Supplementary Material of Parallel Asynchronous Stochastic Variance Reduction for Nonconvex Optimization

Cong Fang, Zhouchen Lin*
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, P. R. China
Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, P. R. China
fangcong@pku.edu.cn, zlin@pku.edu.cn

I. Proofs

For convenience, we will use $\tilde{\mathbf{x}}^s$ to denote $\mathbf{x}_0^s$ in all proofs. To proof convergence, we need to bound the variance. In serial SVRG [1], [2], the variance is bounded through Eq. (4) in the paper. In ASVRG, Eq. (4) changes to Eq. (8) in the paper. We then use Lemma 1 to build the relation between $\mathbf{x}_{j(k)}^s$ and $\mathbf{x}_k^s$. From Lemma 1, we obtain

$$\mathbb{E}\left(\|\mathbf{v}_k^s\|^2\right) \le \mathbb{E}\left(\|\nabla f(\mathbf{x}_{j(k)}^s)\|^2\right) + L^2\mathbb{E}\left(\|\mathbf{x}_{j(k)}^s - \tilde{\mathbf{x}}^s\|^2\right) \le \rho_2\left[\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right) + L^2\mathbb{E}\left(\|\mathbf{x}_k^s - \tilde{\mathbf{x}}^s\|^2\right)\right], \tag{1}$$

where ρ_2 is a constant depending on the step size γ and the delay parameter τ. Comparing Eq. (1) with Eq. (4) in the paper, we can find that the upper bound of the variance in ASVRG is exactly ρ_2 times of the upper bounded in SVRG. So by setting a special γ, we obtain a tight convergence property through the technique of SVRG [2].

Proof of Eq. (4) in the paper. For compleness, we first include the proof of Eq. (4) in the paper. It is taken from [1], [2].

$$\begin{aligned}
& \mathbb{E}_{i_k}\left(\|\mathbf{v}_k^s\|^2\right) \\
= \;& \mathbb{E}_{i_k}\left(\|\nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\tilde{\mathbf{x}}^s) + \mathbf{g}^s - \nabla f(\mathbf{x}_k^s) + \nabla f(\mathbf{x}_k^s)\|^2\right) \\
= \;& \mathbb{E}_{i_k}\left(\|\nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\tilde{\mathbf{x}}^s) + \mathbf{g}^s - \nabla f(\mathbf{x}_k^s)\|^2\right) + \mathbb{E}_{i_k}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right) \\
& +2\mathbb{E}_{i_k}\left(\langle\nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\tilde{\mathbf{x}}^s) + \mathbf{g}^s - \nabla f(\mathbf{x}_k^s), \nabla f(\mathbf{x}_k^s)\rangle\right) \\
= \;& \mathbb{E}_{i_k}\left(\|\nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\tilde{\mathbf{x}}^s) + \mathbf{g}^s - \nabla f(\mathbf{x}_k^s)\|^2\right) + \mathbb{E}_{i_k}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right) \\
= \;& \mathbb{E}_{i_k}\left(\|\nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\tilde{\mathbf{x}}^s) - \mathbb{E}_{i_k}\left(\nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\tilde{\mathbf{x}}^s)\right)\|^2\right) \\
& +\mathbb{E}_{i_k}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right) \\
\le \;& \mathbb{E}_{i_k}\|\left(\nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\tilde{\mathbf{x}}^s)\|^2\right) + \mathbb{E}_{i_k}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right) \\
\le \;& L^2\|\mathbf{x}_k^s - \tilde{\mathbf{x}}^s\|^2 + \|\nabla f(\mathbf{x}_k^s)\|^2,
\end{aligned} \tag{2}$$

where we use the fact that $\mathbb{E}_{i_k}\left(\nabla f_{i_k}(\mathbf{x})\right)) = \nabla f(\mathbf{x})$ in the third and the fourth equalities, and the first inequality in Eq. (2) follows from the fact that $\mathbb{E}(\big\|\xi - \mathbb{E}\left(\xi\right)\|^2) \le \mathbb{E}\left(\|\xi\|^2\right)$.

Proof of Lemma 1

We first analyse $\|\nabla f(\mathbf{x}_k^s)\|^2$ and $\|\mathbf{x}_k^s - \tilde{\mathbf{x}}^s\|^2$, respectively. For $\|\nabla f(\mathbf{x}_k^s)\|^2$, we have

$$\begin{aligned}
& \mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2 - \|\nabla f(\mathbf{x}_{k+1}^s)\|^2\right) \\
\le \;& 2\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|\|\nabla f(\mathbf{x}_k^s) - \nabla f(\mathbf{x}_{k+1}^s)\|\right) \\
\le \;& 2L\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|\|\mathbf{x}_k^s - \mathbf{x}_{k+1}^s\|\right) \\
\le \;& L\gamma\mathbb{E}\left(\frac{1}{C_1}\|\nabla f(\mathbf{x}_k^s)\|^2 + C_1\|\mathbf{v}_{j(k)}^s\|^2\right) \quad (C_1 > 0),
\end{aligned} \tag{3}$$

[*] Corresponding author.

where we use the fact that $\|a\|^2 - \|b\|^2 \leq 2\|a\|\|a-b\|$ [3] in the first inequality and the Cauchy-Schwarz inequality in the third inequality. In the same way, we have

$$\begin{aligned} & \mathbb{E}\left(\|\mathbf{x}_k^s - \tilde{\mathbf{x}}^s\|^2 - \|\mathbf{x}_{k+1}^s - \tilde{\mathbf{x}}^s\|^2\right) \\ \leq\ & 2\mathbb{E}\left(\|\mathbf{x}_k^s - \tilde{\mathbf{x}}^s\|\|\mathbf{x}_k^s - \mathbf{x}_{k+1}^s\|\right) \\ \leq\ & \gamma\mathbb{E}\left(\frac{1}{C_2}\|\mathbf{x}_k^s - \tilde{\mathbf{x}}^s\|^2 + C_2\|\mathbf{v}_{j(k)}^s\|^2\right) \quad (C_2 > 0)\,, \end{aligned} \tag{4}$$

Similarly, we have

$$\begin{aligned} & \mathbb{E}\left(\|\nabla f(\mathbf{x}_{j(k+1)}^s)\|^2 - \|\nabla f(\mathbf{x}_{k+1}^s)\|^2\right) \\ \leq\ & 2\mathbb{E}\left(\|\nabla f(\mathbf{x}_{j(k+1)}^s)\|\|\nabla f(\mathbf{x}_{k+1}^s) - \nabla f(\mathbf{x}_{j(k+1)}^s)\|\right) \\ \leq\ & \frac{L\gamma}{C_3}\mathbb{E}\left(\|\nabla f(\mathbf{x}_{j(k+1)}^s)\|^2\right) + \frac{LC_3}{\gamma}\mathbb{E}\left(\|\mathbf{x}_{k+1}^s - \mathbf{x}_{j(k+1)}^s\|^2\right) (C_3 > 0)\,, \end{aligned} \tag{5}$$

and

$$\begin{aligned} & \mathbb{E}\left(\|\mathbf{x}_{j(k+1)}^s - \tilde{\mathbf{x}}^s\|^2 - \|\mathbf{x}_{k+1}^s - \tilde{\mathbf{x}}^s\|^2\right) \\ \leq\ & 2\mathbb{E}\left(\|\mathbf{x}_{j(k+1)}^s - \tilde{\mathbf{x}}^s\|\|\mathbf{x}_{j(k+1)}^s - \mathbf{x}_{k+1}^s\|\right) \\ \leq\ & \frac{\gamma}{C_4}\mathbb{E}\left(\|\mathbf{x}_{j(k+1)}^s - \tilde{\mathbf{x}}^s\|^2\right) + \frac{C_4}{\gamma}\mathbb{E}\left(\|\mathbf{x}_{j(k+1)}^s - \mathbf{x}_{k+1}^s\|^2\right) \quad (C_4 > 0)\,. \end{aligned} \tag{6}$$

