Supplementary

Anonymous

1. Since we have

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \arg\min_{\boldsymbol{\theta}} \bar{g}_t(\boldsymbol{\theta};\boldsymbol{\theta}_t) \\ &= \arg\min_{\boldsymbol{\theta}} \left\{ g_{\mathcal{S}_2^t}(\boldsymbol{\theta};\boldsymbol{\theta}_t) + \left(-\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \mathcal{V}_{t-1}\right)^\top (\boldsymbol{\theta}-\boldsymbol{\theta}_t) + \frac{\mu}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_t\|^2 \right\}.\end{aligned}$$

The gradient of $\bar{g}_t(\boldsymbol{\theta},\boldsymbol{\theta}_t)$ at $\boldsymbol{\theta}_{t+1}$ satisfies:

$$\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \mathcal{V}_{t-1} + \mu(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t) = 0,$$

then,

$$\begin{aligned}\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t &= -\frac{1}{\mu}(\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) + \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) + \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \mathcal{V}_{t-1}) \\ &= -\frac{1}{\mu}(\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) + \mathcal{V}_t).\end{aligned}$$

2. Suppose $\bar{g}_t(\boldsymbol{\theta},\boldsymbol{\theta}_t)$ is $\bar{\mu}$-smooth, which is reasonable as long as μ is large enough, so we have:

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\| &\le \frac{1}{\bar{\mu}}\|\nabla\bar{g}_t(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - \nabla\bar{g}_t(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t)\| \\ &= \frac{1}{\bar{\mu}}\|\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \mathcal{V}_{t-1} + \mu(\boldsymbol{\theta}_t-\boldsymbol{\theta}_t)\| \\ &= \frac{1}{\bar{\mu}}\|\mathcal{V}_t\|.\end{aligned}$$

1 Proof of Lemmas and Theorems

We aim to bound the iteration steps and gradient computations for attaining the first-order stationary point $\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\| \le \varepsilon$ in non-convex problems:

$$\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\|^2 = \mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi) - \mathcal{V}_\xi + \mathcal{V}_\xi\|^2 \le 2\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi) - \mathcal{V}_\xi\|^2 + 2\mathbb{E}\|\mathcal{V}_\xi\|^2. \tag{1}$$

We first consider the case when $\Phi_i(\boldsymbol{\theta})$ is L-smooth.

By the property that each $\Phi_i(\boldsymbol{\theta})$ has L-Lipschitz continuous gradient, we have:

$$\left\|\nabla\Phi(\boldsymbol{\theta}) - \nabla\Phi(\widetilde{\boldsymbol{\theta}})\right\|^2 = \left\|\mathbb{E}_{i\in\mathcal{S}}\left(\nabla\Phi_i(\boldsymbol{\theta}) - \nabla\Phi_i(\widetilde{\boldsymbol{\theta}})\right)\right\|^2 \leqslant \mathbb{E}_{i\in\mathcal{S}}\left\|\nabla\Phi_i(\boldsymbol{\theta}) - \nabla\Phi_i(\widetilde{\boldsymbol{\theta}})\right\|^2 \leqslant L^2\left\|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right\|^2$$

So $\Phi(\boldsymbol{\theta})$ also has L-Lipschitz continuous gradient.

Lemma 1. *Suppose Assumption 3.1 holds, and a sequence $\{\theta_{n_tp}\}$ is produced by Algorithm 2 after every p iterations. The base surrogate $g_{\mathcal{S}_2^t}(\boldsymbol{\theta};\boldsymbol{\theta}_t)$ is L_f-smooth, $\alpha = \frac{1}{2\mu} - \frac{L_f}{2\mu\bar{\mu}^2} - \frac{L}{2\bar{\mu}^2} - \frac{L^2p}{2\bar{\mu}^2\mu|\mathcal{S}_2|}$, $\mathcal{V}_i = \nabla\bar{g}_i(\boldsymbol{\theta}_i;\boldsymbol{\theta}_i)$. Then the objective function $\Phi(\boldsymbol{\theta})$ after every p iterations is guaranteed to decrease in expectation:*

$$\mathbb{E}\Phi(\boldsymbol{\theta}_{n_tp}) - \mathbb{E}\Phi(\boldsymbol{\theta}_{(n_t-1)p}) \le -\sum_{i=(n_t-1)p}^{n_tp-1} \alpha\mathbb{E}\|\mathcal{V}_i\|^2.$$

The proof of Lemma 1 is part of Lemma 3, we defer it later.

