Proximal Gradient Algorithm for Nonconvex Problems

March 2018

1 Introduction

2 Accelerated Proximal Gradient Method

Algorithm 1 Nonconvex ProxSVRG+

1: **Input:** initial point x_0, batch size B, minibatch size b, epoch length m, step size η
2: **Initialize:** $\tilde{x}^0 = x_0$
3: **for** $s = 1, 2, \ldots, S$ **do**
4: $\quad x_0^s = \widetilde{x}^{s-1}$
5: $\quad \hat{g}^s = \frac{1}{B}\sum_{j \in I_B} \nabla f_j(\widetilde{x}^{s-1})$
6: $\quad$ **for** $t = 1, 2, \ldots, m$ **do**
7: $\quad\quad \hat{v}_{t-1}^s = \frac{1}{b}\sum_{i \in I_b}\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) + \hat{g}^s$
8: $\quad\quad x_t^s = \mathrm{Prox}_{\eta h}(x_{t-1}^s - \eta \hat{v}_{t-1}^s)$
9: $\quad \widetilde{x}^s = x_m^s$
10: **Output:** $\hat{x}$ chosen uniformly from $\{x_{t-1}^s\}_{t \in [m], s \in [S]}$

Theorem 2.1. *If function f is convex, then by choosing $m = O(n)$, ZO-PROXSVRG achieves the following oracle complexity in expectation*

$$O\left(n\sqrt{\frac{F(x^0) - F(x^*)}{\epsilon}} + \sqrt{\frac{nL\|x^0 - x^*\|^2}{\epsilon}}\right).$$

This result implies that ZO-PROXSVRG attains the optimal convergence rate $O(1/T^2)$, where $T = S(m+n)$ is the total number of stochastic iterations.

Proof. We first impose the following constraint on η and θ

$$L\theta + \frac{L\theta}{1-\theta} \le \frac{1}{\eta}, \qquad \text{or equivalently } \eta \le \frac{1-\theta}{L\theta(2-\theta)}. \tag{1}$$

Algorithm 2 ZO-PROXSVRG for convex Optimization

Input: mini-batch size b, S, m and step size $\eta > 0$, parameter θ
Initialize: $\tilde{x}^0 = x_0^1 = x_0 \in \mathbb{R}^d$
for $s = 1, 2, \dots, S$ **do**
    $\mu_s = \hat{\nabla} f(\tilde{x}^s) = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_i(\tilde{x}^s), \theta = \frac{2}{s+4}, \eta = \frac{1}{4L\theta}$
    for $j = 1, \dots, m$ **do**
        Randomly pick up an i_j from $\{1, \dots, n\}$
        $y_{j-1} = \theta x_{j-1}^s + (1-\theta)\tilde{x}_{s-1}$
        $\hat{v}_j^s = \nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1}) + \mu_s$
        $x_j^s = \operatorname{argmin}_x \left\{ \frac{1}{2\eta}\left\|x - x_{j-1}^s\right\|^2 + \left\langle \hat{v}_j^s, x \right\rangle + g(x) \right\}$
    $\tilde{x}^s = \frac{\theta}{m}\sum_{j=1}^{m} x_j^s + (1-\theta)\tilde{x}^{s-1}$
    $x_0^{s+1} = x_m^s$
Output: $\tilde{x}_S$

We start with the convexity of f at y_{j-1}. By definition for any vector $u \in \mathbb{R}^d$, we have

$$f(y_{j-1}) - f(u) \leq \left\langle \nabla f(y_{j-1}), y_{j-1} - i \right\rangle$$
$$\frac{1-\theta}{\theta}\left\langle \nabla f(y_{j-1}), \tilde{x}^{s-1} - y_{j-1} \right\rangle + \left\langle \nabla f(y_{j-1}), x_{j-1} - u \right\rangle, \tag{2}$$

where (a) follows from the fact that $y_{j-1} = \theta x_{j-1} + (1-\theta)\tilde{x}^{s-1}$. Then we further expand $\left\langle \nabla f(y_{j-1}), x_{j-1} - u \right\rangle$ as

$$\left\langle \nabla f(y_{j-1}), x_{j-1} - u \right\rangle = \left\langle \nabla f(y_{j-1}) - \hat{v}_j^s, x_{j-1} - u \right\rangle + \left\langle \hat{v}_j^s, x_{j-1} - x_j \right\rangle + \left\langle \hat{v}_j^s, x_j - u \right\rangle. \tag{3}$$

Using L-smooth of f at (y_j, y_{j-1}), we get

$$\begin{aligned} f(y_j) - f(y_{j-1}) &\leq \left\langle \nabla f(y_{j-1}), y_j - y_{j-1} \right\rangle + \frac{L}{2}\left\|y_j - y_{j-1}\right\|^2 \\ &= \left[\left\langle \nabla f(y_{j-1}) - \hat{v}_j^s, x_j - x_{j-1} \right\rangle + \left\langle \hat{v}_j^s, x_j - x_{j-1} \right\rangle + \frac{L\theta^2}{2}\left\|x_j - x_{j-1}\right\|^2 \right] \end{aligned} \tag{4}$$

Equivalently we obtain

$$\left\langle \hat{v}_j^s, x_j - x_{j-1} \right\rangle \leq \frac{1}{\theta}(f(y_{j-1}) - f(y_j)) + \left\langle \nabla f(y_{j-1}) - \hat{v}_j^s, x_j - x_{j-1} \right\rangle + \frac{L\theta}{2}\left\|x_j - x_{j-1}\right\|^2. \tag{5}$$

Using the constraint (1) we have

$$\left\langle \hat{v}_j^s, x_{j-1} - x_j \right\rangle \leq \frac{1}{\theta}(f(y_{j-1}) - f(y_j)) + \left\langle \nabla f(y_{j-1}) - \hat{v}_j^s, x_j - x_{j-1} \right\rangle + \frac{1}{2\eta}\left\|x_j - x_{j-1}\right\|^2 - \frac{L\theta}{2(1-\theta)}\left\|x_j - x_{j-1}\right\|^2. \tag{6}$$

Then we can combine (2), (3), (6), as well as Lemma 3, which leads to

$$
\begin{aligned}
f(y_{j-1}) - f(u) \le \frac{1-\theta}{\theta}\left\langle \nabla f(y_{j-1}), \tilde{x}^{s-1} - y_{j-1}\right\rangle + \left\langle \nabla f(y_{j-1}) - \hat{v}_j^s, x_j - u\right\rangle + \frac{1}{\theta}(f(y_{j-1}) - f(y_j)) \\
- \frac{L\theta}{2(1-\theta)}\left\|x_j - x_{j-1}\right\|^2 + \frac{1}{2\eta}\left\|x_{j-1} - u\right\|^2 - \frac{1}{2\eta}\left\|x_j - u\right\|^2 + g(u) - g(x_j)
\end{aligned}
\tag{7}
$$

After taking expectation with respect to the sample i_j, we get

$$
\begin{aligned}
f(y_{j-1}) - f(u) \le \frac{1-\theta}{\theta}\left\langle \nabla f(y_{j-1}), \tilde{x}^{s-1} - y_{j-1}\right\rangle + \frac{1}{2\beta}\mathbb{E}_{i_j}[\left\|\nabla f(y_{j-1}) - \hat{v}_j^s\right\|^2] + \frac{\beta}{2}\mathbb{E}_{i_j}[\left\|x_j - u\right\|^2] \\
+ \frac{1}{\theta}(f(y_{j-1}) - f(y_j)) \\
- \frac{L\theta}{2(1-\theta)}\left\|x_j - x_{j-1}\right\|^2 + \frac{1}{2\eta}\left\|x_{j-1} - u\right\|^2 - \frac{1}{2\eta}\left\|x_j - u\right\|^2 + g(u) - g(x_j)
\end{aligned}
\tag{8}
$$

□

3 Convergence Analysis

Lemma 3.1. *Assume that the function $f(x)$ is L-smooth. Let $\hat{\nabla} f(x)$ denote the estimated gradient defined by* **CooSGE**. *Define $f_{\mu_j} = \mathbb{E}_{u\sim U[\mu_j,\mu_j]} f(x + ue_j)$, where $U[-\mu_j, \mu_j]$ denotes the uniform distribution at the interval $[\mu_j, \mu_j]$. Then we have 1) f_{μ_j} is L-smooth, and*

$$
\hat{\nabla} f(x) = \sum_{j=1}^{d} \frac{\partial f_{\mu_j}}{\partial x_j} e_j \tag{9}
$$

where $\partial f/\partial x_j$ denotes the partial derivative with respect to jth coordinate.

