Homework

PB22010344 黄境

2024 年 12 月 11 日

Exercise 1. **Proximal Operator**

Solution. 1. We have

$$\begin{aligned} p(\mathbf{x}_c) &= \underset{\mathbf{x}}{\mathbf{argmin}}\{g(\mathbf{x}) + \frac{L}{2}\|\mathbf{x} - (\mathbf{x}_c - \frac{1}{L}\nabla f(\mathbf{x}_c))\|^2\} \\ &= \underset{\mathbf{x}}{\mathbf{argmin}}\{\frac{g(\mathbf{x})}{L} + \frac{1}{2}\|\mathbf{x} - (\mathbf{x}_c - \frac{1}{L}\nabla f(\mathbf{x}_c))\|^2\} \\ &= \mathrm{prox}_{\frac{g}{L}}(\mathbf{x}_c - \frac{1}{L}\nabla f(\mathbf{x}_c)) \end{aligned}$$

2.(a) $\forall\, \mathbf{x} \in \mathbb{R}^n$, $\mathrm{prox}_f(x)$ exists and is unique if and only if the optimization problem

$$\min_{\mathbf{u}\in \mathbf{dom}\ f} f(\mathbf{u}) + \frac{1}{2}\|\mathbf{x} - \mathbf{u}\|^2$$

has one unique solution.

Since f is convex and close, $f(\mathbf{u}) + \frac{1}{2}\|\mathbf{x} - \mathbf{u}\|^2$ is strongly convex with parameter $1 > 0$. Therefore, from exercise 1.6 in HW4, the problom admits a unique solution, which implies that $\mathrm{prox}_f(x)$ exists and is unique, $\forall\, \mathbf{x} \in \mathbb{R}^n$.

(b) Let $g(\mathbf{u}) = f(\mathbf{u}) + \frac{1}{2}\|\mathbf{x} - \mathbf{u}\|^2$, we have

$$\partial g(\mathbf{u}) = \{\mathbf{g} : \mathbf{g} = \mathbf{g}_f + \mathbf{u} - \mathbf{x},\, \mathbf{g}_f \in \partial f(\mathbf{u})\}.$$

($\Rightarrow$) Since $\mathbf{u} = \mathop{\mathbf{argmin}}\limits_{\mathbf{u}\in \mathbf{dom}\ f} g(\mathbf{u})$, which implies that

$$g(\mathbf{v}) \geq g(\mathbf{u}) = g(\mathbf{u}) + \langle \mathbf{0},\, \mathbf{u}-\mathbf{v}\rangle,\, \forall\, \mathbf{v} \in \mathbf{dom}\ f \quad \Rightarrow \quad \mathbf{0} \in \partial g(\mathbf{u})$$

That is, $\mathbf{x} - \mathbf{u} \in \partial f(\mathbf{u})$

($\Leftarrow$) Similarly, since $\mathbf{x} - \mathbf{u} \in \partial f(\mathbf{u})$, we have $\mathbf{0} \in \partial g(\mathbf{u})$, which implies that $\mathbf{u} = \mathop{\mathbf{argmin}}\limits_{\mathbf{u}\in \mathbf{dom}\ f} g(\mathbf{u}) = \mathrm{prox}_f(\mathbf{x})$.

3.(a) Let $\mathbf{v} = \lambda\mathbf{u} + \mathbf{a}$, we have

$$\begin{aligned}
\mathrm{prox}_h(\mathbf{x}) &= \mathop{\mathbf{argmin}}\limits_{\mathbf{u}\in \mathbf{dom}\ h} \{h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\} \\
&= \mathop{\mathbf{argmin}}\limits_{\lambda\mathbf{u}+\mathbf{a}\in \mathbf{dom}\ f} \{f(\lambda\mathbf{u}+\mathbf{a}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\} \\
&= \frac{1}{\lambda}(\mathop{\mathbf{argmin}}\limits_{\mathbf{v}\in \mathbf{dom}\ f} \{f(\mathbf{v}) + \frac{1}{2}\|\frac{\mathbf{v}-\mathbf{a}}{\lambda} - \mathbf{x}\|^2\} - \mathbf{a}) \\
&= \frac{1}{\lambda}(\mathop{\mathbf{argmin}}\limits_{\mathbf{v}\in \mathbf{dom}\ f} \{\lambda^2 f(\mathbf{v}) + \frac{1}{2}\|\mathbf{v}-\mathbf{a}-\lambda\mathbf{x}\|^2\} - \mathbf{a}) \\
&= \frac{1}{\lambda}(\mathrm{prox}_{\lambda^2 f}(\lambda\mathbf{x}+\mathbf{a}) - \mathbf{a})
\end{aligned}$$

