Supplemental Materials

In this section, we present the complete proofs of the above lemmas and theorems. In the beginning, we give some useful properties of CoordSGE and RandSGE, respectively.

Lemma 10 ((Liu et al. 2018))**.** *Suppose that the function $f(x)$ is L-smooth. Let $\hat{\nabla} f(x)$ denote the estimated gradient defined by CoordSGE. Define $f_\mu = \mathbb{E}_{u\sim U[-\mu,\mu]} f(x + ue_j)$, where $U[-\mu, \mu]$ denotes the uniform distribution on the interval $[-\mu, \mu]$. Then for any $x \in \mathbb{R}^d$ we have*

1. *f_μ is L-smooth, and $\hat{\nabla} f(x) = \sum_{j=1}^{d} \frac{\partial f_\mu(x)}{\partial x_j} e_j$.*
2. *$\left|f_\mu(x) - f(x)\right| \le \frac{L\mu^2}{2}$ and $\left|\frac{\partial f_\mu(x)}{\partial x_j}\right| \le \frac{L\mu^2}{2}$.*
3. *$\left\|\hat{\nabla} f(x) - \nabla f(x)\right\|^2 \le \frac{L^2 d^2 \mu^2}{4}$.*

Lemma 11. *Assume that the function $f(x)$ is L-smooth. Let $\hat{\nabla}_r f(x)$ denote the estimated gradient defined by RandSGE. Define $f_\mu = \mathbb{E}_{u\sim U_S}[f(x + \mu u)]$, where U is uniform distribution over a d-dimensional unit ball S. Then, we have*

1. *For any $x \in \mathbb{R}^d$, $\nabla f_\mu(x) = \mathbb{E}_u[\hat{\nabla}_r f(x, u)]$.*
2. *$\left|f_\mu(x) - f(x)\right| \le \frac{\mu^2 L}{2}$ and $\left\|f_\mu(x) - f(x)\right\| \le \frac{\mu L d}{2}$ for any $x \in \mathbb{R}^d$.*
3. *$\mathbb{E}_u \left\|\hat{\nabla}_r f(x, u) - \hat{\nabla}_r f(y, u)\right\|^2 \le 3dL^2 \|x - y\|^2 + \frac{3L^2 d^2 \mu^2}{2}$.*

Proof. The proof of items 1 and 2 can be found in (Gao, Jiang, and Zhang 2018). Item 3 is due to Lemma 5 in (Ji et al. 2019). □

Lemma 12. *For a given $x \in \mathbb{R}^d$, let $\overline{x} = Prox_{\eta h}(x - \eta v)$, then we have for all $w \in \mathbb{R}^d$*

$$\begin{aligned} F(\overline{x}) \le F(w) + \langle \nabla f(x) - v, \overline{x} - w\rangle - \frac{1}{\eta}\langle \overline{x} - x, \overline{x} - w\rangle \\ + \frac{L}{2}\|\overline{x} - x\|^2 + \frac{L}{2}\|w - x\|^2 \end{aligned} \tag{1}$$

Proof. First, we recall the proximal operator

$$\mathrm{Prox}_{\eta h}(x - \eta v) := \arg\min_{y\in\mathbb{R}^d} \left(h(y) + \frac{1}{2\eta}\|y - x\|^2 + \langle v, y\rangle \right) \tag{2}$$

For the nonsmooth function $h(x)$, for all $w \in \mathbb{R}^d$ we have

$$\begin{aligned} h(\overline{x}) &\le h(w) + \langle p, \overline{x} - w\rangle \\ &= h(w) - \left\langle v + \frac{1}{\eta}(\overline{x} - x), \overline{x} - w \right\rangle \end{aligned} \tag{3}$$

where $p \in \partial h(\overline{x})$ such that $p + \frac{1}{\eta}(\overline{x} - x) + v = 0$ according to the optimality condition of (2), and (3) due to the convexity of h. In addition, since since $f(x)$ is L-Lipschitz continuous, we have

$$f(\overline{x}) \le f(x) + \langle \nabla f(x), \overline{x} - x\rangle + \frac{L}{2}\|\overline{x} - x\|^2 \tag{4}$$

and

$$f(x) \le f(w) + \langle \nabla f(x), x - w\rangle + \frac{L}{2}\|w - x\|^2 \tag{5}$$

This lemma is obtained by adding (3), (4), (5), and using $F(x) = f(x) + h(x)$. □

Lemma 13. *Let $x_t^s = Prox_{\eta h}(x_{t-1}^s - \eta \hat{v}_{t-1}^s)$ and $\overline{x}_t^s$ be the proximal projection using full true gradient, i.e., $\overline{x}_t^s = Prox_{\eta h}(x_{t-1}^s - \eta \nabla f(x_{t-1}^s))$. Then the following inequality holds*

$$\left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \le \eta \left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2$$

