A Proof of Lemma 4.1

Proof. By the definition of $\boldsymbol{v}_{\mathcal{B}}$, we directly have

$$\begin{aligned}\mathbb{E}_{\mathcal{B}}\boldsymbol{v}_{\mathcal{B}} &= \mathbb{E}_{\mathcal{B}}\left(\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\nabla f_i(\boldsymbol{\theta}^{(t-1)}) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\nabla f_i(\widetilde{\boldsymbol{\theta}}) + \nabla\mathcal{F}(\widetilde{\boldsymbol{\theta}})\right)\\ &= \frac{1}{|\mathcal{B}|}\mathbb{E}_{\mathcal{B}}\sum_{i\in\mathcal{B}}\nabla f_i(\boldsymbol{\theta}^{(t-1)}) - \frac{1}{|\mathcal{B}|}\mathbb{E}_{\mathcal{B}}\sum_{i\in\mathcal{B}}\nabla f_i(\widetilde{\boldsymbol{\theta}}) + \nabla\mathcal{F}(\widetilde{\boldsymbol{\theta}})\\ &= \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)}) - \nabla\mathcal{F}(\widetilde{\boldsymbol{\theta}}) + \nabla\mathcal{F}(\widetilde{\boldsymbol{\theta}}) = \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)}).\end{aligned}$$

Thus $\boldsymbol{v}_{\mathcal{B}}$ is an unbiased estimator of $\nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})$. Let i be an index sampled from $\{1,\cdots,n\}$ with equal probability, we define $\boldsymbol{v}_i = \nabla f_i(\boldsymbol{\theta}^{(t-1)}) - \nabla f_i(\widetilde{\boldsymbol{\theta}}) + \widetilde{\boldsymbol{\mu}}$. Since all indices in $\mathcal{B}$ are independently sampled from $\{1,\cdots,n\}$ with equal probability, we have

$$\mathbb{E}_{\mathcal{B}}||\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})||^2 = \frac{1}{|\mathcal{B}|}\mathbb{E}_i||\boldsymbol{v}_i - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})||^2. \tag{A.1}$$

We then proceed to bound R.H.S. of (A.1) from above as follows,

$$\begin{aligned}\mathbb{E}_i||\boldsymbol{v}_i - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})||^2 &\le \mathbb{E}_i||\nabla f_i(\boldsymbol{\theta}^{(t-1)}) - \nabla f_i(\widetilde{\boldsymbol{\theta}}) + \nabla\mathcal{F}(\widetilde{\boldsymbol{\theta}}) - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})||^2\\ &\overset{\text{(i)}}{\le} \mathbb{E}_i||\nabla f_i(\boldsymbol{\theta}^{(t-1)}) - \nabla f_i(\widetilde{\boldsymbol{\theta}})||^2\\ &\overset{\text{(ii)}}{\le} 2\mathbb{E}_i||\nabla f_i(\boldsymbol{\theta}^{(t-1)}) - \nabla f_i(\widehat{\boldsymbol{\theta}})||^2 + 2\mathbb{E}_i||\nabla f_i(\widetilde{\boldsymbol{\theta}}) - \nabla f_i(\widehat{\boldsymbol{\theta}})||^2\\ &= \frac{2}{n}\sum_{i=1}^n||\nabla f_i(\boldsymbol{\theta}^{(t-1)}) - \nabla f_i(\widehat{\boldsymbol{\theta}})||^2 + \frac{2}{n}\sum_{i=1}^n||\nabla f_i(\widetilde{\boldsymbol{\theta}}) - \nabla f_i(\widehat{\boldsymbol{\theta}})||^2,\end{aligned} \tag{A.2}$$

where (i) comes from $\mathbb{E}||\boldsymbol{u}-\mathbb{E}\boldsymbol{u}||^2 = \mathbb{E}||\boldsymbol{u}||^2 - ||\mathbb{E}\boldsymbol{u}||^2 \le \mathbb{E}||\boldsymbol{u}||^2$, and (2) comes from $||\boldsymbol{u}-\boldsymbol{w}||^2 \le 2||\boldsymbol{u}||^2 + 2||\boldsymbol{w}||^2$ for any random vectors $\boldsymbol{u}$ and $\boldsymbol{w}$.

We then define $g_i(\boldsymbol{\theta}) = f_i(\boldsymbol{\theta}) - f_i(\widehat{\boldsymbol{\theta}}) - \nabla f_i(\widehat{\boldsymbol{\theta}})^T(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})$. By Assumption 2.1, we have

$$||\nabla g_i(\boldsymbol{\theta}') - \nabla g_i(\boldsymbol{\theta})|| = ||\nabla f_i(\boldsymbol{\theta}') - \nabla f_i(\boldsymbol{\theta})|| \le T_{\max}||\boldsymbol{\theta}'-\boldsymbol{\theta}||, \tag{A.3}$$

which implies that $\nabla g_i(\boldsymbol{\theta})$ is also Lipschitz continuous. Thus we have

$$\begin{aligned}0 = g_i(\widehat{\boldsymbol{\theta}}) &\le \min_{\alpha} g_i(\boldsymbol{\theta} - \alpha\nabla g_i(\boldsymbol{\theta}))\\ &= \min_{\alpha}[g_i(\boldsymbol{\theta}) - \alpha||\nabla g_i(\boldsymbol{\theta})||^2 + \frac{T_{\max}\alpha^2}{2}||\nabla g_i(\boldsymbol{\theta})||^2,\end{aligned} \tag{A.4}$$

where the first inequality comes from the fact that $\widehat{\boldsymbol{\theta}}$ is the minimizer to (A.3). Minimizing R.H.S. of (A.4), we have $\alpha = 1/T_{\max}$. Then (A.4) can be written as

$$0 \le g_i(\boldsymbol{\theta}) - \frac{1}{2T_{\max}}||\nabla g_i(\boldsymbol{\theta})||^2. \tag{A.5}$$

