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

Pan Zhou [1] **Xiao-Tong Yuan** [2]

Abstract

Stochastic variance-reduced gradient (SVRG) algorithms have been shown to work favorably in solving large-scale learning problems. Despite the remarkable success, the stochastic gradient complexity of SVRG-type algorithms usually scales linearly with data size and thus could still be expensive for huge data. To address this deficiency, we propose a hybrid stochastic-deterministic minibatch proximal gradient (HSDMPG) algorithm for strongly-convex problems that enjoys provably improved data-size-independent complexity guarantees. More precisely, for quadratic loss $F(\boldsymbol{\theta})$ of n components, we prove that HSDMPG can attain an ϵ-optimization-error $\mathbb{E}[F(\boldsymbol{\theta})-F(\boldsymbol{\theta}^*)] \leq \epsilon$ within $\mathcal{O}\Big(\frac{\kappa^{1.5}\epsilon^{0.75}\log^{1.5}(\frac{1}{\epsilon})+1}{\epsilon} \wedge \Big(\kappa\sqrt{n}\log^{1.5}\big(\frac{1}{\epsilon}\big)+ n\log\big(\frac{1}{\epsilon}\big)\Big)\Big)$ stochastic gradient evaluations, where κ is condition number. For generic strongly convex loss functions, we prove a nearly identical complexity bound though at the cost of slightly increased logarithmic factors. For large-scale learning problems, our complexity bounds are superior to those of the prior state-of-the-art SVRG algorithms with or without dependence on data size. Particularly, in the case of $\epsilon = \mathcal{O}\big(1/\sqrt{n}\big)$ which is at the order of intrinsic excess error bound of a learning model and thus sufficient for generalization, the stochastic gradient complexity bounds of HSDMPG for quadratic and generic loss functions are respectively $\mathcal{O}(n^{0.875}\log^{1.5}(n))$ and $\mathcal{O}(n^{0.875}\log^{2.25}(n))$, which to our best knowledge, for the first time achieve optimal generalization in less than a single pass over data. Extensive numerical results demonstrate the computational advantages of our algorithm over the prior ones.

[1]Salesforce Research [2] B-DAT Lab and CICAEET, Nanjing University of Information Science & Technology, Nanjing, 210044, China. Correspondence to: Xiao-Tong Yuan <xtyuan@nuist.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

1. Introduction

We consider the following ℓ_2-regularized empirical risk minimization (ERM) problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} F(\boldsymbol{\theta}) := \frac{1}{n}\sum\nolimits_{i=1}^{n} \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) + \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2, \quad (1)$$

where $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ is a training set; the convex loss function $\ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i)$ measures the discrepancy between the linear prediction $\boldsymbol{\theta}^\top \boldsymbol{x}_i$ and the ground truth $\boldsymbol{y}_i$; and the regularization term $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$ aims at enhancing generalization ability of the linear model. In the field of statistical learning, the formulation (1) encapsulates a vast body of problems including least squares regression, logistic regression and softmax regression, to name a few. In this work, we focus on developing scalable and autonomous first-order optimization methods to solve this fundamental problem, which has been extensively studied with a bunch of efficient algorithms proposed including gradient descent (GD) (Cauchy, 1847), stochastic GD (SGD) (Robbins & Monro, 1951), SDCA (Shalev-Shwartz, 2012), SVRG (Johnson & Zhang, 2013), Catalyst (Lin et al., 2015), SCSG (Lei & Jordan, 2017) and Katyusha (Allen-Zhu, 2017).

Motivation. Despite the remarkable success of the stochastic gradient methods and their variance-reduced extensions, the stochastic gradient evaluation complexity (which usually dominates the computational cost) of these algorithms tends to scale linearly with data size n. Such a linear dependence is not only expensive when data scale is huge but also problematic in online and life-long learning regimes where samples are coming infinitely. As pointed out in (Lei & Jordan, 2017), there are situations in which accurate solutions can be obtained with less than a single pass through the data, *e.g.* for a large-scale dataset with similar and redundant samples. Therefore, developing data-size-independent learning algorithms is of special importance in big data era.

Particularly, we are interested in efficiently optimizing problem (1) to its intrinsic excess error bound which typically scales as $\mathcal{O}(1/\sqrt{n})$. As shown in (Bottou & Bousquet, 2008), the excess error, which measures the expected prediction discrepancy between the optimum model and the learnt model over all possible samples and thus reflects the generalization performance of the model, can be decomposed into model approximation error, estimation error and

Table 1: Comparison of IFO complexity for first-order stochastic algorithms on the μ-strongly-convex problem (1) with condition number κ. The solution $\boldsymbol{\theta}$ with ϵ-optimization-error is measured by sub-optimality $\mathbb{E}\left[F(\boldsymbol{\theta})-F(\boldsymbol{\theta}^*)\right]\leq\epsilon$ with optimum $F(\boldsymbol{\theta}^*)$. Here we define a set of constants for quadratic (generic) loss: $\beta_1=1.5$ (2.25), $\beta_2=1$ (2), $\beta_3=3$ (4.5), $\beta_4=1$ (2.5), $\beta_5=1$ (1.5), $\gamma=1.5$ (2.25). These different constants only affects the logarithm factor $\xi=\log\left(\frac{1}{\epsilon}\right)$. For brevity, we define $\Theta=\frac{\kappa^{1.5}\xi^{\gamma}}{\epsilon^{0.25}}+\frac{1}{\epsilon}$. The third column summarizes the conditions under which HSDMPG has lower IFO complexity than the compared algorithms.

	ϵ-Optimization Error for ERM (1)		$\frac{1}{\sqrt{n}}$-Optimization Error for ERM (1)
	IFO Complexity	Better Zoom of HSDMPG	
SGD	$\mathcal{O}\left(\frac{1}{\mu\epsilon}\right)$	① $\mu\leq 1 \& \mu\kappa^{1.5}\epsilon^{0.75}\xi^{\beta_1}\leq\mathcal{O}(1)$ or ② $\mathcal{O}(n)\leq\frac{1}{\mu\epsilon\xi^{\beta_2}}\wedge\frac{1}{\kappa^2\mu^2\epsilon^2\xi^{\beta_3}}$	$\mathcal{O}(n)$
SVRG, SAGA, APSDCA	$\mathcal{O}\left((n+\kappa)\log\left(\frac{1}{\epsilon}\right)\right)$	① $\Theta\xi^{-1}\leq\mathcal{O}(n)$	$\mathcal{O}(n\log(n))$
APCG	$\mathcal{O}\left(\frac{n}{\sqrt{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$	① $\Theta\mu^{0.5}\xi^{-1}\leq\mathcal{O}(n)$ or ② $\mu^{-1}\kappa^2\xi^{\beta_4}\leq\mathcal{O}(n)$	$\mathcal{O}\left(n^{1.25}\log(n)\right)$
SPDC, Catalyst, Katyusha	$\mathcal{O}\left((n+\sqrt{n\kappa})\log\left(\frac{1}{\epsilon}\right)\right)$	① $\Theta\xi^{-1}\wedge\Theta^2\xi^{-2}\kappa^{-1}\leq\mathcal{O}(n)$	$\mathcal{O}(n\log(n))$
AMSVRG	$\mathcal{O}\left(\left(n+\frac{n\kappa}{n+\sqrt{\kappa}}\right)\log\left(\frac{1}{\epsilon}\right)\right)$	① $\Theta\xi^{-1}\leq\mathcal{O}(n)$	$\mathcal{O}(n\log(n))$
Varag	$\mathcal{O}\left(n\log\left(n\wedge\frac{1}{\epsilon}\right)+\sqrt{n}\left(\frac{1}{\epsilon^{0.5}}\wedge\kappa^{0.5}\log\left(\frac{1}{\epsilon\kappa}\right)\right)\right)$	① $\Theta\log^{-1}(n\wedge\frac{1}{\epsilon})\leq\mathcal{O}(n)$ or ② $\Theta^2(\epsilon\vee\frac{1}{\kappa\log^2(\frac{1}{\epsilon\kappa})})\leq\mathcal{O}(n)$	$\mathcal{O}(n\log(n))$
SCSG	$\mathcal{O}\left((n\wedge\frac{\kappa}{\epsilon}+\kappa)\log\left(\frac{1}{\epsilon}\right)\right)$	① $\Theta\xi^{-1}\leq\mathcal{O}(n)\leq\frac{\kappa}{\epsilon}$ or ② $\kappa\epsilon^{1.5}\xi^{\beta_5}\leq\mathcal{O}(1)\&\frac{\kappa}{\epsilon}\leq\mathcal{O}(n)$	$\mathcal{O}(n\log(n))$
HSDMPG quadratic	$\mathcal{O}\left(\frac{\kappa^{1.5}\epsilon^{0.75}\log^{1.5}(\frac{1}{\epsilon})+1}{\epsilon}\wedge\left(\kappa\sqrt{n}\log^{1.5}\left(\frac{1}{\epsilon}\right)+n\log\left(\frac{1}{\epsilon}\right)\right)\right)$	———	$\mathcal{O}\left(n^{0.875}\log^{1.5}(n)\right)$
HSDMPG generic	$\mathcal{O}\left(\frac{\kappa^{1.5}\epsilon^{0.75}\log^{2.25}(\frac{1}{\epsilon})+1}{\epsilon}\wedge\left(\kappa\sqrt{n}\log^{2.5}\left(\frac{1}{\epsilon}\right)+n\log^2\left(\frac{1}{\epsilon}\right)\right)\right)$	———	$\mathcal{O}\left(n^{0.875}\log^{2.25}(n)\right)$