Proof. Based on inequality (3) we obtain

$$h(x_t^s) \le h(\overline{x}_t^s) - \left\langle \hat{v}_{t-1}^s + \frac{1}{\eta}(x_t^s - x_{t-1}^s), x_t^s - \overline{x}_t^s \right\rangle \tag{6}$$

$$h(\overline{x}_t^s) \le h(x_t^s) - \left\langle \nabla f(x_{t-1}^s) + \frac{1}{\eta}(\overline{x}_t^s - x_{t-1}^s), \overline{x}_t^s - x_t^s \right\rangle \tag{7}$$

By summing (6) and (7), we have

$$\frac{1}{\eta}\left\langle x_t^s - \overline{x}_t^s, x_t^s - \overline{x}_t^s\right\rangle \leq \left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s\right\rangle$$

$$\frac{1}{\eta}\left\|x_t^s - \overline{x}_t^s\right\|^2 \leq \left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\| \left\|x_t^s - \overline{x}_t^s\right\| \tag{8}$$

where (8) holds by Cauchy-Schwarz inequality. Thus, we obtain

$$\left\|x_t^s - \overline{x}_t^s\right\| \leq \eta \left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\| \tag{9}$$

Now the proof is complete using Cauchy-Schwarz inequality and (9). □

Proof of Lemma 1

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2\right] \\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \hat{g}^s\right)\right\|^2\right] \\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1})\right)\right\|^2\right] \\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right) + \left(\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \\
=&\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right) + \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \\
\leq&2\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\hat{\nabla} f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right) + \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \\
&+ 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad (10) \\
=&2\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\hat{\nabla} f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right)\right\|^2\right] \\
&+ 2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] + 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad (11) \\
=&\frac{2\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\left(\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\hat{\nabla} f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right)\right\|^2\right] \\
&+ 2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] + 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad (12) \\
\leq&\frac{2\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right\|^2\right] \\
&+ 2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] + 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad (13) \\
\leq&\frac{3\eta L^2 d^2\mu^2}{b} + \frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] + 2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]
\end{aligned}$$

$$+ 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \tag{14}$$

$$\leq \frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] + 2\frac{I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 + \frac{3\eta L^2 d^2\mu^2}{b} \tag{15}$$

$$\leq \frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] + 2\frac{I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{L^2 d^2\mu^2}{2} + \frac{3\eta L^2 d^2\mu^2}{b} \tag{16}$$

$$\leq \frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] + 2\frac{I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2 d^2\mu^2}{2} \tag{17}$$

where, recalling that a deterministic gradient estimator is employed and the expectations are taking with respect to I_b and $I_{\mathcal{B}}$. The inequality (10) holds by the Jensen's inequality. (11) and (12) are due to $\mathbb{E}[\|x_1 + x_2 + \ldots + x_k\|^2] = \sum_{i=1}^k \mathbb{E}[\|x_i\|^2]$ if $x_1, x_2, \ldots, x_k$ are independent and of mean zero. Recall that I_b and $I_{\mathcal{B}}$ are also independent. (13) applies the fact that for any random variable z, $\mathbb{E}[\|z - \mathbb{E}[z]\|^2] \leq \mathbb{E}[\|z\|^2]$. (14) employs following inequality

$$\begin{aligned}
\mathbb{E}\left\|\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(\tilde{x}^s)\right\|^2 &= \mathbb{E}\left\|\hat{\nabla} f_i(x_t^s) - \nabla f_i(x_t^s) + \nabla f_i(x_t^s) - \hat{\nabla} f_i(\tilde{x}^s) + \nabla f_i(\tilde{x}^s) - \nabla f_i(\tilde{x}^s)\right\|^2 \\
&\leq 3\mathbb{E}\left\|\hat{\nabla} f_i(x_t^s) - \nabla f_i(x_t^s)\right\|^2 + 3\left\|\hat{\nabla} f_i(\tilde{x}^s) - \nabla f_i(\tilde{x}^s)\right\|^2 \\
&\quad + 3\left\|\nabla f_i(x_t^s) - \nabla f_i(\tilde{x}^s)\right\|^2 \\
&\leq \frac{3L^2 d^2\mu^2}{2} + 3L^2\left\|x_t^s - \tilde{x}^s\right\|^2
\end{aligned} \tag{18}$$

where the last inequality used the fact that f_{i,μ_j} is L-smooth. (15) is by Assumption 2 and (16) uses Lemma 10. The proof is now complete. □

