Supplementary Material of Hybrid Stochastic-Deterministic Minibatch Proximal Gradient

Pan Zhou, Xiao-Tong Yuan, *Member, IEEE,* Zhouchen Lin, *Fellow, IEEE,* Steven C.H. Hoi, *Fellow, IEEE*

This supplementary document contains the technical proofs of convergence results and some additional numerical results of the paper entitled "A Hybrid Stochastic-Deterministic Minibatch Proximal Gradient Method for Efficient Optimization and Generalization". It is structured as follows. Appendix A first present several auxiliary lemmas which will be used for subsequent analysis and whose proofs are deferred to Appendix D. Then Appendix B gives the proofs of the main results in Sec. 3, including Theorem 1 which analyzes convergence rate of HSDMPG, Corollaries 1 and 2 which analyze the IFO complexity of HSDMPG on the quadratic problems, and Theorem 2 which analyzes the online problems. Next, Appendix C provides the proofs of the results in Sec. 4, including Theorem 3 which proves the convergence rate of HSDMPG and analyzes its IFO complexity for generic problems, Corollary 3 which gives the IFO complexity of HSDMPG to achieve the intrinsic excess error bound, and Theorem 4 which analyzes the online problems. Then in Appendix D we present the proofs of auxiliary lemmas in Appendix A, including Lemmas 1 $\sim$ 4. Finally, more experimental results are presented in Appendix E.

Appendix A
Some Auxiliary Lemmas

Here we introduce auxiliary lemmas which will be used for proving the results in the manuscript. For the sake of readability, we defer the proofs of some lemmas into Appendix D. The following elementary lemma will be used frequently throughout our analysis.

Lemma 1. *For online setting, suppose that $z_i \in \mathbb{R}^p$ is an arbitrary population vector with $\mathbb{E}_i[z_i]=0$. Let S be a uniform random subset with size n. Then*

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i\in S} z_i\right\|^2 \leq \frac{1}{n}\mathbb{E}_i[\|z_i\|^2].$$

For finite-sum setting, let $z_1,...,z_N \in \mathbb{R}^p$ be an arbitrary population of N vectors with $\sum_{i=1}^N z_i = 0$. Let S be a uniform random subset of $[N]$ with size n. Then

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i\in S} z_i\right\|^2 \leq \frac{\mathbb{1}(n<N)}{n}\frac{1}{N}\sum_{i=1}^N \|z_i\|^2.$$

See its proof in Appendix D.1. The proof is based on the method in [1].

Lemma 2. *Assume that the loss $F(\boldsymbol{\theta})$ is a μ-strongly convex loss, $\sup_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_i\|\boldsymbol{H}^{-1/2}(\nabla F(\boldsymbol{\theta})-\nabla\ell_i(\boldsymbol{\theta}))\|_2^2 \leq \nu^2$ for online setting and $\sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^n\|\boldsymbol{H}^{-1/2}(\nabla F(\boldsymbol{\theta})-\nabla\ell_i(\boldsymbol{\theta}))\|_2^2 \leq \nu^2$ for finite-sum setting. Suppose $\boldsymbol{r}_{t-1}=\nabla F(\boldsymbol{\theta}_{t-1})-\boldsymbol{g}_{t-1}$ where $\boldsymbol{g}_{t-1}=\nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1})$. Then by setting*

$$|\mathcal{S}_t| = \frac{16\nu^2(\mu+2\gamma)^2}{\mu^2}\exp\left(\frac{\mu t}{\mu+2\gamma}\right)\bigwedge n,$$

we have

$$\mathbb{E}\left[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_t\|^2\right] \leq \frac{\mu^2}{16(\mu+2\gamma)^2}\exp\left(-\frac{\mu t}{\mu+2\gamma}\right), \quad \mathbb{E}\left[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_t\|\right] \leq \frac{\mu}{4(\mu+2\gamma)}\exp\left(-\frac{\mu t}{2(\mu+2\gamma)}\right).$$

See its proof in Appendix D.2.

Lemma 3. *For both online and finite-sum settings, suppose $\boldsymbol{H}$ and $\boldsymbol{H}_{\mathcal{S}}$ respectively denote the Hessian matrix of $F(\boldsymbol{\theta})$ and $F_{\mathcal{S}}(\boldsymbol{\theta})$ in problem (1). w.l.o.g., suppose $\|\boldsymbol{x}_i\| \leq r$. Then if $s \geq \frac{28}{3}\log\left(\frac{2d}{\delta}\right)$, with probability at least $1-\delta$, we have*

$$\frac{1}{\frac{3}{2}+\frac{2\gamma}{\mu}} = \frac{2\mu}{3\mu+4\gamma} \leq \|\boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2}\| \leq 2,$$

$$\|\boldsymbol{I}-\lambda\boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2}\| \leq \max\left(1-2\lambda, 1-\frac{\lambda}{\frac{3}{2}+\frac{2\gamma}{\mu}}\right),$$

where $0 \leq \lambda \leq \frac{1}{2}$, *s is the size of* $\mathcal{S}$ *and* $\gamma = \frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d}{\delta}\right) - \mu\right)^+$. *Here* $x^+ = x$ *if* $x > 0$*, otherwise* $x^+ = 0$.

See proof in Appendix D.3.

Lemma 4. *Let* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be two symmetric and positive definite matrices and* $\boldsymbol{B} \succeq \mu I$ *for some* $\mu > 0$*. If* $\|\boldsymbol{A} - \boldsymbol{B}\| \leq \gamma$*, then* $(\boldsymbol{A} + \gamma \boldsymbol{I})^{-1}\boldsymbol{B}$ *is diagonalizable and*

$$\frac{\mu}{\mu + 2\gamma} \leq \left\|\boldsymbol{B}^{1/2}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{B}^{1/2}\right\| \leq 1.$$

Moreover, the following spectral norm bound holds:

$$\|\boldsymbol{I} - \boldsymbol{B}^{1/2}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{B}^{1/2}\| \leq \frac{2\gamma}{\mu + 2\gamma}.$$

See its proof in Appendix D.4.

Appendix B
Proofs for the Results in Section 3

We collect in this appendix section the technical proofs of the results in Section 3 of the main paper.

B.1 Proof of Theorem 1

Proof. We first consider the case where there is a regularization term $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$, namely $\tau = 1$, and then consider the case where there is no regularization term, namely $\tau = 0$. For both cases, their proofs have four steps. To begin with, for brevity, let $\boldsymbol{u}_t = \boldsymbol{H}^{1/2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)$ where $\boldsymbol{H}$ denotes the Hessian matrix of $F(\boldsymbol{\theta})$. In the first step, we establish the relation between $\boldsymbol{u}_t$ and $\boldsymbol{u}_{t-1}$ which will be widely used for subsequent proof. Since for quadratic problems, we have $\mathbb{E}[F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}^*)] = \frac{1}{2}\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}}^2]$. So here we aim to upper bound $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}}^2]$ first, and then use it to upper bound $\mathbb{E}[F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}^*)]$. To bound the second-order moment $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}}^2]$, we need to first bound its first-order moment $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}}]$. So in the second step, we use the result in the first step to upper bound $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}}]$. Then in the third step, we upper bound $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}}^2]$. Finally, we can use above result to upper bound the loss. Please see the proof steps below. Since $F(\boldsymbol{\theta}^*)$ is a constant, without loss of generality, we assume $F(\boldsymbol{\theta}^*) = 0$.

Case 1. there is a regularization term $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$, namely $\tau = 1$. In the following we use the above four steps to prove the desired results for this case.

Step 1. Establish the relation between $\boldsymbol{u}_t$ and $\boldsymbol{u}_{t-1}$.