Combining the definition of $g_i(\boldsymbol{\theta})$ and (A.5), we further have

$$||\nabla f_i(\boldsymbol{\theta}) - \nabla f_i(\widehat{\boldsymbol{\theta}})||^2 \le 2T_{\max}\left[f_i(\boldsymbol{\theta}) - f_i(\widehat{\boldsymbol{\theta}}) - \nabla f_i(\widehat{\boldsymbol{\theta}})^T(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right]. \tag{A.6}$$

Taking the summation of (A.6) over $i = 1, ..., n$, we obtain

$$\frac{1}{n}\sum_{i=1}^n||\nabla f_i(\boldsymbol{\theta}) - \nabla f_i(\widehat{\boldsymbol{\theta}})||^2 \le 2T_{max}\left[\mathcal{F}(\boldsymbol{\theta}) - \mathcal{F}(\widehat{\boldsymbol{\theta}}) - \nabla\mathcal{F}(\widehat{\boldsymbol{\theta}})^T(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right]. \tag{A.7}$$

The optimality condition of $\widehat{\boldsymbol{\theta}}$ implies that there exists some $\boldsymbol{\xi} \in \partial\mathcal{R}(\widehat{\boldsymbol{\theta}})$ such that

$$\nabla\mathcal{F}(\widehat{\boldsymbol{\theta}}) + \boldsymbol{\xi} = \mathbf{0}. \tag{A.8}$$

Combining (A.8) with the convexity of $\mathcal{R}$, we have

$$\nabla\mathcal{F}(\widehat{\boldsymbol{\theta}})^T(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}) = -\boldsymbol{\xi}^T(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}) \le \mathcal{R}(\boldsymbol{\theta}) - \mathcal{R}(\widehat{\boldsymbol{\theta}}). \tag{A.9}$$

Combining (A.1), (A.7), (A.9), and (A.2), we eventually obtain

$$\mathbb{E}_{\mathcal{B}}||\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})||^2 \le \frac{4T_{max}}{|\mathcal{B}|}\left[\mathcal{P}(\boldsymbol{\theta}^{(t-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}}) + \mathcal{P}(\widetilde{\boldsymbol{\theta}}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right]. \tag{A.10}$$

□

B Proof of Theorem 4.2

Before we proceed with the proof, we first define

$$\mathcal{T}_\eta(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}'} \frac{1}{2\eta}||\boldsymbol{\theta}' - \boldsymbol{\theta}||^2 + \mathcal{R}(\boldsymbol{\theta}') = \operatorname*{argmin}_{\boldsymbol{\theta}'} \frac{1}{2\eta}\sum_{j=1}^k ||\boldsymbol{\theta}'_{\mathcal{G}_j} - \boldsymbol{\theta}_{\mathcal{G}_j}||^2 + \sum_{j=1}^k r_j(\boldsymbol{\theta}'_{\mathcal{G}_j}),$$

where the last equality comes from the assumption that $\mathcal{R}(\boldsymbol{\theta}')$ is block separable. Then by Assumption 2.3, we have

$$\mathcal{T}_\eta(\boldsymbol{\theta}) = (\mathcal{T}_\eta^1(\boldsymbol{\theta}_{\mathcal{G}_1})^T, \cdots, \mathcal{T}_\eta^k(\boldsymbol{\theta}_{\mathcal{G}_k})^T)^T.$$

For any vector $\boldsymbol{v} \in \mathbb{R}^d$, we define $\boldsymbol{v}^{\mathcal{G}_j} = (\mathbf{0}^T, ..., \boldsymbol{v}_{\mathcal{G}_j}^T, ...\mathbf{0}^T)^T$. It is easy to verify $\boldsymbol{v} = \sum_{j=1}^k \boldsymbol{v}^{\mathcal{G}_j}$, and $\boldsymbol{v}^{\mathcal{G}_j}$ and $\boldsymbol{v}^{\mathcal{G}_{j'}}$ are orthogonal to each other, for any $j \neq j'$, i.e., $(\boldsymbol{v}^{\mathcal{G}_j})^T\boldsymbol{v}^{\mathcal{G}_{j'}} = 0$. We then define $\bar{\boldsymbol{\theta}} = \mathcal{T}_\eta(\boldsymbol{\theta} - \eta\boldsymbol{v})$ and

$$\bar{\boldsymbol{\theta}}^{\mathcal{G}_j} = \left(\boldsymbol{\theta}_{\mathcal{G}_1}^T, ..., [\mathcal{T}_\eta^j\left(\boldsymbol{\theta}_{\mathcal{G}_j} - \eta\boldsymbol{v}_{\mathcal{G}_j}\right)]^T, ..., \boldsymbol{\theta}_{\mathcal{G}_k}^T\right)^T.$$

We then introduce the following lemma:

Lemma B.1. *Define $\boldsymbol{\delta} = (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})/\eta$ and $\boldsymbol{\delta}^{\mathcal{G}_j} = (\bar{\boldsymbol{\theta}}^{\mathcal{G}_j} - \boldsymbol{\theta})/\eta$. If the block index j is randomly selected from $\{1, ..., k\}$ with equal probability, then we have*

$$\mathbb{E}_j\left[\boldsymbol{\delta}^{\mathcal{G}_j}\right] = \boldsymbol{\delta}/k \quad \textit{and} \quad \mathbb{E}_j||\boldsymbol{\delta}^{\mathcal{G}_j}||^2 = ||\boldsymbol{\delta}||^2/k.$$

Moreover, taking $\eta \leq 1/L_{\max}$, we have

$$\mathbb{E}_j\left[(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T\boldsymbol{\delta}^{\mathcal{G}_j} + \frac{\eta}{2}||\boldsymbol{\delta}^{\mathcal{G}_j}||^2\right] \leq \frac{1}{k}\mathcal{P}(\widehat{\boldsymbol{\theta}}) + \frac{k-1}{k}\mathcal{P}(\boldsymbol{\theta}) - \mathbb{E}_j[\mathcal{P}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j})] + \frac{1}{k}(\boldsymbol{v} - \nabla\mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}).$$

The proof of Lemma B.1 is presented in Appendix D.

Now we proceed with the proof of Theorem 4.2. At the t-th iteration of the inner loop, we randomly sample a mini-batch $\mathcal{B}$ and a block of coordinates $\mathcal{G}_j$. Define $\boldsymbol{\delta}_{\mathcal{B}}^{\mathcal{G}_j} = (\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)})/\eta$. We then have

$$\begin{aligned}
\mathbb{E}_{\mathcal{B},j}||\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{\theta}}||^2 &= \mathbb{E}_{\mathcal{B},j}||\boldsymbol{\theta}^{(t-1)} + \eta\boldsymbol{\delta}_{\mathcal{B}}^{\mathcal{G}_j} - \widehat{\boldsymbol{\theta}}||^2 \\
&= ||\boldsymbol{\theta}^{(t-1)} - \widehat{\boldsymbol{\theta}}||^2 + 2\eta\mathbb{E}_{\mathcal{B},j}[(\boldsymbol{\theta}^{(t-1)} - \widehat{\boldsymbol{\theta}})^T\boldsymbol{\delta}_{\mathcal{B}}^{\mathcal{G}_j}] + \eta^2\mathbb{E}_{\mathcal{B},j}||\boldsymbol{\delta}_{\mathcal{B}}^{\mathcal{G}_j}||^2 \\
&= ||\boldsymbol{\theta}^{(t-1)} - \widehat{\boldsymbol{\theta}}||^2 + 2\eta(\boldsymbol{\theta}^{(t-1)} - \widehat{\boldsymbol{\theta}})^T\mathbb{E}_{\mathcal{B},j}[\boldsymbol{\delta}_{\mathcal{B}}^{\mathcal{G}_j}] + \eta^2\mathbb{E}_{\mathcal{B},j}||\boldsymbol{\delta}_{\mathcal{B}}^{\mathcal{G}_j}||^2 \\
&= ||\boldsymbol{\theta}^{(t-1)} - \widehat{\boldsymbol{\theta}}||^2 + \mathbb{E}_{\mathcal{B}}\left[2\eta(\boldsymbol{\theta}^{(t-1)} - \widehat{\boldsymbol{\theta}})^T\mathbb{E}_j[\boldsymbol{\delta}_{\mathcal{B}}^{\mathcal{G}_j}] + \eta^2\mathbb{E}_j||\boldsymbol{\delta}_{\mathcal{B}}^{\mathcal{G}_j}||^2\right]. \qquad \text{(B.1)}
\end{aligned}$$

Define $\bar{\boldsymbol{\theta}}_{\mathcal{B}} = \mathcal{T}_\eta\left(\boldsymbol{\theta}^{(t-1)} - \eta\boldsymbol{v}_{\mathcal{B}}\right)$ and $\boldsymbol{v}_{\mathcal{B}} = \sum_{i\in\mathcal{B}}\nabla f_i(\boldsymbol{\theta}^{(t-1)}) - \nabla f_i(\widetilde{\boldsymbol{\theta}}) + \widetilde{\boldsymbol{\mu}}$. Then by applying Lemma B.1 to (B.1), we further have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B},j}||\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{\theta}}||^2 - ||\boldsymbol{\theta}^{(t-1)} - \widehat{\boldsymbol{\theta}}||^2 \\
&\leq \frac{2\eta}{k}\mathbb{E}_{\mathcal{B}}[(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}})] + 2\eta\mathbb{E}_{\mathcal{B}}\left[\frac{1}{k}\mathcal{P}(\widehat{\boldsymbol{\theta}}) + \frac{k-1}{k}\mathcal{P}(\boldsymbol{\theta}) - \mathbb{E}_j\mathcal{P}(\bar{\boldsymbol{\theta}}_{\mathcal{B}}^{\mathcal{G}_j})\right] \\
&\leq \frac{2\eta}{k}\mathbb{E}_{\mathcal{B}}[(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}}) \\
&\qquad + 2\eta\mathbb{E}_{\mathcal{B}}\left[\frac{1}{k}\mathcal{P}(\widehat{\boldsymbol{\theta}}) + \frac{k-1}{k}\mathcal{P}(\widehat{\boldsymbol{\theta}}) - \frac{k-1}{k}\mathcal{P}(\widehat{\boldsymbol{\theta}}) + \frac{k-1}{k}\mathcal{P}(\boldsymbol{\theta}) - \mathbb{E}_j\mathcal{P}(\bar{\boldsymbol{\theta}}_{\mathcal{B}}^{\mathcal{G}_j})\right] \\
&\leq \frac{2\eta}{k}\mathbb{E}_{\mathcal{B}}[(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}})] \\
&\qquad - 2\eta\mathbb{E}_{\mathcal{B}}\left[\mathbb{E}_j\mathcal{P}(\bar{\boldsymbol{\theta}}_{\mathcal{B}}^{\mathcal{G}_j}) - \mathcal{P}(\widehat{\boldsymbol{\theta}}) - \frac{k-1}{k}\left(\mathcal{P}(\widehat{\boldsymbol{\theta}}) - \mathcal{P}(\boldsymbol{\theta})\right)\right]. \qquad \text{(B.2)}
\end{aligned}$$

