Lecture Notes of Optimization Theory (2025)

Luo Luo
School of Data Science, Fudan University

August 13, 2025

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

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

3. For scalar variables $x, y \in \mathbb{R}$, 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

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

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

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

which holds

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

then we write its differential as follows

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

For the first term, we have

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

which implies

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

For the second term, we have

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

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

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

which means

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

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* $f(\mathbf{X}) = \mathrm{tr}(\mathbf{A}\mathbf{X}^{-1})$*, 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 λ_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 $i = 1, \dots, p$ when $\mathbf{H} \to \mathbf{0}$. Therefore, it holds

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

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

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

□

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

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

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 ∂ 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

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

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

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

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

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

Then we have

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

and

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

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*

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

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

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

and

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

Therefore, we have

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

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*

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

  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 $1, 1/2, 1, 1/2 \dots$ 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*

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

which converges to $x^* = 0$. 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 0.

3 Convex Analysis

Theorem 3.1. *Let $\mathcal{C}_\theta$ be convex sets indexed by θ, 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 θ, 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. $\square$

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 $\operatorname{proj}_{\mathcal{C}}(\mathbf{y})$.

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. $\square$

Theorem 3.3. *If $\mathbf{y} \notin \mathcal{C}$ for some close and convex set $\mathcal{C} \subseteq \mathbb{R}^d$, then $\mathbf{z} = \operatorname{proj}_{\mathcal{C}}(\mathbf{y})$ 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 $\mathbf{z} = \operatorname{proj}_{\mathcal{C}}(\mathbf{y})$ 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 $\mathbf{z} = \operatorname{proj}_{\mathcal{C}}(\mathbf{y})$. 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 $\langle \mathbf{y} - \mathbf{z}, \mathbf{x} - \mathbf{z} \rangle \leq 0$.

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

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

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 $\operatorname{epi} f$. 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 $\operatorname{epi} f$, which means the epigraph is convex.

Part II: Suppose the epigraph

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

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 $(\mathbf{x}_1, u_1), (\mathbf{x}_2, u_2) \in \operatorname{epi} f$. The convexity of epigraph means

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

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

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

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

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

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. For any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, 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

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

$\square$

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 λ, 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 $\mathbf{h}=\mathbf{y}-\mathbf{x}=[h_1,\dots,h_d]^\top$. *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

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

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 $\|\mathbf{v}\| = 1$. 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. *ℓ_p-norm vs. ℓ_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 $\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq 0$.

$\square$

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 $\mathbf{z} = \lambda\mathbf{y} + (1-\lambda)\mathbf{x} \in \mathcal{C}$. 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 $\mathbb{R}^2$*

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

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 $\partial f(z) = \{(x_1, x_2) : x_1 \geq 0, x_2 = 0\}$ and $g(z) = \{(x_1, x_2) : x_1 \leq 0, x_2 = 0\}$, 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*

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

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

$$\langle \mathbf{y}_{k_l} - \mathbf{z}_{k_l}, \mathbf{x} - \mathbf{z}_{k_l} \rangle \leq 0 \quad \Longleftrightarrow \quad \langle \mathbf{w}_{k_l}, \mathbf{x} - \mathbf{z}_{k_l} \rangle \leq 0 \quad \Longleftrightarrow \quad \langle \mathbf{w}_{k_l}, \mathbf{x} \rangle \leq \langle \mathbf{w}_{k_l}, \mathbf{z}_{k_l} \rangle$$

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

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

for all k_l. Connecting above inequalities, we have

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

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

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

□

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

1. *If any $\mathbf{x} \in \mathrm{dom}\, f$ 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 $\mathrm{dom}\, f$, 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 $(\mathbf{y}, t) \in \operatorname{epi} f$, i.e., $t \geq f(\mathbf{y})$. 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

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

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, taking $t = f(\mathbf{y})$ means 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 $y \geq 0$*. This can not holds because:*

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

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

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

The point $\mathbf{x}^*$ is an optimal solution of the problem if there exists a subgradient $\mathbf{g}^* \in \partial f(\mathbf{x}^*)$ such that for all $\mathbf{y} \in \mathcal{C}$ 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: Recall that Theorem 3.10 says

$$\partial \mathbb{1}_{\mathcal{C}}(\mathbf{x}^*) = \left\{\mathbf{g} \in \mathbb{R}^d : \langle \mathbf{g}, \mathbf{y} - \mathbf{x}^* \rangle \leq 0 \text{ for all } \mathbf{y} \in \mathcal{C}\right\}.$$

Suppose there a subgradient $\mathbf{g}^* \in \partial f(\mathbf{x}^*)$ such that for all $\mathbf{y} \in \mathcal{C}$ satisfies $\langle \mathbf{g}^*, \mathbf{y} - \mathbf{x}^* \rangle \geq 0$, then we have

$$-\mathbf{g}^* \in \left\{\mathbf{g} \in \mathbb{R}^d : \langle \mathbf{g}, \mathbf{y} - \mathbf{x}^* \rangle \leq 0 \text{ for all } \mathbf{y} \in \mathcal{C}\right\} = \partial \mathbb{1}_{\mathcal{C}}(\mathbf{x}^*).$$

Therefore, we have

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

which means $\mathbf{x}^*$ is an optimal solution by following Part I. The last step use the condition $\mathcal{C} \subseteq (\operatorname{dom} f)^\circ$.

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

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

□

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*

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

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 (the last one is based on Remark 3.4)

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

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

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

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

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

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

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

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

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

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

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*

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

which implies $\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|_2\le L\,\|\mathbf{x}-\mathbf{y}\|_2$.

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 ∇f 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 $\tilde{t} \in (0, \hat{t}) \subseteq (0, T]$ such that

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

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

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

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

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

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

which leads to contradiction. Hence, we conclude $\nabla^2 f(\mathbf{x}^*)$ 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 $\|\mathbf{p}\|_2 < \delta$, 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

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

Remark 3.14. *The proof is based on the fact*

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

Let $\mathbf{h}(\mathbf{x}) = \nabla f(\mathbf{x})$. We can write

$$h_i(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_i}.$$

Recall that Remark 3.4 has shown that

$$\lim_{t \to 0} \frac{h_i(\mathbf{x} + t\mathbf{p}) - h_i(\mathbf{x})}{t} = \langle \nabla h_i(\mathbf{x}), \mathbf{p} \rangle = \sum_{j=1}^d \frac{\partial h_i(\mathbf{x})}{\partial x_j} \cdot p_j = \sum_{j=1}^d \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \cdot p_j,$$

which means

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

Let $\mathbf{v}(t) = \mathbf{h}(\mathbf{y} + t\mathbf{p}) = \nabla f(\mathbf{y} + t\mathbf{p})$*, then we have*

$$v_i'(t) = \sum_{j=1}^{d} \frac{\partial h_i(\mathbf{y} + t\mathbf{p})}{\partial(y_j + tp_j)} \cdot \frac{\partial(y_j + tp_j)}{\partial t} = \sum_{j=1}^{d} \left(\nabla^2 f(\mathbf{y} + t\mathbf{p})\right)_{ij} p_j.$$

Therefore, we have $\mathbf{v}'(t) = \nabla^2 f(\mathbf{y} + t\mathbf{p})\mathbf{p}$*.*

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* μ*-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 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 holds

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

Then we have

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

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 ℓ_2-norm on both sides, we obtain

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

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

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 neccesary 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

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

□

Example 3.6. *For unconstrained quadratic problem*

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

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

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

Therefore, the vector $\mathbf{x}$ satisfying $\mathbf{Q}\mathbf{x} = \mathbf{b}$ is the minimizer.

Example 3.7. *Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, the solution of minimization problem*

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

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

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

$$\begin{aligned}
\frac{\partial f(\mathbf{x})}{\partial x_j} =& \frac{1}{n} \sum_{i=1}^n \phi_i'(\mathbf{a}_i^\top \mathbf{x}) \cdot \frac{\partial \mathbf{a}_i^\top \mathbf{x}}{\partial x_j} + \frac{\lambda}{2} \frac{\partial \|\mathbf{x}\|_2^2}{\partial x_j} \\
=& \frac{1}{n} \sum_{i=1}^n \phi_i'(\mathbf{a}_i^\top \mathbf{x}) a_{ij} + \lambda x_j.
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k} =& \frac{\partial}{\partial x_k} \left(\frac{1}{n} \sum_{i=1}^n \phi_i'(\mathbf{a}_i^\top \mathbf{x}) a_{ij} + \lambda x_j \right) \\
=& \frac{1}{n} \sum_{i=1}^n \phi_i''(\mathbf{a}_i^\top \mathbf{x}) \cdot \frac{\partial \mathbf{a}_i^\top \mathbf{x}}{\partial x_k} \cdot a_{ij} + \lambda \mathbb{1}(j = k).
\end{aligned}$$

Therefore, we have

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

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

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

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, but it is NOT strongly convex. Note that we have $\phi''(z) \to 0$ by taking $z \to 0$. This implies there is no $\mu > 0$ such that $\phi''(z) \ge \mu$ for any $z \in \mathbb{R}$.

Remark 3.15. *Consider the function*

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

Let $\phi_i(z) = \frac{1}{2}(z-b_i)^2$, then we have $\phi_i'(z) = z - b_i$ and $\phi_i''(z) = 1$. It corresponds to ridge regression, i.e.,

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

Applications in Matrix Approximation: Given a symmetric positive-definite matrix $\mathbf{K} \in \mathbb{R}^{d\times d}$ and we sample a subset of columns $\mathbf{C} \in \mathbb{R}^{d\times m}$, where $m < d$. We want to find $\mathbf{W} \in \mathbb{R}^{m\times m}$ such that $\mathbf{K} \approx \mathbf{C}^\top\mathbf{W}\mathbf{C}^\top$. We write $\mathbf{C} \in \mathbb{R}^{d\times m}$ and $\mathbf{K} \in \mathbb{R}^{d\times d}$ as

$$\mathbf{C} = \begin{bmatrix}\mathbf{D}\\ \mathbf{E}\end{bmatrix} \qquad \text{and} \qquad \mathbf{K} = \begin{bmatrix}\mathbf{D} & \mathbf{E}^\top\\ \mathbf{E} & \mathbf{F}\end{bmatrix} \approx \mathbf{C}\mathbf{W}\mathbf{C}^\top = \begin{bmatrix}\mathbf{D}\\ \mathbf{E}\end{bmatrix}\mathbf{W}\begin{bmatrix}\mathbf{D} & \mathbf{E}^\top\end{bmatrix} = \begin{bmatrix}\mathbf{DWD} & \mathbf{DWE}^\top\\ \mathbf{EWD} & \mathbf{EWE}^\top\end{bmatrix}.$$

If we only sample columns $\mathbf{C} \in \mathbb{R}^{d\times m}$, the information of $\mathbf{F}$ is missing. Therefore, taking $\mathbf{W} = \mathbf{D}^{-1}$ that leads to

$$\mathbf{K} = \begin{bmatrix}\mathbf{D} & \mathbf{E}^\top\\ \mathbf{E} & \mathbf{F}\end{bmatrix} \approx \mathbf{C}\mathbf{W}\mathbf{C}^\top = \begin{bmatrix}\mathbf{D} & \mathbf{E}^\top\\ \mathbf{E} & \mathbf{E}\mathbf{D}^{-1}\mathbf{E}^\top\end{bmatrix},$$

which recover the information of $\mathbf{D}$, $\mathbf{E}$ and $\mathbf{F}$. This is Nyström method.

Remark 3.16. *The rank of estimator $\tilde{\mathbf{K}} = \mathbf{C}\mathbf{W}\mathbf{C}^\top$ is m, which means it is singular.*

If we can pass the matrix $\mathbf{K} \in \mathbb{R}^{d\times d}$, but still want to establish its approximation by $\mathbf{C} \in \mathbb{R}^{d\times m}$. We can construct the estimator of $\mathbf{K}$ by

$$\mathbf{K} \approx \mathbf{C}\mathbf{U}\mathbf{C}^\top + \delta\mathbf{I}_d,$$

We consider the problem

$$\min_{\mathbf{U}\in\mathbb{R}^{m\times m},\, \delta\in\mathbb{R}} f(\mathbf{U},\delta) \triangleq \left\|\mathbf{K} - (\mathbf{C}\mathbf{U}\mathbf{C}^\top + \delta\mathbf{I}_d)\right\|_F^2.$$

We can write

$$f(\mathbf{U},\delta) = \mathrm{tr}\left((\mathbf{K} - \mathbf{C}\mathbf{U}\mathbf{C}^\top - \delta\mathbf{I}_d)(\mathbf{K} - \mathbf{C}\mathbf{U}\mathbf{C}^\top - \delta\mathbf{I}_d)^\top\right).$$

Taking the gradient of $f(\mathbf{U},\delta)$ with respect to $\mathbf{U}$ be zero, we have

$$\begin{aligned}\frac{\partial f(\mathbf{U},\delta)}{\partial \mathbf{U}} =& \frac{\partial}{\partial \mathbf{U}}\mathrm{tr}\left(\mathbf{C}\mathbf{U}\mathbf{C}^\top\mathbf{C}\mathbf{U}\mathbf{C}^\top - 2\mathbf{K}\mathbf{C}\mathbf{U}\mathbf{C}^\top + 2\delta\mathbf{C}\mathbf{U}\mathbf{C}^\top\right)\\ =& 2(\mathbf{C}^\top\mathbf{C}\mathbf{U}\mathbf{C}^\top\mathbf{C} - \mathbf{C}^\top\mathbf{K}\mathbf{C} + \delta\mathbf{C}^\top\mathbf{C}) = \mathbf{0},\end{aligned}$$

that is

$$\begin{aligned}&\mathbf{C}^\top\mathbf{C}\mathbf{U}\mathbf{C}^\top\mathbf{C} = \mathbf{C}^\top\mathbf{K}\mathbf{C} - \delta\mathbf{C}^\top\mathbf{C}\\ \Longleftrightarrow &\mathbf{U} = (\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{K}\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1} - \delta(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}\\ &\quad = \mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^\top - \delta(\mathbf{C}^\top\mathbf{C})^{-1}.\end{aligned}$$

Taking the derivative of $f(\mathbf{U},\delta)$ with respect to δ be zero, we have

$$\begin{aligned}\frac{\partial f(\mathbf{U},\delta)}{\partial \delta} =& \frac{\partial}{\partial \delta}\mathrm{tr}\left(\delta^2\mathbf{I}_d - 2\delta\mathbf{K} + 2\delta\mathbf{C}\mathbf{U}\mathbf{C}^\top\right)\\ =& 2d\delta - 2\mathrm{tr}(\mathbf{K}) + 2\mathrm{tr}(\mathbf{C}\mathbf{U}\mathbf{C}^\top) = 0,\end{aligned}$$

that is

$$\begin{aligned}\delta =& \frac{1}{d}\left(\mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{C}\mathbf{U}\mathbf{C}^\top)\right)\\ =& \frac{1}{d}\left(\mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{C}\mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^\top\mathbf{C}^\top) + \mathrm{tr}(\delta\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top)\right)\\ =& \frac{1}{d}\left(\mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{C}\mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^\top\mathbf{C}^\top) + \delta m\right).\end{aligned}$$

Consider SVD of $\mathbf{C}$ and the the expression of $\mathbf{C}^\dagger$

$$\mathbf{C} = \mathbf{P}\mathbf{\Sigma}\mathbf{V}^\top \qquad \text{and} \qquad \mathbf{C}^\dagger = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{P}^\top$$

where $\mathbf{P} \in \mathbb{R}^{d\times m}$ is column orthogonal, $\mathbf{\Sigma} \in \mathbb{R}^{m\times m}$ is diagonal (full rank) and $\mathbf{V} \in \mathbb{R}^{m\times m}$ is orthogonal. We have

$$\mathrm{tr}(\mathbf{C}\mathbf{C}^\dagger) = \mathrm{tr}(\mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{Q}\mathbf{\Sigma}^{-1}\mathbf{P}^\top) = \mathrm{tr}(\mathbf{P}\mathbf{P}^\top) = \mathrm{tr}(\mathbf{P}^\top\mathbf{P}) = \mathrm{tr}(\mathbf{I}_m) = m.$$

We also have

$$\begin{aligned}&\mathrm{tr}(\mathbf{C}\mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^\top\mathbf{C}^\top)\\ =&\mathrm{tr}(\mathbf{C}^\top\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{K}(\mathbf{C}^\dagger)^\top)\\ =&\mathrm{tr}(\mathbf{C}^\top\mathbf{K}(\mathbf{C}^\dagger)^\top).\end{aligned}$$

Therefore, we have

$$\delta = \frac{1}{d-m}\left(\mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{C}^\top\mathbf{K}(\mathbf{C}^\dagger)^\top)\right).$$

We can verify

$$\mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{C}^\dagger\mathbf{K}\mathbf{C}) = \mathrm{tr}(\mathbf{K}(\mathbf{I}_d - \mathbf{C}\mathbf{C}^\dagger)) = \mathrm{tr}(\mathbf{K}(\mathbf{I}_d - \mathbf{P}\mathbf{P}^\dagger)) = \mathrm{tr}(\mathbf{K}\mathbf{P}_\perp\mathbf{P}_\perp^\top)) = \mathrm{tr}(\mathbf{P}_\perp^\top\mathbf{K}\mathbf{P}_\perp) \geq 0.$$

where $\mathbf{C}$ has SVD of the form $\mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top$ and $\mathbf{P}_\perp$ is the orthogonal complement of $\mathbf{P}$. Noticing that δ^{ss} is zero if and only if

$$\mathrm{tr}(\mathbf{P}_\perp^\top\mathbf{K}\mathbf{P}_\perp) = 0.$$

Since we assume $\mathbf{K} \succ \mathbf{0}$, the above trace cannot be zero. Hence, we conclude $\delta^{\mathrm{ss}} > 0$. Then the estimator holds

$$
\begin{aligned}
&\mathbf{C}\mathbf{U}^{\mathrm{ss}}\mathbf{C}^\top + \delta^{\mathrm{ss}}\mathbf{I}_d \\
=&\mathbf{C}(\mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^\top - \delta^{\mathrm{ss}}(\mathbf{C}^\top\mathbf{C})^\dagger)\mathbf{C}^\top + \delta^{\mathrm{ss}}\mathbf{I}_d \\
=&\mathbf{C}\mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^\top\mathbf{C}^\top + \delta^{\mathrm{ss}}(\mathbf{I}_d - \mathbf{C}(\mathbf{C}^\top\mathbf{C})^\dagger\mathbf{C}^\top) \\
\succeq&\mathbf{P}\mathbf{P}^\top\mathbf{K}\mathbf{P}\mathbf{P}^\top + \delta^{\mathrm{ss}}\mathbf{P}_\perp\mathbf{P}_\perp^\top \\
=&\begin{bmatrix}\mathbf{P} & \mathbf{P}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{P}^\top\mathbf{K}\mathbf{P} & \mathbf{0} \\ \mathbf{0} & \delta^{\mathrm{ss}}\mathbf{I}_d\end{bmatrix}\begin{bmatrix}\mathbf{P} \\ \mathbf{P}_\perp\end{bmatrix} \succ \mathbf{0}
\end{aligned}
$$

where we use the fact

$$
\begin{aligned}
&\mathbf{I}_d - \mathbf{C}(\mathbf{C}^\top\mathbf{C})^\dagger\mathbf{C}^\top \\
=&\mathbf{I}_d - \mathbf{P}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{V}\mathbf{\Sigma}\mathbf{P}^\top\mathbf{P}\mathbf{\Sigma}\mathbf{V}^\top)^\dagger\mathbf{V}\mathbf{\Sigma}\mathbf{P}^\top \\
=&\mathbf{I}_d - \mathbf{P}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top)^\dagger\mathbf{V}\mathbf{\Sigma}\mathbf{P}^\top \\
=&\mathbf{P}\mathbf{P}^\top + \mathbf{P}_\perp\mathbf{P}_\perp^\top - \mathbf{P}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^{-2}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}\mathbf{P}^\top \\
=&\mathbf{P}_\perp\mathbf{P}_\perp^\top
\end{aligned}
$$

and $\mathbf{K} \succ \mathbf{0}$.

Remark 3.17. *How to show the convexity of f?*

The Woodbury identity means says

$$(\mathbf{L}\mathbf{S}\mathbf{R} + \mathbf{A})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{L}(\mathbf{S}^{-1} + \mathbf{R}\mathbf{A}^{-1}\mathbf{L})^{-1}\mathbf{R}\mathbf{A}^{-1}. \tag{6}$$

We can let

$$\mathbf{A} = \delta^{\mathrm{ss}}\mathbf{I}_n, \qquad \mathbf{L} = \mathbf{C}, \qquad \mathbf{S} = \mathbf{U}^{\mathrm{ss}} \qquad \text{and} \qquad \mathbf{R} = \mathbf{C}^\top,$$

then the matrix $\mathbf{U}^{\mathrm{ss}} \in \mathbb{R}^{k\times k}$ may be singular even if $\mathbf{C}\mathbf{U}^{\mathrm{ss}}\mathbf{C}^\top + \delta^{\mathrm{ss}}\mathbf{I}_n$ is non-singular.
For given $\mathbf{C} \in \mathbb{R}^{d\times m}$, we apply QR on $\mathbf{C} \in \mathbb{R}^{d\times m}$ to obtain

$$\mathbf{C} = \mathbf{Q}\mathbf{R},$$

where $\mathbf{Q} \in \mathbb{R}^{n\times k}$ with $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_n$ and $\mathbf{R} \in \mathbb{R}^{k\times k}$. For column orthogonal $\mathbf{Q} \in \mathbb{R}^{d\times m}$, we have

$$\mathbf{Q}^\dagger = (\mathbf{Q}^\top\mathbf{Q})^{-1}\mathbf{Q}^\top = \mathbf{Q}^\top. \tag{7}$$

We denote $\lambda_i(\cdot)$ as the i-th largest eigenvalue of a matrix and $\mathbf{P} = \mathbf{I}_n - \mathbf{Q}\mathbf{Q}^\top$. We establish the approximation by (replace $\mathbf{C}$ in previous model by $\mathbf{Q}$)

$$(\mathbf{U}^{\mathrm{ss}}, \delta^{\mathrm{ss}}) = \operatorname*{arg\,min}_{\mathbf{U}\in\mathbb{R}^{m\times m},\, \delta\in\mathbb{R}} \left\|\mathbf{K} - (\mathbf{Q}\mathbf{U}\mathbf{Q}^\top + \delta\mathbf{I}_d)\right\|_F^2.$$

We replace the matrix $\mathbf{C}$ with $\mathbf{Q}$ and achieve

$$\delta^{\mathrm{ss}} = \frac{1}{d-m}\left(\mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{Q}^\top\mathbf{K}\mathbf{Q})\right) \tag{8}$$

and

$$
\begin{aligned}
\mathbf{U}^{\mathrm{ss}} =&\mathbf{Q}^\dagger\mathbf{K}(\mathbf{Q}^\dagger)^\top - \delta^{\mathrm{ss}}(\mathbf{Q}^\top\mathbf{Q})^\dagger \\
=&\mathbf{Q}^\top\mathbf{K}\mathbf{Q} - \delta^{\mathrm{ss}}\mathbf{I}_m \\
=&\mathbf{Q}^\top(\mathbf{K} - \delta^{\mathrm{ss}}\mathbf{I}_m)\mathbf{Q}.
\end{aligned} \tag{9}
$$

Applying Woodbury identity with

$$\mathbf{A} = \delta^{\text{ss}}\mathbf{I}_d, \qquad \mathbf{L} = \mathbf{Q}, \qquad \mathbf{S} = \mathbf{I}_m \qquad \text{and} \qquad \mathbf{R} = \mathbf{U}^{\text{ss}}\mathbf{Q}^\top,$$

we have

$$\begin{aligned}(\mathbf{Q}\mathbf{U}^{\text{ss}}\mathbf{Q}^\top + \delta^{\text{ss}}\mathbf{I}_d)^{-1} =&(\delta^{\text{ss}})^{-1}\mathbf{I}_d - (\delta^{\text{ss}})^{-2}\mathbf{Q}(\mathbf{I}_m + (\delta^{\text{ss}})^{-1}\mathbf{U}^{\text{ss}}\mathbf{Q}^\top\mathbf{Q})^{-1}\mathbf{U}^{\text{ss}}\mathbf{Q}^\top \\ =&(\delta^{\text{ss}})^{-1}\mathbf{I}_n - (\delta^{\text{ss}})^{-2}\mathbf{Q}(\mathbf{I}_m + (\delta^{\text{ss}})^{-1}\mathbf{U}^{\text{ss}})^{-1}\mathbf{U}^{\text{ss}}\mathbf{Q}^\top.\end{aligned} \tag{10}$$

Substituting equations (8) and (9) into (10), the term needs to be inverted has the form of

$$\begin{aligned}&\mathbf{I}_k + (\delta^{\text{ss}})^{-1}\mathbf{U}^{\text{ss}} \\ =&\mathbf{I}_k + (\delta^{\text{ss}})^{-1}\mathbf{Q}^\top(\mathbf{K} - \delta^{\text{ss}}\mathbf{I}_m)\mathbf{Q} \\ =&\mathbf{I}_k + (\delta^{\text{ss}})^{-1}\mathbf{Q}^\top\mathbf{K}\mathbf{Q} - (\delta^{\text{ss}})^{-1}\mathbf{Q}^\top \cdot \delta^{\text{ss}}\mathbf{I}_m\mathbf{Q} \\ =&\mathbf{I}_k + (\delta^{\text{ss}})^{-1}\mathbf{Q}^\top\mathbf{K}\mathbf{Q} - \mathbf{I}_m \\ =&(\delta^{\text{ss}})^{-1}\mathbf{Q}^\top\mathbf{K}\mathbf{Q} \succ \mathbf{0},\end{aligned}$$

which is positive-definite if we assume $\mathbf{K}$ is symmetric positive definite.

