Hybrid Stochastic-Deterministic Minibatch Proximal Gradient: Less-Than-Single-Pass Optimization with Nearly Optimal Generalization (Supplementary File)

This supplementary document contains the technical proofs of convergence results and some additional numerical results of the paper entitled "Hybrid Stochastic-Deterministic Minibatch Proximal Gradient: Less-Than-Single-Pass Optimization with Nearly Optimal 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. 3.1, including Theorem 1 which analyzes convergence rate of HSDMPG and Corollaries 1 and 2 which analyze the IFO complexity of HSDMPG on the quadratic problems. Next, Appendix C provides the proofs of the results in Sec. 3.2, including Theorem 2 which proves the convergence rate of HSDMPG and analyzes its IFO complexity for generic problems, and Corollary 3 which gives the IFO complexity of HSDMPG to achieve the intrinsic excess error bound. Then in Appendix D we present the proofs of auxiliary lemmas in Appendix A, including Lemmas 1 $\sim$ 3. Finally, more details of the testing datasets used in the manuscript are presented in Appendix D.4.

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. *Assume that the loss $F(\boldsymbol{\theta})$ is a μ-strongly convex loss, $\sup_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{H}^{-1/2}(\nabla F(\boldsymbol{\theta}) - \nabla \ell_i(\boldsymbol{\theta}))\|_2^2 \leq \nu^2$. 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.1.

Lemma 2. *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 $\|\boldsymbol{x}_i\| \leq r$ $(i = 1, \cdots, n)$ and $\ell(\boldsymbol{\theta}^\top \boldsymbol{x}, \boldsymbol{y})$ is L-smooth w.r.t. $\boldsymbol{\theta}^\top \boldsymbol{x}$. Then we have*

$$\mathbb{E}_{\mathcal{S}}\left[\|\boldsymbol{H}_{\mathcal{S}} - \boldsymbol{H}\|^2\right] \leq \frac{(\sqrt{\log(d)}+\sqrt{2})^2 L^2 r^4}{s} \quad \text{and} \quad \mathbb{E}_{\mathcal{S}}\left[\|\boldsymbol{H}_{\mathcal{S}} - \boldsymbol{H}\|\right] \leq \frac{(\sqrt{\log(d)}+\sqrt{2}) L r^2}{\sqrt{s}},$$

where s is the size of $\mathcal{S}$.

see its proof in Appendix D.2

Lemma 3. *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.3.

B. Proofs for the Results in Section 3.1

We collect in this appendix section the technical proofs of the results in Section 3.1 of the main paper.

B.1. Proof of Theorem 1

Proof. This proof has four steps. To begin with, for brevity, let $\boldsymbol{u}_t = \boldsymbol{H}^{1/2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)$. 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.

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

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_{t-1} - \boldsymbol{H}^{-1}\nabla F(\boldsymbol{\theta}_{t-1}). \tag{6}$$

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

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

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

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

where $\boldsymbol{r}_{t-1} = \nabla F(\boldsymbol{\theta}_{t-1}) - \boldsymbol{g}_{t-1}$. Next plugging Eqn. (6) into the above equation, it establishes

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

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

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

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

$$\boldsymbol{u}_t = (\boldsymbol{I} - \boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2})\boldsymbol{u}_t + \boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\nabla P_{t-1}(\boldsymbol{\theta}_t) + \boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{r}_{t-1}. \tag{7}$$

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

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{u}_t\|] \leq& \mathbb{E}\left[\|\boldsymbol{I} - \boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\|\|\boldsymbol{u}_{t-1}\| + \|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\|\|\boldsymbol{H}^{-1/2}\nabla P_{t-1}(\boldsymbol{\theta}_t)\|\right]\\ &+ \mathbb{E}\left[\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\|\mathbb{E}[\|\boldsymbol{H}^{-1/2}\boldsymbol{r}_{t-1}\|].\right]\end{aligned} \tag{8}$$

From Lemma 1, we know that by setting $|\mathcal{S}_t| = \frac{16\nu^2(\mu+2\gamma)^2}{\mu^2}\exp\left(\frac{\mu t}{\mu+2\gamma}\right) \bigwedge n$, then the inequality always holds

