Faster First-Order Methods for Stochastic Non-Convex Optimization on Riemannian Manifolds (Supplementary File)

This supplementary document contains the technical proofs of convergence results and some additional numerical results of the manuscript entitled "Faster First-Order Methods for Stochastic Non-convex Optimization on Riemannian Manifolds". It is structured as follows. The proof of the key lemma, namely Lemma 1 in Section 3.2, is presented in Appendix A. Then Appendix B.1 provides the proofs of the main results for general finite-sum non-convex problems in Section 3.2, including Theorem 1 and Corollary 1. Next, Appendix B.3 gives the proof of the results for online setting, including Theorem 2 and Corollary 2. For gradient dominated results in Section 3.3, including Theorems 3 and 4, are given in Appendix C.1. Finally, the detailed descriptions of datasets and more experimental results are provided in Appendix D.

A Proofs of Lemma 1

Before proving Lemma 1, we first present an useful lemma from [1]. Let $Q(\boldsymbol{x})$ denote arbitrary determinstic vector and $\xi_k(\boldsymbol{x}_{0:k})$ denote the unbiased estimate $Q(\boldsymbol{x}_k) - Q(\boldsymbol{x}_{k-1})$. Namely, $\mathbb{E}[\xi_k(\boldsymbol{x}_{0:k})|\boldsymbol{x}_{0:k}] = Q(\boldsymbol{x}_k) - Q(\boldsymbol{x}_{k-1})$. Then we aim to use the stochastic differential estimate to approximate $Q(\boldsymbol{x}_k)$ as follows:

$$\tilde{Q}(\boldsymbol{x}_k) = \tilde{Q}(\boldsymbol{x}_0) + \sum_{i=1}^{k} \xi_i(\boldsymbol{x}_{0:i}),$$

where $\tilde{Q}(\boldsymbol{x}_0)$ is the estimation of $Q(\boldsymbol{x}_0)$.

Lemma 2. *[1] For any vector h, we have*

$$\mathbb{E}\|\tilde{Q}(\boldsymbol{x}_k) - Q(\boldsymbol{x}_k)\|^2 \leq \mathbb{E}\|\tilde{Q}(\boldsymbol{x}_0) - Q(\boldsymbol{x}_0)\|^2 + \sum_{i=1}^{k} \mathbb{E}\|\xi_i(\boldsymbol{x}_{0:i}) - (Q(\boldsymbol{x}_i) - Q(\boldsymbol{x}_{i-1}))\|^2.$$

Let $\mathcal{A}_i$ map any vector $\boldsymbol{x}$ to a random vector esimate $\mathcal{A}_i(\boldsymbol{x})$ such that

$$\mathbb{E}[\mathcal{A}_i(\boldsymbol{x}_k) - \mathcal{A}_i(\boldsymbol{x}_{k-1})|\boldsymbol{x}_{0:k}] = \mathcal{V}_k - \mathcal{V}_{k-1}, \tag{3}$$

where $\mathcal{V}_k$ is defined below. Assume $\mathcal{A}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \mathcal{A}_i$ where $\mathcal{S}$ denote the sampled data of sample number $|\mathcal{S}|$. Besides, $\mathcal{A}_i$ satisfies

$$\mathbb{E}_i\|\mathcal{A}_i(\boldsymbol{x}) - \mathcal{A}_i(\boldsymbol{y})\|_2^2 \leq L^2\|\mathrm{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\|^2.$$

Then we define $\mathcal{V}_k = \mathcal{A}_{\mathcal{S}}(\boldsymbol{x}_k) - \mathcal{A}_{\mathcal{S}}(\boldsymbol{x}_{k-1}) + \mathcal{V}_{k-1}$ and $\mathcal{V}_0$ is the estimate of $\mathcal{A}(\boldsymbol{x}_0)$. Based on Lemma 2, we can further conclude:

Lemma 3. *Assume $d(\boldsymbol{x}_{k-1}, \boldsymbol{x}_k) = \|Exp_{\boldsymbol{x}_k}^{-1}(\boldsymbol{x}_{k-1})\| = \rho_{k-1}$. Then we have*

$$\mathbb{E}\|\mathcal{V}_k - \mathcal{A}(\boldsymbol{x}_k)\|^2 \leq \mathbb{E}\|\mathcal{V}_0 - \mathcal{A}(\boldsymbol{x}_0)\|^2 + L^2\sum_{i=1}^{t} \mathbb{I}_{\{|\mathcal{S}_i|<n\}}\frac{\rho_{i-1}^2}{|\mathcal{S}_i|}. \tag{4}$$

Proof. The proof here mimics that of Lemma 4 in [1]. For completeness, we provide the proof. Assume for the

k-th sampling, the seleced sample set is denoted by $\mathcal{S}_k$. Then, we have

$$\begin{aligned}
\mathbb{E}\|\mathcal{V}_k - \mathcal{A}(\boldsymbol{x}_k)\|^2 =& \mathbb{E}\|\mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_k) - \mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_{k-1}) + \mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_k)\|^2 \\
=& \mathbb{E}\|\mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_k) - \mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_{k-1}) - \mathcal{A}(\boldsymbol{x}_k) + \mathcal{A}(\boldsymbol{x}_{k-1}) + \mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 \\
=& \mathbb{E}\|\mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_k) - \mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_{k-1}) - \mathcal{A}(\boldsymbol{x}_k) + \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 + \mathbb{E}\|\mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 \\
&+ \mathbb{E}\langle \mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_k) - \mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_{k-1}) - \mathcal{A}(\boldsymbol{x}_k) + \mathcal{A}(\boldsymbol{x}_{k-1}), \mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\rangle \\
\overset{①}{=}& \mathbb{E}\|\mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_k) - \mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_{k-1}) - \mathcal{A}(\boldsymbol{x}_k) + \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 + \mathbb{E}\|\mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 \\
=& \frac{1}{|\mathcal{S}_k|}\mathbb{E}\|\mathcal{A}_i(\boldsymbol{x}_k) - \mathcal{A}_i(\boldsymbol{x}_{k-1}) - \mathcal{A}(\boldsymbol{x}_k) + \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 + \mathbb{E}\|\mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 \\
\overset{②}{\leq}& \frac{1}{|\mathcal{S}_k|}\mathbb{E}\|\mathcal{A}_i(\boldsymbol{x}_k) - \mathcal{A}_i(\boldsymbol{x}_{k-1})\|^2 + \mathbb{E}\|\mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 \\
\overset{③}{\leq}& \frac{L^2}{|\mathcal{S}_k|}\|\mathrm{Exp}_{\boldsymbol{x}_k}^{-1}(\boldsymbol{x}_{k-1})\|^2 + \mathbb{E}\|\mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 \\
\leq& \frac{L^2\rho_{k-1}^2}{|\mathcal{S}_k|} + \mathbb{E}\|\mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\|^2,
\end{aligned}$$