For convenience, set $B_k = \mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2 + L^2\|\mathbf{x}_k^s - \tilde{\mathbf{x}}^s\|\right)$, which has omitted the superscript s. Then from Eq. (2), we have

$$E\left(\|\mathbf{v}_k^s\|^2\right) \leq B_k. \tag{7}$$

Similarly, we set $B_{j(k)} = \mathbb{E}\left(\|\nabla f(\mathbf{x}_{j(k)}^s)\|^2 + L^2\|\mathbf{x}_{j(k)}^s - \tilde{\mathbf{x}}^s\|\right)$. In the same way, we have

$$E\left(\|\mathbf{v}_{j(k)}^s\|^2\right) \leq B_{j(k)}. \tag{8}$$

Multiplying Eq. (4) by L^2 and then adding Eq. (3), we have

$$\begin{aligned} B_k - B_{k+1} \quad \leq \quad & \mathbb{E}\left(L\gamma\frac{\|\nabla f(\mathbf{x}_k^s)\|^2}{C_1} + L\gamma C_1\|\mathbf{v}_{j(k)}^s\|^2\right) \\ & + \mathbb{E}\left(\frac{L^2\gamma}{C_2}\|\mathbf{x}_k^s - \tilde{\mathbf{x}}^s\|^2 + L^2\gamma C_2\|\mathbf{v}_{j(k)}^s\|^2\right). \end{aligned} \tag{9}$$

Set $C_2 = C_1/L$, we have

$$\begin{aligned} B_k - B_{k+1} \quad \leq \quad & \frac{L\gamma}{C_1}B_k + 2L\gamma C_1\mathbb{E}\left(\|\mathbf{v}_{j(k)}^s\|^2\right) \\ \leq \quad & \frac{L\gamma}{C_1}B_k + 2L\gamma C_1 B_{j(k)}, \end{aligned} \tag{10}$$

where we use Eq. (7) in the first inequality and Eq. (8) in the second inequality.

Now we use induction to prove that

$$B_{k-1} \leq \rho_1 B_k, \tag{11}$$

and

$$B_{j(k)} \leq \rho_2 B_k. \tag{12}$$

Suppose $k = 1$. Since $B_{j(0)} = B_0$. we set C_1 to be $\frac{1}{\sqrt{2}}$. Then from Eq. (10), we have

$$B_0 - B_1 \leq 2\sqrt{2}L\gamma B_0. \tag{13}$$

Simplifying Eq. (13), we have

$$B_0 \leq \frac{1}{1 - 2\sqrt{2}L\gamma}B_1. \tag{14}$$

Recalling the assumption on γ,

$$L\gamma \leq \frac{\rho_1 - 1}{2\sqrt{2}\rho_1\sqrt{\rho_2}} \leq \frac{\rho_1 - 1}{2\sqrt{2}\rho_1} = \frac{1}{2\sqrt{2}}(1 - \frac{1}{\rho_1}) < \frac{1}{2\sqrt{2}}, \tag{15}$$

so

$$B_0 \leq \frac{1}{1-2\sqrt{2}L\gamma}B_1 \leq \rho_1 B_1. \tag{16}$$

On the other hand, multiplying Eq. (6) by L^2 and then adding Eq. (5), we have

$$\begin{aligned} B_{j(k+1)} - B_{k+1} \leq & \frac{L\gamma}{C_3}\mathbb{E}\left(\|\nabla f(\mathbf{x}^s_{j(k+1)})\|^2\right) + \frac{LC_3}{\gamma}\mathbb{E}\left(\mathbf{x}^s_{k+1} - \mathbf{x}^s_{j(k+1)}\|^2\right) \\ & + \frac{L^2\gamma}{C_4}\mathbb{E}\left(\|\mathbf{x}^s_{j(k+1)} - \tilde{\mathbf{x}}^s\|^2\right) + \frac{C_4L^2}{\gamma}\mathbb{E}\left(\|\mathbf{x}^s_{k+1} - \mathbf{x}^s_{j(k+1)}\|^2\right). \end{aligned} \tag{17}$$

Setting $C_4 = C_3/L$, we have

$$B_{j(k+1)} - B_{k+1} \quad \leq \quad \frac{L\gamma}{C_3}B_{j(k+1)} + 2\frac{LC_3}{\gamma}\mathbb{E}\left(\|\mathbf{x}^s_{k+1} - \mathbf{x}^s_{j(k+1)}\|^2\right). \tag{18}$$

When $k = 1$,

$$\begin{aligned} & \mathbb{E}\left(\|\mathbf{x}^s_1 - \mathbf{x}^s_{j(1)}\|^2\right) \\ \leq \quad & \gamma^2\mathbb{E}\left(\|I_{0(0)}(\mathbf{v}^s_0)\|^2\right) \\ \leq \quad & \gamma^2\mathbb{E}\left(\|\mathbf{v}^s_0\|^2\right) \\ \leq \quad & \gamma^2\rho_1 B_1, \end{aligned} \tag{19}$$

where $I_{0(0)}$ is the function that indicates whether the elements of $\mathbf{v}^s_0$ have been written into $\mathbf{x}^s_1$ and we use Eq.(7) in the third inequality. Substituting Eq.(19) into Eq.(18), we have

$$B_{j(1)} - B_1 \leq \frac{L\gamma}{C_3}B_{j(1)} + 2\gamma LC_3\rho_1 B_1. \tag{20}$$

Setting C_3 to be $\frac{1}{\sqrt{2\rho_1}}$,

$$B_{j(1)} - B_1 \leq \sqrt{2\rho_1}L\gamma B_{j(1)} + \sqrt{2\rho_1}L\gamma B_1. \tag{21}$$

Then

$$B_{j(1)} \leq \frac{1+\sqrt{2\rho_1}L\gamma}{1-\sqrt{2\rho_1}L\gamma}B_1. \tag{22}$$

Recalling the assumption on γ, we have

$$L\gamma \leq \frac{\rho_2 - 1}{2\sqrt{2\rho_1}\rho_2^{\frac{3}{2}}\frac{\rho_1^{\frac{\tau}{2}}-1}{\sqrt{\rho_1}-1}} \leq \frac{\rho_2-1}{2\sqrt{2\rho_1}\rho_2}. \tag{23}$$

So we have

$$2\sqrt{2\rho_1}L\gamma \leq 1 - \frac{1}{\rho_2} < 1. \tag{24}$$

Then

$$\begin{aligned} & B_{j(1)} \\ \leq \quad & \frac{1+\sqrt{2\rho_1}L\gamma}{1-\sqrt{2\rho_1}L\gamma}B_1 \\ \leq \quad & \frac{1}{1-2\sqrt{2\rho_1}L\gamma}B_1 \\ \leq \quad & \rho_2 B_1, \end{aligned} \tag{25}$$

where we use the fact that $\frac{1+x}{1-x} \leq \frac{1}{1-2x}$ when $2x < 1$ in the second inequality.