4 Gradient Descent Methods

Theorem 4.1. *For the minimization problem*

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \tag{11}$$

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 $\eta_t = \eta \leq 1/L$. Then we have

$$\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_t) \leq f(\hat{\mathbf{x}}) + \frac{L\left\|\mathbf{x}_0 - \hat{\mathbf{x}}\right\|_2^2}{2T}$$

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

Proof. Theorem 3.19 means

$$\begin{aligned}f(\mathbf{x}_{t+1}) \leq& f(\mathbf{x}_t) + \langle\nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2 \\ \leq& f(\mathbf{x}_t) - \left(\eta - \frac{L\eta^2}{2}\right)\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \\ \leq& f(\mathbf{x}_t) - \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2\end{aligned} \tag{12}$$

For any $\hat{\mathbf{x}} \in \mathbb{R}^d$, we obtain

$$\begin{aligned}&\left\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\right\|_2^2 \\ =&\left\|\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t) - \hat{\mathbf{x}}\right\|_2^2 \\ =&\left\|\mathbf{x}_t - \hat{\mathbf{x}}\right\|_2^2 - 2\eta\langle\nabla f(\mathbf{x}_t), \mathbf{x}_t - \hat{\mathbf{x}}\rangle + \eta^2\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \\ \leq&\left\|\mathbf{x}_t - \hat{\mathbf{x}}\right\|_2^2 + 2\eta(f(\hat{\mathbf{x}}) - f(\mathbf{x}_t)) + \eta^2\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \\ \leq&\left\|\mathbf{x}_t - \hat{\mathbf{x}}\right\|_2^2 + 2\eta(f(\hat{\mathbf{x}}) - f(\mathbf{x}_t)) + 2\eta(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) \\ \leq&\left\|\mathbf{x}_t - \hat{\mathbf{x}}\right\|_2^2 + 2\eta(f(\hat{\mathbf{x}}) - f(\mathbf{x}_{t+1})),\end{aligned} \tag{13}$$

where the first inequality uses the convexity of f such that

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

and the second inequality uses (12). Taking the average over equation (13) with $t = 0, \ldots, T-1$, we obtain

$$\begin{aligned} \frac{1}{T}\left\|\mathbf{x}_T - \hat{\mathbf{x}}\right\|_2^2 &\leq \frac{1}{T}\left\|\mathbf{x}_0 - \hat{\mathbf{x}}\right\|_2^2 + \frac{2\eta}{T}\sum_{t=1}^{T}(f(\hat{\mathbf{x}}) - f(\mathbf{x}_t)) \\ \Longrightarrow \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_t) &\leq f(\hat{\mathbf{x}}) + \frac{1}{2\eta T}\big(\left\|\mathbf{x}_0 - \hat{\mathbf{x}}\right\|_2^2 - \left\|\mathbf{x}_T - \hat{\mathbf{x}}\right\|_2^2\big) \leq f(\hat{\mathbf{x}}) + \frac{L\left\|\mathbf{x}_0 - \hat{\mathbf{x}}\right\|_2^2}{2T} \end{aligned}$$

□

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*

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

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

Remark 4.2. *Applying equation (12), we have*

$$f(\mathbf{x}_T) \leq f(\mathbf{x}_{T-1}) \cdots \leq f(\mathbf{x}_0).$$

Then Theorem 4.1 means

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \frac{L\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{2T}.$$

Remark 4.3 (nonconvex case)**.** *Noticing that inequality (12) holds even if the function is nonconvex, then we have*

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

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

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

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 ϵ-stationary point in expectation.

Theorem 4.2. *Under the setting of Theorem 4.1, we additionally suppose the objective is μ-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

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle + \frac{\mu}{2}\left\|\mathbf{x} - \mathbf{x}^*\right\|_2$$

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

for any $\mathbf{x} \in \mathbb{R}^d$. Using the result of (12), 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 4.4. *We can find $\mathbf{x}_T$ such that*

$$
f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \epsilon
$$

within

$$
\left\lceil \kappa \ln\left(\frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{\epsilon}\right)\right\rceil
$$

first-order oracle complexity, where $\kappa \triangleq L/\mu$ is the condition number. If $\mu \ll \epsilon$, we have

$$
\left\lceil \frac{L}{\mu} \ln\left(\frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{\epsilon}\right)\right\rceil \geq \left\lceil \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\epsilon}\right\rceil
$$

Example 4.1. *For regularized linear regression*

$$
\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{\beta}{2}\|\mathbf{x}\|_2^2
$$

where $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^d$ and $\lambda > 0$. We have

$$
\nabla^2 f(\mathbf{x}) = \mathbf{A}^\top\mathbf{A} + \beta\mathbf{I} \qquad \textit{and} \qquad \kappa = \frac{\lambda_1(\mathbf{A}^\top\mathbf{A}) + \beta}{\lambda_d(\mathbf{A}^\top\mathbf{A}) + \beta} = 1 + \frac{\lambda_1(\mathbf{A}^\top\mathbf{A}) - \lambda_d(\mathbf{A}^\top\mathbf{A})}{\lambda_d(\mathbf{A}^\top\mathbf{A}) + \beta}.
$$

Example 4.2. *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). We consider the problem of minimizing $f(\mathbf{x})$.

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

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

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

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

Based on the analysis for strongly convex case, we also have the linear convergence

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

.

Remark 4.5. *We say $f : \mathbb{R}^d \to \mathbb{R}$ satisfies Polyak–Łojasiewicz (PL) condition if there exists some $\mu > 0$ such that*

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

holds for any $\mathbf{x} \in \mathbb{R}^d$, where $f^* = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

Theorem 4.3. *Let $g : \mathbb{R}^m \to \mathbb{R}$ be smooth and μ-strongly convex and $\mathbf{A} \in \mathbb{R}^{m\times d}$ with $\mathrm{rank}(\mathbf{A}) = m$. 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 $\mu\lambda_m(\mathbf{A}\mathbf{A}^\top)$.*

Proof. We can verify

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

For any $\mathbf{x} \in \mathbb{R}^d$, we have

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

where the first inequality is due to $\mathbf{A}\mathbf{x} \subseteq \mathbb{R}^m$ that leads to

$$g^* = \inf_{\mathbf{y}\in\mathbb{R}^m} g(\mathbf{y}) \leq \inf_{\mathbf{x}\in\mathbb{R}^d} g(\mathbf{A}\mathbf{x}) = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = f^*,$$

and the last inequality requires

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

□

Remark 4.6. *For logistic regression, we have*

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \phi_i(\mathbf{a}^\top\mathbf{x}),$$

where

$$\phi(z) = \ln(1 + \exp(-b_i z)).$$

We can write $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$, where

$$g(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^n \phi_i(y_i).$$

Since the function $\phi_i(\cdot)$ is strongly-convex on compact set, gradient descent also has linear convergence.

Example 4.3. *Nonconvex function may also hold PL condition, such as $f(x) = x^2 + 3(\sin x)^2$. We have*

$$f^* = 0, \qquad f'(x) = 6\cos x \sin x + 2x \qquad \text{and} \qquad f''(x) = -6(\sin x)^2 + 6(\cos x)^2 + 2.$$

We can find $\mu = 0.01$ such that

$$x^2 + 3(\sin x)^2 \le \frac{1}{2\mu}(6\cos x \sin x + 2x)^2$$

for any $x \in \mathbb{R}$.

Remark 4.7. *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

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

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

Theorem 4.4. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and lower bounded. 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*

$$f(\mathbf{x}_t + \alpha_t \mathbf{p}_t) \le f(\mathbf{x}_t) + c_1 \alpha_t \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle, \tag{14}$$

$$\langle \nabla f(\mathbf{x}_t + \alpha_t \mathbf{p}_t), \mathbf{p}_t \rangle \ge c_2 \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle \tag{15}$$

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, since $\phi(\alpha)$ is lower bounded and

$$|\phi'(0)| = |\langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle| > c|\langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle| = |l'(0)|.$$

Let $\alpha' > 0$ be the smallest intersecting value of α, that is

$$f(\mathbf{x}_t + \alpha' \mathbf{p}_t) = f(\mathbf{x}_t) + \alpha' c_1 \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle. \tag{16}$$

Then condition (14) clearly holds for all $\alpha < \alpha'$.

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

$$\phi(0) = \phi(\alpha') + \phi(\alpha'')(0 - \alpha')$$

$$
\begin{aligned}
&\Longleftrightarrow \phi(\alpha') - \phi(0) = \phi(\alpha'')\alpha' \\
&\Longleftrightarrow f(\mathbf{x}_t + \alpha'\mathbf{p}_t) - f(\mathbf{x}_t) = \alpha' \langle \nabla f(\mathbf{x}_t + \alpha''\mathbf{p}_t), \mathbf{p}_t \rangle . \qquad (17)
\end{aligned}
$$

By combining (16) and (17), 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 4.5. *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 α_k satisfies the Wolfe conditions (14)-(15). 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

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

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

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

Combining above relations, we obtain

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

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

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

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

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

Since f is lower bounded, we have

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

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

Remark 4.8. *This result implies*

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

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

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

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

We let $\mathbf{s} = \mathbf{x}_{t+1} - \mathbf{x}_t$ and $\mathbf{y} = \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)$, then we have

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

which leads to

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

We also have

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

which leads to

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

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

5 Acceleration

We first consider the quadratic problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} Q(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{b}^\top\mathbf{x}, \tag{18}$$

where $\mathbf{A}\in\mathbb{R}^{d\times d}$ is positive definite and $\mathbf{b}\in\mathbb{R}^d$.

Theorem 5.1. *Consider the quadratic problem (18). 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

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

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

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

$\square$

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 $1-1/\kappa$.*

Remark 5.2. *Let $f(\mathbf{x})$ be the potential energy at position $\mathbf{x}$. The negative gradient $-\nabla f(\mathbf{x})$ represents the force pushing the system toward lower energy. The continuous-time motion of a ball with mass m, subject to a potential force $-\nabla f(\mathbf{x}(t))$ and and damping (friction) γ is described by*

$$m\ddot{\mathbf{x}}(t) = -\gamma\dot{\mathbf{x}(t)} - \nabla f(\mathbf{x}(t)).$$

We consider the discretizations

$$\dot{\mathbf{x}}(t) \approx \mathbf{x}_t - \mathbf{x}_{t-1} \quad \textit{and} \quad \ddot{\mathbf{x}}(t) \approx \dot{\mathbf{x}}(t+1) - \dot{\mathbf{x}}(t) \approx \mathbf{x}_{t+1} - \mathbf{x}_t - (\mathbf{x}_t - \mathbf{x}_{t-1}) = \mathbf{x}_{t+1} - 2\mathbf{x}_t + \mathbf{x}_{t-1},$$

which leads to

$$\begin{aligned}
&m(\mathbf{x}_{t+1} - 2\mathbf{x}_t + \mathbf{x}_{t-1}) = -\gamma(\mathbf{x}_t - \mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t)\\
\Longleftrightarrow&\mathbf{x}_{t+1} = \mathbf{x}_t + \left(1 - \frac{\gamma}{m}\right)(\mathbf{x}_t - \mathbf{x}_{t-1}) - \frac{1}{m}\nabla f(\mathbf{x}_t).
\end{aligned}$$

Theorem 5.2. *Solving problem (18) 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 $\beta \geq \max\{(1-\sqrt{\eta L}\,)^2, (1-\sqrt{\eta\mu}\,)^2\}$. 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 $t \geq 0$ and some $\mathbf{M}$ with spectral radius of β.

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

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

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

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

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

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

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 5.3. *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

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

Then

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

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

$$\left\|\mathbf{A}^k\right\|_2 = \left\|(\rho(\mathbf{A})+\epsilon)^k\,\mathbf{A}_+^k\right\|_2 = (\rho(\mathbf{A})+\epsilon)^k \left\|\mathbf{A}_+^k\right\|_2 < (\rho(\mathbf{A})+\epsilon)^k. \tag{19}$$

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 5.4. *For heavy ball method, we are interested in the bound (19). 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

$$\eta = \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2 \qquad \text{and} \qquad \beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2,$$

then we have

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

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

Analysis of AGD by Lyapunov Function (Strongly Convex): We first consider the problem

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

where $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth and μ-strongly convex. We study AGD iteration

$$\begin{cases}\mathbf{y}_t = \mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1}),\\ \mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t).\end{cases}$$

where $\mathbf{x}_{-1} = \mathbf{x}_0$ and $\beta_t \in (0,1)$.
We define

$$\Phi_0(\mathbf{x}) = f(\mathbf{x}_0) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{x}_0\|_2^2$$

and

$$\Phi_{t+1}(\mathbf{x}) = \left(1-\frac{1}{\sqrt{\kappa}}\right)\Phi_t(\mathbf{x}) + \frac{1}{\sqrt{\kappa}}\left(f(\mathbf{y}_t) + \langle\nabla f(\mathbf{y}_t), \mathbf{x}-\mathbf{y}_t\rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}_t\|_2^2\right) \quad \text{for } t\geq 0.$$

Recall that the strong convexity implies

$$f(\mathbf{x}) \geq f(\mathbf{y}_t) + \langle\nabla f(\mathbf{y}_t), \mathbf{x}-\mathbf{y}_t\rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}_t\|_2^2,$$

which means

$$\begin{aligned} & \Phi_{t+1}(\mathbf{x}) \leq \left(1-\frac{1}{\sqrt{\kappa}}\right)\Phi_t(\mathbf{x}) + \frac{1}{\sqrt{\kappa}} f(\mathbf{x}) \\ \Longleftrightarrow\quad & \Phi_{t+1}(\mathbf{x}) - f(\mathbf{x}) \leq \left(1-\frac{1}{\sqrt{\kappa}}\right)(\Phi_t(\mathbf{x}) - f(\mathbf{x})) \\ \Longleftrightarrow\quad & \Phi_t(\mathbf{x}) - f(\mathbf{x}) \leq \left(1-\frac{1}{\sqrt{\kappa}}\right)^t(\Phi_0(\mathbf{x}) - f(\mathbf{x})). \end{aligned} \tag{20}$$

We introduce the following lemma (which will be proved later).

Lemma 5.1. *The setting in this paragraph holds that*

$$f(\mathbf{x}_t) \leq \min_{\mathbf{x}\in\mathbb{R}^d}\Phi_t(\mathbf{x}).$$

Applying the result of (20) and Lemma 5.1, we have

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \min_{\mathbf{x}\in\mathbb{R}^d}\Phi_t(\mathbf{x}) - f(\mathbf{x}^*) \\ &\leq \Phi_t(\mathbf{x}^*) - f(\mathbf{x}^*) \\ &\leq \left(1-\frac{1}{\sqrt{\kappa}}\right)^t(\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)) \\ &= \left(1-\frac{1}{\sqrt{\kappa}}\right)^t\left(f(\mathbf{x}_0) + \frac{\mu}{2}\|\mathbf{x}_0-\mathbf{x}^*\|_2^2 - f(\mathbf{x}^*)\right). \end{aligned}$$

This implies achieving $f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \epsilon$ requires $t = \mathcal{O}(\sqrt{\kappa}\ln(1/\epsilon))$.
Now we prove Lemma 5.1.

Proof. We prove this lemma by induction. It is true for $t = 0$ since we have

$$
\begin{aligned}
f(\mathbf{x}_0) =& f(\mathbf{x}_0) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}_0\|_2^2 \\
=& \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}_0) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \\
=& \min_{\mathbf{x}\in\mathbb{R}^d} \Phi_0(\mathbf{x})
\end{aligned}
$$

For $t \geq 1$, the smoothness of f and the update

$$
\mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t)
$$

means

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) \leq& f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_{t+1} - \mathbf{y}_t\rangle + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2 \\
=& f(\mathbf{y}_t) - \frac{1}{2L}\|\nabla f(\mathbf{y}_t)\|_2^2 \\
=& \left(1 - \frac{1}{\sqrt{\kappa}}\right) f(\mathbf{x}_t) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)(f(\mathbf{y}_t) - f(\mathbf{x}_t)) + \frac{1}{\sqrt{\kappa}} f(\mathbf{y}_t) - \frac{1}{2L}\|\nabla f(\mathbf{y}_t)\|_2^2 \\
\leq& \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_t^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\langle \nabla f(\mathbf{y}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \frac{1}{\sqrt{\kappa}} f(\mathbf{y}_t) - \frac{1}{2L}\|\nabla f(\mathbf{y}_t)\|_2^2.
\end{aligned}
$$

Thus, we have to show the last line is smaller or equal to $\Phi_{t+1}^* = \min_{\mathbf{x}\in\mathbb{R}^d} \Phi_{t+1}(\mathbf{x})$. That is

$$
\left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_t^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\langle \nabla f(\mathbf{y}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \frac{1}{\sqrt{\kappa}} f(\mathbf{y}_t) - \frac{1}{2L}\|\nabla f(\mathbf{y}_t)\|_2^2 \leq \Phi_{t+1}^*. \tag{21}
$$

Note that for any t, the function Φ_t is quadratic and the induction implies

$$
\nabla^2 \Phi_t(\mathbf{x}) = \mu \mathbf{I}.
$$

Hence, the function Φ_t has the form of

$$
\Phi_t(\mathbf{x}) = \Phi_t^* + \frac{\mu}{2}\|\mathbf{x} - \mathbf{v}_t\|_2^2
$$

for some $\mathbf{v}_t \in \mathbb{R}^d$, and we have

$$
\nabla \Phi_t(\mathbf{x}) = \mu(\mathbf{x} - \mathbf{v}_t).
$$

Substituting above result into the recursion of (gradient of) Φ_t, we have

$$
\begin{aligned}
\nabla \Phi_{t+1}(\mathbf{x}) &= \left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla \Phi_t(\mathbf{x}) + \frac{1}{\sqrt{\kappa}}\left(\nabla f(\mathbf{y}_t) + \mu(\mathbf{x} - \mathbf{y}_t)\right) \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right)\mu(\mathbf{x} - \mathbf{v}_t) + \frac{1}{\sqrt{\kappa}}\left(\nabla f(\mathbf{y}_t) + \mu(\mathbf{x} - \mathbf{y}_t)\right).
\end{aligned}
$$

Since the minimizer of $\Phi_{t+1}(\mathbf{x})$ is at $\mathbf{v}_{t+1}$, we have

$$
\begin{aligned}
\mathbf{0} =& \nabla \Phi_{t+1}(\mathbf{v}_{t+1}) \\
=& \left(1 - \frac{1}{\sqrt{\kappa}}\right)\mu(\mathbf{v}_{t+1} - \mathbf{v}_t) + \frac{1}{\sqrt{\kappa}}\left(\nabla f(\mathbf{y}_t) + \mu(\mathbf{v}_{t+1} - \mathbf{y}_t)\right) \\
=& \mu \mathbf{v}_{t+1} - \left(1 - \frac{1}{\sqrt{\kappa}}\right)\mu \mathbf{v}_t + \frac{1}{\sqrt{\kappa}}\nabla f(\mathbf{y}_t) - \frac{1}{\sqrt{\kappa}}\mu \mathbf{y}_t \\
\Longrightarrow \quad \mathbf{v}_{t+1} =& \left(1 - \frac{1}{\sqrt{\kappa}}\right)\mathbf{v}_t - \frac{1}{\mu\sqrt{\kappa}}\nabla f(\mathbf{y}_t) + \frac{1}{\sqrt{\kappa}}\mathbf{y}_t.
\end{aligned} \tag{22}
$$

Substituting equations (22) and $\Phi_t(\mathbf{x}) = \Phi_t^* + \frac{\mu}{2}\|\mathbf{x}-\mathbf{v}_t\|_2^2$ into the recursion of $\Phi_{t+1}(\mathbf{x})$, we have

$$\begin{aligned}
&\Phi_{t+1}(\mathbf{y}_t)\\
=&\Phi_{t+1}^* + \frac{\mu}{2}\|\mathbf{y}_t-\mathbf{v}_{t+1}\|_2^2\\
=&\left(1-\frac{1}{\sqrt{\kappa}}\right)\left(\Phi_t^* + \frac{\mu}{2}\|\mathbf{y}_t-\mathbf{v}_t\|_2^2\right) + \frac{1}{\sqrt{\kappa}}\left(f(\mathbf{y}_t) + \langle\nabla f(\mathbf{y}_t), \mathbf{y}_t-\mathbf{y}_t\rangle + \frac{\mu}{2}\|\mathbf{y}_t-\mathbf{y}_t\|_2^2\right)\\
=&\left(1-\frac{1}{\sqrt{\kappa}}\right)\Phi_t^* + \frac{\mu}{2}\left(1-\frac{1}{\sqrt{\kappa}}\right)\|\mathbf{y}_t-\mathbf{v}_t\|_2^2 + \frac{1}{\sqrt{\kappa}}f(\mathbf{y}_t).
\end{aligned}$$

Equation (22) also implies

$$\begin{aligned}
\|\mathbf{y}_t-\mathbf{v}_{t+1}\|_2^2 &= \left\|\mathbf{y}_t - \left(\left(1-\frac{1}{\sqrt{\kappa}}\right)\mathbf{v}_t - \frac{1}{\mu\sqrt{\kappa}}\nabla f(\mathbf{y}_t) + \frac{1}{\sqrt{\kappa}}\mathbf{y}_t\right)\right\|_2^2\\
&= \left\|\left(1-\frac{1}{\sqrt{\kappa}}\right)(\mathbf{y}_t-\mathbf{v}_t) - \frac{1}{\mu\sqrt{\kappa}}\nabla f(\mathbf{y}_t)\right\|_2^2\\
&= \left(1-\frac{1}{\sqrt{\kappa}}\right)^2\|\mathbf{y}_t-\mathbf{v}_t\|_2^2 + \frac{1}{\mu^2\kappa}\|\nabla f(\mathbf{y}_t)\|_2^2 - \frac{2}{\mu\sqrt{\kappa}}\left(1-\frac{1}{\sqrt{\kappa}}\right)\langle\nabla f(\mathbf{y}_t), \mathbf{y}_t-\mathbf{v}_t\rangle.
\end{aligned}$$

