f(\mathbf{x}_t)}\\
=&-\lambda_f(\mathbf{x}_t) + \ln\left(1+\lambda_f(\mathbf{x}_t)\right).
\end{aligned}$$

1. For $\lambda_f(\mathbf{x}_t) \geq 1/4$, we have

   $$-\lambda_f(\mathbf{x}_t) + \ln\left(1+\lambda_f(\mathbf{x}_t)\right) \leq -\frac{1}{4} + \ln\left(\frac{5}{4}\right) \leq -0.0268 < \frac{1}{38}.$$

   Suppose $\lambda_f(\mathbf{x}_t) \geq 1/4$ holds for $t = 0, \ldots, t_0$, then

   $$f(\mathbf{x}_{t_0}) \leq f(\mathbf{x}_0) - \frac{t_0}{38},$$

   which means

   $$t_0 \leq 38(f(\mathbf{x}_0) - f(\mathbf{x}_{t_0})) \leq 38(f(\mathbf{x}_0) - f^*).$$

   Hence, the period of $\lambda_f(\mathbf{x}_t) \geq 1/4$ has at most constant iteration.

2. For $\lambda_f(\mathbf{x}_t) < 1/4$, inequality (32) with $\mathbf{x} = \mathbf{x}_t$ and $\mathbf{y} = \mathbf{x}_{t+1}$ means

   $$\begin{aligned}
   \lambda_f(\mathbf{x}_{t+1}) =&\sqrt{\left\langle \nabla f(\mathbf{x}_{t+1}), \left(\nabla^2 f(\mathbf{x}_{t+1})\right)^{-1}\nabla f(\mathbf{x}_{t+1})\right\rangle}\\
   \leq&\frac{1}{1-\delta}\sqrt{\left\langle \nabla f(\mathbf{x}_{t+1}), \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1}\nabla f(\mathbf{x}_{t+1})\right\rangle}\\
   =&\frac{1}{1-\delta}\left\|\left(\nabla^2 f(\mathbf{x}_t)\right)^{-1/2}\nabla f(\mathbf{x}_{t+1})\right\|_2.
   \end{aligned}$$

   Using inequality (33) with $\mathbf{x} = \mathbf{x}_t$ and $\mathbf{y} = \mathbf{x}_{t+1}$, we achieve

   $$\begin{aligned}
   &\left\|\left(\nabla^2 f(\mathbf{x}_t)\right)^{-1/2}\nabla f(\mathbf{x}_{t+1})\right\|_2\\
   \leq&\left\|\left(\nabla^2 f(\mathbf{x}_t)\right)^{-1/2}\left(\nabla f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1}-\mathbf{x}_t)\right)\right\|_2 + \frac{\delta^2}{1-\delta}\\
   =&\left\|\left(\nabla^2 f(\mathbf{x}_t)\right)^{-1/2}\left(\nabla f(\mathbf{x}_t) - \frac{1}{1+\lambda_f(\mathbf{x}_t)}\nabla f(\mathbf{x}_t)\right)\right\|_2 + \frac{\delta^2}{1-\delta}
   \end{aligned}$$

$$
\begin{aligned}
=&\frac{\lambda_f(\mathbf{x}_t)}{1+\lambda_f(\mathbf{x}_t)}\left\|\left(\nabla^2 f(\mathbf{x}_t)\right)^{-1/2}\nabla f(\mathbf{x}_t)\right\|_2+\frac{\delta^2}{1-\delta}\\
=&\frac{(\lambda_f(\mathbf{x}_t))^2}{1+\lambda_f(\mathbf{x}_t)}+\frac{\delta^2}{1-\delta}.
\end{aligned}
$$

Combining above inequalities, we have

$$
\lambda_f(\mathbf{x}_{t+1}) \leq \frac{(\lambda_f(\mathbf{x}_t))^2}{(1+\lambda_f(\mathbf{x}_t))(1-\delta)}+\frac{\delta^2}{(1-\delta)^2}=2(\lambda_f(\mathbf{x}_t))^2,
$$

where the last step is based on $\delta=\lambda_f(\mathbf{x}_t)/(1+\lambda_f(\mathbf{x}_t))$.

In summary, we require

$$
38(f(\mathbf{x}_0)-f^*)+2\ln\ln\left(\frac{1}{\epsilon}\right)
$$

iterations to find $\mathbf{x}_t$ such that $\lambda_f(\mathbf{x}_t)\leq\epsilon$. □

# 9 Quasi-Newton Methods

**Secant Condition** For general $f(\mathbf{x})$ with $L_2$-Lipschitz continuous Hessian, we have

$$
\begin{aligned}
&\nabla f(\mathbf{x}_{t+1})-\nabla f(\mathbf{x}_t)\\
=&\int_0^1 \nabla^2 f(\mathbf{x}_t+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))(\mathbf{x}_{t+1}-\mathbf{x}_t)\,\mathrm{d}\tau\\
=&\nabla^2 f(\mathbf{x}_{t+1})(\mathbf{x}_{t+1}-\mathbf{x}_t)+\int_0^1 \nabla^2(f(\mathbf{x}_t+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))-\nabla^2 f(\mathbf{x}_{t+1}))(\mathbf{x}_{t+1}-\mathbf{x}_t)\,\mathrm{d}\tau\\
=&\nabla^2 f(\mathbf{x}_{t+1})(\mathbf{x}_{t+1}-\mathbf{x}_t)+o(\|\mathbf{x}_{t+1}-\mathbf{x}_t\|_2),
\end{aligned}
$$

where the last step is because of

$$
\begin{aligned}
&\left\|\int_0^1 \nabla^2(f(\mathbf{x}_t+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))-\nabla^2 f(\mathbf{x}_{t+1}))(\mathbf{x}_{t+1}-\mathbf{x}_t)\,\mathrm{d}\tau\right\|_2\\
\leq&\int_0^1 \left\|\nabla^2(f(\mathbf{x}_t+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))-\nabla^2 f(\mathbf{x}_{t+1}))\right\|_2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|_2\,\mathrm{d}\tau\\
\leq&\int_0^1 L_2(1-\tau)\|\mathbf{x}_t-\mathbf{x}_{t+1})\|_2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|_2\,\mathrm{d}\tau\\
=&\frac{L_2}{2}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|_2^2.
\end{aligned}
$$

For one-dimension case, we consider find the root of $g(x)=0$ (function $g(\cdot)$ can be viewed as gradient of objective). The Newton's method can be written as

$$
\begin{aligned}
x_{t+1}&=x_t-(g'(x_t))^{-1}g(x_t),\\
x_{t+2}&=x_{t+1}-(g'(x_{t+1}))^{-1}g(x_{t+1}),\\
&\cdots\cdots
\end{aligned}
$$

The derivative can be estimated by (when $\Delta=x_t-x_{t+1}\approx 0$)

$$
g'(x_{t+1})=\lim_{\Delta\to 0}\frac{g(x_{t+1}+\Delta)-g(x_{t+1})}{\Delta}\approx\frac{g(x_{t+1})-g(x_t)}{x_{t+1}-x_t}\quad\Longrightarrow\quad g'(x_{t+1})(x_{t+1}-x_t)\approx g(x_{t+1})-g(x_t).
$$

In multivariate case, it implies

$$
\nabla^2 f(x_{t+1})(x_{t+1}-x_t)\approx\nabla f(x_{t+1})-\nabla f(x_t).
$$

**SR1 method** We consider secant condition and rank-1 update

$$\mathbf{y}_t = \mathbf{G}_{t+1}\mathbf{s}_t \tag{35}$$

and

$$\mathbf{G}_{t+1} = \mathbf{G}_t + \mathbf{z}_t\mathbf{z}_t^\top. \tag{36}$$

where $\mathbf{y}_t = \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)$ and $\mathbf{s}_t = \mathbf{x}_{t+1} - \mathbf{x}_t$. Combining above equalities implies

$$\begin{aligned}
&\mathbf{y}_t = (\mathbf{G}_t + \mathbf{z}_t\mathbf{z}_t^\top)\mathbf{s}_t && (37)\\
\Longrightarrow &\mathbf{y}_t = \mathbf{G}_t\mathbf{s}_t + (\mathbf{z}_t^\top\mathbf{s}_t)\mathbf{z}_t && (38)\\
\Longrightarrow &(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top = (\mathbf{z}_t^\top\mathbf{s}_t)^2\mathbf{z}_t\mathbf{z}_t^\top. && (39)
\end{aligned}$$

Left Multiplying $\mathbf{s}_t^\top$ on (37) leads to

$$\mathbf{s}_t^\top\mathbf{y}_t = \mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t + (\mathbf{z}_t^\top\mathbf{s}_t)^2 \quad \Longrightarrow \quad (\mathbf{z}_t^\top\mathbf{s}_t)^2 = \mathbf{s}_t^\top\mathbf{y}_t - \mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t. \tag{40}$$

Combining (36), (39) and (40), we achieve

$$\mathbf{G}_{t+1} = \mathbf{G}_t + \frac{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top}{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{s}_t}.$$

The inverse of Hessian can be obtain by Woodbury matrix identity as follows

$$\begin{aligned}
\mathbf{G}_{t+1}^{-1} =& \mathbf{G}_t^{-1} - \frac{\mathbf{G}^{-1}(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)}{\sqrt{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{s}_t}}\left(1 + \frac{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{G}_t^{-1}(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)}{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{s}_t}\right)^{-1}\frac{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{G}^{-1}}{\sqrt{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{s}_t}}\\
=& \mathbf{G}_t^{-1} - \frac{(\mathbf{G}^{-1}\mathbf{y}_t - \mathbf{s}_t)(\mathbf{G}^{-1}\mathbf{y}_t - \mathbf{s}_t)^\top}{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{s}_t + (\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{G}_t^{-1}(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)}\\
=& \mathbf{G}_t^{-1} + \frac{(\mathbf{s}_t - \mathbf{G}^{-1}\mathbf{y}_t)(\mathbf{s}_t - \mathbf{G}^{-1}\mathbf{y}_t)^\top}{(\mathbf{s}_t - \mathbf{G}_t^{-1}\mathbf{y}_t)^\top\mathbf{y}_t},
\end{aligned}$$

where the last step is because of

$$\begin{aligned}
&(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{s}_t + (\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{G}_t^{-1}(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)\\
=&(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\left(\mathbf{s}_t + \mathbf{G}_t^{-1}(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)\right)\\
=&(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{G}_t^{-1}\mathbf{y}_t\\
=&(\mathbf{G}_t^{-1}\mathbf{y}_t - \mathbf{s}_t)^\top\mathbf{y}_t.
\end{aligned}$$

**BFGS Method** The update for inverse Hessian is

$$\mathbf{G}_{t+1}^{-1} = \left(\mathbf{I} - \frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{G}_t^{-1}\left(\mathbf{I} - \frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right) + \frac{\mathbf{s}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t},$$

If $\mathbf{G}_t$ is positive definite then $\mathbf{G}_{t+1}^{-1}$ is also positive definite. For any non-zero $\mathbf{z} \in \mathbb{R}^d$, we have

$$\mathbf{z}^\top\mathbf{G}_{t+1}^{-1}\mathbf{z} = \mathbf{z}^\top\left(\mathbf{I} - \frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{G}_t^{-1}\left(\mathbf{I} - \frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{z} + \frac{(\mathbf{s}_t^\top\mathbf{z})^2}{\mathbf{y}_t^\top\mathbf{s}_t} > 0.$$

Consider that if the second term is 0, then $\mathbf{s}_t^\top\mathbf{z} = 0$, which implies

$$\mathbf{z}^\top\left(\mathbf{I} - \frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{G}_t^{-1}\left(\mathbf{I} - \frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{z} = \left(\mathbf{z}^\top - \frac{(\mathbf{z}^\top\mathbf{s}_t)\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{G}_t^{-1}\left(\mathbf{z} - \frac{\mathbf{y}_t(\mathbf{s}_t^\top\mathbf{z})}{\mathbf{y}_t^\top\mathbf{s}_t}\right) = \mathbf{z}^\top\mathbf{G}\mathbf{z} > 0.$$

DFP also holds the similar property while SR1 not.

**Remark 9.1.** *BFGS and DFP is always well-defined for strongly-convex objective, while the denominator in SR1 update can vanish.*

**Theorem 9.1.** *The solution of the following matrix optimization problem*

$$\begin{aligned}\min_{\mathbf{H}\in\mathbb{R}^{d\times d}} &\|\mathbf{H}-\mathbf{H}_t\|_{\bar{\mathbf{G}}_t}\\ \text{s.t}\quad &\mathbf{H}=\mathbf{H}^\top,\quad \mathbf{H}\mathbf{y}_t=\mathbf{s}_t,\end{aligned}$$

*is*

$$\left(\mathbf{I}-\frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{H}_t\left(\mathbf{I}-\frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)+\frac{\mathbf{s}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}.$$

*where* $\mathbf{H}_t=\mathbf{G}_t^{-1}$ *and the weighted norm* $\|\cdot\|_{\bar{\mathbf{G}}_t}$ *is defined as*

$$\|\mathbf{A}\|_{\bar{\mathbf{G}}_t}=\left\|\bar{\mathbf{G}}_t^{1/2}\mathbf{A}\bar{\mathbf{G}}_t^{1/2}\right\|_F,\quad \textit{with}\quad \bar{\mathbf{G}}_t=\int_0^1\nabla^2 f(\mathbf{x}_t+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))\,\mathrm{d}\tau.$$

*Proof.* We introduce

$$\hat{\mathbf{H}}=\bar{\mathbf{G}}^{1/2}\mathbf{H}\bar{\mathbf{G}}^{1/2},\quad \hat{\mathbf{H}}_t=\bar{\mathbf{G}}^{1/2}\mathbf{H}_t\bar{\mathbf{G}}^{1/2},\quad \hat{\mathbf{s}}_t=\bar{\mathbf{G}}^{1/2}\mathbf{s}_t\quad\text{and}\quad \hat{\mathbf{y}}_t=\bar{\mathbf{G}}^{-1/2}\mathbf{y}_t.$$

Then we have

$$\|\mathbf{H}-\mathbf{H}_t\|_{\bar{\mathbf{G}}_t}=\left\|\bar{\mathbf{G}}_t^{1/2}(\mathbf{H}-\mathbf{H}_t)\bar{\mathbf{G}}_t^{1/2}\right\|_F=\left\|\hat{\mathbf{H}}-\hat{\mathbf{H}}_t\right\|_F$$

and

$$\begin{cases}\mathbf{H}\mathbf{y}_t=\mathbf{s}_t &\iff (\bar{\mathbf{G}}^{-1/2}\hat{\mathbf{H}}\bar{\mathbf{G}}^{-1/2})\bar{\mathbf{G}}^{1/2}\hat{\mathbf{y}}_t=\bar{\mathbf{G}}^{-1/2}\hat{\mathbf{s}}_t \iff \hat{\mathbf{H}}\hat{\mathbf{y}}_t=\hat{\mathbf{s}}_t,\\ \bar{\mathbf{G}}\mathbf{s}_t=\mathbf{y}_t &\iff \bar{\mathbf{G}}^{1/2}\mathbf{s}_t=\bar{\mathbf{G}}^{-1/2}\mathbf{y}_t \iff \hat{\mathbf{s}}_t=\hat{\mathbf{y}}_t,\end{cases}$$

which means to problem is equivalent to

$$\begin{aligned}\min_{\hat{\mathbf{H}}\in\mathbb{R}^{d\times d}} &\left\|\hat{\mathbf{H}}-\hat{\mathbf{H}}_t\right\|_F\\ \text{s.t}\quad &\hat{\mathbf{H}}=\hat{\mathbf{H}}^\top,\quad \hat{\mathbf{H}}\hat{\mathbf{y}}_t=\hat{\mathbf{y}}_t\end{aligned}$$

and $\hat{\mathbf{y}}_t$ is an eigenvector of $\hat{\mathbf{H}}$ with respect to eigenvalue 1 ($\hat{\mathbf{H}}\mathbf{y}_t=\mathbf{y}_t$). Let $\mathbf{u}=\hat{\mathbf{y}}_t/\left\|\hat{\mathbf{y}}_t\right\|_2\in\mathbb{R}^d$ ($\hat{\mathbf{H}}\mathbf{u}=\mathbf{u}$ and $\mathbf{u}^\top\hat{\mathbf{H}}\mathbf{u}=1$) and

$$\mathbf{U}=\begin{bmatrix}\mathbf{u} & \mathbf{U}_\perp\end{bmatrix}\in\mathbb{R}^{d\times d}$$

be an orthogonal matrix, where $\mathbf{U}_\perp\in\mathbb{R}^{d\times(d-1)}$ is the orthogonal complement to $\mathbf{u}$ such that $\mathbf{u}^\top\mathbf{U}_\perp=\mathbf{0}$. Then we have

$$\mathbf{U}^\top\hat{\mathbf{H}}\mathbf{U}=\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{U}_\perp^\top\hat{\mathbf{H}}\mathbf{U}_\perp\end{bmatrix}.$$

Since the Frobenius norm is unitary invariant, we have

$$\begin{aligned}\left\|\hat{\mathbf{H}}-\hat{\mathbf{H}}_t\right\|_F^2=&\left\|\mathbf{U}^\top\hat{\mathbf{H}}\mathbf{U}-\mathbf{U}^\top\hat{\mathbf{H}}_t\mathbf{U}\right\|_F^2=\left\|\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{U}_\perp^\top\hat{\mathbf{H}}\mathbf{U}_\perp\end{bmatrix}-\begin{bmatrix}\mathbf{u}^\top\hat{\mathbf{H}}_t\mathbf{u} & \mathbf{u}^\top\hat{\mathbf{H}}_t\mathbf{U}_\perp\\ \mathbf{U}_\perp^\top\hat{\mathbf{H}}_t\mathbf{u} & \mathbf{U}_\perp^\top\hat{\mathbf{H}}_t\mathbf{U}_\perp\end{bmatrix}\right\|_F^2\\ =&(1-\mathbf{u}^\top\hat{\mathbf{H}}_t\mathbf{u})^2+\left\|\mathbf{u}^\top\hat{\mathbf{H}}_t\mathbf{U}_\perp\right\|_F^2+\left\|\mathbf{U}_\perp^\top\hat{\mathbf{H}}_t\mathbf{u}\right\|_F^2+\left\|\mathbf{U}_\perp^\top\hat{\mathbf{H}}\mathbf{U}_\perp-\mathbf{U}_\perp^\top\hat{\mathbf{H}}_t\mathbf{U}_\perp\right\|_F^2.\end{aligned}$$

Since matrices $\mathbf{u}$, $\mathbf{U}_\perp$ and $\hat{\mathbf{H}}_t$ (because $\mathbf{u}$ depends on $\hat{\mathbf{y}}_t$) will not change by varying $\hat{\mathbf{H}}$, we only need to minimize the last term in above, which can not be smaller than zero. Hence, we desire

$$\mathbf{U}_\perp^\top\hat{\mathbf{H}}\mathbf{U}_\perp=\mathbf{U}_\perp^\top\hat{\mathbf{H}}_t\mathbf{U}_\perp,$$

which can be hold by taking

$$\hat{\mathbf{H}} = \mathbf{U}\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}\end{bmatrix}\mathbf{U}^\top,$$

since

$$\begin{aligned}\mathbf{U}_\perp^\top \hat{\mathbf{H}}\mathbf{U}_\perp =&\mathbf{U}_\perp^\top \begin{bmatrix}\mathbf{u} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{u}^\top\\ \mathbf{U}_\perp^\top\end{bmatrix}\mathbf{U}_\perp\\ =&\begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix}\\ =&\begin{bmatrix}\mathbf{0} & \mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix}\\ =&\mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{H}}\hat{\mathbf{y}}_t =&\begin{bmatrix}\mathbf{u} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{u}^\top\\ \mathbf{U}_\perp^\top\end{bmatrix}\hat{\mathbf{y}}_t\\ =&\begin{bmatrix}\mathbf{u} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{u}^\top\hat{\mathbf{y}}_t\\ \mathbf{0}\end{bmatrix}\\ =&\begin{bmatrix}\mathbf{u} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{u}^\top\hat{\mathbf{y}}_t\\ \mathbf{0}\end{bmatrix}\\ =&\mathbf{u}(\mathbf{u}^\top\hat{\mathbf{y}}_t) = \hat{\mathbf{y}}_t.\end{aligned}$$