where ① holds since $\mathbb{E}\langle \mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_k) - \mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_{k-1}) - \mathcal{A}(\boldsymbol{x}_k) + \mathcal{A}(\boldsymbol{x}_{k-1}), \mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\rangle = \mathbf{0}$ in which the expectation is taken on the random set $\mathcal{S}_k$ ($\mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})$ is constant); ② holds due to $\mathbb{E}\|\boldsymbol{x} - \mathbb{E}(\boldsymbol{x})\|^2 \leq \mathbb{E}\|\boldsymbol{x}\|^2$; ③ holds since $f_i(\boldsymbol{x})$ is L-gradient Lipschitz, namely $\mathbb{E}_i\|\nabla f_i(\boldsymbol{x}) - \Gamma_{\boldsymbol{y}}^{\boldsymbol{x}}(\nabla f_i(\boldsymbol{y}))\|_2^2 \leq L\|\mathrm{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\|^2$. Notice, when $|\mathcal{S}_k| \geq n$, in ①, we have $\mathbb{E}\|\mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_k) - \mathcal{A}_{\mathcal{S}_k}(\boldsymbol{x}_{k-1}) - \mathcal{A}(\boldsymbol{x}_k) + \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 + \mathbb{E}\|\mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\|^2 = 0$. In this case, we can obtain $\mathbb{E}\|\mathcal{V}_k - \mathcal{A}(\boldsymbol{x}_k)\|^2 = \mathbb{E}\|\mathcal{V}_{k-1} - \mathcal{A}(\boldsymbol{x}_{k-1})\|^2$. Therefore, consider these two cases and sum up $k = 0, 1, \cdots, t$, we have

$$\mathbb{E}\|\mathcal{V}(\boldsymbol{x}_t) - \mathcal{A}(\boldsymbol{x}_t)\|^2 \leq \mathbb{E}\|\mathcal{V}_0 - \mathcal{A}(\boldsymbol{x}_0)\|^2 + L^2\sum_{i=1}^{t}\mathbb{I}_{\{|\mathcal{S}_i|<n\}}\frac{\rho_{i-1}^2}{|\mathcal{S}_i|}.$$

The proof is completed. □

Lemma 4. *Suppose Assumptions 1 and 2 hold. Let $k_0 = \lfloor k/p \rfloor$ and $\widetilde{k}_0 = k_0 p$. Assume that for $k = \lfloor k/p \rfloor \cdot p$, we select $|\mathcal{S}_1|$ samples to estimate $\boldsymbol{v}_k$ and for $k \neq \lfloor k/p \rfloor \cdot p$, we select $|\mathcal{S}_{2,k}|$ samples to estimate $\boldsymbol{v}_k$. Then the estimation error between the full Riemannian gradient $\nabla f(\boldsymbol{x}_k)$ and its estimate $\boldsymbol{v}_k$ in Algorithm 1 is bounded as*

$$\mathbb{E}\left[\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 \mid \boldsymbol{x}_{\widetilde{k}_0}, \cdots, \boldsymbol{x}_{\widetilde{k}_0+p-1}\right] \leq \mathbb{I}_{\{|\mathcal{S}_1|<n\}}\frac{\sigma^2}{|\mathcal{S}_1|} + L^2\sum_{i=\widetilde{k}_0}^{\widetilde{k}_0+p-1}\mathbb{I}_{\{|\mathcal{S}_{2,i+1}|<n\}}\frac{d^2(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})}{|\mathcal{S}_{2,i+1}|},$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})$ is the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$.

Proof. Here we construct an auxiliary sequence

$$\begin{aligned}
\widehat{\boldsymbol{v}}_t =& \sum_{i=1}^{t}\left(\mathrm{P}_{\boldsymbol{x}_i}^{\widehat{\boldsymbol{x}}}(\nabla f_{\mathcal{S}_2}(\boldsymbol{x}_i)) - \mathrm{P}_{\boldsymbol{x}_{i-1}}^{\widehat{\boldsymbol{x}}}(\nabla f_{\mathcal{S}_1}(\boldsymbol{x}_{i-1}))\right) + \mathrm{P}_{\boldsymbol{x}_0}^{\widehat{\boldsymbol{x}}}(\nabla f_{\mathcal{S}_1}(\boldsymbol{x}_0)) \\
=& \mathrm{P}_{\boldsymbol{x}_t}^{\widehat{\boldsymbol{x}}}(\nabla f_{\mathcal{S}_2}(\boldsymbol{x}_t)) - \mathrm{P}_{\boldsymbol{x}_{t-1}}^{\widehat{\boldsymbol{x}}}(\nabla f_{\mathcal{S}_2}(\boldsymbol{x}_{t-1})) + \widehat{\boldsymbol{v}}_{t-1},
\end{aligned}$$

where $\widehat{\boldsymbol{x}}$ is a given point and does not depend on the sequence $\{\boldsymbol{x}_k\}$ and the algorithm, and $\widehat{\boldsymbol{v}}_0 = \mathrm{P}_{\boldsymbol{x}_0}^{\widehat{\boldsymbol{x}}}(\nabla f_{\mathcal{S}_1}(\boldsymbol{x}_0))$. In this way, let $\mathcal{A}_{\mathcal{S}}(\boldsymbol{x}_t) = \mathrm{P}_{\boldsymbol{x}_t}^{\widehat{\boldsymbol{x}}}\left(\nabla f_{\mathcal{S}_2^t}(\boldsymbol{x}_t)\right)$. Then we have $\widehat{\boldsymbol{v}}_t = \mathcal{A}_{\mathcal{S}}(\boldsymbol{x}_t) - \mathcal{A}_{\mathcal{S}}(\boldsymbol{x}_{t-1}) + \widehat{\boldsymbol{v}}_{t-1}$. Accordingly, we can obtain

$$\begin{aligned}
\mathbb{E}_i\|\mathcal{A}_i(\boldsymbol{x}_t) - \mathcal{A}_i(\boldsymbol{x}_{t-1})\|_2^2 =& \mathbb{E}_i\left\|\mathrm{P}_{\boldsymbol{x}_t}^{\widehat{\boldsymbol{x}}}(\nabla f_i(\boldsymbol{x}_t)) - \mathrm{P}_{\boldsymbol{x}_{t-1}}^{\widehat{\boldsymbol{x}}}(\nabla f_i(\boldsymbol{x}_{t-1}))\right\|^2 \\
=& \mathbb{E}_i\left\|\mathrm{P}_{\boldsymbol{x}_{t-1}}^{\widehat{\boldsymbol{x}}}\left(\mathrm{P}_{\boldsymbol{x}_t}^{\boldsymbol{x}_{t-1}}(\nabla f_i(\boldsymbol{x}_t))\right) - \mathrm{P}_{\boldsymbol{x}_{t-1}}^{\widehat{\boldsymbol{x}}}(\nabla f_i(\boldsymbol{x}_{t-1}))\right\|^2 \\
=& \mathbb{E}_i\left\|\mathrm{P}_{\boldsymbol{x}_{t-1}}^{\widehat{\boldsymbol{x}}}\left(\mathrm{P}_{\boldsymbol{x}_t}^{\boldsymbol{x}_{t-1}}(\nabla f_i(\boldsymbol{x}_t)) - \nabla f_i(\boldsymbol{x}_{t-1})\right)\right\|^2 \\
\overset{①}{=}& \mathbb{E}_i\left\|\mathrm{P}_{\boldsymbol{x}_t}^{\boldsymbol{x}_{t-1}}(\nabla f_i(\boldsymbol{x}_t)) - \nabla f_i(\boldsymbol{x}_{t-1})\right\|^2 \\
\leq& L^2\|\mathrm{Exp}_{\boldsymbol{x}_{t-1}}^{-1}(\boldsymbol{x}_t)\|^2,
\end{aligned}$$

where ① holds as the parallel transport $\mathrm{P}_{\boldsymbol{x}}^{\boldsymbol{y}}$ preserves the norm. On the other hand, all $\mathcal{A}_i(\boldsymbol{x}_t)$ $(t=0,\cdots,k)$ are located in the tangent space at the point $\boldsymbol{x}_k$. Thus, Lemma 3 is applicable to the sequence $\widehat{\boldsymbol{v}}_t$.