Since the objective function F is quadratic, namely $F(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, for any $\boldsymbol{\theta}_{t-1}$ the optimal solution $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$ can always be expressed as

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_{t-1} - \boldsymbol{H}^{-1}\nabla F(\boldsymbol{\theta}_{t-1}). \tag{8}$$

Then computing the gradient of P_{t-1} yields

$$\nabla P_{t-1}(\boldsymbol{\theta}_t) = \boldsymbol{g}_{t-1} + \eta\left[\nabla F_{\mathcal{S}}(\boldsymbol{\theta}_t) - \nabla F_{\mathcal{S}}(\boldsymbol{\theta}_{t-1}) + \gamma(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})\right],$$

where $\boldsymbol{g}_{t-1} = \nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1})$. Let $\boldsymbol{H}_{\mathcal{S}}$ denotes the Hessian matrix of the loss on minibatch $\mathcal{S}$. Considering $\boldsymbol{H}_{\mathcal{S}}(\boldsymbol{\theta}_t) \equiv \boldsymbol{H}_{\mathcal{S}}$ holds in the quadratic case, we can obtain $\nabla F_{\mathcal{S}}(\boldsymbol{\theta}_t) - \nabla F_{\mathcal{S}}(\boldsymbol{\theta}_{t-1}) = \boldsymbol{H}_{\mathcal{S}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})$. Thus plugging this results into $\nabla P_{t-1}(\boldsymbol{\theta}_t)$ further yields

$$\begin{aligned}\boldsymbol{\theta}_t =& \boldsymbol{\theta}_{t-1} - \frac{1}{\eta}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{g}_{t-1} + \frac{1}{\eta}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\nabla P_{t-1}(\boldsymbol{\theta}_t)\\ =& \boldsymbol{\theta}_{t-1} - \frac{1}{\eta}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\nabla F(\boldsymbol{\theta}_{t-1}) + \frac{1}{\eta}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\nabla P_{t-1}(\boldsymbol{\theta}_t) + \frac{1}{\eta}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1},\end{aligned}$$

where $\boldsymbol{r}_{t-1} = \nabla F(\boldsymbol{\theta}_{t-1}) - \boldsymbol{g}_{t-1}$. Next plugging Eqn. (8) into the above equation, it establishes

$$\boldsymbol{\theta}_t - \boldsymbol{\theta}^* = \left(\boldsymbol{I} - \frac{1}{\eta}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{H}\right)(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*) + \frac{1}{\eta}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\nabla P_{t-1}(\boldsymbol{\theta}_t) + \frac{1}{\eta}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1}.$$

By multiplying $\boldsymbol{H}^{1/2}$ on both sides of the above recurrent form we have

$$\begin{aligned}\boldsymbol{H}^{1/2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) =& \left(\boldsymbol{I} - \frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right)\boldsymbol{H}^{1/2}(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*)\\ &+ \frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\nabla P_{t-1}(\boldsymbol{\theta}_t) + \frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1}.\end{aligned}$$

Since $\boldsymbol{u}_t = \boldsymbol{H}^{1/2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)$, we have

$$\boldsymbol{u}_t = \left(\boldsymbol{I} - \frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right)\boldsymbol{u}_{t-1} + \frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\nabla P_{t-1}(\boldsymbol{\theta}_t) + \frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1}. \tag{9}$$

Step 2. Upper bound $\mathbb{E}[\|\boldsymbol{u}_t\|]$.
Conditioned on $\boldsymbol{\theta}_{t-1}$ and based on the basic inequality $\|\boldsymbol{T}\boldsymbol{x}\| \leq \|\boldsymbol{T}\|\|\boldsymbol{x}\|$ we get

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{u}_t\|] \leq & \mathbb{E}\left[\left\|\boldsymbol{I}-\frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right\| \|\boldsymbol{u}_{t-1}\|+\frac{1}{\eta}\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\|\|\boldsymbol{H}^{-1/2}\nabla P_{t-1}(\boldsymbol{\theta}_t)\|\right] \\ &+\frac{1}{\eta}\mathbb{E}\left[\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\|\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_{t-1}\|\right].\end{aligned} \tag{10}$$

From Lemma 2, we know that by setting $|\mathcal{S}_t| = \frac{16\nu^2(3\mu+4\gamma)^2}{\mu^2}\exp\left(\frac{\mu t}{3\mu+4\gamma}\right)\bigwedge n$, then the inequality always holds

$$\mathbb{E}\left[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_t\|\right] \leq \frac{\mu}{4(3\mu+4\gamma)}\exp\left(-\frac{\mu t}{2(3\mu+4\gamma)}\right).$$

Suppose $\|\boldsymbol{x}_i\| \leq r$ $(i=1,\cdots,n)$ and $s \geq \frac{28}{3}\log\left(\frac{2d}{\delta}\right)$. Then by using Lemma 3, with probability at least $1-\delta$ we have

$$\begin{aligned}&\frac{1}{\frac{3}{2}+\frac{2\gamma}{\mu}} \leq \|\boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2}\| \leq 2, \\ &\|\boldsymbol{I}-\lambda\boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2}\| \leq \max\left(1-2\lambda, 1-\frac{\lambda}{\frac{3}{2}+\frac{2\gamma}{\mu}}\right),\end{aligned}$$

where $0 \leq \lambda \leq \frac{1}{2}$, s is the size of $\mathcal{S}$ and $\gamma = \frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d}{\delta}\right)-\mu\right)^+$, where $x^+ = x$ if $x > 0$, otherwise $x^+ = 0$.

Similarly, we have $\|\boldsymbol{H}^{-1/2}\nabla P_{t-1}(\boldsymbol{\theta}_t)\| \leq \frac{1}{\sqrt{\mu}}\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\| \leq \frac{\varepsilon_t}{\sqrt{\mu}}$. Now by setting $\lambda = \frac{1}{\eta} = \frac{1}{2}$ we plug the above results into Eqn. (10) and establish

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{u}_t\|] &\overset{\textcircled{1}}{\leq} \left(1-\frac{\mu}{3\mu+4\gamma}\right)\|\boldsymbol{u}_{t-1}\| + \frac{\varepsilon_t}{\sqrt{\mu}} + \mathbb{E}[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_{t-1}\|] \\ &\overset{\textcircled{2}}{\leq} \left(1-\frac{\mu}{3\mu+4\gamma}\right)\|\boldsymbol{u}_{t-1}\| + \frac{\mu}{4(3\mu+4\gamma)}\exp\left(-\frac{\mu(t-1)}{2(3\mu+4\gamma)}\right) + \frac{\mu}{4(3\mu+4\gamma)}\exp\left(-\frac{\mu(t-1)}{2(3\mu+4\gamma)}\right) \\ &= \left(1-\frac{\mu}{3\mu+4\gamma}\right)\|\boldsymbol{u}_{t-1}\| + \frac{\mu}{2(3\mu+4\gamma)}\exp\left(-\frac{\mu(t-1)}{2(3\mu+4\gamma)}\right),\end{aligned}$$

where in the inequality ① we have used $\boldsymbol{H} \succeq \mu\boldsymbol{I}$, ② follows from the condition $\varepsilon_t \leq \frac{\mu^{1.5}}{4(3\mu+4\gamma)}\exp\left(-\frac{\mu(t-1)}{2(3\mu+4\gamma)}\right)$. By taking expectation with respect to $\boldsymbol{\theta}_{t-1}$ we arrive at

$$\mathbb{E}[\|\boldsymbol{u}_t\|] \leq \left(1-\frac{\mu}{3\mu+4\gamma}\right)\mathbb{E}[\|\boldsymbol{u}_{t-1}\|] + \frac{\mu}{2(3\mu+4\gamma)}\exp\left(-\frac{\mu(t-1)}{2(3\mu+4\gamma)}\right).$$

By using induction and the basic fact $(1-a) \leq \exp(-a), \forall a > 0$ and for brevity let $a = \frac{\mu}{2(3\mu+4\gamma)}$, the previous inequality then leads to

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\|_{\boldsymbol{H}}] = \mathbb{E}[\|\boldsymbol{u}_t\|] &\leq (1-2a)\,\mathbb{E}[\|\boldsymbol{u}_{t-1}\|] + a\exp\left(-a(t-1)\right) \\ &= (1-2a)^t\,\mathbb{E}[\|\boldsymbol{u}_0\|] + a\sum_{i=0}^{t-1}(1-2a)^{t-1-i}\exp\left(-ai\right) \\ &\leq \left(\frac{1-2a}{1-a}\right)^t\mathbb{E}[\|\boldsymbol{u}_0\|]\exp(-at) + a\sum_{i=0}^{t-1}\left(\frac{1-2a}{1-a}\right)^{t-1-i}\exp\left(-a(t-1)\right) \\ &\leq \left(\frac{1-2a}{1-a}\right)^t\mathbb{E}[\|\boldsymbol{u}_0\|]\exp(-at) + (1-a)\exp\left(-a(t-1)\right) \\ &\leq \left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}} + (1-a)\exp(a)\right)\exp\left(-at\right) \\ &\leq \left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}} + \exp(2a)\right)\exp\left(-at\right) \\ &\leq \left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}} + e\right)\exp\left(-\frac{\mu t}{2(3\mu+4\gamma)}\right).\end{aligned}$$

This means that for all $\boldsymbol{u}_t$, we have

$$\mathbb{E}[\|\boldsymbol{u}_t\|] \leq \left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}} + e\right)\exp\left(-\frac{\mu t}{2(3\mu+4\gamma)}\right). \tag{11}$$

Since for each iteration, by setting $s \geq \frac{28}{3}\log\left(\frac{2d}{\delta}\right)$ and $\gamma = \frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d}{\delta}\right)-\mu\right)^+$ the result holds with probability at least $1-\delta$, Eqn. (11) holds with probability at least $(1-\delta)^T$, where T is the total iteration number.

