Journal of Machine Learning Research XX (2023) 1-54 Submitted 2022; Published 2023

Sharper Analysis for Minibatch Stochastic Proximal Point Methods: Stability, Smoothness, and Deviation

Xiao-Tong Yuan xtyuan1980@gmail.com
CICAEET, Nanjing University of Information Science and Technology
219 Ningliu Rd, Nanjing, 210044, China

Ping Li pingli98@gmail.com
LinkedIn Ads
700 Bellevue Way NE, Bellevue, WA 98004, USA

Editor: Francis Bach and Julien Mairal

Abstract

The stochastic proximal point (SPP) methods have gained recent attention for stochastic optimization, with strong convergence guarantees and superior robustness to the classic stochastic gradient descent (SGD) methods showcased at little to no cost of computational overhead added. In this article, we study a minibatch variant of SPP, namely M-SPP, for solving convex composite risk minimization problems. The core contribution is a set of novel excess risk bounds of M-SPP derived through the lens of algorithmic stability theory. Particularly under smoothness and quadratic growth conditions, we show that M-SPP with minibatch-size n and iteration count T enjoys an in-expectation fast rate of convergence consisting of an $\mathcal{O}\left(\frac{1}{T^2}\right)$ *bias* decaying term and an $\mathcal{O}\left(\frac{1}{nT}\right)$ *variance* decaying term. In the small-n-large-T setting, this result substantially improves the best known results of SPP-type approaches by revealing the impact of noise level of model on convergence rate. In the complementary small-T-large-n regime, we provide a two-phase extension of M-SPP to achieve comparable convergence rates. Moreover, we derive a near-tight high probability (over the randomness of data) bound on the parameter estimation error of a sampling-without-replacement variant of M-SPP. Numerical evidences are provided to support our theoretical predictions when substantialized to Lasso and logistic regression models.

Keywords: Minibatch stochastic proximal point methods, Convex optimization, Smoothness, Excess risk, Uniform stability, Quadratic growth.

1. Introduction

We consider the following problem of regularized risk minimization over a closed convex subset $\mathcal{W} \subseteq \mathbb{R}^p$:

$$\min_{w\in\mathcal{W}} R(w) := R^\ell(w) + r(w), \quad \text{where} \quad R^\ell(w) := \mathbb{E}_{z\sim\mathcal{D}}[\ell(w;z)], \tag{1}$$

where $\ell : \mathcal{W} \times \mathcal{Z} \mapsto R^+$ is a non-negative convex loss function whose value $\ell(w;z)$ measures the loss of a hypothesis, parameterized by $w \in \mathcal{W}$, evaluated over a data sample $z \in \mathcal{Z}$, $\mathcal{D}$ represents a distribution over $\mathcal{Z}$, and $r : \mathcal{W} \mapsto \mathbb{R}^+$ is a data-independent non-negative convex function whose value $r(w)$ measures certain complexity of the hypothesis. We are particularly interested in the situation where the composite population risk R is strongly convex around its minimizers, though in this setting the terms R^ℓ and r are not necessarily required to be so simultaneously. For an instance, the ℓ_1-norm regularizer $r(w) = \mu\|w\|_1$ or its grouped variants are often used for sparse generalized linear models learning with quadratic or logistic loss functions (Negahban et al., 2012; Ravikumar et al., 2009; Van de Geer, 2008).

In statistical machine learning, it is usually assumed that the estimator only has access to, either as a batch training set or in an online/incremental manner, a collection $S = \{z_i\}_{i=1}^N$ of i.i.d. random data instances drawn from $\mathcal{D}$. The goal is to compute a stochastic estimator $\hat{w}_S$ based on the knowledge of S, hopefully that it generalizes well as a near minimizer of the population risk. More precisely, we aim at deriving a suitable law of large numbers, i.e., a sample size vanishing rate δ_N so that the *excess risk* at $\hat{w}_S$ satisfies $R(\hat{w}_S) - R^* \le \delta_N$ in expectation or with high probability over S where $R^* := \min_{w\in\mathcal{W}} R(w)$ represents the minimal value of composite risk.

In this work, inspired by the recent remarkable success of the stochastic proximal point (SPP) algorithms (Patrascu and Necoara, 2017; Asi and Duchi, 2019a,b; Davis and Drusvyatskiy, 2019) and their minibatch extensions (Wang et al., 2017b; Zhou et al., 2019; Asi et al., 2020), we provide a sharper generalization performance analysis for a class of minibatch SPP methods for solving the stochastic composite risk minimization problem (1).

1.1 Algorithm and Motivation of Study

Minibatch Stochastic Proximal Point Algorithm. Let $S_t = \{z_{i,t}\}_{i=1}^n$ be a minibatch of n i.i.d. samples drawn from distribution $\mathcal{D}$ at time instance $t \ge 1$ and denote

$$R_{S_t}(w) := \frac{1}{n}\sum_{i=1}^n \ell(w; z_{i,t}) + r(w)$$

as the regularized empirical risk over S_t. We consider the Minibatch Stochastic Proximal Point (M-SPP) algorithm, as outlined in Algorithm 1, for composite risk minimization based on a sequence of data minibatches $S = \{S_t\}_{t=1}^T$. The precision value ϵ_t in the algorithm quantifies the sub-optimality of w_t for solving the inner-loop regularized ERM over the minibatch S_t. The M-SPP algorithm is generic and it encompasses several existing SPP methods as special cases. For example in the extreme case when $n = 1$ and $\epsilon_t \equiv 0$ M-SPP reduces to a composite variant of the standard SPP method (Bertsekas, 2011), as formulated in (5). In general, the recursion update formulation (2) can be regarded as a

Algorithm 1: `Minibatch Stochastic Proximal Point (M-SPP)`

Input : Regularization modulus $\{\gamma_t\}_{t\geq 1}$.
Output: $\bar{w}_T$ as a weighted average of $\{w_t\}_{1\leq t\leq T}$.
Initialization *Specify a value of w_0. Typically $w_0 = 0$.*
for $t = 1, 2, ..., T$ **do**

Sample a minibatch $S_t := \{z_{i,t}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}^n$ and estimate w_t satisfying

$$F_t(w_t) \leq \min_{w\in\mathcal{W}} \left\{F_t(w) := R_{S_t}(w) + \frac{\gamma_t}{2}\|w - w_{t-1}\|^2\right\} + \epsilon_t, \tag{2}$$

where $R_{S_t}(w) := \frac{1}{n}\sum_{i=1}^n \ell(w; z_{i,t}) + r(w)$ and $\epsilon_t \geq 0$ measures the sub-optimality of estimation.

end

natural composite extension of the existing minibatch stochastic proximal point methods for statistical estimation (Wang et al., 2017b; Asi et al., 2020).

Prior results and limitations. The present study focuses on the generalization analysis of M-SPP for convex composite risk optimization. Recently, it has been shown by Asi et al. (2020, Theorem 2) that if the instantaneous loss functions are strongly convex with respect to the parameters, then the M-SPP algorithm converges at the rate of $\mathcal{O}\left(\frac{\log(nT)}{nT}\right)$. Prior to that, Wang et al. (2017b, Theorem 5) proved an $\mathcal{O}(\frac{1}{nT})$ rate for M-SPP when the individual loss functions are Lipschitz continuous and strongly convex. There results, among others for SPP (Patrascu and Necoara, 2017; Davis and Drusvyatskiy, 2019), commonly require that each instantaneous loss should be strongly convex which is too stringent to be fulfilled in high-dimensional or infinite spaces. For an instance, the quadratic loss $\ell(w; z) = \frac{1}{2}(w^\top x - y)^2$ over a feature-label pair $z = (x, y)$ is convex but in general not strongly convex, although the population risk $R^\ell(w) = \frac{1}{2}\mathbb{E}(y - w^\top x)^2$ is strongly convex provided that the covariance matrix of random feature x is non-degenerate. In the meanwhile, the Lipschitz-loss assumption made for the analysis (Wang et al., 2017b, Theorem 5) limits its applicability to smooth losses like quadratic loss, not to mention an interaction between Lipschitz continuity and strong convexity (Asi and Duchi, 2019b; Agarwal et al., 2012).

The above mentioned deficiencies of prior results motivate us to investigate the convergence behavior of M-SPP for composite risk minimization beyond the setting where each individual loss is strongly convex and Lipschitz continuous. From the perspective of optimization, smoothness is essential for establishing strong convergence guarantees for solving the inner-loop strongly convex risk minimization subproblems in (6), e.g., with variance reduced stochastic algorithms (Johnson and Zhang, 2013; Xiao and Zhang, 2014) or communication-efficient distributed optimization algorithms (Shamir et al., 2014; Zhang and Lin, 2015; Yuan and Li, 2020). Aiming at covering such an important yet less understood problem regime, we focus our study on analyzing the convergence behavior of M-SPP when the convex loss functions are smooth and the risk function exhibits quadratic growth property (see Assumption 2 for a formal definition).

1.2 Our Contributions and Main Results

The main contribution of the present work is a sharper non-asymptotic convergence analysis of the M-SPP algorithm through the lens of algorithmic stability theory (Bousquet and Elisseeff, 2002; Feldman and Vondrák, 2018). Let $W^* := \{w \in \mathcal{W} : R(w) = R^*\}$ be the set of minimizers of the composite population risk R. We are particularly interested in the regime where the loss function ℓ is convex and smooth but not necessarily Lipschitz (e.g., quadratic loss), while the population risk R satisfies the quadratic growth condition, i.e., $R(w) - R^* \geq \frac{\lambda}{2}\min_{w^* \in W^*}\|w - w^*\|^2$, $\forall w \in \mathcal{W}$, for some $\lambda > 0$, which can be satisfied by strongly convex objectives, and various other statistical estimation problems (see, e.g., Karimi et al., 2016; Drusvyatskiy and Lewis, 2018). For the family of L-smooth loss functions, with $\gamma_t = \mathcal{O}(\lambda\rho t)$ for an arbitrary scalar $\rho \in (0, 0.5]$ and $\epsilon_t \equiv 0$, we show in Theorem 1 that the excess risk at the weighted average output $\bar{w}_T = \frac{2}{T(T+1)}\sum_{t=1}^{T} t w_t$ is in expectation upper bounded by the following bound:

$$R(\bar{w}_T) - R^* \lesssim \frac{\rho\left[R(w_0) - R^*\right]}{T^2} + \frac{LR^*}{\rho\lambda nT}. \tag{3}$$

In this composite bound, the first *bias* component associated with initial gap $R(w_0) - R^*$ has a decaying rate $\mathcal{O}\left(\frac{1}{T^2}\right)$ and the second *variance* component associated with R^* converges at the rate of $\mathcal{O}\left(\frac{1}{\lambda nT}\right)$. The variance decaying rate actually matches the corresponding optimal rates of the SGD-type methods for strongly convex optimization (Rakhlin et al., 2012; Dieuleveut et al., 2017; Woodworth and Srebro, 2021). Also, such an $\mathcal{O}\left(\frac{1}{T^2} + \frac{1}{\lambda nT}\right)$ bounds matches those bounds for SPP (Davis and Drusvyatskiy, 2019) or M-SPP (Wang et al., 2017b) which are in contrast obtained under a substantially stronger assumption that each individual loss function should be strongly convex and Lipschitz as well. In the realizable or near realizable machine learning regimes where R^* equals to or approximates zero, the variance term in (3) would be sharper than those bounds of Wang et al. (2017b); Davis and Drusvyatskiy (2019). To our best knowledge, the bound in (3) for smooth and convex loss functions is *new* to the SPP-type methods. More generally for arbitrary convex risk functions, we present in Theorem 3 an $\mathcal{O}(\frac{1}{\sqrt{nT}})$ excess risk bound for exact M-SPP. Further, as shown in Theorem 4 and Theorem 5 that similar results can be extended to the inexact M-SPP given that the inner-loop sub-optimality is sufficiently small.

In the regime $T \ll n$ which is of special interest for off-line incremental learning with large data batches, setting a near-optimal value $\rho = \sqrt{\frac{T}{n\lambda}}$ in the excess risk bound (3) yields an $\mathcal{O}\left(\frac{1}{T\sqrt{\lambda nT}}\right)$ rate of convergence. This rate, however, is substantially slower than its $\mathcal{O}(\frac{1}{\lambda nT})$ counterpart available for the previous small-n-large-T setup. In order to address such a deficiency, we propose a two-phase variant of M-SSP (see Algorithm 2) to boost its performance in the small-T-large-n regime: in the first phase, M-SPP with sufficiently small minibatch-size is invoked over S_1 to obtain w_1, and then initialized by w_1 the second phase applies M-SPP to the rest minibatches. Then in Theorem 2 we show that the in-expectation excess risk at the output of the second phase can be accelerated to scale as

$$R(\bar{w}_T) - R^* \lesssim \frac{L^2(R(w_0) - R^*)}{\lambda^2 n^2 T^2} + \frac{LR^*}{\lambda nT}, \tag{4}$$

which holds regardless to the mutual strength of minibatch size n and iteration count T.

In addition to the above in-expectation risk bounds, we further derive a high-probability model estimation error bound of M-SPP based on algorithmic stability theory. Our deviation analysis is carried out over a sampling-without-replacement variant of M-SPP (see Algorithm 3). For population risk with quadratic growth property, up to an additive term on the inner-loop sub-optimality ϵ_t, we establish in Theorem 6 the following deviation bound on the estimation error $D(\bar{w}_T, W^*)$ that holds with probability at least $1-\delta$ over S while in expectation over the randomness of sampling:

$$D(\bar{w}_T, W^*) \lesssim \frac{\sqrt{L\log(1/\delta)}\log(T)}{\lambda\sqrt{nT}} + \sqrt{\frac{[R(w_0) - R^*]}{\lambda T^2} + \frac{LR^*}{\rho\lambda^2 nT}}.$$

When $T = \Omega(n)$, up to the logarithmic factors, this above bound matches (in terms of the total sample size $N = nT$) the known minimax lower bounds for statistical estimation even without computational limits (Tsybakov, 2008).

To highlight the core contribution of this work, the following three new insights into M-SPP make our results distinguished from the best known of SPP-type methods for convex optimization:

1. First and for most, the fast rates in (3) and (4) reveal the impact of noise level, as quantified by R^*, to convergence rate which has not been previously known for SPP-type methods. These bounds are valid for smooth losses and thus complement the previous ones for Lipschitz losses (Patrascu and Necoara, 2017; Wang et al., 2017b; Davis and Drusvyatskiy, 2019).

2. Second, the risk bounds in (3) and (4) are established under the quadratic growth condition of population risk. This is substantially weaker than the instantaneous-loss-wise strong convexity assumption commonly imposed by prior analysis to achieve the comparable rates for SPP-type methods (Toulis and Airoldi, 2017; Wang et al., 2017b; Asi et al., 2020).

3. Third, we provide a deviation analysis of M-SPP from the viewpoint of uniform algorithmic stability which to our best knowledge has not yet been addressed in the previous study on SPP-type methods.

While we provide some insights into the numerical aspects of M-SPP through an empirical study, this work is essentially a theoretical contribution.

1.3 Related Work

Our work is situated at the intersection of two lines of machine learning research: stochastic optimization and algorithmic stability theory, both of which have been actively studied with a vast body of beautiful and insightful theoretical results established in literature. We next incompletely review some representative work that are closely relevant to ours.

Stochastic optimization. Stemmed from the pioneering work of Robbins and Monro in 1951 (Robbins and Monro, 1951), stochastic gradient descent (SGD) methods have been extensively studied to approximately solve a simplified version of the problem (1) with $r \equiv 0$ (Bottou et al., 2018; Zhang, 2004; Nemirovski et al., 2009; Rakhlin et al., 2012).

For the composite formulation, a vast body of proximal SGD methods have been developed for efficient optimization in the presence of potentially non-smooth regularizers (Lan, 2012; Ghadimi and Lan, 2012; Duchi et al., 2010; Hu et al., 2009; Kulunchakov and Mairal, 2019). To handle the challenges associated with stepsize selection and numerical instability of SGD (Nemirovski et al., 2009; Bach and Moulines, 2011), a number of more sophisticated methods including implicit stochastic/online learning (Kulis and Bartlett, 2010; Toulis and Airoldi, 2017; Toulis et al., 2016; Crammer et al., 2006) and stochastic proximal point (SPP) methods (Bertsekas, 2011; Patrascu and Necoara, 2017; Asi and Duchi, 2019a,b; Davis and Drusvyatskiy, 2019) have recently been investigated for enhancing stability and adaptivity of stochastic (composite) optimization. For an example, in our considered composite optimization regime, the iteration procedure of vanilla SPP can be expressed as the following recursion form for $i \geq 1$:

$$\hat{w}_i^{\text{spp}} := \arg\min_{w \in \mathcal{W}} \ell(w; z_i) + r(w) + \frac{\gamma_i}{2}\|w - \hat{w}_{i-1}^{\text{spp}}\|^2, \tag{5}$$

where $z_i \sim \mathcal{D}$ is a random data sample, γ_i is a regularization modulus and $\|\cdot\|$ stands for the Euclidean norm. In contrast to standard SGD methods which are simple in per-iteration modeling but brittle to stepsize choice, the SPP methods are more accurate in objective approximation which leads to substantially improved stability to the choice of algorithm hyper-parameters while enjoying optimal guarantees on convergence (Asi and Duchi, 2019a,b).

An attractive feature of these above (proximal) stochastic optimization methods is that their convergence guarantees directly apply to the population risk and the minimax optimal rates of order $\mathcal{O}(\frac{1}{T})$ are achievable after T rounds of iteration for strongly convex problems (Agarwal et al., 2012; Nemirovski et al., 2009; Rakhlin et al., 2012). For large-scale machine learning, the improved memory efficiency is another practical argument in favor of stochastic over batch optimization methods. However, due to the sequential processing nature, the stochastic optimization methods tend to be less efficient for parallelization especially in distributed computing environment where excessive communication between nodes would be required for model update (Bottou et al., 2018).

Empirical risk minimization. At the opposite end of SGD-type and online learning, the following defined (composite) empirical risk minimization (ERM, a.k.a., M-estimation) is another popularly studied formulation for statistical learning (Lehmann and Casella, 2006):

$$\hat{w}_S^{\text{erm}} := \arg\min_{w \in \mathcal{W}} \left\{ R_S(w) := \frac{1}{N}\sum_{i=1}^{N} \ell(w; z_i) + r(w) \right\}.$$

Thanks to the finite-sum structure, a large body of randomized incremental algorithms with linear rates of convergence have been established for ERM including SVRG (Johnson and Zhang, 2013; Xiao and Zhang, 2014), SAGA (Defazio et al., 2014) and Katyusha (Allen-Zhu, 2017), to name a few. From the perspective of distributed computation, one intrinsic advantage of ERM over SGD-type methods lies in that it can better explore the statistical correlation among data samples for designing communication-efficient distributed optimization algorithms (Jaggi et al., 2014; Shamir et al., 2014; Lee et al., 2017; Zhang and Lin, 2015). Unlike stochastic optimization methods, the generalization performances of the batch or incremental algorithms are by nature controlled by that of ERM (Bottou and Bousquet, 2007) which has long been studied with a bunch of insightful results available (Vapnik, 1999; Bartlett et al., 2005; Srebro et al., 2010; Mei et al., 2018). Particularly for strongly convex risk functions, the $\mathcal{O}(\frac{1}{N})$ rate of convergence is possible for ERM (Bartlett et al., 2005; Koltchinskii, 2006; Zhang et al., 2017), though these fast rates are in general dimensionality-dependent for parametric learning models.

It has been recognized that SGD-type and ERM-type approaches cannot dominate each other in terms of generalization, runtime, storage and parallelization efficiency. This motivates a recent trend of trying to propose the so called stochastic model-based methods that can achieve the best of two worlds. Among others, a popular paradigm for such a purpose of combination is *minibatch proximal update* which in each iteration updates the model via (approximately) solving a local ERM over a stochastic minibatch (Li et al., 2014; Wang et al., 2017b; Asi et al., 2020; Deng and Gao, 2021). This strategy can be viewed as a minibatch extension to the SPP algorithm and it has been shown to attain a substantially improved trade-off between computation, communication and memory efficiency for large-scale distributed machine learning (Li et al., 2014; Wang et al., 2017a). Alternatively, a number of online extensions of the incremental finite-sum algorithms, such as streaming SVRG (Frostig et al., 2015) and streaming SAGA (Jothimurugesan et al., 2018), have been proposed for stochastic optimization with competitive guarantees to ERM but at lower cost of computation.