Lemma 2. Under Assumption 1, let $n_t = [t/p]$ such that $(n_t-1)p \le t \le n_t p - 1$, $(n_t-1)p$ is the beginning of epoch n_t. Then the estimator $\mathcal{V}_k$ satisfies

$$\mathbb{E}\|\mathcal{V}_t - \nabla\Phi(\boldsymbol{\theta}_t)\|^2 \le \sum_{i=(n_t-1)p}^{t} \frac{L^2}{|\mathcal{S}_2|}\mathbb{E}\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\|^2 \le \sum_{i=(n_t-1)p}^{t} \frac{L^2}{|\mathcal{S}_2|\bar{\mu}^2}\mathbb{E}\|\mathcal{V}_i\|^2.$$

Proof:

$$\begin{aligned}
&\mathbb{E}\|\mathcal{V}_t - \nabla\Phi(\boldsymbol{\theta}_t)\|^2 \\
&= \mathbb{E}\left\|\nabla\bar{g}_t(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla\Phi(\boldsymbol{\theta}_t)\right\|^2 \\
&= \mathbb{E}\left\|\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_t)\right\|^2 \\
&= \mathbb{E}\|\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \nabla\bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_t) + \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2 \\
&\leqslant \frac{1}{|\mathcal{S}_2^t|}\mathbb{E}\|\nabla g_i(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla g_i(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_t) + \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2 + \|\nabla\bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2 \\
&\quad + 2\sum_{i\in\mathcal{S}_2^t}\mathbb{E} < \nabla g_i(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla g_i(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_t) + \nabla\Phi(\boldsymbol{\theta}_{t-1}), \nabla\bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_{t-1}) > \\
&= \frac{1}{|\mathcal{S}_2^t|}\mathbb{E}\|\nabla g_i(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla g_i(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_t) + \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2 + \|\nabla\bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2 \\
&\leqslant \frac{1}{|\mathcal{S}_2^t|}\mathbb{E}\|\nabla g_i(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla g_i(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1})\|^2 + \|\nabla\bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2 \\
&= \frac{1}{|\mathcal{S}_2^t|}\mathbb{E}\|\nabla\Phi_i(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla\Phi_i(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1})\|^2 + \|\nabla\bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2 \\
&\leqslant \frac{L^2}{|\mathcal{S}_2^t|}\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|^2 + \|\nabla\bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2 \\
&\leqslant \frac{L^2}{\bar{\mu}^2\,|\mathcal{S}_2^t|}\mathbb{E}\|\nabla\bar{g}_t(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t)\|^2 + \|\nabla\bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2 \\
&= \frac{L^2}{\bar{\mu}^2\,|\mathcal{S}_2^t|}\mathbb{E}\|\mathcal{V}_t\|^2 + \|\nabla\bar{g}_{t-1}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) - \nabla\Phi(\boldsymbol{\theta}_{t-1})\|^2
\end{aligned}$$

Since we have $|\mathcal{S}_2^t| = |\mathcal{S}_2^{t-1}| = ... = |\mathcal{S}_2^1| = |\mathcal{S}_2|$, and $\|\nabla\bar{g}_0(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0) - \nabla\Phi(\boldsymbol{\theta}_0)\|^2 = 0$. Telescoping inequality above from $i = t, ..., (n_t-1)p$, we have

$$\mathbb{E}\left\|\mathcal{V}_t - \nabla\Phi(\boldsymbol{\theta}_t)\right\|^2 \le \sum_{i=(n_t-1)p}^{t} \frac{L^2}{|\mathcal{S}_2|}\mathbb{E}\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\|^2 \quad \le \sum_{i=(n_t-1)p}^{t} \frac{L^2}{|\mathcal{S}_2|\,\bar{\mu}^2}\mathbb{E}\|\mathcal{V}_i\|^2$$