(b) Similarly,

$$\begin{aligned}
\mathrm{prox}_h(\mathbf{x}) &= \mathop{\mathbf{argmin}}\limits_{\mathbf{u}\in \mathbf{dom}\ h} \{h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\} \\
&= \mathop{\mathbf{argmin}}\limits_{\frac{\mathbf{u}}{\lambda}\in \mathbf{dom}\ f} \{\lambda f(\frac{\mathbf{u}}{\lambda}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\} \\
&= \frac{1}{\lambda}\mathop{\mathbf{argmin}}\limits_{\mathbf{v}\in \mathbf{dom}\ f} \{\lambda f(\mathbf{v}) + \frac{1}{2}\|\lambda\mathbf{v}-\mathbf{x}\|^2\} \\
&= \frac{1}{\lambda}\cdot\lambda^2 \mathop{\mathbf{argmin}}\limits_{\mathbf{v}\in \mathbf{dom}\ f} \{\lambda^{-1} f(\mathbf{v}) + \frac{1}{2}\|\mathbf{v}-\frac{\mathbf{x}}{\lambda}\|^2\} \\
&= \lambda\,\mathrm{prox}_{\lambda^{-1} f}(\frac{\mathbf{x}}{\lambda})
\end{aligned}$$

(c)

$$
\begin{aligned}
\mathrm{prox}_h(\mathbf{x}) &= \underset{\mathbf{u}\in\mathbf{dom}\ f}{\mathbf{argmin}}\{f(\mathbf{u}) + \mathbf{a}^T\mathbf{u} + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\} \\
&= \underset{\mathbf{u}\in\mathbf{dom}\ f}{\mathbf{argmin}}\{f(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}+\mathbf{a}\|^2 + \mathbf{a}^T\mathbf{x} - \frac{1}{2}\|\mathbf{a}\|^2\} \\
&= \underset{\mathbf{u}\in\mathbf{dom}\ f}{\mathbf{argmin}}\{f(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}+\mathbf{a}\|^2\} \\
&= \mathrm{prox}_f(\mathbf{x}-\mathbf{a})
\end{aligned}
$$

4. (a) Consider two cases:

(i) $\mathbf{x} = \mathbf{0}$, and $\mathrm{prox}_f(\mathbf{x}) = \underset{\mathbf{u}\in\mathbf{dom}\ f}{\mathbf{argmin}}\{\|\mathbf{u}\| + \frac{1}{2}\|\mathbf{u}\|^2\} = \mathbf{0}$.

(ii) $\mathbf{x} \neq \mathbf{0}$. Fix $\|\mathbf{u}\| = t$, let

$$
g_t(\mathbf{x}) = \min_{\substack{\mathbf{u}\in\mathbf{dom}\ f \\ \|\mathbf{u}\|=t}}\{t + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\},
$$

we have

$$
\begin{aligned}
g_t(\mathbf{x}) &= t + \frac{1}{2}\|\frac{t\mathbf{x}}{\|\mathbf{x}\|} - \mathbf{x}\|^2 \\
&= t + \frac{1}{2}(t - \|\mathbf{x}\|)^2 \\
&= \frac{1}{2}(t^2 + 2t(1-\|\mathbf{x}\|) + \|\mathbf{x}\|^2)
\end{aligned}
$$

Therefore

$$
\underset{t}{\mathbf{argmin}}\ g_t(\mathbf{x}) = \|\mathbf{x}\| - 1 \quad \Rightarrow \quad \mathrm{prox}_f(\mathbf{x}) = \frac{\|\mathbf{x}\|-1}{\|\mathbf{x}\|}\mathbf{x}.
$$