Consequently, we achieve

$$\begin{aligned}\hat{\mathbf{H}} =&\begin{bmatrix}\mathbf{u} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{u}^\top\\ \mathbf{U}_\perp\end{bmatrix}\\ =&\begin{bmatrix}\mathbf{u} & \mathbf{U}_\perp\mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{u}^\top\\ \mathbf{U}_\perp^\top\end{bmatrix}\\ =&\mathbf{u}\mathbf{u}^\top + \mathbf{U}_\perp\mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp\mathbf{U}_\perp^\top\\ =&\mathbf{u}\mathbf{u}^\top + (\mathbf{I}-\mathbf{u}\mathbf{u}^\top)\hat{\mathbf{H}}_t(\mathbf{I}-\mathbf{u}\mathbf{u}^\top),\end{aligned}$$

which implies

$$\begin{aligned}\mathbf{H} =&\bar{\mathbf{G}}^{-1/2}\hat{\mathbf{H}}\bar{\mathbf{G}}^{-1/2}\\ =&\bar{\mathbf{G}}^{-1/2}\big(\mathbf{u}\mathbf{u}^\top + (\mathbf{I}-\mathbf{u}\mathbf{u}^\top)\hat{\mathbf{H}}_t(\mathbf{I}-\mathbf{u}\mathbf{u}^\top)\big)\bar{\mathbf{G}}^{-1/2}\\ =&\bar{\mathbf{G}}^{-1/2}\mathbf{u}\mathbf{u}^\top\bar{\mathbf{G}}^{-1/2} + \bar{\mathbf{G}}^{-1/2}(\mathbf{I}-\mathbf{u}\mathbf{u}^\top)\bar{\mathbf{G}}^{1/2}\mathbf{H}_t\bar{\mathbf{G}}^{1/2}(\mathbf{I}-\mathbf{u}\mathbf{u}^\top)\bar{\mathbf{G}}^{-1/2}\\ =&\bar{\mathbf{G}}^{-1/2}\mathbf{u}\mathbf{u}^\top\bar{\mathbf{G}}^{-1/2} + (\mathbf{I}-\bar{\mathbf{G}}^{-1/2}\mathbf{u}\mathbf{u}^\top\bar{\mathbf{G}}^{1/2})\mathbf{H}_t(\mathbf{I}-\bar{\mathbf{G}}^{1/2}\mathbf{u}\mathbf{u}^\top\bar{\mathbf{G}}^{-1/2}).\end{aligned}$$

Since the definition means

$$\bar{\mathbf{G}}^{-1/2}\mathbf{u} = \frac{\bar{\mathbf{G}}^{-1/2}\hat{\mathbf{y}}_t}{\|\hat{\mathbf{y}}_t\|_2} = \frac{\bar{\mathbf{G}}^{-1}\mathbf{y}_t}{\|\bar{\mathbf{G}}^{-1/2}\mathbf{y}_t\|_2} = \frac{\left(\int_0^1 \nabla^2 f(\mathbf{x}_t+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))\,\mathrm{d}\tau\right)^{-1}\mathbf{y}_t}{(\mathbf{y}_t^\top\mathbf{G}_t^{-1/2}\mathbf{y}_t)^{1/2}} = \frac{\mathbf{s}_t}{(\mathbf{y}_t\mathbf{s}_t)^{1/2}}$$

and

$$\bar{\mathbf{G}}^{1/2}\mathbf{u} = \frac{\bar{\mathbf{G}}^{1/2}\hat{\mathbf{y}}_t}{(\mathbf{y}_t\mathbf{s}_t)^{1/2}} = \frac{\mathbf{y}_t}{(\mathbf{y}_t\mathbf{s}_t)^{1/2}},$$

we obtain

$$\mathbf{H} = \frac{\mathbf{s}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t} + \left(\mathbf{I}-\frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{H}_t\left(\mathbf{I}-\frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right).$$

□

**The Broyden Family Update** The Broyden family update is

$$\begin{aligned}\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u}) \triangleq \tau&\left[\mathbf{G}-\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G}+\mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}+\left(\frac{\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}+1\right)\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}\right]\\&+(1-\tau)\left[\mathbf{G}-\frac{(\mathbf{G}-\mathbf{A})\mathbf{u}\mathbf{u}^\top(\mathbf{G}-\mathbf{A})}{\mathbf{u}^\top(\mathbf{G}-\mathbf{A})\mathbf{u}}\right],\end{aligned}$$

where $\mathbf{G}\in\mathbb{R}^{d\times d}$ ,$\mathbf{A}\in\mathbb{R}^{d\times d}$, $\mathbf{u}\in\mathbb{R}^d$ and $\tau\in[0,1]$. Classical quasi-Newton methods correspond to taking

$$\mathbf{G}=\mathbf{G}_t,\qquad \mathbf{A}=\int_0^1\nabla^2 f(\mathbf{x}_t+t(\mathbf{x}_{t+1}-\mathbf{x}_t))\,\mathrm{d}t\qquad\text{and}\qquad \mathbf{u}=\mathbf{x}_{t+1}-\mathbf{x}_t=\mathbf{s}_t.$$

For above setting, we have

$$\mathbf{A}\mathbf{u}=\int_0^1\nabla^2 f(\mathbf{x}_t+t(\mathbf{x}_{t+1}-\mathbf{x}_t))\,\mathrm{d}t(\mathbf{x}_{t+1}-\mathbf{x}_t)=\nabla f(\mathbf{x}_{t+1})-\nabla f(\mathbf{x}_t)=\mathbf{y}_t.$$

If $\tau=0$, then the update $\mathbf{G}_{t+1}=\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u})$ corresponds to SR1 update since

$$\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u})=\mathbf{G}-\frac{(\mathbf{G}-\mathbf{A})\mathbf{u}\mathbf{u}^\top(\mathbf{G}-\mathbf{A})}{\mathbf{u}^\top(\mathbf{G}-\mathbf{A})\mathbf{u}}=\mathbf{G}_t-\frac{(\mathbf{G}_t\mathbf{s}_t-\mathbf{y}_t)(\mathbf{G}_t\mathbf{s}_t-\mathbf{y}_t)^\top}{\mathbf{s}_t^\top(\mathbf{G}_t\mathbf{s}_t-\mathbf{y}_t)}.$$

If $\tau=1$, then the update $\mathbf{G}_{t+1}=\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u})$ corresponds to DFP update since

$$\begin{aligned}\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u})=&\mathbf{G}-\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G}+\mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}+\left(\frac{\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}+1\right)\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}\\=&\mathbf{G}_t-\frac{\mathbf{y}_t\mathbf{s}_t^\top\mathbf{G}_t+\mathbf{G}_t\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}+\left(\frac{\mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t}{\mathbf{s}_t^\top\mathbf{y}_t}+1\right)\frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}\\=&\mathbf{G}_t-\frac{\mathbf{y}_t\mathbf{s}_t^\top\mathbf{G}_t+\mathbf{G}_t\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}+\frac{\mathbf{y}_t\mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t\mathbf{y}_t}{\mathbf{s}_t^\top\mathbf{y}_t}+\frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}\\=&\left(\mathbf{I}-\frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}\right)\mathbf{G}_t\left(\mathbf{I}-\frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}\right)+\frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}.\end{aligned}$$

For $\tau=\dfrac{\mathbf{u}^\top\mathbf{A}\mathbf{u}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}$, then the update $\mathbf{G}_{t+1}=\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u})$ corresponds to BFGS update since

$$\begin{aligned}&\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u})\\=&\frac{\mathbf{u}^\top\mathbf{A}\mathbf{u}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}\left[\mathbf{G}-\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G}+\mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}+\left(\frac{\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}+1\right)\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}\right]\\&+\left(1-\frac{\mathbf{u}^\top\mathbf{A}\mathbf{u}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}\right)\left[\mathbf{G}-\frac{(\mathbf{G}-\mathbf{A})\mathbf{u}\mathbf{u}^\top(\mathbf{G}-\mathbf{A})}{\mathbf{u}^\top(\mathbf{G}-\mathbf{A})\mathbf{u}}\right]\\=&\frac{\mathbf{u}^\top\mathbf{A}\mathbf{u}\mathbf{G}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}-\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G}+\mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}+\left(\frac{\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}+1\right)\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}\\&+\left(1-\frac{\mathbf{u}^\top\mathbf{A}\mathbf{u}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}\right)\mathbf{G}-\frac{\mathbf{u}^\top(\mathbf{G}-\mathbf{A})\mathbf{u}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}\cdot\frac{(\mathbf{G}-\mathbf{A})\mathbf{u}\mathbf{u}^\top(\mathbf{G}-\mathbf{A})}{\mathbf{u}^\top(\mathbf{G}-\mathbf{A})\mathbf{u}}\\=&\frac{\mathbf{u}^\top\mathbf{A}\mathbf{u}\mathbf{G}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}-\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G}+\mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}+\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}+\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}\\&+\mathbf{G}-\frac{\mathbf{u}^\top\mathbf{A}\mathbf{u}\mathbf{G}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}-\frac{\mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{G}-\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G}-\mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{A}+\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}\\=&\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}+\mathbf{G}-\frac{\mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{G}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}=\frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}+\mathbf{G}_t-\frac{\mathbf{G}_t\mathbf{s}_t\mathbf{s}_t^\top\mathbf{G}_t}{\mathbf{s}_t^\top\mathbf{G}\mathbf{s}_t}.\end{aligned}$$

Taking $\mathbf{A} = \nabla^2 f(\mathbf{x}_{t+1})$, SR1 update holds that

$$\begin{aligned}\mathbf{G}_{t+1}\mathbf{u} =& \mathbf{G}_t\mathbf{u} - \frac{(\mathbf{G}_t - \mathbf{A}))\mathbf{u}\mathbf{u}^\top(\mathbf{G}_t - \mathbf{A})\mathbf{u}}{\mathbf{u}^\top(\mathbf{G}_t - \mathbf{A})\mathbf{u}} \\ =& \mathbf{G}_t\mathbf{u} - (\mathbf{G}_t - \nabla^2 f(\mathbf{x}_{t+1}))\mathbf{u} = \nabla^2 f(\mathbf{x}_{t+1})\mathbf{u}\end{aligned}$$

for any $\mathbf{u} \in \mathbb{R}^d$; and DFP update holds that

$$\begin{aligned}\mathbf{G}_{t+1}\mathbf{u} =& \mathbf{G}\mathbf{u} - \frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G}\mathbf{u} + \mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{A}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} + \left(\frac{\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} + 1\right)\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} \\ =& \mathbf{G}\mathbf{u} - \frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} - \mathbf{G}\mathbf{u} + \left(\frac{\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} + 1\right)\mathbf{A}\mathbf{u} \\ =& \mathbf{A}\mathbf{u} = \nabla^2 f(\mathbf{x}_{t+1})\mathbf{u}.\end{aligned}$$

Since Broyden's family update is a convex combination of SR1 update an DFP update, it also holds

$$\mathbf{G}_{t+1}\mathbf{u} = \nabla^2 f(\mathbf{x}_{t+1})\mathbf{u}.$$

**Hessian-Vector Product** The Hessian-vector product can be written as

$$\nabla^2 f(\mathbf{x})\mathbf{v} = \lim_{t\to 0}\frac{\nabla f(\mathbf{x} + t\mathbf{v}) - \nabla f(\mathbf{x})}{t}.$$

For generalized linear model

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \phi(\mathbf{a}_i^\top\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2,$$

we have

$$\nabla^2 f(\mathbf{x})\mathbf{v} = \frac{1}{n}\sum_{i=1}^n \phi''(\mathbf{a}_i^\top\mathbf{x})(\mathbf{a}_i^\top\mathbf{v})\mathbf{a}_i + \lambda\mathbf{v}.$$

Note that it is unnecessary to construct

$$\frac{1}{n}\sum_{i=1}^n \mathbf{a}_i\mathbf{a}_i^\top.$$



**Block Quasi-Newton** Let $\mathbf{G}_+ = \text{SR-}k(\mathbf{G}, \mathbf{A}, \mathbf{U})$ with

$$\text{SR-}k(\mathbf{G}, \mathbf{A}, \mathbf{U}) = \mathbf{G} - (\mathbf{G} - \mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G} - \mathbf{A})\mathbf{U})^{-1}\mathbf{U}^\top(\mathbf{G} - \mathbf{A}),$$

then we have

$$\begin{aligned}\mathbf{G}_+\mathbf{U} =& \big(\mathbf{G} - (\mathbf{G} - \mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G} - \mathbf{A})\mathbf{U})^{-1}\mathbf{U}^\top(\mathbf{G} - \mathbf{A})\big)\mathbf{U} \\ =& \mathbf{G}\mathbf{U} - (\mathbf{G} - \mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G} - \mathbf{A})\mathbf{U})^{-1}\mathbf{U}^\top(\mathbf{G} - \mathbf{A})\mathbf{U} \\ =& \mathbf{G}\mathbf{U} - (\mathbf{G} - \mathbf{A})\mathbf{U} = \mathbf{A}\mathbf{U}.\end{aligned}$$

**Part I: Convergence for Matrix Approximation**

**Lemma 9.1.** *For any positive-definite matrices $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{G} \in \mathbb{R}^{d\times d}$ with $\mathbf{A} \preceq \mathbf{G} \preceq \eta\mathbf{A}$ for some $\eta \geq 1$, we let $\mathbf{G}_+ = \text{SR-}k(\mathbf{G}, \mathbf{A}, \mathbf{U})$ for some full rank matrix $\mathbf{U} \in \mathbb{R}^{d\times k}$. Then it holds that*

$$\mathbf{A} \preceq \mathbf{G}_+ \preceq \eta\mathbf{A}. \tag{41}$$

*Proof.* Define $\mathbf{R} = \mathbf{G} - \mathbf{A} \succeq \mathbf{0}$. According to the update rule, we have

$$\begin{aligned}&\mathbf{G}_+ - \mathbf{A}\\ =&\mathbf{R} - \mathbf{R}\mathbf{U}(\mathbf{U}^\top\mathbf{R}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{R}\\ =&\left(\mathbf{I}_d - \mathbf{R}\mathbf{U}(\mathbf{U}^\top\mathbf{R}\mathbf{U})^{-1}\mathbf{U}^\top\right)\mathbf{R}\left(\mathbf{I}_d - \mathbf{U}(\mathbf{U}^\top\mathbf{R}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{R}\right) \succeq \mathbf{0},\end{aligned}$$

which means $\mathbf{G}_+ \succeq \mathbf{A}$. The condition $\mathbf{G} \preceq \eta\mathbf{A}$ means

$$\mathbf{G}_+ \preceq \eta\mathbf{A} - \underbrace{\mathbf{R}\mathbf{U}(\mathbf{U}^\top\mathbf{R}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{R}}_{\succeq\mathbf{0}} \preceq \eta\mathbf{A},$$

which finish the proof. □

**Theorem 9.2.** *Let*

$$\mathbf{G}_+ = \text{SR-}k(\mathbf{G}, \mathbf{A}, \mathbf{U}) \tag{42}$$

*with $\mathbf{G} \succeq \mathbf{A} \in \mathbb{R}^{d\times d}$ and select $\mathbf{U} \in \mathbb{R}^{d\times k}$ by drawomg each entry of $\mathbf{U}$ according to $\mathcal{N}(0,1)$ independently. Then, we have*

$$\mathbb{E}\left[\tau_{\mathbf{A}}(\mathbf{G}_+)\right] \le \left(1 - \frac{k}{d}\right)\tau_{\mathbf{A}}(\mathbf{G}). \tag{43}$$

**Remark 9.2.** *Before start from the proof, we should accept the concept of singular normal distribution. Since the density function*

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

*does not work of singular $\boldsymbol{\Sigma}$. The mass is concentrated on a low dimensional set $\mathcal{S}$. For any $\mathbf{x} \notin \mathcal{S}$, there exists $\mathcal{B}(\mathbf{x}, r)$ such that $r > 0$ and $\mathcal{B} \cap \mathcal{S} = \emptyset$. If the distribution of $\mathbf{x}$ has density function $f$, then $f(\mathbf{x}) = 0$ holds for any $\mathbf{x} \notin \mathcal{S}$. Since the measure of $\mathcal{S}$ is zero, we have $f(\mathbf{x}) = 0$ almost everywhere, which means the integration of $f(\mathbf{x})$ on the whole space is 0.*

We first provide several lemmas for random matrix and the trace of positive definite matrix.

**Lemma 9.2.** *Assume $\mathbf{P} \in \mathbb{R}^{d\times k}$ is column orthonormal ($k \le d$) and $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}\mathbf{P}^\top)$ is a $d$-dimensional multivariate normal distributed vector. Then we have*

$$\mathbb{E}\left[\frac{\mathbf{p}\mathbf{p}^\top}{\mathbf{p}^\top\mathbf{p}}\right] = \frac{1}{k}\mathbf{P}\mathbf{P}^\top.$$

*Proof.* The distribution $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}\mathbf{P}^\top)$ implies there exists a $k$-dimensional multivariate normal distributed vector $\mathbf{p}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$ such that $\mathbf{p} = \mathbf{P}\mathbf{p}_1$. Thus we have

$$\mathbb{E}\left[\frac{\mathbf{p}\mathbf{p}^\top}{\mathbf{p}^\top\mathbf{p}}\right] = \mathbb{E}\left[\frac{(\mathbf{P}\mathbf{p}_1)(\mathbf{P}\mathbf{p}_1)^\top}{(\mathbf{P}\mathbf{p}_1)^\top(\mathbf{P}\mathbf{p}_1)}\right] = \mathbb{E}\left[\frac{\mathbf{P}\mathbf{p}_1\mathbf{p}_1^\top\mathbf{P}^\top}{\mathbf{p}_1^\top\mathbf{p}_1}\right] = \mathbf{P}\mathbb{E}\left[\frac{\mathbf{p}_1\mathbf{p}_1^\top}{\mathbf{p}_1^\top\mathbf{p}_1}\right]\mathbf{P}^\top = \frac{1}{k}\mathbf{P}\mathbf{P}^\top,$$

where the last step is because of $\mathbf{p}_1/\left\|\mathbf{p}_1\right\|_2 \in \mathbb{R}^k$ is uniform distributed on unit sphere and its covariance matrix is $k^{-1}\mathbf{I}_k$. □

**Remark 9.3.** *Let $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\mathbf{y} = \mathbf{z}/\left\|\mathbf{z}\right\|_2$, then*

$$\mathbf{y} = [y_1, \ldots, y_d]^\top \sim \text{Unif}(\mathcal{S}^{k-1})$$

*The uniformly sphere distribution means*

$$\mathbf{y} = [y_1, \ldots, y_i, \ldots, y_d]^\top \quad \textit{and} \quad \mathbf{y} = [y_1, \ldots, -y_i, \ldots, y_d]^\top$$

*has the same distribution. Hence, we have*

$$\mathbb{E}[y_i] = -\mathbb{E}[y_i],$$

*which implies*

$$\mathbb{E}[y_i y_j] = \mathbb{E}[(-y_i)y_j] = -\mathbb{E}[(y_i)y_j]$$

*for any $i \neq j$. This implies $\mathbb{E}[y_i y_j] = 0$. On the other hand, the variables $y_1, \ldots, y_d$ are symmetric, which means $\mathbb{E}[y_1^2] = \cdots = \mathbb{E}[y_d^2]$. Combing with the fact*

$$\sum_{i=1}^{d} y_i^2 = \sum_{i=1}^{d} \frac{z_i^2}{\sum_{i=1}^{d} z_i^2} = 1,$$

*we achieve $\mathbb{E}[y_1^2] = \cdots = \mathbb{E}[y_d^2] = 1/d$.*

**Lemma 9.3.** *Let $\mathbf{U} \in \mathbb{R}^{d\times k}$ be a random matrix and each of its entry is independent and identically distributed according to $\mathcal{N}(0,1)$, then it holds that*

$$\mathbb{E}\left[\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\right] = \frac{k}{d}\mathbf{I}_d. \tag{44}$$