When B_k satisfies Eq. (11) and Eq. (12), we consider B_{k+1}. From Eq. (10),

$$\begin{aligned} B_k - B_{k+1} \quad \leq \quad & \frac{L\gamma}{C_1}B_k + 2LC_1\gamma B_{j(k)} \\ \leq \quad & \frac{L\gamma}{C_1}B_k + 2LC_1\gamma\rho_2 B_k. \end{aligned} \tag{26}$$

Setting $C_1 = \frac{1}{\sqrt{2\rho_2}}$, we have

$$B_k - B_{k+1} \leq 2\sqrt{2\rho_2}L\gamma B_k. \tag{27}$$

Then

$$B_k \leq \frac{1}{1-2\sqrt{2\rho_2}L\gamma}B_{k+1}. \tag{28}$$

From the assumption on γ, we have $B_k \leq \rho_1 B_{k+1}$. Now we prove $B_{j(k+1)} \leq \rho_2 B_{k+1}$. We first analyse $\|\mathbf{x}_{k+1}^s - \mathbf{x}_{j(k+1)}^s\|$.

$$\begin{aligned}
& \mathbb{E}\left(\|\mathbf{x}_{k+1}^s - \mathbf{x}_{j(k+1)}^s\|^2\right) \\
= \ & \gamma^2\mathbb{E}\left(\|\sum_{l=k-\tau+1}^{k} I_{k(l)}\left(\mathbf{v}_{j(l)}^s\right)\|^2\right) \\
\leq \ & \gamma^2\mathbb{E}\left(\sum_{p=1}^{d}\left(\sum_{l=k-\tau+1}^{k}|\mathbf{v}_{j(l)}^s(p)|\right)^2\right). \\
\leq \ & \gamma^2\mathbb{E}\left(\sum_{p=1}^{d}\left(\sum_{i=0}^{\tau-1}\sum_{z=0}^{\tau-1}|\mathbf{v}_{j(k-i)}^s(p)| \times |\mathbf{v}_{j(k-z)}^s(p)|\right)\right),
\end{aligned} \tag{29}$$

where $\mathbf{v}_k^s(p)$ is the p-th coordinate of vector $\mathbf{v}_k^s$. The first inequality uses the inequality that $(a_1 + a_2 + \cdots + a_\tau)^2 \leq (|a_1| + |a_2| + \cdots + |a_\tau|)^2$ on each dimension. For any $i = 0, 1, \ldots, \tau-1$ and $z = 0, 1, \ldots, \tau-1$, we have

$$\begin{aligned}
& \mathbb{E}\left(\sum_{p=1}^{d}\left(2|\mathbf{v}_{j(k-i)}^s(p)| \times |\mathbf{v}_{j(k-z)}^s(p)|\right)\right) \\
\leq \ & \mathbb{E}\left(\sum_{p=1}^{d}\left(\rho_1^{(z-i)/2}|\mathbf{v}_{j(k-i)}^s(p)|^2 + \rho_1^{(i-z)/2}|\mathbf{v}_{j(k-z)}^s(p)|^2\right)\right) \qquad (30) \\
\leq \ & \mathbb{E}\left(\rho_1^{(z-i)/2}\|\mathbf{v}_{j(k-i)}^s\|^2 + \rho_1^{(i-z)/2}\|\mathbf{v}_{j(k-z)}^s\|^2\right) \\
\leq \ & \rho_1^{(z-i)/2}B_{j(k-i)} + \rho_1^{(i-z)/2}B_{j(k-z)} \\
\leq \ & \rho_2\rho_1^{(z-i)/2}\rho_1^{i}B_k + \rho_2\rho_1^{(i-z)/2}\rho_1^{z}B_k \\
\leq \ & 2\rho_2\rho_1^{(i+z)/2}B_k,
\end{aligned}$$

where we use Cauchy-Schwarz in the first inequality and Eq. (8) in the third inequality. So

$$\begin{aligned}
& \mathbb{E}\left(\|\mathbf{x}_{k+1}^s - \mathbf{x}_{j(k+1)}^s\|^2\right) \\
\leq \ & \gamma^2\mathbb{E}\left(\sum_{p=1}^{d}\left(\sum_{i=0}^{\tau-1}\sum_{z=0}^{\tau-1}|\mathbf{v}_{j(k-i)}^s(p)| \times |\mathbf{v}_{j(k-z)}^s(p)|\right)\right) \\
\leq \ & \gamma^2\rho_2\sum_{i=0}^{\tau-1}\sum_{z=0}^{\tau-1}\rho_1^{(i+z)/2}B_k \\
\leq \ & \gamma^2\rho_2\left(\sum_{i=0}^{k-\tau+1}\rho_1^{i/2}\right)^2 B_k \\
\leq \ & \gamma^2\rho_2\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}B_k.
\end{aligned} \tag{31}$$

Substituting Eq. (31) into Eq. (18), we have

$$\begin{aligned}
& B_{j(k+1)} - B_{k+1} \\
\leq \ & \frac{L\gamma}{C_3}B_{j(k+1)} + 2LC_3\gamma\rho_2\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}B_k \\
\leq \ & \frac{L\gamma}{C_3}B_{j(k+1)} + 2LC_3\gamma\rho_2\rho_1\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}B_{k+1}.
\end{aligned} \tag{32}$$

Setting $C_3 = \frac{1}{\sqrt{2\rho_1\rho_2}\frac{\rho_1^{\tau/2}-1}{\sqrt{\rho_1}-1}}$, we have

$$B_{j(k+1)} - B_{k+1} \leq L\gamma\sqrt{2\rho_1\rho_2}\frac{\rho_1^{\tau/2}-1}{\sqrt{\rho_1}-1}(B_{j(k+1)} + B_{k+1}). \tag{33}$$

Considering the assumption on γ, like Eq. (23), we have

$$2L\gamma\sqrt{2\rho_1\rho_2}\frac{\rho_1^{\tau/2}-1}{\sqrt{\rho_1}-1} \leq 1-\frac{1}{\rho_2} < 1, \tag{34}$$

then like Eq. (25), we have

$$\begin{aligned} & B_{j(k+1)} \\ \leq\; & \frac{1+L\gamma\sqrt{2\rho_1\rho_2}\frac{\rho_1^{\tau/2}-1}{\sqrt{\rho_1}-1}}{1-L\gamma\sqrt{2\rho_1\rho_2}\frac{\rho_1^{\tau/2}-1}{\sqrt{\rho_1}-1}}B_{k+1} \\ \leq\; & \frac{1}{1-2L\gamma\sqrt{2\rho_1\rho_2}\frac{\rho_1^{\tau/2}-1}{\sqrt{\rho_1}-1}}B_{k+1} \\ \leq\; & \rho_2 B_{k+1}. \end{aligned} \tag{35}$$

Proof of Theorem 1

The convergence property of ASVRG is inspired by [2]. We first check the conditions in Lemma 1. Since $\tau \leq n^{\alpha/2}$, and $\gamma = \mu/(Ln^{\alpha})$ $(0 < \mu \leq \frac{1}{8e(e-1)})$, we have $\tau \leq n^{\alpha/2} \leq n^{\alpha}$. Setting $\rho_1^{\tau/2} = e$ and $\rho_2 = 2$, we have

$$\begin{aligned} & \frac{\rho_1-1}{2\sqrt{2}\rho_1\rho_2^{1/2}} \\ =\; & \frac{e^{2/\tau}-1}{4\rho_1^{\frac{1}{2}\times 2}} \geq \frac{e^{2/\tau}-1}{4e^2} \\ \geq\; & \frac{1}{2e^2\tau} \geq \frac{1}{2e^2n^{\alpha}} \geq \frac{\mu}{n^{\alpha}} \\ =\; & L\gamma, \end{aligned} \tag{36}$$

where we use the fact that $e^x - 1 \geq x$ $(x \geq 0)$ in the third inequality. In the same way, we have

$$\begin{aligned} & \frac{\rho_2-1}{2\sqrt{2\rho_1}\rho_2^{\frac{3}{2}}\frac{\rho_1^{\tau/2}-1}{\sqrt{\rho_1}-1}} \\ \geq\; & \frac{\sqrt{\rho_1}-1}{8(\rho_1^{\tau/2}-1)\rho_1^{1/2}} \\ \geq\; & \frac{e^{1/\tau}-1}{8(e-1)e} \\ \geq\; & \frac{1}{8(e-1)e\tau} \\ \geq\; & \frac{1}{8(e-1)en^{\alpha}} \\ \geq\; & L\gamma. \end{aligned} \tag{37}$$

Thus Eq. (11) and Eq. (12) hold.