Step 3. Upper bound $\mathbb{E}[\|\boldsymbol{u}_t\|^2]$.
From Eqn. (9), we can upper bound $\mathbb{E}[\|\boldsymbol{u}_t\|^2]$ as

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{u}_t\|^2] =&\mathbb{E}\left[\left\|\left(\boldsymbol{I}-\frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right)\boldsymbol{u}_{t-1}\right\|^2\right.\\
&\left.+\frac{1}{\eta^2}\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\nabla P_{t-1}(\boldsymbol{\theta}_t)\|^2+\frac{1}{\eta^2}\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1}\|^2\right]\\
&+\frac{2}{\eta}\mathbb{E}\left[\langle(\boldsymbol{I}-\frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2})\boldsymbol{u}_{t-1},\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\nabla P_{t-1}(\boldsymbol{\theta}_t)\rangle\right]\\
&+\frac{2}{\eta}\mathbb{E}\left[\langle(\boldsymbol{I}-\frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2})\boldsymbol{u}_{t-1},\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1}\rangle\right]\\
&+\frac{2}{\eta^2}\mathbb{E}\left[\langle\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\nabla P_{t-1}(\boldsymbol{\theta}_t),\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1}\rangle\right].
\end{aligned}$$

Since $\mathbb{E}_{\mathcal{S}_{t-1}}[\boldsymbol{r}_{t-1}]=0$, it is easy to obtain

$$\begin{aligned}
&\mathbb{E}\left[\langle(\boldsymbol{I}-\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2})\boldsymbol{u}_{t-1},\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1}\rangle\right]\\
=&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{S}_{t-1}}\left[\langle(\boldsymbol{I}-\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2})\boldsymbol{u}_{t-1},\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1}\rangle\right]\\
=&\mathbb{E}_{\mathcal{S}}\left[\langle(\boldsymbol{I}-\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2})\boldsymbol{u}_{t-1},\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\mathbb{E}_{\mathcal{S}_{t-1}}\boldsymbol{r}_{t-1}\rangle\right]=0.
\end{aligned}$$

Conditioned on $\boldsymbol{\theta}_{t-1}$ and based on the basic inequality $\|\boldsymbol{T}\boldsymbol{x}\|\leq\|\boldsymbol{T}\|\|\boldsymbol{x}\|$, we get

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{u}_t\|^2]\\
\leq&\mathbb{E}\left[\|(\boldsymbol{I}-\frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2})\|^2\|\boldsymbol{u}_{t-1}\|^2+\frac{1}{\eta^2}\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\|^2\|\boldsymbol{H}^{-1/2}\nabla P_{t-1}(\boldsymbol{\theta}_t)\|^2\right]\\
&+\frac{1}{\eta^2}\mathbb{E}\left[\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\|^2\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_{t-1}\|^2\right]\\
&+\frac{2}{\eta}\mathbb{E}\left[\|(\boldsymbol{I}-\frac{1}{\eta}\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2})\|\cdot\|\boldsymbol{u}_{t-1}\|\cdot\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\|\cdot\|\boldsymbol{H}^{-1/2}\nabla P_{t-1}(\boldsymbol{\theta}_t)\|\right]\\
&+\frac{2}{\eta}\mathbb{E}\left[\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\|^2\cdot\|\boldsymbol{H}^{-1/2}\nabla P_{t-1}(\boldsymbol{\theta}_t)\|\cdot\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_{t-1}\|\right].
\end{aligned}\tag{12}$$

From Lemma 2, we know that by setting $|\mathcal{S}_t|=\frac{16\nu^2(3\mu+4\gamma)^2}{\mu^2}\exp\left(\frac{\mu t}{3\mu+4\gamma}\right)\bigwedge n$, then the inequality always holds

$$\mathbb{E}\left[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_t\|\right]\leq\frac{\mu}{4(3\mu+4\gamma)}\exp\left(-\frac{\mu t}{2(3\mu+4\gamma)}\right).$$

Then by using Lemma 3, with probability at least $1-\delta$ we have

$$\begin{aligned}
&\frac{1}{\frac{3}{2}+\frac{2\gamma}{\mu}}\leq\|\boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2}\|\leq 2,\\
&\|\boldsymbol{I}-\lambda\boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2}\|\leq\max\left(1-2\lambda,1-\frac{\lambda}{\frac{3}{2}+\frac{2\gamma}{\mu}}\right),
\end{aligned}\tag{13}$$

where $0\leq\lambda\leq\frac{1}{2}$, s is the size of $\mathcal{S}$ and $\gamma=\frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d}{\delta}\right)-\mu\right)^+$.

Similarly, by setting $\lambda=\frac{1}{\eta}=\frac{1}{2}$ we have $\|\boldsymbol{H}^{-1/2}\nabla P_{t-1}(\boldsymbol{\theta}_t)\|\leq\frac{1}{\sqrt{\mu}}\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\|\leq\frac{\varepsilon_t}{\sqrt{\mu}}$. Now we plug the above results into Eqn. (12) and establish

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{u}_t\|^2]\leq&\left(1-\frac{\mu}{3\mu+4\gamma}\right)^2\mathbb{E}[\|\boldsymbol{u}_{t-1}\|^2]+\frac{\varepsilon_t^2}{\mu}+\frac{\mu^2}{16(3\mu+4\gamma)^2}\exp\left(-\frac{\mu t}{3\mu+4\gamma}\right)+2\left(1-\frac{\mu}{3\mu+4\gamma}\right)\frac{\varepsilon_t}{\sqrt{\mu}}\mathbb{E}\left[\|\boldsymbol{u}_{t-1}\|\right]\\
&+\frac{\varepsilon_t}{\sqrt{\mu}}\frac{\mu}{(3\mu+4\gamma)}\exp\left(-\frac{\mu t}{2(3\mu+4\gamma)}\right).
\end{aligned}$$

Finally, by using $\mathbb{E}[\|\boldsymbol{u}_t\|] \leq (\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{\boldsymbol{H}} + e)\exp\left(-\frac{\mu t}{2(3\mu+4\gamma)}\right)$ and $\varepsilon_t \leq \frac{\mu^{1.5}}{4(3\mu+4\gamma)}\exp\left(-\frac{\mu(t-1)}{2(3\mu+4\gamma)}\right)$, we can obtain

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{u}_t\|^2]\\
\leq&\left(1-\frac{\mu}{3\mu+4\gamma}\right)^2\mathbb{E}[\|\boldsymbol{u}_{t-1}\|^2]+\frac{\mu^2}{16(3\mu+4\gamma)^2}\left(1+\exp\left(\frac{\mu}{3\mu+4\gamma}\right)+4\exp\left(\frac{\mu}{2(3\mu+4\gamma)}\right)\right)\exp\left(-\frac{\mu t}{3\mu+4\gamma}\right)\\
&+b\left(1+\frac{2\gamma}{\mu}\right)\frac{\mu^2}{(3\mu+4\gamma)^2}\exp\left(\frac{\mu}{2(3\mu+4\gamma)}\right)\exp\left(-\frac{\mu t}{3\mu+4\gamma}\right)\\
\overset{\textcircled{1}}{\leq}&(1-a)^2\,\mathbb{E}[\|\boldsymbol{u}_{t-1}\|^2]+\frac{a^2}{2}\exp\left(-at\right)+2b\left(1+\frac{2\gamma}{\mu}\right)a^2\exp\left(-at\right)\\
=&(1-a)^2\,\mathbb{E}[\|\boldsymbol{u}_{t-1}\|^2]+a^2\left(\frac{1}{2}+2b\left(1+\frac{2\gamma}{\mu}\right)\right)\exp\left(-at\right),
\end{aligned}$$

where $a=\frac{\mu}{3\mu+4\gamma}$ and $b=(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}+e)$. ① uses $1+\exp\left(\frac{\mu}{3\mu+4\gamma}\right)+4\exp\left(\frac{\mu}{2(3\mu+4\gamma)}\right)\leq 8$ and $\exp\left(\frac{\mu}{2(3\mu+4\gamma)}\right)\leq 2$. By using induction and the basic fact $(1-a)\leq\exp(-a),\forall a>0$ and for brevity letting $c=a^2\left(\frac{1}{2}+2b\left(1+\frac{2\gamma}{\mu}\right)\right)$, the previous inequality then leads to

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\|_{\boldsymbol{H}}^2]=\mathbb{E}[\|\boldsymbol{u}_t\|^2]\leq&(1-a)^2\,\mathbb{E}[\|\boldsymbol{u}_{t-1}\|^2]+c\exp\left(-at\right)\\
=&(1-a)^{2t}\,\mathbb{E}[\|\boldsymbol{u}_0\|^2]+c\sum_{i=1}^{t}(1-a)^{t-i}\exp\left(-ai\right)\\
\leq&\mathbb{E}[\|\boldsymbol{u}_0\|^2]\exp(-2at)+c\exp\left(-at\right)\\
\leq&\left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}^2+a^2\left(\frac{1}{2}+2b\left(1+\frac{2\gamma}{\mu}\right)\right)\right)\exp\left(-\frac{\mu t}{3\mu+4\gamma}\right).
\end{aligned}\tag{14}$$

For each iteration, by setting $s\geq\frac{28}{3}\log\left(\frac{2d}{\delta}\right)$ and $\gamma=\frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d}{\delta}\right)-\mu\right)^+$ the result holds with probability at least $(1-\delta)^{T+1}$, where probability $(1-\delta)^T$ comes from the fact that Eqn. (11) holds, and probability $(1-\delta)$ comes from the fact that Eqn. (13) holds. So Eqn. (14) holds with probability at least $(1-\delta)^{T+1}\geq 1-(T+1)\delta$, where T is the total iteration number. Therefore, by setting $s\geq\frac{28}{3}\log\left(\frac{2d(T+1)}{\delta}\right)$ and $\gamma=\frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d(T+1)}{\delta}\right)-\mu\right)^+$, Eqn. (14) holds with probability at least $1-\delta$.