Combining above results, we have

$$\begin{aligned}
\Phi_{t+1}^* &= \left(1-\frac{1}{\sqrt{\kappa}}\right)\Phi_t^* + \frac{1}{\sqrt{\kappa}}f(\mathbf{y}_t) + \frac{\mu}{2\sqrt{\kappa}}\left(1-\frac{1}{\sqrt{\kappa}}\right)\|\mathbf{y}_t-\mathbf{v}_t\|_2^2\\
&\quad - \frac{1}{2L}\|\nabla f(\mathbf{y}_t)\|_2^2 - \frac{1}{\sqrt{\kappa}}\left(1-\frac{1}{\sqrt{\kappa}}\right)\langle\nabla f(\mathbf{y}_t), \mathbf{y}_t-\mathbf{v}_t\rangle\\
&\geq \left(1-\frac{1}{\sqrt{\kappa}}\right)\Phi_t^* + \frac{1}{\sqrt{\kappa}}f(\mathbf{y}_t)\\
&\quad - \frac{1}{2L}\|\nabla f(\mathbf{y}_t)\|_2^2 - \frac{1}{\sqrt{\kappa}}\left(1-\frac{1}{\sqrt{\kappa}}\right)\langle\nabla f(\mathbf{y}_t), \mathbf{y}_t-\mathbf{v}_t\rangle.
\end{aligned}$$

Compared with equation (21), we only need to show

$$\mathbf{y}_t - \mathbf{v}_t = \sqrt{\kappa}(\mathbf{x}_t - \mathbf{y}_t).$$

This can be proved by induction and equation (22) as follows

$$\begin{aligned}
\mathbf{y}_{t+1} - \mathbf{v}_{t+1} =& \mathbf{y}_{t+1} - \left(\left(1-\frac{1}{\sqrt{\kappa}}\right)\mathbf{v}_t - \frac{1}{\mu\sqrt{\kappa}}\nabla f(\mathbf{y}_t) + \frac{1}{\sqrt{\kappa}}\mathbf{y}_t\right)\\
=&\mathbf{y}_{t+1} - \left(1-\frac{1}{\sqrt{\kappa}}\right)((1+\sqrt{\kappa})\mathbf{y}_t - \sqrt{\kappa}\mathbf{x}_t) + \frac{\sqrt{\kappa}}{L}\nabla f(\mathbf{y}_t) - \frac{1}{\sqrt{\kappa}}\mathbf{y}_t\\
=&\mathbf{y}_{t+1} - \sqrt{\kappa}\left(\mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t)\right) + (\sqrt{\kappa}-1)\mathbf{x}_t\\
=&\mathbf{y}_{t+1} - \sqrt{\kappa}\mathbf{x}_{t+1} + (\sqrt{\kappa}-1)\mathbf{x}_t\\
=&\sqrt{\kappa}\left(\frac{1}{\sqrt{\kappa}}\mathbf{y}_{t+1} - \mathbf{x}_{t+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\mathbf{x}_t\right)\\
=&\sqrt{\kappa}\left(\mathbf{x}_{t+1} - \mathbf{y}_{t+1}\right),
\end{aligned}$$

where the last step holds by taking

$$\beta_t = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \quad\Longrightarrow\quad \mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(\mathbf{x}_{t+1}-\mathbf{x}_t).$$

$\square$

Remark 5.6 (Strong Convexity to Non-Strong Convexity)**.** *We have study AGD iteration*

$$\begin{cases} \mathbf{y}_t = \mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1}), \\ \mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t). \end{cases}$$

where $\mathbf{x}_{-1} = \mathbf{x}_0$ *and* $\beta_t \in (0, 1)$. *For strongly-convex case, we have*

$$\beta_t = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \quad \Longrightarrow \quad f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t \left(f(\mathbf{x}_0) - f(\mathbf{x}^*) + \frac{\mu}{2}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2\right).$$

We can minimize non-strongly convex $f(\cdot)$ *by taking* $\delta = \mathcal{O}(\epsilon)$ *and use AGD to solve the problem*

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

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

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

which implies

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

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

We can let

$$\beta_{t+1} = \frac{1+\lambda_t}{\lambda_{t+1}}, \quad \text{where} \quad \lambda_0 = 0 \quad \text{and} \quad \lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_t^2}}{2} \geq 1 \quad \Longrightarrow \quad \lambda_t^2 = \lambda_{t+1}^2 - \lambda_{t+1}.$$

The smoothness and convexity implies

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

Taking $\mathbf{y} = \mathbf{x}_t$ and $\mathbf{y} = \mathbf{x}^*$, we have

$$\begin{cases} f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq L\left\langle \mathbf{y}_t - \mathbf{x}_{t+1}, \mathbf{y}_t - \mathbf{x}_t \right\rangle - \frac{1}{2L}\left\|\nabla f(\mathbf{y}_t)\right\|_2^2 \\ f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq L\left\langle \mathbf{y}_t - \mathbf{x}_{t+1}, \mathbf{y}_t - \mathbf{x}^* \right\rangle - \frac{1}{2L}\left\|\nabla f(\mathbf{y}_t)\right\|_2^2. \end{cases}$$

Multiplying the first inequality by $\lambda_t - 1$ and adding to the second one, we achieve

$$\begin{aligned}
&(\lambda_t - 1)(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)) + f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \\
=&\lambda_t(f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)) - (\lambda_t - 1)(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\
\leq&L \left\langle \mathbf{y}_t - \mathbf{x}_{t+1}, \lambda_t \mathbf{y}_t - (\lambda_t - 1)\mathbf{x}_t - \mathbf{x}^* \right\rangle - \frac{\lambda_t}{2L} \left\| \nabla f(\mathbf{y}_t) \right\|_2^2.
\end{aligned}$$

Multiplying λ_t leads to

$$\begin{aligned}
&\lambda_t^2(f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)) - \lambda_{t-1}^2(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\
\leq&L \left\langle \lambda_t(\mathbf{y}_t - \mathbf{x}_{t+1}), \lambda_t \mathbf{y}_t - (\lambda_t - 1)\mathbf{x}_t - \mathbf{x}^* \right\rangle - \frac{\lambda_t^2}{2L} \left\| \nabla f(\mathbf{y}_t) \right\|_2^2 \\
\leq&L \left\langle \lambda_t(\mathbf{y}_t - \mathbf{x}_{t+1}), \lambda_t \mathbf{y}_t - (\lambda_t - 1)\mathbf{x}_t - \mathbf{x}^* \right\rangle - \frac{L}{2} \left\| \lambda_t(\mathbf{y}_t - \mathbf{x}_{t+1}) \right\|_2^2 \\
=&\frac{L}{2} \left(\left\| \lambda_t(\mathbf{y}_t - \mathbf{x}_{t+1}) \right\|_2^2 + \left\| \lambda_t \mathbf{y}_t - (\lambda_t - 1)\mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \left\| \lambda_t \mathbf{x}_{t+1} - (\lambda_t - 1)\mathbf{x}_t - \mathbf{x}^* \right\|_2^2 \right) - \frac{L}{2} \left\| \lambda_t(\mathbf{y}_t - \mathbf{x}_{t+1}) \right\|_2^2 \\
=&\frac{L}{2} \left(\left\| \lambda_t \mathbf{y}_t - (\lambda_t - 1)\mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \left\| \lambda_t \mathbf{x}_{t+1} - (\lambda_t - 1)\mathbf{x}_t - \mathbf{x}^* \right\|_2^2 \right) \\
=&\frac{L}{2} \left(\left\| \lambda_t \mathbf{y}_t - (\lambda_t - 1)\mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \left\| \lambda_{t+1} \mathbf{y}_{t+1} - (\lambda_{t+1} - 1)\mathbf{x}_{t+1} - \mathbf{x}^* \right\|_2^2 \right),
\end{aligned}$$

where the last step is because of

$$\begin{aligned}
&\lambda_t \mathbf{x}_{t+1} - (\lambda_t - 1)\mathbf{x}_t = \lambda_{t+1}\mathbf{y}_{t+1} - (\lambda_{t+1} - 1)\mathbf{x}_{t+1} \\
\iff &\lambda_t \mathbf{x}_{t+1} - (\lambda_t - 1)\mathbf{x}_t = \lambda_{t+1}(\mathbf{x}_{t+1} - \beta_{t+1}(\mathbf{x}_{t+1} - \mathbf{x}_t)) - (\lambda_{t+1} - 1)\mathbf{x}_{t+1} \\
\iff &(\beta_{t+1}\lambda_{t+1} - \lambda_t - 1)(\mathbf{x}_{t+1} - \mathbf{x}_t) = \mathbf{0} \\
\iff &\beta_{t+1} = \frac{1 + \lambda_t}{\lambda_{t+1}}.
\end{aligned}$$

Summing over above inequality with $t = 1, \dots, T - 1$, we have

$$\begin{aligned}
&\lambda_{T-1}^2(f(\mathbf{x}_T) - f(\mathbf{x}^*)) - \lambda_0^2(f(\mathbf{x}_1) - f(\mathbf{x}^*)) \\
\leq&\frac{L}{2} \left(\left\| \lambda_1 \mathbf{y}_1 - (\lambda_1 - 1)\mathbf{x}_1 - \mathbf{x}^* \right\|_2^2 - \left\| \lambda_T \mathbf{y}_T - (\lambda_T - 1)\mathbf{x}_T - \mathbf{x}^* \right\|_2^2 \right),
\end{aligned}$$

that is

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{L}{2\lambda_{T-1}^2} \left\| \mathbf{y}_1 - \mathbf{x}^* \right\|_2^2.$$

Consider that $\lambda_0 = 0$, $\lambda_1 = 1$ and

$$\lambda_{t+1} = \frac{1 + \sqrt{1 + 4\lambda_t^2}}{2}.$$

We can prove $\lambda_t \geq (t+1)/2$ for $t \geq 1$, since it holds

$$\frac{1 + \sqrt{1 + 4((t+1)/2)^2}}{2} = \frac{1 + \sqrt{1 + (t+1)^2}}{2} \geq \frac{t+2}{2}.$$

Therefore, we have

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{2L}{T^2} \left\| \mathbf{y}_1 - \mathbf{x}^* \right\|_2^2.$$

Additionally, we have

$$\beta_1 = \frac{1 + \lambda_0}{\lambda_1} = 1, \qquad \mathbf{y}_0 = \mathbf{x}_0 + \beta_0(\mathbf{x}_0 - \mathbf{x}_{-1}) = \mathbf{x}_0, \quad \text{and}$$

$$\mathbf{y}_1 = \mathbf{x}_1 + \beta_1(\mathbf{x}_1 - \mathbf{x}_0) = 2\mathbf{x}_1 - \mathbf{x}_0 = 2\left(\mathbf{x}_0 - \frac{1}{L}\nabla f(\mathbf{x}_0)\right) - \mathbf{x}_0 = \mathbf{x}_0 - \frac{2}{L}\nabla f(\mathbf{x}_0),$$

which means

$$\begin{aligned} f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq & \frac{2L}{T^2}\left\|\mathbf{x}_0 - \frac{2}{L}\nabla f(\mathbf{x}_0) - \mathbf{x}^*\right\|_2^2 \\ \leq & \frac{4L}{T^2}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{4}{L^2}\|\nabla f(\mathbf{x}_0) - \nabla f(\mathbf{x}^*)\|_2^2\right) \\ \leq & \frac{4L}{T^2}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + 4\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right) \\ = & \frac{20L}{T^2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2. \end{aligned}$$

Hence, we can achieve $f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \epsilon$ within

$$T = \mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right).$$

iterations.

Assumption 5.1. *An iterative method $\mathcal{M}$ generates a sequence of test points $\{\mathbf{x}_t\}$ such that*

$$\mathbf{x}_t \in \mathbf{x}_0 + \mathrm{span}\{\nabla f(\mathbf{x}_0), \dots, \nabla f(\mathbf{x}_{t-1})\}.$$

Remark 5.7. *For AGD, we have*

$$\begin{aligned} &\textit{(view as } \mathbf{x}_1\textit{)} \quad && \mathbf{x}_1 = \mathbf{x}_0 - \eta_0\nabla f(\mathbf{x}_0), \\ &\textit{(view as } \mathbf{x}_2\textit{)} \quad && \mathbf{y}_1 = \mathbf{x}_1 + \beta_1(\mathbf{x}_1 - \mathbf{x}_0) \\ & && \quad\; = \mathbf{x}_0 - \eta_0\nabla f(\mathbf{x}_0) - \beta_1\eta_0\nabla f(\mathbf{x}_0), \\ & && \quad\; = \mathbf{x}_0 - (1+\beta_1)\eta_0\nabla f(\mathbf{x}_0), \\ &\textit{(view as } \mathbf{x}_3\textit{)} \quad && \mathbf{x}_2 = \mathbf{y}_1 - \eta_1\nabla f(\mathbf{y}_1) \\ & && \quad\; = \mathbf{x}_0 - (1+\beta_1)\eta_0\nabla f(\mathbf{x}_0) - \eta_1\nabla f(\mathbf{y}_1). \end{aligned}$$

We consider the "worst" functions $f : \mathbb{R}^d \to \mathbb{R}$ such that

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

where $\mathbf{x} = [x_1, \dots, x_d]^\top$ and $d \geq t$. We have

$$\frac{\partial^2 f_t(\mathbf{x})}{\partial(x_i)^2} = \frac{L}{2} \quad \text{for} \quad i = 1, \dots, t \qquad \text{and} \qquad \frac{\partial^2 f_t(\mathbf{x})}{\partial x_i \partial x_{i+1}} = -\frac{L}{4} \quad \text{for} \quad i = 1, \dots, t-1.$$

Smoothness and Convexity: We can verify $\nabla^2 f(\mathbf{x}) = \frac{L}{4}\mathbf{A}_t$ and $f(\mathbf{x}) = \frac{L}{4}\left(\frac{1}{2}\mathbf{x}^\top\mathbf{A}_t\mathbf{x} - \mathbf{e}_1^\top\mathbf{x}\right)$ 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

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

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

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

We also have

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

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

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

$$\nabla f_t(\mathbf{x}) = \frac{L}{4}(\mathbf{A}_t\mathbf{x} - \mathbf{e}_1) = \mathbf{0} \quad \iff \quad \begin{cases} 2x_1 - x_2 = 1 \\ -x_1 + 2x_2 - x_3 = 0 \\ \cdots \\ -x_{t-3} + 2x_{t-2} - x_{t-1} = 0 \\ -x_{t-2} + 2x_{t-1} - x_t = 0 \\ -x_{t-1} + 2x_t = 0 \end{cases}$$

$$\iff \begin{cases} 1 = (t+1)x_t \\ x_1 = tx_t \\ x_2 = (t-1)x_t \\ \cdots \\ x_{t-3} = 4x_t \\ x_{t-2} = 3x_t \\ x_{t-1} = 2x_t \end{cases} \iff \begin{cases} x_t = \dfrac{1}{t+1} \\ x_1 = \dfrac{t}{t+1} \\ x_2 = \dfrac{t-1}{t+1} \\ \cdots \\ x_{t-3} = \dfrac{4}{t+1} \\ x_{t-2} = \dfrac{3}{t+1} \\ x_{t-1} = \dfrac{2}{t+1} \end{cases} \iff x_i = \begin{cases} 1 - \dfrac{i}{t+1}, & i = 1, \dots, t, \\ 0, & i = t+1 \dots, d. \end{cases}$$

Then the optimal function value is

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

We also note that

$$\begin{aligned} \|\mathbf{x}_t^*\|_2^2 &= \sum_{i=1}^{t}\left(1 - \frac{i}{t+1}\right)^2 = t - \frac{2}{t+1}\sum_{i=1}^{t} i + \frac{1}{(t+1)^2}\sum_{i=1}^{t} i^2 \\ &\leq t - \frac{2}{t+1} \cdot \frac{t(t+1)}{2} + \frac{1}{(t+1)^2} \cdot \frac{t(t+1)(2t+1)}{6} \leq \frac{(t+1)^3}{3(t+1)^2} = \frac{t+1}{3}. \end{aligned} \tag{23}$$

Lower Bounds: We define

$$\mathbb{R}^{t,d} = \{\mathbf{x} \in \mathbb{R}^d : x_{t+1} = \cdots = x_d = 0\},$$

that is the subspace of $\mathbb{R}^d$, in which only the first t components of the point can differ from zero.

Lemma 5.2 (zero-chain)**.** *Let* $\mathbf{x}_0 = \mathbf{0}$*. Then for any sequence* $\{\mathbf{x}_1, \dots, \mathbf{x}_t\}$ *satisfying the condition*

$$\mathbf{x}_k \in \mathcal{L}_k = \operatorname{span}\{\nabla f_t(\mathbf{x}_0), \dots, \nabla f_t(\mathbf{x}_{k-1})\},$$

for $k = 1, \dots, t$*, we have* $\mathcal{L}_k \subseteq \mathbb{R}^{k,d}$*.*

Proof. Use induction by considering the tri-diagonal structure of $\mathbf{A}_t$. □

Lemma 5.3. *For all* $\mathbf{x} \in \mathbb{R}^{t,d}$*, we have* $f_t(\mathbf{x}) = f_p(\mathbf{x})$ *for* $p = t, t+1, \dots, d$*.*

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

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

and

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

□

Corollary 5.1. *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* $p = t, t+1, \dots, d$*, where* $f_t^* = \min_{\mathbf{x} \in \mathbb{R}^d} f_t(\mathbf{x})$*.*

Theorem 5.3. *For any* $t \in \mathbb{N}$ *and* $\mathbf{x}_0 \in \mathbb{R}^d$*, there exists an* L*-smooth and convex function* $f : \mathbb{R}^d \to \mathbb{R}$ *with* $t \in [1, (d-1)/2]$ *such that for any first-order algorithm* $\mathcal{M}$ *satisfying*

$$\mathbf{x}_k \in \mathbf{x}_0 + \operatorname{span}\{\nabla f(\mathbf{x}_0), \dots, \nabla f(\mathbf{x}_{k-1})\} = \mathbf{x}_0 + \mathcal{L}_k,$$

for all $k = 1, \dots, t$*, we have*

$$f(\mathbf{x}_t) - f^* \geq \frac{3L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{8(t+1)^2} \qquad \text{and} \qquad \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \geq \frac{1}{4} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2.$$

where $\mathbf{x}^* \in \mathbb{R}^d$ *is the minimizer of* f *and* $f^* = f(\mathbf{x}^*)$*.*

Proof. We apply algorithm $\mathcal{M}$ to minimize function

$$f(\mathbf{x}) \triangleq f_{2t+1}(\mathbf{x})$$

which starts from $\mathbf{x}_0$ generate $\{\mathbf{x}_1, \dots, \mathbf{x}_t\}$. We suppose $\mathbf{x}_0 = \mathbf{0}$, otherwise we just need to consider the problem of minimizing $f(\mathbf{x} + \mathbf{x}_0)$ with initial point $\mathbf{0}$.

Using Corollary 5.1 with $p = 2t+1$, we have

$$f_{2t+1}(\mathbf{x}_t) \geq f_t^* = -\frac{L}{8}\left(1 - \frac{1}{t+1}\right).$$

On the other hand, we have

$$f^* = f^*_{2t+1}(\mathbf{x}_t) = -\frac{L}{8}\left(1 - \frac{1}{2t+2}\right).$$

Combining inequality (23), we have

$$\frac{f(\mathbf{x}_t) - f^*}{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2} = \frac{f_{2t+1}(\mathbf{x}_t) - f^*}{\left\|\mathbf{x}_0 - \mathbf{x}^*_{2t+1}\right\|_2^2} \geq \frac{-\frac{L}{8}\left(1 - \frac{1}{t+1}\right) + \frac{L}{8}\left(1 - \frac{1}{2t+2}\right)}{\frac{(2t+1)+1}{3}} = \frac{3L}{8(t+1)^2}$$

For the distance, recall that Lemma 5.2 implies $\mathbf{x}_t \in \mathbb{R}^{t,d}$, that is

$$x_t^{(t+1)} = x_t^{(t+1)} = \cdots = x_t^{(2t+1)} = 0.$$

Hence, we can bound the distance as follows

$$\begin{aligned}
\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 &= \sum_{i=1}^{d}(x_t^{(i)} - x^*_{2t+1}{}^{(i)})^2 \geq \sum_{i=t+1}^{2t+1}(x_t^{(i)} - x^{*(d)}_{2t+1})^2 = \sum_{i=t+1}^{2t+1}(x^{*(d)}_{2t+1})^2 \\
&= \sum_{i=t+1}^{2t+1}\left(1 - \frac{i}{2t+2}\right)^2 = t+1 - \frac{1}{t+1}\sum_{i=t+1}^{2t+1} i + \frac{1}{4(t+1)^2}\sum_{i=t+1}^{2t+1} i^2 = \frac{2t^2+7t+6}{24(t+1)} \geq \frac{1}{4}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2.
\end{aligned}$$

□

Remark 5.8. *For any $\epsilon > 0$, there exists function $f : \mathbb{R}^d \to \mathbb{R}$ with $d = \Theta\big(\sqrt{L/\epsilon}\big)$ such that finding $\mathbf{x}$ with $f(\mathbf{x}) - f^* \leq \epsilon$ requires at least $\Omega\big(\sqrt{L/\epsilon}\big)$ iterations of first-order methods.*

Now we consider the lower complexity bound for minimizing strongly-convex function. We introduce

$$\begin{aligned}
f(\mathbf{x}) =& \frac{L-\mu}{4}\left(\frac{1}{2}\left(x_1^2 + \sum_{i=1}^{d-1}(x_i - x_{i+1})^2 + \left(1 - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)x_d^2\right) - x_1\right) + \frac{\mu}{2}\|\mathbf{x}\|_2^2 \\
=& \frac{L-\mu}{4}\left(\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{e}_1^\top \mathbf{x}\right) + \frac{\mu}{2}\|\mathbf{x}\|_2^2
\end{aligned}$$

for $t = 1, \dots, d$, where $\mathbf{x} = [x_1, \dots, x_d]^\top$ 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

$$\begin{aligned}
\left\langle \mathbf{s}, \nabla^2 f(\mathbf{x})\mathbf{s}\right\rangle =& \frac{L-\mu}{4}\left(s_1^2 + \sum_{i=1}^{d-1}(s_i - s_{i+1})^2 + \left(1 - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)s_d^2\right) + \mu\|\mathbf{s}\|_2^2 \\
\leq& \frac{L-\mu}{4}\left(s_1^2 + \sum_{i=1}^{d-1}(2s_i^2 + 2s_{i+1}^2) + \left(1 - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)s_d^2\right) + \mu\|\mathbf{s}\|_2^2 \\
\leq& (L-\mu)\|\mathbf{s}\|_2^2 + \mu\|\mathbf{s}\|_2^2 = L\|\mathbf{x}\|_2^2
\end{aligned}$$

and $\left\langle \mathbf{s}, \nabla^2 f(\mathbf{x})\mathbf{s}\right\rangle \geq \mu\|\mathbf{s}\|_2^2$. Hence, the function is L-smooth and μ-strongly convex.