Let $k_0=\lfloor K/p \rfloor$. For simplicity, we use $\mathcal{V}_0,\mathcal{V}_1,\cdots,\mathcal{V}_k$ to respectively denote $\mathcal{V}_{k_0},\mathcal{V}_{k_0+1},\cdots,\mathcal{V}_{k_0+k}$. For $\mathcal{V}_0$, we have $\mathcal{V}_0=\mathrm{P}_{\boldsymbol{x}_{k_0}}^{\widehat{\boldsymbol{x}}}\left(\nabla f_{\mathcal{S}_1}(\boldsymbol{x}_{k_0})\right)$. Then it yields

$$\begin{aligned}\mathbb{E}\|\mathcal{V}_0-\mathcal{A}(\boldsymbol{x}_0)\|^2=&\mathbb{E}\|\mathrm{P}_{\boldsymbol{x}_{k_0}}^{\widehat{\boldsymbol{x}}}\left(\nabla f_{\mathcal{S}_1}(\boldsymbol{x}_{k_0})\right)-\mathrm{P}_{\boldsymbol{x}_{k_0}}^{\widehat{\boldsymbol{x}}}\left(\nabla f(\boldsymbol{x}_{k_0})\right)\|^2\\ =&\mathbb{E}\|\nabla f_{\mathcal{S}_1}(\boldsymbol{x}_{k_0})-\nabla f(\boldsymbol{x}_{k_0})\|^2\\ =&\frac{1}{|\mathcal{S}_1|}\mathbb{E}\|\nabla f_i(\boldsymbol{x}_{k_0})-\nabla f(\boldsymbol{x}_{k_0})\|^2\\ \overset{①}{\leq}&\frac{\sigma^2}{|\mathcal{S}_1|},\end{aligned}$$

where ① holds since the gradient variance is bounded in Assumption 2. On the other hand, since $\boldsymbol{x}_{k+1}=\mathrm{Exp}_{\boldsymbol{x}_k}\left(-\eta_k\frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}\right)$, we have

$$\mathrm{d}^2\left(\boldsymbol{x}_{k+1},\boldsymbol{x}_k\right)=\left\|\mathrm{Exp}_{\boldsymbol{x}_k}^{-1}\left(\boldsymbol{x}_{k+1}\right)\right\|.$$

Therefore, we have

$$\mathbb{E}\|\widehat{\boldsymbol{v}}_t-\mathrm{P}_{\boldsymbol{x}_t}^{\widehat{\boldsymbol{x}}}\left(\nabla f(\boldsymbol{x}_t)\right)\|^2\leq L^2\sum_{i=0}^{t-1}\mathbb{I}_{\{|\mathcal{S}_{2,i+1}|<n\}}\frac{\mathrm{d}^2\left(\boldsymbol{x}_i,\boldsymbol{x}_{i+1}\right)}{|\mathcal{S}_{2,i+1}|}+\frac{\sigma^2}{|\mathcal{S}_1|}.$$

Since the parallel transport preserve the norm, we can further establish

$$\mathbb{E}\|\mathrm{P}_{\widehat{\boldsymbol{x}}}^{\boldsymbol{x}_k}\left(\widehat{\boldsymbol{v}}_t-\mathrm{P}_{\boldsymbol{x}_t}^{\widehat{\boldsymbol{x}}}\left(\nabla f(\boldsymbol{x}_t)\right)\right)\|^2=\mathbb{E}\|\widehat{\boldsymbol{v}}_t-\mathrm{P}_{\boldsymbol{x}_t}^{\widehat{\boldsymbol{x}}}\left(\nabla f(\boldsymbol{x}_t)\right)\|^2\leq L^2\sum_{i=0}^{t-1}\mathbb{I}_{\{|\mathcal{S}_{2,i+1}|<n\}}\frac{\mathrm{d}^2\left(\boldsymbol{x}_i,\boldsymbol{x}_{i+1}\right)}{|\mathcal{S}_{2,i+1}|}+\frac{\sigma^2}{|\mathcal{S}_1|}.$$

By setting $t=k$ and noting $t\leq p$ for each epoch, we establish

$$\mathbb{E}\|\boldsymbol{v}_k-\nabla f(\boldsymbol{x}_k)\|^2=\mathbb{E}\|\mathrm{P}_{\widehat{\boldsymbol{x}}}^{\boldsymbol{x}_k}\left(\widehat{\boldsymbol{v}}_k-\mathrm{P}_{\boldsymbol{x}_k}^{\widehat{\boldsymbol{x}}}\left(\nabla f(\boldsymbol{x}_k)\right)\right)\|^2\leq\frac{\sigma^2}{|\mathcal{S}_1|}+L^2\sum_{i=k_0}^{k_0+p-1}\mathbb{I}_{\{|\mathcal{S}_{2,i+1}|<n\}}\frac{\mathrm{d}^2\left(\boldsymbol{x}_i,\boldsymbol{x}_{i+1}\right)}{|\mathcal{S}_{2,i+1}|}.$$

Notice, when we sample all n samples, we have $\mathbb{E}\|\mathcal{V}_0-\mathcal{A}(x_0)\|^2=0$ and thus

$$\mathbb{E}\|\boldsymbol{v}_k-\nabla f(\boldsymbol{x}_k)\|^2\leq L^2\sum_{i=k_0}^{k_0+p-1}\mathbb{I}_{\{|\mathcal{S}_{2,i+1}|<n\}}\frac{\mathrm{d}^2\left(\boldsymbol{x}_i,\boldsymbol{x}_{i+1}\right)}{|\mathcal{S}_{2,i+1}|}.$$

So by combining the two case together, we can obtain the result in Lemma 1. The proof is completed. □

Now we are ready to prove Lemma 1.

Proof of Lemma 1. To prove Lemma 1, we can directly set $|\mathcal{S}_{2,k}|$ in Lemma 4 as $|\mathcal{S}_2|$ in Lemma 1 and obtain the results in Lemma 1. The proof is completed. □

B Proof of the Results in Section 3.2

B.1 Proof of Theorem 1

Proof. For brevity, let $\widetilde{\eta}_k = \frac{\eta_k}{\|\boldsymbol{v}_k\|}$. Then by using the L-gradient Lipschitz, we have

$$\begin{aligned}
f(\boldsymbol{x}_{k+1}) \leq & f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \operatorname{Exp}_{\boldsymbol{x}_k}^{-1}(\boldsymbol{x}_{k+1}) \rangle + \frac{L}{2} \|\operatorname{Exp}_{\boldsymbol{x}_k}^{-1}(\boldsymbol{x}_{k+1})\|^2 \\
\leq & f(\boldsymbol{x}_k) - \widetilde{\eta}_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{v}_k \rangle + \frac{\widetilde{\eta}_k^2 L}{2} \|\boldsymbol{v}_k\|^2 \\
\leq & f(\boldsymbol{x}_k) - \widetilde{\eta}_k \langle \nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k, \boldsymbol{v}_k \rangle - \widetilde{\eta}_k \left(1 - \frac{\widetilde{\eta}_k L}{2}\right) \|\boldsymbol{v}_k\|^2 \\
\leq & f(\boldsymbol{x}_k) - \widetilde{\eta}_k \langle \nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k, \boldsymbol{v}_k \rangle - \widetilde{\eta}_k \left(1 - \frac{\widetilde{\eta}_k L}{2}\right) \|\boldsymbol{v}_k\|^2 \\
\leq & f(\boldsymbol{x}_k) + \frac{\widetilde{\eta}_k}{2} \|\nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k\|^2 - \frac{\widetilde{\eta}_k}{2} (1 - \widetilde{\eta}_k L) \|\boldsymbol{v}_k\|^2.
\end{aligned} \tag{5}$$