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

Suppose $\|\boldsymbol{x}_i\| \leq r$ $(i = 1, \cdots, n)$ and $\ell(\boldsymbol{\theta}^\top \boldsymbol{x}, \boldsymbol{y})$ is L-smooth w.r.t. $\boldsymbol{\theta}^\top \boldsymbol{x}$. Then by using Lemma 2 we have

$$\mathbb{E}\left[\|\boldsymbol{H}_{\mathcal{S}} - \boldsymbol{H}\|\right] \leq \gamma = \frac{(\sqrt{\log(d)} + \sqrt{2})Lr^2}{\sqrt{s}},$$

where s is the size of $\mathcal{S}$. In this way, by using Lemma 3, we can further establish

$$\frac{\mu}{\mu+2\gamma} \leq \left\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma \boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right\| \leq 1 \quad \text{and} \quad \left\|\boldsymbol{I} - \boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}} + \gamma \boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right\| \leq \frac{2\gamma}{\mu+2\gamma}. \tag{9}$$

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 we plug the above results into Eqn. (8) and establish

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

where in the inequality ① we have used $\boldsymbol{H} \succeq \mu \boldsymbol{I}$, ② follows from the condition $\varepsilon_t \leq \frac{\mu^{1.5}}{4(\mu+2\gamma)}\exp\left(-\frac{\mu(t-1)}{2(\mu+2\gamma)}\right)$.

By taking expectation with respect to $\boldsymbol{\theta}_{t-1}$ we arrive at

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

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

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

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

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

Step 3. Upper bound $\mathbb{E}[\|\boldsymbol{u}_t\|^2]$.

From Eqn. (7), we can upper bound $\mathbb{E}[\|\boldsymbol{u}_t\|^2]$ as

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

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

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

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

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

From Lemma 1, we know that by setting $|\mathcal{S}_t|=\frac{16\nu^2(\mu+2\gamma)^2}{\mu^2}\exp\left(\frac{\mu t}{\mu+2\gamma}\right)\bigwedge n$, then the inequality always holds

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

Suppose $\|\boldsymbol{x}_i\|\leq r$ $(i=1,\cdots,n)$ and $\ell(\boldsymbol{\theta}^\top\boldsymbol{x},\boldsymbol{y})$ is L-smooth w.r.t. $\boldsymbol{\theta}^\top\boldsymbol{x}$. Then by using Lemma 2 we have

$$\mathbb{E}\left[\|\boldsymbol{H}_{\mathcal{S}}-\boldsymbol{H}\|^2\right]\leq\gamma^2=\frac{(\sqrt{\log(d)}+\sqrt{2})^2L^2r^4}{s},$$

where s is the size of $\mathcal{S}$. In this way, by using Lemma 3, we can further establish

$$\frac{\mu^2}{(\mu+2\gamma)^2}\leq\left\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right\|^2\leq1\ \text{ and }\ \left\|\boldsymbol{I}-\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right\|^2\leq\frac{4\gamma^2}{(\mu+2\gamma)^2}.$$

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 we plug the above results and Eqn. (9) into Eqn. (10) and establish

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{u}_t\|^2]\leq&\frac{4\gamma^2}{(\mu+2\gamma)^2}\mathbb{E}[\|\boldsymbol{u}_{t-1}\|^2]+\frac{\varepsilon_t^2}{\mu}+\frac{\mu^2}{16(\mu+2\gamma)^2}\exp\left(-\frac{\mu t}{\mu+2\gamma}\right)+\frac{8\gamma}{\mu+2\gamma}\frac{\varepsilon_t}{\sqrt{\mu}}\mathbb{E}\left[\|\boldsymbol{u}_{t-1}\|\right]\\ &+\frac{\varepsilon_t}{\sqrt{\mu}}\frac{\mu}{2(\mu+2\gamma)}\exp\left(-\frac{\mu t}{2(\mu+2\gamma)}\right).\end{aligned}$$