Lemma 3. Under Assumption 1, our new surrogate is $\bar{\mu}$-strongly convex and the base surrogate is L_f-smooth. If the parameters $\mu, \bar{\mu}, L_f, p$ and $\mathcal{S}_2$ are chosen satisfying

$$\alpha \triangleq \frac{1}{2\mu} - \frac{L_f}{2\mu\bar{\mu}^2} - \frac{L}{2\bar{\mu}^2} - \frac{L^2 p}{2\bar{\mu}^2\mu|\mathcal{S}_2|} > 0,$$

we have

$$\mathbb{E}\|\mathcal{V}_\xi\|^2 = \frac{1}{T}\sum_{i=1}^{T-1}\mathbb{E}\|\mathcal{V}_i\|^2 \le \frac{\Phi(\boldsymbol{\theta}_0) - \Phi^*}{T\alpha}.$$

Proof:

$$
\begin{aligned}
\Phi(\boldsymbol{\theta}_{t+1}) &\leqslant \Phi(\boldsymbol{\theta}_t) + \langle \nabla\Phi(\boldsymbol{\theta}_t), \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t \rangle + \frac{L}{2}\left\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\right\|^2 \\
&\leqslant \Phi(\boldsymbol{\theta}_t) - \frac{1}{\mu}\left\langle \nabla\Phi(\boldsymbol{\theta}_t), g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) + \mathcal{V}_t \right\rangle + \frac{L}{2\bar{\mu}^2}\left\|\mathcal{V}_t\right\|^2 \\
&\leqslant \Phi(\boldsymbol{\theta}_t) - \frac{1}{\mu}\left\langle \nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t + \mathcal{V}_t, g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) + \mathcal{V}_t \right\rangle + \frac{L}{2\bar{\mu}^2}\left\|\mathcal{V}_t\right\|^2 \\
&\leqslant \Phi(\boldsymbol{\theta}_t) - \frac{1}{\mu}[\left\langle \nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t, g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) \right\rangle + \langle \nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t, \mathcal{V}_t \rangle + \\
&\qquad \left\langle \mathcal{V}_t, g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) \right\rangle + \langle \nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t, \mathcal{V}_t \rangle + \langle \mathcal{V}_t, \mathcal{V}_t \rangle] + \frac{L}{2\bar{\mu}^2}\left\|\mathcal{V}_t\right\|^2 \\
&\leqslant \Phi(\boldsymbol{\theta}_t) - \frac{1}{\mu}[\left\langle \nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t, g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) \right\rangle + \langle \nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t, \mathcal{V}_t \rangle + \\
&\qquad \left\langle \mathcal{V}_t, g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) \right\rangle + \langle \nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t, \mathcal{V}_t \rangle] - \frac{1}{\mu}\left\|\mathcal{V}_t\right\|^2 + \frac{L}{2\bar{\mu}^2}\left\|\mathcal{V}_t\right\|^2 \\
&\leqslant \Phi(\boldsymbol{\theta}_t) + \frac{1}{2\mu}[\left\|\nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t\right\|^2 + \left\|\mathcal{V}_t\right\|^2 + \left\|g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t)\right\|^2 - \left\|g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t)\right\|^2] \\
&\qquad + \frac{1}{\mu}\left\|\mathcal{V}_t\right\|^2 + \frac{L}{2\bar{\mu}^2}\left\|\mathcal{V}_t\right\|^2 \\
&\leqslant \Phi(\boldsymbol{\theta}_t) + \frac{1}{2\mu}\left\|\nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t\right\|^2 + \frac{1}{2\mu}\left\|g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t)\right\|^2 - \frac{1}{2\mu}\left\|\mathcal{V}_t\right\|^2 + \frac{L}{2\bar{\mu}^2}\left\|\mathcal{V}_t\right\|^2 \\
&\leqslant \Phi(\boldsymbol{\theta}_t) + \frac{1}{2\mu}\left\|\nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t\right\|^2 + \frac{L_f^2}{2\mu}\left\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\right\|^2 - \frac{1}{2\mu}\left\|\mathcal{V}_t\right\|^2 + \frac{L}{2\bar{\mu}^2}\left\|\mathcal{V}_t\right\|^2 \\
&\leqslant \Phi(\boldsymbol{\theta}_t) + \frac{1}{2\mu}\left\|\nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t\right\|^2 + \frac{L_f^2}{2\mu\bar{\mu}^2}\left\|\mathcal{V}_t\right\|^2 - \frac{1}{2\mu}\left\|\mathcal{V}_t\right\|^2 + \frac{L}{2\bar{\mu}^2}\left\|\mathcal{V}_t\right\|^2 \\
&\leqslant \Phi(\boldsymbol{\theta}_t) + \frac{1}{2\mu}\left\|\nabla\Phi(\boldsymbol{\theta}_t) - \mathcal{V}_t\right\|^2 - (\frac{1}{2\mu} - \frac{L}{2\bar{\mu}^2} - \frac{L_f^2}{2\mu\bar{\mu}^2})\left\|\mathcal{V}_t\right\|^2
\end{aligned}
$$