*Proof.* We prove inequality (44) by induction on $k$. The induction base $k = 1$ is easily verified. Now we assume

$$\mathbb{E}\left[\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\right] = \frac{k}{d}\mathbf{I}_d$$

holds for any $\mathbf{U} \in \mathbb{R}^{d\times k}$ that each of its entries are independently distributed according to $\mathcal{N}(0,1)$. We define the random matrix

$$\bar{\mathbf{U}} = \begin{bmatrix}\mathbf{U} & \mathbf{q}\end{bmatrix} \in \mathbb{R}^{d\times(k+1)},$$

where $\mathbf{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ is independent distributed to $\mathbf{U}$. Then we have

$$\bar{\mathbf{U}}(\bar{\mathbf{U}}^\top\bar{\mathbf{U}})^{-1}\bar{\mathbf{U}}^\top = \begin{bmatrix}\mathbf{U} & \mathbf{q}\end{bmatrix}\left(\begin{bmatrix}\mathbf{U}^\top\\ \mathbf{q}^\top\end{bmatrix}\begin{bmatrix}\mathbf{U} & \mathbf{q}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{U}^\top\\ \mathbf{q}^\top\end{bmatrix} = \mathbf{A} + \frac{(\mathbf{I}_d - \mathbf{A})\mathbf{q}\mathbf{q}^\top(\mathbf{I}_d - \mathbf{A})}{\mathbf{q}^\top(\mathbf{I}_d - \mathbf{A})\mathbf{q}},$$

where $\mathbf{A} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$. Since the rank of projection matrix $\mathbf{I}_d - \mathbf{A}$ is $d-k$, we have $\mathbf{I}_d - \mathbf{A} = \mathbf{Q}\mathbf{Q}^\top$ for some column orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{d\times(d-k)}$. Thus, we achieve

$$\begin{aligned}
\mathbb{E}[\bar{\mathbf{U}}(\bar{\mathbf{U}}^\top\bar{\mathbf{U}})\bar{\mathbf{U}}^\top] &= \frac{k}{d}\mathbf{I}_d + \mathbb{E}_{\mathbf{U}}\left[\mathbb{E}_{\mathbf{q}}\left[\frac{(\mathbf{I}_d - \mathbf{A})\mathbf{q}\mathbf{q}^\top(\mathbf{I}_d - \mathbf{A})}{\mathbf{q}^\top(\mathbf{I}_d - \mathbf{A})\mathbf{q}}\,\middle|\,\mathbf{U}\right]\right]\\
&= \frac{k}{d}\mathbf{I}_d + \mathbb{E}_{\mathbf{U}}\left[\mathbb{E}_{\mathbf{q}}\left[\frac{(\mathbf{Q}\mathbf{Q}^\top\mathbf{q})(\mathbf{q}^\top\mathbf{Q}\mathbf{Q}^\top)}{(\mathbf{q}^\top\mathbf{Q}\mathbf{Q}^\top)(\mathbf{Q}\mathbf{Q}^\top\mathbf{q})}\,\middle|\,\mathbf{U}\right]\right]\\
&= \frac{k}{d}\mathbf{I}_d + \frac{1}{d-k}\mathbb{E}_{\mathbf{U}}[\mathbf{Q}\mathbf{Q}^\top]\\
&= \frac{k}{d}\mathbf{I}_d + \frac{1}{d-k}\mathbb{E}_{\mathbf{U}}[\mathbf{I}_d - \mathbf{A}]\\
&= \frac{k}{d}\mathbf{I}_d + \frac{1}{d-k}\frac{d-k}{d}\mathbf{I}_d\\
&= \frac{k+1}{d}\mathbf{I}_d,
\end{aligned}$$

which completes the induction. In above derivation, the second equality is due to Lemma 9.2 and the fact $\mathbf{Q}\mathbf{Q}^\top\mathbf{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}\mathbf{Q}^\top)$ for given $\mathbf{Q}$; the third equality comes from the inductive hypothesis. □

**Lemma 9.4** (not appear on slides)**.** *For any positive semi-definite matrices* $\mathbf{B}, \mathbf{P}_1, \mathbf{P}_2 \in \mathbb{R}^{d\times d}$ *such that* $\mathbf{P}_1 \succeq \mathbf{P}_2$*, we have*

$$\operatorname{tr}(\mathbf{P}_1\mathbf{B}) \geq \operatorname{tr}(\mathbf{P}_2\mathbf{B}). \tag{45}$$

*Proof.* We denote $\mathbf{R} = \mathbf{P}_1 - \mathbf{P}_2 \succeq \mathbf{0}$, then

$$\mathbf{R}^{1/2}\mathbf{B}\mathbf{R}^{1/2} \succeq \mathbf{0},$$

which means

$$\operatorname{tr}((\mathbf{P}_1 - \mathbf{P}_2)^{1/2}\mathbf{B}(\mathbf{P}_1 - \mathbf{P}_2)^{1/2}) \geq 0.$$

So we have

$$\operatorname{tr}(\mathbf{P}_1\mathbf{B}) - \operatorname{tr}(\mathbf{P}_2\mathbf{B}) = \operatorname{tr}((\mathbf{P}_1 - \mathbf{P}_2)\mathbf{B}) = \operatorname{tr}((\mathbf{P}_1 - \mathbf{P}_2)^{1/2}\mathbf{B}(\mathbf{P}_1 - \mathbf{P}_2)^{1/2}) \geq 0.$$

□

**Lemma 9.5** (not appear on slides)**.** *For positive semi-definite matrix* $\mathbf{S} \in \mathbb{R}^{d\times d}$ *and the column orthonormal matrix* $\mathbf{Q} \in \mathbb{R}^{d\times k}$*, we have*

$$\operatorname{tr}(\mathbf{Q}^\top\mathbf{S}\mathbf{Q}) \leq \operatorname{tr}(\mathbf{S}). \tag{46}$$

*Proof.* Since matrix $\mathbf{Q}$ is column orthonormal, we have

$$\mathbf{Q}\mathbf{Q}^\top = \mathbf{Q}(\mathbf{Q}^\top\mathbf{Q})^{-1}\mathbf{Q}^\top \preceq \mathbf{I}_d.$$

According to Lemma 9.4, we have

$$\operatorname{tr}(\mathbf{Q}^\top\mathbf{S}\mathbf{Q}) = \operatorname{tr}(\mathbf{S}\mathbf{Q}\mathbf{Q}^\top) \leq \operatorname{tr}(\mathbf{S}).$$

□

**Lemma 9.6.** *For positive semi-definite matrix* $\mathbf{B} \in \mathbb{R}^{d\times d}$ *and full rank matrix* $\mathbf{U} \in \mathbb{R}^{d\times k}$ *with* $k \leq d$*, it holds that*

$$\operatorname{tr}(\mathbf{B}\mathbf{U}\left(\mathbf{U}^\top\mathbf{B}\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{B}) \geq \operatorname{tr}(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{B}). \tag{47}$$

*Proof.* We denote SVD of $\mathbf{U}$ as $\mathbf{U} = \mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{Q} \in \mathbb{R}^{d\times k}$, $\mathbf{V} \in \mathbb{R}^{k\times k}$ are (column) orthogonal and $\mathbf{\Sigma} \in \mathbb{R}^{k\times k}$ is diagonal. We have

$$\operatorname{tr}(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{B}) = \operatorname{tr}(\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\left(\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top\right)^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{B}) = \operatorname{tr}(\mathbf{Q}\mathbf{Q}^\top\mathbf{B}),$$

and

$$\begin{aligned}
\operatorname{tr}(\mathbf{B}\mathbf{U}\left(\mathbf{U}^\top\mathbf{B}\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{B}) &= \operatorname{tr}(\mathbf{B}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\left(\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{B}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\right)^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{B}) \\
&= \operatorname{tr}(\mathbf{B}\mathbf{Q}\left(\mathbf{Q}^\top\mathbf{B}\mathbf{Q}\right)^{-1}\mathbf{Q}^\top\mathbf{B}) \\
&\overset{(46)}{\geq} \operatorname{tr}(\mathbf{Q}^\top\mathbf{B}\mathbf{Q}\left(\mathbf{Q}^\top\mathbf{B}\mathbf{Q}\right)^{-1}\mathbf{Q}^\top\mathbf{B}\mathbf{Q}) \\
&= \operatorname{tr}(\mathbf{Q}^\top\mathbf{B}\mathbf{Q}),
\end{aligned}$$

which leads to inequality (47). □

Now, we present the proof of Theorem 9.2.

*Proof.* Combining Lemma 9.3 with Lemma 9.6, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{tr}((\mathbf{U}^\top\mathbf{B}\mathbf{U})^{-1}(\mathbf{U}^\top\mathbf{B}^2\mathbf{U}))\right]\\
=&\mathbb{E}\left[\mathrm{tr}(\mathbf{B}\mathbf{U}(\mathbf{U}^\top\mathbf{B}\mathbf{U})^{-1}(\mathbf{U}^\top\mathbf{B}))\right]\\
\overset{(47)}{\geq}&\mathbb{E}\left[\mathrm{tr}(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{B})\right]\\
=&\mathrm{tr}\left(\mathbb{E}\left[\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\right]\mathbf{B}\right)\\
=&\frac{k}{d}\mathrm{tr}(\mathbf{B}),
\end{aligned}$$

for any positive semi-definite $\mathbf{B}\in\mathbb{R}^{d\times d}$. The update rule of SR-$k$ update leads to

$$\begin{aligned}
\mathbf{G}_+-\mathbf{A}&=\mathbf{G}-\mathbf{A}-(\mathbf{G}-\mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U})^{-1}\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\\
&=\mathbf{B}-\mathbf{B}\mathbf{U}(\mathbf{U}^\top\mathbf{B}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{B},
\end{aligned}$$

where we set $\mathbf{B}=\mathbf{G}-\mathbf{A}\succeq\mathbf{0}$. Then, we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{tr}(\mathbf{G}_+-\mathbf{A})\right]&=\mathbb{E}\left[\mathrm{tr}(\mathbf{B}-\mathbf{B}\mathbf{U}(\mathbf{U}^\top\mathbf{B}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{B})\right]\\
&=\mathrm{tr}(\mathbf{B})-\mathbb{E}\left[\mathrm{tr}(\mathbf{B}\mathbf{U}(\mathbf{U}^\top\mathbf{B}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{B})\right]\\
&=\mathrm{tr}(\mathbf{B})-\mathbb{E}\left[\mathrm{tr}((\mathbf{U}^\top\mathbf{B}\mathbf{U})^{-1}(\mathbf{U}^\top\mathbf{B}^2\mathbf{U}))\right]\\
&\leq\mathrm{tr}(\mathbf{B})-\frac{k}{d}\mathrm{tr}(\mathbf{B})\\
&=\left(1-\frac{k}{d}\right)\mathrm{tr}(\mathbf{B}).
\end{aligned}$$

□



**Part II: Convergence for Optimization Algorithm**

**Lemma 9.7.** *Suppose function $f$ is $M$-strongly self-concordant. Let*

$$r=\|\mathbf{x}-\mathbf{y}\|_{\nabla^2 f(\mathbf{x})}\qquad\text{and}\qquad\mathbf{J}=\int_0^1\nabla f(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x}))\,\mathrm{d}\tau,$$

*then we have*

$$\frac{\nabla^2 f(\mathbf{x})}{1+Mr}\preceq\nabla^2 f(\mathbf{y})\preceq(1+Mr)\,\nabla^2 f(\mathbf{x})\tag{48}$$

*and*

$$\frac{\nabla^2 f(\mathbf{x})}{1+\frac{Mr}{2}}\preceq\mathbf{J}\preceq\left(1+\frac{Mr}{2}\right)\nabla^2 f(\mathbf{x}),\qquad\frac{\nabla^2 f(\mathbf{y})}{1+\frac{Mr}{2}}\preceq\mathbf{J}\preceq\left(1+\frac{Mr}{2}\right)\nabla^2 f(\mathbf{y})\tag{49}$$

**Lemma 9.8.** *Suppose that the twice differentiable function $f:\mathbb{R}^d\to\mathbb{R}$ is strongly self-concordant with constant $M>0$ and the positive definite matrix $\mathbf{G}_t\in\mathbb{R}^{d\times d}$ satisfies*

$$\nabla^2 f(\mathbf{x}_t)\preceq\mathbf{G}_t\preceq\eta_t\nabla^2 f(\mathbf{x}_t)\tag{50}$$

*for some $\eta_t\geq 1$ and $M\lambda_t\leq 2$. Then the update formula*

$$\mathbf{x}_{t+1}=\mathbf{x}_t-\mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t)\tag{51}$$

*holds that*

$$\|\mathbf{x}_{t+1}-\mathbf{x}_t\|_{\nabla^2 f(\mathbf{x}_t)}\leq\lambda_t\qquad\text{and}\qquad\lambda_{t+1}\leq\left(1-\frac{1}{\eta_t}\right)\lambda_t+\frac{M}{2}\lambda_t^2+\frac{M^2}{4\eta_t}\lambda_t^3.\tag{52}$$

*Proof.* Let $r_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\nabla^2 f(\mathbf{x}_t)}$, then relation (50) means

$$\begin{aligned} r_t^2 =& \|\mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t)\|_{\nabla^2 f(\mathbf{x}_t)} \\ =& \nabla f(\mathbf{x}_t)^\top \mathbf{G}_t^{-1}\nabla^2 f(\mathbf{x}_t)\mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t) \\ \leq& \nabla f(\mathbf{x}_t)^\top \mathbf{G}_t^{-1}\mathbf{G}_t\mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)^\top \mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t) \\ \leq& \nabla f(\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t) = \lambda_t^2. \end{aligned} \tag{53}$$

Let $\mathbf{J} = \int_0^1 \nabla f(\mathbf{x}_t + \tau(\mathbf{x}_{t+1} - \mathbf{x}))\,\mathrm{d}\tau$, then Taylor's formula means

$$\nabla f(\mathbf{x}_{t+1}) = \nabla f(\mathbf{x}_t) + \mathbf{J}(\mathbf{x}_{t+1} - \mathbf{x}_t) = \nabla f(\mathbf{x}_t) - \mathbf{J}\mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t) = \mathbf{J}(\mathbf{J}^{-1} - \mathbf{G}_t^{-1})\nabla f(\mathbf{x}_t). \tag{54}$$

In the view of result (49) in Lemma 9.7, we have

$$\frac{\nabla^2 f(\mathbf{x}_t)}{1 + \frac{M\lambda_t}{2}} \preceq \mathbf{J} \preceq \left(1 + \frac{M\lambda_t}{2}\right)\nabla^2 f(\mathbf{x}_t) \qquad \text{and} \qquad \mathbf{J} \preceq \left(1 + \frac{M\lambda_t}{2}\right)\nabla^2 f(\mathbf{x}_{t+1}). \tag{55}$$

This implies

$$\frac{1}{1 + \frac{Mr_t}{2}}\mathbf{J} \overset{(55)}{\preceq} \nabla^2 f(\mathbf{x}_t) \overset{(50)}{\preceq} \mathbf{G}_t \overset{(50)}{\preceq} \eta\nabla^2 f(\mathbf{x}_t) \overset{(55)}{\preceq} \eta\left(1 + \frac{Mr_t}{2}\right)\mathbf{J}.$$

Hence, we have

$$\frac{1}{\eta\left(1 + \frac{Mr_t}{2}\right)}\mathbf{J}^{-1} \preceq \mathbf{G}_t^{-1} \preceq \left(1 + \frac{Mr_t}{2}\right)\mathbf{J}^{-1}$$

and

$$\left(1 - \frac{1}{\eta\left(1 + \frac{Mr_t}{2}\right)}\right)\mathbf{J}^{-1} \preceq \mathbf{G}_t^{-1} - \mathbf{J}^{-1} \preceq \frac{Mr_t}{2}\mathbf{J}^{-1}.$$

Note that

$$1 - \frac{1}{\eta\left(1 + \frac{Mr_t}{2}\right)} \leq 1 - \frac{1 - \frac{Mr_t}{2}}{\eta} = \frac{\eta - 1 + \frac{Mr_t}{2}}{\eta}$$

and $M\lambda_t \leq 2$ means

$$\frac{Mr_t}{2} = 1 - \left(1 - \frac{Mr_t}{2}\right) \leq 1 - \frac{1 - \frac{Mr_t}{2}}{\eta} = \frac{\eta - 1 + \frac{Mr_t}{2}}{\eta},$$

which implies

$$-\frac{\eta - 1 + \frac{Mr_t}{2}}{\eta}\mathbf{J}^{-1} \preceq \mathbf{G}_t^{-1} - \mathbf{J}^{-1} \preceq \frac{\eta - 1 + \frac{Mr_t}{2}}{\eta}\mathbf{J}^{-1}.$$

Consequently

$$(\mathbf{G}_t^{-1} - \mathbf{J}^{-1})\mathbf{J}(\mathbf{G}_t^{-1} - \mathbf{J}^{-1}) \preceq \left(\frac{\eta - 1 + \frac{Mr_t}{2}}{\eta}\right)^2 \mathbf{J}^{-1} \tag{56}$$

Finally, we obtain

$$\begin{aligned} \lambda_{t+1}^2 =& \nabla f(\mathbf{x}_{t+1})^\top (\nabla^2 f(\mathbf{x}_{t+1}))^{-1}\nabla f(\mathbf{x}_{t+1}) \\ \overset{(55)}{\leq}& \left(1 + \frac{M\lambda_t}{2}\right)\nabla f(\mathbf{x}_{t+1})^\top \mathbf{J}^{-1}\nabla f(\mathbf{x}_{t+1}) \end{aligned}$$

$$
\begin{aligned}
&\overset{(54)}{=} \left(1+\frac{M\lambda_t}{2}\right)\nabla f(\mathbf{x}_t)^\top(\mathbf{J}^{-1}-\mathbf{G}_t^{-1})\mathbf{J}(\mathbf{J}^{-1}-\mathbf{G}_t^{-1})\nabla f(\mathbf{x}_t)\\
&\overset{(56)}{\leq} \left(1+\frac{M\lambda_t}{2}\right)\left(\frac{\eta-1+\frac{Mr_t}{2}}{\eta}\right)^2\nabla f(\mathbf{x}_t)^\top\mathbf{J}^{-1}\nabla f(\mathbf{x}_t)\\
&\overset{(55)}{\leq} \left(1+\frac{M\lambda_t}{2}\right)^2\left(\frac{\eta-1+\frac{Mr_t}{2}}{\eta}\right)^2\nabla f(\mathbf{x}_t)^\top(\nabla^2 f(\mathbf{x}_t))^{-1}\nabla f(\mathbf{x}_t)\\
&= \left(1+\frac{M\lambda_t}{2}\right)^2\left(\frac{\eta-1+\frac{Mr_t}{2}}{\eta}\right)^2\lambda_t^2,
\end{aligned}
$$

which finish our proof. □

**Lemma 9.9.** *If twice differentiable function $f:\mathbb{R}^d\to\mathbb{R}$ is $M$-strongly self-concordant and $\mu$-strongly convex and positive definite matrix $\mathbf{G}\in\mathbb{R}^{d\times d}$ and $\mathbf{x}\in\mathbb{R}^d$ satisfy*

$$\nabla^2 f(\mathbf{x})\preceq\mathbf{G}\preceq\eta\nabla^2 f(\mathbf{x}) \tag{57}$$

*for some $\eta\geq 1$, then we have*

$$\nabla^2 f(\mathbf{x}_+)\preceq\tilde{\mathbf{G}}\preceq\eta(1+Mr)^2\nabla^2 f(\mathbf{x}_+)$$

*for any $\mathbf{x}_+\in\mathbb{R}^d$, where $\tilde{\mathbf{G}}=(1+Mr)\mathbf{G}$, $r=\|\mathbf{x}-\mathbf{x}_+\|_{\nabla^2 f(\mathbf{x})}$.*

*Proof.* Using Lemma 9.7, we have

$$\nabla^2 f(\mathbf{x}_+)\overset{(48)}{\preceq}(1+Mr)\nabla^2 f(\mathbf{x})\overset{(57)}{\preceq}(1+Mr)\mathbf{G}=\tilde{\mathbf{G}}$$

and

$$\tilde{\mathbf{G}}=(1+Mr)\mathbf{G}\overset{(57)}{\preceq}\eta(1+Mr)\nabla^2 f(\mathbf{x})\overset{(48)}{\preceq}\eta(1+Mr)^2\nabla^2 f(\mathbf{x}_+).$$

□

**Remark 9.4.** *Using Lemma 9.1 and above lemma leads to*

$$\nabla^2 f(\mathbf{x}_+)\preceq\mathbf{G}_+=\text{SR-}k(\tilde{\mathbf{G}},\nabla^2 f(\mathbf{x}_+))\preceq\eta(1+Mr)^2\nabla^2 f(\mathbf{x}_+)$$

*that is*

$$\nabla^2 f(\mathbf{x}_{t+1})\preceq\mathbf{G}_{t+1}=\text{SR-}k(\tilde{\mathbf{G}}_t,\nabla^2 f(\mathbf{x}_{t+1}))\preceq\eta(1+Mr_t)^2\nabla^2 f(\mathbf{x}_{t+1}).$$

*This says $\tilde{\mathbf{G}}_{t+1}$ increase a little distance to Hessian if $r_t$ is small, while Theorem 9.2 says* SR-$k$ *step decrease distance to Hessian a lot.*