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

$$
\begin{aligned}
&\mathbb{E}[\|\boldsymbol{u}_t\|^2]\\
\leq&\frac{4\gamma^2}{(\mu+2\gamma)^2}\mathbb{E}[\|\boldsymbol{u}_{t-1}\|^2]+\frac{\mu^2}{8(\mu+2\gamma)^2}\left(\frac{1}{2}\left(1+\exp\left(\frac{\mu}{\mu+2\gamma}\right)\right)+\exp\left(\frac{\mu}{2(\mu+2\gamma)}\right)\right)\exp\left(-\frac{\mu t}{\mu+2\gamma}\right)\\
&+\frac{2\mu\gamma b}{(\mu+2\gamma)^2}\exp\left(\frac{\mu}{2(\mu+2\gamma)}\right)\exp\left(-\frac{\mu t}{\mu+2\gamma}\right)\\
\overset{\textcircled{1}}{\leq}&\frac{4\gamma^2}{(\mu+2\gamma)^2}\mathbb{E}[\|\boldsymbol{u}_{t-1}\|^2]+2a^2\exp\left(-2at\right)+\frac{4b\gamma a^2}{\mu}\exp\left(-2at\right)\\
=&\frac{4\gamma^2}{(\mu+2\gamma)^2}\mathbb{E}[\|\boldsymbol{u}_{t-1}\|^2]+2a^2\left(1+\frac{2b\gamma}{\mu}\right)\exp\left(-2at\right),
\end{aligned}
$$

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

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

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 we obtain the desired result:

$$
\begin{aligned}
&\mathbb{E}[F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}^*)] = \frac{1}{2}\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}}^2]\\
\leq&\frac{1}{2}\left(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{\boldsymbol{H}}^2 + \frac{\mu^2}{2(\mu+2\gamma)^2}\left(1+\frac{2\gamma}{\mu}\left(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{\boldsymbol{H}} + e\right)\right)\right)\exp\left(-\frac{\mu t}{\mu+2\gamma}\right)\\
\overset{\textcircled{1}}{\leq}&\frac{1}{2}\left(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{\boldsymbol{H}}^2 + \frac{1}{4}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{\boldsymbol{H}} + \frac{3}{2}\right)\exp\left(-\frac{\mu t}{\mu+2\gamma}\right)\\
=&\left(\frac{1}{2}\left(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{\boldsymbol{H}} + \frac{1}{2}\right)^2 + \frac{5}{8}\right)\exp\left(-\frac{\mu t}{\mu+2\gamma}\right),
\end{aligned}
$$

where ① uses $\frac{\mu^2}{2(\mu+2\gamma)^2} \leq \frac{1}{2}$ and $\frac{\mu\gamma}{(\mu+2\gamma)^2} \leq \frac{1}{4}$. The proof is completed. □

B.2. Proof of Corollary 1

Proof. This proof has 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 $P_t(\boldsymbol{\theta}) = F_{\mathcal{S}}(\boldsymbol{\theta}) + \langle \nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1}) - \nabla F_{\mathcal{S}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}\rangle + \frac{\gamma}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}\|_2^2$ 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 $\|P_t(\boldsymbol{\theta})\| \leq \varepsilon_t$. 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.

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

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

where $\zeta = \frac{1}{2}\left(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{\boldsymbol{H}} + \frac{1}{2}\right)^2 + \frac{5}{8}$ 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

$$T = \frac{\mu + 2\gamma}{\mu} \log\left(\frac{\zeta}{\epsilon}\right).$$

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

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

where in ① we have used the definition of T such that $\exp\left(\frac{\mu T}{\mu+2\gamma}\right) = \frac{\zeta}{\epsilon}$ and the fact $\exp(a) \geq 1 + a, \forall a > 0$. At the same time, we also have

$$\sum_{t=0}^{T-1} |\mathcal{S}_t| \leq nT = \frac{(\mu+2\gamma)n}{\mu} \log\left(\frac{\zeta}{\epsilon}\right).$$

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

$$\frac{16\zeta\nu^2(\mu+2\gamma)^3}{\mu^3\epsilon} \bigwedge \frac{(\mu+2\gamma)n}{\mu} \log\left(\frac{\zeta}{\epsilon}\right) = \mathcal{O}\left(\left(1 + \frac{\kappa^3 \log^{1.5}(d)}{s^{1.5}}\right) \frac{\nu^2}{\epsilon} \bigwedge \left(1 + \frac{\kappa \log^{0.5}(d)}{s^{0.5}}\right) n \log\left(\frac{1}{\epsilon}\right)\right),$$

where we use $\gamma = \frac{(\sqrt{\log(d)} + \sqrt{2})Lr^2}{\sqrt{s}}$ and $\kappa = \frac{L}{\mu}$.

Step 3. Estimate computational complexity of the inner optimization via SVRG.
At each iteration time stamp t, we need to optimize the inner problem $P_t(\boldsymbol{\theta}) = F_{\mathcal{S}}(\boldsymbol{\theta}) + \langle \nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1}) - \nabla F_{\mathcal{S}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta} \rangle + \frac{\gamma}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}\|_2^2$. In $P_t(\boldsymbol{\theta})$, its finites-sum structure comes from $F_{\mathcal{S}}(\boldsymbol{\theta})$ and its gradient.