optimization error. Among them, the model approximation error measures how closely the selected predication model can approximate the optimal model; the estimation error measures the prediction effects of minimizing the empirical risk instead of the population risk; the optimization error denotes the prediction difference between the exact and approximate solutions of ERM. Therefore, to achieve small excess error, one should minimize the three terms jointly. With optimal choice $\mu=\mathcal{O}(1/\sqrt{n})$ to balance empirical risk and generalization gap, the estimation error is known to be at the order of $\mathcal{O}(1/\sqrt{n})$, which implies the excess error is dominated by $\mathcal{O}(1/\sqrt{n})$ (Vapnik, 2006; Shalev-Shwartz et al., 2009; Shalev-Shwartz & Ben-David, 2014). Thus, it is sufficient to optimize the regularized ERM problem (1) to the optimization error $\mathcal{O}(1/\sqrt{n})$ to match the optimal excess error without redundant computation.

Overiew of our contribution. The main contribution of this paper is a novel Hybrid Stochastic-Deterministic Minibatch Proximal Gradient (HSDMPG) algorithm with substantially improved data-size-independent complexity over existing methods. For quadratic problems, the core idea of our method is to recurrently convert the original large-scale ERM problem into a series of minibatch proximal ERM subproblems for efficient minimization and update. Specifically, as a starting point, we uniformly randomly select a minibach $\mathcal{S}$ of components of the risk function F to form a stochastic approximation $F_{\mathcal{S}}$ that will be fixed throughout the algorithm iteration. Next, at each iteration step, we first construct a stochastic surrogate of F by combining the Bregman divergence of $F_{\mathcal{S}}$ at the current iterate and a first-order hybrid stochastic-deterministic approximation of F; and then we invoke existing variance-reduced algorithms, such as SVRG, to minimize this surrogate subproblem to desired optimization error. For quadratic loss, we can provably establish sharper bounds of incremental first order oracle (IFO, see Definition 2) for such a hybrid stochastic-deterministic minibatch proximal update procedure in large-scale settings. To extend the strong efficiency guarantee to generic strongly convex losses, we propose to iteratively convert the non-quadratic problem into a sequence of quadratic subproblems such that the aforementioned method can be readily applied for optimization. In this way, up to logarithmic factors, HSDMPG still enjoys an identical sharp bound of IFO for strongly convex problems.

Table 1 summarizes the computational complexity (measured by IFO) of HSDMPG and several representative baselines, including SGD (Robbins & Monro, 1951; Shamir, 2011), SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014), APSDCA (Shalev-Shwartz & Zhang, 2014), APCG (Lin et al., 2014), SPDC (Zhang & Xiao, 2015), Catalyst (Lin et al., 2015), Varag (Lan et al., 2019), AMSVRG (A. Nitanda, 2016), Katyusha (Allen-Zhu, 2017),

SCSG (Lei & Jordan, 2017). In the following, we highlight the advantages of our method over these prior approaches:

- To achieve ϵ-optimization-error, *i.e.* $\mathbb{E}[F(\boldsymbol{\theta})–F(\boldsymbol{\theta}^*)] \leq \epsilon$, the IFO complexity of HSDMPG on problem (1) is $\mathcal{O}\Big(\frac{\kappa^{1.5}\epsilon^{0.75}\log^{\tau_1}(\frac{1}{\epsilon})+1}{\epsilon} \wedge \Big(\kappa\sqrt{n}\log^{\tau_2}\big(\frac{1}{\epsilon}\big)+n\log^{\tau_3}\big(\frac{1}{\epsilon}\big)\Big)\Big)$ where $\tau_1 = 1.5$, $\tau_2 = 1.5$ and $\tau_3 = 1$ for quadratic loss and $\tau_1 = 2.25$, $\tau_2 = 2.5$ and $\tau_3 = 2$ for generic strongly convex loss. In comparison, the IFO complexity bounds of all the compared algorithms except SGD and SCSG scale linearly w.r.t. the data size n. As specified in the third column of Table 1, HSDMPG is superior to these algorithms in large-scale problem settings which are of central interest in big data applications. Compared with SGD, since in most cases, the condition number κ is at the order of $\mathcal{O}(1/\mu)$, HSDMPG improves over SGD by a factor at least $\mathcal{O}\big(\kappa \wedge \frac{1}{\kappa^{0.5}\epsilon^{0.75}}\big)$ (up to logarithm factors). For SCSG, HSDMPG also shows higher computational efficiency when (1) the optimization error ϵ is small which corresponds to conditions ① or ② in Table 1; and (2) the data size n is large which corresponds to condition ③ in Table 1.
- For the practical setting where $\epsilon = \mathcal{O}(1/\sqrt{n})$ which matches the optimal intrinsic excess error, HSDMPG has the IFO complexity $\mathcal{O}\left(n^{0.875}\log^{1.5}(n)\right)$ for the quadratic loss and $\mathcal{O}\left(n^{0.875}\log^{2.25}(n)\right)$ for the generic strongly convex loss. By ignoring the small logarithm term $\log(n)$, both complexities of HSDMPG are lower than the complexity bound $\mathcal{O}\big(n\big)$ of SGD by a factor $\mathcal{O}\big(n^{0.125}\big)$. Similarly, HSDMPG respectively improves over APCG and other remaining algorithms, such as SVRG, Katyusha, Varag and SCSG, by factors of $\mathcal{O}\big(n^{0.375}\big)$ and $\mathcal{O}\big(n^{0.125}\big)$. These results demonstrate the superior computational efficiency of HSDMPG for attaining near-optimal generalization rate of a statistical learning model.