Since we have $\boldsymbol{x}_{k+1} = \operatorname{Exp}_{\boldsymbol{x}_k}^{-1}\left(-\eta_k \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}\right)$, we can obtain

$$\mathrm{d}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) = \eta_k = \min\left(\frac{\epsilon}{2Ln_0}, \frac{\|\boldsymbol{v}_k\|}{4Ln_0}\right) \leq \frac{\epsilon}{2Ln_0}. \tag{6}$$

Now we consider the two cases: (1) k is not an integer multiple of p; (2) k is an integer multiple of p. We can consider case (1) as follows. If $s = n$, then by Lemma 1 and Eqn. (6), we have

$$\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{L^2}{|\mathcal{S}_2|} \sum_{i=k_0}^{k_0+p-1} \mathrm{d}^2(\boldsymbol{x}_i, \boldsymbol{x}_{i+1}) \leq \frac{pL^2}{|\mathcal{S}_2|} \frac{\epsilon^2}{4L^2 n_0^2} = n_0 s^{\frac{1}{2}} L^2 \frac{n_0}{4s^{\frac{1}{2}}} \frac{\epsilon^2}{4L^2 n_0^2} = \frac{1}{16}\epsilon^2.$$

If $s = \frac{16\sigma^2}{\epsilon^2}$, then Lemma 1 gives

$$\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{pL^2}{|\mathcal{S}_2|} \frac{\epsilon^2}{4L^2 n_0^2} + \frac{\sigma^2}{|\mathcal{S}_1|} = n_0 s^{\frac{1}{2}} L^2 \frac{\epsilon^2}{4L^2 n_0^2} \frac{n_0}{4s^{\frac{1}{2}}} + \frac{\epsilon^2}{16} = \frac{1}{8}\epsilon^2. \tag{7}$$

For case (2), namely when k is an integer multiple of p, we have $\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{pL^2\eta_k^2}{|\mathcal{S}_2|} + \frac{\sigma^2}{|\mathcal{S}_1|} = 0 + \frac{\epsilon^2}{16} \leq \frac{1}{8}\epsilon^2$. At the same time, since $\eta_k = \min\left(\frac{\epsilon}{2Ln_0}, \frac{\|\boldsymbol{v}_k\|}{4Ln_0}\right)$, we have $\widetilde{\eta}_k = \frac{\eta_k}{\|\boldsymbol{v}_k\|} = \min\left(\frac{\epsilon}{2Ln_0\|\boldsymbol{v}_k\|}, \frac{1}{4Ln_0}\right) \leq \frac{1}{4Ln_0}$ and

$$\widetilde{\eta}_k(1 - \widetilde{\eta}_k L)\|\boldsymbol{v}_k\|^2 \geq \frac{3\widetilde{\eta}_k}{4}\|\boldsymbol{v}_k\|^2 = \frac{3}{8}\min\left(\frac{\epsilon}{Ln_0\|\boldsymbol{v}_k\|}, \frac{1}{2Ln_0}\right)\|\boldsymbol{v}_k\|^2 = \frac{3\epsilon^2}{16Ln_0}\min\left(\frac{2\|\boldsymbol{v}_k\|}{\epsilon}, \frac{\|\boldsymbol{v}_k\|^2}{\epsilon^2}\right) \overset{①}{\geq} \frac{3\epsilon(2\|\boldsymbol{v}_k\| - \epsilon)}{16Ln_0},$$

where ① uses $x^2 \geq 2|x| - 1$ for $\forall x$. So by taking expectation, we have

$$\mathbb{E}\left[f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)\right] \leq \frac{1}{2}\frac{1}{4Ln_0}\frac{\epsilon^2}{8} - \frac{1}{2}\frac{3\epsilon(2\|\boldsymbol{v}_k\| - \epsilon)}{16Ln_0} = -\frac{\epsilon}{64Ln_0}\left(12\mathbb{E}\|\boldsymbol{v}_k\| - 7\epsilon\right).$$

In this way, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{v}_k\| \leq \frac{7\epsilon}{64} + \frac{16Ln_0}{3K\epsilon}\mathbb{E}\left[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_K)\right] \leq \frac{7\epsilon}{64} + \frac{16Ln_0\Delta}{3K\epsilon},$$

where we use $\mathbb{E}\left[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_K)\right] \leq \mathbb{E}\left[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*)\right] \leq f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*) \leq \Delta$. It means that after running at most $K = \frac{14Ln_0\Delta}{\epsilon^2}$ iterations, the algorithm will terminate, since

$$\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}})\| = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(\boldsymbol{x}_k)\| \leq \frac{1}{K}\sum_{k=0}^{K-1}\left[\mathbb{E}\|\nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k\| + \mathbb{E}\|\boldsymbol{v}_k\|\right] \overset{①}{\leq} \frac{1}{K}\sum_{k=0}^{K-1}\sqrt{\mathbb{E}\|\nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k\|^2} + \frac{\epsilon}{2} \overset{②}{\leq} \epsilon,$$

where ① uses the Jensen's inequality; ② holds since $\mathbb{E}\|\nabla f(\boldsymbol{x}) - \boldsymbol{v}_k\|^2 \leq \frac{\epsilon^2}{8}$ in Eqn. (7). The proof is completed. The proof is completed. □

B.2 Proof of Corollary 1

Proof. According to Theorem 1, we know that after running at most $K = \frac{14Ln_0\Delta}{\epsilon^2}$ iterations, the algorithm will terminate. In this way, we can compute the stochastic gradient complexity as

$$\mathcal{O}\left(\frac{K}{p}|\mathcal{S}_1| + K|\mathcal{S}_2|\right) = \mathcal{O}\left(\frac{Ln_0\Delta}{\epsilon^2}\left(s\frac{1}{n_0 s^{1/2}} + \frac{s^{1/2}}{2n_0}\right)\right) = \mathcal{O}\left(\min\left(n + \frac{L\Delta\sqrt{n}}{\epsilon^2}, \frac{L\Delta\sigma}{\epsilon^3}\right)\right).$$

The proof is completed. □

B.3 Proof of Theorem 2

Proof. For brevity, let $\widetilde{\eta}_k = \frac{\eta_k}{\|\boldsymbol{v}_k\|}$. From Eqn. (5), we can obtain the following inequality:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \frac{\widetilde{\eta}_k}{2}\|\nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k\|^2 - \frac{\widetilde{\eta}_k}{2}\left(1 - \widetilde{\eta}_k L\right)\|\boldsymbol{v}_k\|^2. \tag{8}$$