For $(\mu+\gamma)$-strongly-convex and $(L+\gamma)$-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 $(\mu+\gamma)$-strongly-convex, we have $\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\|_2 \leq 2(\mu+\gamma)(P_{t-1}(\boldsymbol{\theta}_T) - P_{t-1}(\boldsymbol{\theta}^*))$. In this way, to achieve $\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\|_2 \leq \varepsilon_t = \frac{\mu^{1.5}}{4(\mu+2\gamma)} \exp\left(-\frac{\mu(t-1)}{2(\mu+2\gamma)}\right)$, the expected IFO complexity of SVRG is

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

From above result we know that $\mathbb{E}[F(w^{(t)})] \leq F(w^*) + \epsilon$ 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 $\gamma = \frac{(\sqrt{\log(d)} + \sqrt{2})Lr^2}{\sqrt{s}}$ into the above inner-loop IFO bound to obtain

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

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

$$\begin{aligned}&\mathcal{O}\left(\frac{L\sqrt{s\log(d)}}{\mu}\left(\log\left(\frac{L^{1.5}}{\mu^{1.5}}\sqrt{\frac{\log(d)}{s}}\right)\log\left(\frac{1}{\epsilon}\right)+\log^2\left(\frac{1}{\epsilon}\right)\right)+\left(1+\frac{\kappa^3\log^{1.5}(d)}{s^{1.5}}\right)\frac{\nu^2}{\epsilon}\bigwedge\left(1+\frac{\kappa\log^{0.5}(d)}{s^{0.5}}\right)n\log\left(\frac{1}{\epsilon}\right)\right)\\ =&\mathcal{O}\left(\kappa\sqrt{s\log(d)}\log^2\left(\frac{1}{\epsilon}\right)+\left(1+\frac{\kappa^3\log^{1.5}(d)}{s^{1.5}}\right)\frac{\nu^2}{\epsilon}\bigwedge\left(1+\frac{\kappa\log^{0.5}(d)}{s^{0.5}}\right)n\log\left(\frac{1}{\epsilon}\right)\right),\end{aligned}$$

where $\kappa=\frac{L}{\mu}$.

This competes the proof. □

B.3. Proof of Corollary 2

Proof. The result in Corollary 2 can be easily obtained. Specifically, we plug $\epsilon=\mathcal{O}(\frac{1}{\sqrt{n}})$, $\kappa=\mathcal{O}(\sqrt{n})$ and $s=\mathcal{O}(\frac{\nu n^{0.75}\log^{0.5}(d)}{\log(n)})$ into Corollary 1 and can compute the desired results. □

C. Proofs for the Results in Section 3.2

C.1. Proof of Theorem 2

Proof. This proof has two steps. In the first step, we prove the results in the first part of Theorem 2, 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 see the following detailed steps.

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 $\nabla^2F(\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^n\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top+\mu\boldsymbol{I}\succeq\frac{\sigma}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top+\mu\boldsymbol{I}$ for all $\boldsymbol{\theta}$. In this way, we can lower bound

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

where $\textcircled{1}$ 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

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

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

$$
\begin{aligned}
&\mathcal{O}\left(\kappa\sqrt{s\log(d)}\log^2\left(\frac{1}{\varepsilon'_t}\right)+\left(1+\frac{\kappa^3\log^{1.5}(d)}{s^{1.5}}\right)\frac{\nu^2}{\varepsilon'_t}\bigwedge\left(1+\frac{\kappa\log^{0.5}(d)}{s^{0.5}}\right)n\log\left(\frac{1}{\varepsilon'_t}\right)\right)\\
=&\mathcal{O}\left(\frac{\sigma^2\sqrt{s\log(d)}}{L\mu}t^2+\left(1+\frac{\kappa^3\log^{1.5}(d)}{s^{1.5}}\right)\frac{L\nu^2}{\sigma}\exp\left(\frac{\sigma}{L}t\right)\bigwedge\left(1+\frac{\kappa\log^{0.5}(d)}{s^{0.5}}\right)\frac{Ln}{\sigma}t\right)
\end{aligned}
$$

where $\kappa=\frac{L}{\mu}$ denotes the conditional number and $\varepsilon'_t=\frac{\sigma}{2L}\exp\left(-\frac{\sigma(t-1)}{2L}\right)$.