Now we are to bound $\mathbb{E}\left(f(\mathbf{x}_{k+1}^s)\right)$ and $\mathbb{E}\left(\|\mathbf{x}_{k+1}^s - \tilde{\mathbf{x}}^s\|^2\right)$. Since f have L-Lipschitz continuous gradients, we have

$$\begin{aligned} \mathbb{E}\left(f(\mathbf{x}_{k+1}^s)\right) \leq\; & \mathbb{E}\left(f(\mathbf{x}_k^s) + \left\langle \nabla f(\mathbf{x}_k^s), \mathbf{x}_{k+1}^s - \mathbf{x}_k^s\right\rangle + \frac{L}{2}\|\mathbf{x}_{k+1}^s - \mathbf{x}_k^s\|^2\right) \\ =\; & \mathbb{E}\left(f(\mathbf{x}_k^s) - \gamma\left\langle \nabla f(\mathbf{x}_k^s), \mathbf{v}_{j(k)}^s\right\rangle + \frac{L\gamma^2}{2}\|\mathbf{v}_{j(k)}^s\|^2\right) \\ =\; & \mathbb{E}\left(f(\mathbf{x}_k^s) - \gamma\left\langle \nabla f(\mathbf{x}_k^s), \nabla f(\mathbf{x}_{j(k)}^s)\right\rangle + \frac{L\gamma^2}{2}\|\mathbf{v}_{j(k)}^s\|^2\right) \\ =\; & \mathbb{E}\left(f(\mathbf{x}_k^s) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}_k^s)\|^2 - \frac{\gamma}{2}\|\nabla f(\mathbf{x}_{j(k)}^s)\|^2 + \frac{L\gamma^2}{2}\|\mathbf{v}_{j(k)}^s\|^2\right) \\ & + \frac{\gamma}{2}\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s) - \nabla f(\mathbf{x}_{j(k)}^s)\|^2\right), \end{aligned} \tag{38}$$

where we use $\mathbb{E}_{i(k)}\left(\mathbf{v}^s_{j(k)}\right) = \nabla f(\mathbf{x}^s_{j(k)})$ in the second equality, and in the last equality we apply the equality that $\langle a, b\rangle = \frac{\|a\|^2}{2} + \frac{\|b\|^2}{2} - \frac{\|a-b\|^2}{2}$ [4]. From Eq.(31), we have

$$\begin{aligned}
& \mathbb{E}\left(\|\nabla f(\mathbf{x}^s_k) - \nabla f(\mathbf{x}^s_{j(k)})\|^2\right) \\
\leq \quad & L^2\gamma^2\rho_2\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}B_{k-1} \\
\leq \quad & L^2\gamma^2\rho_1\rho_2\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}B_k. \qquad (39)
\end{aligned}$$

For $\|\mathbf{x}^s_{k+1} - \tilde{\mathbf{x}}^s\|^2$, we have

$$\begin{aligned}
\mathbb{E}\left(\|\mathbf{x}^s_{k+1} - \tilde{\mathbf{x}}^s\|^2\right) &= \mathbb{E}\left(\|\mathbf{x}^s_{k+1} - \mathbf{x}^s_k + \mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right) \\
&= \mathbb{E}\left(\|\mathbf{x}^s_{k+1} - \mathbf{x}^s_k\|^2\right) + \mathbb{E}\left(\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right) \\
&\quad +2\mathbb{E}\left(\langle \mathbf{x}^s_{k+1} - \mathbf{x}^s_k, \mathbf{x}^s_k - \tilde{\mathbf{x}}^s\rangle\right) \\
&= \mathbb{E}\left(\gamma^2\|\mathbf{v}^s_{j(k)}\|^2\right) + \mathbb{E}\left(\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right) \\
&\quad -2\mathbb{E}\left(\left\langle \gamma\mathbf{v}^s_{j(k)}, \mathbf{x}^s_k - \tilde{\mathbf{x}}^s\right\rangle\right) \\
&= \mathbb{E}\left(\gamma^2\|\mathbf{v}^s_{j(k)}\|^2\right) + \mathbb{E}\left(\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right) \\
&\quad -2\mathbb{E}\left(\left\langle \gamma\nabla f(\mathbf{x}^s_{j(k)}), \mathbf{x}^s_k - \tilde{\mathbf{x}}^s\right\rangle\right) \\
&\leq \mathbb{E}\left(\gamma^2\|\mathbf{v}^s_{j(k)}\|^2\right) + \mathbb{E}\left(\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right) \\
&\quad +\gamma\mathbb{E}\left(\frac{1}{C_5}\|\nabla f(\mathbf{x}^s_{j(k)})\|^2 + C_5\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right) \\
&= \gamma^2\mathbb{E}\left(\|\mathbf{v}^s_{j(k)}\|^2\right) + \frac{\gamma}{C_5}\mathbb{E}\left(\|\nabla f(\mathbf{x}^s_{j(k)})\|^2\right) \\
&\quad +(1+C_5\gamma)\mathbb{E}\left(\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right) \quad (C_5 > 0), \qquad (40)
\end{aligned}$$

where we use $\mathbb{E}_{ik}\left(\mathbf{v}^s_{j(k)}\right) = \nabla f(\mathbf{x}^s_{j(k)})$ in the third equality. Set $R^s_k = \mathbb{E}\left(f(\mathbf{x}^s_k) + D_k\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right)$. Multiplying Eq.(40) by D_{k+1}, then adding Eq.(38), we have

$$\begin{aligned}
R^s_{k+1} &\leq \mathbb{E}\left(f(\mathbf{x}^s_k) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}^s_k)\|^2 - \frac{\gamma}{2}\|\nabla f(\mathbf{x}^s_{j(k)})\|^2 + \frac{L\gamma^2}{2}\|\mathbf{v}^s_{j(k)}\|^2\right) \\
&\quad +\frac{L^2\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}B_k \\
&\quad +D_{k+1}\left(\gamma^2\mathbb{E}\left(\|\mathbf{v}^s_{j(k)}\|^2\right) + \frac{\gamma}{C_5}\mathbb{E}\left(\|\nabla f(\mathbf{x}^s_{j(k)})\|^2\right) + (1+C_5\gamma)\mathbb{E}\left(\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right)\right) \\
&\leq \mathbb{E}\left(f(\mathbf{x}^s_k)\right) + D_{k+1}(1+C_5\gamma)\mathbb{E}\left(\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right) \\
&\quad +\mathbb{E}\left(\frac{L\gamma^2\rho_2}{2} + \frac{L^2\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2} + \rho_2 D_{k+1}\gamma^2\right)B_k \\
&\quad -\mathbb{E}\left(\frac{\gamma}{2}\|\nabla f(\mathbf{x}^s_k)\|^2\right) - \mathbb{E}\left((\frac{\gamma}{2} - \frac{D_{k+1}\gamma}{C_5})\|\nabla f(\mathbf{x}^s_{j(k)})\|^2\right), \qquad (41)
\end{aligned}$$

where we use Eq.(39) in the first inequality and applies $\|\mathbf{v}^s_{j(k)}\|^2 \leq B_{j(k)} \leq \rho_2 B_k$ in the second inequality. Substituting $B_k = \mathbb{E}\left(\|\nabla f(\mathbf{x}^s_k)\|^2 + L^2\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|\right)$ into Eq. (41), we have

$$\begin{aligned}
R^s_{k+1} &\leq \mathbb{E}\left(f(\mathbf{x}^s_k)\right) \\
&\quad -\left(\frac{\gamma}{2} - \frac{L\gamma^2\rho_2}{2} - \frac{L^2\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2} - \rho_2 D_{k+1}\gamma^2\right)\mathbb{E}\left(\|\nabla f(\mathbf{x}^s_k)\|^2\right) \\
&\quad -\mathbb{E}\left((\frac{\gamma}{2} - \frac{D_{k+1}\gamma}{C_5})\|\nabla f(\mathbf{x}^s_{j(k)})\|^2\right) + D_{k+1}\left(1 + C_5\gamma + L^2\gamma^2\rho_2\right)\mathbb{E}\left(\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right) \\
&\quad +\left(\frac{L^3\gamma^2\rho_2}{2} + \frac{L^4\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}\right)\mathbb{E}\left(\|\mathbf{x}^s_k - \tilde{\mathbf{x}}^s\|^2\right). \qquad (42)
\end{aligned}$$