Taking expectation on both sides of the above inequality yields that

$$
\begin{aligned}
&\mathbb{E}\Phi(\boldsymbol{\theta}_{t+1}) \\
&\leq \mathbb{E}\Phi\left(\boldsymbol{\theta}_t\right) + \frac{1}{2\mu}\mathbb{E}\left\|\nabla\Phi\left(\boldsymbol{\theta}_t\right) - \mathcal{V}_t\right\|^2 - \left(\frac{1}{2\mu} - \frac{L}{2\bar{\mu}^2} - \frac{L_f^2}{2\mu\bar{\mu}^2}\right)\mathbb{E}\left\|\mathcal{V}_t\right\|^2 \\
&\overset{(i)}{\leq} \mathbb{E}\Phi\left(\boldsymbol{\theta}_t\right) + \frac{1}{2\mu}\sum_{i=(n_t-1)p}^{t}\frac{L^2}{|S_2|}\mathbb{E}\left\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\right\|^2 + \frac{1}{2\mu}\mathbb{E}\left\|\mathcal{V}_{(n_t-1)p} - \nabla\Phi\left(\boldsymbol{\theta}_{(n_t-1)p}\right)\right\|^2 - \left(\frac{1}{2\mu} - \frac{L}{2\bar{\mu}^2} - \frac{L_f^2}{2\mu\bar{\mu}^2}\right)\mathbb{E}\left\|\mathcal{V}_t\right\|^2 \\
&\overset{(ii)}{=} \mathbb{E}\Phi\left(\boldsymbol{\theta}_t\right) + \frac{1}{2\mu\bar{\mu}^2}\sum_{i=(n_k-1)p}^{t}\frac{L^2}{|S_2|}\mathbb{E}\left\|\mathcal{V}_i\right\|^2 - \left(\frac{1}{2\mu} - \frac{L}{2\bar{\mu}^2} - \frac{L_f^2}{2\mu\bar{\mu}^2}\right)\mathbb{E}\left\|\mathcal{V}_t\right\|^2
\end{aligned}
$$

(i),(ii) are due to the result of Lemma 2, $(n_t-1)p$ is the first iteration of this epoch, $\mathbb{E}\|\mathcal{V}_{(n_t-1)p} - \nabla\Phi(\boldsymbol{\theta}_{(n_t-1)p})\|^2 = 0$. Telescoping the inequality above from $t = (n_t-1)p$ to t, $(n_t-1)p$ is the first iteration of n_t epoch, we have