Algorithmic stability and generalization. Since the seminal work of Bousquet and Elisseeff (Bousquet and Elisseeff, 2002), algorithmic stability has been extensively studied with remarkable success achieved in establishing generalization bounds for strongly convex ERM estimators (Zhang, 2003; Mukherjee et al., 2006; Shalev-Shwartz et al., 2010). Particularly, the state-of-the-art risk bounds of strongly convex ERM are offered by approaches based on the notion of uniform stability (Feldman and Vondrák, 2018, 2019; Bousquet et al., 2020; Klochkov and Zhivotovskiy, 2021). It was shown by Hardt et al. (2016) that the solution obtained via (stochastic) gradient descent is stable for smooth convex or non-convex loss functions. For non-smooth convex losses, the stability induced generalization bounds of SGD have been established in expectation (Lei and Ying, 2020) or deviation (Bassily et al., 2020). For learning with sparsity, algorithmic stability theory has been employed to derive the generalization bounds of the popularly used iterative hard thresholding (IHT) algorithm (Yuan and Li, 2022). Through the lens of uniform algorithmic stability, convergence rates of M-SPP have been studied for convex (Wang et al., 2017b) and weakly convex (Deng and Gao, 2021) Lipschitz losses. While sharing a similar spirit to Wang et al. (2017b); Deng and Gao (2021), our analysis customized for smooth convex loss functions is considerably different and the resultant fast rates are of special interest in low-noise statistical settings (Srebro et al., 2010).

1.4 Notation and Paper Organization

Notation. The key quantities and notations frequently used in our analysis are summarized in Table 1.

Notation	Definition
n	minibatch size
T	round of iteration
N	total number of samples visited, i.e., $N = nT$
f	hypothesis
ℓ	loss function
r	regularization term
R^ℓ	population risk: $R^\ell(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_w(x), y)]$
R	composite population risk: $R(w) := R^\ell(w) + r(w)$
R^*	the optimal value of composite risk, i.e., $R^* := \min_{w\in\mathcal{W}} R(w)$
W^*	the optimal solution set of composite risk, i.e., $W^* := \arg\min_{w\in\mathcal{W}} R(w)$
S_t	data minibatch at time instance t
S_I	The union of data minibatch over I, i.e., $S_I := \{S_t\}_{t\in I}$
R^ℓ_S	empirical risk over S, i.e., $R^\ell_S(w) := \frac{1}{\lvert S\rvert}\sum_{(x,y)\in S} \ell(f_w(x, y)$
R_S	composite empirical risk over S, i.e., $R_S(w) := R^\ell_S(w) + r(w)$
ϵ_t	precision of minibatch risk minimization at time instance t
$\\|w\\|_1$	ℓ_1-norm of a vector w, i.e., $\\|w\\|_1 := \sum_i \lvert[w]_i\rvert$
$\\|w\\|$	Euclidean norm of a vector w
$D(w, W^*)$	the distance from w to W^*, i.e., $D(w, W^*) = \min_{w^*\in W^*} \\|w - w^*\\|$
$[T]$	$[T] := \{1, ..., T\}$
$\mathbf{1}_{\{C\}}$	the indicator function of the condition C

Table 1: Table of notation.

Organization. The paper proceeds with the material organized as follows: In Section 2, we analyze the risk bounds of exact M-SPP with convex and smooth loss functions and present a two-phase variant to further improve convergence performance. In Section 3, we extend our analysis to the more realistic setting where inexact M-SPP iteration is allowed. In Section 4, we study the high-probability estimation error bounds of M-SPP. A comprehensive comparison to some closely relevant results is highlighted in Section 5. The numerical study for theory verification and algorithm evaluation is provided in Section 6. The concluding remarks are made in Section 7. All the proofs of main results and some additional results on the iteration stability of M-SPP are relegated to appendix.

2. A Sharper Analysis of M-SPP for Smooth Loss

In this section, we analyze the convergence rate of M-SPP for smooth and convex loss functions using the tools developed in algorithmic stability theory. In what follows, for the sake of notation simplicity and presentation clarity of core ideas, we assume for the time being that the inner-loop composite ERM in the M-SPP iteration procedure (2) has been solved exactly with $\epsilon_t \equiv 0$, i.e.,

$$w_t = \arg\min_{w \in \mathcal{W}} \left\{ F_t(w) := R_{S_t}(w) + \frac{\gamma_t}{2} \|w - w_{t-1}\|^2 \right\}. \tag{6}$$

A full convergence analysis for the inexact variant (i.e., $\epsilon_t > 0$) will be presented in the Section 3 via a slightly more involved perturbation analysis.

2.1 Basic Assumptions

We begin by introducing some basic assumptions that will be used in the analysis to follow. We say a differentiable function $g : \mathcal{W} \mapsto \mathbb{R}$ is L-smooth if $\forall s, t \in \mathbb{R}$,

$$\left| g(w) - g(w') - \langle \nabla g(w), w - w' \rangle \right| \leq \frac{L}{2} |w - w'|^2.$$

As formally stated in the following assumption, we suppose that the individual loss functions are convex and L-smooth which can be satisfied, e.g., by the quadratic loss (for regression) and the logistic loss (for prediction).

Assumption 1 *The loss function ℓ is convex and L-smooth with respect to its first argument. Also, we assume that the regularization term r is convex over $\mathcal{W}$.*

Let us define $D(w, W^*) := \min_{w^* \in W^*} \|w - w^*\|$ as the distance from w to the set W^* of minimizers. The next assumption requires that the population risk has the characterization of quadratic growth away from the set of minimizers (Anitescu, 2000; Karimi et al., 2016).

Assumption 2 *The population risk function R satisfies $R(w) \geq R^* + \frac{\lambda}{2} D^2(w, W^*), \forall w \in \mathcal{W}$ for some $\lambda > 0$.*

Clearly, the quadratic growth property can be implied by the traditional strong convexity condition (around the minimizers) which is satisfied by a number of popular learning models including linear and logistic regression, generalized linear models, smoothed Huber losses, and various other statistical estimation problems. Particularly, Assumption 2 holds when R^ℓ is strongly convex and r is convex. Notice that for risk functions with quadratic growth property, the prior analysis of M-SPP for Lipschitz losses (Wang et al., 2017b) is not generally applicable because Assumption 2 implies that the Lipschitz constant of loss could be arbitrarily large if the infinite distance $\min_{w^* \in W^*} \|w - w^*\| \to \infty$ is allowed.

2.2 Main Results

The following theorem is our main result on the in-expectation rate of convergence of the exact M-SPP with smooth loss and quadratic growth population risk functions. Recall that $N = nT$ is the total number of data points visited up to the iteration counter T.

Theorem 1 *Suppose that Assumptions 1 and 2 hold. Consider $\epsilon_t \equiv 0$ and the weighted average output $\bar{w}_T = \frac{2}{T(T+1)}\sum_{t=1}^{T} t w_t$ in Algorithm 1. Let $\rho \in (0, 0.5]$ be an arbitrary scalar.*

(a) Suppose that $n \geq \frac{64L}{\lambda\rho}$. Set $\gamma_t = \frac{\lambda\rho t}{4}$ for $t \geq 1$. Then for any $T \geq 1$,

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \frac{4\rho\left[R(w_0) - R^*\right]}{T^2} + \frac{2^9 L}{\lambda\rho n T} R^*.$$

(b) Set $\gamma_t = \frac{\lambda\rho t}{4} + \frac{16L}{n}$ for $t \geq 1$. Then for any $T \geq 1$,

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \left(\frac{4\rho}{T^2} + \frac{2^8 L}{\lambda n T}\right)\left[R(w_0) - R^*\right] + \left(\frac{2^{16} L^2}{\lambda^2\rho^2 n^2 T} + \frac{2^9 L}{\lambda\rho n T}\right) R^*,$$

Proof The proof technique is inspired by the uniform stability arguments developed by Wang et al. (2017b) for Lipschitz and instance-wise strongly convex loss, with several new elements along developed for handling smooth loss and quadratic growth of risk function. As a non-trivial ingredient, we show that it is possible to extend those stability arguments to smooth losses in view of a classical result from (Srebro et al., 2010, Lemma 2.1) that allows the derivative of a smooth loss to be bounded in terms of its function value. See Appendix A.1 for a full proof of this result. ■

A few remarks on Theorem 1 are in order.

Remark 1 *In Part (a), the minibatch size is required to be sufficiently large. In this setting, the excess risk bound consists of two components: the first* bias *component associated with initial gap $R(w_0) - R^*$ has a decaying rate $\mathcal{O}(\frac{1}{T^2})$ and the second* variance *component associated with R^* vanishes at a dominate rate of $\mathcal{O}(\frac{1}{\lambda n T})$. The variance term shows that the convergence rate can be improved in the low-noise settings where the factor of R^* is relatively small. Extremely in the separable case with $R^* = 0$, the excess risk bound of Theorem 1 would scale as fast as $\mathcal{O}(\frac{1}{T^2})$.*

Remark 2 *One disadvantage of the result in Part (a) lie in that the minibatch size is required to be sufficiently larger than the condition number of the population risk R. Contrastively, the excess risk bound in Part (b) holds for arbitrary minibatch sizes. The cost, however, is a relatively slower bias decaying term $\mathcal{O}(\frac{1}{T^2} + \frac{1}{\lambda n T})$ which is dominated by $\mathcal{O}(\frac{1}{\lambda n T})$ in the case of $T \gg n$.*

Remark 3 *Let $N = nT$ be the total number of data points accessed. When $T \gg n$, the $\mathcal{O}(\frac{1}{N})$ dominant rates in Theorem 1 match those prior ones for SPP-type methods (Wang et al., 2017b; Davis and Drusvyatskiy, 2019) which are, however, obtained under the assumption that each individual loss function should be Lipschitz continuous and strongly convex. In comparison to the $\mathcal{O}(\frac{1}{N})$ rate established for SGD with smooth loss (Lei and Ying, 2020, Theorem 12), our result in Theorem 1 is stronger and less stringent in the following senses: 1) our bound shows explicitly the impact of R^* which usually represents the noise level of model, and 2) we only require the population risk to have quadratic growth property while the bound of Lei and Ying (2020, Theorem 12) not only requires the loss to be Lipschitz but also assumes the empirical risk to be strongly convex.*

Let us further look into the choice of the scalar ρ in Theorem 1. We focus the discussion on the part (a) and similar observations apply to the part (b). We distinguish the discussion in the following two complementary cases regarding the mutual strength of minibatch-size n and round of iteration T:

- **Case I: Small-n-large-T.** Suppose that $n = \mathcal{O}(1)$ and $T \to \infty$ is allowed. In this case, simply setting $\rho = 0.5$ yields the convergence rate of order $\mathcal{O}\left(\frac{1}{T^2} + \frac{1}{\lambda n T}\right)$ in the part (a).
- **Case II: Small-T-large-n.** Suppose that $T = \mathcal{O}(1)$ and $n \to \infty$ is allowed. In this setup, given that $n \geq \frac{4T}{\lambda}$, then with a roughly optimal choice $\rho = \sqrt{\frac{T}{n\lambda}}$ the excess risk bound in Theorem 1(a) will be of the order $\mathcal{O}\left(\frac{1}{T\sqrt{\lambda n T}}\right)$, which is substantially slower than the previous fast rate in Case I. This is intuitive because M-SPP with large minibatches behaves more like regularized ERM which is known to exhibit slow rate of convergence even for strongly convex problems (Shalev-Shwartz et al., 2010; Srebro et al., 2010). Nevertheless, such a small-T-large-n setup is of special interest for off-line incremental learning with large minibatches and distributed statistical learning (Li et al., 2014; Wang et al., 2017b; You et al., 2020). We will address this critical case in the next subsection.

2.3 A Two-Phase M-SPP Method

Algorithm 2: `Two-Phase M-SPP (M-SPP-TP)`

Input : Dataset $S = \{S_t\}_{t=1}^T$ in which $S_t := \{z_{i,t}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}^n$, regularization modulus $\{\gamma_t > 0\}_{t\in[T]}$.
Output: $\bar{w}_T$ as a weighted average of $\{w_t\}_{2\leq t\leq T}$.
Initialization *Specify a value of* w_0. *Typically* $w_0 = 0$.
`/* Phase-I */`
Divide sample S_1 into disjoint minibatches of equal size m;
Run M-SPP over these minibatches to obtain the output w_1;
`/* Phase-II */`
Initialized with w_1, run M-SPP over data minibatches $\{S_t\}_{2\leq t\leq T}$ with $\{\gamma_t\}_{2\leq t\leq T}$ to obtain the sequence $\{w_t\}_{2\leq t\leq T}$.

To remedy the deficiencies mentioned in the previous discussion, we propose a two-phase variant of M-SSP, as outlined in Algorithm 2, to boost its performance in the small-T-large-n regimes. The procedure can be regarded as sort of a restarting argument (Nemirovskii and Nesterov, 1985; Renegar and Grimmer, 2021; Zhou et al., 2022) for M-SPP. More specifically, the Phase-I serves as an initialization step that invokes M-SPP to a uniform division of S_1 with minibatch size m to obtain w_1. Then starting from w_1, the Phase-II just invokes M-SPP to the consequent large minibatches $\{S_t\}_{t\geq 2}$ which is suitable for large-scale parallelization if applicable. The following theorem is a consequence of Theorem 1 to such a two-phase M-SPP procedure.

Theorem 2 *Suppose that Assumptions 1 and 2 hold. Consider* $\epsilon_t \equiv 0$ *for implementing M-SPP in both Phase-I and Phase-II of Algorithm 2. Consider the weighted average output* $\bar{w}_T = \frac{2}{(T-1)(T+2)} \sum_{t=2}^{T} t w_t$ *in Phase-II.*

(a) Suppose that $n \geq \frac{128L}{\lambda}$*. Set* $m = \frac{128L}{\lambda}$ *in Phase-I and* $\gamma_t = \frac{\lambda t}{8}$ *for implementing M-SPP in both Phase-I and Phase II. Then for any* $T \geq 2$*,* $\bar{w}_T$ *satisfies*

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{L^2 \left[R(w_0) - R^*\right]}{\lambda^2 n^2 T^2} + \frac{L}{\lambda n T} R^*.$$

(b) Set $m = \mathcal{O}(1)$ *in Phase-I and* $\gamma_t = \frac{\lambda t}{8} + \frac{16L}{n}$ *for implementing M-SPP in both Phase-I and Phase-II. Then for any* $T \geq 2$*,* $\bar{w}_T$ *satisfies*

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{L^2 \left[R(w_0) - R^*\right]}{\lambda^2 n T} + \frac{L^3}{\lambda^3 n T} R^*.$$

Proof See Appendix A.2 for a proof of this result. ■

Remark 4 *The part (a) of Theorem 2 suggests that when the minibatch size is sufficiently large, the excess risk bound of two-phase M-SPP has a bias decaying term of scale* $\mathcal{O}\left(\frac{1}{n^2T^2}\right)$ *and a variance term that decays at the rate of* $\mathcal{O}(\frac{1}{nT})$*. The rate is valid even when the scales of* T *relatively small, and thus is stronger than the* $\mathcal{O}\left(\frac{1}{T\sqrt{nT}}\right)$ *rate implied by Theorem 1 for the vanilla M-SPP in the small-*T*-large-*n *regime. It is worth to mention that both the bias and variance components in our bound for M-SPP are faster than those derived for strongly convex ERM (Srebro et al., 2010).*

Remark 5 *The excess risk bound in Part (b) of Theorem 2 is valid for arbitrary minibatch sizes, but at the cost of a relatively slower* $\mathcal{O}(\frac{1}{nT})$ *bias decaying rate.*

2.4 Results for Arbitrary Convex Risks

We further analyze the proposed M-SPP algorithm when the loss function ℓ is convex and smooth, but without requiring that the composite risk R has quadratic growth property. The following is our main result in such a generic setting.

Theorem 3 *Suppose that Assumption 1 holds. Set* $\gamma_t \equiv \gamma \geq \frac{16L}{n}$*. Let* $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ *be the average output of Algorithm 1. Then*

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{\gamma}{T} D^2(w_0, W^*) + \frac{L}{\gamma n} R^*.$$

Particularly for $\gamma = \sqrt{\frac{T}{n}} + \frac{16L}{n}$*, it holds that*

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right) D^2(w_0, W^*) + \frac{L}{\sqrt{nT}} R^*.$$

Proof See Appendix A.3 for a proof of this result. ■

Remark 6 *The first bound of Theorem 3 implies that for any $\epsilon \in (0,1)$, by setting $\gamma = \mathcal{O}\left(\frac{L}{\epsilon n}\right)$, $R(\bar{w}_t)$ converges to $(1+\epsilon)R^*$ at the rate of $\mathcal{O}(\frac{1}{nT\epsilon})$. This bound matches the results of Lei and Ying (2020, Theorem 4) for smooth SGD method. The second bound of Theorem 3 further shows that by setting $\gamma = \mathcal{O}(\sqrt{\frac{T}{n}} + \frac{L}{n})$, the excess risk of $\bar{w}_T$ decays at the rate of $\mathcal{O}(\frac{1}{\sqrt{nT}})$ for both bias and variance terms, which matches in order the corresponding bound derived for Lipschitz-loss (Wang et al., 2017b, Theorem 4). To our knowledge, such a bias-variance composite rate of convergence is new for SPP-type methods with convex and smooth loss functions.*

Analogous to the robustness analysis of SPP (Asi and Duchi, 2019a,b), we have also analyzed the iteration stability of M-SPP for convex losses with respect to the choice of regularization modulus γ_t. The corresponding results, which can be found in Appendix A.4, confirm that the choice of γ_t is insensitive to the gradient scale of loss functions for generating a non-divergent sequence of estimation errors.

3. Perturbation Analysis for Inexact M-SPP

In the preceding section, we have analyzed the convergence rates of M-SPP under the assumption that the inner-loop proximal ERM subproblems constructed in its iteration procedure (2) are solved exactly, i.e., $\epsilon_t \equiv 0$. To make our analysis more practical, we further provide in this section a perturbation analysis of M-SPP when the inner-loop proximal ERM subproblems are only required to be solved approximately up to certain precision $\epsilon_t > 0$. As a starting point, we need to impose the following Lipschitz continuity assumption on the regularization term r.

Assumption 3 *The regularization term r is Lipschitz continuous over $\mathcal{W}$, i.e., $|r(w) - r(w')| \leq G\|w - w'\|$, $\forall w, w' \in \mathcal{W}$.*

For example, the ℓ_1-norm regularizer $r(w) = \mu\|w\|_1$ satisfies this assumption with respect to Euclidean norm as $|r(w) - r(w')| = \mu|\|w\|_1 - \|w'\|_1| \leq \mu\|w - w'\|_1 \leq \mu\sqrt{p}\|w - w'\|$.

The following theorem is our main result on the rate of convergence of the inexact M-SPP for composite stochastic convex optimization with smooth losses.

Theorem 4 *Suppose that Assumptions 1, 2 and 3 hold. Let $\rho \in (0, 1/4]$ be an arbitrary scalar and set $\gamma_t = \frac{\lambda\rho t}{4}$. Suppose that $n \geq \frac{76L}{\lambda\rho}$. Assume that $\epsilon_t \leq \frac{\epsilon}{nt^4}$ for some $\epsilon \in [0,1]$. Then for any $T \geq 1$, the weighted average output $\bar{w}_t = \frac{2}{T(T+1)}\sum_{t=1}^{T} t w_t$ of Algorithm 1 satisfies*

$$\mathbb{E}\left[R(\bar{w}_t) - R^*\right] \lesssim \frac{\rho}{T^2}(R(w_0) - R^*) + \frac{L}{\lambda\rho n T}R^* + \frac{\sqrt{\epsilon}}{T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right).$$

Proof See Appendix B.1 for a proof of this result. We would like to highlight that our perturbation analysis for smooth loss is considerably different from that of Wang et al. (2017b) developed for Lipschitz loss. This is mainly because in the smooth loss case, the change of loss could no longer be upper bounded by the change of prediction, and thus we need to make a more careful treatment to the perturbation caused by inexact minimization of the regularized minibatch empirical risk. ■

We provide in order a few remarks on Theorem 4.