Step 4. Bound $\mathbb{E}[F(\boldsymbol{\theta}_t)-F(\boldsymbol{\theta}^*)]$.

It is easy to check $\mathbb{E}[F(\boldsymbol{\theta}_t)-F(\boldsymbol{\theta}^*)]=\frac{1}{2}\mathbb{E}[\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\|_{\boldsymbol{H}}^2]$ in the quadratic case. So with probability at least $1-\delta$, we have

$$\begin{aligned}
&\mathbb{E}[F(\boldsymbol{\theta}_t)-F(\boldsymbol{\theta}^*)]=\frac{1}{2}\mathbb{E}[\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\|_{\boldsymbol{H}}^2]\\
\leq&\frac{1}{2}\left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}^2+\frac{\mu^2}{(3\mu+4\gamma)^2}\left(\frac{1}{2}+2\left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}+e\right)\left(1+\frac{2\gamma}{\mu}\right)\right)\right)\exp\left(-\frac{\mu t}{3\mu+4\gamma}\right)\\
\leq&\frac{1}{2}\left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}^2+\frac{\mu^2}{(3\mu+4\gamma)^2}\left(\frac{1}{2}+2\left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}+e\right)\left(1+\frac{2\gamma}{\mu}\right)\right)\right)\exp\left(-\frac{\mu t}{3\mu+4\gamma}\right)\\
\overset{\textcircled{1}}{\leq}&\frac{1}{2}\left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}^2+\frac{2}{9}\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}+\frac{3}{4}\right)\exp\left(-\frac{\mu t}{3\mu+4\gamma}\right)\\
\leq&\left(\frac{1}{2}\left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}+\frac{1}{9}\right)^2+\frac{3}{8}\right)\exp\left(-\frac{\mu t}{3\mu+4\gamma}\right),
\end{aligned}$$

where ① uses $\frac{\mu^2}{(3\mu+4\gamma)^2}\leq\frac{1}{9}$ and $\frac{\mu\gamma}{(3\mu+4\gamma)^2}\leq\frac{1}{49}$.

Case 2. there is no regularization term $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$ (*i.e.* $\tau=0$) but the function $F(\boldsymbol{\theta})$ is μ-strongly convex. Here since in the above Step $1\sim4$, we always use the μ-strongly-convexity of the whole function $F(\boldsymbol{\theta})$, we can directly use the same proof to obtain the desired results. The proof is completed. □

B.2 Proof of Corollary 1

Proof. We first consider the case where there is a regularization term $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$, namely $\tau=1$, and then consider the case where there is no regularization term, namely $\tau=0$. For both cases, their proofs have four steps. In the first step, we estimate the smallest iteration number T such that $\mathbb{E}[F(\boldsymbol{\theta}_T)-F(\boldsymbol{\theta}^*)]\leq\epsilon$. Since the IFO complexity comes from two aspects: (1) the outer sampling steps for constructing the proximal function $P_t(\boldsymbol{\theta})=\langle\nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}\rangle+\eta\big[F_{\mathcal{S}}(\boldsymbol{\theta})-\langle\nabla F_{\mathcal{S}}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}\rangle+\frac{\gamma}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1}\|_2^2\big]$ which requires sampling the gradient $\nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1})$; (2) the inner optimization complexity which is produced by SVRG to solve the inner problem $P_t(\boldsymbol{\theta})$ such that $\mathbb{E}\|P_t(\boldsymbol{\theta})\|\leq\varepsilon_t$. So in the second step, we estimate computational complexity of the outer sampling. In the third step, we estimate computational complexity of the inner optimization via SVRG. Finally, we combine these two kinds of complexity together to obtain total IFO bounds. Please see the proof steps below.

Case 1. there is a regularization term $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$, namely $\tau=1$. In the following we use the above four steps to prove the desired results.
Step 1. Estimate the smallest iteration number T such that $\mathbb{E}[F(\boldsymbol{\theta}_T)-F(\boldsymbol{\theta}^*)]\leq\epsilon$.
According to Theorem 1, we have

$$\mathbb{E}[F(\boldsymbol{\theta}_t)-F(\boldsymbol{\theta}^*)]=\frac{1}{2}\mathbb{E}[\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\|_{\boldsymbol{H}}^2]\leq\zeta\exp\left(-\frac{\mu t}{3\mu+4\gamma}\right),$$

where $\zeta=\frac{1}{2}\left(\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_*\|_{\boldsymbol{H}}+\frac{1}{9}\right)^2+\frac{3}{8}$ with $\|\boldsymbol{\theta}\|_{\boldsymbol{H}}=\sqrt{\boldsymbol{\theta}^\top\boldsymbol{H}\boldsymbol{\theta}}$. In this way, to guarantee $\mathbb{E}[F(\boldsymbol{\theta}_t)-F(\boldsymbol{\theta}^*)]\leq\epsilon$, the iteration number T should be satisfies

$$T=\frac{3\mu+4\gamma}{\mu}\log\left(\frac{\zeta}{\epsilon}\right).$$

Step 2. Estimate computational complexity of the outer sampling.
The stochastic gradient estimation complexity up to the time step T is given by

$$\begin{aligned}\sum_{t=0}^{T-1}|\mathcal{S}_t|\leq&\frac{16\nu^2(3\mu+4\gamma)^2}{\mu^2}\sum_{t=0}^{T-1}\exp\left(\frac{\mu t}{3\mu+4\gamma}\right)=\frac{16\nu^2(3\mu+4\gamma)^2}{\mu^2}\frac{\exp\left(\frac{\mu T}{3\mu+4\gamma}\right)-1}{\exp\left(\frac{\mu}{3\mu+4\gamma}\right)-1}\\ \overset{①}{\leq}&\frac{16\nu^2(3\mu+4\gamma)^2}{\mu^2}\frac{3\mu+4\gamma}{2\mu}\frac{\zeta}{\epsilon}=\frac{16\zeta\nu^2(3\mu+4\gamma)^3}{\mu^3\epsilon},\end{aligned}$$

where in ① we have used the definition of T such that $\exp\left(\frac{\mu T}{3\mu+4\gamma}\right)=\frac{\zeta}{\epsilon}$ and the fact $\exp(a)\geq 1+a,\forall a>0$. At the same time, we also have

$$\sum_{t=0}^{T-1}|\mathcal{S}_t|\leq nT=\frac{(3\mu+4\gamma)n}{\mu}\log\left(\frac{\zeta}{\epsilon}\right).$$

By combing the above two inequalities we obtain the computational complexity of the outer sampling as

$$\frac{16\zeta\nu^2(3\mu+4\gamma)^3}{\mu^3\epsilon}\bigwedge\frac{(3\mu+4\gamma)n}{\mu}\log\left(\frac{\zeta}{\epsilon}\right)=\mathcal{O}\left(\left(1+\frac{r^6\log^3\left(\frac{dr^2}{\delta\mu s}\right)}{s^3\mu^3}\right)\frac{\nu^2}{\epsilon}\bigwedge\left(1+\frac{r^2\log\left(\frac{dr^2}{\delta\mu s}\right)}{s\mu}\right)n\log\left(\frac{1}{\epsilon}\right)\right).$$

where we plug $s\geq\frac{28}{3}\log\left(\frac{2d(T+1)}{\delta}\right)$ and $\gamma=\frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d(T+1)}{\delta}\right)-\mu\right)^+$, and ignore $\log(\log(1/\epsilon))$.
Step 3. Estimate computational complexity of the inner optimization via SVRG.
At each iteration time stamp t, we need to optimize the inner problem $P_t(\boldsymbol{\theta})=\langle\nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}\rangle+\eta\big[F_{\mathcal{S}}(\boldsymbol{\theta})-\langle\nabla F_{\mathcal{S}}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}\rangle+\frac{\gamma}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1}\|_2^2\big]$. In $P_t(\boldsymbol{\theta})$, its finites-sum structure comes from $F_{\mathcal{S}}(\boldsymbol{\theta})$ and its gradient, where $\eta=2$.