To bound $\mathbb{E}_{\mathcal{B}}[(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}})]$, we define $\bar{\boldsymbol{\theta}} = \mathcal{T}_\eta(\boldsymbol{\theta}^{(t-1)} - \eta\nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)}))$. Then we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}}(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}}) \\
&\quad= \mathbb{E}_{\mathcal{B}}\left[(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}} + \bar{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}})\right] \\
&\quad= \mathbb{E}_{\mathcal{B}}\left[(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) + (\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}))^T(\bar{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}})\right] \\
&\quad= \mathbb{E}_{\mathcal{B}}\left[(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)}))^T(\bar{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}})\right],
\end{aligned}$$

where the last equality comes from the fact $\mathbb{E}_{\mathcal{B}}[\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})] = \mathbf{0}$. By Cauchy-Schwarz inequality, we further have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}}[(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}})] \\
&\quad\le \mathbb{E}_{\mathcal{B}}||\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})|| \cdot ||\mathcal{T}_\eta(\boldsymbol{\theta} - \eta\nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})) - \mathcal{T}_\eta(\boldsymbol{\theta} - \eta\boldsymbol{v}_{\mathcal{B}})|| \\
&\quad\le \mathbb{E}_{\mathcal{B}}||\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})|| \cdot ||(\boldsymbol{\theta} - \eta\nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})) - (\boldsymbol{\theta} - \eta\boldsymbol{v}_{\mathcal{B}})|| \\
&\quad= \eta\mathbb{E}_{\mathcal{B}}||\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}^{(t-1)})||^2, \qquad \text{(B.3)}
\end{aligned}$$

where the second inequality comes from the non-expansiveness of the proximal operator [11]. Combining (B.2) and (B.3), we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B},j}||\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{\theta}}||^2 - ||\boldsymbol{\theta}^{(t-1)} - \widehat{\boldsymbol{\theta}}||^2 \\
&\le -2\eta\mathbb{E}_{\mathcal{B}}\left[\mathbb{E}_j\mathcal{P}(\bar{\boldsymbol{\theta}}_{\mathcal{B}}^{\mathcal{G}_j}) - \mathcal{P}(\widehat{\boldsymbol{\theta}}) - \frac{k-1}{k}\left(\mathcal{P}(\widehat{\boldsymbol{\theta}}) - \mathcal{P}(\boldsymbol{\theta})\right)\right] + \frac{2\eta^2}{k}\mathbb{E}_{\mathcal{B}}(\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{\mathcal{B}}) \\
&\le -2\eta\mathbb{E}_{\mathcal{B}}\left[\mathbb{E}_j\mathcal{P}(\bar{\boldsymbol{\theta}}_{\mathcal{B}}^{\mathcal{G}_j}) - \mathcal{P}(\widehat{\boldsymbol{\theta}}) - \frac{k-1}{k}\left(\mathcal{P}(\widehat{\boldsymbol{\theta}}) - \mathcal{P}(\boldsymbol{\theta})\right)\right] + \frac{2\eta^2}{k}\mathbb{E}_{\mathcal{B}}||\boldsymbol{v}_{\mathcal{B}} - \nabla\mathcal{F}(\boldsymbol{\theta})||^2 \\
&\le -2\eta\mathbb{E}_{\mathcal{B}}\left[\mathbb{E}_j\mathcal{P}(\bar{\boldsymbol{\theta}}_{\mathcal{B}}^{\mathcal{G}_j}) - \mathcal{P}(\widehat{\boldsymbol{\theta}}) - \frac{k-1}{k}\left(\mathcal{P}(\widehat{\boldsymbol{\theta}}) - \mathcal{P}(\boldsymbol{\theta})\right)\right] \\
&\qquad\qquad + \frac{8\eta^2 T_{max}}{k|\mathcal{B}|}\left(\mathcal{P}(\boldsymbol{\theta}^{(t-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})) + \mathcal{P}(\widetilde{\boldsymbol{\theta}}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right) \\
&\le -2\eta\left[\mathbb{E}_{\mathcal{B},j}\mathcal{P}(\bar{\boldsymbol{\theta}}_{\mathcal{B}}^{\mathcal{G}_j}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right] + 2\eta\frac{k-1}{k}\left(\mathcal{P}(\boldsymbol{\theta}^{(t-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right) \\
&\qquad\qquad + \frac{8T_{max}\eta^2}{k|\mathcal{B}|}\left(\mathcal{P}(\boldsymbol{\theta}^{(k-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})) + \mathcal{P}(\widetilde{\boldsymbol{\theta}}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right), \qquad \text{(B.4)}
\end{aligned}$$

where the third inequality comes from Lemma 4.1.