**Remark 9.5.** *Lemma 9.8 says the convergence of $\lambda_t$ can be measured by how $\mathbf{G}_t$ converges to $\nabla^2 f(\mathbf{x}_t)$ (corresponds to $\eta_t$). Lemma 9.9 implies*

$$\nabla^2 f(\mathbf{x}_t)\preceq\mathbf{G}_t\preceq(1+\delta_t)\nabla^2 f(\mathbf{x}_t),$$

*where*

$$\delta_t=\frac{d\kappa\text{tr}(\mathbf{G}_t-\nabla^2 f(\mathbf{x}_t))}{\text{tr}(\nabla^2 f(\mathbf{x}_t))}.$$

*Using induction on $\lambda_t$ and $\delta_t$, and Theorem 9.2, we obtain the convergence result.*

# 10 Stochastic Gradient Descent

**Theorem 10.1.** *Using stochastic subgradient method in slides to solve*

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \mathbb{E}[F(\mathbf{x};\xi)],$$

*where each $F(\mathbf{x};\xi)$ is convex and $G$-Lipschitz such that $\|\mathbf{g}\|_2 \le G$ for any $\mathbf{g}\in\partial F(\mathbf{x};\xi)$. Then we have*

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] \le f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0-\hat{\mathbf{x}}\|_2^2 + \sum_{t=0}^{T-1} G^2\eta_t^2}{2\sum_{t=0}^{T-1}\eta_t}.$$

*Proof.* Conditioned on $\xi_0,\dots,\xi_{t-1}$, we have

$$\begin{aligned}
&\mathbb{E}_{\xi_t}\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\|_2^2\\
=&\mathbb{E}_{\xi_t}\|\mathbf{x}_{t+1}-\mathbf{x}_t+\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2\\
=&\mathbb{E}_{\xi_t}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|_2^2 + 2\mathbb{E}_{\xi_t}\langle\mathbf{x}_{t+1}-\mathbf{x}_t,\mathbf{x}_t-\hat{\mathbf{x}}\rangle + \mathbb{E}_{\xi_t}\|\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2\\
=&\eta_t^2\mathbb{E}_{\xi_t}\|\mathbf{g}_t\|_2^2 - 2\eta_t\mathbb{E}_{\xi_t}\langle\mathbf{g}_t,\mathbf{x}_t-\hat{\mathbf{x}}\rangle + \|\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2\\
\le&\eta_t^2G^2 - 2\eta_t\langle\tilde{\mathbf{g}}_t,\mathbf{x}_t-\hat{\mathbf{x}}\rangle + \|\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2\\
\le&\eta_t^2G^2 + 2\eta_t(f(\hat{\mathbf{x}})-f(\mathbf{x}_t)) + \|\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2,
\end{aligned}$$

where the first inequality is based on the bounded subgradient assumption and the second one use the definition of subgradient. Here we let

$$\tilde{\mathbf{g}}_t = \mathbb{E}_{\xi_t}[\mathbf{g}_t] \in \partial f(\mathbf{x}_t),$$

which is because of $\mathbf{g}_{t,\xi_1}\in\partial F(\mathbf{x};\xi_1),\dots,\mathbf{g}_{t,\xi_n}\in\partial F(\mathbf{x};\xi_n)$ leads to

$$\sum_{i=1}^n \mathbf{g}_{t,\xi_i} \in \partial\sum_{i=1}^n F(\mathbf{x};\xi_i)$$

We sum above inequality over $t=0,\dots,T-1$ and taking expectation with all the history, then

$$0 \le \mathbb{E}\|\mathbf{x}_T-\hat{\mathbf{x}}\|_2^2 \le \sum_{t=0}^{T-1}\eta_t^2G^2 + 2\sum_{t=0}^{T-1}\eta_t\mathbb{E}[f(\hat{\mathbf{x}})-f(\mathbf{x}_t)] + \|\mathbf{x}_0-\hat{\mathbf{x}}\|_2^2,$$

which implies

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)-f(\hat{\mathbf{x}})] = \frac{\sum_{t=0}^{T-1}\eta_t(f(\mathbf{x}_t)-f(\hat{\mathbf{x}}))}{\sum_{t=0}^{T-1}\eta_t} \le \frac{\|\mathbf{x}_0-\hat{\mathbf{x}}\|_2^2 + \sum_{t=0}^{T-1}G^2\eta_t^2}{2\sum_{t=0}^{T-1}\eta_t}.$$

□

**Remark 10.1.** *Compared with deterministic case, this result is about expectation, and we suppose $G$-Lipschitz and convexity on each stochastic component..*

**Remark 10.2.** *It is $n$ times faster than deterministic algorithm for finite-sum case.*

**Remark 10.3.** *We consider $\mu$-strongly objective. We have*

$$\begin{aligned}
&\langle\mathbf{g}_t,\mathbf{x}_t-\hat{\mathbf{x}}\rangle\\
=&\frac{1}{\eta_t}\langle\mathbf{x}_t-\mathbf{x}_{t+1},\mathbf{x}_t-\hat{\mathbf{x}}\rangle\\
=&\frac{1}{2\eta_t}\left(\|\mathbf{x}_t-\mathbf{x}_{t+1}\|_2^2 + \|\mathbf{x}_t-\hat{\mathbf{x}}\|_2^2 - \|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\|_2^2\right)
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{2\eta_t}\left(\eta_t^2\left\|\mathbf{g}_t\right\|_2^2+\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2-\left\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\right\|_2^2\right)\\
\leq&\frac{1}{2\eta_t}\left(\eta_t^2G^2+\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2-\left\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\right\|_2^2\right)\\
=&\frac{1}{2\eta_t}\left(\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2-\left\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\right\|_2^2\right)+\frac{\eta_tG^2}{2}
\end{aligned}
$$

*Conditioned on $\xi_0,\ldots,\xi_{t-1}$, we obtain*

$$
\langle\mathbb{E}_{\xi_t}[\mathbf{g}_t],\mathbf{x}_t-\hat{\mathbf{x}}\rangle\leq\frac{1}{2\eta_t}\left(\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2-\mathbb{E}_{\xi_t}\left\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\right\|_2^2\right)+\frac{\eta_tG^2}{2}.
$$

*Taking $\eta_t=2/(\mu(t+1))$ and combining with the strong convexity, we obtain*

$$
\begin{aligned}
&f(\mathbf{x}_t)-f(\hat{\mathbf{x}})\\
\leq&\langle\mathbb{E}_{\xi_t}[\mathbf{g}_t],\mathbf{x}_t-\hat{\mathbf{x}}\rangle-\frac{\mu}{2}\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2\\
\leq&\frac{1}{2\eta_t}\left(\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2-\mathbb{E}_{\xi_t}\left\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\right\|_2^2\right)+\frac{\eta_tG^2}{2}-\frac{\mu}{2}\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2\\
=&\frac{\mu(t+1)}{4}\left(\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2-\mathbb{E}_{\xi_t}\left\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\right\|_2^2\right)+\frac{G^2}{\mu(t+1)}-\frac{\mu}{2}\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2\\
=&\frac{\mu(t-1)}{4}\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2-\frac{\mu(t+1)}{4}\mathbb{E}_{\xi_t}\left\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\right\|_2^2+\frac{G^2}{\mu(t+1)},
\end{aligned}
$$

*which implies*

$$
\begin{aligned}
t(f(\mathbf{x}_t)-f(\hat{\mathbf{x}}))\leq&\frac{\mu(t-1)t}{4}\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2-\frac{\mu t(t+1)}{4}\mathbb{E}_{\xi_t}\left\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\right\|_2^2+\frac{G^2t}{\mu(t+1)}\\
\leq&\frac{\mu(t-1)t}{4}\left\|\mathbf{x}_t-\hat{\mathbf{x}}\right\|_2^2-\frac{\mu t(t+1)}{4}\mathbb{E}_{\xi_t}\left\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}\right\|_2^2+\frac{G^2}{\mu}.
\end{aligned}
$$

*We sum over above inequality over $t=0,\ldots,T-1$ and take expectation on all of history, then*

$$
\sum_{t=0}^{T-1}\mathbb{E}[t(f(\mathbf{x}_t)-f(\hat{\mathbf{x}}))]\leq-\frac{\mu(T-1)T}{4}\mathbb{E}\left\|\mathbf{x}_T-\hat{\mathbf{x}}\right\|_2^2+\frac{TG^2}{\mu}\leq\frac{TG^2}{\mu}
$$

*Hence, we have*

$$
\sum_{t=0}^{T-1}\frac{t}{T(T-1)}\mathbb{E}[f(\mathbf{x}_t)]\leq f(\hat{\mathbf{x}})+\frac{2G^2}{\mu(T-1)}.
$$

*Let*

$$
\bar{\mathbf{x}}_T=\sum_{t=0}^{T-1}\frac{t\mathbf{x}_t}{T(T-1)},
$$

*then*

$$
\mathbb{E}[f(\bar{\mathbf{x}}_T)]\leq f(\hat{\mathbf{x}})+\frac{2G^2}{\mu(T-1)}.
$$

**Remark 10.4.** *We can not benefit to large batch, since we only replace $\mathbb{E}\left[\mathbf{g}_{t,\xi_i}\right]\in\partial f(\mathbf{x}_t)$ by*

$$
\mathbb{E}\left[\frac{1}{b}\sum_{j=1}^{b}\mathbf{g}_{t,\xi_j}\right]\in\partial f(\mathbf{x}_t).
$$

**Analysis for Mini-Batch SGD (Smooth and Convex)** Let

$$F(\mathbf{x}_t;\mathcal{S}_t)=\frac{1}{b}\sum_{i=1}^{b}F(\mathbf{x}_t;\xi_{t,i}).$$

We have $\mathbb{E}[F(\mathbf{x}_t;\mathcal{S}_t)]=f(\mathbf{x}_t)$ and $\mathbb{E}[\nabla F(\mathbf{x}_t;\mathcal{S}_t)]=\nabla f(\mathbf{x}_t)$. Conditioned on $\mathcal{S}_0,\dots,\mathcal{S}_{t-1}$, it follows that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t}\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2\\
=&\mathbb{E}_{\mathcal{S}_t}\left\|\mathbf{x}_t-\eta_t\nabla F(\mathbf{x}_t;\mathcal{S}_t)-\mathbf{x}^*\right\|_2^2\\
=&\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta_t\mathbb{E}_{\mathcal{S}_t}\langle\mathbf{x}_t-\mathbf{x}^*,\nabla F(\mathbf{x}_t;\mathcal{S}_t)\rangle+\eta_t^2\mathbb{E}_{\mathcal{S}_t}\left\|\nabla F(\mathbf{x}_t;\mathcal{S}_t)\right\|_2^2\\
=&\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta_t\langle\mathbf{x}_t-\mathbf{x}^*,\nabla f(\mathbf{x}_t)\rangle+\eta_t^2\mathbb{E}_{\mathcal{S}_t}\left\|\nabla F(\mathbf{x}_t;\mathcal{S}_t)\right\|_2^2\\
\le&\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2+2\eta_t\left(f(\mathbf{x}^*)-f(\mathbf{x}_t)\right)+\eta_t^2\underbrace{\mathbb{E}_{\mathcal{S}_t}\left\|\nabla F(\mathbf{x}_t;\mathcal{S}_t)\right\|_2^2}_{C_t},
\end{aligned}\tag{58}$$

where the inequality is because of the strong convexity. Furthermore,

$$\begin{aligned}
C_t=&\mathbb{E}_{\mathcal{S}_t}\left\|\nabla F(\mathbf{x}_t;\mathcal{S}_t)-\nabla F(\mathbf{x}^*;\mathcal{S}_t)+\nabla F(\mathbf{x}^*;\mathcal{S}_t)\right\|_2^2\\
\le&2\mathbb{E}_{\mathcal{S}_t}\left\|\nabla F(\mathbf{x}_t;\mathcal{S}_t)-\nabla F(\mathbf{x}^*;\mathcal{S}_t)\right\|_2^2+2\mathbb{E}_{\mathcal{S}_t}\left\|\nabla F(\mathbf{x}^*;\mathcal{S}_t)\right\|_2^2.
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t}\left\|\nabla F(\mathbf{x}_t;\mathcal{S}_t)-\nabla F(\mathbf{x}^*;\mathcal{S}_t)\right\|_2^2\\
\le&\mathbb{E}_{\mathcal{S}_t}\left[2L(F(\mathbf{x}_t;\mathcal{S}_t)-F(\mathbf{x}^*;\mathcal{S}_t)-\langle\nabla F(\mathbf{x}_t;\mathcal{S}_t),\mathbf{x}_t-\mathbf{x}^*\rangle)\right]\\
=&2L(f(\mathbf{x}_t)-f(\mathbf{x}^*)),
\end{aligned}$$

where the inequality is due to the third statement of Theorem 3.19. Let

$$V^*=\mathbb{E}_\xi\left\|\nabla F(\mathbf{x}^*;\xi)-\nabla f(\mathbf{x}^*)\right\|_2^2.$$

For the second term, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t}\left\|\nabla F(\mathbf{x}^*;\mathcal{S}_t)\right\|_2^2\\
=&\mathbb{E}_{\mathcal{S}_t}\left\|\nabla F(\mathbf{x}^*;\mathcal{S}_t)-\nabla f(\mathbf{x}^*)\right\|_2^2\\
=&\mathbb{E}_{\mathcal{S}_t}\left\|\frac{1}{b}\sum_{i=1}^{b}\left(\nabla F(\mathbf{x}^*;\xi_{t,i})-\mathbb{E}[\nabla F(\mathbf{x}^*;\xi_{t,i})]\right)\right\|_2^2\\
=&\frac{1}{b}\mathbb{E}_{\xi_{t,i}}\left\|\nabla F(\mathbf{x}^*;\xi_{t,i})-\mathbb{E}[\nabla F(\mathbf{x}^*;\xi_{t,i})]\right\|_2^2=\frac{V^*}{b}.
\end{aligned}$$

Hence, we have

$$C_t\le 4L(f(\mathbf{x}_t)-f(\mathbf{x}^*))+\frac{2V^*}{b}$$

and

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_t}\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2&\le\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2+2\eta_t\left(f(\mathbf{x}^*)-f(\mathbf{x}_t)\right)+4\eta_t^2L(f(\mathbf{x}_t)-f(\mathbf{x}^*))+\frac{2\eta_t^2V^*}{b}\\
&=\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2+(2\eta_t-4\eta_t^2L)(f(\mathbf{x}^*)-f(\mathbf{x}_t))+\frac{2\eta_t^2V^*}{b}.
\end{aligned}$$

We sum over above inequality over $t=0,\dots,T-1$ and take expectation on all of history, then

$$\sum_{t=0}^{T-1}2\eta_t(1-2\eta_tL)(\mathbb{E}[f(\mathbf{x}_t)]-f(\mathbf{x}^*))\le\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2^2-\mathbb{E}\left\|\mathbf{x}_T-\mathbf{x}^*\right\|_2^2+\frac{2V^*\sum_{t=0}^{T-1}\eta_t^2}{b}$$

$$\leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{2V^* \sum_{t=0}^{T-1} \eta_t^2}{b}$$

Taking $\eta_t \leq 1/(3L)$, we have $\eta_t(1 - 2\eta_t L) \geq \eta_t/3$. Hence,

$$\sum_{t=0}^{T-1} \frac{2\eta_t}{3}(\mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}^*)) \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{2V^* \sum_{t=0}^{T-1} \eta_t^2}{b}.$$

For fixed $\eta_t = \eta$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}^*) \leq \frac{3\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\eta T} + \frac{3V^* \sum_{t=0}^{T-1} \eta}{bT} = \frac{3\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\eta T} + \frac{3V^*\eta}{b}.$$

Let $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{x}_t$. We can set different parameters.

- For $b = 1$ and $\eta = 1/(L\sqrt{T})$, we have

  $$\mathbb{E}[f(\bar{\mathbf{x}}_T)] - f(\mathbf{x}^*) \leq \frac{3L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\sqrt{T}} + \frac{3V^*}{L\sqrt{T}}.$$

  We require $T = \mathcal{O}(\epsilon^{-2})$ to obtain $\epsilon$-suboptimal solution.

- For general, we set $\eta = 1/(L\sqrt{T/b})$. Then

  $$\mathbb{E}[f(\bar{\mathbf{x}}_T)] - f(\mathbf{x}^*) \leq \frac{3L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\sqrt{bT}} + \frac{3V^*}{L\sqrt{bT}}.$$

  We require $T = \mathcal{O}(\epsilon^{-2}/b)$ to obtain $\epsilon$-suboptimal solution.

**Remark 10.5.** *We consider $\mu$-strongly convex case. Following (58) and setting $\eta_t \leq 1/(2L)$, we have*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \\
=&\mathbb{E}_{\mathcal{S}_t}\|\mathbf{x}_t - \eta_t \nabla F(\mathbf{x}_t; \mathcal{S}_t) - \mathbf{x}^*\|_2^2 \\
=&\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t \mathbb{E}_{\mathcal{S}_t}\langle \mathbf{x}_t - \mathbf{x}^*, \nabla F(\mathbf{x}_t; \mathcal{S}_t)\rangle + \eta_t^2 \mathbb{E}_{\mathcal{S}_t}\|\nabla F(\mathbf{x}_t; \mathcal{S}_t)\|_2^2 \\
=&\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t \langle \mathbf{x}_t - \mathbf{x}^*, \nabla f(\mathbf{x}_t)\rangle + \eta_t^2 C_t \\
\leq&\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t \langle \mathbf{x}_t - \mathbf{x}^*, \nabla f(\mathbf{x}_t)\rangle + \eta_t^2 \left(4L(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{2V^*}{b}\right) \\
\leq&\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t(\mu\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + f(\mathbf{x}_t) - f(\mathbf{x}^*)) + 4L\eta_t^2(f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{2\eta_t^2 V^*}{b} \\
=&(1 - 2\eta\mu)\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - (2\eta_t - 4L\eta_t^2)(f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{2\eta_t^2 V^*}{b} \\
\leq&(1 - 2\eta\mu)\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \frac{2\eta_t^2 V^*}{b}.
\end{aligned}$$

*where the second inequality uses strong convexity and the last one uses $\eta_t \leq 1/(2L)$. For $\eta_t = \eta$, we have*

$$\mathbb{E}\|\mathbf{x}_T - \mathbf{x}^*\|_2^2 \leq (1 - 2\eta\mu)^T\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{2\eta V^*}{2\mu b}.$$

*The larger $\eta$ means faster linear convergence, but we need larger batch to reduce the variance*

**Another analysis for strongly-convex case** Let $\eta_t \leq 1/(4L)$. We have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t} \left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \\
\leq& \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + 2\eta_t \left(f(\mathbf{x}^*) - f(\mathbf{x}_t) - \frac{\mu}{2}\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right) + 4\eta_t^2 L(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{2\eta_t^2 V^*}{b} \\
\leq& (1 - \eta_t\mu)\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 - 4\eta_t^2 L\left(f(\mathbf{x}^*) - f(\mathbf{x}_t)\right) + \frac{2\eta_t^2 V^*}{b} \\
\leq& (1 - \eta_t\mu)\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 - \eta_t\left(f(\mathbf{x}^*) - f(\mathbf{x}_t)\right) + \frac{2\eta_t^2 V^*}{b}.
\end{aligned}$$

Hence,

$$\frac{1}{\eta_t}\mathbb{E}_{\mathcal{S}_t}\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \leq \frac{1 - \eta_t\mu}{\eta_t}\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + \left(f(\mathbf{x}^*) - f(\mathbf{x}_t)\right) + \frac{2\eta_t^2 V^*}{b}.$$

Let $\eta_t = 2/(8L + \mu t)$, then $(1 - \eta_t\mu)/\eta_t = 1/\mu_{t-1}$ and

$$\frac{1}{\eta_t}\mathbb{E}_{\mathcal{S}_t}\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \leq \frac{1}{\eta_{t-1}}\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + \left(f(\mathbf{x}^*) - f(\mathbf{x}_t)\right) + \frac{2\eta_t V^*}{b}.$$

Taking all expectation and summing over above results, we have

$$\frac{1}{\eta_{T-1}}\mathbb{E}\left\|\mathbf{x}_T - \mathbf{x}^*\right\|_2^2 \leq \frac{1}{\eta_{-1}}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2 + \sum_{t=0}^{T-1}\mathbb{E}[f(\mathbf{x}^*) - f(\mathbf{x}_t)] + \sum_{t=0}^{T-1}\frac{2\eta_t V^*}{b},$$

which means

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \leq \frac{8L - \mu}{2T}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2 + \frac{2V^*}{bT}\sum_{t=0}^{T-1}\eta_t \leq \frac{4L}{T}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2 + \frac{2V^*}{bT}\sum_{t=0}^{T-1}\eta_t.$$

Since

$$\sum_{t=0}^{T-1}\eta_t = \sum_{t=0}^{T-1}\frac{2}{8L + \mu t} \leq \int_1^T \frac{2}{8L + \mu t}\mathrm{d}t \leq \frac{2}{\mu}\ln(8L + \mu t)\Big|_1^T = \frac{2}{\mu}\ln\frac{8L + \mu T}{8L + \mu} \leq \frac{2}{\mu}\ln(T + 1),$$

we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \leq \frac{4L}{T}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2 + \frac{4V^*}{bT\mu}\ln(T + 1).$$



# 11 Variance Reduction Methods

Let

$$V_t = \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right\|_2^2.$$

The smoothness means

$$\begin{aligned}
\mathbb{E}_i[f(\mathbf{x}_t - \eta\nabla f_i(\mathbf{x}_t))] \leq& f(\mathbf{x}_t) - \eta_t\mathbb{E}_i[\langle\nabla f(\mathbf{x}_t), \nabla f_i(\mathbf{x}_t)\rangle] + \frac{L\eta_t^2}{2}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_t)\right\|_2^2 \\
=& f(\mathbf{x}_t) - \eta_t\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{L\eta_t^2}{2}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_t)\right\|_2^2 \\
\leq& f(\mathbf{x}_t) - \eta_t\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + L\eta_t^2\mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}_t) - \nabla f(\mathbf{x})\right\|_2^2 + \left\|\nabla f(\mathbf{x})\right\|_2^2\right]
\end{aligned}$$

$$\leq f(\mathbf{x}_t) - \eta_t \|\nabla f(\mathbf{x}_t)\|_2^2 + L(\|\nabla f(\mathbf{x}_t)\|_2^2 + V_t)\eta_t^2.$$

Taking

$$\eta_t = \frac{\|\nabla f(\mathbf{x}_t)\|_2^2}{2L(\|\nabla f(\mathbf{x}_t)\|_2^2 + V_t)}$$

leads to the steepest descent. For $\|\nabla f(\mathbf{x}_t)\|_2^2 \to 0$, we have $\eta_t \to 0$ and the descent

$$-\eta_t \|\nabla f(\mathbf{x}_t)\|_2^2 + L(\|\nabla f(\mathbf{x}_t)\|_2^2 + V_t)\eta_t^2 = -\frac{\|\nabla f(\mathbf{x}_t)\|_2^4}{4L(\|\nabla f(\mathbf{x}_t)\|_2^2 + V_t)}$$

also converges to 0.