2) For $j \in [d]$,

$$
\left|f_{\mu_j}(x) - f(x)\right| \le \frac{L\mu_j^2}{2} \tag{10}
$$

$$
\left|\frac{\partial f_{\mu_j}(x)}{\partial x_j}\right| \le \frac{L\mu_j^2}{2} \tag{11}
$$

3) If $\mu = \mu_j$ for $j \in [d]$, then

$$
\left\|\hat{\nabla} f(x) - \nabla f(x)\right\|^2 \le \frac{L^2 d^2 \mu^2}{4} \tag{12}
$$

Lemma 3.2. *Assume that the function $f(x)$ is L-smooth. Let $\hat{\nabla} f(x)$ denote the estimated gradient defined by* **GauSGE**. *Define $f_\mu = \mathbb{E}_{u\sim\mathcal{N}(\prime,\mathcal{I})}[f(x + \mu u)]$. Then we have 1) For any $x \in \mathbb{R}^d$, $\nabla f_\mu(x) = \mathbb{E}_u[\hat{\nabla} f(x)]$.*

2) For any $x \in \mathbb{R}^d$,

$$\left|f_\mu(x) - f(x)\right| \le \frac{Ld\mu^2}{2} \tag{13}$$

$$\left|\nabla f_\mu(x) - \nabla f(x)\right| \le \frac{L\mu(d+3)^{\frac{3}{2}}}{2} \tag{14}$$

$$\mathbb{E}_u \left\|\hat{\nabla} f(x)\right\|^2 \le 2(d+4)\,\|\nabla f(x)\|^2 + \frac{\mu^2 L^2 (d+6)^3}{2} \tag{15}$$

3) For any $x \in \mathbb{R}^d$,

$$\mathbb{E}_u \left\|\hat{\nabla} f(x) - \nabla f(x)\right\|^2 \le 2(2d+9)\,\|\nabla f(x)\|^2 + \mu^2 L^2 (d+6)^3. \tag{16}$$

Lemma 3.3. *Let $x^+ = Prox_{\eta h}(x - \eta v)$, then the following inequality holds:*

$$\Phi(x^+) \le \Phi(z) + \langle \nabla f(x) - v, x^+ - z\rangle - \frac{1}{\eta}\langle x^+ - x, x^+ - z\rangle + \frac{L}{2}\|x^+ - x\|^2 + \frac{L}{2}\|z - x\|^2, \forall z \in \mathbb{R}^d. \tag{17}$$

Proof. First, we recall the proximal operator

$$\mathrm{Prox}_{\eta h}(x - \eta v) := \arg\min_{y \in \mathbb{R}^d} \left(h(y) + \frac{1}{2\eta}\|y - x\|^2 + \langle v, y\rangle\right) \tag{18}$$

For the nonsmooth function $h(x)$, we have

$$\begin{aligned} h(x^+) &\le h(z) + \langle p, x^+ - z\rangle \\ &= h(z) - \left\langle v + \frac{1}{\eta}(x^+ - x), x^+ - z\right\rangle \end{aligned} \tag{19}$$

where $p \in \partial h(x^+)$ such that $p + \frac{1}{\eta}(x^+ - x) + v = 0$ according to the optimality condition of (18), and (19) due to the convexity of h.

$$f(x^+) \le f(x) + \langle \nabla f(x), x^+ - x\rangle + \frac{L}{2}\|x^+ - x\|^2 \tag{20}$$

$$-f(z) \le -f(x) + \langle -\nabla f(x), z - x\rangle + \frac{L}{2}\|z - x\|^2 \tag{21}$$

where (20) holds since $f(x)$ has L-Lipschitz continuous gradient, and (21) holds since $-f(x)$ has the same L-Lipschitz continuous gradient as $f(x)$.

This lemma is proved by adding (19), (20), (21), and recalling $\Phi(x) = f(x) + h(x)$. □

Lemma 3.4.

Proof.

$$
\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(\widetilde{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(\widetilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\frac{1}{B}\sum_{j\in I_B}\hat{\nabla} f_j(\widetilde{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(\widetilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\hat{\nabla} f(\tilde{x}^{s-1})\right)+\left(\frac{1}{B}\sum_{j\in I_B}\hat{\nabla} f_j(\widetilde{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(\widetilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right)+\frac{1}{B}\sum_{j\in I_B}\left(\hat{\nabla} f_j(\widetilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]\\
=&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(\widetilde{x}^{s-1})\right)-\left(\hat{\nabla} f(x_{t-1}^s)-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right)+\frac{1}{B}\sum_{j\in I_B}\left(\hat{\nabla} f_j(\widetilde{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)\\
=&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(\widetilde{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}{B}\sum_{j\in I_B}\left(\hat{\nabla} f_j(\widetilde{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)\\
=&\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(\widetilde{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}{B}\sum_{j\in I_B}\left(\hat{\nabla} f_j(\widetilde{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 (24)\\
\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(\widetilde{x}^{s-1})\right\|^2\right]+2\eta\mathbb{E}\left[\left\|\frac{1}{B}\sum_{j\in I_B}\left(\hat{\nabla} f_j(\widetilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \qquad (25)\\
&+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad (26)\\
\leq&\frac{2\eta L^2 d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\widetilde{x}^{s-1}\right\|^2\right]+2\frac{I\{B<n\}\eta\sigma^2}{B}+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad (27)\\
\leq&\frac{2\eta L^2 d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\widetilde{x}^{s-1}\right\|^2\right]+2\frac{I\{B<n\}\eta\sigma^2}{B}+\eta\frac{L^2d^2\mu^2}{2} \qquad (28)
\end{aligned}
$$

where the last inequality holds by Lemma 3.1. Using Lemma 3.1, we have

$$\begin{aligned}
\mathbb{E}\left\|\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(\tilde{x}^s)\right\|^2 &= \mathbb{E}\left\|\sum_{j=1}^{d} \frac{f_{i,\mu_j}(x_t^s)}{\partial x_j} e_j - \frac{f_{i,\mu_j}(\tilde{x}^s)}{\partial x_j} e_j\right\|^2 \\
&\le d\sum_{j=1}^{d} \mathbb{E}\left\|\frac{f_{i,\mu_j}(x_t^s)}{\partial x_j} - \frac{f_{i,\mu_j}(\tilde{x}^s)}{\partial x_j}\right\|^2 \\
&\le L^2 d \sum_{j=1}^{d} \mathbb{E}\left\|x_{t,j}^s - \tilde{x}_j^s\right\|^2 = L^2 d \|x_t^s - \tilde{x}^s\|^2
\end{aligned} \tag{29}$$

□

Theorem 3.5. *Let step size $\eta = \frac{1}{6L}$ and b denote the minibatch size. The $\hat{x}$ returned by Algorithm 1 is an ϵ- accurate solution for problem* **??***. We distinguish the following two cases:*

1) We let batch size $B = n$. The number of SFO calls is at most

$$36L(\Phi(x_0) - \Phi(x^*))\left(\frac{B}{\epsilon\sqrt{b}} + \frac{b}{\epsilon}\right) = O\left(\frac{n}{\epsilon\sqrt{b}} + \frac{b}{\epsilon}\right).$$

2) Under Assumption 1, we let batch size $B = \{6\sigma^2/\epsilon, n\}$. The number of SFO calls is at most

$$36L(\Phi(x_0) - \Phi(x^*))\left(\frac{B}{\epsilon\sqrt{b}} + \frac{b}{\epsilon}\right) = O\left((n \wedge \frac{1}{\epsilon})\frac{1}{\epsilon\sqrt{b}} + \frac{b}{\epsilon}\right).$$

where $\wedge$ denotes the minimum.

In both cases, the number of *PO* calls equals to the total number of iterations T, which is at most

$$\frac{36L}{\epsilon}(\Phi(x_0) - \Phi(x^*)) = O\left(\frac{1}{\epsilon}\right).$$

Proof. Now, we are ready to use Lemma 3.3 to prove Theorem 3.5. Let $x_t^s = \mathrm{Prox}_{\eta h}(x_{t-1}^s - \eta v_{t-1}^s)$ and $\overline{x}_t^s = \mathrm{Prox}_{\eta h}(x_{t-1}^s - \eta \nabla f(x_{t-1}^s))$. By letting $x^+ = x_t^s$, $x = x_{t-1}^s$, $v = v_{t-1}^s$ and $z = \overline{x}_t^s$ in (17), we have

$$\Phi(x_t^s) \le \Phi(\overline{x}_t^s) + \left\langle \nabla f(x_{t-1}^s) - v_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle - \frac{1}{\eta}\left\langle x_t^s - x_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle + \frac{L}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 + \frac{L}{2}\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2. \tag{30}$$

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 (17), we have

$$\Phi(\overline{x}_t^s) \le \Phi(x_{t-1}^s) - \frac{1}{\eta}\left\langle \overline{x}_t^s - x_{t-1}^s, \overline{x}_t^s - x_{t-1}^s \right\rangle + \frac{L}{2}\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 = \Phi(x_{t-1}^s) - (\frac{1}{\eta} - \frac{L}{2})\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2. \tag{31}$$

We add (30) and (31) to obtain the key inequality

$$
\begin{aligned}
\Phi(x_t^s) \leq\ & \Phi(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) - 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 \\
=\ & \Phi(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) - v_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\
& - \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) \\
=\ & \Phi(x_{t-1}^s) - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\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) - 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 \\
\leq\ & \Phi(x_{t-1}^s) - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\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) - 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 \\
=\ & \Phi(x_{t-1}^s) - \left(\frac{5}{8\eta} - \frac{L}{2}\right)\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) - v_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\
\leq\ & \Phi(x_{t-1}^s) - \left(\frac{5}{8\eta} - \frac{L}{2}\right)\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) - v_{t-1}^s\right\|^2
\end{aligned} \tag{32}
$$

where the second inequality Young's inequality and the last inequality holds due to the Lemma **??**.