Now we consider the two cases: (1) k is not an integer multiple of p; (2) k is an integer multiple of p. We can consider case (1) as follows. By setting $p = \frac{\sigma n_0}{\epsilon}$, $\eta_k = \min\left(\frac{\epsilon}{2Ln_0}, \frac{\|\boldsymbol{v}_k\|}{4Ln_0}\right)$, $|\mathcal{S}_1| = \frac{16\sigma^2}{\epsilon^2}$, $\mathcal{S}_2 = \frac{\sigma}{2\epsilon n_0}$, where $n_0 \in [1, 2\sigma/\epsilon]$, Lemma 1 gives

$$\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{pL^2\eta_k^2}{|\mathcal{S}_2|} + \frac{\sigma^2}{|\mathcal{S}_1|} = \frac{\sigma n_0}{\epsilon}L^2\frac{\epsilon^2}{4L^2n_0^2}\frac{\epsilon n_0}{4\sigma} + \frac{\epsilon^2}{16} = \frac{1}{8}\epsilon^2. \tag{9}$$

For case (2), namely when k is an integer multiple of p, we have $\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{pL^2\eta^2}{|\mathcal{S}_2|} + \frac{\sigma^2}{|\mathcal{S}_1|} = 0 + \frac{\epsilon^2}{16} \leq \frac{1}{8}\epsilon^2$. Then similar to proof in Sec. B.1, since $\eta_k = \min\left(\frac{\epsilon}{2Ln_0}, \frac{\|\boldsymbol{v}_k\|}{4Ln_0}\right)$, we have $\widetilde{\eta}_k = \frac{\eta_k}{\|\boldsymbol{v}_k\|} = \min\left(\frac{\epsilon}{2Ln_0\|\boldsymbol{v}_k\|}, \frac{1}{4Ln_0}\right) \leq \frac{1}{4Ln_0}$ and

$$\widetilde{\eta}_k(1 - \widetilde{\eta}_k L)\|\boldsymbol{v}_k\|^2 \geq \frac{3\widetilde{\eta}_k}{4}\|\boldsymbol{v}_k\|^2 = \frac{3}{8}\min\left(\frac{\epsilon}{Ln_0\|\boldsymbol{v}_k\|}, \frac{1}{2Ln_0}\right)\|\boldsymbol{v}_k\|^2 = \frac{3\epsilon^2}{16Ln_0}\min\left(\frac{2\|\boldsymbol{v}_k\|}{\epsilon}, \frac{\|\boldsymbol{v}_k\|^2}{\epsilon^2}\right) \overset{①}{\geq} \frac{3\epsilon(2\|\boldsymbol{v}_k\| - \epsilon)}{16Ln_0},$$

where ① uses $x^2 \geq 2|x| - 1$ for $\forall x$. So by taking expectation, we have

$$\mathbb{E}\left[f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)\right] \leq \frac{1}{2}\frac{1}{4Ln_0}\frac{\epsilon^2}{8} - \frac{1}{2}\frac{3\epsilon(2\|\boldsymbol{v}_k\| - \epsilon)}{16Ln_0} = -\frac{\epsilon}{64Ln_0}\left(12\mathbb{E}\|\boldsymbol{v}_k\| - 7\epsilon\right).$$

In this way, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{v}_k\| \leq \frac{7\epsilon}{64} + \frac{16Ln_0}{3K\epsilon}\mathbb{E}\left[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_K)\right] \leq \frac{7\epsilon}{64} + \frac{16Ln_0\Delta}{3K\epsilon},$$

where we use $\mathbb{E}\left[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_K)\right] \leq \mathbb{E}\left[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*)\right] \leq f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*) \leq \Delta$. It means that after running at most $K = \frac{14Ln_0\Delta}{\epsilon^2}$ iterations, the algorithm will terminate, since

$$\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}})\| = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(\boldsymbol{x}_k)\| \leq \frac{1}{K}\sum_{k=0}^{K-1}\left[\mathbb{E}\|\nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k\| + \mathbb{E}\|\boldsymbol{v}_k\|\right] \overset{①}{\leq} \frac{1}{K}\sum_{k=0}^{K-1}\sqrt{\mathbb{E}\|\nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k\|^2} + \frac{\epsilon}{2} \overset{②}{\leq} \epsilon,$$

where ① uses the Jensen's inequality; ② holds since $\mathbb{E}\|\nabla f(\boldsymbol{x}) - \boldsymbol{v}_k\|^2 \leq \frac{\epsilon^2}{8}$ in Eqn. (7). The proof is completed. The proof is completed. □

B.4 Proof of Corollary 2

Proof. We adopt similar proof sketch of Corollary 1. According to Theorem 2, we know that after running at most $K = \frac{14Ln_0\Delta}{\epsilon^2}$ iterations, the algorithm will terminate. In this way, we can compute the stochastic gradient complexity as

$$\mathcal{O}\left(\frac{K}{p}|\mathcal{S}_1| + K|\mathcal{S}_2|\right) = \mathcal{O}\left(\frac{Ln_0\Delta}{\epsilon^2}\left(\frac{\sigma^2}{\epsilon^2}\frac{\epsilon}{\sigma n_0} + \frac{\sigma}{\epsilon n_0}\right)\right) = \mathcal{O}\left(\frac{L\sigma\Delta}{\epsilon^3}\right).$$

The proof is completed. □

C Proofs of the Results in Section 3.3

Before proving Theorems 3 and 4, we first prove Lemma 5 which is a key lemma to prove Theorems 3 and 4.

Lemma 5. *Assume function $f(\boldsymbol{x})$ is τ-gradient dominated. Let $\mathcal{E}$ denotes the event:*

$$\mathcal{E} = \left\{\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}})\|^2 \leq \epsilon^2 \quad \text{and} \quad \mathbb{E}\left[f(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)\right] \leq \tau\epsilon^2.\right\}$$

(1) For online-setting, we have $p = \frac{\sigma n_0}{\epsilon}$, $\eta_k = \frac{\|\boldsymbol{v}_k\|}{2Ln_0}$, $|\mathcal{S}_1| = \frac{32\sigma^2}{\epsilon^2}$, $|\mathcal{S}_{2,k}| = \frac{8\sigma\|\boldsymbol{v}_{k-1}\|^2}{\epsilon^3 n_0}$. To let the event $\mathcal{E}$ happen, Algorithm 1 runs at most $K = \frac{64Ln_0\Delta}{\epsilon^2}$ iterations and the IFO complexity is

$$\mathcal{O}\left(\frac{L\Delta\sigma}{\epsilon^3}\right), \quad \text{where } \widetilde{\Delta} = f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*).$$

(2) For finite-sum setting, we let $s = \min\left(n, \frac{32\sigma^2}{\epsilon^2}\right)$, $p = n_0 s^{\frac{1}{2}}$, $\eta_k = \frac{\|\boldsymbol{v}_k\|}{2Ln_0}$, $|\mathcal{S}_1| = s$, $|\mathcal{S}_{2,k}| = \min\left(\frac{8p\|\boldsymbol{v}_{k-1}\|^2}{n_0^2\epsilon^2}, n\right)$. To let the event $\mathcal{E}$ happen, Algorithm 1 runs at most $K = \frac{64Ln_0\Delta}{\epsilon^2}$ iterations and the IFO complexity is

$$\mathcal{O}\left(\min\left(n + \frac{L\Delta\sqrt{n}}{\epsilon^2}, \frac{L\Delta\sigma}{\epsilon^3}\right)\right), \quad \text{where } \Delta = f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*).$$