Remark 7 *Theorem 4 suggests that the excess risk bound of exact M-SPP in the part (a) of Theorem 1 can be extended to its inexact version, provided that the inner-loop minibatch ERMs (2) are solved to sufficient accuracy, say,* $\epsilon_t \le \mathcal{O}\left(\frac{1}{nt^4}\right)$. *Similarly, the result in the part (b) of Theorem 1 for arbitrary minibatch sizes can also be extended to the inexact M-SPP, which is omitted to avoid redundancy. Since the inner-loop minibatch ERMs are strongly convex and the loss functions are smooth, in average the desired accuracy can be attained in logarithmic time* $\mathcal{O}\left(\log\left(\frac{1}{\epsilon_t}\right)\right)$ *via variance-reduced SGD methods (Xiao and Zhang, 2014).*

Remark 8 *Analogous to the discussions at the end of Section 2.2, by specifying the choice of* ρ *we can derive a direct consequent result of Theorem 4 which more explicitly shows the rate of convergence with respect to* $N = nT$. *Also for the two-phase M-SPP, in view of Theorem 4 we can show that the bound in Theorem 2 can be extended to the inexact setting if the minibatch optimization is sufficiently accurate. These extensions are more or less straightforward and thus are omitted.*

In the following theorem, we provide an excess risk bound for the inexact M-SPP when the composite risk R is convex but not necessarily has quadratic growth property.

Theorem 5 *Suppose that Assumptions 1 and 3 hold. Set* $\gamma_t \equiv \gamma \ge \frac{19L}{n}$. *Assume that* $\epsilon_t \le \min\left\{\frac{\epsilon}{n^2t^5}, \frac{2G^2}{9n^2\gamma}\right\}$ *for some* $\epsilon \in [0,1]$. *Then the average output* $\bar{w}_T = \frac{1}{T}\sum_{t=1}^T w_t$ *of Algorithm 1 satisfies*

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{\gamma}{T}D^2(w_0, W^*) + \frac{L}{\gamma n}R^* + \left(\frac{L}{\gamma n} + \frac{\gamma}{LnT} + \frac{G}{\sqrt{\gamma}nT}\right)\sqrt{\epsilon}.$$

Particularly for $\gamma = \sqrt{\frac{T}{n}} + \frac{19L}{n}$, *it holds that*

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right)D^2(w_0, W^*) + \frac{L}{\sqrt{nT}}R^* + \left(\frac{L+G}{\sqrt{nT}} + \frac{1}{nT}\right)\sqrt{\epsilon}.$$

Proof See Appendix B.2 for a proof of this result. ■

Remark 9 *Theorem 5 confirms that the excess risk bounds established in Theorem 3 for exact M-SPP are tolerant to sufficiently small sub-optimality* $\epsilon_t \le \mathcal{O}(\frac{1}{n^2t^5})$ *of minibatch proximal ERM subproblems.*

4. Performance Guarantees with High Probability

In the previous two sections, we have analyzed the excess risk bounds of M-SPP in expectation. In this section, we move on to study high-probability guarantees of M-SPP with respect to the randomness of training data, still under the notion of algorithmic stability. To this end, we first introduce a variant of M-SPP which carries out the proximal point update via sampling without replacement over the given data minibatches. We then show that the output of the proposed algorithm is uniformly stable in expectation over the randomness of sampling. As a main result of this section, for strongly convex population risk, we establish a near-optimal high probability (with respect to data) bound on the estimation error $\|\bar{w}_t - w^*\|$ that holds in expectation over the randomness of inner-data sampling. Additionally, we provide a high-probability generalization bound for arbitrary convex loss.

4.1 Sampling Without Replacement M-SPP

Algorithm 3: `Sampling Without Replacement M-SPP (M-SPP-SWoR)`

Input : Dataset $S = \{S_t\}_{t=1}^T$ in which $S_t := \{z_{i,t}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}^n$, regularization modulus $\{\gamma_t > 0\}_{t\in[T]}$.

Output: $\bar{w}_T$ as a weighted average of $\{w_t\}_{1\le t\le T}$..

Initialization *Specify a value of* w_0. *Typically* $w_0 = 0$.

for $t = 1, 2, ..., T$ **do**

Uniformly randomly sample an index $\xi_t \in [T]$ without replacement.

Estimate w_t satisfying

$$F_t(w_t) \le \min_{w\in\mathcal{W}} \left\{ F_t(w) := R_{S_{\xi_t}}(w) + \frac{\gamma_t}{2}\|w - w_{t-1}\|^2 \right\} + \epsilon_t, \tag{7}$$

where $\epsilon_t \ge 0$ measures the sub-optimality.

end

Let us consider the M-SPP-SWoR (M-SPP via Sampling Without Replacement) procedure as outlined in Algorithm 3. Given a set S of T data minibatches, at each iteration, the algorithm uniformly randomly samples one minibatch from S without replacement for proximal update. After T rounds of iteration, all the minibatches are used to update the model. Since this procedure is merely a random shuffling variant of M-SPP as presented in Algorithm 1, we can see that all the in-expectation bounds established in the previous sections for M-SPP directly transfer to M-SPP-SWoR under any implementation of shuffling. As we will show shortly in the next subsection that such a random shuffling scheme is beneficial for boosting the on-average algorithmic stability of M-SPP which then leads to strong high-probability guarantees for M-SPP-SWoR.

4.2 A Uniform Stability Analysis

Let $S = \{S_t\}_{t\in[T]}$ and $S' = \{S'_t\}_{t\in[T]}$ be two sets of data minibatches. We denote by $S_t \doteq S'_t$ if S_t and S'_t differ in a single data point, and by $S \doteq S'$ if S and S' differ in a single minibatch and a single data point in that minibatch. We introduce the following concept of uniform stability of M-SPP which substantializes the concept of uniform algorithmic stability that serves as a powerful tool for analyzing generalization bounds of statistical estimators and their learning algorithms (Bousquet and Elisseeff, 2002; Hardt et al., 2016; Feldman and Vondrák, 2019).

Definition 1 (Uniform Stability of M-SPP) *The M-SPP algorithm is said to be ϱ-uniformly stable with respect to a mapping $h : \mathcal{W} \mapsto \mathbb{R}^q$ if $\|h(\bar{w}_T) - h(\bar{w}'_T)\| \le \varrho$ for any pair of data sets $S \doteq S'$.*

The following result gives a uniform stability (with respect to identical mapping) bound of the vanilla M-SPP (Algorithm 1) that holds deterministically, and a corresponding bound for M-SPP-SWoR (Algorithm 3) that holds in expectation over the randomness of minibatch sampling.

Proposition 1 *Suppose that Assumption 1 holds and the loss function is bounded such that $0 \le \ell(y', y) \le M$ for all y, y'. Let $S = \{S_t\}_{t\in[T]}$ and $S' = \{S'_t\}_{t\in[T]}$ be two sets of data minibatches satisfying $S \doteq S'$. Then*

(a) The weighted average output $\bar{w}_T$ and $\bar{w}'_T$ respectively generated by M-SPP (Algorithm 1) over S and S' satisfy

$$\sup_{S,S'} \|\bar{w}_T - \bar{w}'_T\| \le \frac{4\sqrt{2LM}}{n \min_{t\in[T]} \gamma_t} + \sum_{t=1}^{T} 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}.$$

(b) The weighted average output $\bar{w}_T$ and $\bar{w}'_T$ respectively generated by M-SPP-SWoR (Algorithm 3) over S and S' satisfy

$$\sup_{S,S'} \mathbb{E}_{\xi_{[T]}} \left[\|\bar{w}_T - \bar{w}'_T\|\right] \le \sum_{t=1}^{T} \left\{ \frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}} \right\}.$$

Proof See Appendix C.1 for a proof. ■

Remark 10 *Suppose that the sub-optimality $\{\epsilon_t\}_{t\in[T]}$ are sufficiently small. If setting $\gamma_t = \mathcal{O}(t)$ as used for population risks with quadratic growth property, then Proposition 1 shows that M-SPP is $\mathcal{O}\left(\frac{1}{n}\right)$-uniformly stable, while in expectation over the randomness of without-replacement sampling, M-SPP-SWoR has an much improved uniform stability parameter scaling as $\mathcal{O}\left(\frac{\log(T)}{nT}\right)$. If setting $\gamma_t \equiv \sqrt{\frac{T}{n}}$ as used for generic convex loss, then M-SPP will be $\mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$-uniformly stable while M-SPP-SWoR has an identical uniform stability parameter in expectation over sampling.*

In the following theorem, based on the uniform stability bounds in Proposition 1, we derive an upper bound on the estimation error $D(\bar{w}_T, W^*)$ of M-SPP-SWoR that holds with high probability over data distribution while in expectation over randomly sampling the minibatches for update.

Theorem 6 *Suppose that Assumptions 1, 2, 3 hold and the loss function ℓ is bounded in the interval $(0, M]$. Let $\rho \in (0, 1/4]$ be an arbitrary scalar and set $\gamma_t = \frac{\lambda \rho t}{4}$. Suppose that $n \geq \frac{76L}{\lambda\rho}$. Assume that $\epsilon_t \leq \min\left\{\frac{\epsilon}{nt^4}, \frac{LM}{\lambda\rho n^2 T^2 t}\right\}$ for some $\epsilon \in [0, 1]$. Then with probability at least $1 - \delta$ over S, the weighted average output $\bar{w}_T$ of M-SPP-SWoR (Algorithm 3) satisfies*

$$
\begin{aligned}
&\mathbb{E}_{\xi_{[T]}}\left[D(\bar{w}_T, W^*)\right] \\
\lesssim& \frac{\sqrt{LM\log(1/\delta)}\log(T)}{\lambda\rho\sqrt{nT}} + \sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2 \rho n T} R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)}.
\end{aligned}
$$

Proof See Appendix C.2 for a proof of this result. ■

Remark 11 *We comment on the optimality of the bound in Theorem 6. Consider $\rho = \mathcal{O}(1)$. The first term of scale $\mathcal{O}\big(\frac{\sqrt{\log(1/\delta)}\log(T)}{\sqrt{nT}}\big)$ represents the overhead of getting generalization with high probability over data. The second term matches the corresponding in-expectation estimation error bound in Theorem 4, which matches the known optimal rates for strongly convex SGD (Rakhlin et al., 2012; Dieuleveut et al., 2017). In view of the minimax lower bounds for statistical estimation (Tsybakov, 2008), the estimation error bound established in Theorem 6 is near-optimal for strongly convex risk minimization.*

Finally, we provide a high-probability generalization bound of M-SPP for arbitrary convex population risk functions.

Theorem 7 *Suppose that Assumptions 1 and 3 hold and the loss function ℓ is bounded in the interval $[0, M]$. Set $\gamma_t \equiv \sqrt{\frac{T}{n}}$. Assume that $\epsilon_t \leq \frac{LM}{4nT^2\sqrt{nT}}$. Then with probability at least $1 - \delta$ over S, the average output $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ of M-SPP (Algorithm 1) satisfies*

$$
|R(\bar{w}_T) - R_S(\bar{w}_T)| \lesssim \frac{(LM + G\sqrt{LM})\log(N)\log(1/\delta)}{\sqrt{nT}} + M\sqrt{\frac{\log(1/\delta)}{nT}}.
$$

Proof See Appendix C.3 for a proof of this result. ■

We remark in passing that using similar uniform stability argument, the high-probability generalization bound in Theorem 7 can be shown to hold for convex and non-smooth loss functions as well. We omit the detailed analysis as it is out of the scope of this paper focusing on smooth losses.

5. Comparison against Prior Methods

Comparison with M-SPP and SPP methods. The M-SPP algorithm considered in this article is a minibatch extension of the SPP methods. The convergence analysis of SPP has received recent wide attention in stochastic optimization community. Specially for finite-sum optimization over N data points, an incremental SPP method was proposed and analyzed in (Bertsekas, 2011). For learning with linear prediction models and strongly convex Lipschitz-loss, (Toulis et al., 2016) established a set of $\mathcal{O}(\frac{1}{N^\gamma})$ rates of convergence for SPP with suitable $\gamma \in (0.5, 1]$, where N is the iteration counter. For arbitrary convex loss functions, the non-asymptotic convergence performance of SPP was studied with $\mathcal{O}(\frac{1}{\sqrt{N}})$ rate obtained for Lipschitz losses (Patrascu and Necoara, 2017; Davis and Drusvyatskiy, 2019), $\mathcal{O}(\frac{1}{N})$ for strongly convex and Lipschitz (Davis and Drusvyatskiy, 2019) or smooth (Patrascu and Necoara, 2017) losses, or $\mathcal{O}\left(\frac{\log(N)}{N}\right)$ rate for strongly convex non-smooth losses (Asi and Duchi, 2019b). Recently, it has been shown that the $\mathcal{O}\left(\frac{\log(N)}{N}\right)$ rate also extends to M-SPP with strongly convex losses (Asi et al., 2020). The asymptotic and non-asymptotic behaviors of SPP for weakly convex losses (e.g., composite of convex loss with smooth map) have been studied for stochastic optimization with (Duchi and Ruan, 2018) or without (Davis and Drusvyatskiy, 2019) composite structures. Among others, our work is most closely related to the minibatch proximal update method developed for communication-efficient distributed optimization (Wang et al., 2017b). Similarly from the viewpoint of algorithmic stability, the $\mathcal{O}(\frac{1}{N^\gamma})$ rates were established for that method for Lipschitz-loss with arbitrary convexity ($\gamma = 0.5$) or strong convexity ($\gamma = 1$). In comparison to these prior results, our convergence results for M-SPP are new in the following aspects:

- The convergence rates are derived for smooth losses and they explicitly show the impact of noise level of a statistical model, as encoded in R^*, to convergence performance which has not been previously known for SPP-type methods.
- The $\mathcal{O}(N^{-1})$ fast rate attained in this article is valid for population risks with quadratic growth property, without requiring each instantaneous loss to be strongly convex.
- We provide a near-optimal model estimation error bound of a sampling-without-replacement variant of M-SPP that holds with high probability over the randomness of data while in expectation over the randomness of sampling.

Comparison with SGD and ERM. Similar to those in Theorem 1 and Theorem 3, the bias-variance composite rates have been known for accelerated SGD for least squares regression (Dieuleveut et al., 2017), or minibatch SGD (M-SGD) for generic convex and smooth learning problems (Woodworth and Srebro, 2021). While the results are of similar flavor, we came to the path in a distinct algorithmic framework using quite different proof techniques. Particularly, in contrast to Woodworth and Srebro (2021), our analysis neither uses the knowledge of model scale which is typically inaccessible in real problems, nor relies on the restarting arguments for strongly convex problems. Also for SGD with smooth loss functions, a fast rate of $\mathcal{O}(\frac{1}{N})$ has recently been established via stability theory in the ideally clean case where the optimal population risk is zero (Lei and Ying, 2020, Theorem 4).

With $\gamma = \mathcal{O}(\frac{1}{n})$, the first bound of our Theorem 3 matches that bound in the context of M-SPP. For strongly convex problems, our results in Theorem 1 are stronger than (Lei and Ying, 2020, Theorem 12) in the sense that the formers (ours) only require the population risk to have quadratic growth property while the latter requires the loss to be Lipschitz and the empirical risk to be strongly convex. Finally, for convex ERM, similar composite risk bounds have been established by Srebro et al. (2010); Zhang et al. (2017) under somewhat more stringent conditions such as bounded domain of interest and huge sample with $N \gg p$.

Table 2 summaries a comparison of the risk bounds obtained in this work to several prior ones for (M-)SPP, (M-)SGD and ERM.

Method	Literature	Risk Bound	Conditions		
			Loss	R	R_S
M-SPP	Asi et al. (2020)	$\mathcal{O}\left(\frac{\log(N)}{N}\right)$	s.cvx	—	—
	Wang et al. (2017b)	$\mathcal{O}\left(\frac{1}{N}\right)$	Lip & s.cvx	—	—
	Theorem 1 **(our work)**	$\mathcal{O}\left(\frac{1}{T^2} + \frac{R^*}{N}\right)$ or $\mathcal{O}\left(\frac{1}{T^2} + \frac{1+R^*}{N}\right)$	sm & cvx	qg	—
	Theorem 3 **(our work)**	$\mathcal{O}\left(\frac{1}{N} + R^*\right)$ or $\mathcal{O}\left(\frac{1+R^*}{\sqrt{N}}\right)$	sm & cvx	—	—
SPP	Asi and Duchi (2019b)	$\mathcal{O}\left(\frac{\log(N)}{N}\right)$	s.cvx	—	—
	Patrascu and Necoara (2017)	$\mathcal{O}\left(\frac{1}{N}\right)$	sm & s.cvx	—	
	Davis and Drusvyatskiy (2019)	$\mathcal{O}\left(\frac{1}{N^2} + \frac{1}{N}\right)$	Lip & s.cvx	—	
M-SGD	Woodworth and Srebro (2021)	$\mathcal{O}\left(\frac{1}{T^2} + \frac{1}{N} + \sqrt{\frac{R^*}{N}}\right)$	sm & cvx	—	—
		$\mathcal{O}\left(e^{-T} + \frac{R^*}{N}\right)$	sm & cvx	qg	—
SGD	Dieuleveut et al. (2017)	$\mathcal{O}\left(\frac{1}{N^2} + \frac{R^*}{N}\right)$	quadratic	s.cvx	—
	Lei and Ying (2020)	$\mathcal{O}\left(\frac{1}{N} + R^*\right)$ or $\mathcal{O}\left(\frac{1+R^*}{\sqrt{N}}\right)$	sm & cvx	—	s.cvx
	Rakhlin et al. (2012)	$\mathcal{O}\left(\frac{1}{N}\right)$	Lip & sm & cvx	s.cvx	—
ERM	Zhang et al. (2017)	$\mathcal{O}\left(\frac{p}{N} + \frac{R^*}{N}\right)$ or $\mathcal{O}\left(\frac{1}{N^2} + \frac{R^*}{N}\right)$ for $N \gtrsim p$	sm & cvx	Lip & s.cvx	—
	Srebro et al. (2010)	$\mathcal{O}\left(\frac{1}{N} + \sqrt{\frac{R^*}{N}}\right)$	sm & cvx	—	—

Table 2: Comparison of our risk bounds to some prior results for M-SPP and SPP as well as for SGD and ERM. Recall that T is the iteration count and N is the total number of samples accessed. All the listed bounds hold in expectation. Here we have used the following abbreviations: cvx (convex), s.cvx (strongly convex), Lip (Lipschitz continuous), sm (smooth), qg (quadratic growth).

6. Experiments

In this section, we carry out a set of numerical study to demonstrate the convergence performance of minibatch stochastic proximal point methods in (composite) statistical learning problems. The goal of this study is to answer the following three questions associated with the key theory and algorithms established in this article:

- Question 1: *How the size of minibatch and noise level of a statistical learning model affect the convergence speed of M-SPP for smooth loss function?* This question is mainly about verifying Theorem 1 and Theorem 5, and it is answered through a simulation study on Lasso estimation in Section 6.1.

- Question 2: *Can the two-phase variant of M-SPP improve over M-SPP in the small-T-large-n setting?* The simulation results presented in Section 6.1 also answer this question related to the verification of Theorem 2.

- Question 3: *How M-SPP(-TP) methods compare with M-SGD in convergence performance?* The real-data experimental results on logistic regression tasks in Section 6.2 answer this question about algorithm comparison.