2. Related Work

Stochastic gradient algorithms. Gradient descent (GD) (Cauchy, 1847) method has long been applied to solve ERM and enjoys linear convergence rate on strongly convex problems. But it needs to compute full gradient per iteration, leading to huge computation cost on large-scale problems. To improve efficiency, incremental gradient algorithms have been developed via leveraging the finite-sum structure and have witnessed tremendous progress recently. For instance, SGD (Robbins & Monro, 1951; Bottou, 1991) only evaluates gradient of one (or a minibatch) randomly selected sample at each iteration, which greatly reduces the cost of each iteration and shows more appealing efficiency than GD on large-scale problems (Shamir, 2011; A. Nitanda, 2016; Hendrikx et al., 2019; Mohammadi et al., 2019). Along this line of research, a variety of variance-reduced variants, such as SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014), APSDCA (Shamir, 2011), AMSVRG (A. Nitanda, 2016), SCSG (Lei & Jordan, 2017), Catalyst (Lin et al., 2015), Katyusha (Allen-Zhu, 2017), Varag (Lan et al., 2019), are developed and have delivered exciting progress such as linear convergence rates on strongly convex problems as opposed to sublinear rates of vanilla SGD (Shamir, 2011). The hybrid stochastic-deterministic gradient descent method (Friedlander & Schmidt, 2012; Zhou et al., 2018a;b; Mokhtari et al., 2016; Mokhtari & Ribeiro, 2017) iteratively samples an evolving minibatch of samples for gradient estimation or subproblem construction and works favorably in reducing the computational complexity. Our HSDMPG method differs significantly from these prior algorithms. Based on the Bregman-divergence of the minibatch function and a hybrid stochastic-deterministic first-order approximation of the original function, HSDMPG constructs a variance-reduced minibatch proximal function which is provably more efficient. Moreover, HSDMPG can employ any off-the-shelf algorithms to solve the constructed sub-problems in the inner loop and thus is flexible for implementation. HSDMPG shares a similar spirit with the DANE method (Shamir et al., 2014) which also uses a local Bregman-divergence-based function approximation for communication-efficient distributed quadratic loss optimization. The main difference lies in the way of constructing first-order approximation of the risk function: HSDMPG employs a novel hybrid stochastic-deterministic approximation strategy which is substantially more efficient than the deterministic strategy as used by DANE.

Generalization and optimization. In the seminal work of Bottou & Bousquet (2008), it has been demonstrated that the excess error that measures the generalization performance of an ERM model over a function class can be decomposed into three terms in expectation: an *approximation error* that measures how accurate the function class can approximate the underlying optimum model; an *estimation error* that measures the effects of minimizing ERM instead of population risk; and an *optimization error* that represents the difference between the exact solution and the approximate solution of ERM. Particularly, for the ℓ_2-regularized convex ERM with linear models as in (1), its estimation error (or excess risk) has long been studied with a vast body of deep theoretical results established (Shalev-Shwartz & Ben-David, 2014; Hardt et al., 2016; Bach & Moulines, 2013; Dieuleveut et al., 2017; Zhou & Feng, 2018a;b). A simple yet powerful tool for analyzing estimation error is the *stability* of an estimator to the changes of training dataset (Bousquet & Elisseeff, 2002). The ℓ_2-regularized convex ERM has been shown to have uniform stability of order $\mathcal{O}(1/(\mu n))$ (Bousquet & Elisseeff, 2002), which then gives rise to the optimal choice $\mu = \mathcal{O}(1/\sqrt{n})$

to balance empirical loss and generalization gap to achieve estimation error $\mathcal{O}(1/\sqrt{n})$ (Shalev-Shwartz et al., 2009; Feldman & Vondrak, 2019). This implies that the overall excess error is dominated by $\mathcal{O}(1/\sqrt{n})$. In this sense, it suffices to solve the ℓ_2-regularized ERM to optimization error $\mathcal{O}(1/\sqrt{n})$ to match the intrinsic excess error.

3. Hybrid Stochastic-Deterministic Minibatch Proximal Gradient

In this section, we first introduce the hybrid stochastic-deterministic minibatch proximal gradient (HSDMPG) algorithm for quadratic loss function along with convergence rate and computational complexity analysis. Then, we extend HSDMPG and its theoretical analysis to generic strongly convex loss functions.

3.1. The HSDMPG method for quadratic loss

3.1.1. Algorithm

The HSDMPG method is outlined in Algorithm 1. The initial step is to randomly sample a minibatch $\mathcal{S}$ of data points of size s to construct a stochastic approximation

$$F_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{s}\sum\nolimits_{i\in\mathcal{S}} \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) + \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2 \qquad (2)$$

to the original risk function $F(\boldsymbol{\theta})$ in problem (1). $F_{\mathcal{S}}(\boldsymbol{\theta})$ will be fixed throughout the computational procedure to follow. Then in the iteration loop the algorithm iterates between two steps of S1 and S2. In step S1, we uniformly randomly sample a size increasing minibatch $\mathcal{S}_t$ of samples to estimate an inexact function $F_{\mathcal{S}_t}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) + \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$. Let $\mathcal{D}_g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = g(\boldsymbol{\theta}_1) - g(\boldsymbol{\theta}_2) - \langle \nabla g(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\rangle$ denote the Bregman divergence of a function g. Based on $F_{\mathcal{S}}(\boldsymbol{\theta})$ and $F_{\mathcal{S}_t}(\boldsymbol{\theta})$, we construct a variance-reduced minibatch proximal objective $\widetilde{P}_{t-1}(\boldsymbol{\theta})$ to approximate the objective $F(\boldsymbol{\theta})$ in (1), where $\widetilde{P}_{t-1}(\boldsymbol{\theta}) \triangleq$

$$F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1}) + \langle \nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}\rangle + \mathcal{D}_{\widetilde{F}_{\mathcal{S}}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t-1}).$$

Here $\mathcal{D}_{\widetilde{F}_{\mathcal{S}}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t-1})$ is the Bregman divergence of a regularized loss $\widetilde{F}_{\mathcal{S}}(\boldsymbol{\theta}) = F_{\mathcal{S}}(\boldsymbol{\theta}) + \frac{\gamma}{2}\|\boldsymbol{\theta}\|_2^2$ which essentially measures the distance between $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t-1}$ on the current geometry curve estimated on $\widetilde{F}_{\mathcal{S}}(\boldsymbol{\theta})$. We define the next iterate as

$$\boldsymbol{\theta}_t = \arg\min_{\boldsymbol{\theta}} \widetilde{P}_{t-1}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} P_{t-1}(\boldsymbol{\theta}), \qquad (3)$$

where $P_{t-1}(\boldsymbol{\theta}) \triangleq$

$$F_{\mathcal{S}}(\boldsymbol{\theta}) + \langle \nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1}) - \nabla F_{\mathcal{S}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}\rangle + \frac{\gamma}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}\|_2^2.$$

In P_{t-1}, its finite-sum structure comes from the initial stochastic approximation $F_{\mathcal{S}}(\boldsymbol{\theta})$ and its gradient at $\boldsymbol{\theta}_{t-1}$.

Algorithm 1 Hybrid Stochastic-Deterministic Minibatch Proximal Gradient (HSDMPG) for quadratic loss.

Input: initialization $\boldsymbol{\theta}_0$, regularization constant γ in (3), optimization error ε_t.
Initialization: Uniformly randomly sample a data batch $\mathcal{S}$ of size s to form $F_{\mathcal{S}}(\boldsymbol{\theta})$ in (2).
for $t = 1, 2, \ldots, T$ **do**
  (S1) Uniformly randomly sample a minibatch $\mathcal{S}_t$ to form $F_{\mathcal{S}_t}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) + \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$ and compute $\nabla F_{\mathcal{S}_t}(\boldsymbol{\theta}_{t-1})$ to construct loss $P_{t-1}(\boldsymbol{\theta})$ in (3).
  (S2) Optimize the subproblem (3), *e.g.* via SVRG, to obtain $\boldsymbol{\theta}_t$ that satisfies $\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\|_2 \le \varepsilon_t$.
end for
Output: $\boldsymbol{\theta}_T$.