Proof. For brevity, let $\widetilde{\eta}_k = \frac{\eta_k}{\|\boldsymbol{v}_k\|} = \frac{1}{2Ln_0}$. Then similar to Eqn. (5), by using the L-gradient Lipschitz, we have

$$\begin{aligned}
f(\boldsymbol{x}_{k+1}) \leq & f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \mathrm{Exp}_{\boldsymbol{x}_k}^{-1}(\boldsymbol{x}_{k+1})\rangle + \frac{L}{2}\|\mathrm{Exp}_{\boldsymbol{x}_k}^{-1}(\boldsymbol{x}_{k+1})\|^2 \\
\leq & f(\boldsymbol{x}_k) - \widetilde{\eta}_k\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{v}_k\rangle + \frac{\widetilde{\eta}_k^2 L}{2}\|\boldsymbol{v}_k\|^2 \\
\leq & f(\boldsymbol{x}_k) - \widetilde{\eta}_k\langle \nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k, \boldsymbol{v}_k\rangle - \widetilde{\eta}_k\left(1 - \frac{\widetilde{\eta}_k L}{2}\right)\|\boldsymbol{v}_k\|^2 \\
\leq & f(\boldsymbol{x}_k) - \widetilde{\eta}_k\langle \nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k, \boldsymbol{v}_k\rangle - \widetilde{\eta}_k\left(1 - \frac{\widetilde{\eta}_k L}{2}\right)\|\boldsymbol{v}_k\|^2 \\
\leq & f(\boldsymbol{x}_k) + \frac{\widetilde{\eta}_k}{2}\|\nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k\|^2 - \frac{\widetilde{\eta}_k}{2}\left(1 - \widetilde{\eta}_k L\right)\|\boldsymbol{v}_k\|^2 \\
\overset{①}{\leq} & f(\boldsymbol{x}_k) + \frac{1}{4Ln_0}\|\nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k\|^2 - \frac{1}{8Ln_0}\|\boldsymbol{v}_k\|^2,
\end{aligned}$$

where ① holds since $n_0 \geq 1$. By summing up this equation from 0 to $K-1$ and taking expectation, we can obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{v}_k\|^2 \leq \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 + \frac{8Ln_0}{K}\left[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_K)\right] \overset{①}{\leq} \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 + \frac{8Ln_0\Delta}{K},$$

where ① uses $\mathbb{E}\left[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_K)\right] \leq \mathbb{E}\left[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*)\right] \leq f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*) \leq \Delta$.

Now we use Lemma 4 to bound each $\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2$ for both online and finite-sum setting. For online-setting, we have $p = \frac{\sigma n_0}{\epsilon}$, $\eta_k = \frac{\|\boldsymbol{v}_k\|}{2Ln_0}$, $|\mathcal{S}_1| = \frac{32\sigma^2}{\epsilon^2}$, $|\mathcal{S}_{2,k}| = \frac{8\sigma\|\boldsymbol{v}_{k-1}\|^2}{\epsilon^3 n_0}$. From Lemma 4, we can establish

$$\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 \leq \mathbb{I}_{\{|\mathcal{S}_1|<n\}}\frac{\sigma^2}{|\mathcal{S}_1|} + L^2\sum_{i=k_0}^{k_0+p-1}\mathbb{I}_{\{|\mathcal{S}_{2,i+1}|<n\}}\frac{\mathrm{d}^2(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})}{|\mathcal{S}_{2,i+1}|} \leq \sigma^2\frac{\epsilon^2}{32\sigma^2} + L^2\sum_{i=k_0}^{k_0+p-1}\frac{\|\boldsymbol{v}_i\|^2}{4L^2n_0^2}\frac{\epsilon^3 n_0}{8\sigma\|\boldsymbol{v}_i\|^2} \leq \frac{\epsilon^2}{16},$$

where we use $\mathrm{d}^2(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) = \left\|\mathrm{Exp}_{\boldsymbol{x}_k}^{-1}(\boldsymbol{x}_{k+1})\right\|^2 = \eta_k^2$ since $\boldsymbol{x}_{k+1} = \mathrm{Exp}_{\boldsymbol{x}_k}\left(-\eta_k\frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}\right)$. For finite-sum setting, we let $s = \min\left(n, \frac{32\sigma^2}{\epsilon^2}\right)$, $p = n_0 s^{\frac{1}{2}}$, $\eta_k = \frac{\|\boldsymbol{v}_k\|}{2Ln_0}$, $|\mathcal{S}_1| = s$, $|\mathcal{S}_{2,k}| = \min\left(\frac{8p\|\boldsymbol{v}_{k-1}\|^2}{n_0^2\epsilon^2}, n\right)$. In this case, we also have

$$\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 \leq \mathbb{I}_{\{|\mathcal{S}_1|<n\}}\frac{\sigma^2}{|\mathcal{S}_1|} + L^2\sum_{i=k_0}^{k_0+p-1}\mathbb{I}_{\{|\mathcal{S}_{2,i+1}|<n\}}\frac{\mathrm{d}^2(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})}{|\mathcal{S}_{2,i+1}|} \leq \sigma^2\frac{\epsilon^2}{32\sigma^2} + L^2\sum_{i=k_0}^{k_0+p-1}\frac{\|\boldsymbol{v}_i\|^2}{4L^2n_0^2}\frac{n_0^2\epsilon^2}{8p\|\boldsymbol{v}_i\|^2} \leq \frac{\epsilon^2}{16}.$$

Meanwhile, we set $K = \frac{64Ln_0\Delta}{\epsilon^2}$, which gives

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{v}_k\|^2 \leq \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{v}_k - \nabla f(\boldsymbol{x}_k)\|^2 + \frac{8Ln_0\Delta}{K} \leq \frac{\epsilon^2}{4}.$$

It means that after running at most $K = \frac{14Ln_0\Delta}{\epsilon^2}$ iterations, the algorithm will terminate, since

$$\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}})\|^2 = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{1}{K}\sum_{k=0}^{K-1}\left[2\mathbb{E}\|\nabla f(\boldsymbol{x}_k) - \boldsymbol{v}_k\|^2 + 2\mathbb{E}\|\boldsymbol{v}_k\|^2\right] \leq \epsilon^2.$$

Then we use the definition of τ-gradient dominated function, we have

$$\mathbb{E}[f(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)] \leq \frac{\tau}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 = \tau\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}})\|^2 \leq \tau\epsilon^2.$$

Now consider the IFO complexity for both online and finite-sum settings. For online setting, its IFO complexity is

$$\mathcal{O}\left(\frac{K}{p}|\mathcal{S}_1| + \sum_{k=0}^{K-1}\mathbb{E}|\mathcal{S}_{2,k}|\right) = \mathcal{O}\left(\frac{L\Delta\sigma}{\epsilon^3} + \frac{\sigma}{n_0\epsilon^3}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{v}_k\|^2\right) \leq \mathcal{O}\left(\frac{L\Delta\sigma}{\epsilon^3} + \frac{\sigma}{n_0\epsilon^3}K\cdot\frac{\epsilon^2}{4}\right) = \mathcal{O}\left(\frac{L\Delta\sigma}{\epsilon^3}\right).$$

similarly, we can compute the expectation IFO complexity for finite-sum setting:

$$\mathcal{O}\left(\frac{K}{p}|\mathcal{S}_1| + \sum_{k=0}^{K-1}\mathbb{E}|\mathcal{S}_{2,k}|\right) = \mathcal{O}\left(\min\left(n + \frac{L\Delta\sqrt{n}}{\epsilon^2}, \frac{L\Delta\sigma}{\epsilon^3}\right)\right).$$

The proof is completed. □

C.1 Proof of Theorems 3

Now we are ready to prove Theorem 3.