$$
\begin{aligned}
&\mathbb{E}\Phi(\boldsymbol{\theta}_{t+1}) \\
&= \mathbb{E}\Phi\left(\boldsymbol{\theta}_t\right) + \frac{1}{2\mu\bar{\mu}^2} \sum_{j=(n_t-1)p}^{t} \sum_{i=(n_t-1)p}^{j} \frac{L^2}{|S_2|} \mathbb{E}\left\|\mathcal{V}_i\right\|^2 - \left(\frac{1}{2\mu} - \frac{L}{2\bar{\mu}^2} - \frac{L_f^2}{2\mu\bar{\mu}^2}\right) \sum_{j=(n_t-1)p}^{t} \mathbb{E}\left\|\mathcal{V}_j\right\|^2 \\
&\overset{(i)}{\le} \mathbb{E}\Phi\left(\boldsymbol{\theta}_t\right) + \frac{1}{2\mu\bar{\mu}^2} \sum_{j=(n_t-1)p}^{t} \sum_{i=(n_t-1)p}^{t} \frac{L^2}{|S_2|} \mathbb{E}\left\|\mathcal{V}_i\right\|^2 - \left(\frac{1}{2\mu} - \frac{L}{2\bar{\mu}^2} - \frac{L_f^2}{2\mu\bar{\mu}^2}\right) \sum_{j=(n_t-1)p}^{t} \mathbb{E}\left\|\mathcal{V}_j\right\|^2 \\
&\overset{(ii)}{\le} \mathbb{E}\Phi\left(\boldsymbol{\theta}_t\right) + \frac{L^2 p}{2\mu\bar{\mu}^2\left|S_2\right|} \sum_{i=(n_t-1)p}^{t} \mathbb{E}\left\|\mathcal{V}_i\right\|^2 - \left(\frac{1}{2\mu} - \frac{L}{2\bar{\mu}^2} - \frac{L_f^2}{2\mu\bar{\mu}^2}\right) \sum_{j=(n_t-1)p}^{t} \mathbb{E}\left\|\mathcal{V}_j\right\|^2 \\
&= \mathbb{E}\Phi\left(\boldsymbol{\theta}_t\right) - \left(\frac{1}{2\mu} - \frac{L}{2\bar{\mu}^2} - \frac{L_f^2}{2\mu\bar{\mu}^2} - \frac{L^2 p}{2\mu\bar{\mu}^2\left|S_2\right|}\right) \sum_{j=(n_t-1)p}^{k} \mathbb{E}\left\|\mathcal{V}_j\right\|^2 \\
&\overset{(iii)}{=} \mathbb{E}\Phi\left(\boldsymbol{\theta}_t\right) - \alpha \sum_{j=(n_t-1)p}^{t} \mathbb{E}\left\|\mathcal{V}_j\right\|^2
\end{aligned}
$$

where, (i), extends the summation from j to t. (ii) follows from the fact that $t \le n_t p - 1$. (iii) satisfies by setting $\alpha = \frac{1}{2\mu} - \frac{L}{2\bar{\mu}^2} - \frac{L_f^2}{2\mu\bar{\mu}^2} - \frac{L^2 p}{2\mu\bar{\mu}^2|S_2|}$. Then we obtain

$$
\begin{aligned}
&\mathbb{E}\Phi\left(\boldsymbol{\theta}_T\right) - \mathbb{E}\Phi\left(\boldsymbol{\theta}_0\right) \\
&= \left(\mathbb{E}\Phi\left(\boldsymbol{\theta}_p\right) - \mathbb{E}\Phi\left(\boldsymbol{\theta}_0\right)\right) + \left(\mathbb{E}\Phi\left(\boldsymbol{\theta}_{2p}\right) - \mathbb{E}\Phi\left(\boldsymbol{\theta}_p\right)\right) + \cdots + \left(\mathbb{E}\Phi\left(\boldsymbol{\theta}_T\right) - \mathbb{E}\Phi\left(\boldsymbol{\theta}_{(n_t-1)p}\right)\right) \\
&\le -\sum_{i=0}^{p-1}\left(\alpha\mathbb{E}\left\|\mathcal{V}_i\right\|^2\right) - \sum_{i=p}^{2p-1}\left(\alpha\mathbb{E}\left\|\mathcal{V}_i\right\|^2\right) - \cdots - \sum_{i=(n_T-1)p}^{T-1}\left(\alpha\mathbb{E}\left\|\mathcal{V}_i\right\|^2\right) \\
&= -\sum_{i=0}^{T-1} \alpha\mathbb{E}\left\|\mathcal{V}_i\right\|^2
\end{aligned}
$$

We thus have

$$
\begin{aligned}
\sum_{i=0}^{T-1} \alpha\mathbb{E}\left\|\mathcal{V}_i\right\|^2 &\le \Phi\left(\boldsymbol{\theta}_0\right) - \Phi(\boldsymbol{\theta}_T) \\
&\le \Phi\left(\boldsymbol{\theta}_0\right) - \Phi^*
\end{aligned}
$$

Finally, we get

$$
\mathbb{E}\|\mathcal{V}_\xi\|^2 = \frac{1}{T}\sum_{i=1}^{T-1} \mathbb{E}\|\mathcal{V}_i\|^2 \le \frac{\Phi(\boldsymbol{\theta}_0) - \Phi^*}{T\alpha}.
$$

Theorem 1. *Suppose Assumptions 3.1 holds and apply SPI-MM in Algorithm 2. Let* $p = \sqrt{n}$, $\mathcal{S}_2 = \sqrt{n}$ *and* μ *be large enough. Then we have final output satisfying* $\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\| \le \epsilon$ *as long as the total number of iterations* T *satisfies*

$$
T \ge \mathcal{O}\left(\frac{\Phi(\boldsymbol{\theta}_0) - \Phi^*}{\varepsilon^2}\right). \tag{2}
$$

And the total resulting IFO complexity is $\mathcal{O}(\sqrt{n}\varepsilon^{-2} + n)$.