2. The optimal solution should satisfies

$$\frac{L-\mu}{4}(\mathbf{A}\mathbf{x}-\mathbf{e}_1)+\mu\mathbf{x}=\mathbf{0} \qquad \Longleftrightarrow \qquad \left(\mathbf{A}+\frac{4}{\kappa-1}\mathbf{I}\right)\mathbf{x}=\mathbf{e}_1,$$

which leads to

$$\begin{cases} \dfrac{2(\kappa+1)}{\kappa-1}x_1 - x_2 = 1 \\ -x_1 + \dfrac{2(\kappa+1)}{\kappa-1}x_2 - x_3 = 0 \\ \cdots\cdots \\ -x_{d-2} + \dfrac{2(\kappa+1)}{\kappa-1}x_{d-1} - x_d = 0 \\ -x_{d-1} + \dfrac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}x_d = 0 \end{cases} \quad \Longrightarrow \quad x_i = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{i-1} x_1 \quad \Longrightarrow \quad x_i = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{i} = q^i,$$

where we use the fact $2+4/(\kappa-1)=2(\kappa+1)/(\kappa-1)$.

Let $d=2t$. Combining above results with zero-chain property, we have

$$\|\mathbf{x}_t-\mathbf{x}^*\|_2^2 \geq \sum_{i=t+1}^{d}\left\|x^{*(i)}\right\|_2^2 = \sum_{i=t+1}^{d} q^{2i} = \sum_{i=t+1}^{2t} q^{2i}.$$

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

$$\begin{aligned} &\frac{f(\mathbf{x}_t)-f^*}{\|\mathbf{x}_0-\mathbf{x}^*\|_2^2} \geq \frac{\frac{\mu}{2}\|\mathbf{x}_t-\mathbf{x}^*\|_2^2}{\|\mathbf{x}_0-\mathbf{x}^*\|_2^2} = \frac{\mu}{2}\cdot\frac{\sum_{i=t+1}^{2t} q^{2i}}{\sum_{i=1}^{2t} q^{2i}} \\ =&\frac{\mu}{2}\cdot\frac{q^{2t}\sum_{i=1}^{t} q^{2i}}{(1+q^{2t})\sum_{i=1}^{t} q^{2i}} = \frac{\mu}{2}\cdot\frac{q^{2t}}{1+q^{2t}} \\ =&\frac{\mu}{2}\cdot\frac{(\sqrt{\kappa}-1)^{2t}}{(\sqrt{\kappa}+1)^{2t}+(\sqrt{\kappa}-1)^{2t}} \geq \frac{\mu}{4}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} = \frac{\mu}{4}\left(1-\frac{2}{\sqrt{\kappa}+1}\right)^{2t} \end{aligned}$$

Remark 5.9. *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.*

Making the gradient small: Recall that for L-smooth and convex function $f:\mathbb{R}^d\to\mathbb{R}$, we have

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

for any $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$. Taking $\mathbf{x}=\mathbf{x}^*$ and $\mathbf{y}=\mathbf{x}_T$, running AGD on μ-strongly convex f holds that

$$\begin{aligned} &f(\mathbf{x}^*)+\frac{1}{2L}\|\nabla f(\mathbf{x}_T)\|_2^2 \leq f(\mathbf{x}_T) \\ \Longrightarrow &\|\nabla f(\mathbf{x}_T)\|_2^2 \leq 2L(f(\mathbf{x}_T)-f(\mathbf{x}^*)) \leq 2L\left(1-\frac{1}{\sqrt{\kappa}}\right)^t\left(f(\mathbf{x}_0)-f(\mathbf{x}^*)+\frac{\mu}{2}\|\mathbf{x}_0-\mathbf{x}^*\|_2^2\right). \end{aligned}$$

We can achieve $\|\nabla f(\mathbf{x}_T)\|_2\leq\epsilon$ within $\mathcal{O}(\sqrt{\kappa}\ln(L/\epsilon))$ iterations.

For non-strongly convex f, we have

$$\|\nabla f(\mathbf{x}_T)\|_2^2 \leq 2L(f(\mathbf{x}_T)-f(\mathbf{x}^*)) \leq 2L\cdot\frac{20L}{T^2}\|\mathbf{x}_0-\mathbf{x}^*\|_2^2.$$

We can achieve $\|\nabla f(\mathbf{x}_T)\|_2\leq\epsilon$ within $\mathcal{O}(L/\epsilon)$ iterations.

Regularization: We use AGD to solving the problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \hat{f}(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}-\mathbf{x}_0\|_2^2,$$

then we have $\nabla\hat{f}(\mathbf{x}_T) = \nabla f(\mathbf{x}_T) + \lambda(\mathbf{x}_T - \mathbf{x}_0)$ and

$$\|\nabla f(\mathbf{x}_T)\|_2^2 \le 2\left\|\nabla\hat{f}(\mathbf{x}_T)\right\|_2^2 + 2\lambda^2\|\mathbf{x}-\mathbf{x}_0\|_2^2.$$

For the first term, we have

$$\left\|\nabla\hat{f}(\mathbf{x}_T)\right\|_2^2 \le 2(L+\lambda)\left(1-\sqrt{\frac{\lambda}{L+\lambda}}\right)^T\left(\hat{f}(\mathbf{x}_0)-\hat{f}(\mathbf{x}^*)+\frac{\lambda}{2}\|\mathbf{x}_0-\hat{\mathbf{x}}^*\|_2^2\right)$$

Taking $\lambda = \mathcal{O}(\epsilon)$ and $T = \mathcal{O}(\sqrt{L/\epsilon}\ln(L/\epsilon))$ leads to $\|\nabla f(\mathbf{x}_T)\|_2 \le \epsilon$.

Proximal Point Methods for Convex Optimization: We consider the proximal point iterations

$$\mathbf{x}_s \approx \arg\min_{\mathbf{x}\in\mathbb{R}^d} f_s(x), \quad \text{where } \; f_t(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\sigma_s}{2}\|\mathbf{x}-\bar{\mathbf{x}}_s\|_2^2 \;\text{ for some } \bar{\mathbf{x}}\in\mathbb{R}^d.$$

We apply the adaptive regularization in proximal point methods($\bar{\mathbf{x}}_0 = \mathbf{x}_0$ and $\sigma_0 = 0$)

$$\begin{cases}\sigma_s = 4^{s-2}\epsilon/D\\ \gamma_s = 1-\sigma_{s-1}/\sigma_s\\ \bar{\mathbf{x}}_s = (1-\gamma_s)\bar{\mathbf{x}}_{s-1} + \gamma_s\mathbf{x}_{s-1}\\ \mathbf{x}_s = \mathrm{AGD}\left(f_s(\cdot), \mathbf{x}_{s-1}, N_s\right)\end{cases} \quad \text{where} \quad f_s(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\sigma_s}{2}\|\mathbf{x}-\bar{\mathbf{x}}_s\|_2^2 \quad \text{and} \quad D = \min_{x^*\in\mathcal{X}}\|\mathbf{x}_0-\mathbf{x}^*\|_2.$$

We denote

$$\mathbf{x}_s^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} f_s(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\sigma_s}{2}\|\mathbf{x}-\bar{\mathbf{x}}_s\|_2^2.$$

Lemma 5.4. *For all $s \ge 1$, we have*

$$\|\mathbf{x}_{s-1}-\mathbf{x}_s^*\|_2 \le \left\|\mathbf{x}_{s-1}-\mathbf{x}_{s-1}^*\right\|_2, \tag{24}$$

$$\sigma_s\|\bar{\mathbf{x}}_s-\mathbf{x}_s^*\|_2 \le \sum_{i=1}^{s}(\sigma_{i-1}+\sigma_i)\left\|\mathbf{x}_{i-1}^*-\mathbf{x}_{i-1}\right\|_2. \tag{25}$$

Lemma 5.5. *The AGD step in the algorithm holds that*

$$f_s(\mathbf{x}_s) - f_s(\mathbf{x}_s^*) \le \frac{cL}{N_s^2}\|\mathbf{x}_{s-1}-\mathbf{x}_s^*\|_2^2.$$

Theorem 5.4. *The above proximal point iteration can achieve $\|\nabla f(\mathbf{x}_S)\|_2 \le \epsilon$ within $\mathcal{O}(\sqrt{LD/\epsilon})$ gradient calls by taking $S = 1 + \lceil\log_4(LD/\epsilon)\rceil$ and $\sigma_s = 4^{s-2}\epsilon/D$.*

Proof. The optimality of x_s^* means

$$\nabla f_s(\mathbf{x}_s^*) = \nabla f(\mathbf{x}_s^*) + \sigma_s(\mathbf{x}_s^* - \bar{\mathbf{x}}_s) = \mathbf{0}.$$

Then we have (noticing that $\mathbf{x}_0^* = \mathbf{x}^*$ and using inequality (25) in the last step)

$$\begin{aligned}\|\nabla f(\mathbf{x}_s)\|_2 &= \|\nabla f(\mathbf{x}_s) - \nabla f(\mathbf{x}_s^*) - \sigma_s(\mathbf{x}_s^*-\bar{\mathbf{x}}_s)\|_2\\ &\le \|\nabla f(\mathbf{x}_s) - \nabla f(\mathbf{x}_s^*)\|_2 + \sigma_s\|\mathbf{x}_s^*-\bar{\mathbf{x}}_s\|_2\\ &\le L\|\mathbf{x}_s-\mathbf{x}_s^*\|_2 + \sigma_s\|\mathbf{x}_s^*-\bar{\mathbf{x}}_s\|_2\\ &\le L\|\mathbf{x}_s-\mathbf{x}_s^*\|_2 + \sigma_1\|\mathbf{x}^*-\mathbf{x}_0\|_2 + \sum_{i=2}^{s}(\sigma_{i-1}+\sigma_i)\left\|\mathbf{x}_{i-1}^*-\mathbf{x}_{i-1}\right\|_2.\end{aligned} \tag{26}$$

Noting that the function f_s is σ_s-strongly convex and applying Lemma 5.5, we have

$$\begin{aligned} & f_s(\mathbf{x}_s) - f_s(\mathbf{x}_s^*) \geq \frac{\sigma_s}{2} \|\mathbf{x}_s - \mathbf{x}_s^*\|_2^2 \\ \Longrightarrow & \|\mathbf{x}_s - \mathbf{x}_s^*\|_2 \leq \sqrt{\frac{2}{\sigma_s}(f_s(\mathbf{x}_s) - f_s(\mathbf{x}_s^*))} \leq \frac{1}{N_s}\sqrt{\frac{2cL}{\sigma_s}} \|\mathbf{x}_{s-1} - \mathbf{x}_s^*\|_2 . \end{aligned}$$

Taking $N_s = \lceil 8\sqrt{2cL/\sigma_s} \rceil$ and using inequality (24), we have

$$\|\mathbf{x}_s - \mathbf{x}_s^*\|_2 \leq \frac{1}{8} \|\mathbf{x}_{s-1} - \mathbf{x}_s^*\|_2 \leq \frac{1}{8} \left\|\mathbf{x}_{s-1} - \mathbf{x}_{s-1}^*\right\|_2 . \tag{27}$$

The settings $S = 1 + \lceil \log_4(LD/\epsilon) \rceil$ and $\sigma_s = 4^{s-2}\epsilon/D$ leads to $\sigma_S \leq L$ and $\sigma_s = 4\sigma_{s-1}$ for all $s \geq 1$. Combining the results of (26) and (27), we have

$$\begin{aligned} \|\nabla f(\mathbf{x}_S)\|_2 \leq & L \|\mathbf{x}_s - \mathbf{x}_s^*\|_2 + \sigma_1 \|\mathbf{x}^* - \mathbf{x}_0\|_2 + \sum_{i=2}^{s} (\sigma_{i-1} + \sigma_i) \left\|\mathbf{x}_{i-1}^* - \mathbf{x}_{i-1}\right\|_2 \\ \leq & 8^{-S} L \|\mathbf{x}_0 - \mathbf{x}^*\|_2 + \sigma_1 \|\mathbf{x}^* - \mathbf{x}_0\|_2 + \sum_{i=2}^{S} (4^{i-2} + 4^{i-1})\sigma_1 \cdot 8^{-(i+1)} \|\mathbf{x}^* - \mathbf{x}_0\|_2 \\ \leq & \frac{L}{2} \cdot 4^{-S} \|\mathbf{x}^* - \mathbf{x}_0\|_2 + \frac{9\sigma_1}{4} \|\mathbf{x}_0 - \mathbf{x}^*\|_2 . \end{aligned}$$

Then the setting of S and σ_1 results $\|\nabla f(\mathbf{x}_S)\|_2 \leq \epsilon$.

Noticing that $\sigma_s \leq \sigma_S \leq L$ for all $s = 1, \ldots, S$, we have

$$N_s \leq 1 + 8\sqrt{2cL/\sigma_s} \leq \left(1 + 8\sqrt{2c}\right)\sqrt{L/\sigma_s}.$$

Recalling $\sigma_s = 4\sigma_{s-1}$, we have

$$\sum_{s=1}^{S} N_s \leq \left(1 + 8\sqrt{2c}\right)\sqrt{L} \sum_{s=1}^{S} \frac{1}{\sqrt{\sigma_s}} \leq \left(1 + 8\sqrt{2c}\right)\sqrt{\frac{L}{\sigma_1}} \sum_{s=1}^{S} 2^{-(s-1)} \leq 2\left(1 + 8\sqrt{2c}\right)\sqrt{\frac{LD}{\epsilon}}.$$

□

Then we prove Lemma 5.4.

Proof. **Part I:** We can write

$$\mathbf{x}_s^* = \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) + \frac{\sigma_{s-1}}{2} \|\mathbf{x} - \bar{\mathbf{x}}_{s-1}\|_2^2 + \frac{\sigma_s - \sigma_{s-1}}{2} \|\mathbf{x} - \mathbf{x}_{s-1}\|_2^2 ,$$

which is because of

$$\begin{aligned} & \sigma_s(\mathbf{x} - \bar{\mathbf{x}}_s) = \sigma_{s-1}(\mathbf{x} - \bar{\mathbf{x}}_{s-1}) + (\sigma_s - \sigma_{s-1})(\mathbf{x} - \mathbf{x}_{s-1}) \\ \Longleftrightarrow & \sigma_s \mathbf{x} - \sigma_s \bar{\mathbf{x}}_s = \sigma_{s-1}\mathbf{x} - \sigma_{s-1}\bar{\mathbf{x}}_{s-1} + (\sigma_s - \sigma_{s-1})\mathbf{x} - (\sigma_s - \sigma_{s-1})\mathbf{x}_{s-1} \\ \Longleftrightarrow & -\sigma_s \bar{\mathbf{x}}_s = -\sigma_{s-1}\bar{\mathbf{x}}_{s-1} - (\sigma_s - \sigma_{s-1})\mathbf{x}_{s-1} \end{aligned}$$

and

$$\begin{aligned} \sigma_s \bar{\mathbf{x}}_s = & \sigma_s((1 - \gamma_s)\bar{\mathbf{x}}_{s-1} + \gamma_s \mathbf{x}_{s-1}) \\ = & \sigma_s \left(\frac{\sigma_{s-1}}{\sigma_s} \bar{\mathbf{x}}_{s-1} + \left(1 - \frac{\sigma_{s-1}}{\sigma_s}\right) \mathbf{x}_{s-1} \right) \\ = & \sigma_{s-1}\bar{\mathbf{x}}_{s-1} + (\sigma_s - \sigma_{s-1})\mathbf{x}_{s-1}. \end{aligned}$$

The optimality of $\mathbf{x}_{s-1}^*$ and $\mathbf{x}_s^*$ indicates

$$
\begin{aligned}
& f(\mathbf{x}_{s-1}^*) + \frac{\sigma_{s-1}}{2}\left\|\mathbf{x}_{s-1}^* - \bar{\mathbf{x}}_{s-1}\right\|_2^2 + \frac{\sigma_s - \sigma_{s-1}}{2}\left\|\mathbf{x}_s^* - \mathbf{x}_{s-1}\right\|_2^2 \\
\leq & f(\mathbf{x}_s^*) + \frac{\sigma_{s-1}}{2}\left\|\mathbf{x}_s^* - \bar{\mathbf{x}}_{s-1}\right\|_2^2 + \frac{\sigma_s - \sigma_{s-1}}{2}\left\|\mathbf{x}_s^* - \mathbf{x}_{s-1}\right\|_2^2 \\
\leq & f(\mathbf{x}_{s-1}^*) + \frac{\sigma_{s-1}}{2}\left\|\mathbf{x}_{s-1}^* - \bar{\mathbf{x}}_{s-1}\right\|_2^2 + \frac{\sigma_s - \sigma_{s-1}}{2}\left\|\mathbf{x}_{s-1}^* - \mathbf{x}_{s-1}\right\|_2^2
\end{aligned}
$$

which concludes inequality (24) because of $\sigma_s > \sigma_{s-1}$.

Part II: We denote $\alpha_s = \sigma_s - \sigma_{s-1}$, then we have

$$
\gamma_s = 1 - \frac{\sigma_{s-1}}{\sigma_s} = \frac{\alpha_s}{\sigma_s}
$$

and

$$
\sigma_s\bar{\mathbf{x}}_s = \sigma_s(1-\gamma_s)\bar{\mathbf{x}}_{s-1} + \sigma_s\gamma_s\mathbf{x}_{s-1} = (\sigma_s - \alpha_s)\bar{\mathbf{x}}_{s-1} + \alpha_s\mathbf{x}_{s-1} = \sigma_{s-1}\bar{\mathbf{x}}_{s-1} + \alpha_s\mathbf{x}_{s-1}.
$$

Recall that $\sigma_0 = 0$, then above recursion leads to

$$
\sigma_s\bar{\mathbf{x}}_s = \sum_{i=1}^{s}\alpha_i\mathbf{x}_{i-1} \qquad \Longrightarrow \qquad \bar{\mathbf{x}}_s = \sum_{i=1}^{s}\frac{\alpha_i}{\sigma_s}\cdot\mathbf{x}_{i-1},
$$

which implies $\mathbf{x}$ is a convex combination of $\mathbf{x}_0, \dots, \mathbf{x}_{s-1}$ with weights α_i/σ_s since $\sum_{i=1}^{s}\alpha_i = \sigma_s$. Therefore, we have

$$
\begin{aligned}
& \sigma_s(\bar{\mathbf{x}}_s - \mathbf{x}_s^*) = \sum_{i=1}^{s}\alpha_i(\mathbf{x}_{i-1} - \mathbf{x}_s^*) \\
=& \alpha_s(\mathbf{x}_{s-1} - \mathbf{x}_s^*) + \sum_{i=1}^{s-1}\alpha_i(\mathbf{x}_{i-1} - \mathbf{x}_{s-1}^*) + \left(\sum_{i=1}^{s-1}\alpha_i\right)(\mathbf{x}_{s-1}^* - \mathbf{x}_s^*) \\
=& \alpha_s(\mathbf{x}_{s-1} - \mathbf{x}_s^*) + \sigma_{s-1}(\bar{\mathbf{x}}_{s-1} - \mathbf{x}_{s-1}^*) + \sigma_{s-1}(\mathbf{x}_{s-1}^* - \mathbf{x}_{s-1}) + \sigma_{s-1}(\mathbf{x}_{s-1} - \mathbf{x}_s^*) \\
=& \sigma_{s-1}(\bar{\mathbf{x}}_{s-1} - \mathbf{x}_{s-1}^*) + \sigma_{s-1}(\mathbf{x}_{s-1}^* - \mathbf{x}_{s-1}) + \sigma_s(\mathbf{x}_{s-1} - \mathbf{x}_s^*).
\end{aligned}
$$

The above recursion yields

$$
\sigma_s(\bar{\mathbf{x}}_s - \mathbf{x}_s^*) = \sum_{i=1}^{s}\left(\sigma_{i-1}(\mathbf{x}_{i-1}^* - \mathbf{x}_{i-1}) + \sigma_s(\mathbf{x}_{i-1} - \mathbf{x}_i^*)\right).
$$

Taking the norm and applying the result of first part, we have

$$
\begin{aligned}
\left\|\sigma_s(\bar{\mathbf{x}}_s - \mathbf{x}_s^*)\right\|_2 &\leq \sum_{i=1}^{s}\left(\sigma_{i-1}\left\|\mathbf{x}_{i-1}^* - \mathbf{x}_{i-1}\right\|_2 + \sigma_i\left\|\mathbf{x}_{i-1} - \mathbf{x}_i^*\right\|_2\right) \\
&\leq \sum_{i=1}^{s}\left(\sigma_{i-1}\left\|\mathbf{x}_{i-1}^* - \mathbf{x}_{i-1}\right\|_2 + \sigma_i\left\|\mathbf{x}_{i-1} - \mathbf{x}_{i-1}^*\right\|_2\right) \\
&= \sum_{i=1}^{s}(\sigma_{i-1} + \sigma_i)\left\|\mathbf{x}_{i-1}^* - \mathbf{x}_{i-1}\right\|_2.
\end{aligned}
$$

□

Proximal Point Methods for Nonconvex Optimization We have shown that gradient descent can achieves

$$\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}})\right\|_2^2 = \frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \leq \frac{L(f(\mathbf{x}_0)-f^*)}{T},$$

where $\hat{\mathbf{x}}$ is sampled from $\{\mathbf{x}_0,\dots,\mathbf{x}_{T-1}\}$. The complexity $\mathcal{O}(L\epsilon^{-2})$ is optimal to achieve $\left\|\nabla f(\mathbf{x})\right\|_2 \leq \epsilon$ for L-smooth f.

We additionally suppose $f:\mathbb{R}^d\to\mathbb{R}$ is weakly convex, that is

$$f(\mathbf{x}) + \frac{\ell}{2}\left\|\mathbf{x}\right\|_2^2 \text{ is convex} \quad\Longrightarrow\quad f(\mathbf{x}) + \ell\left\|\mathbf{x}\right\|_2^2 \text{ is } \ell\text{-strongly convex.}$$

Consider the perfect iteration

$$\mathbf{x}_s = \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \ell\left\|\mathbf{x}-\mathbf{x}_{s-1}\right\|_2^2.$$

We have

$$f(\mathbf{x}_s) + \ell\left\|\mathbf{x}_s-\mathbf{x}_{s-1}\right\|_2^2 \leq f(\mathbf{x}_{s-1}) + \ell\left\|\mathbf{x}_{s-1}-\mathbf{x}_{s-1}\right\|_2^2 = f(\mathbf{x}_{s-1})$$

and

$$\nabla f(\mathbf{x}_s) + 2\ell(\mathbf{x}_s-\mathbf{x}_{s-1}) = \mathbf{0} \qquad\Longrightarrow\qquad \frac{1}{4\ell}\left\|\nabla f(\mathbf{x}_s)\right\|_2^2 = \ell\left\|\mathbf{x}_s-\mathbf{x}_{s-1}\right\|_2^2.$$

Therefore, it holds

$$\begin{aligned}
& f(\mathbf{x}_s) + \frac{1}{4\ell}\left\|\nabla f(\mathbf{x}_s)\right\|_2^2 \leq f(\mathbf{x}_{s-1}) \\
\Longrightarrow & \left\|\nabla f(\mathbf{x}_s)\right\|_2^2 \leq 4\ell(f(\mathbf{x}_{s-1}) - f(\mathbf{x}_s)) \\
\Longrightarrow & \frac{1}{S}\sum_{s=1}^{S}\left\|\nabla f(\mathbf{x}_s)\right\|_2^2 \leq \frac{4\ell}{S}(f(\mathbf{x}_0)-f^*).
\end{aligned}$$

We require $S=\mathcal{O}(\ell\epsilon^{-2})$ to achieve ϵ-stationary point.