Note that $x_t^s = \mathrm{Prox}_{\eta h}(x_{t-1}^s - \eta v_{t-1}^s)$ is the iterated from in our algorithm. Now, we take expectations with all history for (32).

$$
\mathbb{E}[\Phi(x_t^s)] \leq \mathbb{E}\left[\Phi(x_{t-1}^s) - \left(\frac{5}{8\eta} - \frac{L}{2}\right)\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \eta\left\|\nabla f(x_{t-1}^s) - v_{t-1}^s\right\|^2\right] \tag{33}
$$

Then, we bound the variance term in (33) as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s)-v_{t-1}^s\right\|^2\right]\\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i(x_{t-1}^s)-\nabla f_i(\widetilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-g^s\right)\right\|^2\right]\\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i(x_{t-1}^s)-\nabla f_i(\widetilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\frac{1}{B}\sum_{j\in I_B}\nabla f_j(\widetilde{x}^{s-1})\right)\right\|^2\right]\\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i(x_{t-1}^s)-\nabla f_i(\widetilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\nabla f(\tilde{x}^{s-1})\right)+\left(\frac{1}{B}\sum_{j\in I_B}\nabla f_j(\widetilde{x}^{s-1})-\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(\nabla f_i(x_{t-1}^s)-\nabla f_i(\widetilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\nabla f(\tilde{x}^{s-1})\right)\right)+\frac{1}{B}\sum_{j\in I_B}\left(\nabla f_j(\widetilde{x}^{s-1})-\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(\nabla f_i(x_{t-1}^s)-\nabla f_i(\widetilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\nabla f(\tilde{x}^{s-1})\right)\right)\right\|^2\right]\\
&+\eta\mathbb{E}\left[\left\|\frac{1}{B}\sum_{j\in I_B}\left(\nabla f_j(\widetilde{x}^{s-1})-\nabla f(\tilde{x}^{s-1})\right)\right\|^2\right] \qquad (34)\\
=&\frac{\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\left(\left(\nabla f_i(x_{t-1}^s)-\nabla f_i(\widetilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\nabla f(\tilde{x}^{s-1})\right)\right)\right\|^2\right]\\
&+\eta\mathbb{E}\left[\left\|\frac{1}{B}\sum_{j\in I_B}\left(\nabla f_j(\widetilde{x}^{s-1})-\nabla f(\tilde{x}^{s-1})\right)\right\|^2\right] \qquad (35)\\
\leq&\frac{\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\nabla f_i(x_{t-1}^s)-\nabla f_i(\widetilde{x}^{s-1})\right\|^2\right]+\eta\mathbb{E}\left[\left\|\frac{1}{B}\sum_{j\in I_B}\left(\nabla f_j(\widetilde{x}^{s-1})-\nabla f(\tilde{x}^{s-1})\right)\right\|^2\right] \qquad (36)\\
\leq&\frac{\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\widetilde{x}^{s-1}\right\|^2\right]+\frac{I\{B<n\}\eta\sigma^2}{B} \qquad (37)
\end{aligned}
$$

where the expectations are taking with I_b and I_B, (34) and (35) holds $\mathbb{E}[\|x_1+x_2+\ldots+x_k\|^2]=\sum_{i=1}^k\mathbb{E}[\|x_i\|^2]$ if $x_1,x_2,\ldots,x_k$ are independent and of mean zero (note that I_b and I_B are also independent). (36) uses the fact that $\mathbb{E}[\|x-\mathbb{E}[x]\|^2]\leq\mathbb{E}[\|x\|^2]$, for any random variable x. (37) holds due to (**??**) and Assumption **??**.

Now we plug (37) into (33) to obtain

$$\mathbb{E}[\Phi(x_t^s)]$$

$$\leq \mathbb{E}\left[\Phi(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \frac{\eta L^2}{b}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B < n\}\eta\sigma^2}{B}\right] \tag{38}$$

$$= \mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13L}{4}\left\|x_t^s - x_{t-1}^s\right\|^2 - L\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \frac{L}{6b}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B < n\}\eta\sigma^2}{B}\right] \tag{39}$$

$$= \mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13L}{4}\left\|x_t^s - x_{t-1}^s\right\|^2 - \frac{1}{36L}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + \frac{L}{6b}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B < n\}\eta\sigma^2}{B}\right] \tag{40}$$

$$= \mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13L}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \frac{1}{36L}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + (\frac{L}{6b} + \frac{13L}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B < n\}\eta\sigma^2}{B}\right] \tag{41}$$

$$\mathbb{E}[\Phi(x_t^s)]$$

$$\leq \mathbb{E}\left[\Phi(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \frac{2\eta L^2 d}{b}\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right] \tag{42}$$

$$+ 2\frac{I\{B < n\}\eta\sigma^2}{B} + \eta\frac{L^2 d^2 \mu^2}{2} \tag{43}$$

$$= \mathbb{E}\left[\Phi(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + \frac{2\eta L^2 d}{b}\mathbb{E}\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right] \tag{44}$$

$$+ \frac{2I\{B < n\}\eta\sigma^2}{B} + \eta\frac{L^2 d^2 \mu^2}{2} \tag{45}$$

$$\leq \mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13Ld}{4}\left\|x_t^s - x_{t-1}^s\right\|^2 - \frac{1}{36Ld}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + \frac{Ld}{3b}\mathbb{E}\left[\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right] + \frac{2I\{B < n\}\eta\sigma^2}{B} + \frac{Ld\mu^2}{12}\right] \tag{46}$$

Next, we define an useful Lyapunov function as follows:

$$R_t^s = \mathbb{E}[\Phi(x_t^s) + c_t\left\|x_t^s - \tilde{x}^s\right\|^2] \tag{47}$$

where $\{c_t\}$ is a nonnegative sequence. Considering the upper bound of $\left\|x_t^s - \tilde{x}^{s-1}\right\|^2$, we have

$$\begin{aligned}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 &= \left\|x_t^s - x_{t-1}^s + x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 \\ &= (1 + \frac{1}{\alpha})\left\|x_t^s - x_{t-1}^s\right\|^2 + (1+\alpha)\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\end{aligned} \tag{48}$$

where $\alpha > 0$. Then we have

$$
\begin{aligned}
R_t^s &= \mathbb{E}[\Phi(x_t^s) + c_t \left\| x_t^s - \tilde{x}^{s-1} \right\|^2] \\
&\leq \mathbb{E}[\Phi(x_t^s) + c_t(1+\alpha) \left\| x_t^s - x_{t-1}^s \right\|^2 + c_t(1+\frac{1}{\alpha}) \left\| x_{t-1}^s - \tilde{x}^{s-1} \right\|^2] \qquad (49) \\
&= \mathbb{E}\left[\Phi(x_{t-1}^s) + (c_t(1+\alpha) - (\frac{5}{8\eta} - \frac{L}{2})) \left\| x_t^s - x_{t-1}^s \right\|^2 - \left(\frac{\eta}{3} - L\eta^2 \right) \left\| \mathcal{G}_\eta(x_{t-1}^s) \right\|^2 \right. \qquad (50) \\
&\quad + (\frac{2\eta L^2 d}{b} + c_t(1+\frac{1}{\alpha})) \mathbb{E}\left[\left\| x_{t-1}^s - \widetilde{x}^{s-1} \right\|^2 \right] + 2\frac{I\{B<n\}\eta\sigma^2}{B} + \eta \frac{L^2 d^2 \mu^2}{2} \qquad (51)
\end{aligned}
$$

where $\eta = \frac{\rho}{L}$, $c_{t-1} = \frac{2\rho L d}{b} + c_t(1+\frac{1}{\alpha})$. Let $c_m = 0$, $\alpha = m$, recursing on t. We have

$$
\begin{aligned}
c_t = \frac{2\rho L d}{b} \frac{(1+\frac{1}{\alpha})^{m-t} - 1}{\frac{1}{\alpha}} &= \frac{2\rho L m d}{b} \left((1+\frac{1}{m})^{m-t} - 1 \right) \\
&\leq \frac{2\rho L m d}{b}(e-1) \leq \frac{4\rho L m d}{b}
\end{aligned} \qquad (52)
$$

It follows that

$$
\begin{aligned}
c_t(1+\alpha) + \frac{L}{2} &\leq \frac{4\rho L m d}{b}(1+m) + \frac{L}{2} \\
&\leq \frac{8\rho L m^2 d}{b} + \frac{L}{2} \\
&= 2\frac{L}{2\rho}(\frac{8\rho^2 m^2 d}{b} + \frac{\rho}{2}) \\
&\leq \frac{1}{2\eta} \leq \frac{5}{8\eta}
\end{aligned} \qquad (53)
$$

where $2(\frac{8\rho^2 m^2 d}{b} + \frac{\rho}{2}) \leq 1$.

$$
\begin{aligned}
R_t^s &= \mathbb{E}[\Phi(x_t^s) + c_t \left\| x_t^s - \tilde{x}^{s-1} \right\|^2] \\
&\leq \mathbb{E}\left[\Phi(x_{t-1}^s) + (c_t(1+\alpha) - (\frac{5}{8\eta} - \frac{L}{2})) \left\| x_t^s - x_{t-1}^s \right\|^2 - \left(\frac{\eta}{3} - L\eta^2 \right) \left\| \mathcal{G}_\eta(x_{t-1}^s) \right\|^2 \right] \\
&\quad + (\frac{2\eta L^2 d}{b} + c_t(1+\frac{1}{\alpha})) \mathbb{E}\left[\left\| x_{t-1}^s - \widetilde{x}^{s-1} \right\|^2 \right] + 2\frac{I\{B<n\}\eta\sigma^2}{B} + \eta \frac{L^2 d^2 \mu^2}{2} \\
&= R_{t-1}^s - \left(\frac{\eta}{3} - L\eta^2 \right) \left\| \mathcal{G}_\eta(x_{t-1}^s) \right\|^2 + 2\frac{I\{B<n\}\eta\sigma^2}{B} + \eta \frac{L^2 d^2 \mu^2}{2}
\end{aligned}
$$

Telescoping the above inequality over t from 0 to $m-1$, since $x_0^s = x_m^{s-1} = \tilde{x}^{s-1}$ and $x_m^s = \tilde{x}^s$, we have

$$
\frac{1}{m} \sum_{t=1}^{m} \left\| \mathcal{G}_\eta(x_t^s) \right\|^2 \leq \frac{\mathbb{E}[\Phi(\tilde{x}^{s-1}) - \Phi(\tilde{x}^s)]}{m\gamma} + 2\frac{I\{B<n\}\eta\sigma^2}{B\gamma} + \eta \frac{L^2 d^2 \mu^2}{2\gamma} \qquad (54)
$$

where $\gamma = \frac{\eta}{3} - L\eta^2$. Summing the above inequality from 1 to S, we have

$$\begin{aligned}\frac{1}{T}\sum_{s=1}^{S}\sum_{t=1}^{m}\left\|\mathcal{G}_\eta(x_t^s)\right\|^2 &\leq \frac{\mathbb{E}[\Phi(\tilde{x}^0)-\Phi(\tilde{x}^S)]}{T\gamma} + 2\frac{I\{B<n\}\eta\sigma^2}{B\gamma} + \eta\frac{L^2d^2\mu^2}{2\gamma} \\ &\leq \frac{\mathbb{E}[\Phi(\tilde{x}^0)-\Phi(x^*)]}{T\gamma} + 2\frac{I\{B<n\}\eta\sigma^2}{B\gamma} + \eta\frac{L^2d^2\mu^2}{2\gamma}\end{aligned} \tag{55}$$

where x^* is an optimal solution of problem (**??**).