Since along with more iterations, the size of $\mathcal{S}_t$ increases which indicates that the loss P_{t-1} is a variance-reduced loss and will converge to the original loss $F(\boldsymbol{\theta})$ in problem (1). Then in step S2, we approximately solve problem (3) via a stochastic gradient optimization method such as SVRG. The principle behind this strategy is that for the initial optimization progress, inexact gradient already can well decrease the loss since the current solution is far from the optimum, while along more iterations, the current solution becomes closer to optimum, requiring more accurate gradient for further reducing the loss function. In this way, our proposed method can well balance the converge speed and the computational cost at each iteration and thus has the potential to achieve improved overall computational efficiency. Shamir et al. (2014) has proposed the DANE method which uses a similar local Bregman divergence based regularization for distributed quadratic optimization problems. Our method improves upon DANE in two aspects: 1) we use variance-reduction techniques to reduce the overall computational complexity, and 2) HSDMPG is applicable not only to quadratic problems but also to generic strongly convex problems with about the same computational complexity as discussed in Sec. 3.2.

3.1.2. Convergence and complexity analysis

We first introduce two necessary definitions, namely strong convexity and Lipschitz smoothness, which are conventionally used in the analysis of convex optimization methods (Shamir, 2011; Johnson & Zhang, 2013).

Definition 1 (Strong Convexity and Smoothness)**.** *A differentiable function $g(\boldsymbol{\theta})$ is said to be μ-strongly-convex and L-smooth if $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, it satisfies*

$$\frac{\mu}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 \le \mathcal{D}_g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \le \frac{L}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2.$$

where $\mathcal{D}_g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = g(\boldsymbol{\theta}_1) - g(\boldsymbol{\theta}_2) - \langle \nabla g(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\rangle$.

For brevity, let $\boldsymbol{H}$ be the Hessian matrix of the quadratic function $F(\boldsymbol{\theta})$ and $\ell_i(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) + \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$. Denote

$\|\boldsymbol{\theta}\|_{\boldsymbol{H}} = \sqrt{\boldsymbol{\theta}^\top \boldsymbol{H}\boldsymbol{\theta}}$. In the analysis to follow, we always suppose that $\|\boldsymbol{x}_i\| \leq r, \forall i$, which generally holds for natural data analysis, *e.g.*, in computer vision and signal processing. We summarize our main result in Theorem 1 which shows the linear convergence rate of HSDMPG for quadratic problems. See proof in Appendix B.1.

Theorem 1. *Assume each loss $\ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i)$ is quadratic and L-smooth w.r.t. $\boldsymbol{\theta}^\top \boldsymbol{x}_i$, and $\sup_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{H}^{-1/2}(\nabla F(\boldsymbol{\theta}) - \nabla \ell_i(\boldsymbol{\theta}))\|_2^2 \leq \nu^2$. By setting $\gamma = (\sqrt{\log(d)} + \sqrt{2})Lr^2/\sqrt{s}$, $\varepsilon_t = \frac{\mu^{1.5}}{4(\mu+2\gamma)} \exp\big(-\frac{\mu(t-1)}{2(\mu+2\gamma)}\big)$, $|\mathcal{S}_t| = \frac{16\nu^2(\mu+2\gamma)^2}{\mu^2}\exp\big(\frac{\mu t}{2(\mu+2\gamma)}\big) \wedge n$, where d is the problem dimension, the sequence $\{\boldsymbol{\theta}_t\}$ produced by Algorithm 1 satisfies*

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

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

The main message conveyed by Theorem 1 is that HSDMPG enjoys linear convergence rate on the quadratic loss when we use evolving size of the minibatch $\mathcal{S}_t$. Note here we only assume each loss $\ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i)$ is L-smooth w.r.t. $\boldsymbol{\theta}^\top \boldsymbol{x}_i$. This assumption is much milder than the smoothness assumption on the function $F(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ which is used in other algorithm analysis, such as SGD and SVRG. The assumption that $\sup_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{H}^{-1/2}(\nabla F(\boldsymbol{\theta}) - \nabla \ell_i(\boldsymbol{\theta}))\|_2^2 \leq \nu^2$ in HSDMPG is mild, which requires the variance of stochastic gradient under the Hessian matrix is bounded. Such an assumption is analogous to the one used in analysis of SGD that imposing the bounded-variance assumption on stochastic gradient, namely, $\frac{1}{n}\sum_{i=1}^n \|\nabla F(\boldsymbol{\theta}) - \nabla \ell_i(\boldsymbol{\theta})\|_2^2$.

Based on this result, we further analyze the computational complexity of HSDMPG to better understand its overall efficiency in computation. At each iteration, we use the SVRG method solve the inner-loop subproblem (3) because it only accesses the first-order information of the objective function and is efficient. Following (Johnson & Zhang, 2013; Zhang & Xiao, 2015; Zhou et al., 2019; Shen et al., 2019), we employ the incremental first order oracle (IFO) complexity as the computation complexity metric for solving the finite-sum solving problem (1).

Definition 2. *An IFO takes an index $i \in [n]$ and a point $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, and returns the pair $(\ell_i(\boldsymbol{\theta}), \nabla \ell_i(\boldsymbol{\theta}))$.*

The IFO complexity can accurately reflect the overall computational performance of a first-order algorithm, as objective value and gradient evaluation usually dominate the per-iteration complexity. Based on these preliminaries, we summarize our main result on the computation complexity of HSDMPG in Corollary 1 with proof provided in Appendix B.2.

Corollary 1 (Computation complexity of HSDMPG for quadratic loss)**.** *Suppose that the assumptions in Theorem 1 hold and the inner-loop subproblems are solved via SVRG, then the IFO complexity of* HSDMPG *on the quadratic loss to achieve $\mathbb{E}[F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}^*)] \leq \epsilon$ is of the order $\mathcal{O}\Big(\big(1 + \frac{\kappa^3 \log^{1.5}(d)}{s^{1.5}}\big)\frac{\nu^2}{\epsilon} \bigwedge \big(1 + \frac{\kappa \log^{0.5}(d)}{s^{0.5}}\big) n \log\big(\frac{1}{\epsilon}\big) + \kappa\sqrt{s\log(d)}\log^2\big(\frac{1}{\epsilon}\big)\Big)$, where $\kappa = L/\mu$ denotes the conditional number.*

According to Corollary 1, by choosing s as $s = \frac{\kappa\nu\log^{0.5}(d)}{\epsilon^{0.5}\log(1/\epsilon)} \wedge n$ or $s = \frac{n}{\log(1/\epsilon)}$ and ignoring the constant ν and the logarithm factor $\log(d)$ of the problem dimension d, the IFO complexity of HSDMPG is at the order of

$$\mathcal{O}\Big(\frac{\kappa^{1.5}\epsilon^{0.75}\log^{1.5}(\frac{1}{\epsilon}) + 1}{\epsilon} \wedge \Big(\kappa\sqrt{n}\log^{1.5}\big(\frac{1}{\epsilon}\big) + n\log\big(\frac{1}{\epsilon}\big)\Big)\Big).$$

One may compare such a complexity with the state-of-the-arts listed in Table 1. Compared with those algorithms in the table whose IFO complexity scales linearly with the data size n, *e.g.* SVRG, APCG, Katyusha and AMSVRG, the proposed HSDMPG has data-size-independent IFO complexity and can outperform them for large-scale learning problems where the data size n could be huge. To be more precise, the third column of Table 1 summarizes the conditions under which HSDMPG outperforms these algorithms in terms of computational complexity. For the algorithms whose IFO complexity does not depend on n, namely SGD and SCSG, HSDMPG also enjoys substantially lower complexity in most cases. Concretely, since κ is typically at the order of $\mathcal{O}(1/\mu)$, when $\kappa \leq \epsilon^{1.5}$ which holds for moderately larger κ, HSDMPG improves over SGD by a factor at least $\mathcal{O}\big(\kappa \wedge \frac{1}{\kappa^{0.5}\epsilon^{0.75}}\big)$ (up to the logarithmic factor). As for SCSG, HSDMPG also achieves higher efficiency when (1) the optimization error is small which corresponds to conditions ① in the third column of Table 1, (2) the sampler size n is large which corresponds to condition ②. These results show that HSDMPG is well suited for solving large-scale learning problems.