For $2(\mu+\gamma)$-strongly-convex and $2(L+\gamma)$-smooth problem, it is standardly known that the IFO complexity of the inner-loop SVRG computation to achieve $\mathbb{E}[P_{t-1}(\boldsymbol{\theta}_T)-P_{t-1}(\boldsymbol{\theta}^*)]\leq\varepsilon_t$ can be bounded in expectation by $\mathcal{O}\left(\left(s+\frac{L+\gamma}{\gamma+\mu}\right)\log\left(\frac{1}{\epsilon_t}\right)\right)$, where $\boldsymbol{\theta}^*$ denotes the optimal solution of $P_{t-1}(\boldsymbol{\theta})$. Since $P_{t-1}(\boldsymbol{\theta})$ is $2(\mu+\gamma)$-strongly-convex, we have $\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\|_2\leq 4(\mu+\gamma)(P_{t-1}(\boldsymbol{\theta}_T)-P_{t-1}(\boldsymbol{\theta}^*))$. In this way, to achieve $\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\|_2\leq\varepsilon_t=\frac{\mu^{1.5}}{4(3\mu+4\gamma)}\exp\left(-\frac{\mu(t-1)}{2(3\mu+4\gamma)}\right)$, the expected IFO complexity of SVRG is

$$\begin{aligned}\mathcal{O}\left(\left(s+\frac{L+\gamma}{\gamma+\mu}\right)\log\left(\frac{4(\mu+\gamma)}{\epsilon_t}\right)\right)=&\mathcal{O}\left(\left(s+\frac{L}{\gamma}\right)\log\left(\frac{(\mu+\gamma)^2}{\mu^{1.5}}\exp\left(\frac{\mu(t-1)}{\mu+\gamma}\right)\right)\right)\\=&\mathcal{O}\left(\left(s+\frac{L}{\gamma}\right)\left(\log\left(\frac{(\mu+\gamma)^2}{\mu^{1.5}}\right)+\frac{\mu(t-1)}{\mu+\gamma}\right)\right).\end{aligned}\tag{15}$$

From above result we know that $\mathbb{E}[F(\boldsymbol{\theta}_t)]\leq F(\boldsymbol{\theta}^*)+\epsilon$ after $T=\mathcal{O}\left(\frac{\gamma}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$ rounds of iteration. Therefore the total inner-loop IFO complexity is bounded in expectation by

$$\begin{aligned}&\mathcal{O}\left(\sum_{t=1}^{T}\left\{\left(s+\frac{L}{\gamma}\right)\left(\log\left(\frac{(\mu+\gamma)^2}{\mu^{1.5}}\right)+\frac{\mu(t-1)}{\mu+\gamma}\right)\right\}\right)=\mathcal{O}\left(\left(s+\frac{L}{\gamma}\right)\left(T\log\left(\frac{(\mu+\gamma)^2}{\mu^{1.5}}\right)+\frac{\mu T^2}{\gamma}\right)\right)\\&=\mathcal{O}\left(\left(s+\frac{L}{\gamma}\right)\left(\frac{\gamma}{\mu}\log\left(\frac{(\mu+\gamma)^2}{\mu^{1.5}}\right)\log\left(\frac{1}{\epsilon}\right)+\frac{\gamma}{\mu}\log^2\left(\frac{1}{\epsilon}\right)\right)\right).\end{aligned}$$

We plug $\gamma=\frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d(T+1)}{\delta}\right)-\mu\right)^+$ into the above inner-loop IFO bound to obtain

$$\mathcal{O}\left(\left(\frac{L}{\mu}+\log(d)\right)\log\left(\frac{1}{\epsilon}\right)\left(\log\left(\frac{\log^2\left(\frac{dr^2}{\delta\mu s}\right)}{\mu^{1.5}s^2}\right)+\log\left(\frac{1}{\epsilon}\right)\right)\right).$$

Step 4. Combing inner optimization complexity and outer sampling complexity to obtain total IFO bounds.
Combing the preceding inner-loop optimization complexity and outer sampling complexity yields the following overall computation complexity bound

$$\begin{aligned}&\mathcal{O}\left(\left(\frac{L}{\mu}+\log(d)\right)\log\left(\frac{1}{\epsilon}\right)\left(\log\left(\frac{\log^2\left(\frac{dr^2}{\delta\mu s}\right)}{\mu^{1.5}s^2}\right)+\log\left(\frac{1}{\epsilon}\right)\right)+\left(1+\frac{r^6\log^3\left(\frac{dr^2}{\delta\mu s}\right)}{s^3\mu^3}\right)\frac{\nu^2}{\epsilon}\bigwedge\left(1+\frac{r^2\log\left(\frac{dr^2}{\delta\mu s}\right)}{s\mu}\right)n\log\left(\frac{1}{\epsilon}\right)\right)\\&\overset{①}{=}\mathcal{O}\left((\kappa+\log d)\log^2\left(\frac{1}{\epsilon}\right)+\frac{\nu^2}{\epsilon}\bigwedge n\log\left(\frac{1}{\epsilon}\right)\right),\end{aligned}\tag{16}$$

where ① holds by setting $s\geq\frac{r^2\log\left(\frac{dr^2}{\delta\mu}\right)}{\mu}$.
Case 2. there is no regularization term $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$ (*i.e.* $\tau=0$) but the function $F(\boldsymbol{\theta})$ is μ-strongly convex. For steps 1 and 2, because $F(\boldsymbol{\theta})$ has the same linear convergence rate and the same parameter setting, we directly obtain the same results in steps 1 and 2.

For step 3, the inner problem $P_t(\boldsymbol{\theta})=\langle\nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}\rangle+\eta\left[F_{\mathcal{S}}(\boldsymbol{\theta})-\langle\nabla F_{\mathcal{S}}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}\rangle+\frac{\gamma}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1}\|_2^2\right]$ is 2γ-strongly-convex and $2(L+\gamma)$-smooth. So the IFO complexity of the inner-loop SVRG computation to achieve $\mathbb{E}[P_{t-1}(\boldsymbol{\theta}_T)-P_{t-1}(\boldsymbol{\theta}^*)]\leq\varepsilon_t$ can be bounded in expectation by $\mathcal{O}\left(\left(s+\frac{L+\gamma}{\gamma}\right)\log\left(\frac{1}{\epsilon_t}\right)\right)$, where $\boldsymbol{\theta}^*$ denotes the optimal solution of $P_{t-1}(\boldsymbol{\theta})$. Since $P_{t-1}(\boldsymbol{\theta})$ is 2γ-strongly-convex, we have $\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\|_2\leq 4\gamma(P_{t-1}(\boldsymbol{\theta}_T)-P_{t-1}(\boldsymbol{\theta}^*))$. In this way, to achieve $\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\|_2\leq\varepsilon_t=\frac{\mu^{1.5}}{4(3\mu+4\gamma)}\exp\left(-\frac{\mu(t-1)}{2(3\mu+4\gamma)}\right)$, the expected IFO complexity of SVRG is

$$\begin{aligned}\mathcal{O}\left(\left(s+\frac{L+\gamma}{\gamma}\right)\log\left(\frac{4\gamma}{\epsilon_t}\right)\right)=&\mathcal{O}\left(\left(s+\frac{L}{\gamma}\right)\log\left(\frac{(\mu+\gamma)\gamma}{\mu^{1.5}}\exp\left(\frac{\mu(t-1)}{\mu+\gamma}\right)\right)\right)\\=&\mathcal{O}\left(\left(s+\frac{L}{\gamma}\right)\left(\log\left(\frac{(\mu+\gamma)\gamma}{\mu^{1.5}}\right)+\frac{\mu(t-1)}{\mu+\gamma}\right)\right).\end{aligned}$$

From above result we know that $\mathbb{E}[F(\boldsymbol{\theta}_t)]\leq F(\boldsymbol{\theta}^*)+\epsilon$ after $T=\mathcal{O}\left(\frac{\gamma}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$ rounds of iteration. Therefore the total inner-loop IFO complexity is bounded in expectation by

$$\begin{aligned}&\mathcal{O}\left(\sum_{t=1}^{T}\left\{\left(s+\frac{L}{\gamma}\right)\left(\log\left(\frac{(\mu+\gamma)^2}{\mu^{1.5}}\right)+\frac{\mu(t-1)}{\mu+\gamma}\right)\right\}\right)=\mathcal{O}\left(\left(s+\frac{L}{\gamma}\right)\left(T\log\left(\frac{(\mu+\gamma)^2}{\mu^{1.5}}\right)+\frac{\mu T^2}{\gamma}\right)\right)\\&=\mathcal{O}\left(\left(s+\frac{L}{\gamma}\right)\left(\frac{\gamma}{\mu}\log\left(\frac{(\mu+\gamma)^2}{\mu^{1.5}}\right)\log\left(\frac{1}{\epsilon}\right)+\frac{\gamma}{\mu}\log^2\left(\frac{1}{\epsilon}\right)\right)\right).\end{aligned}$$

We plug $\gamma=\frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d(T+1)}{\delta}\right)-\mu\right)^+$ into the above inner-loop IFO bound to obtain

$$\mathcal{O}\left(\left(\frac{L}{\mu}+\log(d)\right)\log\left(\frac{1}{\epsilon}\right)\left(\log\left(\frac{\log^2\left(\frac{dr^2}{\delta\mu s}\right)}{\mu^{1.5}s^2}\right)+\log\left(\frac{1}{\epsilon}\right)\right)\right).$$

So we have the same complexity for step 3. In this way, we can directly combine results as above step 4 and thus obtain the same computational complexity. This competes the proof. □

B.3 Proof of Corollary 2

Proof. The result in Corollary 2 can be easily obtained. Specifically, we plug $\epsilon=\mathcal{O}(\frac{1}{\sqrt{n}})$, $\kappa=\mathcal{O}(\sqrt{n})$ into Corollary 1 and can compute the desired results. □

B.4 Proof of Theorem 2

Proof. For the linear convergence rate, we can follow the proof framework in the proof of Theorem 1. The difference is that we need to use the results in Lemmas 1, 2, 3 and 4 under online settings. So for this part, our main contribution is to prove Lemmas 1, 2, 3 and 4 under online settings. Please refer to their proof in Sec. D.