6.1 Simulation Study

We first provide a simulation study to verify our theoretical results for smooth losses when substantialize to the widely used Lasso regression model (Wainwright, 2009) with quadratic loss function $\ell(f_w(x), y) = \frac{1}{2}(y - w^\top x)^2$ and $r(f_w) = \mu\|w\|_1$ where μ is the ℓ_1-penalty modulus. Given a model parameter $\bar{w} \in \mathbb{R}^p$ and a feature point $x \in \mathbb{R}^p$ drawn from standard Gaussian distribution $\mathcal{N}(0, I_{p\times p})$, the responses y is generated according to a linear model $y = \bar{w}^\top x + \varepsilon$ with a random Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. In this case, the population risk function can be expressed in a close form as

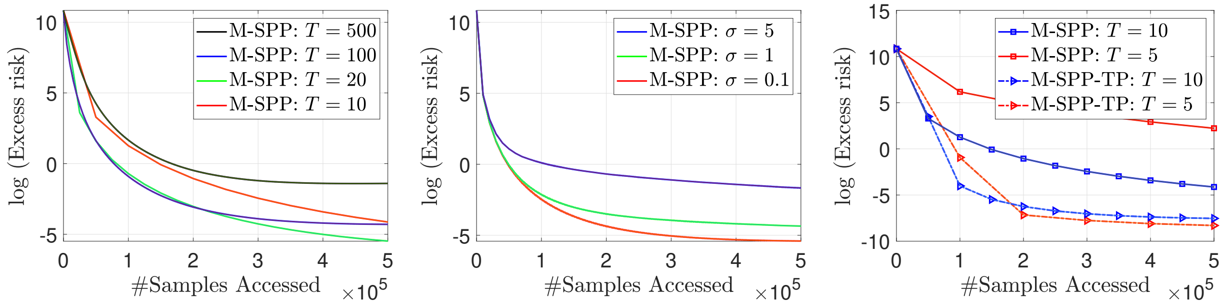
$$R(w) = \frac{1}{2}\|w - \bar{w}\|^2 + \frac{\sigma^2}{2} + \mu\|w\|_1.$$

Given a set of T random n-minibatches $\left\{S_t = \{x_{i,t}, y_{i,t}\}_{i\in[n]}\right\}_{t\in[T]}$ drawn from the above data distribution, we aim at evaluating the convergence performance of M-SPP towards the minimizer of R which can be expressed as

$$w^* = (\bar{w} - \mu)_+ - (-\bar{w} - \mu)_+,$$

where $(\cdot)_+$ is an element-wise function that preserves the positive parts of a vector.

We test with $p = 5000$ and $N = nT = 100p$, and consider a well-specified sparse regression model where the true parameter vector $\bar{w}$ is $\bar{k}$-sparse with $\bar{k} = 0.2p$ and its non-zero entries are sampled from a zero-mean Gaussian distribution. We set $\mu = 10^{-3}$ and initialize $w^{(0)} = 0$. The inner-loop minibatch proximal Lasso subproblems are optimized via a standard proximal gradient descent method, using either of the following two termination criteria: 1) the difference between consecutive objective values is below 10^{-3} and 2) the iteration step reaches 1000.



(a) Results under varying T.

(b) Results under varying σ.

(c) M-SPP *versus* M-SPP-TP

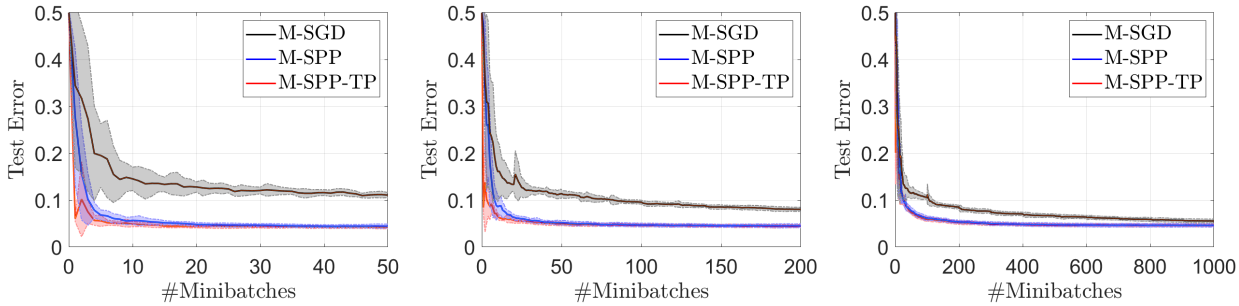
Figure 1: Simulation study on Lasso regression: Convergence performances of M-SPP and M-SPP-TP. The y-axis represents the logarithmic scale of excess risk.

The following two experimental setups are considered for theory verification:

- We fix the noise level $\sigma = 0.1$ and study the impact of varying $T \in \{10, 20, 100, 500\}$ on the convergence of M-SPP. Figure 1(a) shows the evolving curves of excess risk as functions of sample size, in a semi-log layout with y-axis representing the logarithmic scale of excess risk. From this set of curves we can observe a clear trend that in the early stage, M-SPP converges faster when the total number of minibatches is relatively large (say, $T \in \{20, 100\}$). This is consistent with the prediction of Theorem 1 about the impact of T and n on convergence rates. While in the final stage, relatively slower convergence behavior is exhibited under relatively larger T (say, $T \in \{100, 500\}$). This observation can be explained by the inexact analysis in Theorem 4 which shows that to guarantee the desired convergence rate, the inner-loop proximal ERM update needs to be extremely accurate when T is relatively large. Therefore, the question raised in Question 1 on the impact of minibatch size on convergence rate is answered by this group of results.

  Also in this setup, we have compared M-SPP and its two-phase variant M-SPP-TP for $T \in \{5, 10\}$. The related results are shown in Figure 1(c), which indicate that M-SPP-TP significantly improves the convergence of M-SPP in the small-T-large-n cases. This observation supports the result of Theorem 2 and answers Question 2 affirmatively.

- We fix $T = 50$ and study the impact of varying noise level $\sigma \in \{0.1, 1, 5\}$ on the convergence performance. The results are shown in Figure 1(b). From this group of results we can see that faster convergence speed is attained at relatively smaller noise level σ, while the speed becomes insensitive to noise level when σ is sufficiently small (say, $\sigma \leq 1$). This is consistent with the predication by Theorem 1, keeping in mind the fact that $R^* = \frac{1}{2}\|w^* - \bar{w}\|^2 + \frac{\sigma^2}{2} + \mu\|w^*\|_1 \leq \|\bar{w}\|^2 + \frac{1}{2}\sigma^2$. The question raised in Question 1 on the impact of noise level on convergence performance is answered by this group of results.



(a) $n = N/5$ (b) $n = N/20$ (c) $n = N/100$

Figure 2: Real-data results on logistic regression: Test error convergence comparison on `gisette` under varying minibatch size.

6.2 Experiment on Real Data

We further compare our methods with M-SGD for binary prediction problems using the logistic loss $\ell(w^\top x, y) = \log(1 + \exp(-yw^\top x))$. Here the M-SGD method is implemented by an SGD solver from SGDLibrary (Kasai, 2017). For M-SPP and M-SPP-TP, the The inner-loop minibatch proximal ERMs are solved by the same SGD solver applied with a fixed SGD-batch-size 10 and a single epoch of data processing. We initialize $w^{(0)} = 0$ for all the considered methods.

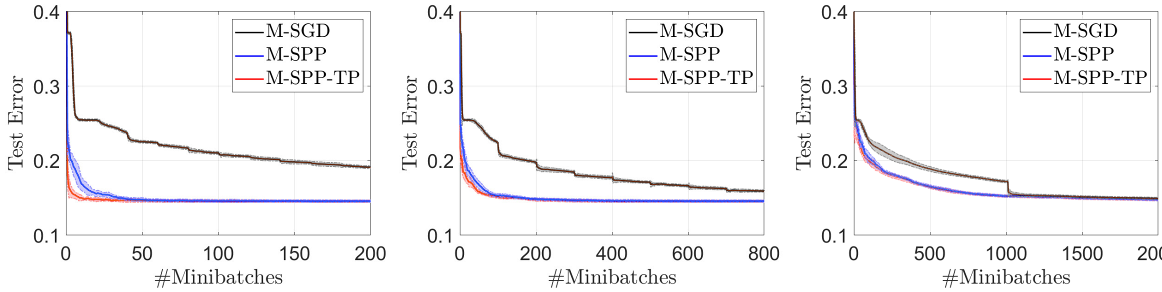
We use two public data sets for evaluation: the `gisette` data (Guyon et al., 2004) with $p = 5000, N = 6000$ and the `covtype.binary` data (Collobert et al., 2002) with $p = 54, N = 581,012$ [1]. For each data set, we use half of the samples as training set and the rest as test set. We are interested in the impact of minibatch-size n on the prediction performance of model measured by test error. All the considered stochastic algorithms are executed with 10 epochs of data processing, and thus the overall number of minibatches is $T = N/n \times 10$. We replicate each experiment 10 times over random split of data and report the results in mean-value along with error bar.

In Figure 2, we show the evolving curves (error bar shaded in color) of test error with respect to the number of minibatches accessed on `gisette`, under varying minibatch size $n \in \{\frac{N}{5}, \frac{N}{20}, \frac{N}{100}\}$. From this set of curves we can observe that:

- Under the same minibatch size, M-SPP and M-SPP-TP converge faster and stabler than M-SGD, especially when the minibatch size is relatively large (see Figure 2(a)). This is as expected because when minibatch size becomes large, M-SGD approaches to gradient descent method while M-SPP approaches ERMs. This answers Question 3 raised at the beginning of the experiment section.

- M-SPP-TP exhibits sharper convergence behavior than M-SPP at the early stage of iteration, especially when the minibatch-size is relatively large. This is consistent with our theoretical results in Theorem 1 and Theorem 2.

1. Both data sets are available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.



(a) $n = N/20$ (b) $n = N/100$ (c) $n = N/1000$

Figure 3: Real-data results on logistic regression: Test error convergence comparison on `covtype.binary` under varying minibatch size.

Figure 3 shows the corresponding results on `covtype` under $n \in \left\{\frac{N}{20}, \frac{N}{100}, \frac{N}{1000}\right\}$. From this set of results we once again see that M-SPP and M-SPP-TP consistently outperform M-SGD under the same minibatch size, and M-SPP-TP converges faster than M-SPP under relatively larger minibatch size (say, $n = \frac{N}{20}$).

7. Conclusions and Future Prospects

In this article, we presented an improved convergence analysis for minibatch stochastic proximal point risk minimization with smooth and convex losses. Under the quadratic growth condition on population risk, we showed that M-SPP with minibatch-size n and iteration count T converges at a composite rate consisting of an $\mathcal{O}(\frac{1}{T^2})$ *bias* decaying component and an $\mathcal{O}(\frac{1}{N})$ *variance* decaying component. In the small-n-large-T case, this result substantially improves the prior relevant results of SPP-type approaches which have usually been restricted to the regime where each instantaneous loss is assumed Lipschitz and/or strongly convex. Complementally in the small-T-large-n setting, we provide a two-phase acceleration of M-SPP which improves the $\mathcal{O}(\frac{1}{T^2})$ bias decaying rate to $\mathcal{O}\left(\frac{\log(N)}{N^2}\right)$. Perhaps the most interesting theoretical finding is that the (dominant) variance decaying term has a factor dependence on the minimal value of population risk, justifying the sharper convergence behavior of M-SPP in low-noise statistical setting as backed up by our numerical evidence. In addition to the in-expectation risk bounds, we have also derived a near-optimal parameter estimation error bound for a random shuffling variant of M-SPP that holds with high probability over data distribution while in expectation over the random shuffling. To conclude, our theory lays a novel and stronger foundation for understanding the convex M-SPP style algorithms that have gained recent significant attention, both in theory and practice, for large-scale machine learning (Li et al., 2014; Wang et al., 2017a; Asi et al., 2020).

There are several key prospects for future investigation of our theory:

- It is still open to derive optimal high-probability excess risk bounds for M-SPP that apply to the (suffix) average or last of iterates over training data.
- Inspired by the recent progresses made towards understanding M-SPP with momentum acceleration (Chadha et al., 2021; Deng and Gao, 2021), it is interesting to provide momentum and weakly-convex extensions of our theory for smooth loss functions.
- Last but not least, we expect that the theory developed in this article can be extended to the setup of non-parametric minibatch stochastic proximal point optimization.

Acknowledgements

The work was in part conducted while both authors were with the Baidu Cognitive Computing Lab. Xiao-Tong Yuan is also funded in part by the National Key Research and Development Program of China under Grant No. 2018AAA0100400 and in part by the Natural Science Foundation of China (NSFC) under Grant No.U21B2049, No.61876090 and No.61936005.

Appendix A. Proofs for the Results in Section 2

In this section, we present the technical proofs for the main results stated in Section 2.

A.1 Proof of Theorem 1

Here we prove Theorem 1 as restated below for convenience.

Theorem 1 *Suppose that Assumptions 1 and 2 hold. Consider $\epsilon_t \equiv 0$ and the weighted average output $\bar{w}_T = \frac{2}{T(T+1)}\sum_{t=1}^T t w_t$ in Algorithm 1. Let $\rho \in (0, 0.5]$ be an arbitrary scalar.*

(a) Suppose that $n \geq \frac{64L}{\lambda\rho}$. Set $\gamma_t = \frac{\lambda\rho t}{4}$ for $t \geq 1$. Then for any $T \geq 1$,

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \frac{4\rho\left[R(w_0) - R^*\right]}{T^2} + \frac{2^9 L}{\lambda\rho n T} R^*.$$

(b) Set $\gamma_t = \frac{\lambda\rho t}{4} + \frac{16L}{n}$ for $t \geq 1$. Then for any $T \geq 1$,

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \left(\frac{4\rho}{T^2} + \frac{2^8 L}{\lambda n T}\right)\left[R(w_0) - R^*\right] + \left(\frac{2^{16}L^2}{\lambda^2\rho^2 n^2 T} + \frac{2^9 L}{\lambda\rho n T}\right) R^*,$$

We first present the following lemma which will be used in the proof. It can be viewed as a straightforward extension of the prior result (Wang et al., 2017b, Lemma 1) to the setup of composite minimization. A proof is included here for the sake of completeness.

Lemma 1 *Assume that the loss function ℓ is convex with respect to its first argument and the regularization function r is convex. Then for any $w \in \mathcal{W}$, we have*

$$R_{S_t}(w_t) - R_{S_t}(w) \leq \frac{\gamma_t}{2}\left(\|w - w_{t-1}\|^2 - \|w - w_t\|^2 - \|w_t - w_{t-1}\|^2\right).$$

Proof Since ℓ and r are both convex, R_{S_t} is convex over $\mathcal{W}$. The optimality of w_t implies that for any $w \in \mathcal{W}$ and $\eta \in (0, 1)$

$$\begin{aligned}
&R_{S_t}(w_t) + \frac{\gamma_t}{2}\|w_t - w_{t-1}\|^2 \leq R_{S_t}((1-\eta)w_t + \eta w) + \frac{\gamma_t}{2}\|(1-\eta)w_t + \eta w - w_{t-1}\|^2 \\
\leq &(1-\eta)R_{S_t}(w_t) + \eta R_{S_t}(w) + \frac{\gamma_t}{2}\left[(1-\eta)\|w_t - w_{t-1}\|^2 + \eta\|w - w_{t-1}\|^2 - \eta(1-\eta)\|w - w_t\|^2\right],
\end{aligned}$$

where in the last inequality we have used the definition of the norm $\|\cdot\|$. Rearranging both sides of the above inequality yields

$$\eta(R_{S_t}(w_t) - R_{S_t}(w)) \leq \frac{\eta\gamma_t}{2}\left[\|w - w_{t-1}\|^2 - (1-\eta)\|w - w_t\|^2 - \|w_t - w_{t-1}\|^2\right],$$

which then implies (keep in mind that $\eta > 0$)

$$R_{S_t}(w_t) - R_{S_t}(w) \leq \frac{\gamma_t}{2}\left[\|w - w_{t-1}\|^2 - (1-\eta)\|w - w_t\|^2 - \|w_t - w_{t-1}\|^2\right].$$

Limiting $\eta \to 0^+$ in the above inequality yields the desired bound. ∎

The following boundedness result for smooth function is due to Srebro et al. (2010, Lemma 3.1).

Lemma 2 *If g is non-negative and L-smooth, then $\|\nabla g(w)\| \le \sqrt{2Lg(w)}$.*

Let $\{\mathcal{F}_t\}_{t\ge 1}$ be the filtration generated by the iterates $\{w_t\}_{t\ge1}$ as $\mathcal{F}_t = \sigma(w_1, w_2, ..., w_t)$. With Lemma 1 and Lemma 2 in place, we can further establish the following key lemma that plays a fundamental role in proving Theorem 1.

Lemma 3 *Suppose that the Assumptions 1 holds. Set $\gamma_t \ge \frac{16L}{n}$. Then we have*

$$\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \le \gamma_t \left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{16L}{\gamma_t n} R^*.$$

Proof Let us consider a sample set $S_t^{(i)}$ which is identical to S_t except that one of the $z_{i,t}$ is replaced by another random sample $z'_{i,t}$. Denote

$$w_t^{(i)} = \arg\min_{w\in\mathcal{W}} \left\{ F_t^{(i)}(w) := R_{S_t^{(i)}}(w) + \frac{\gamma_t}{2}\|w - w_{t-1}\|^2 \right\},$$

where $R_{S_t^{(i)}}(w) := \frac{1}{n}\left(\sum_{j\neq i} \ell(w; z_{j,t}) + \ell(w; z'_{i,t})\right) + r(w)$. Then we can show that

$$\begin{aligned}
&F_t(w_t^{(i)}) - F_t(w_t)\\
=&\frac{1}{n}\sum_{j\neq i}\left(\ell(w_t^{(i)}; z_{j,t}) - \ell(w_t; z_{j,t})\right) + \frac{1}{n}\left(\ell(w_t^{(i)}; z_{i,t}) - \ell(w_t; z_{i,t})\right)\\
&+ r(w_t^{(i)}) - r(w_t) + \frac{\gamma_t}{2}\|w_t^{(i)} - w_{t-1}\|^2 - \frac{\gamma_t}{2}\|w_t - w_{t-1}\|^2\\
=&F_t^{(i)}(w_t^{(i)}) - F_t^{(i)}(w_t) + \frac{1}{n}\left(\ell(w_t^{(i)}; z_{i,t}) - \ell(w_t; z_{i,t})\right) - \frac{1}{n}\left(\ell(w_t^{(i)}; z'_{i,t}) - \ell(w_t; z'_{i,t})\right)\\
\le&\frac{1}{n}\left|\ell(w_t^{(i)}; z_{i,t}) - \ell(w_t; z_{i,t})\right| + \frac{1}{n}\left|\ell(w_t^{(i)}; z'_{i,t}) - \ell(w_t; z'_{i,t})\right|\\
\overset{\zeta_1}{\le}&\frac{\|\nabla \ell(w_t^{(i)}; z_{i,t})\| + \|\nabla \ell(w_t; z'_{i,t})\|}{n}\|w_t^{(i)} - w_t\|\\
\overset{\zeta_2}{\le}&\frac{\sqrt{2L\ell(w_t^{(i)}; z_{i,t})} + \sqrt{2L\ell(w_t; z'_{i,t})}}{n}\|w_t^{(i)} - w_t\|,
\end{aligned}$$

where "ζ_1" is due to the convexity of loss and in "ζ_2"we have used Lemma 2. The bound in Lemma 1 implies

$$F_t(w_t^{(i)}) - F_t(w_t) \ge \frac{\gamma_t}{2}\|w_t^{(i)} - w_t\|^2.$$

Combining the preceding two inequalities yields

$$\frac{\gamma_t}{2}\|w_t^{(i)} - w_t\| \le \frac{\sqrt{2L\ell(w_t^{(i)}; z_{i,t})} + \sqrt{2L\ell(w_t; z'_{i,t})}}{n},$$

which immediately gives

$$\|w_t^{(i)} - w_t\| \le \frac{2\left(\sqrt{2L\ell(w_t^{(i)}; z_{i,t})} + \sqrt{2L\ell(w_t; z'_{i,t})}\right)}{\gamma_t n}. \tag{8}$$

Let us now consider the following population risk and empirical risk over S_t with respect to the loss function ℓ:

$$R^\ell(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(w; z)], \quad R^\ell_{S_t}(w) := \frac{1}{n}\sum_{i=1}^{n} \ell(w; z_{i,t}).$$

Since S_t and $S_t^{(i)}$ are both i.i.d. samples of the data distribution. It follows that

$$\begin{aligned}
&\mathbb{E}_{S_t}\left[R^\ell(w_t) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t; z'_{i,t}) \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}_{S_t^{(i)}}\left[R^\ell(w_t^{(i)}) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{S_t^{(i)}\cup\{z_{i,t}\}}\left[\ell(w_t^{(i)}; z_{i,t}) \mid \mathcal{F}_{t-1}\right] \\
&.
\end{aligned}$$