In practice, we consider the iteration

$$\mathbf{x}_s \approx \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \ell\left\|\mathbf{x}-\mathbf{x}_{s-1}\right\|_2^2.$$

We can solve the sub-problem by AGD such that

$$\left\|\mathbf{x}_s-\mathbf{x}_s^*\right\|_2 \leq \alpha\left\|\mathbf{x}_{s-1}-\mathbf{x}_s^*\right\|_2 \quad \text{for some } \alpha\in(0,1) \quad \text{where } \mathbf{x}_s^* = \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \ell\left\|\mathbf{x}-\mathbf{x}_{s-1}\right\|_2^2.$$

We have

$$\begin{aligned}
& f(\mathbf{x}_s) - f(\mathbf{x}_{s-1}) \\
= & f(\mathbf{x}_s) + \ell\left\|\mathbf{x}_s-\mathbf{s}_{s-1}\right\|_2^2 - (f(\mathbf{x}_{s-1}) + \ell\left\|\mathbf{x}_{s-1}-\mathbf{s}_{s-1}\right\|_2^2) - \ell\left\|\mathbf{x}_s-\mathbf{s}_{s-1}\right\|_2^2 \\
= & f_s(\mathbf{x}_s) - f_s(\mathbf{x}_{s-1}) - \ell\left\|\mathbf{x}_s-\mathbf{s}_{s-1}\right\|_2^2.
\end{aligned}$$

The triangle inequality leads to

$$\left\|\mathbf{x}_s-\mathbf{x}_{s-1}\right\|_2 = \left|\,\left\|\mathbf{x}_s-\mathbf{x}_s^*\right\|_2 - \left\|\mathbf{x}_s^*-\mathbf{s}_{s-1}\right\|_2\,\right| \geq (1-\alpha)\left\|\mathbf{x}_s^*-\mathbf{x}_{s-1}\right\|_2.$$

The optimal conditions implies

$$\nabla f(\mathbf{x}_s^*) + 2\ell(\mathbf{x}_s^*-\mathbf{x}_{s-1}) = \mathbf{0} \quad\Longrightarrow\quad \left\|\nabla f(\mathbf{x}_s^*)\right\|_2 = 2\ell\left\|\mathbf{x}_s^*-\mathbf{x}_{s-1}\right\|_2.$$

Combining above results, we have

$$\begin{aligned}
&f(\mathbf{x}_s) - f(\mathbf{x}_{s-1}) \\
=&f_s(\mathbf{x}_s) - f_s(\mathbf{x}_{s-1}) - \ell \left\| \mathbf{x}_s - \mathbf{x}_{s-1} \right\|_2^2 \\
\leq&f_s(\mathbf{x}_s) - f_s(\mathbf{x}_{s-1}) - (1-\alpha)^2 \ell \left\| \mathbf{x}_s^* - \mathbf{x}_{s-1} \right\|_2^2 \\
\leq& - (1-\alpha)^2 \ell \left\| \mathbf{x}_s^* - \mathbf{x}_{s-1} \right\|_2^2 \\
=& - \frac{(1-\alpha)^2}{4\ell} \left\| \nabla f(\mathbf{x}_s^*) \right\|_2^2 .
\end{aligned} \tag{28}$$

Then we have

$$\sum_{s=1}^{S} \frac{(1-\alpha)^2}{4\ell} \left\| \nabla f(\mathbf{x}_s^*) \right\|_2^2 \leq f(\mathbf{x}_0) - f(\mathbf{x}_S) \quad \Longrightarrow \quad \frac{1}{S} \sum_{s=1}^{S} \left\| \nabla f(\mathbf{x}_s^*) \right\|_2^2 \leq \frac{4\ell(f(\mathbf{x}_0) - f(\mathbf{x}_S))}{(1-\alpha)^2 S}.$$

Taking $\alpha = 1/2$ and $S = \lceil 32\ell\epsilon^{-2} \rceil$, we can achieve some $\mathbf{x}_s^*$ which is an $\epsilon/2$-stationary point in expectation. Noticing that $f_s(\cdot)$ is $(L+2\ell)$-smooth and ℓ-strongly-convex, we can apply AGD to solve

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) + \ell \left\| \mathbf{x} - \mathbf{x}_s \right\|_2^2$$

with the initial point $\mathbf{x}_{s-1}$ and iterations T leads to

$$\begin{aligned}
&\ell \left\| \mathbf{x}_s - \mathbf{x}_s^* \right\|_2^2 \\
\leq&f_s(\mathbf{x}_s) - f_s(\mathbf{x}_s^*) \\
\leq&\left(1 - \frac{1}{\sqrt{(L+2\ell)/\ell}}\right)^T \left(f_s(\mathbf{x}_{s-1}) - f_s(\mathbf{x}^*) + \frac{\ell}{2} \left\| \mathbf{x}_{s-1} - \mathbf{x}_s^* \right\|_2^2 \right) \\
\leq&\left(1 - \frac{1}{\sqrt{(L+2\ell)/\ell}}\right)^T \left(\frac{L+2\ell}{2} \left\| \mathbf{x}_{s-1} - \mathbf{x}_s^* \right\|_2^2 + \frac{\ell}{2} \left\| \mathbf{x}_{s-1} - \mathbf{x}_s^* \right\|_2^2 \right).
\end{aligned}$$

that is

$$\left\| \mathbf{x}_s - \mathbf{x}_s^* \right\|_2^2 \leq \left(1 - \frac{1}{\sqrt{(L+2\ell)/\ell}}\right)^T \frac{L+3\ell}{2\ell} \left\| \mathbf{x}_{s-1} - \mathbf{x}_s^* \right\|_2^2 .$$

We require $T = \mathcal{O}(\sqrt{L/\ell} \ln(L/\ell))$ to achieve $\left\| \mathbf{x}_s - \mathbf{x}_s^* \right\|_2^2 \leq \frac{1}{2} \left\| \mathbf{x}_{s-1} - \mathbf{x}_s^* \right\|_2^2$, then the total complexity is

$$TS = \mathcal{O}(\sqrt{L\ell}\epsilon^{-2} \ln(L/\ell)).$$

Since the solution $\mathbf{x}_s^*$ cannot be achieved directly, we should use AGD to achieve

$$\hat{\mathbf{x}} \approx \mathbf{x}_s^* = \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) + \ell \left\| \mathbf{x} - \mathbf{x}_{s-1} \right\|_2^2 .$$

such that

$$\left\| \nabla f(\hat{\mathbf{x}}) - \nabla f(\mathbf{x}_s^*) \right\|_2 \leq \frac{\epsilon}{2} \quad \Longrightarrow \quad \left\| \nabla f(\hat{\mathbf{x}}) \right\|_2 \leq \left\| \nabla f(\hat{\mathbf{x}}) - \nabla f(\mathbf{x}_s^*) \right\|_2 + \nabla f(\mathbf{x}_s^*) \leq \epsilon.$$

Applying AGD with initial point $\mathbf{x}_{s-1}$ and the iteration number T_s, we have

$$\left\| \nabla f(\hat{x}) - \nabla f(\mathbf{x}_s^*) \right\|_2^2 \leq L^2 \left\| \hat{\mathbf{x}} - \mathbf{x}_s^* \right\|_2^2 \leq L^2 \left(1 - \frac{1}{\sqrt{(L+2\ell)/\ell}}\right)_s^T \frac{L+3\ell}{2\ell} \left\| \mathbf{x}_{s-1} - \mathbf{x}_s^* \right\|_2^2 .$$

In the view of result (28), we have

$$\|\mathbf{x}_{s-1} - \mathbf{x}_s^*\|_2^2 \leq \sum_{i=1}^{s} \|\mathbf{x}_{i-1} - \mathbf{x}_i^*\|_2^2 \leq \frac{f(\mathbf{x}_0) - f(\mathbf{x}_s)}{(1-\alpha)^2 \ell} \leq \frac{4(f(\mathbf{x}_0) - f^*)}{\ell}.$$

Hence, we require $T_s = \mathcal{O}(\sqrt{L/\ell}\ln(\ell\epsilon/L))$. The overall complexity is

$$TS + T_s = TS = \mathcal{O}\big(\sqrt{L\ell}\epsilon^{-2}\ln(L/\ell) + \sqrt{L/\ell}\ln(L/(\ell\epsilon)\big).$$

$$\|\nabla f(\mathbf{x}_s^*)\|_2 = \|\nabla f(\mathbf{x}_s^*) - \nabla f(\mathbf{x})\|_2 \leq \|\nabla f(\mathbf{x}_s^*) - \nabla f(\mathbf{x})\|_2$$

6 Nonsmooth Convex Optimization

Theorem 6.1. *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 $(2L, G^2/(2L))$-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

$$\mathrm{prox}_{\gamma f}(\mathbf{x}) = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \gamma f(\mathbf{z}) + \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_2^2.$$

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

$$\nabla \tilde{f}(\mathbf{x}) = \frac{\mathbf{x} - \mathrm{prox}_{\gamma g}(\mathbf{x})}{\gamma}. \tag{29}$$

For any $\mathbf{x} \in \mathbb{R}^d$, the equation (29) is equivalent to

$$\lim_{\mathbf{y}\to\mathbf{x}} \frac{\tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) - \left\langle \mathbf{y} - \mathbf{x}, \frac{1}{\gamma}(\mathbf{x} - \mathrm{prox}_{\gamma g}(\mathbf{x}))\right\rangle}{\|\mathbf{y} - \mathbf{x}\|_2} = 0.$$

Let $\mathbf{u} = \mathrm{prox}_{\gamma g}(\mathbf{x})$ and $\mathbf{v} = \mathrm{prox}_{\gamma g}(\mathbf{y})$. The optimal condition means

$$\frac{1}{\gamma}(\mathbf{x} - \mathbf{u}) \in \partial f(\mathbf{u}) \qquad \text{and} \qquad \frac{1}{\gamma}(\mathbf{y} - \mathbf{v}) \in \partial f(\mathbf{v}). \tag{30}$$

Then

$$\tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x})$$

$$
\begin{aligned}
&= f(\mathbf{v}) + \frac{1}{2\gamma}\|\mathbf{v}-\mathbf{y}\|_2^2 - \left(f(\mathbf{u}) + \frac{1}{2\gamma}\|\mathbf{u}-\mathbf{x}\|_2^2\right) \\
&= \frac{1}{2\gamma}\left(2\gamma(f(\mathbf{v}) - f(\mathbf{u})) + \|\mathbf{v}-\mathbf{y}\|_2^2 - \|\mathbf{u}-\mathbf{x}\|_2^2\right) \\
&\geq \frac{1}{2\gamma}\left(2\langle \mathbf{x}-\mathbf{u}, \mathbf{v}-\mathbf{u}\rangle + \|\mathbf{v}-\mathbf{y}\|_2^2 - \|\mathbf{u}-\mathbf{x}\|_2^2\right) \\
&= \frac{1}{2\gamma}\left(\|\mathbf{v}-\mathbf{y}-(\mathbf{u}-\mathbf{x})\|_2^2 + 2\langle \mathbf{y}-\mathbf{x}, \mathbf{x}-\mathbf{u}\rangle\right) \\
&\geq \frac{1}{\gamma}\langle \mathbf{y}-\mathbf{x}, \mathbf{x}-\mathbf{u}\rangle,
\end{aligned}
$$

where the first inequality use the fact (30) that implies

$$
f(\mathbf{v}) - f(\mathbf{u}) \geq \left\langle \frac{1}{\gamma}(\mathbf{x}-\mathbf{u}), \mathbf{v}-\mathbf{u}\right\rangle
$$

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

$$
\begin{aligned}
0 &\leq \tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) - \frac{1}{\gamma}\langle \mathbf{y}-\mathbf{x}, \mathbf{x}-\mathbf{u}\rangle \\
&\leq \frac{1}{\gamma}\left(\langle \mathbf{y}-\mathbf{x}, \mathbf{y}-\mathbf{v}\rangle - \langle \mathbf{y}-\mathbf{x}, \mathbf{x}-\mathbf{u}\rangle\right) \\
&= \frac{1}{\gamma}\left(\|\mathbf{y}-\mathbf{x}\|_2^2 - \langle \mathbf{y}-\mathbf{x}, \mathbf{v}-\mathbf{u}\rangle\right) \\
&\leq \frac{1}{\gamma}\|\mathbf{y}-\mathbf{x}\|_2^2,
\end{aligned}
$$

where the last step is because of the result (30) 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 (29) and

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

Therefore $\tilde{f}$ is convex and $2L$-smooth.

3. For given $\mathbf{x} \in \mathbb{R}^d$, let

$$
\mathbf{z}^* = \operatorname*{arg\,min}_{\mathbf{z}\in\mathbb{R}^d}\left(f(\mathbf{z}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|_2^2\right).
$$

Hence, we have

$$
\begin{aligned}
\tilde{f}(\mathbf{x}) &= f(\mathbf{z}^*) + \frac{L}{2}\|\mathbf{z}^*-\mathbf{x}\|_2^2 \\
&\geq f(\mathbf{x}) - G\|\mathbf{x}-\mathbf{z}^*\|_2 + \frac{L}{2}\|\mathbf{z}^*-\mathbf{x}\|_2^2 \\
&= f(\mathbf{x}) + \frac{L}{2}\left(\|\mathbf{x}-\mathbf{z}^*\|_2 - \frac{G}{L}\right)^2 - \frac{G^2}{2L} \\
&\geq f(\mathbf{x}) - \frac{G^2}{2L}.
\end{aligned}
$$

$\square$

Example 6.1. *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

$$\tilde{f}(x) = \begin{cases} |x| - \dfrac{\epsilon}{2}, & |x| \geq \epsilon, \\ \dfrac{x^2}{2\epsilon}, & \textit{otherwise.} \end{cases}$$

We define

$$\mathrm{prox}_{\gamma g}(\mathbf{x}) = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \gamma g(\mathbf{z}) + \frac{1}{2}\|\mathbf{z}-\mathbf{x}\|_2^2.$$

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

$$\phi(\mathbf{y}) = f(\mathbf{y}) + g(\mathbf{y}) \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y}-\mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{y}-\mathbf{x}_t\|_2^2 + g(\mathbf{y}),$$

which is

$$\begin{aligned}
&\arg\min_{\mathbf{y}\in\mathbb{R}^d} f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y}-\mathbf{x}_t\|_2^2 + g(\mathbf{y}) \\
=&\arg\min_{\mathbf{y}\in\mathbb{R}^d} \frac{1}{L}\langle \nabla f(\mathbf{x}_t), \mathbf{y}-\mathbf{x}_t\rangle + \frac{1}{2}\|\mathbf{y}-\mathbf{x}_t\|_2^2 + \frac{1}{L}g(\mathbf{y}) \\
=&\arg\min_{\mathbf{y}\in\mathbb{R}^d} \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}_t + \frac{1}{L}\nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{1}{L}g(\mathbf{y}) \\
=&\,\mathrm{prox}_{\eta g}(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t))
\end{aligned}$$

with $\eta = 1/L$.

Example 6.2. *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*

$$f(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{A}\mathbf{x} - \mathbf{b}\right\|_2^2 \quad \textit{and} \quad g(\mathbf{x}) = \lambda\left\|\mathbf{x}\right\|_1,$$

then we obtain the Lasso problem.

3. *Let $h(x) = \lambda|x|$. Consider the proximal operator*

$$\operatorname{prox}_h(x) = \arg\min_{z\in\mathbb{R}} \left(\frac{1}{2}(z-x)^2 + \lambda|z|\right).$$

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

$$z = \begin{cases} x - \lambda, & x \in (\lambda, +\infty), \\ x + \lambda, & x \in (-\infty, -\lambda), \\ 0, & x \in [-\lambda - x, \lambda - z], \end{cases}$$

that is

$$\arg\min_{\mathbf{z}\in\mathbb{R}} \left(\frac{1}{2}(z-x)^2 + \lambda|x|\right) = \operatorname{sign}(x)\max\left\{|x| - \lambda, 0\right\}.$$

Gradient Mapping The proximal gradient iteration can be written as

$$\begin{aligned}\mathbf{x}_{t+1} =& \operatorname{prox}_{\eta g}(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)) \\ =& \mathbf{x}_t - \eta \cdot \frac{\mathbf{x}_t - \operatorname{prox}_{\eta g}(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)))}{\eta} \\ =& \mathbf{x}_t - \eta\mathcal{G}_{\eta g, f}(\mathbf{x}_t).\end{aligned}$$

If $g(\mathbf{x}) = \mathbf{0}$, then $\mathcal{G}_{\eta g, f}(\mathbf{x}) = \nabla f(\mathbf{x})$.

Lemma 6.1. *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

$$\mathbf{x}^+ = \operatorname{prox}_{\eta g}(\mathbf{x} - \eta\nabla f(\mathbf{x})).$$

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* μ_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* ϕ *is* μ_ϕ*-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. Therefore, there exists some subgradient $\boldsymbol{\zeta}^+ \in \partial Q(\mathbf{x}^+)$ such that $\boldsymbol{\zeta}^+ = \mathbf{0}$, which implies
$$Q(\mathbf{x}) - Q(\mathbf{x}^+) \ge \left\langle\mathbf{x} - \mathbf{x}^+, \boldsymbol{\zeta}^+\right\rangle + \frac{\eta^{-1}+\mu_g}{2}\left\|\mathbf{x} - \mathbf{x}^+\right\|_2^2 = \frac{\eta^{-1}+\mu_g}{2}\left\|\mathbf{x} - \mathbf{x}^+\right\|_2^2. \tag{31}$$
From the smoothness of f, we have
$$\begin{aligned}
&\phi(\mathbf{x}^+) = f(\mathbf{x}^+) + g(\mathbf{x}^+)\\
\le& f(\mathbf{x}) + \left\langle\nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x}\right\rangle + \frac{L}{2}\left\|\mathbf{x}^+ - \mathbf{x}\right\|_2^2 + g(\mathbf{x}^+)\\
=& Q(\mathbf{x}^+) + \frac{L - \eta^{-1}}{2}\left\|\mathbf{x}^+ - \mathbf{x}\right\|_2^2\\
\overset{(31)}{\le}& Q(\mathbf{x}) + \frac{L - \mu_g - 2\eta^{-1}}{2}\left\|\mathbf{x}^+ - \mathbf{x}\right\|_2^2\\
=& \phi(\mathbf{x}) + \frac{(L-\mu_g)\eta^2 - 2\eta}{2}\|\mathcal{G}_{\eta g,f}(\mathbf{x})\|_2^2,
\end{aligned}$$
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 $\hat{\mathbf{x}} \in \mathbb{R}^d$, we have
$$\left\langle\nabla f(\mathbf{x}) + \eta^{-1}(\mathbf{x}^+ - \mathbf{x}) + \boldsymbol{\xi}^+, \hat{\mathbf{x}} - \mathbf{x}^+\right\rangle \ge 0.$$
This implies
$$\phi(\hat{\mathbf{x}}) - \phi(\mathbf{x}^+) - \frac{\mu_\phi}{2}\left\|\mathbf{x}^+ - \hat{\mathbf{x}}\right\|_2^2$$

$$
\begin{aligned}
&\geq \left\langle \nabla f(\mathbf{x}^+) + \boldsymbol{\xi}^+, \hat{\mathbf{x}} - \mathbf{x}^+ \right\rangle \\
&= \left\langle \nabla f(\mathbf{x}^+) - \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}^+ \right\rangle + \left\langle \nabla f(\mathbf{x}^+) + \boldsymbol{\xi}^+, \hat{\mathbf{x}} - \mathbf{x}^+ \right\rangle \\
&\geq \left\langle \nabla f(\mathbf{x}^+) - \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}^+ \right\rangle + \eta^{-1} \left\langle \mathbf{x} - \mathbf{x}^+, \hat{\mathbf{x}} - \mathbf{x}^+ \right\rangle \\
&= \left\langle \nabla \tilde{f}(\mathbf{x}^+) - \nabla \tilde{f}(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}^+ \right\rangle \\
&\geq - \|\nabla \tilde{f}(\mathbf{x}^+) - \nabla \tilde{f}(\mathbf{x})\|_2 \left\|\hat{\mathbf{x}} - \mathbf{x}^+\right\|_2 \\
&\geq - \eta^{-1} \left\|\mathbf{x}^+ - \mathbf{x}\right\|_2 \left\|\hat{\mathbf{x}} - \mathbf{x}^+\right\|_2,
\end{aligned}
$$

where

$$
\tilde{f}(\mathbf{z}) = f(\mathbf{z}) - \frac{1}{2\eta} \|\mathbf{z}\|_2^2
$$

and we can show $\tilde{f}(\mathbf{z})$ is η^{-1} smooth because the smoothness of f means

$$
\begin{aligned}
&0 \leq f(\mathbf{u}) - f(\mathbf{v}) - \langle \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \leq \frac{L}{2} \|\mathbf{u} - \mathbf{v}\|_2^2 \\
\Longleftrightarrow & - \frac{\eta^{-1}}{2} \|\mathbf{u} - \mathbf{v}\|_2^2 \leq f(\mathbf{u}) - \frac{\eta^{-1}}{2} \|\mathbf{u}\|_2^2 - \left(f(\mathbf{v}) - \frac{\eta^{-1}}{2} \|\mathbf{v}\|_2^2 \right) - \left\langle \nabla f(\mathbf{v}) - \eta^{-1}\mathbf{v}, \mathbf{u} - \mathbf{v} \right\rangle \leq \frac{L - \eta^{-1}}{2} \|\mathbf{u} - \mathbf{v}\|_2^2 \\
\Longleftrightarrow & - \frac{\eta^{-1}}{2} \|\mathbf{u} - \mathbf{v}\|_2^2 \leq \tilde{f}(\mathbf{u}) - \tilde{f}(\mathbf{v}) - \langle \nabla \tilde{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \leq \frac{L - \eta^{-1}}{2} \|\mathbf{u} - \mathbf{v}\|_2^2.
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
&\phi(\hat{\mathbf{x}}) - \phi(\mathbf{x}^+) \\
\geq &\frac{\mu_\phi}{2} \left\|\mathbf{x}^+ - \hat{\mathbf{x}}\right\|_2^2 - \eta^{-1} \left\|\mathbf{x}^+ - \mathbf{x}\right\|_2 \left\|\hat{\mathbf{x}} - \mathbf{x}^+\right\|_2 \\
\geq &\inf_{\mathbf{z} \in \mathbb{R}^d} \left(\frac{\mu_\phi}{2} \left\|\mathbf{x}^+ - \mathbf{z}\right\|_2^2 - \eta^{-1} \left\|\mathbf{x}^+ - \mathbf{x}\right\|_2 \left\|\mathbf{z} - \mathbf{x}^+\right\|_2 \right) \\
= & - \frac{1}{2\mu_\phi \eta^2} \left\|\mathbf{x}^+ - \mathbf{x}\right\|_2^2 \\
= & - \frac{1}{2\mu_\phi} \|\mathcal{G}_{\eta g, f}(\mathbf{x})\|_2^2.
\end{aligned}
$$

□

Remark 6.1. *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 $f(\mathbf{x}) \leq f(\mathbf{x}^*) + \frac{1}{2\mu} \|\nabla(\mathbf{x})\|_2^2$.*

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

$$\phi(\mathbf{x}_{t+1}) \leq \phi(\mathbf{x}^*) + \frac{1}{2\mu_\phi} \left\| \mathcal{G}_{\eta g, f}(\mathbf{x}_t) \right\|_2^2 \leq \phi(\mathbf{x}^*) + \frac{L}{\mu_\phi}(\phi(\mathbf{x}_t) - \phi(\mathbf{x}_{t+1})),$$

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 $\mathbf{x}^+ = \mathrm{prox}_{\eta g}(\mathbf{x} - \eta \nabla f(\mathbf{x}))$ means

$$\mathbf{x}^+ - (\mathbf{x} - \eta \nabla f(\mathbf{x})) + \eta \boldsymbol{\xi}^+ = \mathbf{0} \quad \Longleftrightarrow \quad \mathcal{G}_{\eta g, f}(\mathbf{x}) = \frac{\mathbf{x} - \mathbf{x}^+}{\eta} = \nabla f(\mathbf{x}) + \boldsymbol{\xi}^+.$$