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

$$
\begin{aligned}
&\mathcal{O}\left(\sum_{t=1}^{T}\left\{\frac{\sigma^2\sqrt{s\log(d)}}{L\mu}t^2+\left(1+\frac{\kappa^3\log^{1.5}(d)}{s^{1.5}}\right)\frac{L\nu^2}{\sigma}\exp\left(\frac{\sigma}{L}t\right)\bigwedge\left(1+\frac{\kappa\log^{0.5}(d)}{s^{0.5}}\right)\frac{Ln}{\sigma}t\right\}\right)\\
=&\mathcal{O}\left(\frac{\sigma^2\sqrt{s\log(d)}}{L\mu}T^3+\left(1+\frac{\kappa^3\log^{1.5}(d)}{s^{1.5}}\right)\frac{L\nu^2}{\sigma}\exp\left(\frac{\sigma}{L}(T+1)\right)\bigwedge\left(1+\frac{\kappa\log^{0.5}(d)}{s^{0.5}}\right)\frac{Ln}{\sigma}T^2\right)\\
=&\mathcal{O}\left(\frac{L^2\sqrt{s\log(d)}}{\sigma\mu}\log^3\left(\frac{1}{\epsilon}\right)+\left(1+\frac{\kappa^3\log^{1.5}(d)}{s^{1.5}}\right)\frac{L\nu^2}{\sigma\epsilon}\bigwedge\left(1+\frac{\kappa\log^{0.5}(d)}{s^{0.5}}\right)\frac{L^3n}{\sigma^3}\log^2\left(\frac{1}{\epsilon}\right)\right).
\end{aligned}
$$

This proves the desired bound. □

C.2. Proof of Corollary 3

Proof. Based on Theorem 2, the results can be easily obtained. Specifically, we plug $\epsilon=\mathcal{O}(\frac{1}{\sqrt{n}})$, $\kappa=\mathcal{O}(\sqrt{n})$ and $s=\mathcal{O}\big(\frac{\nu n^{0.75}\log^{0.5}(d)}{\log(n)}\big)$ into Theorem 2 and can compute the desired results. □

D. Proof of Auxiliary Lemmas

D.1. Proof of Lemma 1

The following lemma from (Lei & Jordan, 2017) will be used to bound the gradient estimation variance.

Lemma 4. *(Lei & Jordan, 2017) 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*

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

Proof of Lemma 1. Let $\boldsymbol{z}_t^i = \boldsymbol{H}^{-1/2}(\nabla F(\boldsymbol{\theta}_t) - \nabla \ell_i(\boldsymbol{\theta}))$. Then we have $\sum_{i=1}^n \boldsymbol{z}_t^i = 0$, $\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{z}_t^i\|^2 \le \nu^2$ 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 4 we get

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

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

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

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

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

The proof is completed. □

D.2. Proof of Lemma 2

Lemma 5. *(Oliveira, 2010) Suppose* $\{\boldsymbol{A}_i\}_{i=1}^n$ *are deterministic Hermitian matrices and* $\{\varepsilon_i\}_{i=1}^n$ *are independent Bernoulli variables taking values* ± 1 *with probability* $\frac{1}{2}$*. Let* $\boldsymbol{Z} = \sum_{i=1}^n \varepsilon_i \boldsymbol{A}_i$*. Then we have*

$$\mathbb{E}_\varepsilon\left[\|\boldsymbol{Z}\|^2\right] \le (\sqrt{\log(d)}+\sqrt{2})^2\left\|\sum_{i=1}^n \boldsymbol{A}_i^2\right\|.$$

Proof. To begin with, we can compute the Hessian matrix $\boldsymbol{H} = \frac{1}{n}\sum_{i=1}^n \ell''(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \mu\boldsymbol{I}$. In this way, we can formulate

$$\|\boldsymbol{H}_{\mathcal{S}} - \boldsymbol{H}\| = \left\|\frac{1}{s}\sum_{i\in\mathcal{S}} \ell''(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{n}\sum_{i=1}^n \ell''(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|.$$

Assume $\boldsymbol{x}_i$ are drawn from $\mathcal{S}$ and $\bar{\boldsymbol{x}}_i$ are drawn from $\mathcal{S}'$ where $\mathcal{S}'$ is also uniformly sampled from the n samples. In this way, we can establish