At the s-th iteration of the outer loop, we have

$$\boldsymbol{\theta}^{(0)} = \widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}^{(s-1)} \quad \text{and} \quad \widetilde{\boldsymbol{\theta}}^{(s)} = \frac{1}{m}\sum_{t=1}^{m}\boldsymbol{\theta}^{(t)}. \qquad \text{(B.5)}$$

Thus summing (B.4) over $t = 1, ..., m$ and taking expectation with respect to $\mathcal{B}$ and j over all iterations of the inner loop, we obtain

$$\begin{aligned}
&\mathbb{E}||\boldsymbol{\theta}^{(m)} - \widehat{\boldsymbol{\theta}}||^2 - ||\boldsymbol{\theta}^{(0)} - \widehat{\boldsymbol{\theta}}||^2 + 2\eta\sum_{t=1}^{m}\left(\mathbb{E}\mathcal{P}(\boldsymbol{\theta}^{(t)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right) \\
&\le \frac{8T_{max}\eta^2/|\mathcal{B}| + 2\eta(k-1)}{k}\sum_{t=1}^{m-1}\left(\mathbb{E}\mathcal{P}(\boldsymbol{\theta}^{(t)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right) + \frac{8T_{max}\eta^2(m+1)}{k|\mathcal{B}|}\left(\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right) \\
&\le \frac{8T_{max}\eta^2/|\mathcal{B}| + 2\eta(k-1)}{k}\sum_{t=1}^{m}\left(\mathbb{E}\mathcal{P}(\boldsymbol{\theta}^{(t)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right) \\
&\qquad\qquad + \frac{8T_{max}\eta^2(m+1)}{k|\mathcal{B}|}\left(\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right), \qquad \text{(B.6)}
\end{aligned}$$

where the last inequality comes from $\mathbb{E}\mathcal{P}(\boldsymbol{\theta}^{(t)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}}) \geq 0$. By rearranging (B.6), we obtain

$$\begin{aligned}
&2\eta\left(\frac{1-4\eta T_{max}/|\mathcal{B}|}{k}\right)\sum_{t=1}^{m}[\mathbb{E}\mathcal{P}(\boldsymbol{\theta}^{(t)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})] \\
&\leq ||\boldsymbol{\theta}^{(0)} - \widehat{\boldsymbol{\theta}}||^2 + \frac{8T_{max}\eta^2(m+1)}{k|\mathcal{B}|}\left(\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right) \\
&\leq \frac{2}{\mu}(\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})) + \frac{8T_{max}\eta^2(m+1)}{k|\mathcal{B}|}\left(\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})\right), \qquad \text{(B.7)}
\end{aligned}$$

where the last inequality comes from the strong convexity of $\mathcal{P}$ and $\widetilde{\boldsymbol{\theta}}^{(s-1)} = \boldsymbol{\theta}^{(0)}$. By the convexity of $\mathcal{P}$ again, we have

$$\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s)}) \leq \frac{1}{m}\sum_{t=1}^{m}\mathcal{P}(\boldsymbol{\theta}^{(t)}). \qquad \text{(B.8)}$$

Therefore combining (B.7) and (B.8), we obtain

$$\begin{aligned}
&2\eta\left(\frac{1-4\eta T_{max}/|\mathcal{B}|}{k}\right)m[\mathbb{E}\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})] \\
&\qquad\qquad \leq \left(\frac{2}{\mu} + \frac{8T_{max}\eta^2(m+1)}{k|\mathcal{B}|}\right)[\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})]. \qquad \text{(B.9)}
\end{aligned}$$

Define

$$\alpha = \left(\frac{k}{\mu\eta(1-4\eta T_{max}/|\mathcal{B}|)m} + \frac{4\eta T_{max}/|\mathcal{B}|(m+1)}{(1-4\eta T_{max}/|\mathcal{B}|)m}\right).$$

Then (B.9) implies

$$\mathbb{E}\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}}) \leq \alpha[\mathcal{P}(\widetilde{\boldsymbol{\theta}}^{(s-1)}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})]. \qquad \text{(B.10)}$$

By applying (B.10) recursively and setting $|\mathcal{B}| = T_{max}/L$, we complete the proof.

C Proof of Corollary 4.3

Proof. Note that choosing the suggested η, m, and $\mathcal{B}$ guarantees

$$\alpha = \frac{k}{\mu\eta(1-4\eta L_{\max})m} + \frac{4\eta L_{\max}(m+1)}{(1-4\eta L_{\max})m} \leq 2/3.$$

By Markov inequality, we have

$$\mathbb{P}\left(P(\widetilde{\boldsymbol{\theta}}^{(s)}) - P(\widehat{\boldsymbol{\theta}}) \geq \epsilon\right) \overset{\text{(i)}}{\leq} \frac{1}{\epsilon}\left(\mathbb{E}P(\widetilde{\boldsymbol{\theta}}^{(s)}) - P(\widehat{\boldsymbol{\theta}})\right) \overset{\text{(ii)}}{\leq} \frac{1}{\epsilon}(2/3)^s\left(P(\widetilde{\boldsymbol{\theta}}^{(0)}) - P(\boldsymbol{\theta}_*)\right) \leq \rho.$$

where (i) comes from Theorem 4.2, and (ii) comes from choosing the suggested s. □