Given $m = [n^{\frac{1}{3}}]$, $b = [n^{\frac{2}{3}}]$ and $\rho = \frac{1}{6}$, it is easily verified that $2(\frac{8\rho^2m^2}{b} + \frac{\rho}{2}) = \frac{11}{18} < 1$. Using $d \geq 1$, we have $\gamma = \frac{\eta}{3} - L\eta^2 = \frac{1}{18L} - \frac{1}{36L} = \frac{1}{36L}$.

where (92) uses $\eta = \frac{1}{6L}$ and (98) uses the definition of gradient mapping $\mathcal{G}_\eta(x_{t-1}^s)$ and recall $\overline{x}_t^s := \mathrm{Prox}_{\eta h}(x_{t-1}^s - \eta\nabla f(x_{t-1}^s))$. (41) uses $\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 \leq (1+\frac{1}{\alpha})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + (1+\alpha)\left\|x_t^s - x_{t-1}^s\right\|^2$ by choosing $\alpha = 2t-1$.

Now, adding (41) for all iterations $1 \leq t \leq m$ in epoch s and recalling that $x_m^s = \tilde{x}^s$ and $x_0^s = \tilde{x}^{s-1}$, we get

$$\begin{aligned}&\mathbb{E}[\Phi(\tilde{x}^s)] \\ \leq&\mathbb{E}\Bigg[\Phi(\tilde{x}^{s-1}) - \frac{1}{36L}\sum_{t=1}^{m}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 - \sum_{t=1}^{m}\frac{13L}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 \\ &+ \sum_{t=1}^{m}(\frac{L}{6b} + \frac{13L}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \sum_{t=1}^{m}\frac{I\{B<n\}\eta\sigma^2}{B}\Bigg] \\ \leq&\mathbb{E}\Bigg[\Phi(\tilde{x}^{s-1}) - \frac{1}{36L}\sum_{t=1}^{m}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 - \sum_{t=1}^{m-1}\frac{13L}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 \\ &+ \sum_{t=2}^{m}(\frac{L}{6b} + \frac{13L}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \sum_{t=1}^{m}\frac{I\{B<n\}\eta\sigma^2}{B}\Bigg] \qquad (56) \\ =&\mathbb{E}\Bigg[\Phi(\tilde{x}^{s-1}) - \sum_{t=1}^{m}\frac{1}{36L}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 - \sum_{t=1}^{m-1}(\frac{13L}{8t} - \frac{L}{6b} - \frac{13L}{8t+4})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 + \sum_{t=1}^{m}\frac{I\{B<n\}\eta\sigma^2}{B}\Bigg] \\ \leq&\mathbb{E}\Bigg[\Phi(\tilde{x}^{s-1}) - \sum_{t=1}^{m}\frac{1}{36L}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 - \sum_{t=1}^{m-1}(\frac{L}{2t^2} - \frac{L}{6b})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 + \sum_{t=1}^{m}\frac{I\{B<n\}\eta\sigma^2}{B}\Bigg] \\ \leq&\mathbb{E}\Bigg[\Phi(\tilde{x}^{s-1}) - \sum_{t=1}^{m}\frac{1}{36L}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + \sum_{t=1}^{m}\frac{I\{B<n\}\eta\sigma^2}{B}\Bigg] \qquad (57)\end{aligned}$$

$$
\begin{aligned}
&\mathbb{E}[\Phi(\tilde{x}^s)]\\
\leq&\mathbb{E}\left[\Phi(\tilde{x}^{s-1})-\frac{1}{36Ld}\sum_{t=1}^{m}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2-\sum_{t=1}^{m}\frac{13Ld}{8t}\left\|x_t^s-\tilde{x}^{s-1}\right\|^2\right.\\
&\left.+\sum_{t=1}^{m}(\frac{Ld}{3b}+\frac{13Ld}{8t-4})\mathbb{E}\left[\left\|x_{t-1}^s-\widetilde{x}^{s-1}\right\|^2\right]+\sum_{t=1}^{m}\frac{2I\{B<n\}\eta\sigma^2}{B}+\sum_{t=1}^{m}\frac{Ld\mu^2}{12}\right]\\
\leq&\mathbb{E}\left[\Phi(\tilde{x}^{s-1})-\frac{1}{36Ld}\sum_{t=1}^{m}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2-\sum_{t=1}^{m-1}\frac{13Ld}{8t}\left\|x_t^s-\tilde{x}^{s-1}\right\|^2\right.\\
&\left.+\sum_{t=2}^{m}(\frac{Ld}{3b}+\frac{13Ld}{8t-4})\mathbb{E}\left[\left\|x_{t-1}^s-\widetilde{x}^{s-1}\right\|^2\right]+\sum_{t=1}^{m}\frac{2I\{B<n\}\eta\sigma^2}{B}+\sum_{t=1}^{m}\frac{Ld\mu^2}{12}\right] &(58)\\
=&\mathbb{E}\left[\Phi(\tilde{x}^{s-1})-\sum_{t=1}^{m}\frac{1}{36Ld}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2-\sum_{t=1}^{m-1}(\frac{13Ld}{8t}-\frac{Ld}{3b}-\frac{13Ld}{8t+4})\left\|x_t^s-\tilde{x}^{s-1}\right\|^2\right. &(59)\\
&\left.+\sum_{t=1}^{m}\frac{2I\{B<n\}\eta\sigma^2}{B}+\sum_{t=1}^{m}\frac{Ld\mu^2}{12}\right]\\
\leq&\mathbb{E}\left[\Phi(\tilde{x}^{s-1})-\sum_{t=1}^{m}\frac{1}{36Ld}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2-\sum_{t=1}^{m-1}(\frac{Ld}{2t^2}-\frac{Ld}{3b})\left\|x_t^s-\tilde{x}^{s-1}\right\|^2+\sum_{t=1}^{m}\frac{2I\{B<n\}\eta\sigma^2}{B}+\sum_{t=1}^{m}\frac{Ld\mu^2}{12}\right]\\
\leq&\mathbb{E}\left[\Phi(\tilde{x}^{s-1})-\sum_{t=1}^{m}\frac{1}{36Ld}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2+\sum_{t=1}^{m}\frac{2I\{B<n\}\eta\sigma^2}{B}+\sum_{t=1}^{m}\frac{Ld\mu^2}{12}\right] &(60)
\end{aligned}
$$

where (58) holds since $\|.\|^2$ always be non-negative and $x_0^s=\tilde{x}^{s-1}$, and (60) holds since $m=\sqrt{b}$. Thus, $\frac{L}{2t^2}-\frac{L}{6b}\geq 0$ $\frac{Ld}{2t^2}-\frac{Ld}{6b}\geq 0$ for all $1\leq t<m$.

Now we sum up (60) for all epochs $1\leq s\leq S$ to finish the proof as follows:

$$
0\leq\mathbb{E}[\Phi(\tilde{x}^S)-\Phi(x^*)]\leq\mathbb{E}\left[\Phi(\tilde{x}^0)-\Phi(x^*)-\sum_{s=1}^{S}\sum_{t=1}^{m}\frac{1}{36L}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2+\sum_{s=1}^{S}\sum_{t=1}^{m}\frac{I\{B<n\}\eta\sigma^2}{B}\right]
$$

$$
\begin{aligned}
\mathbb{E}[\left\|\mathcal{G}_\eta(\hat{x})\right\|^2]&\leq\frac{36L\,(\Phi(x_0)-\Phi(x^*))}{Sm}+\frac{I\{B<n\}36L\eta\sigma^2}{B} &(61)\\
&=\frac{36L\,(\Phi(x_0)-\Phi(x^*))}{Sm}+\frac{I\{B<n\}6\sigma^2}{B}=2\epsilon &(62)
\end{aligned}
$$

$$
0\leq\mathbb{E}[\Phi(\tilde{x}^S)-\Phi(x^*)]\leq\mathbb{E}\left[\Phi(\tilde{x}^0)-\Phi(x^*)-\sum_{s=1}^{S}\sum_{t=1}^{m}\frac{1}{36Ld}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2+\sum_{s=1}^{S}\sum_{t=1}^{m}(\frac{I\{B<n\}\eta\sigma^2}{B}+\frac{Ld\mu^2}{12})\right]
$$

$$
\begin{aligned}
\mathbb{E}[\left\|\mathcal{G}_\eta(\hat{x})\right\|^2]&\leq\frac{36Ld\,(\Phi(x_0)-\Phi(x^*))}{Sm}+\frac{I\{B<n\}36Ld\eta\sigma^2}{B}+3L^2d^2\mu^2 &(63)\\
&=\frac{36Ld\,(\Phi(x_0)-\Phi(x^*))}{Sm}+\frac{I\{B<n\}6\sigma^2}{B}+3L^2d^2\mu^2=3\epsilon &(64)
\end{aligned}
$$

where (63) holds since $\hat{x}$ is chosen uniformly randomly from $\{x_{t-1}^s\}_{t\in[m],s\in[S]}$, and (64) uses $\eta = \frac{1}{6L}$. Now we obtain the total number of iterations $T = Sm = S\sqrt{b} = \frac{36L(\Phi(x_0)-\Phi(x^*))}{\epsilon}$. The proof is finished since the number of SFO call equals to $Sn + Smb = 36L\,(\Phi(x_0) - \Phi(x^*))\,(\frac{n}{\epsilon\sqrt{b}} + \frac{b}{\epsilon})$ if $B = n$ (i.e., the second term in (64) is 0 and thus assumption **??** is not needed), or equals to $Sn + Smb = 36L\,(\Phi(x_0) - \Phi(x^*))\,(\frac{B}{\epsilon\sqrt{b}} + \frac{b}{\epsilon})$ if $B < n$ (note that $\frac{I\{B<n\}6\sigma^2}{B} \le \epsilon$ since $B \ge 5\sigma^2/\epsilon$). □

4 Convergence Under PL Condition

In this section, we provide the global linear convergence rate for nonconvex functions under the Polyak-Lojasiewicz (PL) condition [Polyak, 1963]. The original form of PL condition is

$$\exists \mu > 0, \text{such that } \|\nabla f(x)\|^2 \ge 2\mu(f(x) - f^*),\ \forall x, \tag{65}$$

where f^* denotes the (global) optimal function value. It is worth noting that f satisfies PL condition when f is μ-strongly convex.