Proof. We first consider the t iteration in Algorithm 2. By Lemma 5, we obtain that by using ϵ_{t-1} with proper other parameters, the IFO complexity of Algorithm 1 for computing $\mathbb{E}[\|\nabla f(\widetilde{\boldsymbol{x}}_t)\|^2] \leq \epsilon_{t-1}^2$ is

$$\mathcal{O}\big(\min\big(n + \frac{L\Delta_t\sqrt{n}}{\epsilon_{t-1}^2}, \frac{L\Delta_t\sigma}{\epsilon_{t-1}^3}\big)\big),$$

when the parameters satisfy $s_t = \min\big(n, \frac{32\sigma^2}{\epsilon_{t-1}^2}\big)$, $p^t = n_0^t s_t^{\frac{1}{2}}$, $\eta_k^t = \frac{\|\boldsymbol{v}_k^t\|}{2Ln_0}$, $|\mathcal{S}_1^t| = s_t$, $|\mathcal{S}_{2,k}^t| = \min(\frac{8p^t\|\boldsymbol{v}_{k-1}^t\|^2}{(n_0^t)^2\epsilon_{t-1}^2}, n)$ and $K^t = \frac{64Ln_0^t\Delta^t}{\epsilon_{t-1}^2}$. Then the initial point $\boldsymbol{x}_0$ at the t iteration is the output $\widetilde{\boldsymbol{x}}_{t-1}$ of the $(t-1)$-th iteration, which gives the distance $\Delta_t = \mathbb{E}[f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*)] = \mathbb{E}[f(\widetilde{\boldsymbol{x}}_{t-1}) - f(\boldsymbol{x}_*)] \leq \tau\epsilon_{t-2}^2$ by using Lemma 5. On the other hand, $\epsilon_t = \frac{\epsilon_0}{2^t}$. So the IFO complexity of the t-th iteration is

$$\mathcal{O}\big(\min\big(n + \frac{L\Delta_t\sqrt{n}}{\epsilon_{t-1}^2}, \frac{L\Delta_t\sigma}{\epsilon_{t-1}^3}\big)\big) = \mathcal{O}\big(\min\big(n + \frac{L\tau\epsilon_{t-2}^2\sqrt{n}}{\epsilon_{t-1}^2}, \frac{L\sigma\tau\epsilon_{t-2}^2}{\epsilon_{t-1}^3}\big)\big) = \mathcal{O}\left(\min\left(n + \tau L\sqrt{n}, \frac{\tau L\sigma}{\epsilon_{t-1}}\right)\right).$$

So to achieve $\epsilon_T \leq \frac{\epsilon_0}{2^T} \leq \epsilon$, T satisfies $T \geq \log\left(\frac{\epsilon_0}{\epsilon}\right)$. So for the T iterations, the total complexity is

$$\mathcal{O}\left(\min\left(\left(n + \tau L\sqrt{n}\right)\log\left(\frac{1}{\epsilon}\right), \tau L\sigma\sum_{t=1}^{T}\frac{1}{\epsilon_{t-1}}\right)\right) = \mathcal{O}\left(\min\left(\left(n + \tau L\sqrt{n}\right)\log\left(\frac{1}{\epsilon}\right), \frac{\tau L\sigma}{\epsilon}\right)\right).$$

Meanwhile, we can obtain

$$\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}}_t)\| \leq \sqrt{\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}}_t)\|^2} \leq \epsilon_{t-1} = \frac{\epsilon_0}{2^{t-1}} = \frac{1}{2^t}\sqrt{\frac{\Delta}{\tau}} \quad \text{and} \quad \mathbb{E}\left[f(\widetilde{\boldsymbol{x}}_t) - f(\boldsymbol{x}_*)\right] \leq \tau\epsilon_{t-1}^2 = \frac{\tau\epsilon_0^2}{4^{t-1}} = \frac{\Delta}{4^t},$$

where we set $\epsilon_0 = \frac{1}{2}\sqrt{\frac{\Delta}{\tau}}$. The proof is completed. □

C.2 Proof of Theorem 4

Proof. The proof here is very similar to the strategy in Section C.1 for proving Theorem 3. The main idea is to use the result in Lemma 5, to achieve

$$\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}})\|^2 \leq \epsilon^2 \quad \text{and} \quad \mathbb{E}\left[f(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)\right] \leq \tau\epsilon^2,$$

the IFO complexity is

$$\mathcal{O}\left(\frac{L\Delta\sigma}{\epsilon^3}\right), \quad \text{where } \widetilde{\Delta} = f(\boldsymbol{x}_0) - f(\boldsymbol{x}_*).$$

Then following the proof in Section C.1 for proving Theorem 3, we can obtain the IFO complexity for achieving $\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}}_t)\|^2 \leq \epsilon_{t-1}^2$:

$$\mathcal{O}\left(\frac{\tau L\sigma}{\epsilon}\right),$$

when the parameters obey $p_t = \frac{\sigma n_0^t}{\epsilon_{t-1}}$, $\eta_k^t = \frac{\|\boldsymbol{v}_k^t\|}{2Ln_0^t}$, $|\mathcal{S}_1^t| = \frac{32\sigma^2}{\epsilon_{t-1}^2}$, $|\mathcal{S}_{2,k}^t| = \frac{8\sigma\|\boldsymbol{v}_{k-1}^t\|^2}{\epsilon_{t-1}^3 n_0^t}$, and $K^t = \frac{64Ln_0^t\Delta^t}{\epsilon_{t-1}^2}$.

Meanwhile, we can obtain

$$\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}}_t)\| \leq \sqrt{\mathbb{E}\|\nabla f(\widetilde{\boldsymbol{x}}_t)\|^2} \leq \epsilon_{t-1} = \frac{\epsilon_0}{2^{t-1}} = \frac{1}{2^t}\sqrt{\frac{\Delta}{\tau}} \quad \text{and} \quad \mathbb{E}\left[f(\widetilde{\boldsymbol{x}}_t) - f(\boldsymbol{x}_*)\right] \leq \tau\epsilon_{t-1}^2 = \frac{\tau\epsilon_0^2}{4^{t-1}} = \frac{\Delta}{4^t},$$

where we set $\epsilon_0 = \frac{1}{2}\sqrt{\frac{\Delta}{\tau}}$. The proof is completed. □

D More Experimental Results

D.1 Descriptions of Testing Datasets

We first briefly introduce the ten testing datasets in the manuscript. Among them, there are six datasets, including a9a, satimage, covtype, protein, ijcnn1 and epsilon, that are provided in the LibSVM website[1]. We also evaluate our algorithms on the three datasets: YaleB [2], AR [3] and PIE [4], which are very commonly used face classification datasets. Finally, we also test those algorithms on a movie recommendation dataset, namely MovieLens-1M[2]. Their detailed information is summarized in Table 2. From it we can observe that these datasets are different from each other due to their feature dimension, training samples, and class numbers, *etc.*

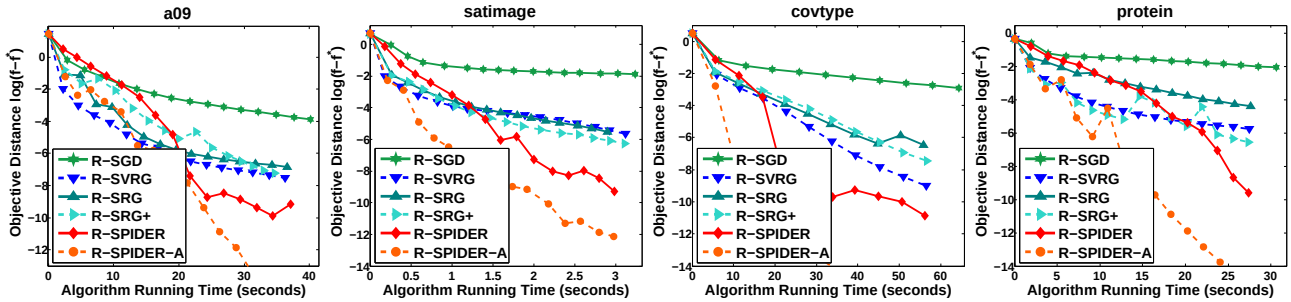
Table 2: Descriptions of the ten testing datasets.