D Proof of Lemma B.1

Proof. Since $\boldsymbol{\delta} = \sum_{j=1}^{k}\boldsymbol{\delta}^{\mathcal{G}_j}$, we directly have

$$\mathbb{E}_j[\boldsymbol{\delta}^{\mathcal{G}_j}] = \boldsymbol{\delta}/k. \qquad \text{(D.1)}$$

Since $\boldsymbol{v}^{\mathcal{G}_j}$ and $\boldsymbol{v}^{\mathcal{G}_j{}'}$ are orthogonal to each other for any $j \neq j'$, we have $||\boldsymbol{\delta}||^2 = \sum_{j=1}^{k}||\boldsymbol{\delta}^{\mathcal{G}_j}||^2$. Taking expectation over the block index j, we directly have

$$\mathbb{E}_j||\boldsymbol{\delta}^{\mathcal{G}_j}||^2 = ||\boldsymbol{\delta}||^2/k. \qquad \text{(D.2)}$$

Since $\mathcal{F}$ and $\mathcal{R}$ are convex, we have

$$\mathcal{R}(\widehat{\boldsymbol{\theta}}) \geq \mathcal{R}(\bar{\boldsymbol{\theta}}) + \boldsymbol{\xi}^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}), \qquad \text{(D.3)}$$

where $\boldsymbol{\xi} \in \partial\mathcal{R}(\bar{\boldsymbol{\theta}})$, and

$$\mathcal{F}(\widehat{\boldsymbol{\theta}}) \geq \mathcal{F}(\boldsymbol{\theta}) + \nabla\mathcal{F}(\boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{D.4}$$

Combining (D.3) and (D.4), we have

$$\mathcal{P}(\widehat{\boldsymbol{\theta}}) = \mathcal{F}(\widehat{\boldsymbol{\theta}}) + \mathcal{R}(\widehat{\boldsymbol{\theta}}) \geq \mathcal{F}(\boldsymbol{\theta}) + \nabla\mathcal{F}(\boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + R(\bar{\boldsymbol{\theta}}) + \boldsymbol{\xi}^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}). \tag{D.5}$$

Since $\mathcal{R}$ is block separable, we have

$$\begin{aligned}
\mathbb{E}_j\mathcal{R}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) &= \frac{1}{k}\sum_{j=1}^{k}\mathcal{R}(\boldsymbol{\theta} + \eta\boldsymbol{\delta}^{\mathcal{G}_j}) \\
&= \frac{1}{k}\sum_{j=1}^{k}\left(\sum_{l\neq j} r_l(\boldsymbol{\theta}_{\mathcal{G}_l}) + r_j(\boldsymbol{\theta}_{\mathcal{G}_j} + \eta\boldsymbol{\delta}_{\mathcal{G}_j})\right) \\
&= \frac{1}{k}\left[(k-1)\sum_{j=1}^{k} r_j(\boldsymbol{\theta}_{\mathcal{G}_j}) + \sum_{j=1}^{k} r_j(\boldsymbol{\theta}_{\mathcal{G}_j} + \eta\boldsymbol{\delta}_{\mathcal{G}_j})\right] \\
&= \frac{k-1}{k}\mathcal{R}(\boldsymbol{\theta}) + \frac{1}{k}\mathcal{R}(\bar{\boldsymbol{\theta}}).
\end{aligned} \tag{D.6}$$

By Assumption 2.1, we have

$$\begin{aligned}
\mathbb{E}_j\mathcal{F}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) &\overset{\text{(i)}}{\leq} \mathcal{F}(\boldsymbol{\theta}) + \mathbb{E}_j\left(\nabla\mathcal{F}(\boldsymbol{\theta})^T(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j} - \boldsymbol{\theta}) + \frac{L_{max}}{2}||\bar{\boldsymbol{\theta}}^{\mathcal{G}_j} - \boldsymbol{\theta}||^2\right) \\
&\overset{\text{(ii)}}{=} \mathcal{F}(\boldsymbol{\theta}) + \mathbb{E}_j\left(\eta\nabla\mathcal{F}(\boldsymbol{\theta})^T\boldsymbol{\delta}^{\mathcal{G}_j} + \frac{\eta^2 L_{max}}{2}||\boldsymbol{\delta}^{\mathcal{G}_j}||^2\right) \\
&\overset{\text{(iii)}}{=} \mathcal{F}(\boldsymbol{\theta}) + \frac{1}{k}\left(\eta\nabla\mathcal{F}(\boldsymbol{\theta})^T\boldsymbol{\delta} + \frac{\eta^2 L_{max}}{2}||\boldsymbol{\delta}||^2\right) \\
&= \frac{k-1}{k}\mathcal{F}(\boldsymbol{\theta}) + \frac{1}{k}\left(\mathcal{F}(\boldsymbol{\theta}) + \eta\nabla\mathcal{F}(\boldsymbol{\theta})^T\boldsymbol{\delta} + \frac{\eta^2 L_{max}}{2}||\boldsymbol{\delta}||^2\right),
\end{aligned} \tag{D.7}$$

where (i) comes from the fact that $\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}$ and $\boldsymbol{\theta}$ are identical except the j-th block of coordinates, (ii) comes from the definition of $\boldsymbol{\delta}^{\mathcal{G}_j}$, and (iii) comes from (D.1) and (D.2).