Since the above holds for all $i = 1, ..., n$, we can further show that

$$\begin{aligned}
&\mathbb{E}_{S_t}\left[R^\ell(w_t) \mid \mathcal{F}_{t-1}\right] \\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t^{(i)}\cup\{z_{i,t}\}}\left[\ell(w_t^{(i)}; z_{i,t}) \mid \mathcal{F}_{t-1}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t^{(i)}; z_{i,t}) \mid \mathcal{F}_{t-1}\right] \\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t; z'_{i,t}) \mid \mathcal{F}_{t-1}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t^{(i)}\cup\{z_{i,t}\}}\left[\ell(w_t; z'_{i,t}) \mid \mathcal{F}_{t-1}\right].
\end{aligned} \tag{9}$$

Regarding the empirical case, we find that

$$\begin{aligned}
&\mathbb{E}_{S_t}\left[R^\ell_{S_t}(w_t) \mid \mathcal{F}_{t-1}\right] \\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t}\left[\ell(w_t; z_{i,t}) \mid \mathcal{F}_{t-1}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t; z_{i,t}) \mid \mathcal{F}_{t-1}\right].
\end{aligned}$$

Combining the preceding two equalities gives that

$$
\begin{aligned}
&|\mathbb{E}_{S_t}[R(w_t) - R_{S_t}(w_t) \mid \mathcal{F}_{t-1}]| \\
=&\left|\mathbb{E}_{S_t}\left[R^\ell(w_t) - R^\ell_{S_t}(w_t) \mid \mathcal{F}_{t-1}\right]\right| \\
=&\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t^{(i)}; z_{i,t}) - \ell(w_t; z_{i,t}) \mid \mathcal{F}_{t-1}\right]\right| \\
\le&\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\left|\ell(w_t^{(i)}; z_{i,t}) - \ell(w_t; z_{i,t})\right| \mid \mathcal{F}_{t-1}\right] \\
\le&\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\sqrt{2L\ell(w_t^{(i)}; z_{i,t})}\|w_t^{(i)} - w_t\| \mid \mathcal{F}_{t-1}\right] \\
\overset{(8)}{\le}&\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t^{(i)}\cup\{z_{i,t}\}}\left[\frac{4L\ell(w_t^{(i)}; z_{i,t})}{\gamma_t n} + \frac{4L\sqrt{\ell(w_t^{(i)}; z_{i,t})\ell(w_t; z'_{i,t})}}{\gamma_t n} \mid \mathcal{F}_{t-1}\right] \\
\overset{\zeta_1}{\le}&\left(\frac{L}{\gamma_t n}\right)\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[6\ell(w_t^{(i)}; z_{i,t}) + 2\ell(w_t; z'_{i,t}) \mid \mathcal{F}_{t-1}\right] \\
\overset{(9)}{=}&\frac{8L}{\gamma_t n}\mathbb{E}_{S_t}\left[R^\ell(w_t) \mid \mathcal{F}_{t-1}\right] \le \frac{8L}{\gamma_t n}\mathbb{E}_{S_t}[R(w_t) \mid \mathcal{F}_{t-1}],
\end{aligned}
$$

where in "ζ_1" we have used the fact $a^2 + b^2 \ge 2ab$ and the last inequality is due to the fact $r \ge 0$.

Let us now denote $w_t^* = \arg\min_{w\in W^*} \|w - w_t\|$. Conditioned on $\mathcal{F}_{t-1}$, taking expectation on both sides of the bound in Lemma 1 for $w = w_{t-1}^*$ yields

$$
\begin{aligned}
&\mathbb{E}_{S_t}[R_{S_t}(w_t) - R^* \mid \mathcal{F}_{t-1}] \\
\le&\frac{\gamma_t}{2}\mathbb{E}_{S_t}\left[\|w_{t-1}^* - w_{t-1}\|^2 - \|w_{t-1}^* - w_t\|^2 - \|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\le&\frac{\gamma_t}{2}\left(\|w_{t-1}^* - w_{t-1}\|^2 - \mathbb{E}_{S_t}\left[\|w_t^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right).
\end{aligned}
$$

Combining the preceding two inequalities yields

$$
\begin{aligned}
&\mathbb{E}_{S_t}[R(w_t) - R^* \mid \mathcal{F}_{t-1}] \\
=&\mathbb{E}_{S_t}[R(w_t) - R_{S_t}(w_t) + R_{S_t}(w_t) - R^* \mid \mathcal{F}_{t-1}] \\
\le&|\mathbb{E}_{S_t}[R(w_t) - R_{S_t}(w_t) \mid \mathcal{F}_{t-1}]| + \mathbb{E}_{S_t}[R_{S_t}(w_t) - R^* \mid \mathcal{F}_{t-1}] \\
\le&\frac{\gamma_t}{2}\left(\|w_{t-1}^* - w_{t-1}\|^2 - \mathbb{E}_{S_t}\left[\|w_t^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{8L}{\gamma_t n}\mathbb{E}_{S_t}[R(w_t) \mid \mathcal{F}_{t-1}] \\
=&\frac{\gamma_t}{2}\left(\|w_{t-1}^* - w_{t-1}\|^2 - \mathbb{E}_{S_t}\left[\|w_t^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{8L}{\gamma_t n}\mathbb{E}_{S_t}[R(w_t) - R^* \mid \mathcal{F}_{t-1}] + \frac{8L}{\gamma_t n}\mathbb{E}_{S_t}[R^*] \\
\le&\frac{\gamma_t}{2}\left(\|w_{t-1}^* - w_{t-1}\|^2 - \mathbb{E}_{S_t}\left[\|w_t^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{1}{2}\mathbb{E}_{S_t}[R(w_t) - R^* \mid \mathcal{F}_{t-1}] + \frac{8L}{\gamma_t n}R^*,
\end{aligned}
$$

where in the last inequality we have used the condition $\gamma_t \geq \frac{52L}{n}$. After rearranging the terms in the above inequality we obtain

$$\begin{aligned}\mathbb{E}_{S_t}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \leq & \gamma_t \left(\|w^*_{t-1} - w_{t-1}\|^2 - \mathbb{E}_{S_t}\left[\|w^*_t - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{16L}{\gamma_t n} R^* \\ = & \gamma_t \left(D^2(w_{t-1}, W^*) - \mathbb{E}_{S_t}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{16L}{\gamma_t n} R^*.\end{aligned}$$

This implies the desired bound. ■

The following lemma is a direct consequence of Lemma 3.

Lemma 4 *Suppose that the Assumptions 1 holds. Set $\gamma_t \geq \frac{16L}{n}$. Then the following holds for all $t \geq 1$:*

$$\mathbb{E}\left[D^2(w_t, W^*)\right] \leq D^2(w_0, W^*) + \sum_{\tau=1}^{t} \frac{16L}{\gamma_\tau^2 n} R^*.$$

Proof Since $R(w_t) \geq R^*$ and $\gamma_t \geq \frac{52L}{n}$, the bound in Lemma 3 immediately implies that

$$\mathbb{E}_{S_t}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] \leq D^2(w_{t-1}, W^*) + \frac{16L}{\gamma_t^2 n} R^*. \tag{10}$$

By unfolding the above recurrent from time instance t to zero we obtain that for all $t \geq 1$,

$$\mathbb{E}\left[D^2(w_t, W^*)\right] \leq D^2(w_0, W^*) + \sum_{\tau=1}^{t} \frac{16L}{\gamma_\tau^2 n} R^*.$$

This proves the desired bound. ■

With all these lemmas in place, we are now ready to prove the main result in Theorem 1.

Proof [of Theorem 1] **Part (a):** Note that the condition on n implies $\gamma_t = \frac{\lambda\rho t}{4} \geq \frac{\lambda\rho}{4} \geq \frac{16L}{n}$. Applying Lemma 3 along with the condition $R(w_t) - R^* \geq \frac{\lambda}{2} D^2(w_t, W^*)$ yields

$$\begin{aligned}& (1-\rho)\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\ \leq & \gamma_t D^2(w_{t-1}, W^*) - \left(\gamma_t + \frac{\lambda\rho}{2}\right)\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{2^4 L}{\gamma_t n} R^* \\ \leq & \frac{\lambda\rho t}{4} D^2(w_{t-1}, W^*) - \frac{\lambda\rho(t+2)}{4}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{2^6 L}{\lambda\rho t n} R^* \\ \leq & \frac{\lambda\rho t}{4} D^2(w_{t-1}, W^*) - \frac{\lambda\rho(t+2)}{4}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{2^7 L}{\lambda\rho(t+1)n} R^*,\end{aligned}$$

where in the last inequality we have used $\frac{1}{t} \leq \frac{2}{t+1}$ for $t \geq 1$. The above inequality implies

$$\begin{aligned}& t\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\ \leq & (t+1)\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\ \leq & \frac{\lambda\rho t(t+1)}{4(1-\rho)} D^2(w_{t-1}, W^*) - \frac{\lambda\rho(t+1)(t+2)}{4(1-\rho)}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{2^7 L}{\lambda n\rho(1-\rho)} R^*.\end{aligned}$$

Then based on the law of total expectation and after proper rearrangement we obtain

$$
\begin{aligned}
& t\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq & \frac{\lambda\rho t(t+1)}{4(1-\rho)}\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \frac{\lambda\rho(t+1)(t+2)}{4(1-\rho)}\mathbb{E}\left[D^2(w_t, W^*)\right] + \frac{2^7 L}{\lambda n\rho(1-\rho)}R^*.
\end{aligned} \tag{11}
$$

By summing the above inequality from $t = 1, ..., T$ and after normalization we obtain

$$
\begin{aligned}
\frac{2}{T(T+1)}\sum_{t=1}^{T} t\mathbb{E}\left[R(w_t) - R^*\right] \leq & \frac{\lambda\rho}{T(T+1)(1-\rho)}D^2(w_0, W^*) + \frac{2^8 L}{\lambda\rho(1-\rho)(T+1)n}R^* \\
\leq & \frac{2\lambda\rho}{T(T+1)}D^2(w_0, W^*) + \frac{2^9 L}{\lambda\rho(T+1)n}R^*,
\end{aligned}
$$

where in the last inequality we have used $\rho \leq 0.5$. Consider the weighted output $\bar{w}_T = \frac{2}{T(T+1)}\sum_{t=1}^{T} tw_t$. In view of the above inequality and the convexity and quadratic growth property of the risk function R we have

$$
\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \frac{4\rho\left[R(w_0) - R^*\right]}{T(T+1)} + \frac{2^9 L}{\lambda\rho n(T+1)}R^*,
$$

which then implies the desired bound in part (a).

Part (b): Note that $\gamma_t = \frac{\lambda\rho t}{4} + \frac{16L}{n} \geq \frac{16L}{n}$ for all $t \geq 1$. According to Lemma 4 we have the following holds for all $t \geq 1$:

$$
\begin{aligned}
& \mathbb{E}\left[D^2(w_t, W^*)\right] \\
\leq & D^2(w_0, W^*) + \sum_{\tau=1}^{t}\frac{16L}{\gamma_\tau^2 n}R^* \leq D^2(w_0, W^*) + \frac{2^8 L}{\lambda^2\rho^2 n}R^*\sum_{\tau=1}^{t}\frac{1}{\tau^2} \leq D^2(w_0, W^*) + \frac{2^9 L}{\lambda^2\rho^2 n}R^*.
\end{aligned} \tag{12}
$$

Similar to the argument in part (a), applying Lemma 3 along with the quadratic growth condition $R(w_t) - R^* \geq \frac{\lambda}{2}D^2(w_t, W^*)$ and $\rho \leq 0.5$ yields

$$
\begin{aligned}
& \frac{1}{2}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq & (1-\rho)\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq & \gamma_t D^2(w_{t-1}, W^*) - \left(\gamma_t + \frac{\lambda\rho}{2}\right)\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{2^4 L}{\gamma_t n}R^* \\
\leq & \frac{\lambda\rho t}{4}D^2(w_{t-1}, W^*) - \frac{\lambda\rho(t+2)}{4}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] \\
& + \frac{16L}{n}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{2^6 L}{\lambda\rho t n}R^*,
\end{aligned}
$$

where in the second inequality we have used $\gamma_t \geq \frac{52L}{n}$, and in the last inequality we have used $\gamma_t \geq \frac{\lambda\rho t}{4}$. Then based on the law of total expectation and after proper rearrangement we have

$$\begin{aligned}
&\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&\frac{\lambda\rho t}{2}\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \frac{\lambda\rho(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*)\right] \\
&+ \frac{2^5 L}{n}\left(\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \mathbb{E}\left[D^2(w_t, W^*)\right]\right) + \frac{2^7 L}{\lambda t n\rho}R^*,
\end{aligned}$$

which implies that

$$\begin{aligned}
&t\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&(t+1)\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&\frac{\lambda\rho t(t+1)}{2}\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \frac{\lambda\rho(t+1)(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*)\right] \\
&+ \frac{2^5 L(t+1)}{n}\left(\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \mathbb{E}\left[D^2(w_t, W^*)\right]\right) + \frac{2^7 L(t+1)}{\lambda t n\rho}R^* \\
\leq&\frac{\lambda\rho t(t+1)}{2}\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \frac{\lambda\rho(t+1)(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*)\right] \\
&+ \frac{2^6 L t}{n}\left(\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \mathbb{E}\left[D^2(w_t, W^*)\right]\right) + \frac{2^8 L}{\lambda n\rho}R^*,
\end{aligned}$$

where in the last inequality we have used the fact $t+1 \leq 2t$ as $t \geq 1$. By summing the above inequality from $t = 1, ..., T$ and after normalization we obtain

$$\begin{aligned}
&\frac{2}{T(T+1)}\sum_{t=1}^{T} t\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&\frac{2\lambda\rho}{T(T+1)}D^2(w_0, W^*) + \frac{2^7 L}{nT(T+1)}\sum_{t=1}^{T} D^2(w_{t-1}, W^*) + \frac{2^9 L}{\lambda\rho(T+1)n}R^* \\
\leq&\frac{2\lambda\rho}{T(T+1)}D^2(w_0, W^*) + \frac{2^7 L}{nT(T+1)}\sum_{t=1}^{T}\left(D^2(w_0, W^*) + \frac{2^9 L}{\lambda^2\rho^2 n}R^*\right) + \frac{2^9 L}{\lambda\rho(T+1)n}R^* \\
=&\left(\frac{2\lambda\rho}{T(T+1)} + \frac{2^7 L}{n(T+1)}\right)D^2(w_0, W^*) + \left(\frac{2^{16} L^2}{\lambda^2\rho^2 n^2(T+1)} + \frac{2^9 L}{\lambda\rho n(T+1)}\right)R^*,
\end{aligned}$$

where in the last inequality we have used (12). Using the convexity and quadratic growth property in the above inequality yields

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \left(\frac{4\rho}{T(T+1)} + \frac{2^8 L}{\lambda n(T+1)}\right)[R(w_0) - R^*] + \left(\frac{2^{16} L^2}{\lambda^2\rho^2 n^2(T+1)} + \frac{2^9 L}{\lambda\rho n(T+1)}\right)R^*,$$

which then implies the desired bound in part (b). The proof is concluded. ■

A.2 Proof of Theorem 2

In this subsection we prove Theorem 2 which is restated below.

Theorem 2 *Suppose that Assumptions 1 and 2 hold. Consider $\epsilon_t \equiv 0$ for implementing M-SPP in both Phase-I and Phase-II of Algorithm 2. Consider the weighted average output $\bar{w}_T = \frac{2}{(T-1)(T+2)}\sum_{t=2}^T t w_t$ in Phase-II.*

(a) Suppose that $n \geq \frac{128L}{\lambda}$. Set $m = \frac{128L}{\lambda}$ in Phase-I and $\gamma_t = \frac{\lambda t}{8}$ for implementing M-SPP in both Phase-I and Phase II. Then for any $T \geq 2$, $\bar{w}_T$ satisfies

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{L^2\left[R(w_0) - R^*\right]}{\lambda^2 n^2 T^2} + \frac{L}{\lambda n T} R^*.$$

(b) Set $m = \mathcal{O}(1)$ in Phase-I and $\gamma_t = \frac{\lambda t}{8} + \frac{16L}{n}$ for implementing M-SPP in both Phase-I and Phase-II. Then for any $T \geq 2$, $\bar{w}_T$ satisfies

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{L^2\left[R(w_0) - R^*\right]}{\lambda^2 n T} + \frac{L^3}{\lambda^3 n T} R^*.$$

Proof Part (a): In Phase-I, by invoking the first part of Theorem 1 with $\rho = 1/2$ and $T = n/m \geq 1$ (with slight abuse of notation) we get immediately that

$$\mathbb{E}_{S_1}\left[R(w_1) - R^*\right] \leq \frac{2m^2\left[R(w_0) - R^*\right]}{n^2} + \frac{2^{10}L}{\lambda n} R^*. \tag{13}$$

In Phase-II, conditioned on $\mathcal{F}_1$, summing the recursion form (11) from $t = 2, ..., T$ with $\rho = 1/2$ and proper normalization yields

$$\begin{aligned}
&\frac{2}{(T-1)(T+2)}\sum_{t=2}^T t\mathbb{E}_{S_{2:t}}\left[R(w_t) - R^* \mid \mathcal{F}_1\right] \\
\leq& \frac{6\lambda D^2(w_1, W^*)}{(T-1)(T+2)} + \frac{2^{10}L}{\lambda n(T+2)} R^* \leq \frac{3\left(R(w_1) - R^*\right)}{(T-1)(T+2)} + \frac{2^{10}L}{\lambda n(T+2)} R^*,
\end{aligned}$$

where in the last inequality we have used the quadratic growth property. Consider the weighted average output $\bar{w}_T = \frac{2}{(T-1)(T+2)}\sum_{t=2}^T t w_t$. Based on the above inequality and law of total expectation we must have

$$\begin{aligned}
\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq& \frac{6\mathbb{E}_{S_1}\left[R(w_1) - R^*\right]}{(T-1)(T+2)} + \frac{2^{10}L}{\lambda n(T+2)} R^* \\
\leq& \frac{6\mathbb{E}_{S_1}\left[R(w_1) - R^*\right]}{T^2} + \frac{2^{102}L}{\lambda n T} R^* \\
\leq& \frac{12m^2\left[R(w_0) - R^*\right]}{n^2 T^2} + \frac{2^{13}L}{\lambda n T} R^* \\
\leq& \frac{2^{22}L^2\left[R(w_0) - R^*\right]}{\lambda^2 n^2 T^2} + \frac{2^{13}L}{\lambda n T} R^*,
\end{aligned}$$

where we have used the fact $T \geq 2$ in multiple places and in the last but one step we have used (13). This immediately implies the desired bound in Part (a).

Part (b): In Phase-I, by applying second part of Theorem 1 (with $\rho = 1/2$ and $T = n/m \geq 1$) and preserving the leading terms we obtain that

$$\begin{aligned}\mathbb{E}_{S_1}\left[R(w_1) - R^*\right] &\lesssim \left(\frac{m^2}{n^2} + \frac{L}{\lambda n}\right)\left[R(w_0) - R^*\right] + \left(\frac{L^2}{\lambda^2 mn} + \frac{L}{\lambda n}\right) R^* \\ &\lesssim \frac{L}{\lambda n}\left[R(w_0) - R^*\right] + \frac{L^2}{\lambda^2 n} R^*.\end{aligned} \tag{14}$$

In Phase-II, based on the proof argument of the part (b) of Theorem 1 we can show that the weighted average output $\bar{w}_T = \frac{2}{(T-1)(T+2)}\sum_{t=2}^{T} t w_t$ satisfies

$$\begin{aligned}\mathbb{E}\left[R(\bar{w}_T) - R^*\right] &\lesssim \left(\frac{1}{T^2} + \frac{L}{\lambda n T}\right)\mathbb{E}_{S_1}\left[R(w_1) - R^*\right] + \left(\frac{L^2}{\lambda^2 n^2 T} + \frac{L}{\lambda n T}\right) R^* \\ &\lesssim \left(\frac{L}{\lambda n T^2} + \frac{L^2}{\lambda^2 n^2 T}\right)\left[R(w_0) - R^*\right] + \left(\frac{L^3}{\lambda^3 n^2 T} + \frac{L^2}{\lambda^2 n T}\right) R^* \\ &\lesssim \frac{L^2}{\lambda^2 n T}\left[R(w_0) - R^*\right] + \frac{L^3}{\lambda^3 n T} R^*,\end{aligned}$$

where in the second step we have used (14). This proves the desired bound in Part (b). ■

A.3 Proof of Theorem 3

In this subsection, we prove Theorem 3 as following restated.