Proof: From Lemma 2 and Lemma 3, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathcal{V}_\xi - \nabla\Phi(\boldsymbol{\theta}_\xi)\right\|^2 &\leqslant \sum_{i=(n_\xi-1)p}^{\xi} \frac{L^2}{|\mathcal{S}_2|\,\bar{\mu}^2}\mathbb{E}\|\mathcal{V}_i\|^2 \\
&\leqslant \sum_{i=(n_\xi-1)p}^{\min\{n_\xi p-1,T-1\}} \frac{L^2}{|\mathcal{S}_2|\,\bar{\mu}^2}\mathbb{E}\|\mathcal{V}_i\|^2 \\
&\leqslant \sum_{i=0}^{T-1} \frac{pL^2}{T\,|\mathcal{S}_2|\,\bar{\mu}^2}\mathbb{E}\|\mathcal{V}_i\|^2 \\
&\leqslant \frac{L^2 p}{T|\mathcal{S}_2|\alpha\bar{\mu}^2}\left(\Phi(\boldsymbol{\theta}_0) - \Phi^*\right)
\end{aligned}$$

Substituting $\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\|^2$ and $\mathbb{E}\|\mathcal{V}_\xi\|^2$ in inequality 1, we obtain,

$$\begin{aligned}
\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\|^2 &\leq 2\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\|^2 + 2\mathbb{E}\|\mathcal{V}_\xi\|^2 \\
&\leq \frac{2\Phi(\boldsymbol{\theta}_0) - \Phi^*}{T\alpha} + \frac{L^2 p}{T|\mathcal{S}_2|\alpha\bar{\mu}^2}\left(\Phi(\boldsymbol{\theta}) - \Phi^*\right) \\
&= \frac{2}{T\alpha}(1 + \frac{L^2 p}{|\mathcal{S}_2|\bar{\mu}^2})(\Phi(\boldsymbol{\theta}_0) - \Phi^*)
\end{aligned} \tag{3}$$

Choose $\mathcal{S}_2 = \sqrt{n}, p = \sqrt{n}$, and $\mu = L$, we have $\bar{\mu} = 2L$ and $\alpha = \frac{1}{8L} > 0$. Plugging related parameters into 3, we have,

$$\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\|^2 \leq \frac{40L}{T}(\Phi(\boldsymbol{\theta}_0) - \Phi^*)$$

Since we have $(\mathbb{E}\|\Phi(\boldsymbol{\theta}_\xi)\|)^2 \leq \mathbb{E}\|\Phi(\boldsymbol{\theta}_\xi)\|^2$ due to Jensen's inequality, we bound $\mathbb{E}\|\Phi(\boldsymbol{\theta}_\xi)\|^2 \leq \varepsilon^2$ in order to ensure $\mathbb{E}\|\Phi(\boldsymbol{\theta}_\xi)\| \leq \varepsilon$. Thus, $\frac{20L}{T}(\Phi(\boldsymbol{\theta}_0) - \Phi^*) \leq \varepsilon^2$, the total number of iterations T satisfies

$$T \geq \frac{20L}{\varepsilon^2}(\Phi(\boldsymbol{\theta}_0) - \Phi^*) = \mathcal{O}\left(\frac{\Phi(\boldsymbol{\theta}_0) - \Phi^*}{\varepsilon^2}\right)$$

The total *IFO* complexity is:

$$\left\lceil \frac{T}{p} \right\rceil \cdot n + T \cdot \mathcal{S}_2 \leq (T + p) \cdot \frac{n}{p} + T \cdot \mathcal{S}_2 = T\sqrt{n} + n + T\sqrt{n} = \mathcal{O}(\sqrt{n}\varepsilon^{-2} + n)$$