By rearranging (D.7), we have

$$\mathcal{F}(\boldsymbol{\theta}) \geq k\mathbb{E}_j\mathcal{F}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{F}(\boldsymbol{\theta}) - \nabla\mathcal{F}(\boldsymbol{\theta})^T(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \frac{\eta^2 L_{max}}{2}||\boldsymbol{\delta}||^2. \tag{D.8}$$

Combining (D.5), (D.6), and (D.8), we further have

$$\begin{aligned}
\mathcal{P}(\widehat{\boldsymbol{\theta}}) &\overset{\text{(i)}}{\geq} k\mathbb{E}_j\mathcal{F}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{F}(\boldsymbol{\theta}) - \nabla\mathcal{F}(\boldsymbol{\theta})^T(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \frac{\eta^2 L_{max}}{2}||\boldsymbol{\delta}||^2 \\
&\qquad + \nabla\mathcal{F}(\boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \mathcal{R}(\bar{\boldsymbol{\theta}}) + \boldsymbol{\xi}(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) \\
&= k\mathbb{E}_j\mathcal{F}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{F}(\boldsymbol{\theta}) + \nabla\mathcal{F}(\boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) - \frac{\eta^2 L_{max}}{2}||\boldsymbol{\delta}||^2 + \mathcal{R}(\bar{\boldsymbol{\theta}}) + \boldsymbol{\xi}(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) \\
&\overset{\text{(ii)}}{=} k\mathbb{E}_j\mathcal{F}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{F}(\boldsymbol{\theta}) + \nabla\mathcal{F}(\boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) - \frac{\eta^2 L_{max}}{2}||\boldsymbol{\delta}||^2 \\
&\qquad + k\mathbb{E}_j\mathcal{R}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{R}(\boldsymbol{\theta}) + \boldsymbol{\xi}(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) \\
&= k\mathbb{E}_j\mathcal{P}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)P(\boldsymbol{\theta}) - \frac{\eta^2 L_{max}}{2}||\boldsymbol{\delta}||^2 + (\nabla\mathcal{F}(\boldsymbol{\theta}) + \boldsymbol{\xi})^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) \\
&\overset{\text{(iii)}}{\geq} k\mathbb{E}_j\mathcal{P}(\bar{\boldsymbol{\theta}}_{ij}) - (k-1)P(\boldsymbol{\theta}) - \frac{\eta}{2}||\boldsymbol{\delta}||^2 + (\nabla\mathcal{F}(\boldsymbol{\theta}) + \boldsymbol{\xi})^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})
\end{aligned} \tag{D.9}$$

where (ii) comes from (D.8), (ii) comes from (D.6), and (iii) comes from $\eta \leq 1/L_{max}$.

By the definition of $\bar{\boldsymbol{\theta}}$, we have

$$\bar{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}'}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{\theta}' - (\boldsymbol{\theta} - \eta \boldsymbol{v})||^2 + \eta \mathcal{R}(\boldsymbol{\theta}'). \tag{D.10}$$

The optimality condition of (D.10) implies that there exists some $\boldsymbol{\xi} \in \partial \mathcal{R}(\bar{\boldsymbol{\theta}})$ satisfying

$$\bar{\boldsymbol{\theta}} - (\boldsymbol{\theta} - \eta \boldsymbol{v}) + \eta \boldsymbol{\xi} = \mathbf{0},$$

which implies $\boldsymbol{\xi} = -\boldsymbol{\delta} - \boldsymbol{v}$. Then by (D.9), we have

$$\begin{aligned}
\mathcal{P}(\widehat{\boldsymbol{\theta}}) &\geq k\mathbb{E}_j \mathcal{P}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{P}(\boldsymbol{\theta}) - \frac{\eta}{2}||\boldsymbol{\delta}||^2 + (\nabla \mathcal{F}(\boldsymbol{\theta}) + \boldsymbol{\xi})(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) \\
&= k\mathbb{E}_j \mathcal{P}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{P}(\boldsymbol{\theta}) - \frac{\eta}{2}||\boldsymbol{\delta}||^2 + (\nabla \mathcal{F}(\boldsymbol{\theta}) - \boldsymbol{\delta} - \boldsymbol{v})^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) \\
&= k\mathbb{E}_j \mathcal{P}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{P}(\boldsymbol{\theta}) - \frac{\eta}{2}||\boldsymbol{\delta}||^2 \\
&\qquad - (\boldsymbol{v} - \nabla \mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) - \boldsymbol{\delta}^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \boldsymbol{\delta}^T(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}).
\end{aligned} \tag{D.11}$$

Since $\boldsymbol{\theta} - \bar{\boldsymbol{\theta}} = \eta \boldsymbol{\delta}$, (D.11) further implies

$$\begin{aligned}
\mathcal{P}(\widehat{\boldsymbol{\theta}}) &\leq k\mathbb{E}_j \mathcal{P}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{P}(\boldsymbol{\theta}) - \frac{\eta}{2}||\boldsymbol{\delta}||^2 \\
&\qquad - (\boldsymbol{v} - \nabla \mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) - \boldsymbol{\delta}^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \eta||\boldsymbol{\delta}||^2 \\
&= k\mathbb{E}_j \mathcal{P}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{P}(\boldsymbol{\theta}) + \frac{\eta}{2}||\boldsymbol{\delta}||^2 - (\boldsymbol{v} - \nabla \mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) - \boldsymbol{\delta}^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
&= k\mathbb{E}_j \mathcal{P}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) - (k-1)\mathcal{P}(\boldsymbol{\theta}) + \frac{k\eta}{2}\mathbb{E}_j||\boldsymbol{\delta}^{\mathcal{G}_j}||^2 \\
&\qquad - (\boldsymbol{v} - \nabla \mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) - k\mathbb{E}_j(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\boldsymbol{\delta}^{\mathcal{G}_j},
\end{aligned} \tag{D.12}$$