Theorem 3 *Suppose that Assumption 1 holds. Set $\gamma_t \equiv \gamma \geq \frac{16L}{n}$. Let $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ be the average output of Algorithm 1. Then*

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{\gamma}{T} D^2(w_0, W^*) + \frac{L}{\gamma n} R^*.$$

Particularly for $\gamma = \sqrt{\frac{T}{n}} + \frac{16L}{n}$, it holds that

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right) D^2(w_0, W^*) + \frac{L}{\sqrt{nT}} R^*.$$

Proof Since $\gamma_t \equiv \gamma \geq \frac{16L}{n}$, the bound in Lemma 3 is valid. Based on law of total expectation and by summing that inequality from $t = 1, ..., T$ with proper normalization we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[R(w_t) - R^*\right] \leq \frac{\gamma}{T} D^2(w_0, W^*) + \frac{16L}{\gamma n} R^*.$$

Consider $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$. In view of the above inequality and convexity of R we have

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \frac{\gamma}{T} D^2(w_0, W^*) + \frac{16L}{\gamma n} R^*.$$

This proves the first desired bound. The second bound follows immediately by substituting $\gamma = \sqrt{\frac{T}{n}} + \frac{16L}{n} > \frac{16L}{n}$ into the above bound. The proof is concluded. ■

A.4 On the (Iteration) Stability of M-SPP

In this appendix subsection, we further provide a sensitivity analysis of M-SPP to the choice of regularization modulus $\{\gamma_t\}_{t\geq 1}$, under the following notion of iteration stability essentially introduced by Asi and Duchi (2019a,b).

Definition 2 *A stochastic optimization algorithm generating iterates $\{w_t\}_{t\geq 1}$ for minimizing the population risk $R(w)$ is staid to be stable if*

$$\sup_{t\geq 1} D(w_t, W^*) < \infty, \quad \textit{with probability 1.}$$

Before presenting the main results on the iteration stability of M-SPP, we first recall the Robbins-Siegmund nonnegative almost supermartingale convergence lemma which is typically used for establishing the stability and convergence of stochastic optimization methods including SPP (Asi and Duchi, 2019b).

Lemma 5 (Robbins and Siegmund (1971)) *Consider fours sequences of nonnegative random variables $\{U_t\}, \{V_t\}, \{\alpha_t\}, \{\beta_t\}$ that are measurable over a filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Suppose that $\sum_t \alpha_t < \infty$, $\sum_t \beta_t < \infty$, and*

$$\mathbb{E}[U_{t+1} \mid \mathcal{F}_t] \leq (1+\alpha_t)U_t + \beta_t - V_t.$$

Then there exits U_∞ such that $U_t \xrightarrow{a.s.} U_\infty$ and $\sum_t V_t < \infty$ with probability 1.

The following proposition shows that the sequence of estimation error $\{\|w_t - w^*\|\}$ is non-divergent in expectation and it converges to some finite value and is bounded with probability 1.

Proposition 2 *Suppose that the Assumptions 1 holds. Assume that $\gamma_t \geq \frac{16L}{n}$ and $\sum_{t\geq 1} L\gamma_t^{-2} < \infty$. Then we have the following hold:*

(a) $\mathbb{E}\left[D(w_t, W^*)\right] < \infty$;

(b) $D(w_t, W^*)$ converges to some finite value and $\sup_{t\geq 1} D(w_t, W^*) < \infty$ with probability 1.

Proof Applying Lemma 4 yields that for all $t \geq 1$

$$\mathbb{E}\left[D^2(w_t, W^*)\right] \lesssim D^2(w_0, W^*) + \sum_{\tau=1}^{t} \frac{L}{\gamma_\tau^2 n} R^* < \infty,$$

where we have used the given conditions on γ_t. This proves the part (a). To show the part (b), invoking Lemma 5 with $\alpha_t = V_t \equiv 0$ and $\beta_t = \frac{16L}{\gamma_t^2 n} R^*$ to (10) yields $D(w_t, W^*)$ converges to some finite value and thus $\sup_{t\geq 1} D(w_t, W^*) < \infty$ almost surely. ■

Remark 12 *Proposition 2 shows that in contrast to minibatch SGD, the choice of γ_t in M-SPP is insensitive to the gradient scale of loss functions for generating a non-divergent sequence of estimation errors.*

Appendix B. Proofs for the Results in Section 3

In this section, we present the technical proofs for the main results stated in Section 3.

B.1 Proof of Theorem 4

In this subsection, we prove Theorem 4 which is restated below.

Theorem 4 *Suppose that Assumptions 1, 2 and 3 hold. Let $\rho \in (0, 1/4]$ be an arbitrary scalar and set $\gamma_t = \frac{\lambda\rho t}{4}$. Suppose that $n \geq \frac{76L}{\lambda\rho}$. Assume that $\epsilon_t \leq \frac{\epsilon}{nt^4}$ for some $\epsilon \in [0, 1]$. Then for any $T \geq 1$, the weighted average output $\bar{w}_t = \frac{2}{T(T+1)} \sum_{t=1}^{T} t w_t$ of Algorithm 1 satisfies*

$$\mathbb{E}\left[R(\bar{w}_t) - R^*\right] \lesssim \frac{\rho}{T^2}(R(w_0) - R^*) + \frac{L}{\lambda\rho n T} R^* + \frac{\sqrt{\epsilon}}{T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right).$$

Preliminaries. In what follows, we denote by $\tilde{w}_t := \arg\min_{w \in \mathcal{W}} F_t(w)$ the exact solution of the inner-loop minibatch ERM optimization, which plays the same role as w_t in Section 2. We first present the following lemma that upper bounds the discrepancy between the inexact minimizer w_t and the exact minimizer $\tilde{w}_t$.

Lemma 6 *Assume that the loss function ℓ is convex with respect to its first argument and r is convex. Then for any $w \in \mathcal{W}$, we have*

$$\|w_t - \tilde{w}_t\| \leq \sqrt{\frac{2\epsilon_t}{\gamma_t}}.$$

Proof Using arguments identical to those of Lemma 1 we can show that for all $w \in \mathcal{W}$,

$$R_{S_t}(\tilde{w}_t) - R_{S_t}(w) \leq \frac{\gamma_t}{2}\left(\|w - w_{t-1}\|^2 - \|w - \tilde{w}_t\|^2 - \|\tilde{w}_t - w_{t-1}\|^2\right). \tag{15}$$

Setting $w = w_t$ in the above yields

$$\frac{\gamma_t}{2}\|w_t - \tilde{w}_t\|^2 \leq F_t(w_t) - F_t(\tilde{w}_t) \leq \epsilon_t,$$

which directly implies $\|w_t - \tilde{w}_t\| \leq \sqrt{2\epsilon_t/\gamma_t}$. This proves the second desired bound. ■

The following lemma as an extension of Lemma 3 to the inexact setting.

Lemma 7 *Suppose that the Assumptions 1, 2 and 3 hold. Assume that $\gamma_t \geq \frac{19L}{n}$. Then the following bound holds for any $\rho \in (0, 1)$:*

$$\begin{aligned}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \leq & \gamma_t\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right) D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) \\ & + \frac{19L}{\gamma_t n} R^* + \left(3n + \frac{4\gamma_t}{\rho\lambda}\right)\epsilon_t + 3G\sqrt{\frac{2\epsilon_t}{\gamma_t}}.\end{aligned}$$

Alternatively, for any $w^* \in W^*$, under Assumptions 1 and 3 we have

$$\begin{aligned}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \leq & \gamma_t \left(\|w_{t-1} - w^*\|^2 - \mathbb{E}\left[\|w_t - w^*\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{19L}{\gamma_t n} R^* \\ & + 3n\epsilon_t + \left(2\sqrt{2\gamma_t}\mathbb{E}\left[\|w_t - w^*\| \mid \mathcal{F}_{t-1}\right] + 3G\sqrt{\frac{2}{\gamma_t}}\right)\sqrt{\epsilon_t}.\end{aligned}$$

Proof Let us decompose $\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right]$ into the following three terms:

$$\begin{aligned}&\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\ =&\underbrace{\mathbb{E}\left[R(w_t) - R(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right]}_{A} + \underbrace{\mathbb{E}\left[R(\tilde{w}_t) - R_{S_t}(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right]}_{B} + \underbrace{\mathbb{E}\left[R_{S_t}(\tilde{w}_t) - R^* \mid \mathcal{F}_{t-1}\right]}_{C}.\end{aligned}$$

We next bound these three terms respectively. To bound the term A, we can show that

$$\begin{aligned}&|A| := \left|\mathbb{E}\left[R(w_t) - R(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right]\right| \\ =&\left|\mathbb{E}\left[R^\ell(w_t) - R^\ell(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[r(w_t) - r(\tilde{w}_t)\right] \mid \mathcal{F}_{t-1}\right| \\ \leq&\mathbb{E}\left[\mathbb{E}_z|\ell(w_t;z) - \ell(\tilde{w}_t;z)| \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[|r(w_t) - r(\tilde{w}_t)| \mid \mathcal{F}_{t-1}\right] \\ \overset{\zeta_1}{\leq}&\mathbb{E}\left[\mathbb{E}_z\left[\sqrt{2L\ell(w_t;z)}\|w_t - \tilde{w}_t\|\right] \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[G\|w_t - \tilde{w}_t\| \mid \mathcal{F}_{t-1}\right] \\ \leq&\mathbb{E}\left[\mathbb{E}_z\left[\frac{L}{\gamma_t n}\ell(w_t;z) + \frac{\gamma_t n}{2}\|w_t - \tilde{w}_t\|^2\right] \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[G\|w_t - \tilde{w}_t\| \mid \mathcal{F}_{t-1}\right] \\ =&\mathbb{E}\left[\frac{L}{\gamma_t n}R^\ell(w_t) \mid \mathcal{F}_{t-1}\right] + \mathbb{E}_{S_t}\left[\frac{\gamma_t n}{2}\|w_t - \tilde{w}_t\|^2 + G\|w_t - \tilde{w}_t\| \mid \mathcal{F}_{t-1}\right] \\ \leq&\mathbb{E}\left[\frac{L}{\gamma_t n}R(w_t) \mid \mathcal{F}_{t-1}\right] + n\epsilon_t + G\sqrt{\frac{2\epsilon_t}{\gamma_t}},\end{aligned}$$

where in "ζ_1" we have used the convexity of loss and Lemma 2 and the Assumption 3 and in the last inequality we have used $r > 0$ and the perturbation bound of Lemma 6.

To bound the term B, using about the same proof arguments as for Lemma 3 we can show that

$$\begin{aligned}B :=&\mathbb{E}\left[R(\tilde{w}_t) - R_{S_t}(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right] \\ \leq&\frac{8L}{\gamma_t n}\mathbb{E}\left[R(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right] \\ =&\frac{8L}{\gamma_t n}\mathbb{E}\left[R(\tilde{w}_t) - R(w_t)\right] + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] \\ \leq&\frac{1}{2}|A| + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right],\end{aligned}$$

where we have used the condition on minibatch size γ_t.

To bound the term C, based on the definition of $\tilde{w}_t$ and by invoking Lemma 1 with $w = w^*_{t-1}$ we can verify that

$$
\begin{aligned}
C :=& \mathbb{E}\left[R_{S_t}(\tilde{w}_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq& \frac{\gamma_t}{2}\mathbb{E}\left[\|w^*_{t-1} - w_{t-1}\|^2 - \|w^*_{t-1} - \tilde{w}_t\|^2 - \|\tilde{w}_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\leq& \frac{\gamma_t}{2}\mathbb{E}\left[\|w^*_{t-1} - w_{t-1}\|^2 - \|w^*_{t-1} - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
=& \frac{\gamma_t}{2}\left(\|w^*_{t-1} - w_{t-1}\|^2 - \mathbb{E}\left[\|w^*_{t-1} - w_t + w_t - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
=& \frac{\gamma_t}{2}\left(\|w^*_{t-1} - w_{t-1}\|^2 - \mathbb{E}\left[\|w^*_{t-1} - w_t\|^2 + 2\langle w^*_{t-1} - w_t, w_t - \tilde{w}_t\rangle + \|w_t - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\leq& \frac{\gamma_t}{2}\left(\|w^*_{t-1} - w_{t-1}\|^2 - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)\|w^*_{t-1} - w_t\|^2 - \frac{2\gamma_t}{\rho\lambda}\|w_t - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\leq& \frac{\gamma_t}{2}\left(\|w^*_{t-1} - w_{t-1}\|^2 - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)\|w^*_{t} - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{2\gamma_t\epsilon_t}{\rho\lambda} \\
=& \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{2\gamma_t\epsilon_t}{\rho\lambda}.
\end{aligned}
$$

Combining the above three bounds yields

$$
\begin{aligned}
&\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] = A + B + C \\
\leq& \frac{3}{2}|A| + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] \\
&+ \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{2\gamma_t\epsilon_t}{\rho\lambda} \\
\leq& \mathbb{E}\left[\frac{3L}{2\gamma_t n}R(w_t) \mid \mathcal{F}_{t-1}\right] + \frac{3n}{2}\epsilon_t + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}} + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] \\
&+ \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{2\gamma_t\epsilon_t}{\rho\lambda} \\
\leq& \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{9.5L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] \\
&+ \left(\frac{3n}{2} + \frac{2\gamma_t}{\rho\lambda}\right)\epsilon_t + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}} \\
=& \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{9.5L}{\gamma_t n}\mathbb{E}\left[R^*\right] + \frac{9.5L}{\gamma_t n}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
&+ \left(\frac{3n}{2} + \frac{2\gamma_t}{\rho\lambda}\right)\epsilon_t + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}} \\
\leq& \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{9.5L}{\gamma_t n}\mathbb{E}\left[R^*\right] + \frac{1}{2}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
&+ \left(\frac{3n}{2} + \frac{2\gamma_t}{\rho\lambda}\right)\epsilon_t + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}},
\end{aligned}
$$

where in the last inequality we have used the condition $\gamma_t \geq \frac{19L}{n}$. After rearranging the terms in the above inequality we obtain the first desired bound.

To derive the second bound, for any fixed $w^* \in W^*$, we note that the term C can be alternatively bounded as

$$\begin{aligned}
C \leq & \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 + 2\langle w^* - w_t, w_t - \tilde{w}_t\rangle + \|w_t - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\leq & \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 - 2\|w_t - w^*\|\|w_t - \tilde{w}_t\| \mid \mathcal{F}_{t-1}\right]\right) \\
\leq & \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \sqrt{2\gamma_t\epsilon_t}\mathbb{E}\left[\|w^* - w_t\| \mid \mathcal{F}_{t-1}\right].
\end{aligned}$$

Similar to the proof of the first bound, we can derive that

$$\begin{aligned}
& \mathbb{E}_{S_t}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] = A + B + C \\
\leq & \frac{3}{2}|A| + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] + \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \sqrt{2\gamma_t\epsilon_t}\mathbb{E}\left[\|w^* - w_t\| \mid \mathcal{F}_{t-1}\right] \\
\leq & \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{9.5L}{\gamma_t n}R^* \\
& + \frac{1}{2}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] + \frac{3n}{2}\epsilon_t + \sqrt{2\gamma_t\epsilon_t}\mathbb{E}\left[\|w^* - w_t\| \mid \mathcal{F}_{t-1}\right] + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}}.
\end{aligned}$$

After rearranging the terms in the above inequality we obtain the second desired bound. ■

With the above preliminary results in hand, we are now in the position to prove the main result of Theorem 4.

Proof [of Theorem 4] Since by assumption $R(w_t) - R^* \geq \frac{\lambda}{2}D^2(w_t, W^*)$ and $\gamma_t = \frac{\lambda\rho t}{4} \geq \frac{\lambda\rho}{4} \geq \frac{19L}{n}$, based on the first bound in Lemma 7 we can show that

$$\begin{aligned}
& (1 - 2\rho)\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq & \gamma_t D^2(w_{t-1}, W^*) - \left(\gamma_t + \frac{\rho\lambda}{2}\right)\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{19L}{\gamma_t n}R^* + \left(3n + \frac{4\gamma_t}{\rho\lambda}\right)\epsilon_t + 3G\sqrt{\frac{2\epsilon_t}{\gamma_t}} \\
\leq & \frac{\lambda\rho t}{4}D^2(w_{t-1}, W^*) - \frac{\rho\lambda(t+2)}{4}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{76L}{\lambda\rho n t}R^* + (3n + t)\epsilon_t + 6G\sqrt{\frac{2\epsilon_t}{\lambda\rho t}}.
\end{aligned}$$

Now suppose that $\epsilon_t \leq \frac{\epsilon}{nt^4}$ for some $\epsilon \in [0, 1]$. Since $\rho \leq 1/4$, the above implies

$$\begin{aligned}
& \mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq & \frac{\lambda\rho t}{2}D^2(w_{t-1}, W^*) - \frac{\rho\lambda(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{152L}{\lambda\rho n t}R^* + \left(\frac{6}{t^4} + \frac{2}{t^3} + 12G\sqrt{\frac{2}{\lambda\rho t^5}}\right)\sqrt{\epsilon}.
\end{aligned}$$

The above inequality then implies

$$\begin{aligned}
t\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \leq & (t+1)\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq & \frac{\lambda\rho t(t+1)}{2}D^2(w_{t-1}, W^*) - \frac{\lambda\rho(t+1)(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] \\
& + \frac{304L}{\lambda\rho n}R^* + \left(\frac{12}{t^3} + \frac{4}{t^2} + \frac{24G}{t}\sqrt{\frac{2}{\lambda\rho t}}\right)\sqrt{\epsilon},
\end{aligned}$$

where we have used the fact $\frac{t+1}{t} \leq 2$ for $t \geq 1$. In view of the law of total expectation, summing the above inequality from $t = 1, ..., T$ with natural normalization yields

$$\begin{aligned}
&\frac{2}{T(T+1)}\sum_{t=1}^{T} t\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq &\frac{2\lambda\rho}{T(T+1)}D^2(w_0, W^*) + \frac{608L}{\lambda\rho(T+1)n}R^* + \frac{\sqrt{\epsilon}}{T(T+1)}\left(64 + 192G\sqrt{\frac{2}{\lambda\rho}}\right) \\
\leq &\frac{4\rho}{T(T+1)}(R(w_0) - R^*) + \frac{608L}{\lambda\rho(T+1)n}R^* + \frac{\sqrt{\epsilon}}{T(T+1)}\left(64 + 192G\sqrt{\frac{2}{\lambda\rho}}\right),
\end{aligned}$$

which then immediately leads to the desired bound. The proof is concluded. ■

B.2 Proof of Theorem 5

In this subsection, we prove Theorem 5 as following restated.

Theorem 5 *Suppose that Assumptions 1 and 3 hold. Set $\gamma_t \equiv \gamma \geq \frac{19L}{n}$. Assume that $\epsilon_t \leq \min\left\{\frac{\epsilon}{n^2 t^5}, \frac{2G^2}{9n^2\gamma}\right\}$ for some $\epsilon \in [0, 1]$. Then the average output $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ of Algorithm 1 satisfies*

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{\gamma}{T}D^2(w_0, W^*) + \frac{L}{\gamma n}R^* + \left(\frac{L}{\gamma n} + \frac{\gamma}{LnT} + \frac{G}{\sqrt{\gamma}nT}\right)\sqrt{\epsilon}.$$

Particularly for $\gamma = \sqrt{\frac{T}{n}} + \frac{19L}{n}$, it holds that

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right)D^2(w_0, W^*) + \frac{L}{\sqrt{nT}}R^* + \left(\frac{L+G}{\sqrt{nT}} + \frac{1}{nT}\right)\sqrt{\epsilon}.$$

The following lemma, which can be proved by induction (see, e.g., Schmidt et al., 2011), will be used to prove the main result.