Due to the nonsmooth term $h(x)$ in problem (**??**), we use the gradient mapping to define a more general form of PL condition as follows

$$\exists \mu > 0, \text{such that} \left\|G_\eta(x)\right\|^2 \ge 2\mu(\Phi(x) - \Phi^*),\ \forall x. \tag{66}$$

Recall that if $h(x)$ is a constant function, the gradient mapping reduces to $G_\eta(x) = \nabla f(x)$.

We want to point out that [] used the following form of PL condition

$$\exists \mu > 0, \text{such that} D_h(x,\alpha) \ge 2\mu(\Phi(x) - \Phi^*),\ \forall x. \tag{67}$$

where $D_h(x,\alpha) := -2\alpha \min_y\{\langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2 + h(y) - h(x)\}$. Our PL condition is arguably more natural.

Theorem 4.1. *Let step size $\eta = \frac{1}{6L}$ and b denote the minibatch size. Then the final iteration point $\tilde{x}^S$ in Algorithm* **??** *satisfies $\mathbb{E}[\Phi(\tilde{x}^S) - \Phi^*] \le \epsilon$ under PL condition. We distinguish the following two cases:*

1) We let batch size $B = n$. The number of SFO calls is bounded by

$$O\left(\frac{n}{\mu\sqrt{b}}\log\frac{1}{\epsilon} + \frac{b}{\mu}\log\frac{1}{\epsilon}\right).$$

2) Under Assumption 1, we let batch size $B = \min\{\frac{6\sigma^2}{\mu\epsilon}, n\}$. The number of SFO calls is bounded by

$$O\left((n \wedge \frac{1}{\mu\epsilon})\frac{1}{\mu\sqrt{b}}\log\frac{1}{\epsilon} + \frac{b}{\mu}\log\frac{1}{\epsilon}\right).$$

where $\wedge$ denotes the minimum.

3) In both cases, the number of PO calls equals to the total number of iterations T which is bounded by

$$O\left(\frac{1}{\mu}\log\frac{1}{\epsilon}\right).$$

Proof. First, we recall a key inequality (41) from the proof of Theorem 1, i.e.,

$$\mathbb{E}\Phi(x_t^s)$$
$$\mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13L}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \frac{1}{36L}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + (\frac{L}{6b} + \frac{13L}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B<n\}\eta\sigma^2}{B}\right] \tag{68}$$

$$\begin{aligned}&\mathbb{E}[\Phi(x_t^s)]\\ \leq&\mathbb{E}\Bigg[\Phi(x_{t-1}^s) - \frac{1}{36Ld}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 - \frac{13Ld}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2\\ &+(\frac{Ld}{6b} + \frac{13Ld}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{Ld\mu^2}{12}\Bigg]\end{aligned} \tag{69}$$

Then, we plug the following PL inequality

$$\left\|G_\eta(x)\right\|^2 \geq 2\mu(\Phi(x) - \Phi^*) \tag{70}$$

into (69) to get

$$\mathbb{E}\Phi(x_t^s)$$
$$\leq \mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13L}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \frac{\mu}{18L}(\Phi(x_{t-1}^s) - \Phi^*) + (\frac{L}{6b} + \frac{13L}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B<n\}\eta\sigma^2}{B}\right] \tag{71}$$

$$\begin{aligned}&\mathbb{E}[\Phi(x_t^s)]\\ \leq&\mathbb{E}\Bigg[\Phi(x_{t-1}^s) - \frac{\mu}{18Ld}(\Phi(x_{t-1}^s) - \Phi^*) - \frac{13Ld}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2\\ &+(\frac{Ld}{6b} + \frac{13Ld}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{Ld\mu^2}{12}\Bigg]\end{aligned} \tag{72}$$

Then, we obtain

$$\mathbb{E}[\Phi(x_t^s) - \Phi^*]$$
$$\leq \mathbb{E}\left[\left(1 - \frac{\mu}{18L}\right)(\Phi(x_{t-1}^s) - \Phi^*) - \frac{13L}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 + (\frac{L}{6b} + \frac{13L}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B<n\}\eta\sigma^2}{B}\right] \tag{73}$$

$$\begin{aligned}&\mathbb{E}[\Phi(x_t^s) - \Phi^*]\\ \leq&\mathbb{E}\Bigg[\left(1 - \frac{\mu}{18Ld}\right)(\Phi(x_{t-1}^s) - \Phi^*) - \frac{13Ld}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2\\ &+(\frac{Ld}{6b} + \frac{13Ld}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{Ld\mu^2}{12}\Bigg]\end{aligned} \tag{74}$$

Let $\alpha := 1 - \frac{\mu}{18L}$ $\alpha := 1 - \frac{\mu}{18Ld}$ and $\Psi_t^s := \frac{\mathbb{E}[\Phi(x_t^s) - \Phi^*]}{\alpha^t}$. Plugging them into (74), we have

$$
\begin{aligned}
&\Psi_t^s \\
&\leq \Psi_{t-1}^s - \mathbb{E}\left[\frac{13L}{8t\alpha^t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \frac{1}{\alpha^t}(\frac{L}{6b} + \frac{13L}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 - \frac{1}{\alpha^t}\frac{I\{B<n\}\eta\sigma^2}{B}\right]
\end{aligned}
\tag{75}
$$

$$
\begin{aligned}
&\Psi_t^s \\
&\leq \Psi_{t-1}^s - \mathbb{E}\left[\frac{13Ld}{8t\alpha^t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right. \\
&\quad \left. - \frac{1}{\alpha^t}(\frac{Ld}{6b} + \frac{13Ld}{8t-4})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 - \frac{1}{\alpha^t}\frac{2I\{B<n\}\eta\sigma^2}{B} - \frac{1}{\alpha^t}\frac{Ld\mu^2}{12}\right]
\end{aligned}
\tag{76}
$$

Now, adding (110) from all iterations $1 \leq t \leq m$ in epoch s and recalling that $x_m^s = \tilde{x}^s$

and $x_0^s = \tilde{x}^{s-1}$, we have

$$
\begin{aligned}
& \mathbb{E}[\Phi(\tilde{x}^s) - \Phi^*] \\
\leq\, & \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1}) - \Phi^*)\right] + \alpha^m \sum_{t=1}^{m} \frac{1}{\alpha^t} \frac{I\{B < n\}\eta\sigma^2}{B} \\
& -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{13L}{8t\alpha^t} \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^{m} \frac{1}{\alpha^t} (\frac{L}{6b} + \frac{13L}{8t-4}) \left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
\leq\, & \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1}) - \Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{I\{B < n\}\eta\sigma^2}{B} \\
& -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{13L}{8t\alpha^t} \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^{m} \frac{1}{\alpha^t} (\frac{L}{6b} + \frac{13L}{8t-4}) \left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
\leq\, & \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1}) - \Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{I\{B < n\}\eta\sigma^2}{B} \\
& -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{13L}{8t\alpha^t} \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=2}^{m} \frac{1}{\alpha^t} (\frac{L}{6b} + \frac{13L}{8t-4}) \left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \qquad (77) \\
\leq\, & \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1}) - \Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{I\{B < n\}\eta\sigma^2}{B} \\
& -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{1}{\alpha^{t+1}} \left(\frac{13L\alpha}{8t} - \frac{L}{6b} - \frac{13L}{8t+4}\right) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
\leq\, & \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1}) - \Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{I\{B < n\}\eta\sigma^2}{B} \\
& -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{1}{\alpha^{t+1}} \left(\frac{13L}{8t}(1 - \frac{1}{18\sqrt{n}}) - \frac{L}{6b} - \frac{13L}{8t+4}\right) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \qquad (78) \\
\leq\, & \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1}) - \Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{I\{B < n\}\eta\sigma^2}{B} \\
& -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{L}{\alpha^{t+1}} \left(\frac{1}{2t^2} - \frac{1}{8\sqrt{n}t} - \frac{1}{6b}\right) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
\leq\, & \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1}) - \Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{I\{B < n\}\eta\sigma^2}{B} \qquad (79)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}[\Phi(\tilde{x}^s)-\Phi^*] \\
&\leq \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1})-\Phi^*)\right] + \alpha^m \sum_{t=1}^{m} \frac{1}{\alpha^t} \frac{2I\{B<n\}\eta\sigma^2}{B} + \alpha^m \sum_{t=1}^{m} \frac{1}{\alpha^t} \frac{Ld\mu^2}{12} \\
&\quad -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{13Ld}{8t\alpha^t} \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^{m} \frac{1}{\alpha^t} (\frac{Ld}{6b} + \frac{13Ld}{8t-4}) \left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
&\leq \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1})-\Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{1-\alpha^m}{1-\alpha} \frac{Ld\mu^2}{12} \\
&\quad -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{13Ld}{8t\alpha^t} \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^{m} \frac{1}{\alpha^t} (\frac{Ld}{6b} + \frac{13Ld}{8t-4}) \left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
&\leq \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1})-\Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{1-\alpha^m}{1-\alpha} \frac{Ld\mu^2}{12} \\
&\quad -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{13Ld}{8t\alpha^t} \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=2}^{m} \frac{1}{\alpha^t} (\frac{Ld}{6b} + \frac{13Ld}{8t-4}) \left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \qquad (80) \\
&\leq \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1})-\Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{1-\alpha^m}{1-\alpha} \frac{Ld\mu^2}{12} \\
&\quad -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{1}{\alpha^{t+1}} \left(\frac{13Ld\alpha}{8t} - \frac{Ld}{6b} - \frac{13Ld}{8t+4}\right) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
&\leq \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1})-\Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{1-\alpha^m}{1-\alpha} \frac{Ld\mu^2}{12} \\
&\quad -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{1}{\alpha^{t+1}} \left(\frac{13Ld}{8t}(1-\frac{1}{18\sqrt{n}}) - \frac{Ld}{6b} - \frac{13Ld}{8t+4}\right) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \qquad (81) \\
&\leq \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1})-\Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{1-\alpha^m}{1-\alpha} \frac{Ld\mu^2}{12} \\
&\quad -\alpha^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{Ld}{\alpha^{t+1}} \left(\frac{1}{2t^2} - \frac{1}{8\sqrt{n}t} - \frac{1}{6b}\right) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
&\leq \alpha^m \mathbb{E}\left[(\Phi(\tilde{x}^{s-1})-\Phi^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{1-\alpha^m}{1-\alpha} \frac{Ld\mu^2}{12} \qquad (82)
\end{aligned}
$$