**Variance Reduction** We define the auxiliary function

$$\tilde{f}_i(\mathbf{x}) = f_i(\mathbf{x}) - \langle \nabla f_i(\tilde{\mathbf{x}}) - \tilde{\boldsymbol{\mu}}, \mathbf{x} \rangle,$$

then

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\tilde{\mathbf{x}}).$$

We apply SGD to finite-sum on $\tilde{f}_i(\mathbf{x})$ and obtain

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla \tilde{f}_i(\mathbf{x}_t) = \mathbf{x}_t - \eta_t(\nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}}) + \tilde{\boldsymbol{\mu}}).$$

The compassion of SAG, SVRG and SAGA

1. SAG (biased, 1 IFO):

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \left( \frac{\nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{x}_{i,t})}{n} + \frac{1}{n}\sum_{j=1}^n \nabla f_j(\mathbf{x}_{j,t}) \right).$$

2. SAGA (unbiased, 1 IFO):

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \left( \nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{x}_{i,t}) + \frac{1}{n}\sum_{j=1}^n \nabla f_j(\mathbf{x}_{j,t}) \right).$$

3. SVRG (unbiased, 2 IFO):

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \left( \nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}}) + \frac{1}{n}\sum_{j=1}^n \nabla f_j(\tilde{\mathbf{x}}) \right).$$

**Convergence Analysis of SVRG** The smoothness and convexity of $f_i$ means (Lemma 3.19)

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \leq 2L(f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle)$$

for any $\mathbf{x} \in \mathbb{R}^d$. Summing over $i = 1, \ldots, n$ and using $\nabla f(\mathbf{x}^*) = \mathbf{0}$, we obtain

$$\mathbb{E}_{i_t} \|\nabla f_{i_t}(\mathbf{x}) - \nabla f_{i_t}(\mathbf{x}^*)\|_2^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2$$

$$
\begin{aligned}
\leq & \frac{1}{n}\sum_{i=1}^{n} 2L(f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle) \\
\leq & 2L(f(\mathbf{x}) - f(\mathbf{x}^*)).
\end{aligned}
$$

Let

$$
\mathbf{v}_t = \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \tilde{\boldsymbol{\mu}}.
$$

Conditioned on $\mathbf{x}_t$, we take expectation on $i_t$ and obtain

$$
\begin{aligned}
& \mathbb{E}_{i_t} \|\mathbf{v}_t\|_2^2 \\
\leq & 2\mathbb{E}_{i_t} \|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}^*)\|_2^2 + 2\mathbb{E}_{i_t} \|\nabla f_{i_t}(\mathbf{x}^*) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla f(\tilde{\mathbf{x}})\|_2^2 \\
= & 2\mathbb{E}_{i_t} \|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}^*)\|_2^2 + 2\mathbb{E}_{i_t} \|\nabla f_{i_t}(\mathbf{x}^*) - \nabla f_{i_t}(\tilde{\mathbf{x}}) - \mathbb{E}[\nabla f_{i_t}(\mathbf{x}^*) - \nabla f_{i_t}(\tilde{\mathbf{x}})]\|_2^2 \\
= & 2\mathbb{E}_{i_t} \|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}^*)\|_2^2 + 2\mathbb{E}_{i_t} \|\nabla f_{i_t}(\mathbf{x}^*) - \nabla f_{i_t}(\tilde{\mathbf{x}})\|_2^2 \\
\leq & 4L(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + 4L(f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)).
\end{aligned}
$$

We also have $\mathbb{E}[\mathbf{v}_t] = \nabla f(\mathbf{x}_t)$. Hence,

$$
\begin{aligned}
& \mathbb{E}_{i_t} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \\
= & \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta\mathbb{E}_{i_t}[\langle \mathbf{x}_t - \mathbf{x}^*, \mathbf{v}_t \rangle] + \eta^2 \mathbb{E}_{i_t} \|\mathbf{v}_t\|_2^2 \\
\leq & \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta\langle \mathbf{x}_t - \mathbf{x}^*, \nabla f(\mathbf{x}_t) \rangle + 4L\eta^2(f(\mathbf{x}_t) - f(\mathbf{x}^*) + f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)) \\
\leq & \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + 4L\eta^2(f(\mathbf{x}_t) - f(\mathbf{x}^*) - f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)) \\
= & \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - (2\eta - 4\eta^2 L)(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + 4L\eta^2(f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)).
\end{aligned}
$$

For stage the $s$-th stage, we let $\tilde{\mathbf{x}} = \mathbf{x}^{(s)}$ and $\mathbf{x}^{(s+1)}$ is sampled from $\{\mathbf{x}_0, \ldots, \mathbf{x}_{m-1}\}$. Summing above over $t = 0, \ldots, m-1$ and taking expectation with all the history, we have

$$
\begin{aligned}
\mathbb{E} \|\mathbf{x}_m - \mathbf{x}^*\|_2^2 \leq & \mathbb{E} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - 2\eta(1 - 2\eta L)\mathbb{E} \sum_{i=0}^{m-1} (f(\mathbf{x}_t) - f(\mathbf{x}^*)) + 4Lm\eta^2 \mathbb{E}[f(\tilde{\mathbf{x}}^{(s)}) - f(\mathbf{x}^*)] \\
= & \mathbb{E} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - 2\eta(1 - 2\eta L)m\mathbb{E}[f(\mathbf{x}^{(s+1)}) - f(\mathbf{x}^*)] + 4Lm\eta^2 \mathbb{E}[f(\tilde{\mathbf{x}}^{(s)}) - f(\mathbf{x}^*)].
\end{aligned}
$$

which means

$$
\begin{aligned}
& \mathbb{E} \|\mathbf{x}_m - \mathbf{x}^*\|_2^2 + 2\eta(1 - 2\eta L)m\mathbb{E}[f(\mathbf{x}^{(s+1)}) - f(\mathbf{x}^*)] \\
\leq & \mathbb{E} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + 4Lm\eta^2 \mathbb{E}[f(\tilde{\mathbf{x}}^{(s)}) - f(\mathbf{x}^*)] \\
= & \mathbb{E}\|\mathbf{x}^{(s)} - \mathbf{x}^*\|_2^2 + 4Lm\eta^2 \mathbb{E}[f(\tilde{\mathbf{x}}^{(s)}) - f(\mathbf{x}^*)] \\
\leq & \frac{2}{\mu}\mathbb{E}[f(\tilde{\mathbf{x}}^{(s)}) - f(\mathbf{x}^*)] + 4Lm\eta^2 \mathbb{E}[f(\tilde{\mathbf{x}}^{(s)}) - f(\mathbf{x}^*)] \\
\leq & \left(\frac{2}{\mu} + 4Lm\eta^2\right) \mathbb{E}[f(\tilde{\mathbf{x}}^{(s)}) - f(\mathbf{x}^*)].
\end{aligned}
$$

Thus we obtain

$$
\mathbb{E}[f(\mathbf{x}^{(s+1)}) - f(\mathbf{x}^*)] \leq \left(\frac{1}{\mu\eta(1 - 2\eta L)m} + \frac{2L\eta}{1 - 2\eta L}\right) \mathbb{E}[f(\mathbf{x}^{(s)}) - f(\mathbf{x}^*)]
$$

**Remark 11.1.** *For $\eta = \Theta(1/L)$ and $m = \Theta(\kappa)$, we have $\rho = \Theta(1) < 1$. Hence, achieving the $\epsilon$-suboptimal solution requires $S = \log(1/\epsilon)$ and IFO complexity is $S(m + n) = \mathcal{O}((n + \kappa)\log(1/\epsilon))$.*

**Convergence Analysis of L-SVRG** The update rule means

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2 \\
=&\mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\langle\mathbf{v}_t,\mathbf{x}_t-\mathbf{x}^*\rangle+\eta^2\left\|\mathbf{v}_t\right\|_2^2\right] \\
=&\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\langle\nabla f(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^*\rangle+\eta^2\mathbb{E}\left\|\mathbf{v}_t\right\|_2^2 \\
\leq&\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)+\frac{\mu}{2}\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right)+\eta^2\mathbb{E}\left\|\mathbf{v}_t\right\|_2^2 \\
=&(1-\mu\eta)\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)+\frac{\mu}{2}\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right)+\eta^2\mathbb{E}\left\|\mathbf{v}_t\right\|_2^2.
\end{aligned} \tag{59}$$

We bound the last term as follows

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{v}_t\right\|_2 \\
=&\mathbb{E}\left\|\nabla f_{i_t}(\mathbf{x}_t)-\nabla f_{i_t}(\mathbf{w}_t)+\nabla f(\mathbf{w}_t)\right\|_2^2 \\
=&\mathbb{E}\left\|\nabla f_{i_t}(\mathbf{x}_t)-\nabla f_{i_t}(\mathbf{x}^*)-(\nabla f_{i_t}(\mathbf{w}_t)-\nabla f_{i_t}(\mathbf{x}^*)-(\nabla f(\mathbf{w}_t)-\nabla f(\mathbf{x}^*)))\right\|_2^2 \\
\leq&2\mathbb{E}\left\|\nabla f_{i_t}(\mathbf{x}_t)-\nabla f_{i_t}(\mathbf{x}^*)\right\|_2^2+2\mathbb{E}\left\|\nabla f(\mathbf{w}_t)-\nabla f(\mathbf{x}^*)-(\nabla f_{i_t}(\mathbf{w}_t)-\nabla f_{i_t}(\mathbf{x}^*))\right\|_2^2 \\
\leq&2\mathbb{E}\left\|\nabla f_{i_t}(\mathbf{x}_t)-\nabla f_{i_t}(\mathbf{x}^*)\right\|_2^2+2\mathbb{E}\left\|\nabla f_{i_t}(\mathbf{w}_t)-\nabla f_{i_t}(\mathbf{x}^*)\right\|_2^2 \\
\leq&\frac{2}{n}\sum_{j=1}^n\left(2L(f_j(\mathbf{x}_t)-f_j(\mathbf{x}^*)-\langle\nabla f_j(\mathbf{x}^*),\mathbf{x}_t-\mathbf{x}^*\rangle)\right)+\frac{pD_k}{2\eta^2} \\
=&4L(f(\mathbf{x}_t)-f(\mathbf{x}^*))+\frac{pD_k}{2\eta^2},
\end{aligned} \tag{60}$$

where

$$D_k=\frac{4\eta^2}{pn}\sum_{j=1}^n\left\|\nabla f_j(\mathbf{w}_t)-\nabla f_j(\mathbf{x}^*)\right\|_2^2$$

and the last inequality is due to Lemma 3.19. We also have

$$\begin{aligned}
&\mathbb{E}[D_{t+1}] \\
=&\mathbb{E}\left[\frac{4\eta^2}{pn}\sum_{j=1}^n\left\|\nabla f_j(\mathbf{w}_{t+1})-\nabla f_j(\mathbf{x}^*)\right\|_2^2\right] \\
=&(1-p)\cdot\frac{4\eta^2}{pn}\sum_{j=1}^n\left\|\nabla f_j(\mathbf{w}_t)-\nabla f_j(\mathbf{x}^*)\right\|_2^2+p\cdot\frac{4\eta^2p}{n}\sum_{j=1}^n\left\|\nabla f_j(\mathbf{x}_t)-\nabla f_j(\mathbf{x}^*)\right\|_2^2 \\
\leq&(1-p)D_t+\frac{4\eta^2}{n}\sum_{j=1}^n2L(f_j(\mathbf{x}_t)-f_j(\mathbf{x}^*)-\langle\nabla f_j(\mathbf{x}^*),\mathbf{x}_t-\mathbf{x}^*\rangle) \\
=&(1-p)D_t+8\eta^2L(f(\mathbf{x}_t)-f(\mathbf{x}^*)),
\end{aligned} \tag{61}$$

where the inequality is due to Lemma 3.19. Let

$$\Phi_t=\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2+D_k.$$

Combining inequalities (59), (60) and (61), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\Phi_{t+1}\right]=\mathbb{E}\left[\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2+D_{t+1}\right] \\
\leq&(1-\mu\eta)\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)\right)+\eta^2\left(4L(f(\mathbf{x}_t)-f(\mathbf{x}^*))+\frac{pD_t}{2\eta^2}\right) \\
&+(1-p)D_k+8\eta^2L(f(\mathbf{x}_t)-f(\mathbf{x}^*))
\end{aligned}$$

$$\leq (1-\mu\eta)\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-\left(2\eta-4\eta^2L-8\eta^2L\right)\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)\right)+\left(\frac{p}{2}+1-p\right)D_k$$
$$\leq (1-\mu\eta)\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-\left(2\eta-12\eta^2L\right)\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)\right)+\left(1-\frac{p}{2}\right)D_t.$$

We let $\eta=1/(6L)$ and $p=1/n$, then we have

$$\mathbb{E}\left[\Phi_{t+1}\right]\leq\left(1-\frac{\mu}{6L}\right)\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2+\left(1-\frac{1}{2n}\right)D_t\leq\max\left\{1-\frac{\mu}{6L},1-\frac{1}{2n}\right\}\Phi_t.$$

Hence, we need to take

$$T=\mathcal{O}\left(\max\left\{\frac{6L}{\mu},2n\right\}\log\left(\frac{\Phi_0}{\epsilon}\right)\right)=\mathcal{O}\left((n+\kappa)\log\left(\frac{1}{\epsilon}\right)\right)$$

to guarantee

$$\mathbb{E}\left\|x_T-\mathbf{x}^*\right\|_2^2\leq\mathbb{E}\left\|\Phi_T\right\|_2^2\leq\left(\max\left\{1-\frac{\mu}{6L},1-\frac{1}{2n}\right\}\right)^K\Phi_0\leq\epsilon.$$

The total number of IFO is $(1-p+pn)K\leq\mathcal{O}(K)=\mathcal{O}\left((n+\kappa)\log\left(1/\epsilon\right)\right)$ in expectation.

**Theorem 11.1.** *Let $\alpha_0=\sqrt{q}$ with $q=\mu/(\mu+\beta)$ and $\epsilon_t=\frac{2}{9}(f(\mathbf{x}_0)-f^*)(1-\rho)^{t+1}$ with $\rho<\sqrt{q}$. Then Catalyst algorithm generates $\{\mathbf{x}_t\}$ such that*

$$f(\mathbf{x}_t)-f^*\leq\frac{8(1-\rho)^t}{(\sqrt{q}-\rho)^2}(f(\mathbf{x}_0)-f^*).$$

**Remark 11.2.** *We let $\rho=\Theta(\sqrt{q})$. For example, take $\rho=0.9\sqrt{q}=0.9\sqrt{\mu/(\mu+\beta)}$. The total iteration number is*

$$\tilde{\mathcal{O}}\left(\frac{1}{\rho}\right)=\tilde{\mathcal{O}}\left(\sqrt{\frac{\mu+\beta}{\mu}}\right).$$

1. *For GD ($\kappa\gg1$), each sub-problem calls require*

$$\tilde{\mathcal{O}}\left(\frac{L+\beta}{\mu+\beta}\right).$$

*Hence, the total number of FO calls is*

$$\sqrt{\frac{\mu+\beta}{\mu}}\cdot\frac{L+\beta}{\mu+\beta}=\frac{1}{\sqrt{\mu}}\frac{L+\beta}{\sqrt{\mu+\beta}}=\frac{1}{\sqrt{\mu}}\left(\frac{L-\mu}{\sqrt{\mu+\beta}}+\sqrt{\mu+\beta}\right)\geq\frac{2\sqrt{L-\mu}}{\sqrt{\mu}}=\mathcal{O}(\sqrt{\kappa}),$$

*where we require*

$$\frac{L-\mu}{\sqrt{\mu+\beta}}=\sqrt{\mu+\beta}\Longrightarrow\beta=L-2\mu=\Omega(L).$$

2. *For SVRG ($\kappa\geq\Omega(n)$), each sub-problem calls require*

$$\tilde{\mathcal{O}}\left(n+\frac{L+\beta}{\mu+\beta}\right).$$

*IFO calls. Hence, the total number of IFO calls is*

$$\sqrt{\frac{\mu+\beta}{\mu}}\cdot\left(n+\frac{L+\beta}{\mu+\beta}\right)=\frac{1}{\sqrt{\mu}}\left(n\sqrt{\mu+\beta}+\frac{L-\mu+\mu+\beta}{\sqrt{\mu+\beta}}\right)$$

$$=\frac{1}{\sqrt{\mu}}\left((n+1)\sqrt{\mu+\beta}+\frac{L-\mu}{\sqrt{\mu+\beta}}\right)\geq\frac{2}{\sqrt{\mu}}\sqrt{(n+1)(L-\mu)}=\Omega(\sqrt{\kappa n}\,),$$

*where we require*

$$(n+1)\sqrt{\mu+\beta}=\frac{L-\mu}{\sqrt{\mu+\beta}}\Longrightarrow(n+1)(\mu+\beta)=L-\mu\Longrightarrow\beta=\frac{L-\mu}{n+1}-\mu.$$

**Remark 11.3.** *For $n\geq\Omega(\kappa)$, the acceleration provide no benefit. Hence, we desire the total IFO complexity of $\mathcal{O}((n+\sqrt{\kappa n}\,)\log(1/\epsilon))$.*

**SGD for Nonconvex Optimization** We consider the SGD iteration

$$\mathbf{x}_{t+1}=\mathbf{x}_t-\eta_t\cdot\frac{1}{b}\sum_{i=1}^{b}\nabla F(\mathbf{x}_t;\xi_{t_i}).$$

We suppose $f(\cdot)$ is $L$-smooth and lower bounded by $f^*$, and there exists $\sigma>0$ such that

$$\mathbb{E}\left\|\nabla F(\mathbf{x};\xi)-\nabla f(\mathbf{x})\right\|_2^2\leq\sigma^2$$

for any $\mathbf{x}\in\mathbb{R}^d$. It implies

$$\mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla F(\mathbf{x};\xi_i)-\nabla f(\mathbf{x})\right\|_2^2=\frac{1}{b}\mathbb{E}\left\|\nabla F(\mathbf{x};\xi_i)-\nabla f(\mathbf{x})\right\|_2^2\leq\frac{\sigma^2}{b}.$$