From the perspective of generalization, we are particularly interested in the computational complexity of HSDMPG for optimizing the ℓ_2-ERM model (1) to its intrinsic excess error bound which characterizes the generalization performance of the model. As reviewed in Section 2, the excess error of the considered ℓ_2-ERM model is typically of order $\mathcal{O}(1/\sqrt{n})$. Accordingly, one only needs to solve the optimization problem to the optimization error $\epsilon = \mathcal{O}(1/\sqrt{n})$ (Bottou & Bousquet, 2008; Shalev-Shwartz et al., 2009). Moreover, to accord with this intrinsic excess error bound, the regularization constant μ should also be at the order of $\mathcal{O}(\frac{1}{\sqrt{n}})$. In this way, the condition number κ could scale as large as $\mathcal{O}(\sqrt{n})$. Based on these results and Corollary 1, we can derive the IFO complexity bound of HSDMPG for this case in Corollary 2.

Corollary 2. *Suppose that the assumptions in Corol-*

Algorithm 2 Hybrid Stochastic-Deterministic Minibatch Proximal Gradient (HSDMPG) on the generic loss.

Input: Regularization constant γ and initialization $\boldsymbol{\theta}_0$.
for $t = 1, 2, \ldots, T$ **do**
(S1) Construct a finite-sum quadratic function $\boldsymbol{Q}_{t-1}(\boldsymbol{\theta})$ in Eqn. (4) to approximate $F(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_{t-1}$.
(S2) Run Algorithm 1 with regularization constant γ and initialization $\boldsymbol{\theta}_{t-1}$ to minimize the finite-sum function $\boldsymbol{Q}_{t-1}(\boldsymbol{\theta})$ such that $\boldsymbol{Q}_{t-1}(\boldsymbol{\theta}_t) \leq \min_{\boldsymbol{\theta}} \boldsymbol{Q}_{t-1}(\boldsymbol{\theta}) + \varepsilon'_t$.
end for
Output: $\boldsymbol{\theta}_T$.

lary 1 hold. By setting $s = \mathcal{O}\big(\frac{\nu n^{0.75}\log^{0.5}(d)}{\log(n)}\big)$, *the IFO complexity of* HSDMPG *on the quadratic loss to achieve* $\mathbb{E}[F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}^*)] \leq \frac{1}{\sqrt{n}}$ *is at the order of* $\mathcal{O}\big(\nu^{0.5} n^{0.875}\log^{0.75}(d)\log^{1.5}(n) + \nu^2 n^{0.5}\big)$.

See its proof in Appendix B.3. From Corollary 2, one can observe that the IFO complexity of HSDMPG for quadratic problems is at the order of $\mathcal{O}\left(n^{0.875}\log^{1.5}(n)\right)$. It means that HSDMPG can reach the intrinsic excess error $\mathcal{O}\big(1/\sqrt{n}\big)$ with strictly less than a single pass over the entire training dataset. In comparison, we can observe from Table 1 that in the same practical setting, SGD and APCG have IFO complexity $\mathcal{O}(n)$ and $\mathcal{O}\left(n^{1.25}\log(n)\right)$ respectively. By ignoring the logarithm factor $\log(n)$ which is much smaller than n for large-scale learning problems, HSDMPG improves over these two methods by factors $\mathcal{O}\big(n^{0.125}\big)$ and $\mathcal{O}\big(n^{0.375}\big)$, respectively. The IFO complexity of all other algorithms in Table 1, including SVRG, SCSG, SPDC, APSDCA, AMSVRG, Catalyst, Katyusha and Varag, are all at the order of $\mathcal{O}(n\log(n))$. Similarly, by ignoring the logarithmic factors, HSDMPG has lower IFO complexity than these algorithms by a factor $\mathcal{O}\big(n^{0.125}\big)$. To summarize this group of results comparison, HSDMPG would be significantly superior to all these state-of-the-arts when solving quadratic optimization problems to intrinsic excess error.

3.2. Algorithm for generic convex loss function

The computational complexity guarantees established in the previous section are only applicable to quadratic loss function. In order to extend these results to non-quadratic convex loss function, we apply a quadratic approximation strategy to convert the original non-quadratic problem into a sequence of quadratic optimization sub-problems such that each of the subproblem can be optimized by HSDMPG. More specifically, suppose that the loss function $\ell(\boldsymbol{\theta}^\top\boldsymbol{x}, \boldsymbol{y})$ is twice differentiable w.r.t. $\boldsymbol{\theta}^\top\boldsymbol{x}$ and is L-smooth w.r.t. $\boldsymbol{\theta}^\top\boldsymbol{x}$. Then we can verify that $\nabla^2 F(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell''(\boldsymbol{\theta}^\top\boldsymbol{x}_i, \boldsymbol{y}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \mu\boldsymbol{I} \preceq \bar{\boldsymbol{H}} \triangleq \frac{L}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top + \mu\boldsymbol{I}$ for all $\boldsymbol{\theta}$. Therefore, at each iteration, we construct an upper bound of the second-order Taylor expansion of F at $\boldsymbol{\theta}_{t-1}$ as expressed by $\boldsymbol{Q}_{t-1}(\boldsymbol{\theta}) \triangleq$

$$F(\boldsymbol{\theta}_{t-1}) + \langle \nabla F(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}\rangle + \Delta_{t-1}(\boldsymbol{\theta}), \quad (4)$$

where $\Delta_{t-1}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})^\top \bar{\boldsymbol{H}}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})$. The finite-sum structure in $\boldsymbol{Q}_{t-1}(\boldsymbol{\theta})$ comes from $\nabla F(\boldsymbol{\theta}_{t-1}) = \frac{1}{n}\sum_{i=1}^n \nabla\ell(\boldsymbol{\theta}^\top\boldsymbol{x}_i, \boldsymbol{y}_i) + \mu\boldsymbol{\theta}$ and $\bar{\boldsymbol{H}}$. Thus we can estimate $\boldsymbol{\theta}_t$ by applying HSDMPG to the quadratic function $\boldsymbol{Q}_{t-1}(\boldsymbol{\theta})$ with a warm-start initialization $\boldsymbol{\theta}_{t-1}$ such that

$$\boldsymbol{Q}_{t-1}(\boldsymbol{\theta}_t) \leq \min_{\boldsymbol{\theta}} \boldsymbol{Q}_{t-1}(\boldsymbol{\theta}) + \varepsilon'_t. \quad (5)$$

The above nested-loop computation procedure is summarized in Algorithm 2. We remark that when computing the gradient of $\boldsymbol{Q}_{t-1}(\boldsymbol{\theta})$, we can compute the gradient associated with $\bar{\boldsymbol{H}}$ at the point $\boldsymbol{\theta}$ as $\bar{\boldsymbol{H}}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) = \frac{L}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}))\boldsymbol{x}_i + \mu(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})$ which only computes the inner-product $\boldsymbol{x}_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})$ without explicitly computing $\bar{\boldsymbol{H}}$. In this way, the computational cost of each stochastic gradient associated with $\bar{\boldsymbol{H}}$ is actually much cheaper than that of computing stochastic gradient of $\nabla F(\boldsymbol{\theta}_{t-1})$, since the former only involves vector products and the later one is usually complicated, *e.g.* involving the exponential computation in logistic regression. Then we establish Theorem 2 to guarantee the convergence of Algorithm 2 and analyze its computational complexity. See Appendix C.1 for a proof of this main result.