Lemma 8 *Assume that the nonnegative sequence $\{u_\tau\}_{\tau\geq 1}$ satisfies the following recursion for all $t \geq 1$:*

$$u_t^2 \leq S_t + \sum_{\tau=1}^{t} \alpha_\tau u_\tau,$$

with $\{S_\tau\}_{\tau\geq 1}$ an increasing sequence, $S_0 \geq u_0^2$ and $\alpha_\tau \geq 0$ for all τ. Then, the following bound holds for all $t \geq 1$:

$$u_t \leq \sqrt{S_t} + \sum_{\tau=1}^{t} \alpha_\tau.$$

The following lemma gives an upper bound on the expected estimation error $\mathbb{E}\left[\|w_0^* - w_t\|\right]$.

Lemma 9 *Under the conditions of Theorem 5, the following bound holds for all $t \ge 1$:*

$$\mathbb{E}\left[\|w_t - w_0^*\|\right] \le \|w_0 - w_0^*\| + \sqrt{\frac{t}{\gamma} R^*} + \frac{6tG}{\gamma}.$$

Proof Recall that $w_0^* = \arg\min_{w \in W^*} \|w_0 - w\|$. Since $\gamma_t \equiv \gamma \ge \frac{19L}{n}$, the second bound in Lemma 7 is valid. For any $t \in [T]$, by summing that inequality with $w^* = w_0^*$ from $\tau = 1, ..., t$ we obtain

$$\begin{aligned} &\sum_{\tau=1}^{t} \mathbb{E}\left[R(w_\tau) - R^*\right] + \gamma \mathbb{E}\left[\|w_t - w_0^*\|^2\right] \\ \le& \gamma\|w_0 - w_0^*\|^2 + \frac{19L}{\gamma n} t R^* + 3n\sum_{\tau=1}^{t}\epsilon_\tau + \sum_{\tau=1}^{t}\left(2\sqrt{2\gamma}\mathbb{E}\left[\|w_0^* - w_\tau\|\right] + 3G\sqrt{\frac{2}{\gamma}}\right)\sqrt{\epsilon_\tau}. \end{aligned} \tag{16}$$

Dropping the non-negative term $\sum_{\tau=1}^{t} \mathbb{E}_{S_{[\tau]}}\left[R(w_\tau) - R^*\right]$ from the above inequality yields

$$\begin{aligned} &\underbrace{\mathbb{E}\left[\|w_t - w_0^*\|^2\right]}_{u_t^2} \\ \le& \|w_0 - w_0^*\|^2 + \frac{19L}{\gamma^2 n} t R^* + \frac{3n}{\gamma}\sum_{\tau=1}^{t}\epsilon_\tau + \sum_{\tau=1}^{t}\left(2\sqrt{\frac{2}{\gamma}}\mathbb{E}\left[\|w_0^* - w_\tau\|\right] + 3G\frac{\sqrt{2}}{\gamma\sqrt{\gamma}}\right)\sqrt{\epsilon_\tau} \\ \overset{\zeta_1}{\le}& \|w_0 - w_0^*\|^2 + \frac{t}{\gamma}R^* + \sum_{\tau=1}^{t}\left(\frac{3n}{\gamma}\epsilon_\tau + \frac{3G\sqrt{2}}{\gamma\sqrt{\gamma}}\sqrt{\epsilon_\tau}\right) + \sum_{\tau=1}^{t}\left(2\sqrt{\frac{2\epsilon_\tau}{\gamma}}\sqrt{\mathbb{E}\left[\|w_0^* - w_\tau\|^2\right]}\right) \\ \le& \underbrace{\|w_0 - w_0^*\|^2 + \frac{t}{\gamma}R^* + \sum_{\tau=1}^{t}\frac{4G\sqrt{2\epsilon_\tau}}{\gamma\sqrt{\gamma}}}_{S_t} + \sum_{\tau=1}^{t}\left(\underbrace{2\sqrt{\frac{2\epsilon_\tau}{\gamma}}}_{\alpha_\tau}\underbrace{\sqrt{\mathbb{E}\left[\|w_0^* - w_\tau\|^2\right]}}_{u_\tau}\right), \end{aligned}$$

where in "ζ_1" we have used $\gamma \ge \frac{19L}{n}$ and the basic inequality $\mathbb{E}^2[X] \le \mathbb{E}[X^2]$, and in the last inequality we have used the condition $\epsilon_\tau \le \frac{2G^2}{9n^2\gamma}$ for all $\tau \ge 1$. By invoking Lemma 8 to the above recursion form we can derive that for all $t \ge 1$,

$$\begin{aligned} \sqrt{\mathbb{E}\left[\|w_t - w_0^*\|^2\right]} \le& \sqrt{\|w_0 - w_0^*\|^2 + \frac{t}{\gamma}R^* + \sum_{\tau=1}^{t}\frac{4G\sqrt{2\epsilon_\tau}}{\gamma}} + \sum_{\tau=1}^{t} 2\sqrt{\frac{2\epsilon_\tau}{\gamma}} \\ \le& \|w_0 - w_0^*\| + \sqrt{\frac{t}{\gamma}R^*} + \sum_{\tau=1}^{t}\sqrt{\frac{4G\sqrt{2\epsilon_\tau}}{\gamma\sqrt{\gamma}}} + \sum_{\tau=1}^{t} 2\sqrt{\frac{2\epsilon_\tau}{\gamma}} \\ \le& \|w_0 - w_0^*\| + \sqrt{\frac{t}{\gamma}R^*} + \frac{6Gt}{\gamma}, \end{aligned}$$

where the last inequality is due to the condition $\epsilon_\tau \le \frac{2G^2}{9\gamma}$ for all $\tau \ge 1$. The above inequality then directly implies the desired bound for all $t \in [T]$. ■

Now we are ready to prove the main result of Theorem 5.

Proof [of Theorem 5] Dropping non-negative term $\gamma\mathbb{E}\left[\|w_t - w^*\|^2\right]$ in (16) followed by natural normalization yields

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[R(w_t)-R^*\right] \\
\leq &\frac{\gamma}{T}\|w_0-w_0^*\|^2+\frac{19L}{\gamma n}R^*+\frac{3n}{T}\sum_{t=1}^{T}\epsilon_t+\frac{1}{T}\sum_{t=1}^{T}\left(2\sqrt{2\gamma}\mathbb{E}\left[\|w_t-w_0^*\|\right]+3G\sqrt{\frac{2}{\gamma}}\right)\sqrt{\epsilon_t} \\
\overset{\zeta_1}{\leq} &\frac{\gamma}{T}\|w_0-w_0^*\|^2+\frac{19L}{\gamma n}R^*+\frac{3n}{T}\sum_{t=1}^{T}\epsilon_t \\
&+\frac{1}{T}\sum_{t=1}^{T}\left(2\sqrt{2}\left(\sqrt{\gamma}\|w_0-w_0^*\|+\sqrt{tR^*}+\frac{6Gt}{\sqrt{\gamma}}\right)+3G\sqrt{\frac{2}{\gamma}}\right)\sqrt{\epsilon_t} \\
\overset{\zeta_2}{\leq} &\frac{\gamma}{T}\|w_0-w_0^*\|^2+\frac{19L}{\gamma n}R^*+\frac{1}{T}\sum_{t=1}^{T}\left(3n\epsilon_t+2\sqrt{2\gamma\epsilon_t}\|w_0-w_0^*\|+2\sqrt{2tR^*\epsilon_t}+\frac{15\sqrt{2\epsilon_t}Gt}{\sqrt{\gamma}}\right) \\
\overset{\zeta_3}{\leq} &\frac{\gamma}{T}\|w_0-w_0^*\|^2+\frac{19L}{\gamma n}R^* \\
&+\frac{1}{T}\sum_{t=1}^{T}\left(3n\epsilon_t+\frac{\gamma\|w_0-w_0^*\|^2}{t^2}+2t^2\epsilon_t+\frac{2LR^*}{\gamma n}+\frac{\gamma n t\epsilon_t}{L}+\frac{15\sqrt{2\epsilon_t}Gt}{\sqrt{\gamma}}\right) \\
\leq &\frac{3\gamma}{T}\|w_0-w_0^*\|^2+\frac{21L}{\gamma n}R^*+\frac{1}{T}\sum_{t=1}^{T}\left(3n\epsilon_t+2t^2\epsilon_t+\frac{\gamma n t\epsilon_t}{L}+\frac{15\sqrt{2\epsilon_t}Gt}{\sqrt{\gamma}}\right),
\end{aligned}
$$

where "ζ_1"follows from Lemma 9, "ζ_2" is due to $t \geq 1$ and "ζ_3" is due to $ab \leq (a^2+b^2)/2$. Now consider $\epsilon_t \leq \frac{\epsilon}{n^2 t^5}$ for some $\epsilon \in [0,1]$. Then it follows from the preceding inequality that

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[R(w_t)-R^*\right] \\
\leq &\frac{3\gamma}{T}\|w_0-w_0^*\|^2+\frac{21L}{\gamma n}R^*+\frac{1}{T}\sum_{t=1}^{T}\left(\frac{3}{nt^5}+\frac{2}{nt^3}+\frac{\gamma}{nLt^4}+\frac{15\sqrt{2}G}{nt^{1.5}\sqrt{\gamma}}\right)\sqrt{\epsilon} \\
\leq &\frac{3\gamma}{T}\|w_0-w_0^*\|^2+\frac{21L}{\gamma n}R^*+\frac{1}{T}\left(\frac{6}{n}+\frac{4}{n}+\frac{2\gamma}{nL}+\frac{45\sqrt{2}G}{n\sqrt{\gamma}}\right)\sqrt{\epsilon}.
\end{aligned}
$$

Let $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$. Combined with the convexity of R, the above inequality implies

$$
\mathbb{E}\left[R(\bar{w}_T)-R^*\right] \lesssim \frac{\gamma}{T}D^2(w_0,W^*)+\frac{L}{\gamma n}R^*+\left(\frac{1}{nT}+\frac{\gamma}{LnT}+\frac{G}{nT\sqrt{\gamma}}\right)\sqrt{\epsilon}.
$$

This proves the first bound. Substituting $\gamma = \sqrt{\frac{T}{n}} + \frac{19L}{n} > \frac{19L}{n}$ into the above bound and preserving the leading terms yields the following second desired bound:

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right) D^2(w_0, W^*) + \frac{L}{\sqrt{nT}} R^* + \left(\frac{L+G}{\sqrt{nT}} + \frac{1}{nT}\right)\sqrt{\epsilon}.$$

The proof is concluded. ■

Appendix C. Proofs for the Results in Section 4

In this section, we present the proofs for the high probability estimation error bounds stated in Section 4.

C.1 Proof of Proposition 1

In this subsection, we prove Proposition 1 as below restated .

Proposition 1 *Suppose that Assumption 1 holds and the loss function is bounded such that $0 \le \ell(y', y) \le M$ for all y, y'. Let $S = \{S_t\}_{t\in[T]}$ and $S' = \{S'_t\}_{t\in[T]}$ be two sets of data minibatches satisfying $S \doteq S'$. Then*

(a) The weighted average output $\bar{w}_T$ and $\bar{w}'_T$ respectively generated by M-SPP (Algorithm 1) over S and S' satisfy

$$\sup_{S,S'} \|\bar{w}_T - \bar{w}'_T\| \le \frac{4\sqrt{2LM}}{n \min_{t\in[T]} \gamma_t} + \sum_{t=1}^{T} 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}.$$

(b) The weighted average output $\bar{w}_T$ and $\bar{w}'_T$ respectively generated by M-SPP-SWoR (Algorithm 3) over S and S' satisfy

$$\sup_{S,S'} \mathbb{E}_{\xi_{[T]}}\left[\|\bar{w}_T - \bar{w}'_T\|\right] \le \sum_{t=1}^{T} \left\{\frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}\right\}.$$

We first need to show the following preliminary result which is about the expansion property of M-SPP update when performed over identical or different minibatches.

Lemma 10 *Suppose that Assumptions 1 holds and the loss function ℓ is bounded in the interval $[0, M]$. From $w_0 = w'_0$, let us define the sequences $\{w_t\}_{t\in[T]}$ and $\{w'_t\}_{t\in[T]}$ that are respectively generated over $\{S_t\}_{t\in[T]}$ and $\{S'_t\}_{t\in[T]}$ according to*

$$F_t(w_t) \le \min_{w\in\mathcal{W}} \left\{F_t(w) := R_{S_t}(w) + \frac{\gamma_t}{2}\|w - w_{t-1}\|^2\right\} + \epsilon_t,$$
$$F'_t(w'_t) \le \min_{w\in\mathcal{W}} \left\{F'_t(w) := R_{S'_t}(w) + \frac{\gamma_t}{2}\|w - w'_{t-1}\|^2\right\} + \epsilon_t.$$

Assume that either $S_t = S'_t$ or $S_t \doteq S'_t$ for all $t \in [T]$. Let $\beta_t = \mathbf{1}_{\{S_t \ne S'_t\}}$. Then the following bound holds for all $t \in [T]$,

$$\|w_t - w'_t\| \le \sum_{\tau=1}^{t} \left\{\beta_\tau \frac{4\sqrt{2LM}}{n\gamma_\tau} + 2\sqrt{\frac{2\epsilon_\tau}{\gamma_\tau}}\right\}.$$

Proof Let $w_t^* = \arg\min_w F_t(w)$ and $w_t'^* = \arg\min_w F_t'(w)$. It follows from Lemma 1 that

$$\begin{aligned}
R_{S_t}(w_t^*) - R_{S_t}(w_t'^*) \leq & \frac{\gamma_t}{2}\left(\|w_t'^* - w_{t-1}\|^2 - \|w_t'^* - w_t^*\|^2 - \|w_t^* - w_{t-1}\|^2\right) \\
R_{S_t'}(w_t'^*) - R_{S_t'}(w_t^*) \leq & \frac{\gamma_t}{2}\left(\|w_t^* - w_{t-1}'\|^2 - \|w_t'^* - w_t^*\|^2 - \|w_t'^* - w_{t-1}'\|^2\right).
\end{aligned}$$

Summing both sides of the above two inequalities yields

$$\begin{aligned}
& R_{S_t}(w_t^*) - R_{S_t}(w_t'^*) + R_{S_t'}(w_t'^*) - R_{S_t'}(w_t^*) \\
\leq & \frac{\gamma_t}{2}\left(\|w_t'^* - w_{t-1}\|^2 - \|w_t^* - w_{t-1}\|^2 + \|w_t^* - w_{t-1}'\|^2 - \|w_t'^* - w_{t-1}'\|^2 - 2\|w_t'^* - w_t^*\|^2\right) \\
= & \frac{\gamma_t}{2}\left(2\langle w_t^* - w_t'^*, w_{t-1} - w_{t-1}'\rangle - 2\|w_t'^* - w_t^*\|^2\right) \\
\leq & \frac{\gamma_t}{2}\left(\|w_{t-1} - w_{t-1}'\|^2 - \|w_t^* - w_t'^*\|^2\right)
\end{aligned}.$$

We need to distinguish the following two complementary cases.

Case I: $S_t = S_t'$. In this case, the previous inequality immediately leads to

$$\|w_t^* - w_t'^*\| \leq \|w_{t-1} - w_{t-1}'\|.$$

By using triangle inequality and Lemma 6 we obtain

$$\|w_t - w_t'\| \leq \|w_t - w_t^*\| + \|w_t^* - w_t'^*\| + \|w_t' - w_t'^*\| \leq \|w_{t-1} - w_{t-1}'\| + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}. \tag{17}$$

Case II: S_t and S_t' differ in a single element. In this case, we have

$$\begin{aligned}
& \|w_t^* - w_t'^*\|^2 \\
\leq & \|w_{t-1} - w_{t-1}'\|^2 + \frac{2}{\gamma_t}\left(R_{S_t}(w_t'^*) - R_{S_t}(w_t^*) + R_{S_t'}(w_t^*) - R_{S_t'}(w_t'^*)\right) \\
= & \|w_{t-1} - w_{t-1}'\|^2 + \frac{2}{\gamma_t}\left(R^\ell_{S_t}(w_t'^*) - R^\ell_{S_t}(w_t^*) + R^\ell_{S_t'}(w_t^*) - R^\ell_{S_t'}(w_t'^*)\right) \\
= & \|w_{t-1} - w_{t-1}'\|^2 + \frac{2}{\gamma_t}\left(\frac{1}{|S_t|}\sum_{z\in S_t}(\ell(w_t'^*; z) - \ell(w_t^*; z) + \frac{1}{|S_t'|}\sum_{z\in S_t'}(\ell(w_t^*; z) - \ell(w_t'^*; z))\right) \\
\leq & \|w_{t-1} - w_{t-1}'\|^2 + \frac{4\sqrt{2LM}}{n\gamma_t}\|w_t^* - w_t'^*\|.
\end{aligned}$$

where in the last inequality we have used $\ell(\cdot;\cdot)$ is $\sqrt{2LM}$-Lipschitz with respect to its first argument which is implied by Lemma 2, and S_t and S_t' differ in a single element as well. Since $x^2 \leq y^2 + ax$ implies $x \leq y + a$ for all $x, y, a > 0$, we can derive from the above that

$$\|w_t^* - w_t'^*\| \leq \|w_{t-1} - w_{t-1}'\| + \frac{4\sqrt{2LM}}{n\gamma_t}.$$

Then based on triangle inequality and Lemma 6 we have

$$\|w_t - w_t'\| \leq \|w_t - w_t^*\| + \|w_t^* - w_t'^*\| + \|w_t' - w_t'^*\| \leq \|w_{t-1} - w_{t-1}'\| + \frac{4\sqrt{2LM}}{n\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}. \tag{18}$$

Let $\beta_t = \mathbf{1}_{\{S_t \neq S'_t\}}$ where $\mathbf{1}_{\{C\}}$ is the indicator function of the condition C. Based on the recursion forms (17) and (18) and the condition $w_0 = w'_0$ we can show that for all $t \in [T]$

$$\|w_t - w'_t\| \le \sum_{\tau=1}^{t} \left\{ \frac{4\beta_\tau \sqrt{2LM}}{n\gamma_\tau} + 2\sqrt{\frac{2\epsilon_\tau}{\gamma_\tau}} \right\},$$

which is the desired bound. ■

Now we are in the position to prove the main result in Proposition 1.

Proof [of Proposition 1] Consider a fixed pair of minibatch sets $S \doteq S'$.

Part (a): Let $\{w_t\}_{t\in[T]}$ and $\{w'_t\}_{t\in[T]}$ be two solution sequences that are respectively generated over $\{S_t\}_{t\in[T]}$ and $\{S'_t\}_{t\in[T]}$ by Algorithm 1. At each time instance t, define random variable $\beta_t := \mathbf{1}_{\{S_t \neq S'_t\}}$. Since by assumption S and S' differ only in a single minibatch, there must exist one and only one $t \in [T]$ such that $\beta_t = 1$ and $\beta_j = 0$ for all $j \in [T], j \neq t$. Then in the worst case of $\beta_\tau = 1$ for $\tau = \arg\min_{i\in[t]} \gamma_i$, it follows from Lemma 10 that for all $t \in [T]$,

$$\|w_t - w'_t\| \le \frac{4\sqrt{2LM}}{n \min_{i\in[t]} \gamma_i} + \sum_{i=1}^{t} 2\sqrt{\frac{2\epsilon_i}{\gamma_i}} \le \frac{4\sqrt{2LM}}{n \min_{i\in[T]} \gamma_i} + \sum_{i=1}^{T} 2\sqrt{\frac{2\epsilon_i}{\gamma_i}}.$$

Then the convex combination nature of $\bar{w}_T$ and $\bar{w}'_T$ implies that

$$\|\bar{w}_T - \bar{w}'_T\| \le \frac{\sum_t \gamma_t \|w_t - w'_t\|}{\sum_t \gamma_t} \le \frac{4\sqrt{2LM}}{n \min_{t\in[T]} \gamma_t} + \sum_{t=1}^{T} 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}.$$

The desired result follows immediately as the above bound holds for any pair $\{S, S'\}$.