Setting $D_m = 0$ and

$$D_k = D_{k+1}\left(1 + C_5\gamma + L^2\gamma^2\rho_2\right) + \frac{L^3\gamma^2\rho_2}{2} + \frac{L^4\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}, \tag{43}$$

we have

$$\begin{aligned} R_{k+1}^s - R_k^s \quad &\leq \quad -\left(\frac{\gamma}{2} - \frac{L\gamma^2\rho_2}{2} - \frac{L^2\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2} - \rho_2 D_{k+1}\gamma^2\right)\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right) \\ &- \quad \left(\frac{\gamma}{2} - \frac{D_{k+1}\gamma}{C_5}\right)\mathbb{E}\left(\|\nabla f(\mathbf{x}_{j(k)}^s)\|^2\right). \end{aligned} \tag{44}$$

Since D_k is monotone decreasing, we have

$$\begin{aligned} R_{k+1}^s - R_k^s \quad &\leq \quad -\left(\frac{\gamma}{2} - \frac{L\gamma^2\rho_2}{2} - \frac{L^2\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2} - \rho_2 D_0\gamma^2\right)\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right) \\ &- \quad (\frac{\gamma}{2} - \frac{D_0\gamma}{C_5})\mathbb{E}\left(\|\nabla f(\mathbf{x}_{j(k)}^s)\|^2\right). \end{aligned} \tag{45}$$

Now we bound D_0. From Eq. (43), we have $D_0 = l_m\frac{(1+\theta)^m-1}{\theta}$, where $\theta = L^2\gamma^2\rho_2 + C_5\gamma$ and

$$\begin{aligned} l_m \quad &= \quad \frac{L^3\gamma^2\rho_2}{2} + \frac{L^4}{2}\gamma^3\rho_1\rho_2\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2} \\ &\leq \quad L^3\gamma^2 + \rho_1(e-1)^2L^4\gamma^3\frac{1}{(e^{1/\tau}-1)^2} \\ &\leq \quad L^3\gamma^2 + e^2(e-1)^2L^4\gamma^3\frac{1}{(e^{1/\tau}-1)^2} \\ &\leq \quad L^3\gamma^2 + e^2(e-1)^2L^4\gamma^3\tau^2 \\ &= \quad L^3\gamma^2 + (e^2(e-1)^2L\gamma\tau^2)L^3\gamma^2 \\ &= \quad L^3\gamma^2 + (e^2(e-1)^2L\frac{\mu}{Ln^\alpha}n^\alpha)L^3\gamma^2 \\ &\leq \quad L^3\gamma^2 + (e^2(e-1)^2\mu)L^3\gamma^2 \\ &\leq \quad 2L^3\gamma^2, \end{aligned} \tag{46}$$

where we use $\rho_1 \leq (\rho_1^{\tau/2})^2 \leq e^2$ in the second inequality. Set $C_5 = L/(\mu n^{\alpha/2})$. Then for θ, we have

$$\theta = \rho_2L^2\gamma^2 + C_5\gamma = \frac{2\mu^2}{n^{2\alpha}} + \frac{1}{n^{3\alpha/2}} \leq \frac{33}{32n^{3\alpha/2}}. \tag{47}$$

The above inequality holds since $\mu \leq 1/8$ and $n \geq 1$. Then

$$\begin{aligned} D_0 \quad &= \quad l_m\frac{(1+\theta)^m-1}{\theta} \\ &\leq \quad \frac{2L\mu^2}{n^{2\alpha}}\frac{(1+\theta)^m-1}{\theta} \\ &= \quad \frac{2L\mu^2}{n^{2\alpha}}\frac{(1+\theta)^m-1}{\frac{2\mu^2}{n^{2\alpha}}+\frac{1}{n^{3\alpha/2}}} \\ &= \quad \frac{2\mu^2L\left((1+\theta)^m-1\right)}{2\mu^2+n^{\alpha/2}} \\ &\leq \quad \frac{2\mu^2L\left(\left(1+\frac{33}{32n^{3\alpha/2}}\right)^{\left(1/\frac{1}{n^{3\alpha/2}}\right)}-1\right)}{2\mu^2+n^{\alpha/2}} \\ &\leq \quad 2n^{-\alpha/2}\mu^2L(e^{\frac{33}{32}}-1). \end{aligned} \tag{48}$$

Now we are to bound $\frac{\gamma}{2}-\frac{L\gamma^2\rho_2}{2}-\frac{L^2\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}-\rho_2 D_0\gamma^2$ and $\frac{\gamma}{2}-\frac{D_0\gamma}{C_5}$ in Eq. (45), respectively.

$$\begin{aligned}
& \frac{\gamma}{2}-\frac{L\gamma^2\rho_2}{2}-\frac{L^2\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}-\rho_2 D_0\gamma^2 \\
\geq\ & \frac{\gamma}{2}-L\gamma^2-L\gamma^2-2D_0\gamma^2 \\
=\ & \left(\frac{1}{2}-2L\gamma-2D_0\gamma\right)\gamma \\
\geq\ & \left(\frac{1}{2}-2\mu-4(e^{\frac{33}{32}}-1)\mu^3\right)\gamma \\
\geq\ & \frac{\nu}{Ln^{\alpha}},
\end{aligned} \tag{49}$$

where $\nu=\frac{\mu}{3}$ and we have used Eq. (46) to bound $\frac{L^2\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2}-1)^2}{(\sqrt{\rho_1}-1)^2}$ in the first inequality. For $\frac{\gamma}{2}-\frac{D_0\gamma}{C_5}$, we have

$$\begin{aligned}
& \frac{\gamma}{2}-\frac{D_0\gamma}{C_5} \\
\geq\ & \frac{\gamma}{2}\Big(1-\frac{4n^{-\alpha/2}\mu^2L(e^{\frac{33}{32}}-1)}{L/(\mu n^{\alpha/2})}\Big) \\
\geq\ & \frac{\gamma}{2}(1-4\mu^3(e^{\frac{33}{32}}-1)) \\
\geq\ & 0.
\end{aligned} \tag{50}$$

Substituting Eq. (49) and Eq. (50) into Eq. (45) , we have

$$\frac{\nu}{Ln^{\alpha}}\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right)\leq R_k^s-R_{k+1}^s \tag{51}$$

Summing k from 0 to $m-1$, we have

$$\frac{\nu}{Ln^{\alpha}}\sum_{k=0}^{m-1}\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right)\leq R_0^s-R_m^s. \tag{52}$$

Since $D_m=0$, we have

$$R_m^s=f(\mathbf{x}_m^s)=f(\mathbf{x}_0^{s+1}). \tag{53}$$

And for $\mathbf{x}_0^s=\tilde{\mathbf{x}}^s$, we have

$$R_0^s=f(\mathbf{x}_0^s). \tag{54}$$

So Eq. (52) can be rewritten as

$$\frac{\nu}{Ln^{\alpha}}\sum_{k=0}^{m-1}\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right)\leq f(\mathbf{x}_0^s)-f(\mathbf{x}_0^{s+1}). \tag{55}$$

Then summing s from 0 to $S-1$, we have

$$\frac{1}{K}\sum_{s=0}^{S-1}\sum_{k=0}^{m-1}\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right)\leq\frac{n^{\alpha}(f(\mathbf{x}_0^0)-f(\mathbf{x}_0^S))}{K\nu}\leq\frac{n^{\alpha}(f(\mathbf{x}_0^0)-f(\mathbf{x}^*))}{K\nu}, \tag{56}$$

where $K=mS$ and $f(\mathbf{x}^*)$ is the minimal value of $f(\mathbf{x})$.