Conditioned on $\mathbf{x}_t$, we have

$$\begin{aligned}
\mathbb{E}_t[f(\mathbf{x}_{t+1})]\leq&f(\mathbf{x}_t)-\mathbb{E}_t\langle\nabla f(\mathbf{x}_t),\mathbf{x}_{t+1}-\mathbf{x}_t\rangle+\frac{L}{2}\mathbb{E}_t\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2\\
=&f(\mathbf{x}_t)-\eta\mathbb{E}_t\left\langle\nabla f(\mathbf{x}_t),\frac{1}{b}\sum_{i=1}^{b}\nabla F(\mathbf{x}_t;\xi_{t_i})\right\rangle+\frac{L\eta^2}{2}\mathbb{E}_t\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla F(\mathbf{x}_t;\xi_{t_i})\right\|_2^2\\
\leq&f(\mathbf{x}_t)-\eta\left\|\nabla f(\mathbf{x}_t)\right\|_2^2+L\eta^2\left(\left\|\nabla f(\mathbf{x}_t)\right\|_2^2+\mathbb{E}_t\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla F(\mathbf{x}_t;\xi_i)-\nabla f(\mathbf{x}_t)\right\|_2^2\right)\\
\leq&f(\mathbf{x}_t)-(\eta-L\eta^2)\left\|\nabla f(\mathbf{x}_t)\right\|_2^2+\frac{L\eta^2\sigma^2}{b}.
\end{aligned}$$

Let $\eta=1/(2L)$ and $b=2\sigma^2\epsilon^{-2}$, then

$$\mathbb{E}_t[f(\mathbf{x}_{t+1})]\leq f(\mathbf{x}_t)-\frac{1}{4L}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2+\frac{\epsilon^2}{8L}\quad\Longrightarrow\quad\left\|\nabla f(\mathbf{x}_t)\right\|_2^2\leq4L(f(\mathbf{x}_t)-\mathbb{E}_t[f(\mathbf{x}_{t+1})])+\frac{\epsilon^2}{2}.$$

Let $\mathbf{x}_{\text{out}}=\mathbf{x}_j$ with $j$ uniformly sampled from $\{0,\ldots,T-1\}$ and $T=\lceil 8L(f(\mathbf{x}_0)-f^*)\epsilon^{-2}\rceil$. Taking expectation on all of history and averaging over $t=0,\ldots,T-1$, we have

$$\begin{aligned}
\mathbb{E}\left\|\nabla f(\mathbf{x}_{out})\right\|_2^2\leq&\frac{4L(f(\mathbf{x}_0)-\mathbb{E}[f(\mathbf{x}_T)])}{T}+\frac{\epsilon^2}{2}\\
\leq&\frac{4L(f(\mathbf{x}_0)-f^*)}{T}+\frac{\epsilon^2}{2}\\
\leq&\frac{\epsilon^2}{2}+\frac{\epsilon^2}{2}=\epsilon^2.
\end{aligned}$$

**PAGE** We consider the $L$-average smooth function, i.e., there exists $L > 0$ such that

$$\mathbb{E}\left\|\nabla f(\mathbf{x};\xi) - \nabla f(\mathbf{y};\xi)\right\|_2^2 \leq L^2 \left\|\mathbf{x}-\mathbf{y}\right\|_2^2$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

**Remark 11.4.** *Using Jensen's inequality, we have*

$$\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\|_2^2 = \left\|\mathbb{E}[\nabla f(\mathbf{x};\xi) - \nabla f(\mathbf{y};\xi)]\right\|_2^2 \leq \mathbb{E}\left\|\nabla f(\mathbf{x};\xi) - \nabla f(\mathbf{y};\xi)\right\|_2^2 \leq L^2 \left\|\mathbf{x}-\mathbf{y}\right\|_2^2.$$

**Lemma 11.1.** *For $L$-smooth function $f(\cdot)$, let $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$ for some $\eta > 0$. Then we have*

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2 + \frac{\eta}{2}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2. \tag{62}$$

*Proof.* Let $\bar{\mathbf{x}}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$. In view of $L$-smoothness of $f$, we have

$$\begin{aligned}
&f(\mathbf{x}_{t+1}) \\
\leq& f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2 \\
=& f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t) - \mathbf{v}_t, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \langle \mathbf{v}_t, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2 \\
=& f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t) - \mathbf{v}_t, -\eta \mathbf{v}_t\rangle - \left(\frac{1}{\eta} - \frac{L}{2}\right)\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2 \\
=& f(\mathbf{x}_t) + \eta\left\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\right\|_2^2 - \eta\langle \nabla f(\mathbf{x}_t) - \mathbf{v}_t, \nabla f(\mathbf{x}_t)\rangle - \left(\frac{1}{\eta} - \frac{L}{2}\right)\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2 \\
=& f(\mathbf{x}_t) + \eta\left\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\right\|_2^2 - \frac{\eta}{2}\Big(\left\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\right\|_2^2 + \left\|\nabla f(\mathbf{x}_t)\right\|_2^2 - \frac{1}{\eta^2}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2\Big) - \left(\frac{1}{\eta} - \frac{L}{2}\right)\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2 \\
=& f(\mathbf{x}_t) - \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2 + \frac{\eta}{2}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2.
\end{aligned}$$

□

**Lemma 11.2.** *For update rule*

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \frac{1}{b}\sum_{\xi\in\mathcal{S}_{t+1}}\left(\nabla F(\mathbf{x}_{t+1};\xi) - \nabla F(\mathbf{x}_t;\xi)\right)$$

*in SARAH, we have*

$$\mathbb{E}\left\|\mathbf{v}_{t+1} - \nabla f(\mathbf{x}_{t+1})\right\|_2^2 \leq (1-p)\mathbb{E}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{(1-p)L^2}{b}\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2 + \frac{p\sigma^2}{b_0}.$$

*Proof.* We first consider the case of

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \frac{1}{b}\sum_{\xi\in\mathcal{S}_{t+1}}\left(\nabla F(\mathbf{x}_{t+1};\xi) - \nabla F(\mathbf{x}_t;\xi)\right).$$

Conditioned on $\mathbf{x}_0, \ldots, \mathbf{x}_{t+1}$ and $\mathbf{v}_0, \ldots, \mathbf{v}_t$, we have

$$\mathbb{E}_{\mathcal{S}_{t+1}}\left[\mathbf{v}_{t+1} - \mathbf{v}_t\right] = \mathbb{E}_{\mathcal{S}_{t+1}}\left[\frac{1}{b}\sum_{\xi\in\mathcal{S}_{t+1}}\left(\nabla F(\mathbf{x}_{t+1};\xi) - \nabla F(\mathbf{x}_t;\xi)\right)\right] = \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t).$$

Hence, we obtain

$$\mathbb{E}\left\|\mathbf{v}_{t+1} - \nabla f(\mathbf{x}_{t+1})\right\|_2^2$$

$$
\begin{aligned}
=&\mathbb{E}\left\|\mathbf{v}_t+\frac{1}{b}\sum_{\xi\in\mathcal{S}_{t+1}}(\nabla F(\mathbf{x}_{t+1};\xi)-\nabla F(\mathbf{x}_t;\xi))-\nabla f(\mathbf{x}_{t+1})\right\|_2^2\\
=&\mathbb{E}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)+\frac{1}{b}\sum_{\xi\in\mathcal{S}_{t+1}}(\nabla F(\mathbf{x}_{t+1};\xi)-\nabla F(\mathbf{x}_t;\xi))-(\nabla f(\mathbf{x}_{t+1})-\nabla f(\mathbf{x}_t))\right\|_2^2\\
=&\mathbb{E}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2+\mathbb{E}\left\|\frac{1}{b}\sum_{\xi\in\mathcal{S}_{t+1}}(\nabla F(\mathbf{x}_{t+1};\xi)-\nabla F(\mathbf{x}_t;\xi))-(\nabla f(\mathbf{x}_{t+1})-\nabla f(\mathbf{x}_t))\right\|_2^2\\
&+\left\langle\mathbf{v}_t-\nabla f(\mathbf{x}_t),\ \frac{1}{b}\sum_{\xi\in\mathcal{S}_{t+1}}(\nabla F(\mathbf{x}_{t+1};\xi)-\nabla F(\mathbf{x}_t;\xi))-(\nabla f(\mathbf{x}_{t+1})-\nabla f(\mathbf{x}_t))\right\rangle\\
=&\mathbb{E}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2+\mathbb{E}\left\|\frac{1}{b}\sum_{\xi\in\mathcal{S}_{t+1}}(\nabla F(\mathbf{x}_{t+1};\xi)-\nabla F(\mathbf{x}_t;\xi))-(\nabla f(\mathbf{x}_{t+1})-\nabla f(\mathbf{x}_t))\right\|_2^2\\
=&\mathbb{E}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2+\frac{1}{b}\mathbb{E}\left\|\nabla F(\mathbf{x}_{t+1};\xi)-\nabla F(\mathbf{x}_t;\xi)-(\nabla f(\mathbf{x}_{t+1})-\nabla f(\mathbf{x}_t))\right\|_2^2\\
\le&\mathbb{E}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2+\frac{1}{b}\mathbb{E}\left\|\nabla F(\mathbf{x}_{t+1};\xi)-\nabla F(\mathbf{x}_t;\xi)\right\|_2^2\\
\le&\mathbb{E}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2+\frac{L^2}{b}\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2.
\end{aligned}
$$

For the other case, we have

$$
\mathbf{v}_{t+1}=\frac{1}{b_0}\sum_{\xi\in\mathcal{S}_{t+1}}\nabla F(\mathbf{x}_{t+1};\xi),
$$

which implies

$$
\mathbb{E}\left\|\mathbf{v}_{t+1}-\nabla f(\mathbf{x}_{t+1})\right\|_2^2=\mathbb{E}\left\|\frac{1}{b_0}\sum_{\xi\in\mathcal{S}_{t+1}}\nabla F(\mathbf{x}_{t+1};\xi)-\nabla f(\mathbf{x}_{t+1})\right\|_2^2\le\frac{\sigma^2}{b_0}.
$$

Hence, we have

$$
\mathbb{E}\left\|\mathbf{v}_{t+1}-\nabla f(\mathbf{x}_{t+1})\right\|_2^2=(1-p)\left(\mathbb{E}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2+\frac{L^2}{b}\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2\right)+\frac{p\sigma^2}{b_0}.
$$

□

Let

$$
\Phi_t=f(\mathbf{x}_t)+\frac{\eta}{2p}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2
$$

Using above two lemmas, we have

$$
\begin{aligned}
\mathbb{E}[\Phi_{t+1}]=&\mathbb{E}\left[f(\mathbf{x}_{t+1})+\frac{\eta}{2p}\left\|\mathbf{v}_{t+1}-\nabla f(\mathbf{x}_{t+1})\right\|_2^2\right]\\
\le&\mathbb{E}\Bigg[f(\mathbf{x}_t)-\frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2-\left(\frac{1}{2\eta}-\frac{L}{2}\right)\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2+\frac{\eta}{2}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2\\
&+\frac{\eta}{2p}\left((1-p)\mathbb{E}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2+\frac{(1-p)L^2}{b}\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2\right)+\frac{p\sigma^2}{b_0}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\leq & \mathbb{E}\left[f(\mathbf{x}_t)+\frac{\eta}{2p}\mathbb{E}\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_2^2-\frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2-\left(\frac{1}{2\eta}-\frac{L}{2}-\frac{(1-p)L^2\eta}{2pb}\right)\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2+\frac{\eta\sigma^2}{2b_0}\right] \\
\leq & \mathbb{E}\left[\Phi_t-\frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2+\frac{\eta\sigma^2}{2b_0}\right],
\end{aligned}
$$

where we take the parameters satisfying

$$
\frac{1}{2\eta}-\frac{L}{2}-\frac{(1-p)L^2\eta}{2pb}\geq 0,
$$

which can be obtained by taking $(1-p)/(bp)\leq 1$ and $\eta=1/(2L)$. It implies

$$
\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2\leq\frac{2}{\eta}\mathbb{E}\left[\Phi_t-\Phi_{t+1}+\frac{\eta\sigma^2}{2b_0}\right].
$$

Taking the average over $t=0,\ldots,T-1$, we obtain

$$
\mathbb{E}\left\|\nabla f(\mathbf{x}_{\text{out}})\right\|_2^2=\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2\right]\leq\frac{2}{\eta T}\mathbb{E}[\Phi_0-\Phi_T]+\frac{\sigma^2}{b_0}.
$$

We also have

$$
\begin{aligned}
& \Phi_0-\Phi_T \\
= & f(\mathbf{x}_0)+\frac{\eta}{2p}\left\|\mathbf{v}_0-\nabla f(\mathbf{x}_0)\right\|_2^2-\left(f(\mathbf{x}_T)+\frac{\eta}{2p}\left\|\mathbf{v}_T-\nabla f(\mathbf{x}_T)\right\|_2^2\right) \\
\leq & f(\mathbf{x}_0)-f^*+\frac{\eta}{2p}\left\|\mathbf{v}_0-\nabla f(\mathbf{x}_0)\right\|_2^2 \\
\leq & f(\mathbf{x}_0)-f^*+\frac{\eta\sigma^2}{2pb_0},
\end{aligned}
$$

which means (taking $b_0=2\sigma^2\epsilon^{-2}$)

$$
\begin{aligned}
\mathbb{E}\left\|\nabla f(\mathbf{x}_{\text{out}})\right\|_2^2\leq & \frac{2}{\eta T}\left(f(\mathbf{x}_0)-f^*+\frac{\eta\sigma^2}{2pb_0}\right)+\frac{\sigma^2}{b_0} \\
\leq & \frac{2(f(\mathbf{x}_0)-f^*)}{\eta T}+\frac{\sigma^2}{pb_0T}+\frac{\sigma^2}{b_0} \\
= & \frac{4L(f(\mathbf{x}_0)-f^*)}{T}+\frac{\epsilon^2}{2pT}+\frac{\epsilon^2}{2}.
\end{aligned}
$$

We desire RHS be $\epsilon^2$, which leads to $\mathbb{E}\left\|\nabla f(\mathbf{x}_{\text{out}})\right\|_2\leq\sqrt{\mathbb{E}\left\|\nabla f(\mathbf{x}_{\text{out}})\right\|_2^2}\leq\epsilon$. We take

$$
T=16L\epsilon^{-2}(f(\mathbf{x}_0)-f^*)+\frac{2}{p}\qquad\text{and}\qquad\eta=\frac{1}{2L}
$$

then

$$
\mathbb{E}\left\|\nabla f(\mathbf{x}_{\text{out}})\right\|_2^2\leq\frac{2(f(\mathbf{x}_0)-f^*)}{\eta}\cdot\frac{\epsilon^2}{16L(f(\mathbf{x}_0)-f^*)}+\frac{\epsilon^2}{2p}\cdot\frac{p}{2}+\frac{\epsilon^2}{2}=\frac{\epsilon^2}{4}+\frac{\epsilon^2}{4}+\frac{\epsilon^2}{2}=\epsilon^2.
$$

The condition $(1-p)/(bp)\leq 1$ can be attained by taking $b=\lceil\sigma\epsilon^{-1}\rceil$ and $p=1/b$. The expected total SFO complexity is

$$
\begin{aligned}
b_0+T(b_0p+b(1-p))\leq & 2\sigma^2\epsilon^{-2}+\left(16L\epsilon^{-2}(f(\mathbf{x}_0)-f^*)+\frac{2}{p}\right)\left(\frac{2\sigma^2\epsilon^{-2}}{\sigma\epsilon^{-1}}+\sigma\epsilon^{-1}\right) \\
\leq & 2\sigma^2\epsilon^{-2}+\left(16L\epsilon^{-2}(f(\mathbf{x}_0)-f^*)+2\sigma\epsilon^{-1}\right)3\sigma\epsilon^{-1} \\
\leq & \mathcal{O}(\sigma^2\epsilon^{-2}+L\sigma\epsilon^{-3})
\end{aligned}
$$

**Remark 11.5.** *The value of $b$ can be selected by minimizing $b_0 p + b$ with constraint $bp = 1$. That is*

$$b_0 p + b = \frac{b_0}{b} + b \geq 2\sqrt{b_0},$$

*where the equality is taken by $b = \sqrt{b_0}$.*

**Remark 11.6.** *Similarly, we take $b_0 = n$ and $b = \Theta(\sqrt{n}\,)$ for finite-sum case.*

# 12 Zeroth-Order Optimization

Given continuous function $f : \mathbb{R}^d \to \mathbb{R}$, we define its Gaussian smoothing as

$$f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}}[f(\mathbf{x} + \delta\mathbf{u})] = \int \frac{1}{(2\pi)^{d/2}} f(\mathbf{x} + \delta\mathbf{u}) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u}$$

for $\delta > 0$, where $\mathbf{u} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$. Since the term $f(\mathbf{x} + \delta\mathbf{u}) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right)$ is continuous w.r.t $\mathbf{x}$ and $\mathbf{u}$, the function $f_\delta(\mathbf{x})$ is differentiable. By introducing $\mathbf{y} = \mathbf{x} + \delta\mathbf{u}$, we have

$$\begin{aligned} f_\delta(\mathbf{x}) &= \int \frac{1}{(2\pi)^{d/2}} f(\mathbf{x} + \delta\mathbf{u}) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} \\ &= \int \frac{1}{(2\pi)^{d/2}} f(\mathbf{y}) \exp\left(-\frac{1}{2\delta^2}\|\mathbf{y} - \mathbf{x}\|_2^2\right) \frac{1}{\delta^d} \mathrm{d}\mathbf{y}, \end{aligned}$$

where we use the fact that Jacobian of $\mathbf{u} = (\mathbf{y} - \mathbf{x})/\delta$ is $\delta^{-1}\mathbf{I}$ and $\det(\delta^{-1}\mathbf{I}) = \delta^{-d}$. The continuity means

$$\begin{aligned} \nabla f_\delta(\mathbf{x}) =& \int \frac{1}{(2\pi)^{d/2}\delta^d} f(\mathbf{y}) \cdot \frac{\partial}{\partial \mathbf{x}} \left(\exp\left(-\frac{1}{2\delta^2}\|\mathbf{y} - \mathbf{x}\|_2^2\right)\right) \mathrm{d}\mathbf{y} \\ =& \int \frac{1}{(2\pi)^{d/2}\delta^d} f(\mathbf{y}) \cdot \exp\left(-\frac{1}{2\delta^2}\|\mathbf{y} - \mathbf{x}\|_2^2\right) \cdot \frac{1}{\delta^2}(\mathbf{y} - \mathbf{x}) \mathrm{d}\mathbf{y} \\ =& \int \frac{1}{(2\pi)^{d/2}\delta^{d+2}} f(\mathbf{y}) \cdot \exp\left(-\frac{1}{2\delta^2}\|\mathbf{y} - \mathbf{x}\|_2^2\right) (\mathbf{y} - \mathbf{x}) \mathrm{d}\mathbf{y} \\ =& \int \frac{1}{(2\pi)^{d/2}\delta^{d+2}} f(\mathbf{x} + \delta\mathbf{u}) \cdot \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \delta\mathbf{u} \cdot \delta^d \mathrm{d}\mathbf{u} \\ =& \int \frac{1}{(2\pi)^{d/2}} \frac{f(\mathbf{x} + \delta\mathbf{u})}{\delta} \cdot \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathbf{u} \, \mathrm{d}\mathbf{u} \\ =& \int \frac{1}{(2\pi)^{d/2}} \frac{f(\mathbf{x} + \delta\mathbf{u}) - f(\mathbf{x})}{\delta} \cdot \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathbf{u} \, \mathrm{d}\mathbf{u} \\ =& \mathbb{E}\left[\frac{f(\mathbf{x} + \delta\mathbf{u}) - f(\mathbf{x})}{\delta} \cdot \mathbf{u}\right]. \end{aligned} \tag{63}$$

The symmetric of $\mathbf{u}$ means we also have

$$\nabla f_\delta(\mathbf{x}) = \mathbb{E}\left[\frac{f(\mathbf{x}) - f(\mathbf{x} - \delta\mathbf{u})}{\delta} \cdot \mathbf{u}\right], \tag{64}$$

which implies

$$\nabla f_\delta(\mathbf{x}) = \mathbb{E}\left[\frac{f(\mathbf{x} + \delta\mathbf{u}) - f(\mathbf{x} - \delta\mathbf{u})}{2\delta} \cdot \mathbf{u}\right]. \tag{65}$$

**Theorem 12.1** (estimation)**.** *We can bound the estimation error of between Gaussian smoothing as follows:*

1. *If $f(\cdot)$ is $G$-Lipschitz continuous, then we have $|f_\delta(\mathbf{x}) - f(\mathbf{x})| \leq \delta G\sqrt{d}$ for any $\mathbf{x} \in \mathbb{R}^d$.*

*2. If If $f(\cdot)$ is $L$-smooth, we have*

$$|f_\delta(\mathbf{x}) - f(\mathbf{x})| \leq \frac{L\delta^2 d}{2} \qquad \textit{and} \qquad \|\nabla f_\delta(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \leq \frac{L\delta(d+3)^{3/2}}{2}$$

*for any $\mathbf{x} \in \mathbb{R}^d$.*

*Proof.* **Part I:** For any $\mathbf{x} \in \mathbb{R}^d$, we have

$$|f_\delta(\mathbf{x}) - f(\mathbf{x})| = \left|\mathbb{E}[f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x})]\right| \leq \mathbb{E}\left[|f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x})|\right] \leq \delta G \mathbb{E}\left\|\mathbf{u}\right\|_2. \tag{66}$$

Using Jensen's inequality, we have

$$\left(\mathbb{E}\left\|\mathbf{u}\right\|_2\right)^2 \leq \mathbb{E}\left\|\mathbf{u}\right\|_2^2 = \mathbb{E}[\mathrm{tr}(\mathbf{u}^\top \mathbf{u})] = \mathrm{tr}(\mathbb{E}[\mathbf{u}\mathbf{u}^\top]) = \mathrm{tr}(\mathbf{I}) = d. \tag{67}$$

Connecting inequalities (66) and (67), we obtain $|f_\delta(\mathbf{x}) - f(\mathbf{x})| \leq \delta G\sqrt{d}$.