Part (b): Recall that $\{\xi_t\}_{t\in[T]}$ are the uniform random indices for iteratively selecting data minibatches from S and S'. Let $\{w_t\}_{t\in[T]}$ and $\{w'_t\}_{t\in[T]}$ be two solution sequences that are respectively generated over $\{S_{\xi_t}\}_{t\in[T]}$ and $\{S'_{\xi_t}\}_{t\in[T]}$ by Algorithm 3. Define random variable $\beta_t := \mathbf{1}_{\left\{S_{\xi_t} \neq S'_{\xi_t}\right\}}$. Since by assumption S and S' differ only in a single minibatch, under without-replacement sampling scheme, there must exist one and only one $t \in [T]$ such that $\beta_t = 1$ and $\beta_j = 0$ for all $j \in [T], j \neq t$. Let us define the event $\mathcal{E}_t := \{\beta_t = 1 \text{ and } \beta_{j\neq t, j\in[T]} = 0\}$ for all $t \in [T]$. Then the uniform randomness of ξ_t implies that

$$R(\mathcal{E}_t) = \frac{1}{T}, \quad t \in [T].$$

Given $t \in [T]$, suppose that $\mathcal{E}_\tau$ occurs for some $\tau \in [t]$. Then it follows from Lemma 10 that

$$\|w_t - w'_t\| \le \frac{4\sqrt{2LM}}{n\gamma_\tau} + \sum_{i=1}^{t} 2\sqrt{\frac{2\epsilon_i}{\gamma_i}}.$$

Suppose that $\mathcal{E}_\tau$ occurs for some $\tau \in \{t+1, t+2, ..., T\}$, again it follows from Lemma 10 that

$$\|w_t - w'_t\| \le \sum_{i=1}^{t} 2\sqrt{\frac{2\epsilon_i}{\gamma_i}}.$$

Then we have

$$\begin{aligned}
\mathbb{E}_{\xi_{[t]}}\left[\|w_t - w_t'\|\right] =& \sum_{\tau=1}^{T} R\left(\mathcal{E}_\tau\right)\left[\|w_t - w_t'\| \mid \mathcal{E}_\tau\right] \\
\leq& \sum_{\tau=1}^{t}\left\{\frac{4\sqrt{2LM}}{nT\gamma_t} + \sum_{i=1}^{t}\frac{2}{T}\sqrt{\frac{2\epsilon_i}{\gamma_i}}\right\} + \sum_{\tau=t+1}^{T}\left\{\sum_{i=1}^{t}\frac{2}{T}\sqrt{\frac{2\epsilon_t}{\gamma_t}}\right\} \\
=& \sum_{\tau=1}^{t}\left\{\frac{4\sqrt{2LM}}{nT\gamma_\tau} + 2\sqrt{\frac{2\epsilon_\tau}{\gamma_\tau}}\right\} \leq \sum_{t=1}^{T}\left\{\frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}\right\}.
\end{aligned}$$

It follows that

$$\mathbb{E}_{\xi_{[T]}}\left[\|\bar{w}_T - \bar{w}_T'\|\right] \leq \frac{\sum_t \gamma_t \mathbb{E}_{\xi_{[t]}}\left[\|w_t - w_t'\|\right]}{\sum_t \gamma_t} \leq \sum_{t=1}^{T}\left\{\frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}\right\}.$$

The desired result follows immediately as the above bound holds for any pair $\{S, S'\}$. ■

C.2 Proof of Theorem 6

In this subsection, we prove Theorem 6 that is restated below.

Theorem 6 *Suppose that Assumptions 1, 2, 3 hold and the loss function ℓ is bounded in the interval $(0, M]$. Let $\rho \in (0, 1/4]$ be an arbitrary scalar and set $\gamma_t = \frac{\lambda\rho t}{4}$. Suppose that $n \geq \frac{76L}{\lambda\rho}$. Assume that $\epsilon_t \leq \min\left\{\frac{\epsilon}{nt^4}, \frac{LM}{\lambda\rho n^2T^2t}\right\}$ for some $\epsilon \in [0,1]$. Then with probability at least $1-\delta$ over S, the weighted average output $\bar{w}_T$ of M-SPP-SWoR (Algorithm 3) satisfies*

$$\begin{aligned}
&\mathbb{E}_{\xi_{[T]}}\left[D(\bar{w}_T, W^*)\right] \\
\lesssim& \frac{\sqrt{LM\log(1/\delta)}\log(T)}{\lambda\rho\sqrt{nT}} + \sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2\rho nT}R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)}.
\end{aligned}$$

To show this result, we need to use the following restated McDiarmid's inequality (McDiarmid, 1989) which is also known as bounded difference inequality.

Lemma 11 (McDiarmid's/Bounded differences inequality) *Let $X_1, X_2, ..., X_N$ be independent random variables valued in $\mathcal{X}$. Suppose that the function $h : \mathcal{X}^N \mapsto \mathbb{R}$ satisfies the bounded differences property, i.e., the following inequality holds for any $i \in [N]$ and any $x_1, ..., x_N, x_i'$:*

$$|h(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_N) - h(x_1, ..., x_{i-1}, x_i', x_{i+1}, ..., x_N)| \leq c_i.$$

Then for any $\varepsilon > 0$,

$$\mathbb{P}\left(h(X_1, ..., X_N) - \mathbb{E}\left[h(X_1, ..., X_N)\right] \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^N c_i^2}\right).$$

Now we are ready to prove Theorem 6.

Proof [of Theorem 6] Let $S = \{S_t\}_{t\in[T]}$ and $S' = \{S'_t\}_{t\in[T]}$ be two sets of data minibatches such that $S \doteq S'$. Then according to Proposition 1 the weighted average output $\bar{w}_T$ and $\bar{w}'_T$ respectively generated by Algorithm 3 over S and S' satisfy

$$\sup_{S,S'} \mathbb{E}_{\xi_{[T]}} \left[\|\bar{w}_T - \bar{w}'_T\|\right] \le \sum_{t=1}^{T} \left\{ \frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}} \right\} \le \sum_{t=1}^{T} \left\{ \frac{5\sqrt{2LM}}{nT\gamma_t} \right\} \le \frac{20\sqrt{2LM}(1+\log(T))}{\lambda\rho nT},$$

where in the last but one inequality we have used the condition $\epsilon_t \le \frac{LM}{4n^2T^2\gamma_t} = \frac{LM}{\lambda\rho N^2 t}$. It follows from the triangle inequality and the above bound that

$$\sup_{S,S'} \mathbb{E}_{\xi_{[T]}} \left[\left|D(\bar{w}_T, W^*) - D(\bar{w}'_T, W^*)\right|\right] \le \sup_{S,S'} \mathbb{E}_{\xi_{[T]}} \left[\|\bar{w}_T - \bar{w}'_T\|\right] \le \frac{20\sqrt{2LM}(1+\log(T))}{\lambda\rho nT}.$$

Since $\xi_{[T]}$ are independent on S, as a direct consequence of applying McDiarmid's inequality with $c_i \equiv c = \frac{20\sqrt{2LM}(1+\log(T))}{\lambda\rho nT}$ to $h(S) := D(\bar{w}_T, W^*)$, we can show that with probability at least $1-\delta$ over the randomness of S,

$$\mathbb{E}_{\xi_{[T]}} \left[D(\bar{w}_T, W^*) - \mathbb{E}_S \left[\mathbb{E}_{\xi_{[T]}} \left[D(\bar{w}_T, W^*)\right]\right]\right] \le c\sqrt{\frac{nT\log(1/\delta)}{2}} = \frac{20\sqrt{LM\log(1/\delta)}(1+\log(T))}{\lambda\rho\sqrt{nT}}.$$

We next derive a bound for $\mathbb{E}_S\left[D(\bar{w}_T, W^*)\right]$. In view of Jensen's inequality and the quadratic growth property of F we have

$$\begin{aligned}
\mathbb{E}_S \left[\mathbb{E}_{\xi_{[T]}} \left[D(\bar{w}_T, W^*)\right]\right] =& \mathbb{E}_{\xi_{[T]}} \left[\mathbb{E}_S \left[D(\bar{w}_T, W^*)\right]\right] \\
\le& \mathbb{E}_{\xi_{[T]}} \left[\sqrt{\mathbb{E}_S \left[D^2(\bar{w}_T, W^*)\right]}\right] \\
\le& \mathbb{E}_{\xi_{[T]}} \left[\sqrt{\frac{2}{\lambda}\mathbb{E}_S \left[R(\bar{w}_T) - R^*\right]}\right] \\
\lesssim& \mathbb{E}_{\xi_{[T]}} \left[\sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2\rho nT}R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)}\right] \\
=& \sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2\rho nT}R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)},
\end{aligned}$$

where in the last inequality we have invoked Theorem 4. Therefore, based on the previous two inequalities we obtain that with probability at least $1-\delta$ over S,

$$\begin{aligned}
&\mathbb{E}_{\xi_{[T]}} \left[D(\bar{w}_T, W^*)\right] \\
\lesssim& \frac{\sqrt{LM\log(1/\delta)}\log(T)}{\lambda\rho\sqrt{nT}} + \sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2\rho nT}R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)},
\end{aligned}$$

which gives the desired bound. ■

C.3 Proof of Theorem 7

Here we prove the following restated Theorem 7.

Theorem 7 *Suppose that Assumptions 1 and 3 hold and the loss function ℓ is bounded in the interval $[0, M]$. Set $\gamma_t \equiv \sqrt{\frac{T}{n}}$. Assume that $\epsilon_t \leq \frac{LM}{4nT^2\sqrt{nT}}$. Then with probability at least $1-\delta$ over S, the average output $\bar{w}_T = \frac{1}{T}\sum_{t=1}^T w_t$ of M-SPP (Algorithm 1) satisfies*

$$|R(\bar{w}_T) - R_S(\bar{w}_T)| \lesssim \frac{(LM + G\sqrt{LM})\log(N)\log(1/\delta)}{\sqrt{nT}} + M\sqrt{\frac{\log(1/\delta)}{nT}}.$$

We need the following lemma essentially from Bousquet et al. (2020, Corollary 8) that gives a near-tight generalization bound for a learning algorithm that is uniformly stable with respect to loss function.

Lemma 12 (Bousquet et al. (2020)) *Suppose that a learning algorithm $\mathcal{A}_w$, parameterized by w, satisfies $|\ell(\mathcal{A}_{w_S}(x), y) - \ell(\mathcal{A}_{w_{S'}}(x), y)| \leq \varrho$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $S \doteq S'$. Assume the loss function satisfies $0 \leq \ell(y', y) \leq M$ for all y, y'. Then for any $\delta \in (0, 1)$, with probability at least $1-\delta$ over an i.i.d. data set S of size N,*

$$|R(\mathcal{A}_{w_S}) - R_S(\mathcal{A}_{w_S})| \lesssim \varrho \log(N) \log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{N}}.$$

With this lemma in place, we can prove the main result in Theorem 7

Proof [of Theorem 7] Let $S = \{S_t\}_{t\in[T]}$ and $S' = \{S'_t\}_{t\in[T]}$ be two sets of data minibatches satisfying $S \doteq S'$. Note that $\gamma_t \equiv \gamma = \sqrt{\frac{T}{n}}$. Then according to Proposition 1 the average output $\bar{w}_T$ and $\bar{w}'_T$ respectively generated by Algorithm 1 over S and S' satisfy

$$\sup_{S,S'} \|\bar{w}_T - \bar{w}'_T\| \leq \frac{4\sqrt{2LM}}{n\gamma} + \sum_{t=1}^T 2\sqrt{\frac{2\epsilon_t}{\gamma}} \leq \frac{5\sqrt{2LM}}{n\gamma} = \frac{5\sqrt{2LM}}{\sqrt{nT}}.$$

where in the last but one inequality we have used the condition $\epsilon_t \leq \frac{LM}{4nT^2\sqrt{N}}$. It follows that

$$|\ell(\bar{w}_T; z) - \ell(\bar{w}'_T; z)| \leq \sqrt{2ML}\|\bar{w}_T - \bar{w}'_T\| \leq \frac{10LM}{\sqrt{nT}},$$

where we have used $\ell(\cdot;\cdot)$ is $\sqrt{2LM}$-Lipschitz with respect to its first argument (which is implied by Lemma 2). In view of Assumption 3 we have

$$|r(\bar{w}_T) - r(\bar{w}'_T)| \leq G\|\bar{w}_T - \bar{w}'_T\| \leq \frac{5G\sqrt{2LM}}{\sqrt{nT}}.$$

This preceding two inequalities indicate that M-SPP is $\frac{10LM+5G\sqrt{2LM}}{\sqrt{nT}}$-uniformly stable with respect to the composite loss function $\ell + r$. By invoking Lemma 12 to M-SPP we obtain that

$$|R(w_S) - R_S(w_S)| \lesssim \frac{(LM + G\sqrt{LM})\log(nT)}{\sqrt{nT}} \log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{nT}}.$$

The proof is concluded. ■

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012. 3, 6

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18:221:1–221:51, 2017. 6

Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM J. Optim.*, 10(4):1116–1135, 2000. 9

Hilal Asi and John C Duchi. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019a. 2, 6, 13, 34

Hilal Asi and John C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM J. Optim.*, 29(3):2257–2290, 2019b. 2, 3, 6, 13, 18, 19, 34

Hilal Asi, Karan N. Chadha, Gary Cheng, and John C. Duchi. Minibatch stochastic approximate proximal point methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020. 2, 3, 5, 7, 18, 19, 23

Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 451–459, Granada, Spain, 2011. 6

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. 7

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020. 7

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163, 2011. 2, 6, 18

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 161–168, Vancouver, Canada, 2007. 7

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018. 5, 6

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002. 4, 7, 16

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 610–626, Virtual Event [Graz, Austria], 2020. 7, 47

Karan Chadha, Gary Cheng, and John C Duchi. Accelerated, optimal, and parallel: Some results on model-based stochastic optimization. *arXiv preprint arXiv:2101.02696*, 2021. 24

Ronan Collobert, Samy Bengio, and Yoshua Bengio. A parallel mixture of svms for very large scale problems. *Neural Comput.*, 14(5):1105–1114, 2002. 22

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, 2006. 6

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019. 2, 3, 4, 5, 6, 10, 18, 19

Aaron Defazio, Francis R. Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1646–1654, Montreal, Canada, 2014. 6

Qi Deng and Wenzhi Gao. Minibatch and momentum model-based methods for stochastic weakly convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 23115–23127, virtual, 2021. 7, 24

Aymeric Dieuleveut, Nicolas Flammarion, and Francis R. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *J. Mach. Learn. Res.*, 18:101:1–101:51, 2017. 4, 17, 18, 19

Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 43(3):919–948, 2018. 4

John C. Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM J. Optim.*, 28(4):3229–3259, 2018. 18

John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In Adam Tauman Kalai and Mehryar Mohri, editors, *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pages 14–26, Haifa, Israel, June 27-29, 2010, 2010. 6

Vitaly Feldman and Jan Vondrák. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9770–9780, Montréal, Canada, 2018. 4, 7

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 1270–1279, Phoenix, AZ, 2019. 7, 16

Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, pages 728–763, Paris, France, 2015. 7

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM J. Optim.*, 22(4):1469–1492, 2012. 6

Isabelle Guyon, Steve R. Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems (NIPS)*, pages 545–552, Vancouver, Canada, 2004. 22

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 1225–1234, New York City, NY, 2016. 7, 16

Chonghai Hu, James T. Kwok, and Weike Pan. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 781–789, Vancouver, Canada, 2009. 6

Martin Jaggi, Virginia Smith, Martin Takác, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I. Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3068–3076, Montreal, Canada, 2014. 6

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, Lake Tahoe, NV, 2013. 3, 6

Ellango Jothimurugesan, Ashraf Tahmasbi, Phillip B. Gibbons, and Srikanta Tirthapura. Variance-reduced stochastic gradient descent on streaming data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9928–9937, Montréal, Canada, 2018. 7

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Part I*, pages 795–811, Riva del Garda, Italy, 2016. 4, 9

Hiroyuki Kasai. SGDLibrary: A MATLAB library for stochastic optimization algorithms. *J. Mach. Learn. Res.*, 18:215:1–215:5, 2017. 22

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5065–5076, virtual, 2021. 7

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. 7

Brian Kulis and Peter L. Bartlett. Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 575–582, Haifa, Israel, 2010. 6

Andrei Kulunchakov and Julien Mairal. A generic acceleration framework for stochastic composite optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12556–12567, Vancouver, Canada, 2019. 6

Guanghui Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133(1-2):365–397, 2012. 6

Jason D. Lee, Qihang Lin, Tengyu Ma, and Tianbao Yang. Distributed stochastic variance reduced gradient methods by sampling extra data with replacement. *J. Mach. Learn. Res.*, 18:122:1–122:43, 2017. 6

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006. 6

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5809–5819, Virtual Event, 2020. 7, 10, 13, 18, 19

Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J. Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 661–670, New York, NY, 2014. 7, 11, 23

C. McDiarmid. Surveys in combinatorics, 1989: On the method of bounded differences. 1989. 45

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018. 7

Sayan Mukherjee, Partha Niyogi, Tomaso A. Poggio, and Ryan M. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.*, 25(1-3):161–193, 2006. 7

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012. 2

Arkadi Nemirovski, Anatoli B. Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4): 1574–1609, 2009. 5, 6

Arkaddii S Nemirovskii and Yu E Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985. 11

Andrei Patrascu and Ion Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *J. Mach. Learn. Res.*, 18:198:1–198:42, 2017. 2, 3, 5, 6, 18, 19

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, UK, 2012. 4, 5, 6, 17, 19

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009. 2

James Renegar and Benjamin Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of Computational Mathematics*, pages 1–46, 2021. 11

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951. 5

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. 1971. 34

Mark Schmidt, Nicolas Le Roux, and Francis R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1458–1466, Granada, Spain, 2011. 39

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, 2010. 7, 11

Ohad Shamir, Nathan Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pages 1000–1008, Beijing, China, 2014. 3, 6

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2199–2207, Vancouver, Canada, 2010. 7, 10, 11, 12, 19, 26

Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017. 5, 6

Panos Toulis, Dustin Tran, and Edoardo M. Airoldi. Towards stability and optimality in stochastic gradient descent. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1290–1298, Cadiz, Spain, 2016. 6, 18

Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Science & Business Media, 2008. 5, 17

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008. 2

Vladimir Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999. 7

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Transactions on Informaiton Theory*, 55(5):2183–2202, 2009. 20

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3636–3645, Sydney, Australia, 2017a. 7, 23

Jialei Wang, Weiran Wang, and Nathan Srebro. Memory and communication efficient distributed stochastic optimization with minibatch prox. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 1882–1919, Amsterdam, The Netherlands, 2017b. 2, 3, 4, 5, 7, 9, 10, 11, 13, 18, 19, 25

Blake E Woodworth and Nathan Srebro. An even more optimal stochastic optimization algorithm: minibatching and interpolation learning. *Advances in Neural Information Processing Systems*, 34:7333–7345, 2021. 4, 18, 19

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim.*, 24(4):2057–2075, 2014. 3, 6, 14

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2020. 11

Xiao-Tong Yuan and Ping Li. On convergence of distributed approximate newton methods: Globalization, sharper bounds and beyond. *Journal of Machine Learning Research*, 21: 1–51, 2020. 3

Xiao-Tong Yuan and Ping Li. Stability and risk bounds of iterative hard thresholding. *IEEE Transactions on Information Theory*, 2022. 7

Lijun Zhang, Tianbao Yang, and Rong Jin. Empirical risk minimization for stochastic convex optimization: $o(1/n)$- and $o(1/n^2)$-type of risk bounds. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 1954–1979, Amsterdam, The Netherlands, 2017. 7, 19

Tong Zhang. Leave-one-out bounds for kernel methods. *Neural Comput.*, 15(6):1397–1437, 2003. 7

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)*, Banff, Canada, 2004. 5

Yuchen Zhang and Xiao Lin. DiSCO: Distributed optimization for self-concordant empirical loss. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 362–370, Lille, France, 2015. 3, 6

Kaiwen Zhou, Lai Tian, Anthony Man-Cho So, and James Cheng. Practical schemes for finding near-stationary points of convex finite-sums. In *International Conference on Artificial Intelligence and Statistics*, pages 3684–3708. PMLR, 2022. 11

Pan Zhou, Xiao-Tong Yuan, Huan Xu, Shuicheng Yan, and Jiashi Feng. Efficient meta learning via minibatch proximal update. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1532–1542, Vancouver, Canada, 2019. 2