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

$$\begin{aligned}
&\phi(\mathbf{x}^+) = \phi(\mathbf{x} - \eta \mathcal{G}_{\eta g, f}(\mathbf{x})) \\
=& f(\mathbf{x} - \eta \mathcal{G}_{\eta g, f}(\mathbf{x})) + g(\mathbf{x} - \eta \mathcal{G}_{\eta g, f}(\mathbf{x})) \\
\leq& f(\mathbf{x}) - \eta \left\langle \nabla f(\mathbf{x}), \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\rangle + \frac{L\eta^2}{2} \left\| \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\|_2^2 + g(\mathbf{x} - \eta \mathcal{G}_{\eta g, f}(\mathbf{x})) \\
\leq& f(\mathbf{z}) + \left\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} \right\rangle - \eta \left\langle \nabla f(\mathbf{x}), \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\rangle + \frac{L\eta^2}{2} \left\| \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\|_2^2 + g(\mathbf{z}) - \left\langle \boldsymbol{\xi}^+, \mathbf{z} - (\mathbf{x} - \eta \mathcal{G}_{\eta g, f}(\mathbf{x})) \right\rangle \\
=& \phi(\mathbf{z}) + \left\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} - \eta \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\rangle + \frac{L\eta^2}{2} \left\| \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\|_2^2 - \left\langle \boldsymbol{\xi}^+, \mathbf{z} - (\mathbf{x} - \eta \mathcal{G}_{\eta g, f}(\mathbf{x})) \right\rangle \\
=& \phi(\mathbf{z}) + \left\langle \nabla f(\mathbf{x}) + \boldsymbol{\xi}^+, \mathbf{x} - \mathbf{z} - \eta \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\rangle + \frac{L\eta^2}{2} \left\| \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\|_2^2 \\
=& \phi(\mathbf{z}) + \left\langle \mathcal{G}_{\eta g, f}(\mathbf{x}), \mathbf{x} - \mathbf{z} - \eta \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\rangle + \frac{L\eta^2}{2} \left\| \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\|_2^2 \\
=& \phi(\mathbf{z}) + \left\langle \mathcal{G}_{\eta g, f}(\mathbf{x}), \mathbf{x} - \mathbf{z} \right\rangle - \eta \left(1 - \frac{L\eta}{2}\right) \left\| \mathcal{G}_{\eta g, f}(\mathbf{x}) \right\|_2^2,
\end{aligned} \tag{32}$$

where the first inequality uses smoothness of f; the second inequality uses the convexity of f and g. Applying equation (32) with $\mathbf{x}^+ = \mathbf{x}_{t+1}$, $\mathbf{x} = \mathbf{x}_t$, $\mathbf{z} = \mathbf{x}^*$, and $\eta = 1/L$, we achieve

$$\begin{aligned}
\phi(\mathbf{x}_{t+1}) \leq& \phi(\mathbf{x}^*) + \left\langle \mathcal{G}_{\eta g, f}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \right\rangle - \eta \left(1 - \frac{L\eta}{2}\right) \left\| \mathcal{G}_{\eta g, f}(\mathbf{x}_t) \right\|_2^2 \\
=& \phi(\mathbf{x}^*) + \left\langle \mathcal{G}_{\eta g, f}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \right\rangle - \frac{1}{2L} \left\| \mathcal{G}_{\eta g, f}(\mathbf{x}_t) \right\|_2^2 \\
=& \phi(\mathbf{x}^*) + \frac{L}{2} \left(\left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \left\| \mathbf{x}_t - \frac{1}{L} \mathcal{G}_{\eta g, f}(\mathbf{x}_t) - \mathbf{x}^* \right\|_2^2 \right) \\
=& \phi(\mathbf{x}^*) + \frac{L}{2} \left(\left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \left\| \mathbf{x}_{t+1} - \mathbf{x}^* \right\|_2^2 \right)
\end{aligned}$$

Summing over above inequality with $t = 0, \dots, T-1$, 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 6.1 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 6.1 still holds with $\mu_g = 0$. Let $\eta = 1/L$, then it implies

$$\|\mathcal{G}_{\eta g,f}(\mathbf{x}_t)\|_2^2 \leq 2L(\phi(\mathbf{x}_t) - \phi(\mathbf{x}_{t+1})).$$

Summing over above inequality with $t = 0, \ldots, T-1$, we obtain

$$\begin{aligned} \mathbb{E}\,\|\mathcal{G}_{\eta g,f}(\hat{\mathbf{x}})\|_2^2 &= \frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_{\eta g,f}(\mathbf{x}_t)\|_2^2 \\ \leq \frac{2L(\phi(\mathbf{x}_0) - \phi(\mathbf{x}_T))}{T} &\leq \frac{2L(\phi(\mathbf{x}_0) - \phi^*)}{T}, \end{aligned}$$

where $\hat{\mathbf{x}}$ is uniformly sampled from $\{\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}\}$.

Example 6.3. *Consider the function*

$$f(x) = |x|.$$

The optimal solution is $x = 0$. For any constant learning rate $\eta_t = \eta$, if we take $x_0 = \eta/2$, then

$$x_1 = -\frac{\eta}{2}, \quad x_2 = \frac{\eta}{2}, \quad x_3 = -\frac{\eta}{2} \ldots$$

Therefore the algorithm does not converge with a constant step size.

Example 6.4. *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 6.2. *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

$$\begin{cases} \tilde{\mathbf{x}}_{t+1} = \mathbf{x}_t - \eta_t\mathbf{g}_t, \\ \mathbf{x}_{t+1} = \operatorname{proj}_{\mathcal{C}}(\tilde{\mathbf{x}}_{t+1}) \end{cases}$$

for $t = 0, 1\ldots$ with $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ and

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

satisfies

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

Proof. Given $\hat{\mathbf{x}} \in \mathcal{C}$, we have

$$\begin{aligned}
&\|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \\
=&\|\mathbf{x}_t - \eta_t \mathbf{g}_t - \hat{\mathbf{x}}\|_2^2 \\
=&\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - 2\eta_t \langle \mathbf{g}_t, \mathbf{x}_t - \hat{\mathbf{x}} \rangle + \eta_t^2 \|\mathbf{g}_t\|_2^2 \\
\leq&\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - 2\eta_t \langle \mathbf{g}_t, \mathbf{x}_t - \hat{\mathbf{x}} \rangle + \eta_t^2 G^2 \\
\leq&\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - 2\eta_t (f(\mathbf{x}_t) - f(\hat{\mathbf{x}})) + \eta_t^2 G^2,
\end{aligned}$$

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

$$\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 \leq \|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \leq \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - 2\eta_t (f(\mathbf{x}_t) - f(\hat{\mathbf{x}})) + \eta_t^2 G^2.$$

We sum above inequality over $t = 0, \dots, T-1$ and obtain

$$0 \leq \|\mathbf{x}_T - \hat{\mathbf{x}}\|_2^2 \leq \sum_{t=0}^{T-1} \eta_t^2 G^2 - 2 \sum_{t=0}^{T-1} \eta_t (f(\mathbf{x}_t) - f(\hat{\mathbf{x}})) + \|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2,$$

which implies the desired result. □

Remark 6.2. *We consider two types of stepsizes to understand the convergence rate:*

1. *Taking $\eta_t = \eta_0/\sqrt{T}$ leads to*

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\hat{\mathbf{x}}) \leq \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \eta_0^2 G^2}{2\eta_0\sqrt{T}}.$$

   Hence, the complexity to find ϵ-suboptimal solution requires $\mathcal{O}(1/\varepsilon^2)$ subgradient oracle complexity. If we further suppose the domain $\mathcal{C}$ is bounded by R, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\hat{\mathbf{x}}) \leq \frac{R^2 + \eta_0^2 G^2}{2\eta_0\sqrt{T}}.$$

   Minimizing the upper bound with respect to η_0 leads to $\eta_0 = R/G$.

2. *If we do not know T a prior, we can take $\eta_t = \eta_0/(\sqrt{t+1} + \sqrt{t})$, which leads to*

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta_t = &\sum_{t=0}^{T-1} \frac{\eta_0}{\sqrt{t+1} + \sqrt{t}} \\
=&\eta_0 \sum_{t=0}^{T-1} \left(\sqrt{t+1} - \sqrt{t}\right) \\
=&\eta_0 \sqrt{T}
\end{aligned}$$

   and

$$\begin{aligned}
&\sum_{t=0}^{T-1} \eta_t^2 = \sum_{t=0}^{T-1} \frac{\eta_0^2}{\left(\sqrt{t+1} + \sqrt{t}\right)^2} \\
\leq&\sum_{t=0}^{T-1} \frac{\eta_0^2}{2t+1} = \eta_0^2 + \eta_0^2 \sum_{t=1}^{T-1} \frac{1}{2t+1} \\
\leq&\eta_0^2 + \eta_0^2 \int_0^{T-1} \frac{1}{2x+1}\,\mathrm{d}x
\end{aligned}$$

$$=\eta_0^2 + \frac{\eta_0^2}{2}\ln(2x+1)\Big|_0^{T-1}$$
$$=\eta_0^2 + \frac{\eta_0^2}{2}\ln(2T-1).$$

Let

$$\bar{\mathbf{x}}_T = \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbf{x}_t.$$

Combining above results and Theorem 6.2, we have

$$f(\bar{\mathbf{x}}_T) - f(\hat{\mathbf{x}}) \leq \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t(f(\mathbf{x}_t) - f(\hat{\mathbf{x}})) \leq \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \eta_0^2(\ln(2T-1)+2)G^2/2}{2\eta_0\sqrt{T}}.$$

3. *The recent proposed technique distance over gradient (DoG) apply*

$$\begin{cases} \bar{r}_t = \max\{\bar{r}_{t-1}, \|\mathbf{x}_t - \mathbf{x}_0\|_2\} \\ G_t = G_{t-1} + \|\mathbf{g}_t\|_2^2, \\ \eta_t = \dfrac{\bar{r}_t}{\sqrt{G_t}}, \end{cases}$$

where $G_0 = 0$ and $r_{-1} \in (0, D_{\mathcal{X}}]$. The complexity only contains additional term of $\log(R/r_{-1})$.

Theorem 6.3. *Under the settings of Theorem 6.2, we suppose f is μ-strongly convex and set*

$$\eta_t = \frac{2}{\mu(t+1)}.$$

Then

$$\sum_{t=0}^{T-1}\frac{t}{T(T-1)}f(\mathbf{x}_t) \leq f(\hat{\mathbf{x}}) + \frac{2G^2}{\mu(T-1)}.$$

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

$$\begin{aligned}
&f(\mathbf{x}_t) - f(\hat{\mathbf{x}}) \\
\leq&\langle \mathbf{g}_t, \mathbf{x}_t - \hat{\mathbf{x}}\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{2\eta_t}\left(\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2\right) + \frac{\eta_t G^2}{2} - \frac{\mu}{2}\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\
&= \frac{\mu(t+1)}{4}\left(\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2\right) + \frac{G^2}{\mu(t+1)} - \frac{\mu}{2}\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\
&= \frac{\mu(t-1)}{4}\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \frac{\mu(t+1)}{4}\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 + \frac{G^2}{\mu(t+1)},
\end{aligned}
$$

which implies

$$
\begin{aligned}
t(f(\mathbf{x}_t) - f(\hat{\mathbf{x}})) &\le \frac{\mu(t-1)t}{4}\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \frac{\mu t(t+1)}{4}\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 + \frac{G^2 t}{\mu(t+1)} \\
&\le \frac{\mu(t-1)t}{4}\|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 - \frac{\mu t(t+1)}{4}\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 + \frac{G^2}{\mu}.
\end{aligned}
$$

We sum over above inequality over $t = 0, \dots, T-1$ and obtain

$$
\sum_{t=0}^{T-1} t(f(\mathbf{x}_t) - f(\hat{\mathbf{x}})) \le -\frac{\mu(T-1)T}{4}\|\mathbf{x}_T - \hat{\mathbf{x}}\|_2^2 + \frac{TG^2}{\mu} \le \frac{TG^2}{\mu}
$$

Hence, we have

$$
\sum_{t=0}^{T-1} \frac{t}{T(T-1)} f(\mathbf{x}_t) \le f(\hat{\mathbf{x}}) + \frac{2G^2}{\mu(T-1)}.
$$

□

Remark 6.3. *If the domain is unbounded, Lipschitz continuous function cannot be strongly convex. For mu-strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have*

$$
\begin{aligned}
&f(\mathbf{y}) - f(\mathbf{x}) \\
\ge &\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \\
\ge &- \|\mathbf{g}\|_2 \|\mathbf{y} - \mathbf{x}\|_2 + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \\
= &\|\mathbf{y} - \mathbf{x}\|_2 \left(- \|\mathbf{g}\|_2 + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|_2\right)
\end{aligned}
$$

where $\mathbf{g} \in \partial f(\mathbf{x})$. We fix $\mathbf{x}$ and take $\mathbf{y}$ such that $\|\mathbf{y} - \mathbf{x}\|_2 \to \infty$, then the value of $(f(\mathbf{y}) - f(\mathbf{x}))/\|\mathbf{y} - \mathbf{x}\|_2$ can be arbitrary large.

7 Newton's Method

Recall that optimizing smooth function $f(\mathbf{x})$ by gradient descent

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

is based on minimizing RHS of

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

In a local region, we can minimize the RHS of

$$
f(\mathbf{y}) \approx f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t \rangle + \frac{1}{2}\left\langle \mathbf{y} - \mathbf{x}_t, \nabla^2 f(\mathbf{x}_t)(\mathbf{y} - \mathbf{x}_t) \right\rangle.
$$

Suppose $\nabla^2 f(\mathbf{x}_t)$ is non-singular, then we achieve Newton's method

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

Theorem 7.1. *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 $\|\mathbf{x}_t - \mathbf{x}^*\|_2 \le \mu/(2L_2)$ holds that

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \le \frac{L_2}{\mu} \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 .$$

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

$$\left\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}^*)\right\|_2 \le L_2 \|\mathbf{x} - \mathbf{x}^*\|_2$$

which means

$$\begin{aligned}
&|\lambda_i(\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}^*))| \le L_2 \|\mathbf{x} - \mathbf{x}^*\|_2 \\
\Longleftrightarrow & - L_2 \|\mathbf{x} - \mathbf{x}^*\|_2 \le \lambda_i(\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}^*)) \le L_2 \|\mathbf{x} - \mathbf{x}^*\|_2 \\
\Longleftrightarrow & - L_2 \|\mathbf{x} - \mathbf{x}^*\|_2 \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}^*) \preceq L_2 \|\mathbf{x} - \mathbf{x}^*\|_2 \mathbf{I} \\
\Longleftrightarrow & \nabla^2 f(\mathbf{x}^*) - L_2 \|\mathbf{x} - \mathbf{x}^*\|_2 \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L_2 \|\mathbf{x} - \mathbf{x}^*\|_2 \mathbf{I} + \nabla^2 f(\mathbf{x}^*) \\
\Longrightarrow & \nabla^2 f(\mathbf{x}) \succeq (\mu - L_2 \|\mathbf{x} - \mathbf{x}^*\|_2)\mathbf{I}.
\end{aligned}$$

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 $\|\mathbf{x}_t - \mathbf{x}^*\|_2 \le \mu/(2L_2)$, then we obtain

$$\nabla^2 f(\mathbf{x}_t) \succeq (\mu - L_2 \|\mathbf{x}_t - \mathbf{x}^*\|_2)\mathbf{I} \succeq \frac{\mu}{2}\mathbf{I}$$

and

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \\
=&\left\|(\nabla^2 f(\mathbf{x}_t))^{-1} \left(\int_0^1 \left(\nabla^2 f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}^* + \tau(\mathbf{x}_t - \mathbf{x}^*))\right)(\mathbf{x}_t - \mathbf{x}^*)\, \mathrm{d}\tau\right)\right\|_2 \\
\le&\left\|(\nabla^2 f(\mathbf{x}_t))^{-1}\right\|_2 \int_0^1 \left\|\nabla^2 f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}^* + \tau(\mathbf{x}_t - \mathbf{x}^*))\right\|_2 \|\mathbf{x}_t - \mathbf{x}^*\|_2 \, \mathrm{d}\tau \\
\le&\frac{2}{\mu} \int_0^1 L_2(1-\tau) \|\mathbf{x}_t - \mathbf{x}^*\|_2 \|\mathbf{x}_t - \mathbf{x}^*\|_2 \, \mathrm{d}\tau \\
=&\frac{L_2}{\mu} \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \\
\le&\frac{1}{2} \|\mathbf{x}_t - \mathbf{x}^*\|_2 .
\end{aligned}$$

Hence, the quadratic convergence holds if $\|\mathbf{x}_0 - \mathbf{x}^*\|_2 \le \mu/(2L_2)$. □

Remark 7.1. *The quadratic convergence means*

$$\frac{L_2}{\mu} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \le \left(\frac{L_2}{\mu} \|\mathbf{x}_t - \mathbf{x}^*\|_2\right)^2 \quad \Longrightarrow \quad \frac{L_2}{\mu} \|\mathbf{x}_T - \mathbf{x}^*\|_2 \le \left(\frac{L_2}{\mu} \|\mathbf{x}_0 - \mathbf{x}^*\|_2\right)^{2^T}.$$

In the local region, Newton's method requires $T = \mathcal{O}(\ln\ln(1/\epsilon))$ iterations to achieve $\|\mathbf{x}_T - \mathbf{x}^*\|_2$. 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

$$\begin{cases}\tilde{\mathbf{x}}_{t+1}=\mathbf{x}_t-(\nabla f(\mathbf{x}_t))^{-1}\nabla f(\mathbf{x}_t),\\ \mathbf{x}_{t+1}=\mathrm{proj}_{\mathcal{C}}(\tilde{\mathbf{x}}_{t+1}),\end{cases}\tag{33}$$

which is not reasonable because of Newton's methods do not depends on Euclidean norm. The correct update should be

$$\begin{aligned}\mathbf{x}_{t+1}=&\arg\min_{\mathbf{x}\in\mathcal{C}}\left(f(\mathbf{x}_t)+\langle\nabla f(\mathbf{x}_t),\mathbf{x}-\mathbf{x}_t\rangle+\frac{1}{2}\left\langle\nabla^2 f(\mathbf{x}_t)(\mathbf{x}-\mathbf{x}_t),\mathbf{x}-\mathbf{x}_t\right\rangle\right)\\ =&\arg\min_{\mathbf{x}\in\mathcal{C}}\frac{1}{2}\left\|\mathbf{x}-\left(\mathbf{x}_t-(\nabla^2 f(\mathbf{x}_t))^{-1}\nabla f(\mathbf{x}_t)\right)\right\|^2_{\nabla^2 f(\mathbf{x}_t)},\end{aligned}$$

which is the projection with respect to $\nabla^2 f(\mathbf{x}_t)$-norm. The proximal Newton methods is similar.

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 $\mathbf{x}_t=\mathbf{A}\mathbf{y}_t$ (or $\mathbf{y}_t=\mathbf{A}^{-1}\mathbf{x}_t$), 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

$$\begin{aligned}\mathbf{y}_{t+1}=&\mathbf{y}_t-\eta\nabla\phi(\mathbf{y}_t)\\ =&\mathbf{y}_t-\eta\mathbf{A}^\top\nabla f(\mathbf{A}\mathbf{y}_t)\\ =&\mathbf{A}^{-1}\left(\mathbf{A}\mathbf{y}_t-\eta\mathbf{A}\mathbf{A}^\top\nabla f(\mathbf{A}\mathbf{y}_t)\right)\\ =&\mathbf{A}^{-1}\left(\mathbf{x}_t-\eta\mathbf{A}\mathbf{A}^\top\nabla f(\mathbf{x}_t)\right)\neq\mathbf{A}^{-1}\mathbf{x}_{t+1}.\end{aligned}$$

Definition 7.1. *We say $f:\mathbb{R}^d\to\mathbb{R}$ is M-strongly self-concordant, if it is twice differentiable and holds*

$$\nabla^2 f(\mathbf{x})-\nabla^2 f(\mathbf{y})\preceq M\|\mathbf{x}-\mathbf{y}\|_{\nabla^2 f(\mathbf{z})}\nabla^2 f(\mathbf{w}),$$

for any $\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}\in\mathbb{R}^d$ and some $M>0$.

Remark 7.2. *If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and has L_2-Lipschitz continuous Hessian, then it is M-strongly self-concordant with $M = L_2/\mu^{3/2}$.*

Lemma 7.1. *If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-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 $\left\|\cdot\right\|_{\mathbf{z}}$ is the weighted norm with respect to $\nabla^2 f(\mathbf{z})$. □

Damped Newton method The damped Newton method is based on

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{1 + M_f \lambda_f(\mathbf{x}_t)} \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t),$$

where $M_f = M/2$ and

$$\lambda_f(\mathbf{x}_t) = \sqrt{\left\langle \nabla f(\mathbf{x}_t), \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t) \right\rangle}.$$

Without loss of generality, we assume $M_f = 1$ and consider

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{1 + \lambda_f(\mathbf{x}_t)} \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t),$$

where

$$\lambda_f(\mathbf{x}_t) = \sqrt{\left\langle \nabla f(\mathbf{x}_t), \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t) \right\rangle}.$$

Then we have

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\lambda_f(\mathbf{x}_t) + \ln\left(1 + \lambda_f(\mathbf{x}_t)\right).$$

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 $t = 0, \dots, t_0$, 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 $\lambda_f(\mathbf{x}_t) < 1/4$, We have

   $$\lambda_f(\mathbf{x}_{t+1}) \leq 2(\lambda_f(\mathbf{x}_t))^2.$$

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 $\lambda_f(\mathbf{x}_t) \leq \epsilon$. Please see Nesterov's book for detailed proofs.

8 Quasi-Newton Methods

Secant Condition For general $f(\mathbf{x})$ with L_2-Lipschitz continuous Hessian, we have

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

where the last term is a high-order term as follows

$$
\begin{aligned}
&\left\|\int_0^1 \nabla^2 (f(\mathbf{x}_t + \tau(\mathbf{x}_{t+1} - \mathbf{x}_t)) - \nabla^2 f(\mathbf{x}_{t+1}))(\mathbf{x}_{t+1} - \mathbf{x}_t)\,\mathrm{d}\tau\right\|_2 \\
\leq &\int_0^1 \left\|\nabla^2 (f(\mathbf{x}_t + \tau(\mathbf{x}_{t+1} - \mathbf{x}_t)) - \nabla^2 f(\mathbf{x}_{t+1}))\right\|_2 \left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2 \mathrm{d}\tau \\
\leq &\int_0^1 L_2(1-\tau)\left\|\mathbf{x}_t - \mathbf{x}_{t+1})\right\|_2 \left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2 \mathrm{d}\tau = \frac{L_2}{2}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2.
\end{aligned}
$$

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

$$
x_{t+1} = x_t - (g'(x_t))^{-1} g(x_t), \qquad x_{t+2} = x_{t+1} - (g'(x_{t+1}))^{-1} g(x_{t+1}), \qquad \cdots\cdots
$$

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

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

SR1 method We consider secant condition and rank-1 update (only $\mathbf{G}_{t+1}$ and $\mathbf{z}_t\mathbf{z}_t^\top$ are unknown)

$$
\mathbf{y}_t = \mathbf{G}_{t+1}\mathbf{s}_t \tag{34}
$$

and

$$
\mathbf{G}_{t+1} = \mathbf{G}_t + \mathbf{z}_t\mathbf{z}_t^\top. \tag{35}
$$

where

$$
\mathbf{y}_t = \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) \qquad \text{and} \qquad \mathbf{s}_t = \mathbf{x}_{t+1} - \mathbf{x}_t.
$$

Combining above equalities implies

$$
\begin{aligned}
&\mathbf{y}_t = (\mathbf{G}_t + \mathbf{z}_t\mathbf{z}_t^\top)\mathbf{s}_t && (36)\\
\Longrightarrow &\mathbf{y}_t = \mathbf{G}_t\mathbf{s}_t + (\mathbf{z}_t^\top\mathbf{s}_t)\mathbf{z}_t && (37)\\
\Longrightarrow &(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top = (\mathbf{z}_t^\top\mathbf{s}_t)^2\mathbf{z}_t\mathbf{z}_t^\top. && (38)
\end{aligned}
$$

Left multiplying $\mathbf{s}_t^\top$ on (36) leads to

$$
\mathbf{s}_t^\top\mathbf{y}_t = \mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t + (\mathbf{z}_t^\top\mathbf{s}_t)^2 \quad \Longrightarrow \quad (\mathbf{z}_t^\top\mathbf{s}_t)^2 = \mathbf{s}_t^\top\mathbf{y}_t - \mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t. \tag{39}
$$

Combining (35), (38) and (39), 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 of Hessian estimator is

$$\mathbf{G}_{t+1} = \mathbf{G}_t - \frac{\mathbf{G}_t\mathbf{s}_t\mathbf{s}_t^\top\mathbf{G}_t}{\mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t} + \frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}.$$

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.