**Part II:** The smoothness means

$$-\frac{\delta^2 L}{2}\left\|\mathbf{u}\right\|_2^2 \leq f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \delta\mathbf{u}\rangle \leq \frac{\delta^2 L}{2}\left\|\mathbf{u}\right\|_2^2.$$

Taking expectation on above results and using the fact $\mathbb{E}[\mathbf{u}] = \mathbf{0}$, we have

$$-\frac{\delta^2 L}{2}\mathbb{E}\left\|\mathbf{u}\right\|_2^2 \leq \mathbb{E}\left[f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x})\right] \leq \frac{\delta^2 L}{2}\mathbb{E}\left\|\mathbf{u}\right\|_2^2, \tag{68}$$

which means

$$\left|\mathbb{E}[f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x})]\right| \leq \frac{\delta^2 L}{2}\mathbb{E}\left\|\mathbf{u}\right\|_2^2 \leq \frac{\delta^2 L}{2}\mathbb{E}\left[\mathrm{tr}(\mathbf{u}^\top\mathbf{u})\right] = \frac{\delta^2 L}{2}\mathrm{tr}\left(\mathbb{E}[\mathbf{u}\mathbf{u}^\top]\right) = \frac{\delta^2 L}{2}\mathrm{tr}(\mathbf{I}) = \frac{\delta^2 L d}{2}.$$

**Part III:** Finally, we consider the difference of gradients. The equation (63) means

$$\nabla f(\mathbf{x}) = \lim_{\delta\to 0^+}\nabla f_\delta(\mathbf{x}) = \lim_{\delta\to 0^+}\mathbb{E}\left[\frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x})}{\delta}\cdot\mathbf{u}\right] = \mathbb{E}\left[\langle\nabla f(\mathbf{x}), \mathbf{u}\rangle\cdot\mathbf{u}\right],$$

which implies

$$\begin{aligned}
&\left\|\nabla f_\delta(\mathbf{x}) - \nabla f(\mathbf{x})\right\|_2 \\
=&\left\|\mathbb{E}\left[\frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x})}{\delta}\cdot\mathbf{u} - \langle\nabla f(\mathbf{x}), \mathbf{u}\rangle\cdot\mathbf{u}\right]\right\|_2 \\
\leq&\mathbb{E}\left\|\frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x}) - \langle\nabla f(\mathbf{x}), \delta\mathbf{u}\rangle}{\delta}\cdot\mathbf{u}\right\|_2 \\
\leq&\mathbb{E}\left[\frac{L\delta^2\left\|\mathbf{u}\right\|_2^2}{2\delta}\cdot\left\|\mathbf{u}\right\|_2\right] \leq \frac{L\delta}{2}\mathbb{E}\left\|\mathbf{u}\right\|_2^3 \leq \frac{L\delta(d+3)^{3/2}}{2},
\end{aligned}$$

where the last step is because of $d^{p/2} \leq \mathbb{E}\left\|\mathbf{u}\right\|_2^p \leq (p+d)^{p/2}$ for $p \geq 2$. □

**Remark 12.1.** *Let $M_p = \mathbb{E}\left\|\mathbf{u}\right\|_2^p$. Fix some $\tau \in [0,1]$, we can verify (compute derivative of $t$)*

$$t^p \exp\left(-\frac{\tau t^2}{2}\right) \leq \left(\frac{p}{\tau \mathrm{e}}\right)^{p/2}$$

*for any $t \geq 0$. Taking $t = \left\|\mathbf{u}\right\|_2$, we have*

$$M_p = \int \frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}\left\|\mathbf{u}\right\|_2^2\right)\left\|\mathbf{u}\right\|_2^p \mathrm{d}\mathbf{u}$$

$$
\begin{aligned}
&= \int \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\tau}{2}\|\mathbf{u}\|_2^2\right) \exp\left(-\frac{1-\tau}{2}\|\mathbf{u}\|_2^2\right) \|\mathbf{u}\|_2^p \,\mathrm{d}\mathbf{u} \\
&\le \int \frac{1}{(2\pi)^{d/2}} \left(\frac{p}{\tau \mathrm{e}}\right)^{p/2} \exp\left(-\frac{1-\tau}{2}\|\mathbf{u}\|_2^2\right) \|\mathbf{u}\|_2^p \,\mathrm{d}\mathbf{u} \\
&= \left(\frac{p}{\tau \mathrm{e}}\right)^{p/2} \frac{1}{(1-\tau)^{d/2}}.
\end{aligned}
$$

*The minimum of right-hand side in $\tau \in (0,1)$ is attained at $\tau = p/(p+d)$. Thus,*

$$
\begin{aligned}
M_p &\le \left(\frac{p}{\tau \mathrm{e}}\right)^{p/2} \frac{1}{(1-\tau)^{d/2}} \le \left(\frac{p+d}{\mathrm{e}}\right)^{p/2} \left(\frac{p+d}{d}\right)^{d/2} \\
&\le (p+d)^{p/2} \left(\frac{1}{\mathrm{e}}\right)^{p/2} \left(\left(1+\frac{p}{d}\right)^{d/p}\right)^{p/2} \le (p+d)^{p/2}.
\end{aligned} \tag{69}
$$

**Theorem 12.2** (smoothness)**.** *Gaussian smoothing has the following properties:*

1. *If $f(\cdot)$ is $G$-Lipschitz continuous, then $f_\delta(\cdot)$ is $G$-Lipschitz continuous and $G\sqrt{d}/\delta$-smooth.*
2. *If $f(\cdot)$ is $L$-smooth, then $f_\delta(\cdot)$ is $L$-smooth.*

*Proof.* **Part I:** For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$
\begin{aligned}
|f_\delta(\mathbf{x}) - f_\delta(\mathbf{y})| &= \left| \frac{1}{(2\pi)^{d/2}} \int (f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{y}+\delta\mathbf{u})) \exp\left(-\frac{\|\mathbf{u}\|_2^2}{2}\right) \mathrm{d}\mathbf{u} \right| \\
&\le \frac{1}{(2\pi)^{d/2}} \int \left| (f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{y}+\delta\mathbf{u})) \exp\left(-\frac{\|\mathbf{u}\|_2^2}{2}\right) \right| \mathrm{d}\mathbf{u} \\
&\le \frac{1}{(2\pi)^{d/2}} \int G\|\mathbf{x}-\mathbf{y}\|_2 \exp\left(-\frac{\|\mathbf{u}\|_2^2}{2}\right) \mathrm{d}\mathbf{u} = G\|\mathbf{x}-\mathbf{y}\|_2.
\end{aligned}
$$

We also have

$$
\begin{aligned}
&\|\nabla f_\delta(\mathbf{x}) - \nabla f_\delta(\mathbf{y})\|_2 = \left\| \mathbb{E}\left[ \frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{y}+\delta\mathbf{u})}{\delta} \cdot \mathbf{u} \right] \right\|_2 \\
&= \left\| \int \frac{1}{(2\pi)^{d/2}} \frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{y}+\delta\mathbf{u})}{\delta} \cdot \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathbf{u} \right\|_2 \mathrm{d}\mathbf{u} \\
&\le \int \frac{1}{(2\pi)^{d/2}} \left\| \frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{y}+\delta\mathbf{u})}{\delta} \right\|_2 \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \|\mathbf{u}\|_2 \,\mathrm{d}\mathbf{u} \\
&\le \frac{G\|\mathbf{x}-\mathbf{y}\|_2}{\delta} \int \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \|\mathbf{u}\|_2 \,\mathrm{d}\mathbf{u} \le \frac{G\sqrt{d}\,\|\mathbf{x}-\mathbf{y}\|_2}{\delta},
\end{aligned}
$$

where the last step uses inequity (67).

**Part II:** For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$
\nabla f_\delta(\mathbf{x}) = \int \frac{1}{(2\pi)^{d/2}} \nabla f(\mathbf{x}+\delta\mathbf{u}) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} \quad \text{and} \quad \nabla f_\delta(\mathbf{y}) = \int \frac{1}{(2\pi)^{d/2}} \nabla f(\mathbf{y}+\delta\mathbf{u}) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u},
$$

then

$$
\begin{aligned}
&\|\nabla f_\delta(\mathbf{x}) - \nabla f_\delta(\mathbf{y})\|_2 \\
&\le \left\| \int \frac{1}{(2\pi)^{d/2}} (\nabla f(\mathbf{y}+\delta\mathbf{u}) - \nabla f(\mathbf{y}+\delta\mathbf{u})) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} \right\|_2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \int \frac{1}{(2\pi)^{d/2}} \|\nabla f(\mathbf{y}+\delta\mathbf{u}) - \nabla f(\mathbf{y}+\delta\mathbf{u})\|_2 \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} \\
&\leq L\|\mathbf{x}-\mathbf{y}\|_2 \int \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} \\
&= L\|\mathbf{x}-\mathbf{y}\|_2 .
\end{aligned}
$$

□

**Theorem 12.3.** *If $f(\cdot)$ is convex, then $f_\delta(\cdot)$ is convex and $f_\delta(\mathbf{x}) \geq f(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$.*

*Proof.* For any $t \in [0,1]$, we have

$$
\begin{aligned}
f_\delta(t\mathbf{x}+(1-t)\mathbf{y}) &= \int \frac{1}{(2\pi)^{d/2}} f(t\mathbf{x}+(1-t)\mathbf{y}+\delta\mathbf{u}) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} \\
&\leq \int \frac{1}{(2\pi)^{d/2}} \left(tf(\mathbf{x}+\delta\mathbf{u}) + (1-t)f(\mathbf{y}+\delta\mathbf{u})\right) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} \\
&= tf_\delta(\mathbf{x}) + (1-t)f_\delta(\mathbf{y}).
\end{aligned}
$$

We also have

$$
\begin{aligned}
f_\delta(\mathbf{x}) &= \int \frac{1}{(2\pi)^{d/2}} f(\mathbf{x}+\delta\mathbf{u}) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} \\
&\geq \int \frac{1}{(2\pi)^{d/2}} \left(f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \delta\mathbf{u}\rangle\right) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} \\
&= \int \frac{1}{(2\pi)^{d/2}} f(\mathbf{x}) \exp\left(-\frac{1}{2}\|\mathbf{u}\|_2^2\right) \mathrm{d}\mathbf{u} = f(\mathbf{x}).
\end{aligned}
$$

□

**Theorem 12.4.** *Given $f : \mathbb{R}^d \to \mathbb{R}$, we define*

$$
\mathbf{g}_\delta(\mathbf{x};\mathbf{u}) = \frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x})}{\delta} \cdot \mathbf{u}.
$$

1. *If $f(\cdot)$ is $G$-Lipschitz continuous, then $\mathbb{E}\|\mathbf{g}_\delta(\mathbf{x};\mathbf{u})\|_2^2 \leq G^2(d+4)^2$.*
2. *If $f(\cdot)$ is $L$-smooth, then $\mathbb{E}\|\mathbf{g}_\delta(\mathbf{x};\mathbf{u})\|_2^2 \leq \frac{L^2\delta^2(d+6)^3}{2} + 2(d+4)\|\nabla f(\mathbf{x})\|_2^2$.*

*Proof.* **Part I:** We have

$$
\begin{aligned}
\mathbb{E}\|\mathbf{g}_\delta(\mathbf{x};\mathbf{u})\|_2^2 &= \frac{1}{\delta^2}\mathbb{E}\left[(f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x}))^2 \|\mathbf{u}\|_2^2\right] \\
&\leq \frac{1}{\delta^2}\mathbb{E}\left[(G\delta\|\mathbf{u}\|_2)^2 \|\mathbf{u}\|_2^2\right] \\
&= G^2\mathbb{E}\|\mathbf{u}\|_2^4 \leq G^2(d+4)^2,
\end{aligned}
$$

where the last step is because of inequality (69).

**Part II:** We have

$$
\begin{aligned}
\mathbb{E}\|\mathbf{g}_\delta(\mathbf{x};\mathbf{u})\|_2^2 &= \frac{1}{\delta^2}\mathbb{E}\left[(f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x}))^2 \|\mathbf{u}\|_2^2\right] \\
&= \frac{1}{\delta^2}\mathbb{E}\left[(f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x}) - \langle\nabla f(\mathbf{x}), \delta\mathbf{u}\rangle + \langle\nabla f(\mathbf{x}), \delta\mathbf{u}\rangle)^2 \|\mathbf{u}\|_2^2\right] \\
&\leq \frac{1}{\delta^2}\mathbb{E}\left[2(f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x}) - \langle\nabla f(\mathbf{x}), \delta\mathbf{u}\rangle)^2 \|\mathbf{u}\|_2^2 + 2(\langle\nabla f(\mathbf{x}), \delta\mathbf{u}\rangle)^2 \|\mathbf{u}\|_2^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{\delta^2}\mathbb{E}\left[2\left(\frac{L\delta^2\|\mathbf{u}\|_2^2}{2}\right)^2\|\mathbf{u}\|_2^2 + 2\delta^2\|\nabla f(\mathbf{x})\|_2^2\|\mathbf{u}\|_2^2\|\mathbf{u}\|_2^2\right] \\
&= \frac{L^2\delta^2}{2}\mathbb{E}\|\mathbf{u}\|_2^6 + 2\|\nabla f(\mathbf{x})\|_2^2\mathbb{E}\|\mathbf{u}\|_2^4 \\
&\leq \frac{L^2\delta^2(d+6)^3}{2} + 2(d+4)\|\nabla f(\mathbf{x})\|_2^2.
\end{aligned}
$$

□

**Remark 12.2.** *Recall that equation (63) means*

$$\mathbb{E}[\mathbf{g}_\delta(\mathbf{x};\mathbf{u})] = \nabla f_\delta(\mathbf{x}).$$

*Hence, the iteration with $\mathbf{g}_\delta(\mathbf{x};\mathbf{u})$ can be viewed as running SGD for minimizing $f_\delta(\mathbf{x})$.*

**Theorem 12.5** (nonsmooth and convex)**.** *Suppose $f:\mathbb{R}^d\to\mathbb{R}$ is convex and $G$-Lipschitz. The iteration*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t\mathbf{g}_\delta(\mathbf{x}_t;\mathbf{u}_t)$$

*holds that*

$$\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}[(f(\mathbf{x}_t) - f(\mathbf{x}^*)] \leq \delta G\sqrt{d} + \frac{1}{2\sum_{t=0}^{T-1}\eta_t}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + G^2(d+4)^2\sum_{t=0}^{T-1}\eta_t^2\right).$$

*Proof.* We have

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 = \|\mathbf{x}_t - \eta_t\mathbf{g}_\delta(\mathbf{x}_t;\mathbf{u}_t) - \mathbf{x}^*\|_2^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t\mathbb{E}[\langle\mathbf{g}_\delta(\mathbf{x}_t;\mathbf{u}_t), \mathbf{x}_t - \mathbf{x}^*\rangle] + \eta_t^2\mathbb{E}\|\mathbf{g}_\delta(\mathbf{x}_t;\mathbf{u}_t)\|_2^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t\langle\nabla f_\delta(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle + \eta_t^2G^2(d+4)^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t(f_\delta(\mathbf{x}_t) - f_\delta(\mathbf{x}^*))) + \eta_t^2G^2(d+4)^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t\big(f(\mathbf{x}_t) - f(\mathbf{x}^*) - \delta G\sqrt{d}\,)\big) + \eta_t^2G^2(d+4)^2,
\end{aligned}
$$

where the first inequality uses the fact $\mathbb{E}[\mathbf{g}_\delta(\mathbf{x};\mathbf{u})] = \nabla f_\delta(\mathbf{x})$ and Theorem 12.4; the second inequality comes from the convexity of $f(\cdot)$; the third inequality uses Theorem 12.1 and 12.3. Summing over above inequality with $t = 0,\ldots,T-1$, we have

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}_T - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - 2\sum_{t=0}^{T-1}\eta_t\big(f(\mathbf{x}_t) - f(\mathbf{x}^*)\big)\big) + 2\delta G\sqrt{d}\sum_{t=0}^{T-1}\eta_t + G^2(d+4)^2\sum_{t=0}^{T-1}\eta_t^2 \\
\Longrightarrow\ &\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \leq \frac{1}{2\sum_{t=0}^{T-1}\eta_t}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \delta G\sqrt{d} + \frac{G^2(d+4)^2}{2\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t^2 \\
\Longleftrightarrow\ &\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \leq \delta G\sqrt{d} + \frac{1}{2\sum_{t=0}^{T-1}\eta_t}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + G^2(d+4)^2\sum_{t=0}^{T-1}\eta_t^2\right).
\end{aligned}
$$

□

**Remark 12.3.** *Let*

$$R = \|\mathbf{x}_0 - \mathbf{x}^*\|_2, \quad \bar{\mathbf{x}}_T = \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbf{x}_t, \quad \delta = \frac{\epsilon}{2G\sqrt{d}}, \quad \eta_t = \frac{R}{(d+4)G\sqrt{T}} \quad \text{and} \quad T = \frac{4(d+4)^2G^2R^2}{\epsilon^2},$$

*then $\mathbb{E}[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*)] \leq \epsilon$.*

**Remark 12.4.** *The term $(d+4)^2$ can be improved. Please see Remark 12.7.*

**Theorem 12.6** (smooth and convex)**.** *Suppose $f:\mathbb{R}^d\to\mathbb{R}$ is convex and $L$-smooth. The iteration*

$$\mathbf{x}_{t+1}=\mathbf{x}_t-\eta\mathbf{g}_\delta(\mathbf{x}_t;\mathbf{u}_t)$$

*with $\eta=1/(4L(d+4))$ holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t)-f(\mathbf{x}^*))\le\frac{4L(d+4)\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2^2}{T}+\frac{9L\delta^2(d+4)^2}{25}.$$