Since $t \geq 0$, we can conclude that

$$
\mathrm{prox}_f(\mathbf{x}) = \begin{cases} \frac{\|\mathbf{x}\|-1}{\|\mathbf{x}\|}\mathbf{x},\ \mathbf{x} > 1 \\ \mathbf{0},\ \|\mathbf{x}\| \leq 1 \end{cases}
$$

(b) Let

$$
g(\mathbf{u}) = I_C(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2,
$$

we have

$$g(\mathbf{u}) = \begin{cases} 1 + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2, \; \mathbf{u} \in C \\ \\ +\infty, \; \mathbf{u} \notin C \end{cases}$$

Therefore $\text{prox}_f(\mathbf{x}) = \underset{\mathbf{u}\in C}{\mathbf{argmin}} \; \|\mathbf{u}-\mathbf{x}\|^2 = \mathbf{P}_C(\mathbf{x})$ is the projection of $\mathbf{x}$ on C.

Exercise 2. **Proximal Gradient**

Solution. 1. Since $\mathbf{x} \in \mathbf{int}\,(\mathbf{dom}\; F)$, $\partial F(\mathbf{x})$ is nonempty. Therefore, we have

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y}-\mathbf{x}\rangle, \; \forall \mathbf{y} \in \mathbf{dom}\; F,$$

which implies that

$$F(\mathbf{y}) - F(\mathbf{x}) \geq \langle \mathbf{g}, \mathbf{y}-\mathbf{x}\rangle \geq 0, \; \forall \mathbf{y} \in \mathbf{dom}\; F,$$

that is, $\mathbf{x}$ is optimal.

3. Consider the situation that $\mathbf{x} \in \partial\, \mathbf{dom}\; F$. To simplify the question, let $F(x) : \mathbb{R} \to \mathbb{R}$, we have

$$\partial F(x) = \emptyset \quad \Leftrightarrow \quad \forall\, g \in \mathbb{R}, \; \exists\, y \in \mathbf{dom}\; F, \;\; \text{s.t. } F(y) < F(x) + g(y-x).$$

Thus, we can set $F(y) = -e^y$, $\mathbf{dom}\; F = \mathbb{R}^+ \cup \{0\}$, $x = 0$.

4. Since f is twice continuously differentiable, by the Taylor's Theorem with Lagrange's form of remainder, we have

$$f(\mathbf{y}) = f(\mathbf{x}) + \mathbf{J}f(\mathbf{x})\cdot(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T\mathbf{H}f(\boldsymbol{\xi})(\mathbf{y}-\mathbf{x}), \; \boldsymbol{\xi} \in \overline{\mathbf{xy}}.$$

Since $\mathbf{x}^T\mathbf{H}\mathbf{x} \leq \lambda_{max}\|\mathbf{x}\|^2$, where λ_{max} is the largest eigenvalue of the symmetric matrix $\mathbf{H}$, we have

$$\begin{aligned} f(\mathbf{y}) &\leq f(\mathbf{x}) + \langle \nabla f, \, \mathbf{y}-\mathbf{x}\rangle + \frac{1}{2}\lambda_{max}\|\mathbf{y}-\mathbf{x}\|^2 \\ &\leq f(\mathbf{x}) + \langle \nabla f, \, \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 \end{aligned}$$

5. From exercise 1.6(d) in HW4, we only need to prove that $Q(\mathbf{x};\mathbf{x}_c)$ is strongly convex with the parameter L, i.e. $g(\mathbf{x}) + \langle \nabla f(\mathbf{x}),\, \mathbf{x}\rangle$ is convex, which is obvious.