Theorem 8.1. *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

$$\min_{\hat{\mathbf{H}} \in \mathbb{R}^{d\times d}} \left\|\hat{\mathbf{H}} - \hat{\mathbf{H}}_t\right\|_F \qquad \text{s.t} \quad \hat{\mathbf{H}} = \hat{\mathbf{H}}^\top, \quad \hat{\mathbf{H}}\hat{\mathbf{y}}_t = \hat{\mathbf{y}}_t$$

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 $\mathbf{u} = \hat{\mathbf{y}}_t / \|\hat{\mathbf{y}}_t\|_2 \in \mathbb{R}^d$ ($\hat{\mathbf{H}}\mathbf{u} = \mathbf{u}$ 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

$$\mathbf{U}^\top\hat{\mathbf{H}}\mathbf{U} = \begin{bmatrix}\mathbf{u}^\top\hat{\mathbf{H}}\mathbf{u} & \mathbf{u}^\top\hat{\mathbf{H}}\mathbf{U} \\ \mathbf{U}^\top\hat{\mathbf{H}}\mathbf{u} & \mathbf{U}^\top\hat{\mathbf{H}}\mathbf{U}\end{bmatrix} = \begin{bmatrix}1 & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_\perp^\top\hat{\mathbf{H}}\mathbf{U}_\perp\end{bmatrix}.$$

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 $\mathbf{u} = \hat{\mathbf{y}}_t / \|\hat{\mathbf{y}}_t\|_2$ 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 (check the constraints by using $\mathbf{u} = \hat{\mathbf{y}}_t / \|\hat{\mathbf{y}}_t\|_2$)

$$
\begin{aligned}
\hat{\mathbf{H}}\hat{\mathbf{y}}_t &= \begin{bmatrix} \mathbf{u} & \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{u}^\top \\ \mathbf{U}_\perp^\top \end{bmatrix} \hat{\mathbf{y}}_t \\
&= \begin{bmatrix} \mathbf{u} & \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_\perp^\top \hat{\mathbf{H}}_t \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{u}^\top \hat{\mathbf{y}}_t \\ \mathbf{0} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{u} & \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{u}^\top \hat{\mathbf{y}}_t \\ \mathbf{0} \end{bmatrix} \\
&= \mathbf{u}(\mathbf{u}^\top \hat{\mathbf{y}}_t) = \hat{\mathbf{y}}_t.
\end{aligned}
$$

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 (we use $\hat{\mathbf{y}}_t = \bar{\mathbf{G}}^{-1/2}\mathbf{y}_t$ and $\mathbf{y}_t = \bar{\mathbf{G}}\mathbf{s}_t$)

$$
\begin{aligned}
\bar{\mathbf{G}}^{-1/2}\mathbf{u} &= \frac{\bar{\mathbf{G}}^{-1/2}\hat{\mathbf{y}}_t}{\|\hat{\mathbf{y}}_t\|_2} = \frac{\bar{\mathbf{G}}^{-1}\mathbf{y}_t}{\left\|\bar{\mathbf{G}}^{-1/2}\mathbf{y}_t\right\|_2} \\
&= \frac{\left(\int_0^1 \nabla^2 f(\mathbf{x}_t + \tau(\mathbf{x}_{t+1} - \mathbf{x}_t))\,\mathrm{d}\tau\right)^{-1}\mathbf{y}_t}{(\mathbf{y}_t^\top\bar{\mathbf{G}}_t^{-1}\mathbf{y}_t)^{1/2}} = \frac{\mathbf{s}_t}{(\mathbf{y}_t^\top\mathbf{s}_t)^{1/2}}
\end{aligned}
$$

and (we use $\mathbf{u} = \hat{\mathbf{y}}_t / \|\hat{\mathbf{y}}_t\|_2$ and $\|\hat{\mathbf{y}}_t\|_2^2 = \mathbf{y}_t^\top\bar{\mathbf{G}}_t^{-1}\mathbf{y}_t = \mathbf{y}_t^\top\mathbf{s}_t$)

$$
\bar{\mathbf{G}}^{1/2}\mathbf{u} = \frac{\bar{\mathbf{G}}^{1/2}\hat{\mathbf{y}}_t}{\|\hat{\mathbf{y}}_t\|_2} = \frac{\bar{\mathbf{G}}^{1/2}\hat{\mathbf{y}}_t}{(\mathbf{y}_t^\top\mathbf{s}_t)^{1/2}} = \frac{\mathbf{y}_t}{(\mathbf{y}_t^\top\mathbf{s}_t)^{1/2}},
$$

we obtain

$$
\mathbf{H} = \frac{\mathbf{s}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t} + \left(\mathbf{I} - \frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{H}_t\left(\mathbf{I} - \frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right).
$$

□

Remark 8.1. *BFGS and DFP is always well-defined for strongly-convex objective, while the denominator in SR1 update can vanish.*

DFP Method Let $\mathbf{G}_{t+1}$ be the solution of following matrix optimization problem

$$
\min_{\mathbf{G}\in\mathbb{R}^{d\times d}} \|\mathbf{G} - \mathbf{G}_t\|_{\bar{\mathbf{G}}_t^{-1}}
$$

$$\text{s.t} \quad \mathbf{G} = \mathbf{G}^\top, \quad \mathbf{G}\mathbf{s}_t = \mathbf{y}_t,$$

where the weighted norm $\|\cdot\|_{\bar{\mathbf{G}}_t}$ is defined as

$$\|\mathbf{A}\|_{\bar{\mathbf{G}}_t} = \left\|\bar{\mathbf{G}}_t^{-1/2}\mathbf{A}\bar{\mathbf{G}}_t^{-1/2}\right\|_F \quad \text{with} \quad \bar{\mathbf{G}}_t = \int_0^1 \nabla^2 f(\mathbf{x}_t + \tau(\mathbf{x}_{t+1} - \mathbf{x}_t))\,\mathrm{d}\tau.$$

It implies DFP update

$$\mathbf{G}_{t+1} = \left(\mathbf{I} - \frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{G}_t\left(\mathbf{I} - \frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right) + \frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}.$$

The corresponding update to inverse of Hessian estimator is

$$\mathbf{G}_{t+1}^{-1} = \mathbf{G}_t^{-1} - \frac{\mathbf{G}_t^{-1}\mathbf{y}_t\mathbf{y}_t^\top\mathbf{G}_t^{-1}}{\mathbf{y}_t^\top\mathbf{G}_t^{-1}\mathbf{y}_t} + \frac{\mathbf{s}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}.$$

Remark 8.2. *The superlinear convergence of classical quasi-Newton methods have been established in 1970's, while the convergence rates are established until 2020's. The BFGS/DFP has the rates of* $\mathcal{O}\big((d\kappa/t)^{t/2}\big)$ *and the rate of SR1 is* $\mathcal{O}\big((d\ln(\kappa)/t)^{t/2}\big)$

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 $\tau \in [0, 1]$.
The 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 $\mathbf{G}\mathbf{u} = \mathbf{G}_t\mathbf{s}_t$ and

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

If $\tau = 0$, then the update $\mathbf{G}_{t+1} = \mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u})$ corresponds to SR1 update since

$$\begin{aligned}\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u}) =& \mathbf{G} - \frac{(\mathbf{G}-\mathbf{A})\mathbf{u}\mathbf{u}^\top(\mathbf{G}-\mathbf{A})}{\mathbf{u}^\top(\mathbf{G}-\mathbf{A})\mathbf{u}} \\ =& \mathbf{G}_t - \frac{(\mathbf{G}_t\mathbf{s}_t - \mathbf{y}_t)(\mathbf{G}_t\mathbf{s}_t - \mathbf{y}_t)^\top}{\mathbf{s}_t^\top(\mathbf{G}_t\mathbf{s}_t - \mathbf{y}_t)}.\end{aligned}$$

If $\tau = 1$, then the update $\mathbf{G}_{t+1} = \mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u})$ corresponds to DFP update since

$$\begin{aligned}\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{A},\mathbf{u}) =& \mathbf{G} - \frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G} + \mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} + \left(\frac{\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} + 1\right)\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} \\ =& \mathbf{G}_t - \frac{\mathbf{y}_t\mathbf{s}_t^\top\mathbf{G}_t + \mathbf{G}_t\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t} + \left(\frac{\mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t}{\mathbf{s}_t^\top\mathbf{y}_t} + 1\right)\frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t} \\ =& \mathbf{G}_t - \frac{\mathbf{y}_t\mathbf{s}_t^\top\mathbf{G}_t + \mathbf{G}_t\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t} + \frac{\mathbf{y}_t\mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t\mathbf{y}_t}{\mathbf{s}_t^\top\mathbf{y}_t} + \frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}\end{aligned}$$

$$=\left(\mathbf{I}-\frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}\right)\mathbf{G}_t\left(\mathbf{I}-\frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}\right)+\frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{s}_t^\top\mathbf{y}_t}.$$

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 $\frac{1}{n}\sum_{i=1}^n\mathbf{a}_i\mathbf{a}_i^\top$.

9 Minimax Optimization

We consider the minimax problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^d} f(\mathbf{x},\mathbf{y}), \tag{40}$$

where $f(\mathbf{x},\mathbf{y})$ is L-smooth, μ-strongly-convex in $\mathbf{x}$ and μ-strongly-concave in $\mathbf{y}$, which implies

$$\begin{aligned} &\|g(\mathbf{z}_1) - g(\mathbf{z}_2)\|_2 \le L\,\|\mathbf{z} - \mathbf{z}'\|_2\,, \\ &\langle g(\mathbf{z}_1) - g(\mathbf{z}_2), \mathbf{z}_1 - \mathbf{z}_2\rangle \ge \mu\,\|\mathbf{z}_1 - \mathbf{z}_2\|_2^2\,. \end{aligned} \tag{41}$$

where $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{y}_1)$, $\mathbf{z}_2 = (\mathbf{x}_2, y_2)$, and $\mathbf{g}(\mathbf{z}) = (\nabla_x f(\mathbf{x},\mathbf{y}), -\nabla_y f(\mathbf{x},\mathbf{y}))$.

Remark 9.1. *The equation (41) is called the monotone property. We can prove it by consider the convexity and concavity. For the convexity on $\mathbf{x}$, we have*

$$f(\mathbf{x}_2,\mathbf{y}_1) \ge f(\mathbf{x}_1,\mathbf{y}_1) + \langle \nabla_x f(\mathbf{x}_1,\mathbf{y}_1), \mathbf{x}_2 - \mathbf{x}_1\rangle + \frac{\mu}{2}\|\mathbf{x}_2 - \mathbf{x}_1\|^2, \tag{42}$$

$$f(\mathbf{x}_1,\mathbf{y}_2) \ge f(\mathbf{x}_2,\mathbf{y}_2) + \langle \nabla_x f(\mathbf{x}_2,\mathbf{y}_2), \mathbf{x}_1 - \mathbf{x}_2\rangle + \frac{\mu}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2. \tag{43}$$

Similarly, the μ-strongly-concavity with respect to the second variable y means

$$-f(\mathbf{x}_1,\mathbf{y}_2) \ge -f(\mathbf{x}_1,\mathbf{y}_1) + \langle -\nabla_y f(\mathbf{x}_1,\mathbf{y}_1), \mathbf{y}_2 - \mathbf{y}_1\rangle + \frac{\mu}{2}\|\mathbf{y}_2 - \mathbf{y}_1\|^2, \tag{44}$$

$$-f(\mathbf{x}_2,\mathbf{y}_1) \ge -f(\mathbf{x}_2,\mathbf{y}_2) + \langle -\nabla_y f(\mathbf{x}_2,\mathbf{y}_2), \mathbf{y}_1 - \mathbf{y}_2\rangle + \frac{\mu}{2}\|\mathbf{y}_1 - \mathbf{y}_2\|^2. \tag{45}$$

Sum all above inequalities equation (42), equation (43), equation (44) and equation (45), we have

$$\begin{aligned} 0 \ge &\langle \nabla_x f(\mathbf{x}_1,\mathbf{y}_1) - \nabla_x f(\mathbf{x}_2,\mathbf{y}_2), \mathbf{x}_2 - \mathbf{x}_1\rangle - \langle \nabla_y f(\mathbf{x}_1,\mathbf{y}_1) - \nabla_y f(\mathbf{x}_2,\mathbf{y}_2), \mathbf{y}_2 - \mathbf{y}_1\rangle \\ &+ \mu\|\mathbf{x}_2 - \mathbf{x}_1\|^2 + \mu\|\mathbf{y}_2 - \mathbf{y}_1\|^2, \end{aligned}$$

which is equivalent to the desired result.

We let $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$ be the solution of problem (40) such that for all $\mathbf{x}, \mathbf{y}$ holds that

$$f(\mathbf{x}^*,\mathbf{y}) \le f(\mathbf{x}^*,\mathbf{y}^*) \le f(x, y^*).$$

Gradient Descent Ascent We study the extragradient method as follows

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta\mathbf{g}(\mathbf{z}_t).$$

We have

$$\begin{aligned} &\|\mathbf{z}_{t+1} - \mathbf{z}^*\|_2^2 = \|\mathbf{z}_t - \eta\mathbf{g}(\mathbf{z}_t) - \mathbf{z}^*\|_2^2 \\ =&\|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - \eta\,\langle \mathbf{g}(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}^*\rangle + \eta^2\,\|\mathbf{g}(\mathbf{z}_t)\|_2^2 \\ \le&\|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - \eta\mu\,\|\mathbf{z}_t - \mathbf{z}^*\|_2^2 + \eta^2 L^2\,\|\mathbf{z}_t - \mathbf{z}^*\|_2^2 \\ =&(1 - \eta\mu + \eta^2 L^2)\,\|\mathbf{z}_t - \mathbf{z}^*\|_2^2\,. \end{aligned}$$

Taking $\eta = \mu/(2L^2)$, we have

$$\|\mathbf{z}_{t+1} - \mathbf{z}^*\|_2^2 \le \left(1 - \frac{\mu^2}{4L^2}\right)\|\mathbf{z}_t - \mathbf{z}^*\|_2^2\,.$$

Extragradient We study the extragradient method as follows

$$\begin{cases} \mathbf{z}_{t+1/2} = \mathbf{z}_t - \eta \mathbf{g}(\mathbf{z}_t), \\ \mathbf{z}_{t+1} = \mathbf{z}_t - \eta \mathbf{g}(\mathbf{z}_{t+1/2}). \end{cases}$$

where $\eta = \mathcal{O}(1/L)$ is the stepsize.
Consider the basic equality

$$2\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a} + \mathbf{b}\|_2^2 - \|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2,$$

which means

$$2\langle \mathbf{z}_t - \mathbf{z}_{t+1}, \mathbf{z}_{t+1} - \mathbf{z}^* \rangle = \|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - \|\mathbf{z}_t - \mathbf{z}_{t+1}\|_2^2 - \|\mathbf{z}_{t+1} - \mathbf{z}^*\|_2^2, \tag{46}$$

$$2\left\langle \mathbf{z}_t - \mathbf{z}_{t+1/2}, \mathbf{z}_{t+1/2} - \mathbf{z}_{t+1} \right\rangle = \|\mathbf{z}_t - \mathbf{z}_{t+1}\|_2^2 - \left\|\mathbf{z}_t - \mathbf{z}_{t+1/2}\right\|_2^2 - \left\|\mathbf{z}_{t+1/2} - \mathbf{z}_{t+1}\right\|_2^2. \tag{47}$$

Hence, we have

$$\begin{aligned} & 2\eta \left\langle g(\mathbf{z}_{t+1/2}), \mathbf{z}_{t+1/2} - \mathbf{z}^* \right\rangle \\ =& 2\eta \left\langle g(\mathbf{z}_{t+1/2}), \mathbf{z}_{t+1} - \mathbf{z}^* \right\rangle + 2\eta \left\langle g(\mathbf{z}_{t+1/2}), \mathbf{z}_{t+1/2} - \mathbf{z}_{t+1} \right\rangle \\ =& 2\langle \mathbf{z}_t - \mathbf{z}_{t+1}, \mathbf{z}_{t+1} - \mathbf{z}^* \rangle + 2\left\langle \mathbf{z}_t - \mathbf{z}_{t+1/2}, \mathbf{z}_{t+1/2} - \mathbf{z}_{t+1} \right\rangle + 2\left\langle \mathbf{z}_{t+1/2} - \mathbf{z}_{t+1}, \mathbf{z}_{t+1/2} - \mathbf{z}_{t+1} \right\rangle \\ =& \|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - \|\mathbf{z}_t - \mathbf{z}_{t+1}\|_2^2 - \|\mathbf{z}_{t+1} - \mathbf{z}^*\|_2^2 + \|\mathbf{z}_t - \mathbf{z}_{t+1}\|_2^2 - \left\|\mathbf{z}_t - \mathbf{z}_{t+1/2}\right\|_2^2 - \left\|\mathbf{z}_{t+1/2} - \mathbf{z}_{t+1}\right\|_2^2 \\ & + 2\eta \left\langle g(\mathbf{z}_{t+1/2}) - g(\mathbf{z}_t), \mathbf{z}_{t+1/2} - \mathbf{z}_{t+1} \right\rangle \\ \le& \|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - \|\mathbf{z}_{t+1} - \mathbf{z}^*\|_2^2 - \left\|\mathbf{z}_t - \mathbf{z}_{t+1/2}\right\|_2^2 - \left\|\mathbf{z}_{t+1/2} - \mathbf{z}_{t+1}\right\|_2^2 \\ & + 4\eta^2 \left\|g(\mathbf{z}_{t+1/2}) - g(\mathbf{z}_t)\right\|_2^2 + \left\|\mathbf{z}_{t+1/2} - \mathbf{z}_{t+1}\right\|_2^2 \\ \le& \|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - \|\mathbf{z}_{t+1} - \mathbf{z}^*\|_2^2 - (1 - 4\eta^2 L^2)\left\|\mathbf{z}_{t+1/2} - \mathbf{z}_t\right\|_2^2 \end{aligned} \tag{48}$$

where the second equality is based on the update rule; the third one is based on (46), (47) and the update rule; the first inequality is due to $2\langle a, b \rangle \le \|a\|_2^2 + \|b\|_2^2$; the last step use the smoothness of f. We also have

$$2\eta \left\langle g(\mathbf{z}_{t+1/2}), \mathbf{z}_{t+1/2} - \mathbf{z}^* \right\rangle \ge 2\eta\mu \left\|\mathbf{z}_{t+1/2} - \mathbf{z}^*\right\|_2^2 \ge \eta\mu \|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - 2\eta\mu \left\|\mathbf{z}_{kt} - \mathbf{z}_{t+1/2}\right\|_2^2 \tag{49}$$

Connecting inequalities (48) and (49), we have

$$\begin{aligned} & \eta\mu \|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - 2\eta\mu \left\|\mathbf{z}_t - \mathbf{z}_{t+1/2}\right\|_2^2 \\ \le& 2\eta \left\langle g(\mathbf{z}_{t+1/2}), \mathbf{z}_{t+1/2} - \mathbf{z}^* \right\rangle \\ \le& \|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - \|\mathbf{z}_{t+1} - \mathbf{z}^*\|_2^2 - (1 - 4\eta^2 L^2)\left\|\mathbf{z}_{t+1/2} - \mathbf{z}_t\right\|_2^2 \end{aligned}$$

which means

$$\|\mathbf{z}_{t+1} - \mathbf{z}^*\|_2^2 \le (1 - \eta\mu)\|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - (1 - 2\eta\mu - 4\eta^2 L^2)\left\|\mathbf{z}_{t+1/2} - \mathbf{z}_t\right\|_2^2$$

Let $\eta = 1/(4L)$ and $\kappa = L/\mu$, then

$$1 - 2\eta\mu - 4\eta^2 L^2 \ge 1 - 2\eta L - 4\eta^2 L^2 = 1 - \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

and

$$\|\mathbf{z}_{t+1} - \mathbf{z}^*\|_2^2 \le \left(1 - \frac{\mu}{4L}\right)\|\mathbf{z}_t - \mathbf{z}^*\|_2^2 = \left(1 - \frac{1}{4\kappa}\right)\|\mathbf{z}_t - \mathbf{z}^*\|_2^2.$$

Hence, we needs $t = \mathcal{O}(\kappa \log(1/\varepsilon))$ number of iterations to obtain $\|\mathbf{z}_t - \mathbf{z}^*\|_2^2 \le \varepsilon$.