where (80) holds since $\|.\|^2$ always be non-negative and $x_0^s = \tilde{x}^{s-1}$. (81) holds since $\alpha = 1 - \frac{\mu}{18L}$ and the assumption $L/\mu > \sqrt{n}$. (111) holds since it is sufficient to show that $\Gamma_t \leq 0$ for all $1 \leq t < m$, where $\Gamma_t = \frac{1}{2t^2} - \frac{1}{8\sqrt{n}t} - \frac{1}{6b}$. Taking a derivative for Γ_t, we get $\Gamma_t' = -\frac{1}{t^3} + \frac{1}{8\sqrt{n}t^2} = -\frac{8\sqrt{n}-t}{8\sqrt{n}t^3} < 0$ since $t < m = \sqrt{b} \leq \sqrt{n}$. Thus, Γ_t decreases in t. We only need to show that $\Gamma_m = \Gamma_{\sqrt{b}} \geq 0$, i.e., $\frac{1}{2b} - \frac{1}{8\sqrt{nb}} - \frac{1}{6b} = \frac{1}{3b} - \frac{1}{8\sqrt{nb}} \geq 0$. It is easy to see that this inequality holds since $b \leq n$.

Similarly, let $\tilde{\alpha} = \alpha^m$ and $\tilde{\Psi}^s = \frac{\mathbb{E}[\Phi(\tilde{x}^s)-\Phi^*]}{\tilde{\alpha}^s}$. Plugging them into (111), we have

$$\widetilde{\Psi}^s \leq \widetilde{\Psi}^{s-1} + \frac{1}{\tilde{\alpha}^s}\frac{1-\tilde{\alpha}}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B} \tag{83}$$

$$\widetilde{\Psi}^s \leq \widetilde{\Psi}^{s-1} + \frac{1}{\tilde{\alpha}^s}\frac{1-\tilde{\alpha}}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B} + \frac{1}{\tilde{\alpha}^s}\frac{1-\tilde{\alpha}}{1-\alpha}\frac{Ld\mu^2}{12} \tag{84}$$

Now, we sum up (112) for all epochs $1 \leq s \leq S$ to finish the proof as follows:

$$\begin{aligned}
\mathbb{E}[\Phi(\widetilde{x}^S)-\Phi^*] &\leq \widetilde{\alpha}^S\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \widetilde{\alpha}^S\sum_{s=1}^{S}\frac{1}{\widetilde{\alpha}^s}\frac{1-\widetilde{\alpha}}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B} \\
&= \alpha^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \frac{1-\widetilde{\alpha}^S}{1-\widetilde{\alpha}}\frac{1-\widetilde{\alpha}}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B} \\
&\leq \alpha^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \frac{1}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B} \\
&= \left(1-\frac{\mu}{18L}\right)^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \frac{I\{B<n\}18L\eta\sigma^2}{\mu B} \qquad (85) \\
&= \left(1-\frac{\mu}{18L}\right)^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \frac{I\{B<n\}3\sigma^2}{\mu B} = 2\epsilon \qquad (86)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\Phi(\widetilde{x}^S)-\Phi^*] &\leq \widetilde{\alpha}^S\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \widetilde{\alpha}^S\sum_{s=1}^{S}\frac{1}{\widetilde{\alpha}^s}\frac{1-\widetilde{\alpha}}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B} + \widetilde{\alpha}^S\sum_{s=1}^{S}\frac{1}{\widetilde{\alpha}^s}\frac{1-\widetilde{\alpha}}{1-\alpha}\frac{Ld\mu^2}{12} \\
&= \alpha^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \frac{1-\widetilde{\alpha}^S}{1-\widetilde{\alpha}}\frac{1-\widetilde{\alpha}}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B} + \frac{1-\widetilde{\alpha}^S}{1-\widetilde{\alpha}}\frac{1-\widetilde{\alpha}}{1-\alpha}\frac{Ld\mu^2}{12} \\
&\leq \alpha^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \frac{1}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B} + \frac{1}{1-\alpha}\frac{Ld\mu^2}{12} \\
&= \left(1-\frac{\mu}{18Ld}\right)^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \frac{I\{B<n\}18Ld\eta\sigma^2}{\mu B} + \frac{18L^2d^2\mu}{12} \qquad (87) \\
&= \left(1-\frac{\mu}{18Ld}\right)^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*] + \frac{I\{B<n\}3\sigma^2}{\mu B} + \frac{3L^2d^2\mu}{2} = 3\epsilon \qquad (88)
\end{aligned}$$

where (113) holds since $\alpha = 1-\frac{\mu}{18L}$ $\alpha = 1-\frac{\mu}{18Ld}$, and (114) uses $\eta = \frac{1}{6L}\eta = \frac{1}{6Ld}$.

From (114), we obtain the total number of iterations $T = Sm = S\sqrt{b} = O(\frac{1}{\mu}\log\frac{1}{\epsilon})$. The number of PO calls equals to $T = Sm = O(\frac{1}{\mu}\log\frac{1}{\epsilon})$. The number of SFO calls equals to $Sn + Smb = O(\frac{n}{\mu\sqrt{b}}\log\frac{1}{\epsilon} + \frac{b}{\mu}\log\frac{1}{\epsilon})$ if $B = n$, or equals to $Sn + Smb = O(\frac{B}{\mu\sqrt{b}}\log\frac{1}{\epsilon} + \frac{b}{\mu}\log\frac{1}{\epsilon})$ if $B < n$(note that $\frac{I\{B<n\}3\sigma^2}{\mu B} \leq \epsilon$ since $B \geq 6\sigma^2/\mu\epsilon$). $\mu \leq \frac{2\epsilon}{3L^2d^2}$ □

$$
\begin{aligned}
&\mathbb{E}[\Phi(x_t^s)] \\
\le& \mathbb{E}\left[\Phi(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \frac{2\eta L^2 d}{b}\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right] \quad (89)\\
&+ 2\frac{I\{B<n\}\eta\sigma^2}{B} + \eta\frac{L^2 d^2 \mu^2}{2} \quad (90)\\
=& \mathbb{E}\left[\Phi(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + \frac{2\eta L^2 d}{b}\mathbb{E}\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right] \quad (91)\\
&+ \frac{2I\{B<n\}\eta\sigma^2}{B} + \eta\frac{L^2 d^2 \mu^2}{2} \quad (92)\\
\le& \mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13Ld}{4}\left\|x_t^s - x_{t-1}^s\right\|^2 - \frac{1}{36Ld}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + \frac{Ld}{3b}\mathbb{E}\left[\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right] + \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{Ld\mu^2}{12}\right] \quad (93)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\mathbb{E}[\Phi(x_t^s)] \\
\le& \mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13Ld}{4}\left\|x_t^s - x_{t-1}^s\right\|^2 - \frac{1}{36Ld}\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + \frac{Ld}{3b}\mathbb{E}\left[\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right]\right. \quad (94)\\
&\left. + \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{Ld\mu^2}{12}\right] \quad (95)
\end{aligned}
$$

Then, we plug the following PL inequality

$$
\left\|G_\eta(x)\right\|^2 \ge 2\mu(\Phi(x) - \Phi^*) \quad (96)
$$

$$
\begin{aligned}
&\mathbb{E}[\Phi(x_t^s) - \Phi^*] \\
\le& \mathbb{E}\left[(1 - \frac{\mu}{18Ld})(\Phi(x_{t-1}^s) - \Phi^*) - \frac{13Ld}{4}\left\|x_t^s - x_{t-1}^s\right\|^2 + \frac{Ld}{3b}\mathbb{E}\left[\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right]\right. \quad (97)\\
&\left. + \frac{2I\{B<n\}\eta\sigma^2}{B} + \frac{Ld\mu^2}{12}\right] \quad (98)
\end{aligned}
$$