Theorem 2 (Convergence rate and computation complexity of HSDMPG for generic loss)**.** *Suppose that each loss function* $\ell(\boldsymbol{\theta}^\top\boldsymbol{x}, \boldsymbol{y})$ *is* L*-smooth and* σ*-strongly convex w.r.t.* $\boldsymbol{\theta}^\top\boldsymbol{x}$. *By setting* $\varepsilon'_t = \frac{\sigma}{2L}\exp\left(-\frac{\sigma}{2L}t\right)$, *the sequence* $\{\boldsymbol{\theta}_t\}$ *produced by Algorithm 2 satisfies*

$$F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}^*) \leq \exp\Big(-\frac{\sigma t}{2L}\Big)\big(1 + F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}^*)\big).$$

Suppose the assumptions in Corollary 1 hold. Then by setting $\kappa = \frac{L}{\mu}$ *the IFO complexity of Algorithm 2 to achieve* $\mathbb{E}\left[F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}^*)\right] \leq \epsilon$ *is at the order of* $\mathcal{O}\Big(\Big(1 + \frac{\kappa^3\log^{1.5}(d)}{s^{1.5}}\Big)\frac{L\nu^2}{\sigma\epsilon} \bigwedge \Big(1 + \frac{\kappa\log^{0.5}(d)}{s^{0.5}}\Big)\frac{L^3 n}{\sigma^3}\log^2\big(\frac{1}{\epsilon}\big) + \frac{L^2\sqrt{s\log(d)}}{\sigma\mu}\log^3\big(\frac{1}{\epsilon}\big)\Big)$.

Theorem 2 suggests that the objective $F(\boldsymbol{\theta}_t)$ converges linearly to the optimum $F(\boldsymbol{\theta}^*)$ with rate $\exp(-\frac{\sigma}{2L})$. Note that σ is the strong convexity parameter of the loss function $\ell(\boldsymbol{\theta}^\top\boldsymbol{x}, \boldsymbol{y})$ w.r.t. $\boldsymbol{\theta}^\top\boldsymbol{x}$ instead of $\boldsymbol{\theta}$ which is usually not relying on data scale for widely used loss functions such as the logistic loss (Yuan & Li, 2019) and thus leads to fast outer-loop convergence rate. In contrast, the strong convexity parameter μ of the risk function F is typically set at the order of $\mathcal{O}\big(1/\sqrt{n}\big)$ so as to match the intrinsic excess error.

In terms of computational complexity, by choosing the proper value of s from $s = \frac{\nu\kappa\log^{0.5}(d)}{\epsilon^{0.5}\log^{1.5}(1/\epsilon)} \wedge n$ and $s =$

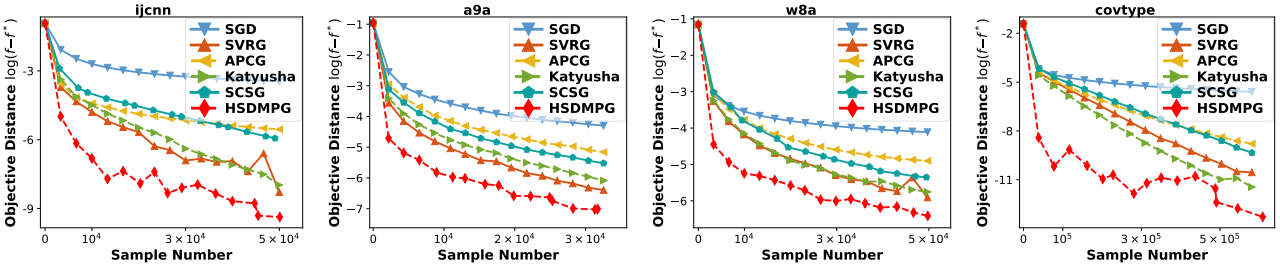

Figure 1: Single-epoch processing: stochastic gradient algorithms process data a single pass on quadratic problems.

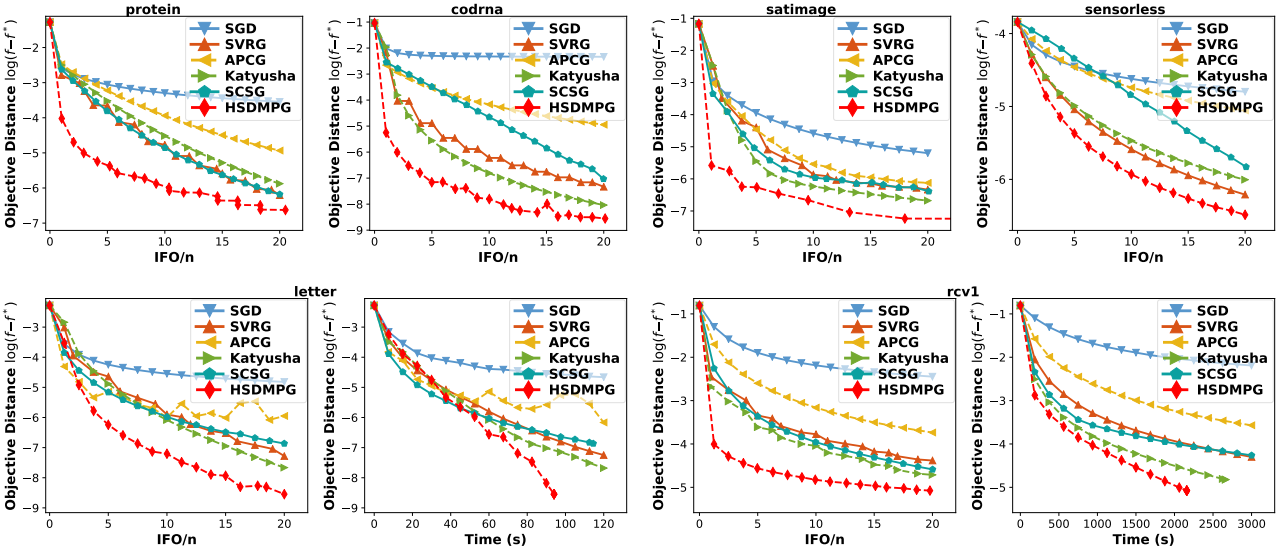

Figure 2: Multi-epoch processing: stochastic gradient algorithms process data multiple pass on quadratic problems.

$\frac{\mu L\kappa n}{\sigma^2\log(1/\epsilon)} \wedge n$ in Algorithm 1, the IFO complexity of HSDMPG for generic convex loss can be shown to scale as

$$\mathcal{O}\Big(\frac{\kappa^{1.5}\epsilon^{0.75}\log^{2.25}(\frac{1}{\epsilon})+1}{\epsilon}\wedge\Big(\kappa\sqrt{n}\log^{2.5}(\frac{1}{\epsilon})+n\log^2(\frac{1}{\epsilon})\Big)\Big)$$

Compared with the methods listed in Table 1, one can observe that for generic strongly convex problems, HSDMPG enjoys lower computational complexity than all the compared algorithms except SGD and SCSG for large-scale learning problems where the sample number n is sufficiently large to satisfy the conditions in the third column of Table 1. Similar to the results on quadratic loss, HSDMPG improves over SGD by a factor at least $\mathcal{O}\big(\kappa\wedge\frac{1}{\kappa^{0.5}\epsilon^{0.75}}\big)$. So when the optimization error ϵ is very small or the condition number κ is large, HSDMPG will be much more efficient than SGD. For SCSG, HSDMPG is of higher efficiency in two regimes, namely (1) the optimization error is small which corresponds to conditions ① or ② in Table 1, and (2) the sampler number n is large which corresponds to condition ③. These results show the advantages HSDMPG in solving large-scale strongly-convex learning problems.

Finally we consider a realistic case where the optimization error of problem (1) matches the intrinsic excess error bound $\mathcal{O}(1/\sqrt{n})$. For this case, as discussed at the end of Section 3.1.2 that the regularization parameter should be set at the scale of $\mu=\mathcal{O}(1/\sqrt{n})$ with balanced impact against the guarantees on estimation error. As a result, the condition number κ could scale as large as $\mathcal{O}(\sqrt{n})$. The following corollary substantializes the IFO complexity bound in Theorem 2 to such a setting. See Appendix C.2 for a proof of this result.