6. $g(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$, $f(\mathbf{w}) = \frac{1}{n}\|\mathbf{y} - X\mathbf{w}\|_2^2$, therefore,

$$\mathbf{w}^+ = p(\mathbf{w}_k) = \underset{\mathbf{w}}{\mathbf{argmin}}\,\{\lambda\|\mathbf{w}\|_1 + \frac{L}{2}\|\mathbf{w} - (\mathbf{w}_k - \frac{1}{L}\nabla f(\mathbf{w}_k))\|^2\}.$$

Since $\mathbf{0} \in \partial p(\mathbf{w}^+)$, we have

$$\mathbf{0} \in \partial\lambda\|\mathbf{w}^+\|_1 + \frac{L}{2}\nabla\|\mathbf{w}^+ - \mathbf{z}\|^2 \quad \Rightarrow \quad \frac{L}{\lambda}(\mathbf{z} - \mathbf{w}^+) \in \partial\lambda\|\mathbf{w}^+\|_1.$$

Since

$$\partial\lambda\|\mathbf{w}^+\|_1 = \left\{\mathbf{v} \in \mathbb{R}^n, v_i = \begin{cases} 1,\ w_i > 0 \\ [-1,1],\ w_i = 0 \\ -1,\ w_i < 0 \end{cases}\right\},$$

we have

$$w_i^+ = \begin{cases} z_i + \dfrac{\lambda}{L},\ z_i < -\dfrac{\lambda}{L} \\ 0,\ |z_i| \le \dfrac{\lambda}{L} \\ z_i - \dfrac{\lambda}{L},\ z_i > \dfrac{\lambda}{L} \end{cases}$$

Exercise 3. **ISTA with Backtracking**

Solution. 1. Let $\mathbf{x}_k = p_{L_k}(\mathbf{x}_{k-1})$, we have

$$\mathbf{0} \in \partial Q_{L_k}(\mathbf{x}_k;\, \mathbf{x}_{k-1}) \Rightarrow -\nabla f(\mathbf{x}_k) - L_k(\mathbf{x}_k - \mathbf{x}_{k-1}) \in \partial g(\mathbf{x}_k).$$

Denote $-\nabla f(\mathbf{x}_k) - L_k(\mathbf{x}_k - \mathbf{x}_{k-1})$ by $\mathbf{g}$. Therefore,

$$g(\mathbf{y}) - g(\mathbf{x}_k) \ge \langle \mathbf{g},\, \mathbf{y} - \mathbf{x}_k\rangle,\ \forall\, \mathbf{y} \in \mathbf{dom}\ g.$$

Since f is convex, we have

$$f(\mathbf{y}) \geq f(\mathbf{x}_{k-1}) + \langle \nabla f(\mathbf{x}_{k-1}),\, \mathbf{y} - \mathbf{x}_{k-1} \rangle,\ \forall\, \mathbf{y} \in \mathbf{dom}\ f.$$

Set $\mathbf{y} = \mathbf{x}_k$, all together we obtain

$$\begin{aligned}
F(\mathbf{x}_{k-1}) - F(\mathbf{x}_k) &\geq F(\mathbf{x}_{k-1}) - Q_{L_k}(\mathbf{x}_k;\, \mathbf{x}_{k-1}) \\
&= f(\mathbf{x}_{k-1}) - f(\mathbf{x}_{k-1}) + g(\mathbf{x}_{k-1}) - g(\mathbf{x}_k) \\
&\quad + \langle \nabla f(\mathbf{x}_{k-1}),\, \mathbf{x}_{k-1} - \mathbf{x}_k \rangle - \frac{L_k}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2 \\
&\geq \langle \mathbf{g} + \nabla f(\mathbf{x}_{k-1}),\, \mathbf{x}_{k-1} - \mathbf{x}_k \rangle - \frac{L_k}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2 \\
&= L_k \langle \mathbf{x}_{k-1} - \mathbf{x}_k,\, \mathbf{x}_{k-1} - \mathbf{x}_k \rangle - \frac{L_k}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2 \\
&= \frac{L_k}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2 \geq 0
\end{aligned}$$

which implies that $F(\mathbf{x}_k)$ is non-increasing.