Proof of Theorem 2

We still first check the conditions in Lemma 1. Recall $\gamma=\mu/Ln^{\alpha}\tau^{\beta}$ with $0<\mu\leq\frac{1}{8(e-1)e}$, $0<\alpha\leq 1$, and $0<\beta\leq 1$, and $m=n^{\frac{3\alpha}{2}}\tau^{\frac{3\beta-1}{2}}$. We still set $\rho_1^{\tau/2}=e$ and $\rho_2=2$. Since $\tau\leq m$, we have

$$\tau\leq n^{\frac{3\alpha}{2}}\tau^{\frac{3\beta-1}{2}}. \tag{57}$$

Rearranging the terms in Eq. (57), we have

$$\tau^3\leq n^{3\alpha}\tau^{3\beta}. \tag{58}$$

So

$$\frac{1}{\tau}\geq\frac{1}{n^{\alpha}\tau^{\beta}}. \tag{59}$$

Further, we have

$$\mu \geq L\tau\gamma. \tag{60}$$

Now we analyse the condition of γ. Like Eq. (36) and Eq. (37), we have

$$\begin{aligned} & \frac{\rho_1 - 1}{2\sqrt{2}\rho_1\rho_2^{1/2}} \\ \geq\ & \frac{e^{2/\tau} - 1}{4\rho_1^{\frac{1}{2}\times 2}} \geq \frac{e^{2/\tau} - 1}{4e^2} \\ \geq\ & \frac{1}{2e^2\tau} \geq \frac{1}{2e^2\mu} L\gamma \\ \geq\ & L\gamma, \end{aligned} \tag{61}$$

and

$$\begin{aligned} & \frac{\rho_2 - 1}{2\sqrt{2\rho_1}\rho_2^{\frac{3}{2}} \frac{\rho_1^{\tau/2} - 1}{\sqrt{\rho_1} - 1}} \\ \geq\ & \frac{\sqrt{\rho_1} - 1}{8(\rho_1^{\tau/2} - 1)\rho_1^{1/2}} \\ \geq\ & \frac{e^{1/\tau} - 1}{8(e-1)e} \\ \geq\ & \frac{1}{8(e-1)e\tau} \\ \geq\ & \frac{L\gamma}{8(e-1)e\mu} \\ =\ & L\gamma, \end{aligned} \tag{62}$$

Thus Eq. (11) and Eq. (12) hold. Since Eq. (43) and Eq. (45) hold, we bound D_0, where $D_0 = l_m \frac{(1+\theta)^m - 1}{\theta}$ and $\theta = L^2\rho_2\gamma^2 + C_5\gamma$. We set $C_5 = \frac{L}{\mu n^{\frac{\alpha}{2}} \tau^{\frac{\beta - 1}{2}}}$. Then we have

$$\theta = 2L^2\gamma^2 + C_5\gamma = \frac{2\mu^2}{n^{2\alpha}\tau^{2\beta}} + \frac{1}{n^{3\alpha/2}\tau^{\frac{3\beta-1}{2}}} \leq \frac{33}{32n^{3\alpha/2}\tau^{\frac{3\beta-1}{2}}}. \tag{63}$$

For l_m, we have

$$\begin{aligned} l_m &= \frac{L^3\gamma^2\rho_2}{2} + \frac{L^4}{2}\gamma^3\rho_1\rho_2 \frac{(\rho_1^{\tau/2} - 1)^2}{(\sqrt{\rho_1} - 1)^2} \\ &\leq L^3\gamma^2 + \rho_1(e-1)^2 L^4\gamma^3 \frac{1}{(e^{1/\tau} - 1)^2} \\ &\leq L^3\gamma^2 + e^2(e-1)^2 L^4\gamma^3 \frac{1}{(e^{1/\tau} - 1)^2} \\ &\leq L^3\gamma^2 + e^2(e-1)^2 L^4\gamma^3\tau^2 \\ &\leq L^3\gamma^2 + \mu e^2(e-1)^2 L^3\gamma^2\tau \\ &\leq L^3\gamma^2\tau + L^3\gamma^2\tau = 2L^3\gamma^2\tau \\ &= \frac{2L\mu^2\tau}{n^{2\alpha}\tau^{2\beta}}, \end{aligned} \tag{64}$$

where we use Eq. (60) in the fourth inequality. So

$$\begin{aligned}
D_0 &= l_m \frac{(1+\theta)^m - 1}{\theta} \\
&\leq \frac{2L\mu^2\tau}{n^{2\alpha}\tau^{2\beta}} \frac{(1+\theta)^m - 1}{\theta} \\
&= \frac{2L\mu^2\tau}{n^{2\alpha}\tau^{2\beta}} \frac{(1+\theta)^m - 1}{\frac{2\mu^2}{n^{2\alpha}\tau^{2\beta}} + \frac{1}{n^{3\alpha/2}\tau^{\frac{3\beta-1}{2}}}} \\
&= \frac{2\mu^2 L\tau\left((1+\theta)^m - 1\right)}{2\mu^2 + n^{\alpha/2}\tau^{\frac{\beta+1}{2}}} \\
&\leq \frac{2\mu^2\tau L\left(\left(1 + \frac{33\mu}{32n^{3\alpha/2}\tau^{\frac{3\beta-1}{2}}}\right)^{\left(1/\frac{\mu}{n^{3\alpha/2}\tau^{\frac{3\beta-1}{2}}}\right)} - 1\right)}{2\mu^2 + n^{\alpha/2}\tau^{\frac{\beta+1}{2}}} \\
&\leq \frac{2\mu^2 L(e^{\frac{33}{32}} - 1)}{n^{\frac{\alpha}{2}}\tau^{\frac{\beta-1}{2}}}.
\end{aligned} \tag{65}$$

Then

$$\begin{aligned}
&\frac{\gamma}{2} - \frac{L\gamma^2\rho_2}{2} - \frac{L^2\gamma^3\rho_1\rho_2}{2}\frac{(\rho_1^{\tau/2} - 1)^2}{(\sqrt{\rho_1} - 1)^2} - \rho_2 D_0\gamma^2 \\
\geq\ & \frac{\gamma}{2} - L\gamma^2 - L^2\gamma^3 e^2(e-1)^2\tau^2 - 2D_0\gamma^2 \\
\geq\ & \left(\frac{1}{2} - L\gamma - e^2(e-1)^2\mu^2 - 2D_0\gamma\right)\gamma \\
\geq\ & \left(\frac{1}{2} - L\gamma\tau - e^2(e-1)^2\mu^2 - \frac{4\mu^2 L(e^{\frac{33}{32}} - 1)}{n^{\frac{\alpha}{2}}\tau^{\frac{\beta-1}{2}}}\gamma\right)\gamma \\
\geq\ & \left(\frac{1}{2} - L\gamma\tau - e^2(e-1)^2\mu^2 - \frac{4\mu^2(e^{\frac{33}{32}} - 1)L\tau\gamma}{n^{\frac{\alpha}{2}}\tau^{\frac{\beta}{2}}}\right)\gamma \\
\geq\ & \left(\frac{1}{2} - \mu - e^2(e-1)^2\mu^2 - 4\mu^3(e^{\frac{33}{32}} - 1)\right)\gamma \\
\geq\ & \frac{1}{3}\gamma = \frac{\nu}{Ln^\alpha\tau^\beta},
\end{aligned} \tag{66}$$

where $\nu = \frac{\mu}{3}$. The second and the fifth inequalities use Eq. (60). We substitute Eq. (65) into the third inequality. The fourth inequality follows the fact that $\tau^{\frac{1}{2}} \leq \tau$. Besides,

$$\begin{aligned}
&\frac{\gamma}{2} - \frac{D_0\gamma}{C_5} \\
\geq\ & \frac{\gamma}{2}\left(1 - 2\frac{2\mu L(e^{\frac{33}{32}} - 1)}{n^{\frac{\alpha}{2}}\tau^{\frac{\beta-1}{2}}}\frac{n^{\frac{\alpha}{2}}\tau^{\frac{\beta-1}{2}}}{L}\right) \\
\geq\ & \frac{\gamma}{2}(1 - 4\mu(e^{\frac{33}{32}} - 1)) \\
\geq\ & 0.
\end{aligned} \tag{67}$$

So like Eq. (51), we have

$$\frac{\nu}{Ln^\alpha\tau^\beta}\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right) \leq R_k^s - R_{k+1}^s. \tag{68}$$

Then like Eq. (56), we get

$$\frac{1}{K}\sum_{s=0}^{S-1}\sum_{k=0}^{m-1}\mathbb{E}\left(\|\nabla f(\mathbf{x}_k^s)\|^2\right) \leq \frac{n^\alpha\tau^\beta(f(\mathbf{x}_0^0) - f(\mathbf{x}_0^S))}{K\nu} \leq \frac{n^\alpha\tau^\beta(f(\mathbf{x}_0^0) - f(\mathbf{x}^*))}{K\nu}, \tag{69}$$

where $K = mS$ and $f(\mathbf{x}^*)$ is the minimal value of $f(\mathbf{x})$.