Corollary 3. *Suppose the assumptions in Theorem 2 hold. By setting* $s=\mathcal{O}\big(\frac{\nu n^{0.75}\log^{0.5}(d)}{\log(n)}\big)$, *the IFO complexity of* HSDMPG *on the generic loss to achieve* $\mathbb{E}[F(\boldsymbol{\theta}_t)-F(\boldsymbol{\theta}^*)]\leq\frac{1}{\sqrt{n}}$ *is of order* $\mathcal{O}\big(\nu^{0.5}n^{0.875}\log^{0.75}(d)\log^{2.25}(n)+\nu^2n^{0.5}\big)$.

Corollary 3 shows that for generic convex loss, the IFO complexity of HSDMPG to attain the $\mathcal{O}\big(1/\sqrt{n}\big)$ intrinsic excess error is of the order $\mathcal{O}\left(n^{0.875}\log^{2.25}(n)\right)$. This shows that HSDMPG is able to achieve nearly optimal generalization with less than a single pass over data. Compared with the complexity bound for the quadratic loss, such a more general IFO complexity bound of HSDMPG only comes at the cost of a slightly increased overhead on the logarithmic factor, *i.e.*, from $\log^{1.5}(n)$ for the quadratic case to the $\log^{2.25}(n)$ for generic convex loss. Similar to the ob-

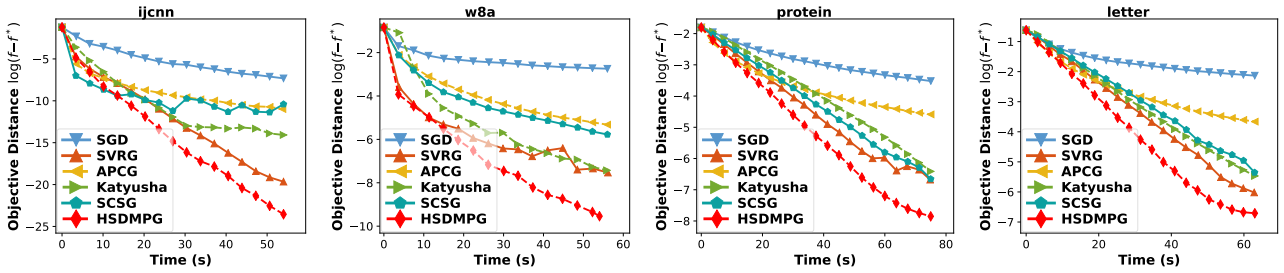

Figure 3: Multi-epoch processing (about 8 epochs): stochastic gradient algorithms process data multiple pass on logistic regression problems (ijcnn and w08) and softmax regression problems (protein and letter).

servations in the quadratic case, from results in Table 1 one can observe that all the considered state-of-the-art methods need to process the entire data at least one pass to achieve the desired optimization error for generic convex loss. All in all, the established theoretical results for both quadratic and non-quadratic loss functions showcase the benefit of HSDMPG for efficient optimization of large-scale learning problems with near-optimal generalization.

4. Experiments

In this section, we carry out experiments to compare the numerical performance of HSDMPG with several representative stochastic gradient optimization algorithms, including SGD (Robbins & Monro, 1951), SVRG (Johnson & Zhang, 2013), APCG (Lin et al., 2014), Katyusha (Allen-Zhu, 2017) and SCSG (Lei & Jordan, 2017). We evaluate all the considered algorithms on two sets of strongly-convex learning tasks. The first set is for ridge regression with least squared loss $\ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) = \frac{1}{2}\|\boldsymbol{\theta}^\top \boldsymbol{x}_i - \boldsymbol{y}_i\|_2^2$, where $\boldsymbol{y}_i$ is the target output of sample $\boldsymbol{x}_i$. In the second setting we consider two classification models: logistic regression with loss $\ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) = \log\left(1+\exp(-\boldsymbol{y}_i \boldsymbol{\theta}^\top \boldsymbol{x}_i)\right)$ and multi-class softmax regression with k-classification loss $\ell(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \boldsymbol{y}_i) = \sum_{j=1}^k \mathbf{1}\{\boldsymbol{y}_i = j\} \log\left(\frac{\exp(\boldsymbol{\theta}_j^\top \boldsymbol{x}_i)}{\sum_{s=1}^k \exp(\boldsymbol{\theta}_s^\top \boldsymbol{x}_i)}\right)$. We run simulations on ten datasets whose details are described in Appendix D.4. For HSDMPG, we set the size s of $\mathcal{S}$ around $n^{0.75}$. For the minibatch for inner problems, we set initial minibatch size $|\mathcal{S}_1| = 50$ and then follow our theory to exponentially expand size of $\mathcal{S}_t$ with proper exponential rate. The regularization constant in the subproblem (3) is set to be $\gamma = \sqrt{\log(d)/s}$ as suggested by our theory. The optimization error ε_t in (3) is controlled by respectively allowing SVRG to run 3 epochs and 10 epochs on the two sets of tasks. Similarly, we control the optimization error ε_t' in (5) by running SVRG with 3 epochs. Since there is no ground truth on real data, we run FGD sufficiently long until $\|\nabla F(\tilde{\boldsymbol{\theta}})\|_2 \leq 10^{-10}$ and take $F(\tilde{\boldsymbol{\theta}})$ as an approximate optimal value $F(\boldsymbol{\theta}^*)$ for sub-optimality estimation.

4.1. Results for the quadratic loss

Single-epoch evaluation results. Here we first evaluate well-conditioned quadratic problems such that moderately accurate solution can be obtained after only one epoch of data pass. Such a one epoch setting usually occurs in online learning. Towards this goal, we set the regularization parameter $\mu = 0.01$ to make the quadratic problems well-conditioned. From Figure 1, one can observe that HSDMPG exhibits much sharper convergence behavior than the considered baselines, though most algorithms can achieve small optimization error after one epoch processing of data. This confirms the theoretical predictions in Corollaries 1 and 2 that HSDMPG is cheaper in IFO complexity than SGD and variance-reduced algorithms, *e.g.* SVRG and SCSG, when the data scale is large.

Multi-epoch evaluation results. For more challenging problems, an algorithm usually requires multiple cycles of data processing to achieve accurate optimization. Here we reset the regularization strength parameter in quadratic problems as $\mu = 10^{-4}$ for generating more challenging optimization tasks. As shown in Figure 2, one can again observe that HSDMPG converges faster than all the compared algorithms in terms of IFO complexity. Particularly, we compare both IFO complexity and wall-clock running time on the letter and rcv11 datasets. The convergence curves under these two metrics consistently show the superior computational efficiency of HSDMPG to the considered state-of-the-arts on large-scale learning tasks, which well support the theoretical predictions in Corollaries 1 and 2.

4.2. Results for the non-quadratic loss

Finally, we investigate the convergence performance of the proposed HSDMPG on non-quadratic convex loss functions. Specifically, we evaluate all the compared algorithms on logistic regression and its multi-classes version, *i.e.* softmax regression, in which their regularization modulus parameters are set as $\mu = 0.01$. Figure 3 reports the running time evolving curves which can accurately reflects the efficiency of an algorithm. These results show that HSDMPG converges significantly faster than the baseline

algorithms for the considered non-quadratic loss functions, which well support the predictions in Theorem 2 and Corollary 3 that HSDMPG has lower IFO complexity than the state-of-the-arts in the regimes where data scale is large. This set of results also demonstrates the effectiveness of our sequential quadratic-approximation approach for extending the attractive computational complexity guarantees on quadratic loss to generic convex loss.