Proof of Lemma 2

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s) - \tilde{v}_{t-1}^s\right\|^2\right] \\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_{t-1}^s) - \hat{\nabla}_r f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \hat{g}^s\right)\right\|^2\right] \\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_{t-1}^s) - \hat{\nabla}_r f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1})\right)\right\|^2\right] \\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_{t-1}^s) - \hat{\nabla}_r f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right) + \left(\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \\
=&\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla}_r f_i(x_{t-1}^s) - \hat{\nabla}_r f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right) + \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \\
\leq&2\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla}_r f_i(x_{t-1}^s) - \hat{\nabla}_r f_i(\tilde{x}^{s-1})\right) - \left(\hat{\nabla} f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right) + \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \\
&+ 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad (19) \\
=&2\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla}_r f_i(x_{t-1}^s) - \hat{\nabla}_r f_i(\tilde{x}^{s-1})\right) - \left(\hat{\nabla} f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right)\right\|^2\right] \\
&+ 2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] + 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad (20)
\end{aligned}$$

$$
\begin{aligned}
=&\frac{2\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\left(\left(\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right)-\left(\hat{\nabla}f(x_{t-1}^s)-\hat{\nabla}f(\tilde{x}^{s-1})\right)\right)\right\|^2\right] \\
&+2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla}f_j(\tilde{x}^{s-1})-\hat{\nabla}f(\tilde{x}^{s-1})\right)\right\|^2\right]+2\eta\mathbb{E}\left\|\hat{\nabla}f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad (21)\\
\leq&\frac{2\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right\|^2\right]+2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla}f_j(\tilde{x}^{s-1})-\hat{\nabla}f(\tilde{x}^{s-1})\right)\right\|^2\right] \\
&+2\eta\mathbb{E}\left\|\hat{\nabla}f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad (22)\\
\leq&\frac{3\eta L^2d^2\mu^2}{b}+\frac{6\eta L^2d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla}f_j(\tilde{x}^{s-1})-\hat{\nabla}f(\tilde{x}^{s-1})\right)\right\|^2\right] \\
&+2\eta\mathbb{E}\left\|\hat{\nabla}f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad (23)\\
\leq&\frac{6\eta L^2d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+2\eta\mathbb{E}\left\|\hat{\nabla}f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \\
&+\frac{3\eta L^2d^2\mu^2}{b} \qquad (24)\\
\leq&\frac{6\eta L^2d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} \\
&+\eta\frac{L^2d^2\mu^2}{2}+\frac{3\eta L^2d^2\mu^2}{b} \qquad (25)\\
\leq&\frac{6\eta L^2d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\eta\frac{7L^2d^2\mu^2}{2} \qquad (26)
\end{aligned}
$$

where, the expectations are taking with respect to I_b and $I_{\mathcal{B}}$ and random directions $\{u_i\}$ in (2). The inequality (19) holds by the Jensen's inequality. (20) and (21) are based on $\mathbb{E}[\|x_1+x_2+\ldots+x_k\|^2]=\sum_{i=1}^k\mathbb{E}[\|x_i\|^2]$ if $x_1,x_2,\ldots,x_k$ are independent and of mean zero. Note that I_b and $I_{\mathcal{B}}$ are also independent. (22) uses the fact that $\mathbb{E}[\|x-\mathbb{E}[x]\|^2]\leq\mathbb{E}[\|x\|^2]$, for any random variable x. (23) holds due to Lemma 11. (24) is by Assumption 2 and (25) is by Lemma 10. (26) uses $b\geq 1$. The proof is now complete. □

In the sequel we frequently use the following inequality

$$
\|x-z\|^2\leq(1+\frac{1}{\alpha})\|x-y\|^2+(1+\alpha)\|y-z\|^2,\forall\alpha>0 \qquad (27)
$$

Proof of Theorem 3

Proof. Now, we apply Lemma 12 to prove Theorem 3. Let $x_t^s=\mathrm{Prox}_{\eta h}(x_{t-1}^s-\eta\hat{v}_{t-1}^s)$ and $\overline{x}_t^s=\mathrm{Prox}_{\eta h}(x_{t-1}^s-\eta\nabla f(x_{t-1}^s))$. By letting $x^+=x_t^s$, $x=x_{t-1}^s$, $v=\hat{v}_{t-1}^s$ and $z=\overline{x}_t^s$ in (1), we have

$$
\begin{aligned}
F(x_t^s)\leq&F(\overline{x}_t^s)+\left\langle\nabla f(x_{t-1}^s)-\hat{v}_{t-1}^s,x_t^s-\overline{x}_t^s\right\rangle-\frac{1}{\eta}\left\langle x_t^s-x_{t-1}^s,x_t^s-\overline{x}_t^s\right\rangle \\
&+\frac{L}{2}\left\|x_t^s-x_{t-1}^s\right\|^2+\frac{L}{2}\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2. \qquad (28)
\end{aligned}
$$

Besides, by letting $x^+=\overline{x}_t^s$, $x=x_{t-1}^s$, $v=\nabla f(x_{t-1}^s)$ and $z=x=x_{t-1}^s$ in (1), we have

$$
\begin{aligned}
F(\overline{x}_t^s)\leq&F(x_{t-1}^s)-\frac{1}{\eta}\left\langle\overline{x}_t^s-x_{t-1}^s,\overline{x}_t^s-x_{t-1}^s\right\rangle+\frac{L}{2}\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2 \\
=&F(x_{t-1}^s)-(\frac{1}{\eta}-\frac{L}{2})\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2. \qquad (29)
\end{aligned}
$$