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\left[\left\|\frac{1}{s}\sum_{i\in\mathcal{S}}\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top-\frac{1}{n}\sum_{i=1}^{n}\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|^2\right]\\
=&\mathbb{E}_{\mathcal{S}}\left[\left\|\frac{1}{s}\sum_{i=0}^{s}\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top-\mathbb{E}_{\mathcal{S}'}\frac{1}{s}\sum_{i=0}^{s}\ell''(\boldsymbol{\theta}^\top\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{y}}_i)\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top\right\|^2\right]\\
\overset{\text{①}}{\leq}&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{S}'}\left[\left\|\frac{1}{s}\sum_{i=0}^{s}\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top-\frac{1}{s}\sum_{i=0}^{s}\ell''(\boldsymbol{\theta}^\top\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{y}}_i)\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top\right\|^2\right]\\
\overset{\text{②}}{=}&\mathbb{E}_{\varepsilon}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{S}'}\left[\left\|\frac{1}{s}\sum_{i=1}^{s}\varepsilon_i\left(\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top-\ell''(\boldsymbol{\theta}^\top\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{y}}_i)\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top\right)\right\|^2\right]\\
\leq&4\mathbb{E}_{\varepsilon}\mathbb{E}_{\mathcal{S}}\left[\left\|\frac{1}{s}\sum_{i=1}^{s}\varepsilon_i\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|^2\right]
\end{aligned}
$$

where ① uses the Jensen's Inequality; in ② the variable ε has two values ± 1 with probability $\frac{1}{2}$. From Lemma 5, we have

$$
\mathbb{E}_{\varepsilon}\left[\left\|\sum_{i=1}^{s}\varepsilon_i\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|^2\right]\leq L^2\mathbb{E}_{\varepsilon}\left[\left\|\sum_{i=1}^{s}\varepsilon_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|^2\right]\leq(\sqrt{\log(d)}+\sqrt{2})^2L^2\left\|\sum_{i=1}^{s}(\boldsymbol{x}_i\boldsymbol{x}_i^\top)^2\right\|.
$$

W.l.o.g., suppose $\|\boldsymbol{x}_i\|\leq r$. Then we can obtain

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\left\|\frac{1}{s}\sum_{i\in\mathcal{S}}\ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top-\frac{1}{n}\sum_{i=1}^{n}\ell''(\boldsymbol{\theta}^\top\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{y}}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|^2\right]\leq&\frac{(\sqrt{\log(d)}+\sqrt{2})^2L^2}{s}\mathbb{E}_{\mathcal{S}}\left\|\frac{1}{s}\sum_{i=1}^{s}(\boldsymbol{x}_i\boldsymbol{x}_i^\top)^2\right\|\\
\leq&\frac{(\sqrt{\log(d)}+\sqrt{2})^2r^4L^2}{s}.
\end{aligned}
$$

Therefore, we can further obtain

$$
\mathbb{E}_{\mathcal{S}}\left[\|\boldsymbol{H}_{\mathcal{S}}-\boldsymbol{H}\|^2\right]\leq\frac{(\sqrt{\log(d)}+\sqrt{2})^2L^2r^4}{s}.
$$

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

$$
\mathbb{E}\left[\|\boldsymbol{H}_{\mathcal{S}}-\boldsymbol{H}\|\right]\leq\sqrt{\mathbb{E}\left[\|\boldsymbol{H}_{\mathcal{S}}-\boldsymbol{H}\|^2\right]}\leq\frac{(\sqrt{\log(d)}+\sqrt{2})Lr^2}{\sqrt{s}}.
$$

The proof is completed. □

D.3. Proof of Lemma 3

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. □

D.4. Descriptions of Testing Datasets

We first briefly introduce the ten testing datasets in the manuscript including including ijcnn, a09, w8a, covtype, protein, codrna, satimage, sensorless, letter, rcv1. All these datasets are provided in the LibSVM website[1]. Their detailed information is summarized in Table 1. 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 1: Descriptions of the ten testing datasets.

	#class	#sample	#feature		#class	#sample	#feature
ijcnn1	2	49,990	22	codrna	2	59,535	8
a09	2	32,561	123	satimage	6	4,435	36
w8a	2	49,749	300	sensorless	11	58,509	48
covtype	2	581,012	54	rcv1	2	20,242	47,236
protein	3	14,895	357	letter	26	10,500	16

References

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

Oliveira, R. Sums of random hermitian matrices and an inequality by rudelson. *Electronic Communications in Probability*, 15:203–212, 2010.

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/