5. Conclusions

We proposed HSDMPG as a hybrid stochastic-deterministic minibach proximal gradient method for ℓ_2-regularized ERM problems. For quadratic loss, we showed that HSDMPG enjoys provably lower computational complexity than prior state-of-the-art SVRG algorithms in large-scale settings. Particularly, to attain the optimization error $\epsilon = \mathcal{O}\big(1/\sqrt{n}\big)$ at the order of intrinsic excess error bound of ERM which is sufficient for generalization, the stochastic gradient complexity of HSDMPG is dominated by $\mathcal{O}(n^{0.875})$ (up to logarithmic factors). To our best knowledge, HSDMPG for the first time achieves nearly optimal generalization in less than a single pass over data. Almost identical computational complexity guarantees hold for an extension of HSDMPG to generic strongly convex loss functions via sequential quadratic approximation. Extensive numerical results demonstrate the substantially improved computational efficiency of HSDMPG over the prior methods. We expect that the algorithms and computational learning theory developed in this paper for ℓ_2-regularized ERM can be extended to stochastic convex optimization problems. Also, it is worthwhile to explore the opportunity of using first-order acceleration techniques to further improve the computational complexity guarantees of HSDMPG.

Acknowledgements

The authors sincerely thank the anonymous reviewers for their constructive comments on this work. Xiao-Tong Yuan is supported in part by National Major Project of China for New Generation of AI under Grant No.2018AAA0100400 and in part by Natural Science Foundation of China (NSFC) under Grant No.61876090 and No.61936005.

References

A. Nitanda, A. Accelerated stochastic gradient descent for minimizing finite sums. In *Artificial Intelligence and Statistics*, pp. 195–203, 2016.

Allen-Zhu, Z. Katyusha: The First Direct Acceleration of Stochastic Gradient Methods. In *ACM SIGACT Symposium on Theory of Computing*, 2017.

Bach, F. and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Proc. Conf. Neural Information Processing Systems*, pp. 773–781, 2013.

Bottou, L. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12, 1991.

Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. In *Proc. Conf. Neural Information Processing Systems*, pp. 161–168, 2008.

Bousquet, O. and Elisseeff, A. Stability and generalization. *J. of Machine Learning Research*, 2(Mar):499–526, 2002.

Cauchy, M. A. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptesrendus des séances de l'Académie des sciences de Paris*, 25:536–538, 1847.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Proc. Conf. Neural Information Processing Systems*, pp. 1646–1654, 2014.

Dieuleveut, A., Flammarion, N., and Bach, F. Harder, better, faster, stronger convergence rates for least-squares regression. *J. of Machine Learning Research*, 18(1):3520–3570, 2017.

Feldman, V. and Vondrak, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conf. on Learning Theory*, pp. 1270–1279, 2019.

Friedlander, M. P. and Schmidt, M. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *Proc. Int'l Conf. Machine Learning*, 2016.

Hendrikx, H., Bach, F., and Massoulié, L. Asynchronous accelerated proximal stochastic gradient for strongly convex distributed finite sums. *arXiv preprint arXiv:1901.09865*, 2019.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Proc. Conf. Neural Information Processing Systems*, pp. 315–323, 2013.

Lan, G., Li, Z., and Zhou, Y. A unified variance-reduced accelerated gradient method for convex optimization. In *Proc. Conf. Neural Information Processing Systems*, pp. 10462–10472, 2019.

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

Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. In *Proc. Conf. Neural Information Processing Systems*, pp. 3384–3392, 2015.

Lin, Q., Lu, Z., and Xiao, L. An accelerated proximal coordinate gradient method. In *Proc. Conf. Neural Information Processing Systems*, pp. 3059–3067, 2014.

Mohammadi, H., Razaviyayn, M., and Jovanović, M. Robustness of accelerated first-order algorithms for strongly convex optimization problems. *arXiv preprint arXiv:1905.11011*, 2019.

Mokhtari, A. and Ribeiro, A. First-order adaptive sample size methods to reduce complexity of empirical risk minimization. In *Proc. Conf. Neural Information Processing Systems*, pp. 2060–2068, 2017.

Mokhtari, A., Daneshmand, H., Lucchi, A., Hofmann, T., and Ribeiro, A. Adaptive newton method for empirical risk minimization to statistical accuracy. In *Proc. Conf. Neural Information Processing Systems*, pp. 4062–4070, 2016.

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

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *Proc. Int'l Conf. Machine Learning*, pp. 64–72, 2014.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Stochastic convex optimization. In *Conf. on Learning Theory*, 2009.

Shamir, O. Making gradient descent optimal for strongly convex stochastic optimization. *CoRR abs/1109.5647*, 2011.

Shamir, O., Srebro, N., and Zhang, T. Communication-efficient distributed optimization using an approximate newton-type method. In *Proc. Int'l Conf. Machine Learning*, pp. 1000–1008, 2014.

Shen, Z., Zhou, P., Fang, C., and Ribeiro, A. A stochastic trust region method for non-convex minimization. *arXiv preprint arXiv:1903.01540*, 2019.

Vapnik, V. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Yuan, X. and Li, P. On convergence of distributed approximate newton methods: Globalization, sharper bounds and beyond. *arXiv preprint arXiv:1908.02246*, 2019.

Zhang, Y. and Xiao, L. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *Proc. Int'l Conf. Machine Learning*, pp. 353–361, 2015.

Zhou, P. and Feng, J. Understanding generalization and optimization performance of deep cnns. In *Proc. Int'l Conf. Machine Learning*, 2018a.

Zhou, P. and Feng, J. Empirical risk landscape analysis for understanding deep neural networks. In *Int'l Conf. Learning Representations*, 2018b.

Zhou, P., Yuan, X., and Feng, J. Efficient stochastic gradient hard thresholding. In *Proc. Conf. Neural Information Processing Systems*, 2018a.

Zhou, P., Yuan, X., and Feng, J. New insight into hybrid stochastic gradient descent: Beyond with-replacement sampling and convexity. In *Proc. Conf. Neural Information Processing Systems*, pp. 1234–1243, 2018b.

Zhou, P., Yuan, X., and Feng, J. Faster first-order methods for stochastic non-convex optimization on riemannian manifolds. In *Int'l Conf. Artificial Intelligence and Statistics*, 2019.

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*

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

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 $\|\boldsymbol{T}\boldsymbol{x}\| \le \|\boldsymbol{T}\|\|\boldsymbol{x}\|$ we get

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

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

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

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

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

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

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

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

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

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

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 $\|\boldsymbol{T}\boldsymbol{x}\|\le\|\boldsymbol{T}\|\|\boldsymbol{x}\|$, we get

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

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

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

Suppose $\|\boldsymbol{x}_i\|\le r$ $(i=1,\cdots,n)$ 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

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

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

$$\frac{\mu^2}{(\mu+2\gamma)^2}\le\left\|\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right\|^2\le1\ \text{ and }\ \left\|\boldsymbol{I}-\boldsymbol{H}^{1/2}(\boldsymbol{H}_{\mathcal{S}}+\gamma\boldsymbol{I})^{-1}\boldsymbol{H}^{1/2}\right\|^2\le\frac{4\gamma^2}{(\mu+2\gamma)^2}.$$

Similarly, we have $\|\boldsymbol{H}^{-1/2}\nabla P_{t-1}(\boldsymbol{\theta}_t)\|\le\frac{1}{\sqrt{\mu}}\|\nabla P_{t-1}(\boldsymbol{\theta}_t)\|\le\frac{\varepsilon_t}{\sqrt{\mu}}$. Now we plug the above results and Eqn. (9) into Eqn. (10) and establish

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

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

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

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:

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

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 $s=\mathcal{O}\big(\frac{\nu n^{0.75}\log^{0.5}(d)}{\log(n)}\big)$ 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

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

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

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

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

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

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 $s = \mathcal{O}\left(\frac{\nu n^{0.75}\log^{0.5}(d)}{\log(n)}\right)$ 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*

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

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 $\sum_{i=1}^n \boldsymbol{z}_t^i = 0$, $\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{z}_t^i\|^2 \leq \nu^2$ 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 4 we get

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

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

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

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*

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

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

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

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 2. From it we can observe that these datasets are different from each other due to their feature dimension, training samples, and class numbers, *etc*.

Table 2: Descriptions of the ten testing datasets.

	#class	#sample	#feature		#class	#sample	#feature
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

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