Combining (28) and (29) we have

$$
\begin{aligned}
F(x_t^s)\leq&F(x_{t-1}^s)+\frac{L}{2}\left\|x_t^s-x_{t-1}^s\right\|^2-\left(\frac{1}{\eta}-L\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+\left\langle\nabla f(x_{t-1}^s)-\hat{v}_{t-1}^s,x_t^s-\overline{x}_t^s\right\rangle \\
&-\frac{1}{\eta}\left\langle x_t^s-x_{t-1}^s,x_t^s-\overline{x}_t^s\right\rangle \\
=&F(x_{t-1}^s)+\frac{L}{2}\left\|x_t^s-x_{t-1}^s\right\|^2-\left(\frac{1}{\eta}-L\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+\left\langle\nabla f(x_{t-1}^s)-\hat{v}_{t-1}^s,x_t^s-\overline{x}_t^s\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&-\frac{1}{2\eta}\left(\left\|x_t^s-x_{t-1}^s\right\|^2+\|x_t^s-\overline{x}_t^s\|^2-\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2\right)\\
=&F(x_{t-1}^s)-(\frac{1}{2\eta}-\frac{L}{2})\left\|x_t^s-x_{t-1}^s\right\|^2-\left(\frac{1}{2\eta}-L\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+\left\langle\nabla f(x_{t-1}^s)-\hat{v}_{t-1}^s,x_t^s-\overline{x}_t^s\right\rangle\\
&-\frac{1}{2\eta}\|x_t^s-\overline{x}_t^s\|^2\\
\le&F(x_{t-1}^s)-(\frac{1}{2\eta}-\frac{L}{2})\left\|x_t^s-x_{t-1}^s\right\|^2-\left(\frac{1}{2\eta}-L\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+\left\langle\nabla f(x_{t-1}^s)-\hat{v}_{t-1}^s,x_t^s-\overline{x}_t^s\right\rangle\\
&-\frac{1}{8\eta}\left\|x_t^s-x_{t-1}^s\right\|^2+\frac{1}{6\eta}\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2 \qquad (30)\\
=&F(x_{t-1}^s)-(\frac{5}{8\eta}-\frac{L}{2})\left\|x_t^s-x_{t-1}^s\right\|^2-\left(\frac{1}{3\eta}-L\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+\left\langle\nabla f(x_{t-1}^s)-\hat{v}_{t-1}^s,x_t^s-\overline{x}_t^s\right\rangle\\
\le&F(x_{t-1}^s)-(\frac{5}{8\eta}-\frac{L}{2})\left\|x_t^s-x_{t-1}^s\right\|^2-\left(\frac{1}{3\eta}-L\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+\eta\left\|\nabla f(x_{t-1}^s)-\hat{v}_{t-1}^s\right\|^2 \qquad (31)
\end{aligned}
$$

where the second inequality uses (27) with $\alpha=3$ and the last inequality holds due to the Lemma 13.

Note that $x_t^s=\mathrm{Prox}_{\eta h}(x_{t-1}^s-\eta\hat{v}_{t-1}^s)$ is the iterated form in our algorithm. By taking the expectation with respect to all random variables in (31) we obtain

$$
\mathbb{E}[F(x_t^s)]\le\mathbb{E}\left[F(x_{t-1}^s)-(\frac{5}{8\eta}-\frac{L}{2})\left\|x_t^s-x_{t-1}^s\right\|^2-\left(\frac{1}{3\eta}-L\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+\eta\left\|\nabla f(x_{t-1}^s)-\hat{v}_{t-1}^s\right\|^2\right] \qquad (32)
$$

In (32), we further bound $\eta\left\|\nabla f(x_{t-1}^s)-\hat{v}_{t-1}^s\right\|^2$ using Lemma 1 to obtain

$$
\begin{aligned}
&\mathbb{E}[F(x_t^s)]\\
\le&\mathbb{E}\left[F(x_{t-1}^s)-(\frac{5}{8\eta}-\frac{L}{2})\left\|x_t^s-x_{t-1}^s\right\|^2-\left(\frac{1}{3\eta}-L\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2\right]\\
&+\frac{6\eta L^2}{b}\mathbb{E}\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2+\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\eta\frac{7L^2d^2\mu^2}{2}\\
=&\mathbb{E}\left[F(x_{t-1}^s)-(\frac{5}{8\eta}-\frac{L}{2})\left\|x_t^s-x_{t-1}^s\right\|^2-\left(\frac{\eta}{3}-L\eta^2\right)\left\|g_\eta(x_{t-1}^s)\right\|^2\right]\\
&+\frac{6\eta L^2}{b}\mathbb{E}\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2+\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\eta\frac{7L^2d^2\mu^2}{2} \qquad (33)\\
\le&\mathbb{E}\left[F(x_{t-1}^s)-\frac{1}{2t}(\frac{5}{8\eta}-\frac{L}{2})\left\|x_t^s-\tilde{x}^{s-1}\right\|^2-\left(\frac{\eta}{3}-L\eta^2\right)\left\|g_\eta(x_{t-1}^s)\right\|^2\right]\\
&+(\frac{6\eta L^2}{b}+\frac{1}{2t-1}(\frac{5}{8\eta}-\frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2+\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\eta\frac{7L^2d^2\mu^2}{2} \qquad (34)
\end{aligned}
$$