Next, we define an useful Lyapunov function as follows:

$$
R_t^s = \mathbb{E}[\Phi(x_t^s) - \Phi^* + c_t\left\|x_t^s - \tilde{x}^s\right\|^2] \quad (99)
$$

where $\{c_t\}$ is a nonnegative sequence. Considering the upper bound of $\left\|x_t^s - \tilde{x}^{s-1}\right\|^2$, we have

$$
\begin{aligned}
\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 &= \left\|x_t^s - x_{t-1}^s + x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 \\
&= (1 + \frac{1}{\alpha})\left\|x_t^s - x_{t-1}^s\right\|^2 + (1+\alpha)\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2
\end{aligned} \quad (100)
$$

where $\alpha > 0$. Then we have

$$
\begin{aligned}
R_t^s &= \mathbb{E}[\Phi(x_t^s) - \Phi^* + c_t \left\|x_t^s - \tilde{x}^{s-1}\right\|^2] \\
&\leq \mathbb{E}[\Phi(x_t^s) - \Phi^* + c_t(1+\alpha)\left\|x_t^s - x_{t-1}^s\right\|^2 + c_t(1+\frac{1}{\alpha})\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2] && (101) \\
&= \mathbb{E}\left[(\Phi(x_{t-1}^s) - \Phi^*) - \left(\frac{\eta}{3} - L\eta^2\right)\left\|\mathcal{G}_\eta(x_{t-1}^s)\right\|^2 + (c_t(1+\alpha) - (\frac{5}{8\eta} - \frac{L}{2}))\left\|x_t^s - x_{t-1}^s\right\|^2\right. && (102) \\
&\quad + (\frac{2\eta L^2 d}{b} + c_t(1+\frac{1}{\alpha}))\mathbb{E}\left[\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right] + 2\frac{I\{B<n\}\eta\sigma^2}{B} + \eta\frac{L^2 d^2 \mu^2}{2} && (103) \\
&= \mathbb{E}\left[(1-2\mu\gamma)(\Phi(x_{t-1}^s) - \Phi^*) + (c_t(1+\alpha) - (\frac{5}{8\eta} - \frac{L}{2}))\left\|x_t^s - x_{t-1}^s\right\|^2\right. && (104) \\
&\quad + (\frac{2\eta L^2 d}{b} + c_t(1+\frac{1}{\alpha}))\mathbb{E}\left[\left\|x_{t-1}^s - \widetilde{x}^{s-1}\right\|^2\right] + 2\frac{I\{B<n\}\eta\sigma^2}{B} + \eta\frac{L^2 d^2 \mu^2}{2} && (105)
\end{aligned}
$$

where $\gamma = \frac{\eta}{3} - L\eta^2$, $\eta = \frac{\rho}{L}$, $c_{t-1} = \frac{2\rho L d}{b} + c_t(1+\frac{1}{\alpha})$. Let $c_m = 0$, $\alpha = m$, recursing on t. We have

$$
\begin{aligned}
c_t = \frac{2\rho L d}{b}\frac{(1+\frac{1}{\alpha})^{m-t} - 1}{\frac{1}{\alpha}} &= \frac{2\rho L m d}{b}\left((1+\frac{1}{m})^{m-t} - 1\right) \\
&\leq \frac{2\rho L m d}{b}(e-1) \leq \frac{4\rho L m d}{b}
\end{aligned} \tag{106}
$$

It follows that

$$
\begin{aligned}
c_t(1+\alpha) + \frac{L}{2} &\leq \frac{4\rho L m d}{b}(1+m) + \frac{L}{2} \\
&\leq \frac{8\rho L m^2 d}{b} + \frac{L}{2} \\
&= 2\frac{L}{2\rho}(\frac{8\rho^2 m^2 d}{b} + \frac{\rho}{2}) \\
&\leq \frac{1}{2\eta} \leq \frac{5}{8\eta}
\end{aligned} \tag{107}
$$

where $2(\frac{8\rho^2 m^2 d}{b} + \frac{\rho}{2}) \leq 1$. where $\eta = \frac{\rho}{L}$, $\beta c_{t-1} = \frac{2\rho L d}{b} + c_t(1+\frac{1}{\alpha})$. Let $c_m = 0$, $\alpha = 2$, recursing on t. We have

$$
\begin{aligned}
c_t = \frac{2\rho L d}{b\beta}\frac{(\frac{1}{\beta} + \frac{1}{\beta\alpha})^{m-t} - 1}{\frac{1}{\alpha\beta} + \frac{1}{\beta} - 1} = \frac{2\rho L d}{b\beta}\frac{(\frac{1}{\beta} + \frac{1}{\beta\alpha})^{m-t} - 1}{\frac{1}{\alpha\beta}} &= \frac{2\rho L d}{b\beta}\left((1+\frac{1}{\beta})^{m-t} - 1\right) \\
&\leq \frac{2\rho L d}{b\beta}\left((1+\frac{1}{\beta})^{m-t} - 1\right) \leq \frac{4\rho L d 3^{m-t}}{b}
\end{aligned} \tag{108}
$$

It follows that

$$\begin{aligned} c_t(1+\alpha)+\frac{L}{2} &\le \frac{4\rho L d 3^{m+1-t}}{b}+\frac{L}{2} \\ &= 2\frac{L}{2\rho}\left(\frac{4\rho^2 3^{m+1-t} d}{b}+\frac{\rho}{2}\right) \\ &\le \frac{1}{2\eta} \le \frac{5}{8\eta} \end{aligned} \tag{109}$$

where $2(\frac{4\rho^2 3^{m+1-t} d}{b}+\frac{\rho}{2}) \le 1$. Be carfull make $c_{m+1}=0$, $0<\beta\le\alpha$

$$\begin{aligned} R_t^s &= \mathbb{E}[\Phi(x_t^s)-\Phi^*+c_t\left\|x_t^s-\tilde{x}^{s-1}\right\|^2] \\ &\le \mathbb{E}\left[(1-\frac{\mu}{18Ld})(\Phi(x_{t-1}^s)-\Phi^*)+(c_t(1+\alpha)-(\frac{5}{8\eta}-\frac{L}{2}))\left\|x_t^s-x_{t-1}^s\right\|^2\right] \\ &\quad +(\frac{2\eta L^2 d}{b}+c_t(1+\frac{1}{\alpha}))\mathbb{E}\left[\left\|x_{t-1}^s-\widetilde{x}^{s-1}\right\|^2\right]+2\frac{I\{B<n\}\eta\sigma^2}{B}+\eta\frac{L^2d^2\mu^2}{2} \\ &= \beta(\Phi(x_{t-1}^s)-\Phi^*)+\beta c_{t-1}\left\|x_{t-1}^s-\widetilde{x}^{s-1}\right\|^2+2\frac{I\{B<n\}\eta\sigma^2}{B}+\eta\frac{L^2d^2\mu^2}{2} \\ &= \beta R_{t-1}^s+2\frac{I\{B<n\}\eta\sigma^2}{B}+\eta\frac{L^2d^2\mu^2}{2} \end{aligned}$$

Thus, we obtain,

$$\Psi_t^s \le \Psi_{t-1}^s-\mathbb{E}\left[-\frac{1}{\alpha^t}\frac{2I\{B<n\}\eta\sigma^2}{B}-\frac{1}{\alpha^t}\eta\frac{L^2d^2\mu^2}{2}\right] \tag{110}$$

with $\Psi_t^s := \frac{\mathbb{E}[\Phi(x_t^s)-\Phi^*]+c_t\left\|x_t^s-\tilde{x}^{s-1}\right\|^2}{\alpha^t}$. Telescoping the above inequality over t from 0 to $m-1$, since $x_0^s=x_m^{s-1}=\tilde{x}^{s-1}$ and $x_m^s=\tilde{x}^s$, we have

$$\begin{aligned} &\mathbb{E}[\Phi(\tilde{x}^s)-\Phi^*]+c_m\left\|\tilde{x}^s-\tilde{x}^{s-1}\right\|^2 \\ &\le \alpha^m\mathbb{E}\left[(\Phi(\tilde{x}^{s-1})-\Phi^*)\right]+\alpha^m\sum_{t=1}^{m}\frac{1}{\alpha^t}\frac{2I\{B<n\}\eta\sigma^2}{B}+\alpha^m\sum_{t=1}^{m}\frac{1}{\alpha^t}\eta\frac{L^2d^2\mu^2}{2} \\ &\le \alpha^m\mathbb{E}\left[(\Phi(\tilde{x}^{s-1})-\Phi^*)\right]+\frac{1-\alpha^m}{1-\alpha}\frac{2I\{B<n\}\eta\sigma^2}{B}+\frac{1-\alpha^m}{1-\alpha}\eta\frac{L^2d^2\mu^2}{2} \end{aligned} \tag{111}$$

$$\widetilde{\Psi}^s+\frac{c_m}{\tilde{\alpha}^s}\left\|\tilde{x}^s-\tilde{x}^{s-1}\right\|^2 \le \widetilde{\Psi}^{s-1}+\frac{1}{\tilde{\alpha}^s}\frac{1-\tilde{\alpha}}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B}+\frac{1}{\tilde{\alpha}^s}\frac{1-\tilde{\alpha}}{1-\alpha}\eta\frac{L^2d^2\mu^2}{2} \tag{112}$$

with $\widetilde{\Psi}^s := \frac{\mathbb{E}[\Phi(\tilde{x}^s)-\Phi^*]}{\alpha^s}$. Now, we sum up (112) for all epochs $1\le s\le S$ to finish the proof as follows:

$$\mathbb{E}[\Phi(\widetilde{x}^S)-\Phi^*]+\sum_{s=1}^{S}\frac{c_m}{\widetilde{\alpha}^s}\left\|\tilde{x}^s-\tilde{x}^{s-1}\right\|^2 \le \widetilde{\alpha}^S\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*]+\widetilde{\alpha}^S\sum_{s=1}^{S}\frac{1}{\widetilde{\alpha}^s}\frac{1-\widetilde{\alpha}}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B}+\widetilde{\alpha}^S\sum_{s=1}^{S}\frac{1}{\widetilde{\alpha}^s}\frac{1-\widetilde{\alpha}}{1-\alpha}\eta\frac{L^2d^2\mu^2}{2}$$

$$= \alpha^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*]+\frac{1-\widetilde{\alpha}^S}{1-\widetilde{\alpha}}\frac{1-\widetilde{\alpha}}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B}+\frac{1-\widetilde{\alpha}^S}{1-\widetilde{\alpha}}\frac{1-\widetilde{\alpha}}{1-\alpha}\eta\frac{L^2d^2\mu^2}{2}$$

$$\le \alpha^{Sm}\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*]+\frac{1}{1-\alpha}\frac{I\{B<n\}\eta\sigma^2}{B}+\frac{1}{1-\alpha}\eta\frac{L^2d^2\mu^2}{2}$$

$$= (1-2\mu\gamma)^{Sm}\,\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*]+\frac{I\{B<n\}\eta\sigma^2}{2\mu\gamma B}+\eta\frac{L^2d^2\mu^2}{2\gamma\mu} \tag{113}$$

$$= (1-2\mu\gamma)^{Sm}\,\mathbb{E}[\Phi(\widetilde{x}^0)-\Phi^*]+\frac{I\{B<n\}\eta\sigma^2}{2\mu\gamma B}+\eta\frac{L^2d^2\mu^2}{2\gamma\mu}=3\epsilon \tag{114}$$

where (113) holds since $\alpha=1-\frac{\mu}{18L}$ $\alpha=1-\frac{\mu}{18Ld}$, and (114) uses $\eta=\frac{1}{6L}\eta=\frac{1}{6Ld}$.