where the last equality comes from (D.1) and (D.2). By rearranging (D.12), we obtain

$$\begin{aligned}
\mathbb{E}_j(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T\boldsymbol{\delta}^{\mathcal{G}_j} + \frac{\eta}{2}\mathbb{E}_j||\boldsymbol{\delta}^{\mathcal{G}_j}||^2 &\leq \frac{1}{k}\mathcal{P}(\widehat{\boldsymbol{\theta}}) - \mathbb{E}_j \mathcal{P}(\bar{\boldsymbol{\theta}}^{\mathcal{G}_j}) \\
&\qquad + \frac{k-1}{k}\mathcal{P}(\boldsymbol{\theta}) + \frac{1}{k}(\boldsymbol{v} - \nabla \mathcal{F}(\boldsymbol{\theta}))^T(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}),
\end{aligned} \tag{D.13}$$

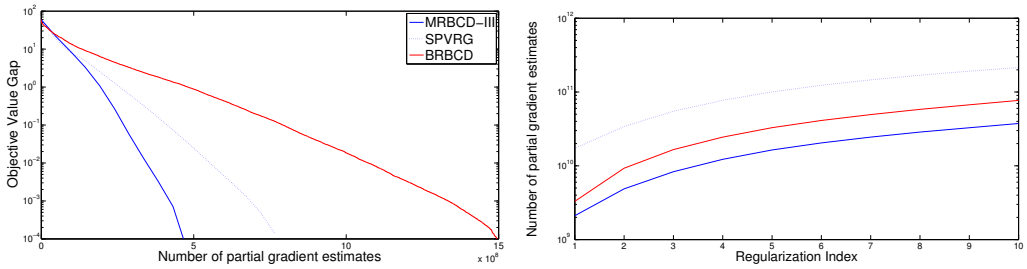
which completes the proof. □

E Numerical Simulations

F Real Data Experimental Results

Table E.1: Quantitive comparison of different methods on the simulated dataset for a sequence of regularization parameters. All three methods attains similar objective values for each regularization parameter, but MRBCD-III requires fewer partial gradient estimates than SPVRG and BRBCD.

MRBCD	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
# of P.G.	29.32e5	61.12e5	93.81e5	126.42e5	159.1e5	192.2e5	225.0e5	260.1e5	300.6e5	343.2e5
O.V.G.	9.23e-14	7.10e-14	7.45e-14	7.99e-14	7.81e-14	4.97e-14	4.61e-14	6.39e-14	4.26e-14	3.90e-14
Reg.	λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	λ_{20}
# of P.G .	387.9e5	433.4e5	478.0e5	522.7e5	566.9e5	610.2e5	653.0e5	695.5e5	738.0e5	780.0e5
O.V.G.	1.77e-14	2.48e-14	1.42e-14	3.55e-15	3.67e-15	4.67e-15	5.46e-15	5.57e-15	2.66e-15	1.78e-15
SPVRG	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
# of P.G.	270.6e5	548.4e5	817.8e5	1074e5	1328e5	1586e5	1845e5	2133e5	2441e5	2776e5
O.V.G.	8.57-14	9.43e-14	6.65e-14	9.12e-14	6.39e-14	4.97e-14	4.61e-14	6.39e-14	4.26e-14	0.461e-14
Reg.	λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	λ_{20}
# of P.G .	3113e5	3454e5	3791e5	4124e5	4456e5	4782e5	5106e5	5425e5	5741e5	6053e5
O.V.G.	3.90e-14	1.42e-14	3.19e-14	1.42e-14	8.88e-15	5.33e-15	3.55e-15	7.57e-15	4.44e-15	2.66e-15
BRBCD	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
# of P.G.	43.50e5	95.80e5	153.2e5	209.5e5	264.8e5	320.4e5	375.7e5	435.4e5	508.8e5	585.2e5
O.V.G.	5.68e-14	8.52e-14	7.81e-14	7.10e-14	7.10e-14	3.90e-14	4.26e-14	3.55e-14	5.68e-14	3.19e-14
Reg.	λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	λ_{20}
# of P.G .	663.8e5	743.2e5	820.8e5	897.2e5	974.2e5	1050e5	1126e5	1201e5	1275e5	1356e5
O.V.G.	7.11e-15	2.48e-14	1.42e-14	5.33e-15	3.55e-15	5.33e-15	3.55e-15	4.44e-15	1.78e-15	1.78e-15



(a) Comparison between different methods for a single regularization parameter.

(b) Comparison between different methods for a sequence of regularization parameters.

Figure F.1: [a] The vertical axis corresponds to objective value gaps $\mathcal{P}(\boldsymbol{\theta}) - \mathcal{P}(\widehat{\boldsymbol{\theta}})$ in log scale. The horizontal axis corresponds to numbers of partial gradient estimates. [b] The horizontal axis corresponds to indices of regularization parameters. The vertical axis corresponds to numbers of partial gradient estimates in log scale. We see that MRBCD attains the best performance among all methods for both settings