where recalling $\overline{x}_t^s:=\mathrm{Prox}_{\eta h}(x_{t-1}^s-\eta\nabla f(x_{t-1}^s))$, (33) is based on the definition of gradient mapping $g_\eta(x_{t-1}^s)$. (34) uses (27) by choosing $\alpha=2t-1$.

Taking a telescopic sum for $t=1,2,\ldots,m$ in epoch s from (34) and recalling that $x_m^s=\tilde{x}^s$ and $x_0^s=\tilde{x}^{s-1}$, we obtain

$$
\begin{aligned}
&\mathbb{E}[F(\tilde{x}^s)]\\
\le&\mathbb{E}\left[F(\tilde{x}^{s-1})-\sum_{t=1}^m\frac{1}{2t}(\frac{5}{8\eta}-\frac{L}{2})\left\|x_t^s-\tilde{x}^{s-1}\right\|^2-\left(\frac{\eta}{3}-L\eta^2\right)\sum_{t=1}^m\left\|g_\eta(x_{t-1}^s)\right\|^2\right]\\
&+\sum_{t=1}^m(\frac{6\eta L^2}{b}+\frac{1}{2t-1}(\frac{5}{8\eta}-\frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\\
&+\sum_{t=1}^m\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\sum_{t=1}^m\eta\frac{7L^2d^2\mu^2}{2}\\
\le&\mathbb{E}\left[F(\tilde{x}^{s-1})-\sum_{t=1}^{m-1}\frac{1}{2t}(\frac{5}{8\eta}-\frac{L}{2})\left\|x_t^s-\tilde{x}^{s-1}\right\|^2-\left(\frac{\eta}{3}-L\eta^2\right)\sum_{t=1}^m\left\|g_\eta(x_{t-1}^s)\right\|^2\right]\\
&+\sum_{t=2}^m(\frac{6\eta L^2}{b}+\frac{1}{2t-1}(\frac{5}{8\eta}-\frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{t=1}^{m}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \qquad (35)\\
=&\mathbb{E}\left[F(\tilde{x}^{s-1})-\left(\frac{\eta}{3}-L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right]\\
&-\sum_{t=1}^{m-1}\left((\frac{1}{2t}-\frac{1}{2t+1})(\frac{5}{8\eta}-\frac{L}{2})-\frac{6\eta L^2}{b})\right)\mathbb{E}\left\|x_t^s-\tilde{x}^{s-1}\right\|^2\\
&+\sum_{t=1}^{m}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2}\\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1})-\left(\frac{\eta}{3}-L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right]\\
&-\sum_{t=1}^{m-1}\left(\frac{1}{6t^2}(\frac{5}{8\eta}-\frac{L}{2})-\frac{6\eta L^2}{b})\right)\mathbb{E}\left\|x_t^s-\tilde{x}^{s-1}\right\|^2\\
&+\sum_{t=1}^{m}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2}\\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1})-\frac{\eta}{6}\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right]+\sum_{t=1}^{m}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \qquad (36)
\end{aligned}
$$

where (35) holds since norm is always non-negative and $x_0^s=\tilde{x}^{s-1}$. In (36) we have used the fact that $(\frac{1}{6t^2}(\frac{5}{8\eta}-\frac{L}{2})-\frac{6\eta L^2}{b})\geq 0$ for all $1\leq t\leq m$ and $\frac{\eta}{6}\leq\frac{\eta}{3}-L\eta^2$ since $\eta=\min\{\frac{1}{8L},\frac{\sqrt{b}}{12mL}\}$. Telescoping the sum for $s=1,2,\ldots,S$ in (36), we obtain

$$
\begin{aligned}
&0\leq\mathbb{E}[F(\tilde{x}^S)-F(x^*)]\\
&\leq\mathbb{E}\left[F(\tilde{x}^0)-F(x^*)-\sum_{s=1}^{S}\sum_{t=1}^{m}\frac{\eta}{6}\left\|g_\eta(x_{t-1}^s)\right\|^2+\sum_{s=1}^{S}\sum_{t=1}^{m}(\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\eta\frac{7L^2d^2\mu^2}{2})\right]
\end{aligned}
$$

Thus, we have

$$
\mathbb{E}[\left\|g_\eta(\hat{x})\right\|^2]\leq\frac{6\left(F(x_0)-F(x^*)\right)}{\eta Sm}+\frac{I(\mathcal{B}<n)12\sigma^2}{\mathcal{B}}+21L^2d^2\mu^2 \qquad (37)
$$

where (37) holds since we choose $\hat{x}$ uniformly randomly from $\{x_{t-1}^s\}_{t\in[m],s\in[S]}$. □