10 Stochastic Gradient Descent

Theorem 10.1. *Consider the stochastic problem*

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) \triangleq \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{D}}[F(\mathbf{x};\xi)],$$

where $\mathcal{C}$ is convex and compact each $F(\mathbf{x};\xi)$ is convex and G-Lipschitz such that $\|\mathbf{g}\|_2 \leq G$ for any $\mathbf{g} \in \partial F(\mathbf{x};\xi)$. The update

$$\begin{cases} \textit{draw} \ \ \xi \sim \mathcal{D}, \\ \mathbf{g}_t \in \partial F(\mathbf{x}_t;\xi), \\ \tilde{\mathbf{x}}_{t+1} = \mathbf{x} - \eta_t \mathbf{g}_t, \\ \mathbf{x}_{t+1} = \mathrm{proj}_{\mathcal{C}}(\tilde{\mathbf{x}}_{t+1}), \end{cases}$$

holds that

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] \leq f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \sum_{t=0}^{T-1} G^2\eta_t^2}{2\sum_{t=0}^{T-1}\eta_t}.$$

Proof. Conditioned on $\xi_0, \ldots, \xi_{t-1}$, we have

$$\begin{aligned} & \mathbb{E}_{\xi_t} \|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \\ =& \mathbb{E}_{\xi_t} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_t + \mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\ =& \mathbb{E}_{\xi_t} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|_2^2 + 2\mathbb{E}_{\xi_t} \langle \tilde{\mathbf{x}}_{t+1} - \mathbf{x}_t, \mathbf{x}_t - \hat{\mathbf{x}} \rangle + \mathbb{E}_{\xi_t} \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\ =& \eta_t^2 \mathbb{E}_{\xi_t} \|\mathbf{g}_t\|_2^2 - 2\eta_t \mathbb{E}_{\xi_t} \langle \mathbf{g}_t, \mathbf{x}_t - \hat{\mathbf{x}} \rangle + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\ \leq& \eta_t^2 G^2 - 2\eta_t \langle \tilde{\mathbf{g}}_t, \mathbf{x}_t - \hat{\mathbf{x}} \rangle + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \\ \leq& \eta_t^2 G^2 + 2\eta_t (f(\hat{\mathbf{x}}) - f(\mathbf{x}_t)) + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2 \end{aligned}$$

for all $\hat{\mathbf{x}} \in \mathcal{C}$, where the first inequality is based on the bounded subgradient assumption and the second one use the definition of subgradient. Here we let

$$\tilde{\mathbf{g}}_t = \mathbb{E}_{\xi_t}[\mathbf{g}_t] \in \partial f(\mathbf{x}_t),$$

and the last step is because of taking expectation on the inequality

$$F(\mathbf{y};\xi) \geq F(\mathbf{x};\xi) + \langle \mathbf{g}_t, \mathbf{y} - \mathbf{x} \rangle.$$

Using Theorem 3.3, we obtain

$$\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}\|_2^2 \leq \|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}\|_2^2 \leq \eta_t^2 G^2 + 2\eta_t (f(\hat{\mathbf{x}}) - f(\mathbf{x}_t)) + \|\mathbf{x}_t - \hat{\mathbf{x}}\|_2^2.$$

We sum above inequality over $t = 0, \ldots, T-1$ and taking expectation with all the history, then

$$0 \leq \mathbb{E}\|\mathbf{x}_T - \hat{\mathbf{x}}\|_2^2 \leq \sum_{t=0}^{T-1} \eta_t^2 G^2 + 2\sum_{t=0}^{T-1} \eta_t \mathbb{E}[f(\hat{\mathbf{x}}) - f(\mathbf{x}_t)] + \|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2,$$

which implies

$$\mathbb{E}[f(\bar{\mathbf{x}}_T) - f(\hat{\mathbf{x}})] = \frac{\sum_{t=0}^{T-1} \eta_t (f(\mathbf{x}_t) - f(\hat{\mathbf{x}}))}{\sum_{t=0}^{T-1} \eta_t} \leq \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \sum_{t=0}^{T-1} G^2 \eta_t^2}{2\sum_{t=0}^{T-1} \eta_t}.$$

□

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.*

Analysis for Mini-Batch SGD (Smooth and Convex) Suppose each component $F(\cdot, \xi)$ is L-smooth and convex and let $\mathcal{C} = \mathbb{R}^d$. We denote

$$F(\mathbf{x}_t; \mathcal{S}_t) = \frac{1}{b}\sum_{i=1}^{b} F(\mathbf{x}_t; \xi_{t,i}).$$

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

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \\
=&\mathbb{E}_{\mathcal{S}_t} \|\mathbf{x}_t - \eta_t \nabla F(\mathbf{x}_t; \mathcal{S}_t) - \mathbf{x}^*\|_2^2 \\
=& \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t \mathbb{E}_{\mathcal{S}_t} \langle \mathbf{x}_t - \mathbf{x}^*, \nabla F(\mathbf{x}_t; \mathcal{S}_t)\rangle + \eta_t^2 \mathbb{E}_{\mathcal{S}_t} \|\nabla F(\mathbf{x}_t; \mathcal{S}_t)\|_2^2 \\
=& \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t \langle \mathbf{x}_t - \mathbf{x}^*, \nabla f(\mathbf{x}_t)\rangle + \eta_t^2 \mathbb{E}_{\mathcal{S}_t} \|\nabla F(\mathbf{x}_t; \mathcal{S}_t)\|_2^2 \\
\leq& \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + 2\eta_t (f(\mathbf{x}^*) - f(\mathbf{x}_t)) + \eta_t^2 \underbrace{\mathbb{E}_{\mathcal{S}_t} \|\nabla F(\mathbf{x}_t; \mathcal{S}_t)\|_2^2}_{C_t},
\end{aligned} \tag{50}$$

where the inequality is because of the strong convexity. Furthermore,

$$\begin{aligned}
C_t =& \mathbb{E}_{\mathcal{S}_t} \|\nabla F(\mathbf{x}_t; \mathcal{S}_t) - \nabla F(\mathbf{x}^*; \mathcal{S}_t) + \nabla F(\mathbf{x}^*; \mathcal{S}_t)\|_2^2 \\
\leq& 2\mathbb{E}_{\mathcal{S}_t} \|\nabla F(\mathbf{x}_t; \mathcal{S}_t) - \nabla F(\mathbf{x}^*; \mathcal{S}_t)\|_2^2 + 2\mathbb{E}_{\mathcal{S}_t} \|\nabla F(\mathbf{x}^*; \mathcal{S}_t)\|_2^2.
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t} \|\nabla F(\mathbf{x}_t; \mathcal{S}_t) - \nabla F(\mathbf{x}^*; \mathcal{S}_t)\|_2^2 \\
\leq& \mathbb{E}_{\mathcal{S}_t} [2L(F(\mathbf{x}_t; \mathcal{S}_t) - F(\mathbf{x}^*; \mathcal{S}_t) - \langle \nabla F(\mathbf{x}^*; \mathcal{S}_t), \mathbf{x}_t - \mathbf{x}^*\rangle)] \\
=& 2L(f(\mathbf{x}_t) - f(\mathbf{x}^*)),
\end{aligned}$$

where the inequality is due to the third statement of Theorem 3.19. Let

$$V^* = \mathbb{E}_\xi \|\nabla F(\mathbf{x}^*; \xi) - \nabla f(\mathbf{x}^*)\|_2^2.$$

For the second term, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t} \|\nabla F(\mathbf{x}^*; \mathcal{S}_t)\|_2^2 \\
=& \mathbb{E}_{\mathcal{S}_t} \|\nabla F(\mathbf{x}^*; \mathcal{S}_t) - \nabla f(\mathbf{x}^*)\|_2^2 \\
=& \mathbb{E}_{\mathcal{S}_t} \left\| \frac{1}{b}\sum_{i=1}^{b} \left(\nabla F(\mathbf{x}^*; \xi_{t,i}) - \mathbb{E}[\nabla F(\mathbf{x}^*; \xi_{t,i})]\right)\right\|_2^2 \\
=& \frac{1}{b}\mathbb{E}_{\xi_{t,i}} \|\nabla F(\mathbf{x}^*; \xi_{t,i}) - \mathbb{E}[\nabla F(\mathbf{x}^*; \xi_{t,i})]\|_2^2 \\
=& \frac{V^*}{b}.
\end{aligned}$$

Hence, we have

$$C_t \leq 4L(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{2V^*}{b}$$

and

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \\
\leq& \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + 2\eta_t (f(\mathbf{x}^*) - f(\mathbf{x}_t)) + 4\eta_t^2 L(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{2\eta_t^2 V^*}{b} \\
=& \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + (2\eta_t - 4\eta_t^2 L)(f(\mathbf{x}^*) - f(\mathbf{x}_t)) + \frac{2\eta_t^2 V^*}{b}.
\end{aligned}$$

We sum over above inequality over $t = 0, \dots, T-1$ and take expectation on all of history, then

$$
\begin{aligned}
&\sum_{t=0}^{T-1} 2\eta_t(1 - 2\eta_t L)(\mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}^*)) \\
\leq &\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2 - \mathbb{E}\left\|\mathbf{x}_T - \mathbf{x}^*\right\|_2^2 + \frac{2V^* \sum_{t=0}^{T-1} \eta_t^2}{b} \\
\leq &\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2 + \frac{2V^* \sum_{t=0}^{T-1} \eta_t^2}{b}
\end{aligned}
$$

Taking $\eta_t \leq 1/(3L)$, we have $\eta_t(1 - 2\eta_t L) \geq \eta_t/3$. Hence,

$$
\sum_{t=0}^{T-1} \frac{2\eta_t}{3}(\mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}^*)) \leq \left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2 + \frac{2V^* \sum_{t=0}^{T-1} \eta_t^2}{b}.
$$

For fixed $\eta_t = \eta$, we have

$$
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}^*) \leq \frac{3\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{2\eta T} + \frac{3V^* \sum_{t=0}^{T-1} \eta}{bT} = \frac{3\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{2\eta T} + \frac{3V^*\eta}{b}.
$$

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

$$
\mathbb{E}[f(\bar{\mathbf{x}}_T)] - f(\mathbf{x}^*) \leq \frac{3L\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{2\sqrt{T}} + \frac{3V^*}{L\sqrt{T}}.
$$

  We require $T = \mathcal{O}(\epsilon^{-2})$ to obtain ϵ-suboptimal solution.

- For general, we set $\eta = 1/(L\sqrt{T/b})$. Then

$$
\mathbb{E}[f(\bar{\mathbf{x}}_T)] - f(\mathbf{x}^*) \leq \frac{3L\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{2\sqrt{bT}} + \frac{3V^*}{L\sqrt{bT}}.
$$

  We require $T = \mathcal{O}(\epsilon^{-2}/b)$ to obtain ϵ-suboptimal solution.

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

$$
\begin{aligned}
\mathbb{E}_i[f(\mathbf{x}_t - \eta\nabla f_i(\mathbf{x}_t))] \leq& f(\mathbf{x}_t) - \eta_t\mathbb{E}_i[\langle\nabla f(\mathbf{x}_t), \nabla f_i(\mathbf{x}_t)\rangle] + \frac{L\eta_t^2}{2}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_t)\right\|_2^2 \\
=& f(\mathbf{x}_t) - \eta_t\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{L\eta_t^2}{2}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_t)\right\|_2^2 \\
\leq& f(\mathbf{x}_t) - \eta_t\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + L\eta_t^2\mathbb{E}\big[\left\|\nabla f_i(\mathbf{x}_t) - \nabla f(\mathbf{x})\right\|_2^2 + \left\|\nabla f(\mathbf{x})\right\|_2^2\big] \\
\leq& f(\mathbf{x}_t) - \eta_t\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + L(\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + V_t)\eta_t^2.
\end{aligned}
$$

Taking

$$
\eta_t = \frac{\left\|\nabla f(\mathbf{x}_t)\right\|_2^2}{2L(\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + V_t)}
$$

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)

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

for any $\mathbf{x} \in \mathbb{R}^d$. Summing over $i = 1, \dots, n$ and using $\nabla f(\mathbf{x}^*) = \mathbf{0}$, we obtain

$$\begin{aligned}
&\mathbb{E}_{i_t} \|\nabla f_{i_t}(\mathbf{x}) - \nabla f_{i_t}(\mathbf{x}^*)\|_2^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \\
\le& \frac{1}{n}\sum_{i=1}^n 2L(f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle) \\
\le& 2L(f(\mathbf{x}) - f(\mathbf{x}^*)).
\end{aligned}$$

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

$$
\begin{aligned}
&\mathbb{E}_{i_t}\left\|\mathbf{v}_t\right\|_2^2 \\
\leq&2\mathbb{E}_{i_t}\left\|\nabla f_{i_t}(\mathbf{x}_t)-\nabla f_{i_t}(\mathbf{x}^*)\right\|_2^2+2\mathbb{E}_{i_t}\left\|\nabla f_{i_t}(\mathbf{x}^*)-\nabla f_{i_t}(\tilde{\mathbf{x}})+\nabla f(\tilde{\mathbf{x}})\right\|_2^2 \\
=&2\mathbb{E}_{i_t}\left\|\nabla f_{i_t}(\mathbf{x}_t)-\nabla f_{i_t}(\mathbf{x}^*)\right\|_2^2+2\mathbb{E}_{i_t}\left\|\nabla f_{i_t}(\mathbf{x}^*)-\nabla f_{i_t}(\tilde{\mathbf{x}})-\mathbb{E}[\nabla f_{i_t}(\mathbf{x}^*)-\nabla f_{i_t}(\tilde{\mathbf{x}})]\right\|_2^2 \\
=&2\mathbb{E}_{i_t}\left\|\nabla f_{i_t}(\mathbf{x}_t)-\nabla f_{i_t}(\mathbf{x}^*)\right\|_2^2+2\mathbb{E}_{i_t}\left\|\nabla f_{i_t}(\mathbf{x}^*)-\nabla f_{i_t}(\tilde{\mathbf{x}})\right\|_2^2 \\
\leq&4L(f(\mathbf{x}_t)-f(\mathbf{x}^*))+4L(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)).
\end{aligned}
$$

We also have $\mathbb{E}[\mathbf{v}_t]=\nabla f(\mathbf{x}_t)$. Hence,

$$
\begin{aligned}
&\mathbb{E}_{i_t}\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2 \\
=&\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\mathbb{E}_{i_t}[\langle\mathbf{x}_t-\mathbf{x}^*,\mathbf{v}_t\rangle]+\eta^2\mathbb{E}_{i_t}\left\|\mathbf{v}_t\right\|_2^2 \\
\leq&\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta\left\langle\mathbf{x}_t-\mathbf{x}^*,\nabla f(\mathbf{x}_t)\right\rangle+4L\eta^2(f(\mathbf{x}_t)-f(\mathbf{x}^*)+f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)) \\
\leq&\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-2\eta(f(\mathbf{x}_t)-f(\mathbf{x}^*))+4L\eta^2(f(\mathbf{x}_t)-f(\mathbf{x}^*)-f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)) \\
=&\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2-(2\eta-4\eta^2L)(f(\mathbf{x}_t)-f(\mathbf{x}^*))+4L\eta^2(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)).
\end{aligned}
$$

For stage the s-th stage, we let $\tilde{\mathbf{x}}=\mathbf{x}^{(s)}$ and $\mathbf{x}^{(s+1)}$ is sampled from $\{\mathbf{x}_0,\dots,\mathbf{x}_{m-1}\}$. Summing above over $t=0,\dots,m-1$ and taking expectation with all the history, we have

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_m-\mathbf{x}^*\right\|_2^2\leq&\mathbb{E}\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2^2-2\eta(1-2\eta L)\mathbb{E}\sum_{i=0}^{m-1}(f(\mathbf{x}_t)-f(\mathbf{x}^*))+4Lm\eta^2\mathbb{E}[f(\tilde{\mathbf{x}}^{(s)})-f(\mathbf{x}^*)] \\
=&\mathbb{E}\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2^2-2\eta(1-2\eta L)m\mathbb{E}[f(\mathbf{x}^{(s+1)})-f(\mathbf{x}^*)]+4Lm\eta^2\mathbb{E}[f(\tilde{\mathbf{x}}^{(s)})-f(\mathbf{x}^*)].
\end{aligned}
$$

which means

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}_m-\mathbf{x}^*\right\|_2^2+2\eta(1-2\eta L)m\mathbb{E}[f(\mathbf{x}^{(s+1)})-f(\mathbf{x}^*)] \\
\leq&\mathbb{E}\left\|\mathbf{x}_0-\mathbf{x}^*\right\|_2^2+4Lm\eta^2\mathbb{E}[f(\tilde{\mathbf{x}}^{(s)})-f(\mathbf{x}^*)] \\
=&\mathbb{E}\|\mathbf{x}^{(s)}-\mathbf{x}^*\|_2^2+4Lm\eta^2\mathbb{E}[f(\tilde{\mathbf{x}}^{(s)})-f(\mathbf{x}^*)] \\
\leq&\frac{2}{\mu}\mathbb{E}[f(\tilde{\mathbf{x}}^{(s)})-f(\mathbf{x}^*)]+4Lm\eta^2\mathbb{E}[f(\tilde{\mathbf{x}}^{(s)})-f(\mathbf{x}^*)] \\
\leq&\left(\frac{2}{\mu}+4Lm\eta^2\right)\mathbb{E}[f(\tilde{\mathbf{x}}^{(s)})-f(\mathbf{x}^*)].
\end{aligned}
$$

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 ϵ-suboptimal solution requires $S=\log(1/\epsilon)$ and IFO complexity is $S(m+n)=\mathcal{O}((n+\kappa)\log(1/\epsilon))$.*

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

$$\begin{aligned}
\mathbb{E}_t[f(\mathbf{x}_{t+1})] \leq & f(\mathbf{x}_t) - \mathbb{E}_t\left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\right\rangle + \frac{L}{2}\mathbb{E}_t\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2 \\
= & f(\mathbf{x}_t) - \eta\mathbb{E}_t\left\langle \nabla f(\mathbf{x}_t), \frac{1}{b}\sum_{i=1}^{b}\nabla F(\mathbf{x}_t;\xi_{t_i})\right\rangle + \frac{L\eta^2}{2}\mathbb{E}_t\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla F(\mathbf{x}_t;\xi_{t_i})\right\|_2^2 \\
\leq & f(\mathbf{x}_t) - \eta\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + L\eta^2\left(\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + \mathbb{E}_t\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla F(\mathbf{x}_t;\xi_i) - \nabla f(\mathbf{x}_t)\right\|_2^2\right) \\
\leq & f(\mathbf{x}_t) - (\eta - L\eta^2)\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{L\eta^2\sigma^2}{b}.
\end{aligned}$$

Let $\eta = 1/(2L)$ and $b = 2\sigma^2\epsilon^{-2}$, then

$$\mathbb{E}_t[f(\mathbf{x}_{t+1})] \leq f(\mathbf{x}_t) - \frac{1}{4L}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{\epsilon^2}{8L} \quad\Longrightarrow\quad \left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \leq 4L(f(\mathbf{x}_t) - \mathbb{E}_t[f(\mathbf{x}_{t+1})]) + \frac{\epsilon^2}{2}.$$

Let $\mathbf{x}_{\text{out}} = \mathbf{x}_j$ with j uniformly sampled from $\{0, \dots, T-1\}$ and $T = \lceil 8L(f(\mathbf{x}_0) - f^*)\epsilon^{-2}\rceil$. Taking expectation on all of history and averaging over $t = 0, \dots, T-1$, 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 11.2. *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*

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2 + \frac{\eta}{2}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2. \tag{51}$$

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

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

□

Lemma 11.2. *For update rule*

$$
\mathbf{v}_{t+1} = \mathbf{v}_t + \frac{1}{b} \sum_{\xi \in \mathcal{S}_{t+1}} (\nabla F(\mathbf{x}_{t+1}; \xi) - \nabla F(\mathbf{x}_t; \xi))
$$

in SARAH, we have

$$
\mathbb{E} \|\mathbf{v}_{t+1} - \nabla f(\mathbf{x}_{t+1})\|_2^2 \leq (1-p)\mathbb{E} \|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|_2^2 + \frac{(1-p)L^2}{b} \mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 + \frac{p\sigma^2}{b_0}.
$$

Proof. We first consider the case of

$$
\mathbf{v}_{t+1} = \mathbf{v}_t + \frac{1}{b} \sum_{\xi \in \mathcal{S}_{t+1}} (\nabla F(\mathbf{x}_{t+1}; \xi) - \nabla F(\mathbf{x}_t; \xi)) .
$$

Conditioned on $\mathbf{x}_0, \dots, \mathbf{x}_{t+1}$ and $\mathbf{v}_0, \dots, \mathbf{v}_t$, we have

$$
\mathbb{E}_{\mathcal{S}_{t+1}} [\mathbf{v}_{t+1} - \mathbf{v}_t] = \mathbb{E}_{\mathcal{S}_{t+1}} \left[\frac{1}{b} \sum_{\xi \in \mathcal{S}_{t+1}} (\nabla F(\mathbf{x}_{t+1}; \xi) - \nabla F(\mathbf{x}_t; \xi)) \right] = \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t).
$$

Hence, we obtain

$$
\begin{aligned}
& \mathbb{E} \|\mathbf{v}_{t+1} - \nabla f(\mathbf{x}_{t+1})\|_2^2 \\
= & \mathbb{E} \left\| \mathbf{v}_t + \frac{1}{b} \sum_{\xi \in \mathcal{S}_{t+1}} (\nabla F(\mathbf{x}_{t+1}; \xi) - \nabla F(\mathbf{x}_t; \xi)) - \nabla f(\mathbf{x}_{t+1}) \right\|_2^2 \\
= & \mathbb{E} \left\| \mathbf{v}_t - \nabla f(\mathbf{x}_t) + \frac{1}{b} \sum_{\xi \in \mathcal{S}_{t+1}} (\nabla F(\mathbf{x}_{t+1}; \xi) - \nabla F(\mathbf{x}_t; \xi)) - (\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)) \right\|_2^2 \\
= & \mathbb{E} \|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|_2^2 + \mathbb{E} \left\| \frac{1}{b} \sum_{\xi \in \mathcal{S}_{t+1}} (\nabla F(\mathbf{x}_{t+1}; \xi) - \nabla F(\mathbf{x}_t; \xi)) - (\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)) \right\|_2^2 \\
& + \left\langle \mathbf{v}_t - \nabla f(\mathbf{x}_t), \frac{1}{b} \sum_{\xi \in \mathcal{S}_{t+1}} (\nabla F(\mathbf{x}_{t+1}; \xi) - \nabla F(\mathbf{x}_t; \xi)) - (\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)) \right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2 + \mathbb{E}\left\|\frac{1}{b}\sum_{\xi\in\mathcal{S}_{t+1}}\left(\nabla F(\mathbf{x}_{t+1};\xi) - \nabla F(\mathbf{x}_t;\xi)\right) - \left(\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)\right)\right\|_2^2 \\
&= \mathbb{E}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{1}{b}\mathbb{E}\left\|\nabla F(\mathbf{x}_{t+1};\xi) - \nabla F(\mathbf{x}_t;\xi) - \left(\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)\right)\right\|_2^2 \\
&\leq \mathbb{E}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{1}{b}\mathbb{E}\left\|\nabla F(\mathbf{x}_{t+1};\xi) - \nabla F(\mathbf{x}_t;\xi)\right\|_2^2 \\
&\leq \mathbb{E}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{L^2}{b}\mathbb{E}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2.
\end{aligned}
$$

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

$$
\mathbb{E}\left\|\mathbf{v}_{t+1} - \nabla f(\mathbf{x}_{t+1})\right\|_2^2 = \mathbb{E}\left\|\frac{1}{b_0}\sum_{\xi\in\mathcal{S}_{t+1}}\nabla F(\mathbf{x}_{t+1};\xi) - \nabla f(\mathbf{x}_{t+1})\right\|_2^2 \leq \frac{\sigma^2}{b_0}.
$$

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

$$
\begin{aligned}
\mathbb{E}[\Phi_{t+1}] &= \mathbb{E}\left[f(\mathbf{x}_{t+1}) + \frac{\eta}{2p}\left\|\mathbf{v}_{t+1} - \nabla f(\mathbf{x}_{t+1})\right\|_2^2\right] \\
&\leq \mathbb{E}\Bigg[f(\mathbf{x}_t) - \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2 + \frac{\eta}{2}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2 \\
&\qquad + \frac{\eta}{2p}\left((1-p)\mathbb{E}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{(1-p)L^2}{b}\mathbb{E}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2\right) + \frac{p\sigma^2}{b_0}\Bigg] \\
&\leq \mathbb{E}\left[f(\mathbf{x}_t) + \frac{\eta}{2p}\mathbb{E}\left\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\right\|_2^2 - \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 - \left(\frac{1}{2\eta} - \frac{L}{2} - \frac{(1-p)L^2\eta}{2pb}\right)\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2 + \frac{\eta\sigma^2}{2b_0}\right] \\
&\leq \mathbb{E}\left[\Phi_t - \frac{\eta}{2}\left\|\nabla f(\mathbf{x}_t)\right\|_2^2 + \frac{\eta\sigma^2}{2b_0}\right],
\end{aligned}
$$

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 $t = 0, \dots, T-1$, 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 ϵ^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 11.3. *The value of b can be selected by minimizing $b_0p + b$ with constraint $bp = 1$. That is*

$$b_0p + b = \frac{b_0}{b} + b \geq 2\sqrt{b_0},$$

where the equality is taken by $b = \sqrt{b_0}$.

Remark 11.4. *Similarly, we take $b_0 = n$ and $b = \Theta(\sqrt{n}\,)$ for finite-sum case.*