Proof of Theorem 3

Theorem 3 can be obtained by the Markov inequality directly. For condition 1, when $K \geq \frac{n^{\alpha}\left(f(\mathbf{x}^0)-f(\mathbf{x}^*)\right)}{\nu\epsilon\eta}$, from the Markov's inequality, we have

$$\begin{aligned} & \mathcal{P}\left(\frac{1}{K}\sum_{s=0}^{S-1}\sum_{k=0}^{m-1}\|\nabla F(\mathbf{x}_k^s)\|^2 \geq \epsilon\right) \\ \leq & \ \epsilon^{-1}\frac{1}{K}\sum_{s=0}^{S-1}\sum_{k=0}^{m-1}\mathbb{E}\left(\|\nabla F(\mathbf{x}_k^s)\|^2\right) \\ \leq & \ \eta. \end{aligned} \tag{70}$$

Case 2 can be obtained similarly.

II. Experimental Results of ASVRG-atom

In this section, we demonstrate the speedup property of ASVRG-atom. The curves of objective loss against iterations and running time on MNIST and CIFAR10 are drawn in Figures 1, 2, 3, and 4, respectively. We report their speedup in Tables I and II, respectively. The results on one core SVRG, one core SGD and 12 cores SGD are directly taken from the experiment of ASVRG-wild.

TABLE I
Iteration and running time speedup of ASVRG-atom on MNIST.

	thread-1	thread-4	thread-8	thread-12	thread-16	thread-20
iteration	1	4.13	7.53	11.92	15.76	19.01
time	1	3.41	5.99	8.28	9.76	11.6

TABLE II
Iteration and running time speedup of ASVRG-atom on CIFAR10.

	thread-1	thread-4	thread-8	thread-12	thread-16	thread-20
iteration	1	4.11	7.73	12.06	16.01	19.44
time	1	3.49	5.78	9.10	10.98	11.75

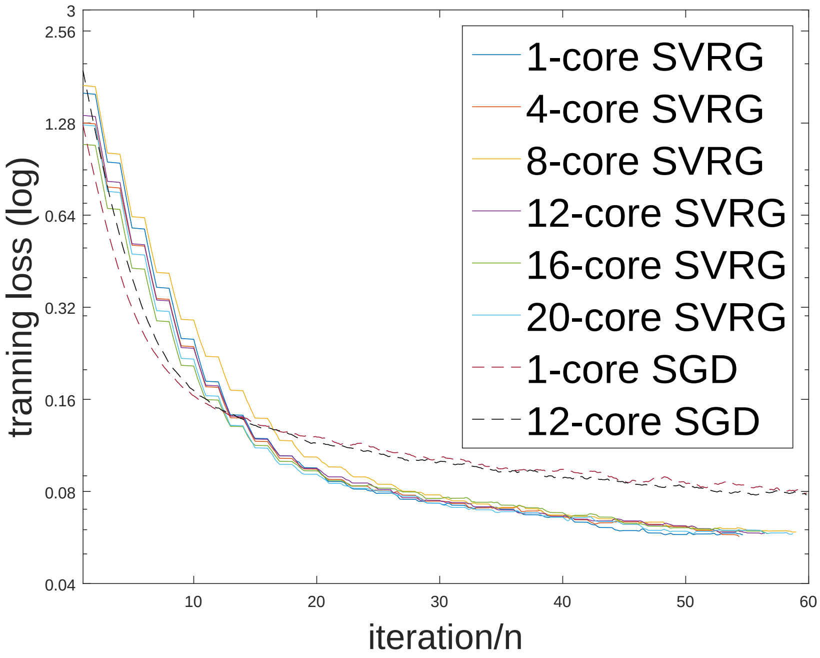

Fig. 1. The curves of loss against iteration on MNIST in the speedup experiment of ASVRG-atom. The horization axis is the number of effective pass through the data, which has included the cost of calculating full gradients for SVRG.

From the results, we conclude

1) The linear speedup is achievable in ASVRG-atom through iteration speedup. What influences the speed most is still the hardware.

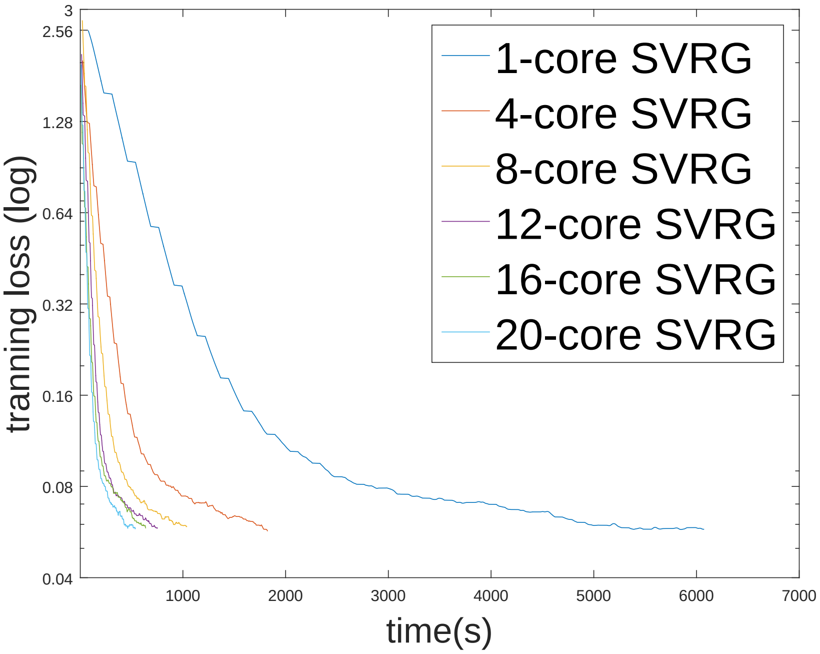

Fig. 2. The curves of loss against time on MNIST in the speedup experiment of ASVRG-atom.

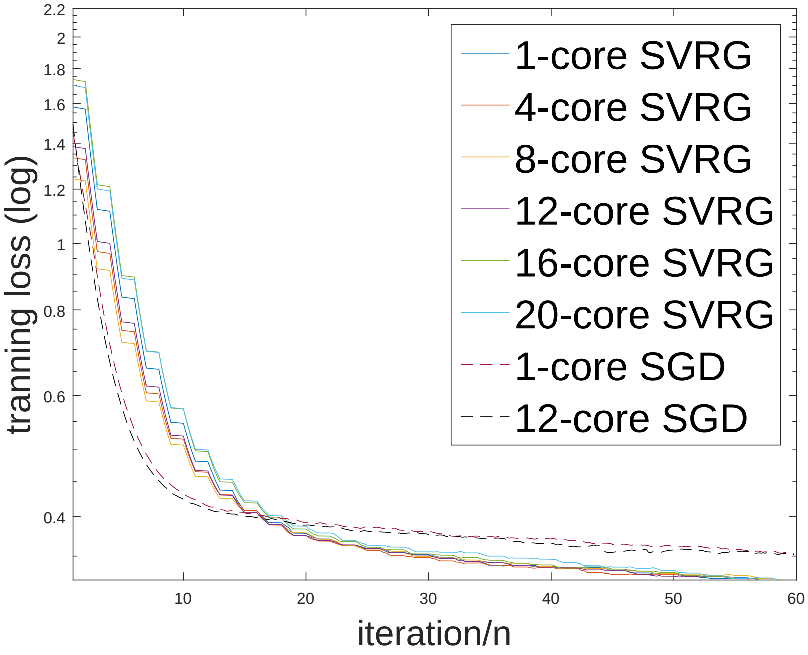

Fig. 3. The curves of loss against iteration on CIFAR10 in the speedup experiment of ASVRG-atom. The horization axis is the number of effective pass through the data, which has included the cost of calculating full gradients for SVRG.