Proof of Corollary 4

Proof. Using Theorem 3, we have $\eta=\min\{\frac{1}{8L},\frac{\sqrt{b}}{12mL}\}$

$$
\begin{aligned}
\mathbb{E}[\left\|g_\eta(\hat{x})\right\|^2]\leq&\frac{6\left(F(x_0)-F(x^*)\right)}{\eta Sm}\\
&+\frac{I(\mathcal{B}<n)12\sigma^2}{\mathcal{B}}+21L^2d^2\mu^2=3\epsilon \qquad (38)
\end{aligned}
$$

Now we derive the total number of iterations $T=Sm=\frac{6(F(x_0)-F(x^*))}{\epsilon\eta}$. Since $\mu\leq\frac{\sqrt{\epsilon}}{5dL}$, and for $\mathcal{B}=n$, the second term in the bound (38) is 0, the proof is completed as the number of SZO call equals to $Sn+Smb=6\left(F(x_0)-F(x^*)\right)(\frac{n}{\epsilon\eta m}+\frac{b}{\epsilon\eta})$. If $\mathcal{B}<n$ the number of SZO calls equal to $d(S\mathcal{B}+Smb)=6d\left(F(x_0)-F(x^*)\right)(\frac{\mathcal{B}}{\epsilon\eta m}+\frac{b}{\epsilon\eta})$ by noting that $\frac{I(\mathcal{B}<n)12\sigma^2}{\mathcal{B}}\leq\epsilon$ due to $\mathcal{B}\geq 12\sigma^2/\epsilon$. The second part of corollary is obtained by setting $m=\sqrt{b}$ in the first part. □

Corollary 14. *We set the batch size $\mathcal{B}=\min\{12\sigma^2/\epsilon,n\}$ and the smoothing parameter $\mu\leq\frac{\sqrt{\epsilon}}{5dL}$. Suppose $\hat{x}$ returned by Algorithm 1 is an ϵ-accurate solution for problem* (1)*. Recalling that CoordSGE and RandSGE require $O(d)$ and $O(1)$ function queries respectively, the number of SZO calls is at most*

$$
\begin{aligned}
(dS\mathcal{B}+Smb)=&6\left(F(x_0)-F(x^*)\right)(\frac{\mathcal{B}d}{\epsilon\eta m}+\frac{b}{\epsilon\eta})\\
=&O\left(\frac{\mathcal{B}d}{\epsilon\eta m}+\frac{b}{\epsilon\eta}\right) \qquad (39)
\end{aligned}
$$

and the number of PO calls is equal to $T = Sm = \frac{6(F(x_0)-F(x^*))}{\epsilon\eta} = O\left(\frac{1}{\epsilon\eta}\right)$. *In particular, by setting* $m = \sqrt{b}$ *and* $\eta = \frac{1}{12L\sqrt{d}}$, *the number of ZO calls is at most*

$$\begin{aligned} &72L(F(x_0) - F(x^*))\left(\frac{\mathcal{B}d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right) \\ &= O\left(s_n\frac{d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right) \end{aligned} \tag{40}$$

where $s_n = \min\{n, \frac{1}{\epsilon}\}$ *and the number of PO queries is equal to* $T = Sm = S\sqrt{b} = \frac{72L\sqrt{d}(F(x_0)-F(x^*))}{\epsilon} = O\left(\frac{\sqrt{d}}{\epsilon}\right)$.

Proof of Theorem 7

Proof. We start by recalling inequality (33) from the proof of Theorem 3, i.e.,

$$\begin{aligned} &\mathbb{E}[F(x_t^s)] \\ \leq&\mathbb{E}\left[F(x_{t-1}^s) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\ &+ (\frac{2\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{L^2d^2\mu^2}{2} \\ \leq&\mathbb{E}\left[F(x_{t-1}^s) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \frac{\eta}{6}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\ &+ (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \end{aligned} \tag{41}$$

where in (41) inequality we applied $\eta L \leq \frac{1}{6}$. Moreover, substituting PL inequality, i.e.,

$$\left\|g_\eta(x)\right\|^2 \geq 2\lambda(F(x) - F^*) \tag{42}$$

into (41), we obtain

$$\begin{aligned} &\mathbb{E}[F(x_t^s)] \\ \leq&\mathbb{E}\left[F(x_{t-1}^s) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \lambda\frac{\eta}{3}(F(x_{t-1}^s) - F^*)\right] \\ &+ (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \end{aligned} \tag{43}$$

Thus, we have

$$\begin{aligned} &\mathbb{E}[F(x_t^s)] \\ \leq&\mathbb{E}\left[(1 - \lambda\frac{\eta}{3})(F(x_{t-1}^s) - F^*) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\ &+ (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \end{aligned} \tag{44}$$