*Proof.* We have

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2=\left\|\mathbf{x}_t-\eta\mathbf{g}_\delta(\mathbf{x}_t;\mathbf{u}_t)-\mathbf{x}^*\right\|_2^2\\
&=\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\mathbb{E}[\langle\mathbf{g}_\delta(\mathbf{x}_t;\mathbf{u}_t),\mathbf{x}_t-\mathbf{x}^*\rangle]+\eta^2\mathbb{E}\left\|\mathbf{g}_\delta(\mathbf{x}_t;\mathbf{u}_t)\right\|_2^2\\
&\le\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\langle\nabla f_\delta(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^*\rangle+\eta^2\left(\frac{L^2\delta^2(d+6)^3}{2}+2(d+4)\left\|\nabla f(\mathbf{x}_t)\right\|_2^2\right)\\
&\le\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta(f_\delta(\mathbf{x}_t)-f_\delta(\mathbf{x}^*)))+\eta^2\left(\frac{L^2\delta^2(d+6)^3}{2}+4(d+4)L(f(\mathbf{x}_t)-f(\mathbf{x}^*))\right)\\
&\le\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\big(f(\mathbf{x}_t)-f(\mathbf{x}^*)-\frac{L\delta^2d}{2})\big)+\eta_t^2G^2(d+4)^2+\eta^2\left(\frac{L^2\delta^2(d+6)^3}{2}+4(d+4)L(f(\mathbf{x}_t)-f(\mathbf{x}^*))\right)\\
&\le\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta(1-2(d+4)L\eta)\big(f(\mathbf{x}_t)-f(\mathbf{x}^*)\big)+L\eta\delta^2d+\frac{L^2\eta^2\delta^2(d+6)^3}{2}\\
&=\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-\frac{f(\mathbf{x}_t)-f(\mathbf{x}^*)}{4L(d+4)}+\frac{\delta^2}{4}\left(\frac{d}{d+4}+\frac{(d+6)^3}{8(d+4)^2}\right)\\
&\le\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-\frac{f(\mathbf{x}_t)-f(\mathbf{x}^*)}{4L(d+4)}+\frac{9\delta^2(d+4)}{100}
\end{aligned}$$

where the first inequality uses the fact $\mathbb{E}[\mathbf{g}_\delta(\mathbf{x};\mathbf{u})]=\nabla f_\delta(\mathbf{x})$ and Theorem 12.4; the second inequality comes from the convexity of $f(\cdot)$ and Theorem 3.19; the third inequality uses Theorem 12.1 and 12.3; the last two line is based on $\eta=1/(4L(d+4))$. Taking expectation on all of the history and setting $\bar{\mathbf{x}}_T=\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}_t$, we have

$$\begin{aligned}
&\frac{f(\mathbf{x}_t)-f(\mathbf{x}^*)}{4L(d+4)}\le\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2+\frac{9\delta^2(d+4)}{100}\\
\iff&f(\mathbf{x}_t)-f(\mathbf{x}^*)\le4L(d+4)\left(\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2\right)+\frac{9L\delta^2(d+4)^2}{25}\\
\implies&\mathbb{E}[f(\bar{\mathbf{x}}_T)-f(\mathbf{x}^*)]\le\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t)-f(\mathbf{x}^*))\le\frac{4L(d+4)\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2^2}{T}+\frac{9L\delta^2(d+4)^2}{25}.
\end{aligned}$$

□

**Remark 12.5.** *Let*

$$T=8L(d+4)\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2^2\epsilon^{-1}\qquad\text{and}\qquad\delta=\frac{5\epsilon}{3\sqrt{2L}\,(d+4)},$$

*then $\mathbb{E}[f(\bar{\mathbf{x}}_T)-f(\mathbf{x}^*)]\le\epsilon$.*

**Remark 12.6.** *If we additionally suppose $f(\cdot)$ is $\mu$-strongly convex and let*

$$\Delta=\frac{18\delta^2L(d+4)^2}{25\mu},$$

*then*

$$\mathbb{E}\left[\left\|\mathbf{x}_T-\mathbf{x}^*\right\|_2^2-\Delta\right]\le\left(1-\frac{\mu}{8L(d+4)}\right)^T\left(\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2^2-\Delta\right).$$

*We set*

$$T = \frac{8L(d+4)}{\mu} \log\left(\frac{2\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{\epsilon}\right) \qquad \textit{and} \qquad \delta = \frac{5}{6(d+4)}\sqrt{\frac{\mu\epsilon}{L}},$$

*then* $\mathbb{E}\left\|\mathbf{x}_T - \mathbf{x}^*\right\|_2^2 \leq \epsilon$.

**Theorem 12.7** (Smooth Nonconvex)**.** *Suppose* $f : \mathbb{R}^d \to \mathbb{R}$ *is* $L$*-smooth. The iteration*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{g}_\delta(\mathbf{x}_t; \mathbf{u}_t)$$

*with*

$$\eta = \frac{1}{4L(d+4)}, \quad T = 16L(d+4)(f(\mathbf{x}_0) - f^*)\epsilon^{-2} \quad \textit{and} \quad \delta = \frac{2\epsilon}{L}\sqrt{\frac{1}{(d+4)(d+16)}}$$

*leads to*

$$\mathbb{E}\left\|\nabla f(\mathbf{x}_{\text{out}})\right\|_2^2 \leq \epsilon^2$$

*where* $\mathbf{x}_{\text{out}}$ *is uniformly sampled from* $\{\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}\}$.

*Proof.* We have

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_{t+1})] \leq & f(\mathbf{x}_t) - \eta\mathbb{E}[\langle \mathbf{g}_\delta(\mathbf{x}_t; \mathbf{u}_t), \nabla f(\mathbf{x}_t)\rangle] + L\eta^2\mathbb{E}\left\|\mathbf{g}_\delta(\mathbf{x}_t; \mathbf{u}_t)\right\|_2^2 \\
\leq & f(\mathbf{x}_t) - \eta\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + L\eta^2\left(\frac{L^2\delta^2(d+6)^3}{2} + 2(d+4)\left\|\nabla f(\mathbf{x})\right\|_2^2\right) \\
= & f(\mathbf{x}_t) - \eta\left(1 - 2L\eta(d+4)\right)\left\|\nabla f(\mathbf{x})\right\|_2^2 + \frac{L^3\eta^2\delta^2(d+6)^3}{2}.
\end{aligned}$$

Let $\eta = 1/(4L(d+4))$, we have

$$\begin{aligned}
& \eta\left(1 - 2L\eta(d+4)\right)\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \leq f(\mathbf{x}_t) - \mathbb{E}[f(\mathbf{x}_{t+1})] + \frac{L^3\eta^2\delta^2(d+6)^3}{2} \\
\Longrightarrow & \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \leq f(\mathbf{x}_t) - \mathbb{E}[f(\mathbf{x}_{t+1})] + \frac{L\delta^2(d+6)^3}{32(d+4)^2} \leq f(\mathbf{x}_t) - \mathbb{E}[f(\mathbf{x}_{t+1})] + \frac{L\delta^2(d+16)}{32}.
\end{aligned}$$

Taking the average, we obtain

$$\begin{aligned}
\mathbb{E}\left\|\nabla f(\mathbf{x}_{\text{out}})\right\|_2^2 = & \frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \leq \frac{2(f(\mathbf{x}_0) - \mathbb{E}[f(\mathbf{x}_T))]}{\eta T} + \frac{L\delta^2(d+16)}{16\eta T} \\
\leq & \frac{8L(d+4)(f(\mathbf{x}_0) - f^*)}{T} + \frac{L^2\delta^2(d+4)(d+16)}{4}.
\end{aligned}$$

Hence, taking

$$T = 16L(d+4)(f(\mathbf{x}_0) - f^*)\epsilon^{-2} \quad \text{and} \quad \delta = \frac{2\epsilon}{L}\sqrt{\frac{1}{(d+4)(d+16)}}$$

leads to $\mathbb{E}\left\|\nabla f(\mathbf{x}_{\text{out}})\right\|_2^2 \leq \epsilon^2$. □

**Lemma 12.1.** *For any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ *and some* $\delta > 0$*, it holds that:*

1. *If* $f(\cdot)$ *is* $G$*-Lipschitz continuous, then* $\mathbb{E}\left\|\mathbf{g}_\delta(\mathbf{x}; \mathbf{u}) - \mathbf{g}_\delta(\mathbf{y}; \mathbf{u})\right\|_2^2 \leq \dfrac{2G^2 d\left\|\mathbf{x} - \mathbf{y}\right\|_2^2}{\delta}$.

*2. If $f(\cdot)$ is L-smooth, then $\mathbb{E}\left\|\mathbf{g}_\delta(\mathbf{x};\mathbf{u})-\mathbf{g}_\delta(\mathbf{y};\mathbf{u})\right\|_2^2 \leq \dfrac{3L^2\delta^2(d+6)^3}{2}+3L^2(d+4)\left\|\mathbf{x}-\mathbf{y}\right\|_2^2$.*

*Proof.* **Part I:** We have

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{g}_\delta(\mathbf{x};\mathbf{u})-\mathbf{g}_\delta(\mathbf{y};\mathbf{u})\right\|_2^2=\mathbb{E}\left\|\frac{f(\mathbf{x}+\delta\mathbf{u})-f(\mathbf{x})}{\delta}\cdot\mathbf{u}-\frac{f(\mathbf{y}+\delta\mathbf{u})-f(\mathbf{y})}{\delta}\cdot\mathbf{u}\right\|_2^2\\
\leq&2\mathbb{E}\left\|\frac{f(\mathbf{x}+\delta\mathbf{u})-f(\mathbf{y}+\delta\mathbf{u})}{\delta}\cdot\mathbf{u}\right\|_2^2+\mathbb{E}\left\|\frac{f(\mathbf{y})-f(\mathbf{x})}{\delta}\cdot\mathbf{u}\right\|_2^2\\
\leq&\frac{G^2\left\|\mathbf{x}-\mathbf{y}\right\|_2^2}{\delta^2}\mathbb{E}\left\|\mathbf{u}\right\|_2^2+\frac{G^2\left\|\mathbf{x}-\mathbf{y}\right\|_2^2}{\delta^2}\mathbb{E}\left\|\mathbf{u}\right\|_2^2\\
\leq&\frac{2G^2d\left\|\mathbf{x}-\mathbf{y}\right\|_2^2}{\delta}.
\end{aligned}$$

**Part II:** We have

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{g}_\delta(\mathbf{x};\mathbf{u})-\mathbf{g}_\delta(\mathbf{y};\mathbf{u})\right\|_2^2=\mathbb{E}\left\|\frac{f(\mathbf{x}+\delta\mathbf{u})-f(\mathbf{x})}{\delta}\cdot\mathbf{u}-\frac{f(\mathbf{y}+\delta\mathbf{u})-f(\mathbf{y})}{\delta}\cdot\mathbf{u}\right\|_2^2\\
=&\mathbb{E}\left\|\frac{f(\mathbf{x}+\delta\mathbf{u})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\delta\mathbf{u}\rangle}{\delta}\cdot\mathbf{u}-\frac{f(\mathbf{y}+\delta\mathbf{u})-f(\mathbf{y})-\langle\nabla f(\mathbf{y}),\delta\mathbf{u}\rangle}{\delta}\cdot\mathbf{u}+\frac{\langle\nabla f(\mathbf{x})-\nabla f(\mathbf{y}),\delta\mathbf{u}\rangle\cdot\mathbf{u}}{\delta}\right\|_2^2\\
\leq&3\mathbb{E}\left[\left\|\frac{f(\mathbf{x}+\delta\mathbf{u})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\delta\mathbf{u}\rangle}{\delta}\cdot\mathbf{u}\right\|_2^2+\left\|\frac{f(\mathbf{y}+\delta\mathbf{u})-f(\mathbf{y})-\langle\nabla f(\mathbf{y}),\delta\mathbf{u}\rangle}{\delta}\cdot\mathbf{u}\right\|_2^2+\left\|\langle\nabla f(\mathbf{x})-\nabla f(\mathbf{y}),\mathbf{u}\rangle\cdot\mathbf{u}\right\|_2^2\right]\\
\leq&3\mathbb{E}\left[\frac{\left\|f(\mathbf{x}+\delta\mathbf{u})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\delta\mathbf{u}\rangle\right\|_2^2\left\|\mathbf{u}\right\|_2^2}{\delta^2}+\frac{\left\|f(\mathbf{y}+\delta\mathbf{u})-f(\mathbf{y})-\langle\nabla f(\mathbf{y}),\delta\mathbf{u}\rangle\right\|_2^2\left\|\mathbf{u}\right\|_2^2}{\delta^2}+(d+4)\left\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\right\|_2^2\right]\\
\leq&3\mathbb{E}\left[\frac{L^2\left\|\delta\mathbf{u}\right\|_2^4\left\|\mathbf{u}\right\|_2^2}{4\delta^2}+\frac{L^2\left\|\delta\mathbf{u}\right\|_2^4\left\|\mathbf{u}\right\|_2^2}{4\delta^2}+L^2(d+4)\left\|\mathbf{x}-\mathbf{y}\right\|_2^2\right]\\
=&3\mathbb{E}\left[\frac{L^2\delta^2\left\|\mathbf{u}\right\|_2^6}{2}+L^2(d+4)\left\|\mathbf{x}-\mathbf{y}\right\|_2^2\right]\leq\frac{3L^2\delta^2(d+6)^3}{2}+3L^2(d+4)\left\|\mathbf{x}-\mathbf{y}\right\|_2^2,
\end{aligned}$$

where we use the fact

$$\mathbb{E}\left\|\langle\mathbf{a},\mathbf{u}\rangle\mathbf{u}\right\|_2^2\leq\mathbb{E}\left[\langle\mathbf{a},\mathbf{u}\rangle^2\left\|\mathbf{u}\right\|_2^2\right]\leq(d+4)\left\|\mathbf{a}\right\|_2^2.$$

□

**Remark 12.7.** *We have ($\mathbf{u}=\mathbf{v}/\sqrt{1-\tau}$)*

$$\begin{aligned}
&\mathbb{E}\left[\langle\mathbf{a},\mathbf{u}\rangle^2\left\|\mathbf{u}\right\|_2^2\right]=\int\frac{1}{(2\pi)^{d/2}}\langle\mathbf{a},\mathbf{u}\rangle^2\left\|\mathbf{u}\right\|_2^2\exp\left(-\frac{1}{2}\left\|\mathbf{u}\right\|_2^2\right)\mathrm{d}\mathbf{u}\\
=&\frac{1}{(2\pi)^{d/2}}\int\langle\mathbf{a},\mathbf{u}\rangle^2\left\|\mathbf{u}\right\|_2^2\exp\left(-\frac{\tau}{2}\left\|\mathbf{u}\right\|_2^2\right)\exp\left(-\frac{1-\tau}{2}\left\|\mathbf{u}\right\|_2^2\right)\mathrm{d}\mathbf{u}\\
\leq&\frac{1}{(2\pi)^{d/2}}\int\langle\mathbf{a},\mathbf{u}\rangle^2\cdot\frac{2}{\tau\mathrm{e}}\cdot\exp\left(-\frac{1-\tau}{2}\left\|\mathbf{u}\right\|_2^2\right)\mathrm{d}\mathbf{u}\\
=&\frac{2}{(2\pi)^{d/2}\tau\mathrm{e}}\int\frac{\langle\mathbf{a},\mathbf{v}\rangle^2}{1-\tau}\cdot\exp\left(-\frac{1}{2}\left\|\mathbf{v}\right\|_2^2\right)\mathrm{d}\frac{\mathbf{v}}{(1-\tau)^{d/2}}\\
=&\frac{2}{(2\pi)^{d/2}\tau\mathrm{e}}\int\frac{\langle\mathbf{a},\mathbf{v}\rangle^2}{(1-\tau)^{d/2+1}}\cdot\exp\left(-\frac{1}{2}\left\|\mathbf{v}\right\|_2^2\right)\mathrm{d}\mathbf{v}\\
=&\frac{2}{\tau(1-\tau)^{d/2+1}\mathrm{e}}\int\langle\mathbf{a},\mathbf{v}\rangle^2\cdot\frac{1}{(2\pi)^{d/2}}\cdot\exp\left(-\frac{1}{2}\left\|\mathbf{v}\right\|_2^2\right)\mathrm{d}\mathbf{v}\\
=&\frac{2\left\|\mathbf{a}\right\|_2^2}{\tau(1-\tau)^{d/2+1}\mathrm{e}}.
\end{aligned}$$

*where the first inequality use the fact*

$$t^p \exp\left(-\frac{\tau t^2}{2}\right) \leq \left(\frac{p}{\tau \mathrm{e}}\right)^{p/2}$$

*with* $t = \|\mathbf{u}\|_2$ *and* $p = 2$ *and the last step is because of* $\langle \mathbf{a}, \mathbf{v}\rangle \sim \mathcal{N}(\mathbf{0}, \mathbf{a}^\top \mathbf{a})$ *for* $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. *The minimum in* $\tau$ *is attained for* $\tau = 2/(d+4)$, *then*

$$\tau(1-\tau)^{d/2+1} = \frac{2}{d+4}\left(\frac{1}{1+2/(d+2)}\right)^{(d+2)/2} > \frac{2}{(d+4)\mathrm{e}}.$$

*Hence, we have* $\mathbb{E}\,\|\langle \mathbf{a}, \mathbf{u}\rangle \mathbf{u}\|_2^2 \leq (d+4)\,\|\mathbf{a}\|_2^2$.

**Remark 12.8.** *We have show that*

$$\mathbb{E}[\mathbf{g}_\delta(\mathbf{x}; \mathbf{u})] = \nabla f_\delta(\mathbf{x})$$

*then*

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b} \mathbf{g}_\delta(\mathbf{x}; \mathbf{u}_i) - \nabla f_\delta(\mathbf{x})\right\|_2^2 = \frac{1}{b}\mathbb{E}\,\|\mathbf{g}_\delta(\mathbf{x}; \mathbf{u}) - \nabla f_\delta(\mathbf{x})\|_2^2 \\
=&\frac{1}{b}\left(\mathbb{E}\,\|\mathbf{g}_\delta(\mathbf{x}; \mathbf{u})\|_2^2 - \left(\mathbb{E}\,\|\mathbf{g}_\delta(\mathbf{x}; \mathbf{u})\|_2\right)^2\right) \leq \frac{1}{b}\mathbb{E}\,\|\mathbf{g}_\delta(\mathbf{x}; \mathbf{u})\|_2^2 .
\end{aligned}$$

*Since Theorem 12.1 says sufficient small* $\delta$ *leads to the value (or gradient) of* $f(\cdot)$ *and* $f_\delta(\cdot)$ *be arbitrary close, above inequality implies we can use mini-batch algorithms like stochastic first-order optimization.*

**Remark 12.9.** *Since* $\mathbf{u}$ *is symmetric, we have*

$$\mathbb{E}\left[\frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x})}{\delta}\cdot \mathbf{u}\right] = \mathbb{E}\left[\frac{f(\mathbf{x}+\delta\mathbf{u})}{\delta}\cdot \mathbf{u}\right] = \mathbb{E}\left[\frac{f(\mathbf{x}-\delta\mathbf{u})}{\delta}\cdot (-\mathbf{u})\right],$$

*which leads to the above one is equal to*

$$\mathbb{E}\left[\frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x}-\delta\mathbf{u})}{2\delta}\cdot \mathbf{u}\right].$$

*Hence, we can also replace* $\mathbf{g}_\delta(\mathbf{x}; \mathbf{u})$ *by*

$$\hat{\mathbf{g}}_\delta(\mathbf{x}; \mathbf{u}) = \frac{f(\mathbf{x}+\delta\mathbf{u}) - f(\mathbf{x}-\delta\mathbf{u})}{2\delta}\cdot \mathbf{u}.$$

**Nonsmooth Nonconvex Optimization** Consider the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous, but possibly nonsmooth and nonconvex. We typically introduce the notion of $(\delta, \epsilon)$-Goldstein stationary point to describe the convergence of the algorithms. We say $x$ is an $(\delta, \epsilon)$-Goldstein stationary point of Lipschitz continuous function $f$ if it satisfies

$$\mathrm{dist}(0, \partial_\delta f(x)) \leq \epsilon, \tag{70}$$

where

$$\partial_\delta f(x) := \mathrm{conv}\left\{\cup_{y\in\mathbb{B}_\delta(x)} \partial f(y)\right\},$$

is the Goldstein $\delta$-subdifferential and

$$\partial f(x) := \operatorname{conv}\left\{g : g = \lim_{x_k \to x} \nabla f(x_k)\right\}.$$

is the Clarke subdifferential. The popular zeroth-order optimization algorithms approximate the objective $f(\mathbf{x})$ by

$$\hat{f}_\delta(\mathbf{x}) \triangleq \mathbb{E}[f(\mathbf{x} + \delta\mathbf{u})],$$

where the $d$-dimensional random vector $\mathbf{u}$ is uniformly distributed on unit ball $\mathcal{B}^d = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\|_2 \leq 1\}$. Finding $(\delta, \epsilon)$-Goldstein stationary point of $f$ can be reduced to find the approximate stationary point of $\hat{f}_\delta(\mathbf{x})$, which is smooth and its unbiased gradient estimator can be established by

$$\hat{\mathbf{g}}_\delta(\mathbf{x}; \mathbf{w}) = \frac{d(f(\mathbf{x} + \delta\mathbf{w}) - f(\mathbf{x} - \delta\mathbf{w}))}{2\delta} \cdot \mathbf{w},$$

where the $d$-dimensional random vector $\mathbf{w}$ is uniformly sampled from unit sphere $\mathcal{S}^{d-1} = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\|_2 = 1\}$.