For the computational complexity, we can follow the analysis framework in proof of Corollary 1. Note by using our method, the optimization at each iteration actually becomes the finite-sum optimization problem. Specifically, $F_{\mathcal{S}}(\boldsymbol{\theta})$ in Eqn. (2) and the subproblem (5) both have finite-sum structure. So we can use the same computational complexity analysis method in proof of Corollary 1. The proof is completed. □

Appendix C
Proofs for the Results in Section 4

C.1 Proof of Theorem 3

Proof. Here we consider finite-sum setting, which includes two cases (1) there is a regularization $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$ and (2) there is no regularization $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$ but $F(\boldsymbol{\theta})$ is μ-strongly-convex. For both cases, their proofs have two steps. In the first step, we prove the results in the first part of Theorem 3, namely the linearly convergence of $F(\boldsymbol{\theta})$ on the generic loss functions. Then in the second step, we analyze the computational complexity of HSDMPG on the generic loss functions. Please refer to the following detailed steps.

Case (1) in finite-sum setting where where there is a regularization $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$.

Step 1. Establish linearly convergence of $F(\boldsymbol{\theta})$.

To begin with, by using the smoothness property of each individual loss function $\ell(\boldsymbol{\theta}^\top\boldsymbol{x},\boldsymbol{y})$ we can obtain

$$F(\boldsymbol{\theta}_t) \leq \boldsymbol{Q}_{t-1}(\boldsymbol{\theta}_t) = F(\boldsymbol{\theta}_{t-1}) + \langle \nabla F(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\rangle + \Delta_{t-1}(\boldsymbol{\theta}_t),$$

where $\Delta_{t-1}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1})^\top \bar{\boldsymbol{H}}(\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1})$ with $\bar{\boldsymbol{H}} = \frac{L}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top + \mu\boldsymbol{I}$.

On the other hand, from our optimization rule, we can establish for any $z\in[0,1]$

$$\begin{aligned}
&\boldsymbol{Q}_{t-1}(\boldsymbol{\theta}_t) \leq \boldsymbol{Q}_{t-1}((1-z)\boldsymbol{\theta}_t + z\boldsymbol{\theta}^*) + \varepsilon_t'\\
=&F(\boldsymbol{\theta}_{t-1}) + z\langle \nabla F(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\rangle + \frac{Lz^2}{2}(\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1})^\top\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{\mu}{L}\boldsymbol{I}\right)(\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1}) + \varepsilon_t'.
\end{aligned}$$

Next, from the σ-strongly convexity of each loss $\ell(\boldsymbol{\theta}^\top\boldsymbol{x},\boldsymbol{y})$, we can obtain $\nabla^2 F(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \mu\boldsymbol{I} \succeq \frac{\sigma}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top + \mu\boldsymbol{I}$ for all $\boldsymbol{\theta}$. In this way, we can lower bound

$$\begin{aligned}
F(\boldsymbol{\theta}^*) \geq& F(\boldsymbol{\theta}_{t-1}) + \langle \nabla F(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\rangle + \frac{\sigma}{2}(\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1})^\top\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{\mu}{\sigma}\boldsymbol{I}\right)(\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1})\\
\overset{\text{①}}{\geq}& F(\boldsymbol{\theta}_{t-1}) + \langle \nabla F(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\rangle + \frac{\sigma}{2}(\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1})^\top\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{\mu}{L}\boldsymbol{I}\right)(\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1})
\end{aligned}$$

where ① we use $L\geq\sigma$. By setting $z=\frac{\sigma}{L}$ and combining all results together, we have

$$\begin{aligned}
&F(\boldsymbol{\theta}_t) \leq \boldsymbol{Q}_{t-1}(\boldsymbol{\theta}_t)\\
\leq& F(\boldsymbol{\theta}_{t-1}) + \frac{\sigma}{L}\left[\langle \nabla F(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\rangle + \frac{\sigma}{2}(\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1})^\top\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{\mu}{L}\boldsymbol{I}\right)(\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1})\right] + \varepsilon_t'\\
\leq& F(\boldsymbol{\theta}_{t-1}) + \frac{\sigma}{L}\left[F(\boldsymbol{\theta}^*) - F(\boldsymbol{\theta}_{t-1})\right] + \varepsilon_t'.
\end{aligned}$$

Then by using the basic fact $(1-a)\leq\exp(-a), \forall a>0$ and $\varepsilon_t' = \frac{\sigma}{2L}\exp\left(-\frac{\sigma(t-1)}{2L}\right)$ we rewrite this equation and obtain

$$\begin{aligned}
F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}^*) \leq& \left(1-\frac{\sigma}{L}\right)\left(F(\boldsymbol{\theta}_{t-1}) - F(\boldsymbol{\theta}^*)\right) + \frac{\sigma}{2L}\exp\left(-\frac{\sigma(t-1)}{2L}\right)\\
\overset{\text{①}}{=}& (1-2a)^t\left(F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^*)\right) + a\sum_{i=1}^t (1-2a)^{t-i}\exp\left(-a(i-1)\right)\\
\overset{\text{②}}{\leq}& \left(\frac{1-2a}{1-a}\right)^t\left(F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^*)\right)\exp(-at) + a\sum_{i=1}^t\left(\frac{1-2a}{1-a}\right)^{t-i}\exp\left(-a(t-1)\right)\\
=& \left(\frac{1-2a}{1-a}\right)^t\left(F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^*)\right)\exp(-at) + (1-a)\exp\left(-a(t-1)\right)\\
\leq& \left(F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^*) + (1-a)\exp(a)\right)\exp(-at)\\
\leq& \left(F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^*) + 1\right)\exp(-at),
\end{aligned}$$

where in ① we let $a=\frac{\sigma}{2L}$ for brevity; ② uses $(1-a)^k\leq\exp(-ak)$ for $a>0$.

Step 2. Establish computational complexity of HSDMPG for achieving $\mathbb{E}[F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}^*)]\leq\epsilon$.

It follows immediately that $\mathbb{E}[F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}^*)]\leq\epsilon$ is valid when

$$t \geq \frac{2L}{\sigma}\log\left(\frac{F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^*) + 1}{\epsilon}\right).$$

At each iteration time stamp t, the leading terms in Theorem 1 suggest that the IFO complexity of the inner-loop HSDMPG computation to achieve ε_t'-sub-optimality of $\boldsymbol{Q}_t$ can be bounded in expectation by

$$\mathcal{O}\left(\kappa\sqrt{s\log(d)}\log^2\left(\frac{1}{\varepsilon_t'}\right)+\kappa^3\left(\frac{\log(d)}{s}\right)^{1.5}\frac{\nu^2}{\varepsilon_t'}\right)=\mathcal{O}\left(\frac{\sigma^2\sqrt{s\log(d)}}{L\mu}t^2+\left(\frac{L}{\mu}\sqrt{\frac{\log(d)}{s}}\right)^3\frac{L\nu^2}{\sigma}\exp\left(\frac{\sigma}{L}t\right)\right)$$

where $\kappa=\frac{L}{\mu}$ denotes the conditional number.

$$\mathcal{O}\left((\kappa+\log d)\log^2\left(\frac{1}{\varepsilon_t'}\right)+\frac{\nu^2}{\varepsilon_t'}\bigwedge n\log\left(\frac{1}{\varepsilon_t'}\right)\right)=\mathcal{O}\left(\frac{\sigma^2(\kappa+\log d)}{L^2}t^2+\frac{L\nu^2}{\sigma}\exp\left(\frac{\sigma}{L}t\right)\bigwedge\frac{\sigma n}{L}t\right)$$

where we use $\varepsilon_t'=\frac{\sigma}{2L}\exp\left(-\frac{\sigma(t-1)}{2L}\right)$.