2) ASVRG-atom is slower than ASVRG-wild, which meets our common sense.
3) ASVRG-atom also has an obvious speedup when compared with serial SVRG, e.g., there are 11 times speedup when there are 20 cores.

III. Experimental Results on a Deeper Neural Networks

In this section, we test ASVRG-wild on a deeper neural network. We train a neural network with 7 layers ($784 \times 100 \times 100 \times 100 \times 100 \times 100 \times 10$). We choose the step size as 0.1 for all the algorithms. The curves of objective loss against iterations and running time on MNIST are drawn in Figures 5 and 6. The speedup is reported in Tables III.

This experiment empirically demonstrates that the critical point obtained from ASVRG is not worse than serial SVRG when the optimization function is highly non-convex.

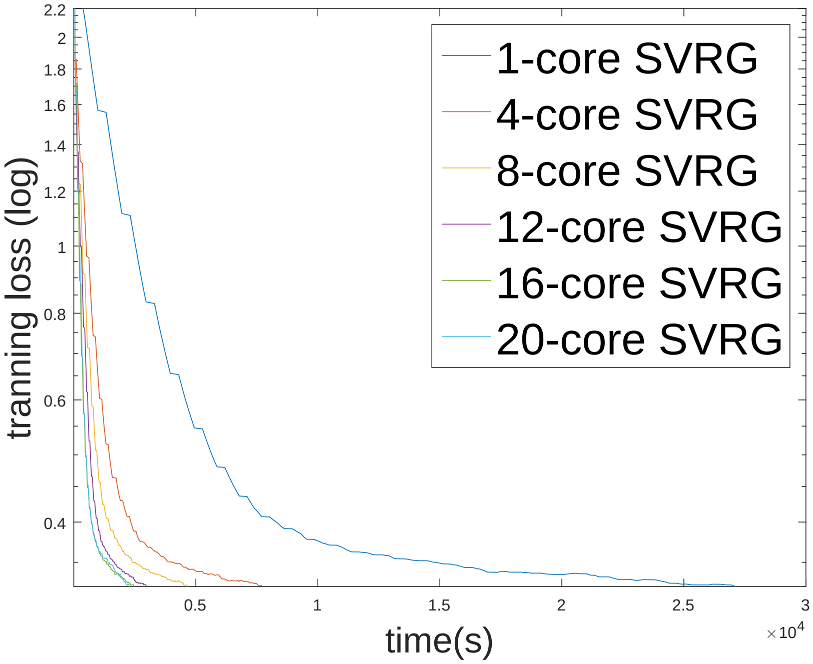

Fig. 4. The curves of loss against time on CIFAR10 in the speedup experiment of ASVRG-atom.

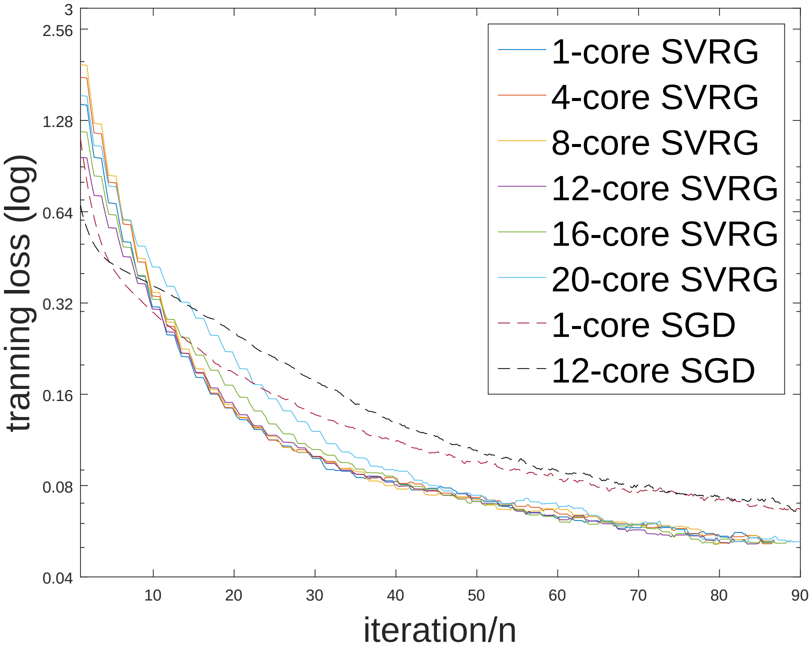

Fig. 5. The curves of loss against iteration on MNIST of ASVRG-wild training a deep neural network. The horization axis is the number of effective pass through the data, which has included the cost of calculating full gradients for SVRG.

TABLE III
ITERATION AND RUNNING TIME SPEEDUP OF ASVRG-WILD ON MNIST.

	thread-1	thread-4	thread-8	thread-12	thread-16	thread-20
iteration	1	4.01	8.10	12.13	15.63	19.07
time	1	3.25	5.52	8.17	11.03	11.99

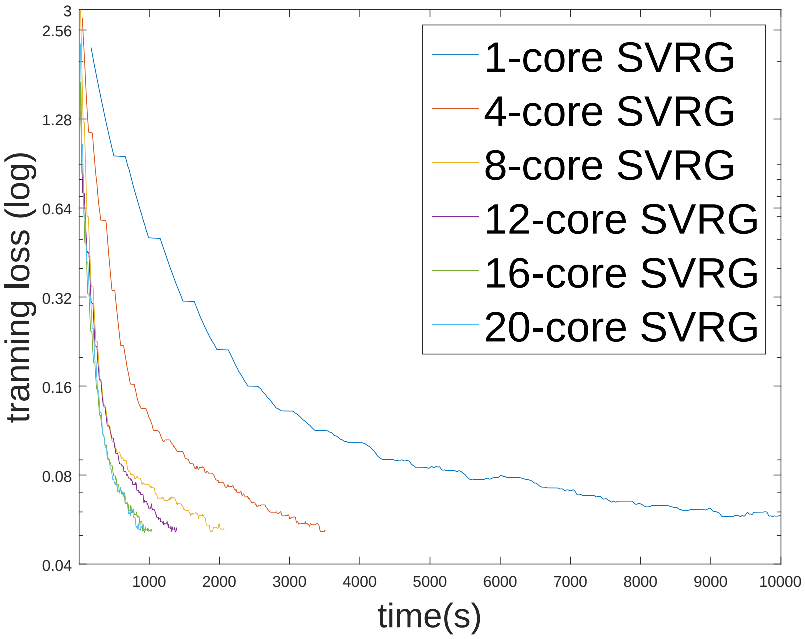

Fig. 6. The curves of loss against time on MNIST of ASVRG-wild training a deep neural network.

References

[1] Z. Allen-Zhu and E. Hazan, "Variance reduction for faster non-convex optimization," in *Proc. Int'l. Conf. on Machine Learning*, 2016.

[2] S. J. Reddi, A. Hefny, S. Sra, B. Póczós, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *Proc. Int'l. Conf. on Machine Learning*, 2016.

[3] J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar, "An asynchronous parallel stochastic coordinate descent algorithm," *Journal of Machine Learning Research*, vol. 16, no. 285-322, pp. 1–5, 2015.

[4] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Advances in Neural Information Processing Systems*, 2015.