2. Since

$$f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1}) + \langle \nabla f(\mathbf{x}_{k-1}),\, \mathbf{x}_k - \mathbf{x}_{k-1} \rangle + \frac{L}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2,$$

where $\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2 \geq 0$, we have

$$\begin{aligned}
F_{p_{\widetilde{L}}}(\mathbf{x}_{k-1}) &= f(\mathbf{x}_k) + g(\mathbf{x}_k) \\
&\leq f(\mathbf{x}_{k-1}) + \langle \nabla f(\mathbf{x}_{k-1}),\, \mathbf{x}_k - \mathbf{x}_{k-1} \rangle + \frac{L}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2 + g(\mathbf{x}_k) \\
&\leq f(\mathbf{x}_{k-1}) + \langle \nabla f(\mathbf{x}_{k-1}),\, \mathbf{x}_k - \mathbf{x}_{k-1} \rangle + \frac{\widetilde{L}}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2 + g(\mathbf{x}_k) \\
&= Q_{\widetilde{L}}(\mathbf{x}_k;\, \mathbf{x}_{k-1}),\ \forall\, \widetilde{L} \geq L
\end{aligned}$$

Therefore, inequality (3) is satisfied for any $\widetilde{L} \geq L$.

Moreover, $L_k = \eta^{\sum_{i=1}^{k} i_k} L_0 = \eta^{j_k} L_0$, where $j_k = \sum_{i=1}^{k} i_k$ is the smallest integer which satisfies inequality (3). Set $r_k = [\frac{\ln L - \ln L_0}{\ln \eta}] + 1 > 1$, we have

$$L \leq \eta^{r_k} L_0 \leq \eta L,$$

which implies that $\eta^{r_k} L_0$ satisfies inequality (3). Since j_k is the smallest, we have $j_k \le r_k \Rightarrow L_k \le \eta^{r_k} L_0 \le \eta L$.

3. From exercise 3.1, we have

$$F(\mathbf{x}^*) - F(\mathbf{x}_k) \ge \frac{L_k}{2}(\|\mathbf{x}^* - \mathbf{x}_k\|^2 - \|\mathbf{x}^* - \mathbf{x}_{k-1}\|^2).$$

Since $F(\mathbf{x}_k)$ is non-increasing and $L_k \le \eta L$, summing up and we obtain

$$\begin{aligned}
\frac{2k}{\eta L}(F(\mathbf{x}_k) - F(\mathbf{x}^*)) &\le \sum_{i=1}^{k} \frac{2}{\eta L}(F(\mathbf{x}_i) - F(\mathbf{x}^*)) \\
&\le \sum_{i=1}^{k} \frac{2}{L_i}(F(\mathbf{x}_i) - F(\mathbf{x}^*)) \\
&\le \sum_{i=1}^{k} \|\mathbf{x}^* - \mathbf{x}_{i-1}\|^2 - \|\mathbf{x}^* - \mathbf{x}_i\|^2 \\
&= \|\mathbf{x}^* - \mathbf{x}_0\|^2 - \|\mathbf{x}^* - \mathbf{x}_k\|^2 \\
&\le \|\mathbf{x}^* - \mathbf{x}_0\|^2
\end{aligned}$$

Therefore, we have

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \le \frac{\eta L}{2k}\|\mathbf{x}^* - \mathbf{x}_0\|^2$$

Exercise 4. **Naive Bayes Classifier**

Solution. 3. We can convert the product to a sum by calculating the logarithm to avoid data overflow.

4. The result is shown below.

```
Training Naive Bayes: 100.00%

Testing Naive Bayes: 100.00%Confusion Matrix:
 [[391   0]
 [  2 191]]
Accuracy: 0.9965753424657534
Precision: 1.0
Recall: 0.9896373056994818
F1-score: 0.9947916666666666
Found a wrong prediction in position:543
Found a wrong prediction in position:566
```

5. The result is shown below. We can see that Laplace smoothing technique is useful.

```
Testing Naive Bayes: 100.00%Confusion Matrix:
 [[391   0]
 [112  81]]
Accuracy: 0.8082191780821918
Precision: 1.0
Recall: 0.41968911917098445
F1-score: 0.5912408759124088
```