From above result, we know that $\mathbb{E}[F(\boldsymbol{\theta})-F(\boldsymbol{\theta}^*)]\leq\epsilon$ after $T=\mathcal{O}\left(\frac{L}{\sigma}\log\left(\frac{1}{\epsilon}\right)\right)$ rounds of iteration. Therefore the total inner-loop IFO complexity (w.r.t. the quadratic sub-problem) is bounded in expectation by

$$\begin{aligned}\mathcal{O}\left(\sum_{t=1}^{T}\left\{\frac{\sigma^2(\kappa+\log d)}{L^2}t^2+\frac{L\nu^2}{\sigma}\exp\left(\frac{\sigma}{L}t\right)\bigwedge\frac{\sigma n}{L}t\right\}\right)&=\mathcal{O}\left(\frac{\sigma^2(\kappa+\log d)}{L^2}T^3+\frac{L\nu^2}{\sigma}\exp\left(\frac{\sigma}{L}(T+1)\right)\bigwedge\frac{\sigma n}{L}T^2\right)\\&=\mathcal{O}\left(\frac{L(\kappa+\log d)}{\sigma}\log^3\left(\frac{1}{\epsilon}\right)+\frac{L\nu^2}{\sigma\epsilon}\bigwedge\frac{Ln}{\sigma}\log^2\left(\frac{1}{\epsilon}\right)\right).\end{aligned}$$

This proves the desired bound.

Case (2) where there is no regularization $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$ but $F(\boldsymbol{\theta})$ is μ-strongly-convex. For this case, $\Delta_{t-1}(\boldsymbol{\theta})=\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1})^\top\bar{\boldsymbol{H}}(\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1})$ with $\bar{\boldsymbol{H}}=\frac{L}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top$. Moreover, $\nabla^2F(\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^n\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\succeq\frac{\sigma}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top$ and $\nabla^2F(\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^n\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\succeq\mu\boldsymbol{I}$ for all $\boldsymbol{\theta}$. Then by using the same techniques in Step 1, we can obtain the linear convergence. Moreover, the inner problem

$$\boldsymbol{Q}_{t-1}(\boldsymbol{\theta})\triangleq F(\boldsymbol{\theta}_{t-1})+\langle\nabla F(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1}\rangle+\Delta_{t-1}(\boldsymbol{\theta})$$

is also μ-strongly convex. Thus for Step 2, we can directly apply the computational complexity of HSDMPG on strongly convex problems and thus obtain the desired results. The proof is completed. □

C.2 Proof of Corollary 3

Proof. Based on Theorem 3, the results can be easily obtained. Specifically, we plug $\epsilon=\mathcal{O}(\frac{1}{\sqrt{n}})$, $\kappa=\mathcal{O}(\sqrt{n})$ into Theorem 3 and can compute the desired results. □

C.3 Proof of Theorem 4

Proof. For online-sum setting, it contains two cases (1) there is a regularization $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$ and (2) there is no regularization $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$ but $F(\boldsymbol{\theta})$ is μ-strongly-convex. Here we follow the proof of finite-sum settings in Appendix C.1. Specifically, for case (1), $\Delta_{t-1}(\boldsymbol{\theta})=\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1})^\top\bar{\boldsymbol{H}}(\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1})$ with $\bar{\boldsymbol{H}}=L\mathbb{E}_i[\boldsymbol{x}_i\boldsymbol{x}_i^\top]+\mu\boldsymbol{I}$. Moreover, $\nabla^2F(\boldsymbol{\theta})=\mathbb{E}[\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top]+\mu\boldsymbol{I}\succeq\sigma\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top]+\mu\boldsymbol{I}$ for all $\boldsymbol{\theta}$. Then by using the same techniques in Step 1 in Appendix C.1, we can obtain the linear convergence. Moreover, the inner problem $\boldsymbol{Q}_{t-1}(\boldsymbol{\theta})$ is also μ-strongly convex. Thus for Step 2 in Appendix C.1, we can directly apply the computational complexity of HSDMPG on strongly convex problems and thus obtain the desired results.

For case (2), we have $\Delta_{t-1}(\boldsymbol{\theta})=\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1})^\top\bar{\boldsymbol{H}}(\boldsymbol{\theta}-\boldsymbol{\theta}_{t-1})$ with $\bar{\boldsymbol{H}}=L\mathbb{E}_i[\boldsymbol{x}_i\boldsymbol{x}_i^\top]$. Moreover, $\nabla^2F(\boldsymbol{\theta})=\mathbb{E}[\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top]\succeq\sigma\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top]$ and $\nabla^2F(\boldsymbol{\theta})=\mathbb{E}[\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top]\succeq\mu\boldsymbol{I}$ for all $\boldsymbol{\theta}$. Then by using the same techniques in Step 1 in Appendix C.1, we can obtain the linear convergence. Moreover, the inner problem $\boldsymbol{Q}_{t-1}(\boldsymbol{\theta})$ is also μ-strongly convex. Thus for Step 2 in Appendix C.1, we can directly apply the computational complexity of HSDMPG on strongly convex problems and thus obtain the desired results. The proof is completed. □

Appendix D
Proof of Auxiliary Lemmas

D.1 Proof of Lemma 1

Proof. We first consider online setting. Under this setting, we have

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i\in S}z_i\right\|^2=\frac{1}{n^2}\mathbb{E}\left[\sum_{i\in S}\|z_i\|^2+\sum_{i,j\in S,i\neq j}\langle z_i,z_j\rangle\right]\overset{①}{=}\frac{1}{n^2}\mathbb{E}\left[\sum_{i\in S}\|z_i\|^2\right]=\frac{1}{n}\mathbb{E}\left[\|z_i\|^2\right]$$

where ① holds since z_i and z_j are independent.

For finite-sum setting, we can use the same method to prove the result in Lemma 1 or directly use the result in [1]. The proof is completed. □

D.2 Proof of Lemma 2

Proof. Let $\boldsymbol{z}_t^i = \boldsymbol{H}^{-1/2}(\nabla F(\boldsymbol{\theta}_t) - \nabla \ell_i(\boldsymbol{\theta}_t))$. We first consider finite-sum setting. To begin with, we have $\sum_{i=1}^n \boldsymbol{z}_t^i = 0$, $\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{z}_t^i\|^2 \leq \nu^2$ and $\boldsymbol{H}^{-1/2}\boldsymbol{r}_t = \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} \boldsymbol{z}_t^i$. By invoking Lemma 1 we get

$$\mathbb{E}\left[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_t\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t}\boldsymbol{z}_t^i\right\|^2\right] \leq \frac{\nu^2 \mathbb{1}(|\mathcal{S}_t| < n)}{|\mathcal{S}_t|}.$$

Provided that

$$|\mathcal{S}_t| = \frac{16\nu^2(\mu+2\gamma)^2}{\mu^2}\exp\left(\frac{\mu t}{\mu+2\gamma}\right)\bigwedge n,$$

then the following condition always holds

$$\mathbb{E}\left[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_t\|^2\right] \leq \frac{\mu^2}{16(\mu+2\gamma)^2}\exp\left(-\frac{\mu t}{\mu+2\gamma}\right).$$

Next, by using Jensen's Inequality, we can obtain

$$\mathbb{E}\left[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_t\|\right] \leq \sqrt{\mathbb{E}\left[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_t\|^2\right]} = \sqrt{\mathbb{E}\left[\left\|\frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t}\boldsymbol{z}_t^i\right\|^2\right]} \leq \frac{\mu}{4(\mu+2\gamma)}\exp\left(-\frac{\mu t}{2(\mu+2\gamma)}\right).$$

For online setting, we have $\mathbb{E}_i\boldsymbol{z}_t^i = 0$, $\mathbb{E}_i\|\boldsymbol{z}_t^i\|^2 \leq \nu^2$ and $\boldsymbol{H}^{-1/2}\boldsymbol{r}_t = \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t}\boldsymbol{z}_t^i$. By invoking Lemma 1 we get the desired results. The proof is completed. □

D.3 Proof of Lemma 3

Lemma 5. *[2] For both online and finite-sum settings, suppose $\boldsymbol{H}$ and $\boldsymbol{H}_{\mathcal{S}}$ respectively denote the Hessian matrix of $F(\boldsymbol{\theta})$ and $F_{\mathcal{S}}(\boldsymbol{\theta})$ in problem (1). Assume the individual loss in $F(\boldsymbol{\theta})$ is quadratic loss. W.l.o.g., suppose $\|\boldsymbol{x}_i\| \leq r$. Then if $s \geq \frac{28}{3}\log\left(\frac{2d}{\delta}\right)$, with probability at least $1-\delta$, we have*

$$\frac{1}{\frac{3}{2}+\frac{2\gamma}{\mu}}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right) \preceq \boldsymbol{H} \preceq 2\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right),$$

where s is the size of $\mathcal{S}$ and $\gamma = \frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d}{\delta}\right)-\mu\right)^+$.