Let $\beta := 1 - \lambda\frac{\eta}{3}$ and $\Psi_t^s := \frac{\mathbb{E}[F(x_t^s)-F^*]}{\beta^t}$. Combining these definitions with (44), we have

$$\begin{aligned} &\Psi_t^s \\ \leq&\Psi_{t-1}^s - \frac{1}{\beta^t}\mathbb{E}\left[\frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\ &+ \frac{1}{\beta^t}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1}{\beta^t}\eta\frac{7L^2d^2\mu^2}{2} \end{aligned} \tag{45}$$

Similar to the proof of Theorem 3, summing (45) for $t = 1, 2, \ldots, m$ in epoch s and remembering that $x_m^s = \tilde{x}^s$ and $x_0^s = \tilde{x}^{s-1}$, we have

$$\mathbb{E}[F(\tilde{x}^s) - F^*]$$

$$
\begin{aligned}
&\leq \beta^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \beta^m \sum_{t=1}^{m} \frac{1}{\beta^t} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \beta^m \sum_{t=1}^{m} \frac{1}{\beta^t} \eta \frac{7L^2 d^2 \mu^2}{2} \\
&-\beta^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{1}{2t\beta^t} (\frac{5}{8\eta} - \frac{L}{2}) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^{m} \frac{1}{\beta^t} (\frac{6\eta L^2}{b} + \frac{1}{2t-1} (\frac{5}{8\eta} - \frac{L}{2})) \left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
&\leq \beta^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^m}{1-\beta} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^m}{1-\beta} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&-\beta^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{1}{2t\beta^t} (\frac{5}{8\eta} - \frac{L}{2}) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^{m} \frac{1}{\beta^t} (\frac{6\eta L^2}{b} + \frac{1}{2t-1} (\frac{5}{8\eta} - \frac{L}{2})) \left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
&\leq \beta^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^m}{1-\beta} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^m}{1-\beta} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&-\beta^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{1}{2t\beta^t} (\frac{5}{8\eta} - \frac{L}{2}) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
&+\beta^m \mathbb{E}\left[\sum_{t=2}^{m} \frac{1}{\beta^t} (\frac{6\eta L^2}{b} + \frac{1}{2t-1} (\frac{5}{8\eta} - \frac{L}{2})) \left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \qquad (46) \\
&\leq \beta^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^m}{1-\beta} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^m}{1-\beta} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&-\beta^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{1}{\beta^{t+1}} \left((\frac{\beta}{2t} - \frac{1}{2t+1})(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b}\right) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
&\leq \beta^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^m}{1-\beta} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^m}{1-\beta} \frac{7\eta L^2 d^2 \mu^2}{2} \qquad (47)
\end{aligned}
$$

where (46) since $\|.\|^2$ always is non-negative and $x_0^s = \tilde{x}^{s-1}$. (47) holds since it is sufficient to show $(\frac{\beta}{2t} - \frac{1}{2t+1})(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b} \geq 0$, for all $t = 1, 2, \ldots, m$. It is easy to see that this inequality is valid due to $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{\gamma b}}{12mL}\}$, where $\gamma = 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} > 0$. Similarly, let $\tilde{\beta} = \beta^m$ and $\tilde{\Psi}^s = \frac{\mathbb{E}[F(\tilde{x}^s) - F^*]}{\tilde{\beta}^s}$. Substituting these definitions into (47), we have

$$
\tilde{\Psi}^s \leq \tilde{\Psi}^{s-1} + \frac{1}{\tilde{\beta}^s} \frac{1-\tilde{\beta}}{1-\beta} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1}{\tilde{\beta}^s} \frac{1-\tilde{\beta}}{1-\beta} \frac{7\eta L^2 d^2 \mu^2}{2} \qquad (48)
$$

Taking a telescopic sum from (48) for all epochs $1 \leq s \leq S$, we obtain

$$
\begin{aligned}
\mathbb{E}[F(\tilde{x}^S) - F^*] &\leq \tilde{\beta}^S \mathbb{E}[F(\tilde{x}^0) - F^*] + \tilde{\beta}^S \sum_{s=1}^{S} \frac{1}{\tilde{\beta}^s} \frac{1-\tilde{\beta}}{1-\beta} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \tilde{\beta}^S \sum_{s=1}^{S} \frac{1}{\tilde{\beta}^s} \frac{1-\tilde{\beta}}{1-\beta} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&= \beta^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{1-\tilde{\beta}^S}{1-\tilde{\beta}} \frac{1-\tilde{\beta}}{1-\beta} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\tilde{\beta}^S}{1-\tilde{\beta}} \frac{1-\tilde{\beta}}{1-\beta} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&\leq \beta^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{1}{1-\beta} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1}{1-\beta} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&= \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I(\mathcal{B} < n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2 d^2 \mu^2}{2\lambda} \qquad (49)
\end{aligned}
$$