Exercise 5. **Logistic Regression and Newton's Method**

Solution. 1.(a) Since

$$L(\mathbf{w}) = \frac{1}{n}\left(\sum_{i\in I^+} \ln(1 + e^{-\mathbf{w}^T\bar{\mathbf{x}}_i}) + \sum_{i\in I^-} \ln(1 + e^{\mathbf{w}^T\bar{\mathbf{x}}_i})\right),$$

set $\mathbf{w}_n = n\hat{\mathbf{w}}$, $n \in \mathbb{N}^+$, we have

$$\langle \mathbf{w}_n, \bar{\mathbf{x}}_i\rangle = n\langle \hat{\mathbf{w}}, \bar{\mathbf{x}}_i\rangle > 0, \forall\, i \in I^+,$$

$$\langle \mathbf{w}_n, \bar{\mathbf{x}}_i\rangle = n\langle \hat{\mathbf{w}}, \bar{\mathbf{x}}_i\rangle < 0, \forall\, i \in I^-,$$

therefore $L(\mathbf{w})$ is decreasing when $n \to \infty$, which implies that problem (4) has no solution on $\mathbb{R}^{d+1}$

(b) By the expression of $L(\mathbf{w})$ in exercise 4.1(a), we obtain that $L(\mathbf{w})$ is continuous on $\mathbb{R}^{d+1}$. Therefore, if problem (4) has no solution, we must have

$$\lim_{\|\mathbf{w}\|\to\infty} L(\mathbf{w}) = -\infty$$

On the other hand, let $\mathbf{w}_0 \in \mathbb{R}^{d+1}$, $\mathbf{w}_0 \neq \mathbf{0}$. WOLG, let $\langle \mathbf{w}_0,\, \bar{\mathbf{x}}_{i_0}\rangle < 0$, $i_0 \in I^+$. Therefore, we have

$$\begin{aligned} L(n\mathbf{w}_0) &= \frac{1}{n}(\sum_{i\in I^+} \ln(1+e^{-n\mathbf{w}_0^T\bar{\mathbf{x}}_i}) + \sum_{i\in I^-} \ln(1+e^{n\mathbf{w}_0^T\bar{\mathbf{x}}_i})) \\ &\geq \ln(1+e^{-n\mathbf{w}_0^T\bar{\mathbf{x}}_{i_0}}) \to \infty,\, n\to\infty, \end{aligned}$$

which leads to contradiction.

2. Since $\nabla^2 L(\mathbf{w}) = \bar{\mathbf{X}}\mathbf{D}\bar{\mathbf{X}}^T$, where

$$\mathbf{D} = diag(\frac{1+e^{-\mathbf{w}^T\bar{\mathbf{x}}_1}}{(1+e^{-\mathbf{w}^T\bar{\mathbf{x}}_1})^2}, \frac{1+e^{-\mathbf{w}^T\bar{\mathbf{x}}_2}}{(1+e^{-\mathbf{w}^T\bar{\mathbf{x}}_2})^2}, \ldots, \frac{1+e^{-\mathbf{w}^T\bar{\mathbf{x}}_n}}{(1+e^{-\mathbf{w}^T\bar{\mathbf{x}}_n})^2},).$$

Therefore, $\nabla^2 L(\mathbf{w})$ is positive definite, which implies that $L(\mathbf{w})$ is strictly convex.

Exercise 6. **Convergence of Stochastic Gradient Descent for Convex Function**

Solution. 1. Since F is strongly convex with parameter μ, set

$$G(\mathbf{u},\, \mathbf{v}) = F(\mathbf{v}) + \langle \nabla F(\mathbf{v}),\, \mathbf{u}-\mathbf{v}\rangle + \frac{\mu}{2}\|\mathbf{u}-\mathbf{v}\|^2 \leq F(\mathbf{u}),$$

we have

$$\nabla_{\mathbf{u}} G(\mathbf{u},\, \mathbf{v}) = \nabla F(\mathbf{v}) + \mu(\mathbf{u}-\mathbf{v}),$$

and

$$\nabla^2_{\mathbf{u}} G(\mathbf{u}) = \mu I,$$

which implies that $G(\mathbf{u}, \mathbf{v})$ is convex, and

$$\min_{\mathbf{u}} G(\mathbf{u}, \mathbf{v}) = G(\mathbf{v} - \frac{\nabla F(\mathbf{v})}{\mu}, \mathbf{v}) = F(\mathbf{v}) - \frac{1}{2\mu}\|\nabla F(\mathbf{v})\|^2.$$