Proof. From Lemma 5, we know that with probability at least $1-\delta$, we have

$$\frac{1}{\frac{3}{2}+\frac{2\gamma}{\mu}}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right) \preceq \boldsymbol{H} \preceq 2\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right),$$

where s is the size of $\mathcal{S}$ and $\gamma = \frac{1}{2}\left(\frac{28r^2}{3s}\log\left(\frac{2d}{\delta}\right)-\mu\right)^+$. Based on this result, we can easily obtain

$$\begin{aligned}&\frac{1}{\frac{3}{2}+\frac{2\gamma}{\mu}}\boldsymbol{I} \preceq \boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2} \preceq 2\boldsymbol{I},\\ &(1-2\lambda)\boldsymbol{I} \preceq \boldsymbol{I} - \lambda\boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2} \preceq \left(1-\frac{\lambda}{\frac{3}{2}+\frac{2\gamma}{\mu}}\right)\boldsymbol{I},\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\frac{1}{\frac{3}{2}+\frac{2\gamma}{\mu}} \leq \|\boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2}\| \leq 2,\\ &\|\boldsymbol{I} - \lambda\boldsymbol{H}^{1/2}\left(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{1/2}\| \leq \max\left(1-2\lambda, 1-\frac{\lambda}{\frac{3}{2}+\frac{2\gamma}{\mu}}\right),\end{aligned}$$

where $0 \leq \lambda \leq \frac{1}{2}$. The proof is completed. □

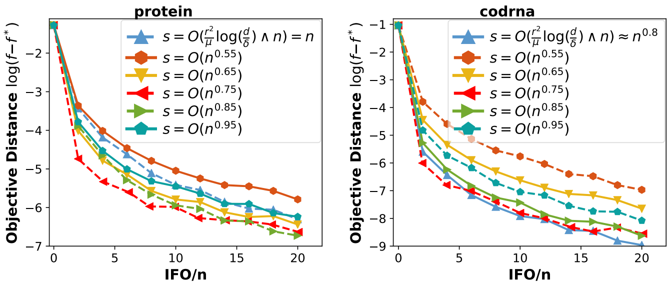

Fig. 7: Investigation of the minibatch size s for constructing function $F_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{s}\sum_{i\in\mathcal{S}} \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) + \frac{\tau\mu}{2}\|\boldsymbol{\theta}\|_2^2$ on the quadratic problems.

D.4 Proof of Lemma 4

Proof. Since both $\boldsymbol{A} + \gamma\boldsymbol{I}$ and $\boldsymbol{B}$ are symmetric and positive definite, it is known that the eigenvalues of $(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{B}$ are positive real numbers and identical to those of $(A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2}$. Let us consider the following eigenvalue decomposition of $(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}\boldsymbol{B}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}$:

$$(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}\boldsymbol{B}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2} = \boldsymbol{Q}^\top \Lambda \boldsymbol{Q},$$

where $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$ and Λ is a diagonal matrix with eigenvalues as diagonal entries. It is then implied that

$$(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{B} = (\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}\boldsymbol{Q}^\top \Lambda \boldsymbol{Q}(\boldsymbol{A} + \gamma\boldsymbol{I})^{1/2},$$

which is a diagonal eigenvalue decomposition of $(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{B}$. Thus $(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{B}$ is diagonalizable.

To prove the eigenvalue bounds of $(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{B}$, it suffices to prove the same bounds for $(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}\boldsymbol{B}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}$. Since $\|\boldsymbol{A} - \boldsymbol{B}\| \leq \gamma$, we have $\boldsymbol{B} \preceq \boldsymbol{A} + \gamma\boldsymbol{I}$ which implies $(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}\boldsymbol{B}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2} \preceq I$ and hence $\mathbb{E}\left[\lambda_{\max}((\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}\boldsymbol{B}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2})\right] \leq 1$. Moreover, since $\boldsymbol{B} \succeq \mu\boldsymbol{I}$, it holds that $\frac{2\gamma}{\mu}\boldsymbol{B} - \gamma\boldsymbol{I} \succeq \gamma\boldsymbol{I} \succeq \mathbb{E}_{\boldsymbol{A}}\boldsymbol{A} - \boldsymbol{B}$. Then we get $(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}\boldsymbol{B}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2} \succeq \frac{\mu}{\mu+2\gamma}\boldsymbol{I}$ which implies $\lambda_{\min}((\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}\boldsymbol{B}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1/2}) \geq \frac{\mu}{\mu+2\gamma}$. Similarly, we can show that $\frac{\mu}{\mu+2\gamma}I \preceq \boldsymbol{B}^{1/2}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{B}^{1/2} \preceq I$, implying $\|\boldsymbol{I} - \boldsymbol{B}^{1/2}(\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{B}^{1/2}\| \leq \frac{2\gamma}{\mu+2\gamma}$. The proof is competed. □

Appendix E
More Experiments

In this section, we investigate the effects of the minibatch s for constructing function $F_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{s}\sum_{i\in\mathcal{S}} \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) + \frac{\tau\mu}{2}\|\boldsymbol{\theta}\|_2^2$. To begin with, we first compare the convergence performance of HSDMPG when using our empirical setting $s = \mathcal{O}\left(n^{0.75}\right)$ and our theory suggested one $s = \mathcal{O}\left(r^2\log(d/\delta)/\mu \wedge n\right)$. In the experiments, we estimate the smallest eigenvalue of the Hessian of $\boldsymbol{F}(\boldsymbol{\theta})$ as μ, and then use it to estimate the corresponding $s = \mathcal{O}\left(r^2\log(d/\delta)/\mu \wedge n\right) = n$ on the protein dataset and $s = \mathcal{O}\left(r^2\log(d/\delta)/\mu \wedge n\right) \approx n^{0.8}$ on cordna. From Figure 7, one can observe that when $s = \mathcal{O}\left(r^2\log(d/\delta)/\mu \wedge n\right)$ is at the order of n on the protein dataset, then using $s = \mathcal{O}\left(n^{0.75}\right)$ can achieve slightly better convergence performance; when $s = \mathcal{O}\left(r^2\log(d/\delta)/\mu \wedge n\right)$ is at the order $n^{0.8}$ and thus smaller than n on the codrna dataset, then using $s = \mathcal{O}\left(n^{0.75}\right)$ provides very similar convergence performance. This can be explained as follows. When μ is very small, e.g. $\mu \approx 10^{-4}$, the estimation $s = \mathcal{O}\left(r^2\log(d/\delta)/\mu \wedge n\right)$ could be as large as the data scale n in practice. In this way, at the beginning, we sample the whole data as $\mathcal{S}$ for constructing the function $F_{\mathcal{S}}(\boldsymbol{\theta})$. But this could be not necessary, since actually we only need partial data to estimate the local Bregman divergence $\mathcal{D}_{\widetilde{F}_{\mathcal{S}}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t-1})$ where $\widetilde{F}_{\mathcal{S}}(\boldsymbol{\theta}) = F_{\mathcal{S}}(\boldsymbol{\theta}) + \frac{\gamma}{2}\|\boldsymbol{\theta}\|_2^2$. Note we do not require very precise Bregman divergence estimation, which can be observed from the regularization term $\frac{\gamma}{2}\|\boldsymbol{\theta}\|_2^2$ that could indeed reduce the precision of Bregman divergence estimation. So we empirically set s around $n^{0.75}$ to approximately and efficiently estimate the local Bregman divergence. Moreover, another reason that we empirically set $s = \mathcal{O}\left(n^{0.75}\right)$ is that it is not easy to estimate the strong convexity parameter μ which actually is larger than the regularization constant λ. This is also a challenge for other stochastic algorithms which derive optimal parameters which however involve some factors that are not easy to be computed or controlled. For instance, the optimal minibatch size B in SCSG should be $B = \left\lceil \frac{c(h^2 \vee L\epsilon)}{\mu\epsilon} \wedge n \right\rceil$ where $c > \frac{9}{2}$ is a constant and $h^2 = \frac{1}{n}\sum_{i=1}^n \|\ell_i'(\boldsymbol{\theta}^*)\|_2^2$. But in their experiments, they select B from the set $\{0.01n, 0.05n, 0.25n\}$. Therefore, using a unified and easily-controlled parameter setting, i.e. $s = \mathcal{O}\left(n^{0.75}\right)$, is a good choice in practice.

Then we further investigate the effects of the minibatch s to the convergence performance of HSDMPG. We respectively set $s = \{n^{0.55}, n^{0.65}, n^{0.75}, n^{0.85}, n^{0.95}\}$ in HSDMPG and report the results in Figure 7. One can observe that when $s \in \{n^{0.75}, n^{0.85}\}$, HSDMPG achieves better performance. This is because as aforementioned, we do not need to use all data to estimate a very precise local local Bregman divergence $\mathcal{D}_{\widetilde{F}_{\mathcal{S}}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t-1})$ of the optimization $\boldsymbol{F}(\boldsymbol{\theta})$.

References

[1] L. Lei and M. Jordan, "Less than a single pass: Stochastically controlled stochastic gradient," in *Artificial Intelligence and Statistics*, 2017, pp. 148–156.

[2] Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, and Laurent Massoulie, "Statistically preconditioned accelerated gradient method for distributed optimization," in *Proc. Int'l Conf. Machine Learning*, 2020.