5 Proof Under Form 8

First, similar to [Reddi et al., 2016b], we need the following inequality:

$$\Phi(\overline{x}_t^s)=f(\overline{x}_t^s)+h(\overline{x}_t^s)+h(x_{t-1}^s)-h(x_{t-1}^s)$$

$$\le f(x_{t-1}^s)+\left\langle\nabla f(x_{t-1}^s),\overline{x}_t^s-x_{t-1}^s\right\rangle+\frac{L}{2}\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+h(\overline{x}_t^s)+h(x_{t-1}^s)-h(x_{t-1}^s) \tag{115}$$

$$=\Phi(x_{t-1}^s)+\left\langle\nabla f(x_{t-1}^s),\overline{x}_t^s-x_{t-1}^s\right\rangle+\frac{L}{2}\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+h(\overline{x}_t^s)-h(x_{t-1}^s) \tag{116}$$

$$\le\Phi(x_{t-1}^s)+\left\langle\nabla f(x_{t-1}^s),\overline{x}_t^s-x_{t-1}^s\right\rangle+\frac{1}{2\eta}\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2+h(\overline{x}_t^s)-h(x_{t-1}^s) \tag{117}$$

$$=\Phi(x_{t-1}^s)-\frac{\eta}{2}D_h(x_{t-1}^s,\frac{1}{\eta}) \tag{118}$$

$$\le\Phi(x_{t-1}^s)-\eta\mu(\Phi(x_{t-1}^s)-\Phi^*) \tag{119}$$

where (115) holds since f has L-Lipschitz continuous gradient, (117) holds due to $\eta=\frac{1}{6L}<\frac{1}{L}$, (118) follows from the definition of D_h and recall $\overline{x}_t^s=\mathrm{Prox}_{\eta h}(x_{t-1}^s-\eta\nabla f(x_{t-1}^s))$, and (119) follows from the definition of PL condition with from (**??**).

Then, adding $\frac{9}{11}$ times (31) and $\frac{2}{11}$ times (119), we have

$$\Phi(\overline{x}_t^s)\le\Phi(x_{t-1}^s)-\frac{9}{11}\left(\frac{1}{\eta}-\frac{L}{2}\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2-\frac{2}{11}\eta\mu(\Phi(\overline{x}_{t-1}^s)-\Phi^*)$$

$$\le\Phi(x_{t-1}^s)-\left(\frac{9}{11\eta}-\frac{9L}{22}\right)\left\|\overline{x}_t^s-x_{t-1}^s\right\|^2-\frac{2\eta\mu}{11}(\Phi(\overline{x}_{t-1}^s)-\Phi^*) \tag{120}$$

We add (120) and (30) to obtain the following inequality:

$$\begin{aligned}
\Phi(x_t^s) &\le \Phi(x_{t-1}^s) + \frac{L}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{9}{11\eta} - \frac{9L}{22} - \frac{L}{2}\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 - \frac{2\eta\mu}{11}(\Phi(x_{t-1}^s) - \Phi^*) \\
&- \frac{1}{\eta}\left\langle x_t^s - x_{t-1}^s, x_t^s - \overline{x}_t^s\right\rangle + \left\langle \nabla f(x_{t-1}^s) - v_{t-1}^s, x_t^s - \overline{x}_t^s\right\rangle \\
&= \Phi(x_{t-1}^s) + \frac{L}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{9}{11\eta} - \frac{9L}{22} - \frac{L}{2}\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 - \frac{2\eta\mu}{11}(\Phi(x_{t-1}^s) - \Phi^*) \\
&- \frac{1}{2\eta}(\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) + \left\langle \nabla f(x_{t-1}^s) - v_{t-1}^s, x_t^s - \overline{x}_t^s\right\rangle \\
&= \Phi(x_{t-1}^s) - (\frac{1}{2\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{7}{22\eta} - \frac{10L}{11}\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 - \frac{2\eta\mu}{11}(\Phi(x_{t-1}^s) - \Phi^*) \\
&- \frac{1}{2\eta}\|x_t^s - \overline{x}_t^s\|^2 + \left\langle \nabla f(x_{t-1}^s) - v_{t-1}^s, x_t^s - \overline{x}_t^s\right\rangle \\
&\le \Phi(x_{t-1}^s) - (\frac{1}{2\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{7}{22\eta} - \frac{10L}{11}\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 - \frac{2\eta\mu}{11}(\Phi(x_{t-1}^s) - \Phi^*) \\
&- \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 + \left\langle \nabla f(x_{t-1}^s) - v_{t-1}^s, x_t^s - \overline{x}_t^s\right\rangle \qquad (121) \\
&= \Phi(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{5}{33\eta} - \frac{10L}{11}\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 - \frac{2\eta\mu}{11}(\Phi(x_{t-1}^s) - \Phi^*) \\
&+ \left\langle \nabla f(x_{t-1}^s) - v_{t-1}^s, x_t^s - \overline{x}_t^s\right\rangle \\
&\le \Phi(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{5}{33\eta} - \frac{10L}{11}\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 - \frac{2\eta\mu}{11}(\Phi(x_{t-1}^s) - \Phi^*) \\
&+ \eta\left\|\nabla f(x_{t-1}^s) - v_{t-1}^s\right\|^2 \qquad (122)
\end{aligned}$$

In the same way as (**??**) and (32), (121) uses Young's inequality (**??**) (choose $\alpha = 3$) and (122) follows from Lemma **??**.

Now, we take expectations for (122) and then plug the variance bound (37) into it to obtain

$$\begin{aligned}
&\mathbb{E}[\Phi(x_t^s)] \\
&\le \mathbb{E}\Bigg[\Phi(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{5}{33\eta} - \frac{10L}{11}\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 - \frac{2\eta\mu}{11}(\Phi(x_{t-1}^s) - \Phi^*) \\
&+ \frac{\eta L^2}{b}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B < n\}\eta\sigma^2}{B}\Bigg] \\
&= \mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13L}{4}\left\|x_t^s - x_{t-1}^s\right\|^2 - \frac{\mu}{33L}(\Phi(x_{t-1}^s) - \Phi^*) + \frac{L}{6b}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B < n\}\eta\sigma^2}{B}\right] \qquad (123) \\
&\le \mathbb{E}\left[\Phi(x_{t-1}^s) - \frac{13L}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \frac{\mu}{33L}(\Phi(x_{t-1}^s) - \Phi^*) + \left(\frac{L}{6b} + \frac{13L}{8t-4}\right)\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B < n\}\eta\sigma^2}{B}\right] \qquad (124)
\end{aligned}$$

where (123) uses $\eta = \frac{1}{6L}$, and (124) uses Young's inequality by choosing $\alpha = 2t - 1$.

Now, according to (124), we obtain the following key inequality

$$\mathbb{E}[\Phi(x_t^s) - \Phi^*] \tag{125}$$

$$\leq \mathbb{E}\left[(1 - \frac{\mu}{33L})(\Phi(x_{t-1}^s) - \Phi^*) - \frac{13L}{8t}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 + \left(\frac{L}{6b} + \frac{13L}{8t-4}\right)\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{I\{B<n\}\eta\sigma^2}{B}\right] \tag{126}$$

The remaining proof is exactly the same as our proof in Appendix B.1 from (74) to the end.

6 Strongly Convex with Momentum Acceleration

Algorithm 3 ZO-PROXSVRG for convex Optimization

1: **Input:** initial point x_0, batch size B, minibatch size b, epoch length m, step size η
2: **Initialize:** $\tilde{x}^0 = x_0$
3: **for** $s = 1, 2, \ldots, S$ **do**
4: $\quad x_0^s = x_m^{s-1}$
5: $\quad \hat{g}^s = \frac{1}{B}\sum_{j\in I_B} \hat{\nabla} f_j(\widetilde{x}^{s-1})$
6: $\quad$ **for** $t = 1, 2, \ldots, m$ **do**
7: $\quad\quad y_{t-1} = \theta x_{t-1}^s + (1-\theta)\widetilde{x}^{s-1}$
8: $\quad\quad v_{t-1}^s = \frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(y_{t-1}) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) + \hat{g}^s$
9: $\quad\quad x_t^s = \mathrm{Prox}_{\eta h}(x_{t-1}^s - \eta\hat{v}_{t-1}^s)$
10: $\quad \tilde{x}^s = \frac{\theta}{m}\sum_{j=1}^m x_j^s + (1-\theta)\tilde{x}^{s-1}$
11: **Output:** $\tilde{x}_S$