Therefore, we have

$$F^* = F(\mathbf{w}^*) \geq G(\mathbf{w}^*, \mathbf{w}) \geq F(\mathbf{w}) - \frac{1}{2\mu}\|\nabla F(\mathbf{w})\|^2,$$

for all $\mathbf{w} \in \mathbf{dom}\ F$. Thus,

$$F(\mathbf{w}) - F^* \leq \frac{1}{2\mu}\|\nabla F(\mathbf{w})\|^2.$$

The strong convexity makes it easy to estimate the distance between $F(\mathbf{w}$ and F^* with gradient.

2. From Part 5.2 in Lecture 11, replace $\mathbf{g}(\xi_k) = f_{i_k}(\mathbf{w}_k)$ with $\mathbf{g}(\xi_k) = \frac{1}{n_m}\sum_{i\in\mathbf{S}_k} f_i(\mathbf{w}_k)$, we obtain

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k)] \leq -\alpha(1 - \frac{L}{2}\alpha)\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2}\alpha^2\mathbb{D}_{\xi_k}[\mathbf{g}(\xi_k)],$$

since

$$\|\mathbb{E}[\mathbf{g}(\xi_k)]\|^2 = \|\frac{1}{n_m}\sum_{i\in\mathbf{S}_k}\mathbb{E}[f_i(\mathbf{w}_k)]\|^2 = \|\nabla F(\mathbf{w}_k)\|^2.$$

With the Assumption 5, we have

$$\begin{aligned}
\mathbb{D}_{\xi_k}[\mathbf{g}(\xi_k)] &= \mathbb{D}_{\xi_k}[\frac{1}{n_m}\sum_{i\in\mathbf{S}_k} f_i(\mathbf{w}_k)] \\
&= \frac{1}{n_m^2}\sum_{i\in\mathbf{S}_k}\mathbb{D}_{\xi_k}[f_i(\mathbf{w}_k)] \\
&= \frac{1}{n_m^2}\cdot n_m\mathbb{D}_{\xi_k}[f_i(\mathbf{w}_k)] \\
&\leq \frac{1}{n_m}(M + M_V\|\nabla F(\mathbf{w})\|^2)
\end{aligned}$$

Follow the same steps in Lemma 3, set $M_G = M_V + n_m$, we have

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k)] \leq -\alpha(1 - \frac{L}{2n_m}M_G\alpha)\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2n_m}M\alpha^2.$$

Therefore, follow the same steps in Theorem 1, set $0 < \alpha < \frac{n_m}{LM_G}$, we have

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F^*] \leq F(\mathbf{w}_k) - F^* - \frac{\alpha}{2}\|\nabla F(\mathbf{w})\|^2 + \frac{L}{2n_m}M\alpha^2,$$

combining with Lemma 4 leads to

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F^*] \leq (1 - \mu\alpha)(F(\mathbf{w}_k) - F^*) + \frac{L}{2n_m}M\alpha^2,$$

that is,

$$\mathbb{E}_{\xi_{k-1}}[F(\mathbf{w}_k) - F^* - \frac{LM}{2\mu n_m}\alpha] \leq (1 - \mu\alpha)(F(\mathbf{w}_{k-1}) - F^* - \frac{LM}{2\mu n_m}\alpha).$$

Take the expectation with respect to ξ_{k-1}, $\ldots$, ξ_0, we obtain

$$\mathbb{E}_{\xi_0:\xi_{k-1}}[F(\mathbf{w}_k) - F^* - \frac{LM}{2\mu n_m}\alpha] \leq (1 - \mu\alpha)^k(F(\mathbf{w}_0) - F^* - \frac{LM}{2\mu n_m}\alpha),$$

which is the required inequality.

Since the speed of convergence in SGD is roughly $\frac{LM}{2}\alpha$, we can see that the number of step in SGD is roughly n_m times than which mini-batch SGD needs.