	#class	#sample	#feature		#class	#sample	#feature
a9a	2	32,561	123	epsilon	2	40,000	2000
satimage	6	4,435	36	YaleB	38	2,414	2,016
covtype	2	581,012	54	AR	100	2,600	1,200
protein	3	14,895	357	PIE	64	11,554	1,024
ijcnn1	2	49,990	22	MovieLens-1M	—	6,040	3,706

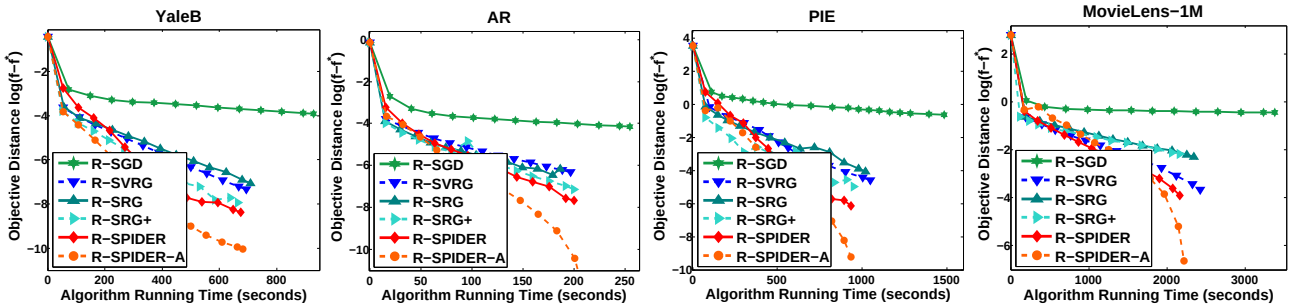
D.2 Comparison of Algorithm Running Time

In this subsection, we present more experimental results to show the algorithm running time comparison among the compared algorithms in the manuscript. The experimental results in Figure 1 only provides the algorithm running time comparison of the ijcnn and epsilon datasets. Here we provide the comparison of all remaining datasets in Figure 5 which respond to Figures 1 and 3 in the manuscript. From the curves of comparison of optimality gap vs. algorithm running time, one can observe that our R-SPIDER-A is the fastest method and R-SPIDER can also quickly converge to a relatively high accuracy, *e.g.* 10^{-8}. We have discussed these results in the manuscript. Besides, all these results are consistent with the curves of the comparison of optimality gap vs. IFO, since the IFO complexity can comprehensively reflect the overall computational performance of a first-order Riemannian algorithm.

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/
[2] https://grouplens.org/datasets/movielens/1m/



(a) Comparison among Riemannian stochastic gradient algorithms on k-PCA problem.



(b) Comparison among Riemannian stochastic gradient algorithms on low-rank matrix completion problem.

Figure 5: Comparison of algorithm running time of Riemannian stochastic gradient algorithms.

D.3 Comparison between Riemannian Stochastic Gradient Algorithms with Adaptive Learning Rate

Here we provide more comparison among our proposed R-SPIDER-A, R-SRG-A and R-SRG+A. R-SRG-A and R-SRG+A are respectively the counterparts of R-SRG and R-SRG+ with adaptive learning rate of formulation $\eta_k = \alpha(1 + \alpha\lambda_\alpha \lfloor \frac{k}{p} \rfloor)$ [5]. Notice, the reason that we do not compare all algorithms together is to avoid too many curves in one figure, leading to poor readability.

By observing Figure 6, we can find that the algorithm with adaptive learning rate usually outperforms the vanilla counterpart, which demonstrates the effectiveness of the strategy of adaptive learning rate. Moreover, R-SPIDER-A also consistently shows sharpest convergence behaviors compared with R-SRG-A and R-SRG+A. All these results are consistent with the experimental results in the manuscript. All results shows the advantages of our proposed R-SPIDER and R-SPIDER-A.

References

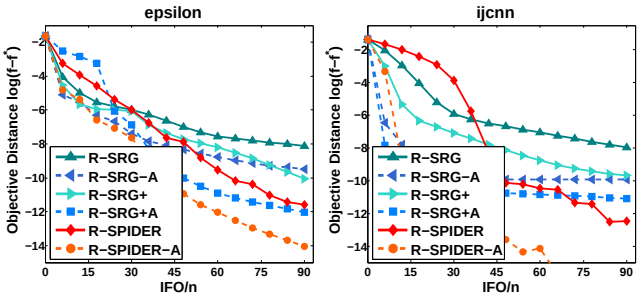
[1] C. Fang, C. Li, Z. Lin, and T. Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018. 1

[2] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23:643–660, Jun. 2001. 8

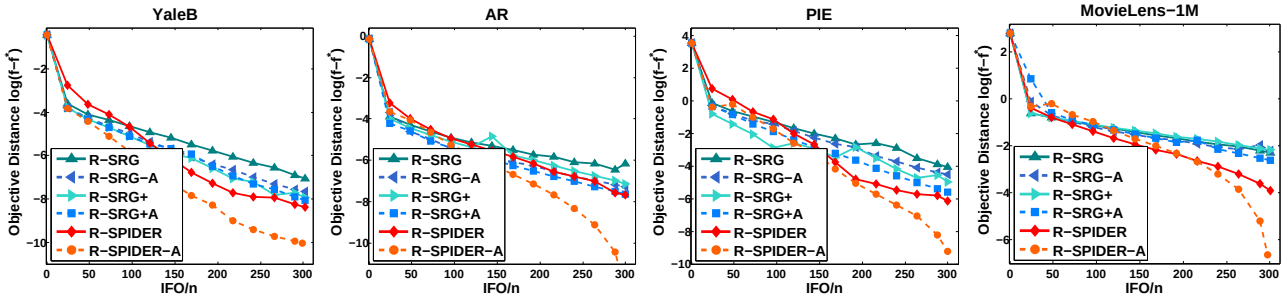
[3] A. Martinez and R. Benavente. The AR face database. *CVC Tech. Rep. 24*, Jun. 1998. 8

[4] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25:1615–1618, Dec. 2003. 8

[5] H. Kasai, H. Sato, and B. Mishra. Riemannian stochastic recursive gradient algorithm with retraction and vector transport and its convergence analysis. In *Proc. Int'l Conf. Machine Learning*, pages 2521–2529, 2018. 9



(a) Comparison among Riemannian stochastic gradient algorithms on k-PCA problem.



(b) Comparison among Riemannian stochastic gradient algorithms on low-rank matrix completion problem.

Figure 6: More comparison between R-SPIDER and R-SRG with adaptive learning rates.