where in (49) we recall that $\beta = 1 - \frac{\lambda\eta}{3}$. □

Proof of Corollary 8

Proof. From Theorem 7, we have

$$
\begin{aligned}
\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} &\mathbb{E}[F(\tilde{x}^0) - F^*] \\
&+ \frac{6I(\mathcal{B} < n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2 d^2 \mu^2}{2\lambda} = 3\epsilon
\end{aligned}
$$

which gives the total number of iterations $T = Sm = O(\frac{1}{\lambda\eta} \log \frac{1}{\epsilon})$ and is equal to the number of PO calls. Since $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4Ld}$, we have $\frac{21L^2 d^2 \mu^2}{2\lambda} \leq \epsilon$. The number of SZO calls is equal to $d(S\mathcal{B} + Smb) = O(\frac{\mathcal{B}d}{\lambda\eta m} \log \frac{1}{\epsilon} + \frac{bd}{\lambda\eta} \log \frac{1}{\epsilon})$. Note that if $\mathcal{B} < n$ then $\frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}} \leq \epsilon$ since

$\mathcal{B} \geq 6\sigma^2/\lambda\epsilon$. With $m = \sqrt{b}$ and $\eta = \frac{\sqrt{\gamma}}{12L}$, the number of PO queries to $T = Sm = O(\frac{1}{\lambda\sqrt{\gamma}} \log \frac{1}{\epsilon})$ and the number of SZO calls is equal to $d(S\mathcal{B} + Smb) = O(\frac{\mathcal{B}d}{\lambda\sqrt{\gamma}m} \log \frac{1}{\epsilon} + \frac{bd}{\lambda\sqrt{\gamma}} \log \frac{1}{\epsilon})$. □

Corollary 15

Corollary 15. *Suppose the final iteration point $\tilde{x}^S$ in Algorithm 1 satisfies $\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \epsilon$ under PL condition. Under Assumptions 1 and 2, we set batch size $\mathcal{B} = \min\{\frac{6\sigma^2}{\lambda\epsilon}, n\}$ and the smoothing parameter $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4Ld}$. The number of SZO calls is bounded by*

$$(S\mathcal{B}d + Smb) = O(\frac{s_n d}{\lambda\eta m} \log \frac{1}{\epsilon} + \frac{b}{\lambda\eta} \log \frac{1}{\epsilon})$$

where $s_n = \min\{n, \frac{1}{\lambda\epsilon}\}$. The number of PO calls is equal to the total number of iterations T which is given by

$$T = Sm = O(\frac{1}{\lambda\eta} \log \frac{1}{\epsilon})$$

In particular, given the setting $m = \sqrt{b}$ and $\eta = \frac{\sqrt{\gamma}}{12L\sqrt{d}}$, the number of SZO calls simplifies to $(S\mathcal{B}d + Smb) = O(\frac{\mathcal{B}d\sqrt{d}}{\lambda\sqrt{\gamma}m} \log \frac{1}{\epsilon} + \frac{b\sqrt{d}}{\lambda\sqrt{\gamma}} \log \frac{1}{\epsilon})$.

Proof. From Theorem 7, we have

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I(\mathcal{B} < n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2d^2\mu^2}{2\lambda} = 3\epsilon \tag{50}$$

which gives the total number of iterations $T = Sm = O(\frac{1}{\lambda\eta} \log \frac{1}{\epsilon})$ and equals to the number of PO calls. Since $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4Ld}$, we have $\frac{21L^2d^2\mu^2}{2\lambda} \leq \epsilon$. The number of SZO calls equals to $(S\mathcal{B}d + Smb) = O(\frac{\mathcal{B}d}{\lambda\eta m} \log \frac{1}{\epsilon} + \frac{b}{\lambda\eta} \log \frac{1}{\epsilon})$. Note that if $\mathcal{B} < n$ then $\frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}} \leq \epsilon$ since $\mathcal{B} \geq 6\sigma^2/\lambda\epsilon$. With $m = \sqrt{b}$ and $\eta = \frac{\sqrt{\gamma}}{12L\sqrt{d}}$, the number of PO calls equals to $T = Sm = O(\frac{\sqrt{d}}{\lambda\sqrt{\gamma}} \log \frac{1}{\epsilon})$ and the number of SZO calls equals to $(S\mathcal{B}d + Smb) = O(\frac{\mathcal{B}d\sqrt{d}}{\lambda\sqrt{\gamma}m} \log \frac{1}{\epsilon} + \frac{b\sqrt{d}}{\lambda\sqrt{\gamma}} \log \frac{1}{\epsilon})$. □

References

Gao, X.; Jiang, B.; and Zhang, S. 2018. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing* 76(1):327–363.

Ji, K.; Wang, Z.; Zhou, Y.; and Liang, Y. 2019. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, 3100–3109.

Liu, S.; Kailkhura, B.; Chen, P.-Y.; Ting, P.; Chang, S.; and Amini, L. 2018. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 3727–3737.