Learning Rates for Stochastic Gradient Descent with Nonconvex Objectives

Yunwen Lei and Ke Tang, *Senior Member, IEEE*

Abstract—Stochastic gradient descent (SGD) has become the method of choice for training highly complex and nonconvex models since it can not only recover good solutions to minimize training errors but also generalize well. Computational and statistical properties are separately studied to understand the behavior of SGD in the literature. However, there is a lacking study to jointly consider the computational and statistical properties in a nonconvex learning setting. In this paper, we develop novel learning rates of SGD for nonconvex learning by presenting *high-probability* bounds for both computational and statistical errors. We show that the complexity of SGD iterates grows in a controllable manner with respect to the iteration number, which sheds insights on how an implicit regularization can be achieved by tuning the number of passes to balance the computational and statistical errors. As a byproduct, we also slightly refine the existing studies on the uniform convergence of gradients by showing its connection to Rademacher chaos complexities.

Index Terms—Stochastic Gradient Descent, Learning Rates, Nonconvex Optimization, Early Stopping

1 Introduction

Stochastic algorithms such as stochastic gradient descent (SGD) have become one of the workhorses behind many machine learning tasks [1–3]. As an iterative algorithm, SGD iteratively moves models along the reverse direction of an unbiased gradient estimate built on one or several training examples. Despite its simplicity, SGD has found great success in training highly complex and nonconvex models by the ability to identify good solutions to minimize training errors [3–6]. Surprisingly, the models trained in this way also generalize well to testing examples even in the case with significantly more parameters than training examples [7, 8], which can not be explained well by the traditional generalization analysis based only on the complexity of models [9, 10]. This is especially the case for nonconvex learning problems: while there are multiple local minima for empirical risks, SGD is prone to identify one with good generalization ability. With the recent extensive studies, it is gradually realized that the generalization performance of models depends not only on the complexity of models but also on the optimization algorithms used to train models [7]. In this spirit, there is a growing interest on the theoretical work to understand the success of SGD by considering either its computational or statistical properties.

Roughly speaking, the computational property is related to how the learning algorithm minimizes an empirical risk, while the statistical property is related to the gap between the empirical and population risks on the trained model. It is the interaction of these two factors that determines the learning performance of the trained model [5]. Therefore, it is necessary to take both factors into account for a full understanding on the behavior of the algorithm, which, however, is lacking in the nonconvex setting. Existing theoretical results on SGD in a nonconvex setting are mainly studied for computational errors [11–14]. However, our primary interests in machine learning is the generalization behavior of the trained model on testing examples, which may differ largely from the empirical behavior on training examples due to the possible noises in the data. Motivated by this observation, the generalization gap between empirical and population risks is studied by analyzing the sensitivity of SGD to a small perturbation of the training set, in the framework of algorithmic stability [15–17]. However, the step sizes there need to be very small to enjoy a good stability in the nonconvex setting, which requires an exponential number of iterations for a moderate decay of computational errors. In this way, the resulting learning rates are not quite satisfactory. Very recently, the generalization gap of nonconvex learning is studied via the uniform convergence for the gradients of empirical risks to their population counterparts [18–20]. However, these discussions ignore the interaction between the corresponding training algorithm and the training examples to produce the trained model.

In this paper, we study the learning performance of SGD for nonconvex learning problems from a joint perspective of computational and statistical properties. For the computational properties, we provide a high-probability bound on the decay rate for the gradient of empirical risks (Lemma 4). For the statistical properties, we show that the complexity of models grows in a controllable manner along the optimization process by presenting a high-probability bound on the norm of SGD iterates (Lemma 9). This joint

This work was supported in part by the Guangdong Provincial Key Laboratory under Grant 2020B121201001, the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2017ZT07X386, the National Natural Science Foundation of China under Grant 61806091, the Science and Technology Commission of Shanghai Municipality under Grant 19511120600, and the MOE University Scientific-Technological Innovation Plan Program. (Corresponding author: Ke Tang.)

Yunwen Lei was with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China. He is now with the School of Computer Science, University of Birmingham, Birmingham B15 2TT, United Kingdom (e-mail: y.lei@bham.ac.uk).

Ke Tang is with the Guangdong Key Laboratory of Brain-Inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China, and also with the Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: tangk3@sustech.edu.cn).

perspective sheds intuitive insights on how the computational and statistical errors should be balanced by choosing an optimal iteration number to achieve a good learning rate. As compared to the existing stability-based approach requiring a fast-decaying step-size sequence, our statistical errors are controlled for a general polynomially-decaying step-size sequence for which the computational errors would decay significantly faster. We consider both general nonconvex and gradient-dominated objective functions, for each of which our analysis can achieve, to our best knowledge, the first learning rates that can match the associated ones in the convex setting. As a byproduct, we also derive slightly better uniform convergence rates between gradients of empirical risks and the corresponding population risks by employing the tool of Rademacher chaos complexities [21].

The paper is structured as follows. We formulate the problem in Section 2, and study uniform convergence of gradients in Section 3. We present learning rates in Section 4, and discuss related work in Section 5. We report experimental results in Section 6, and conclude the paper in Section 7. All the proofs can be found in the Supplementary Material.

2 Problem Formulation

Let ρ be a probability measure defined over a compact sample space $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}\subset\mathbb{R}^d\times\mathbb{R}$, where $d\in\mathbb{N}$ is the input dimension. We aim to learn a prediction rule from the input space $\mathcal{X}$ to the output space $\mathcal{Y}$ in a hypothesis space indexed by $\mathcal{W}\subseteq\mathbb{R}^d$. The error of a model at a single example z can be quantified by a loss function $f:\mathcal{W}\times\mathcal{Z}\mapsto\mathbb{R}_+$, from which we can define the population risk (generalization error) by $F:\mathcal{W}\mapsto\mathbb{R}_+$ to measure the performance of a model on testing examples. That is, $F(\mathbf{w})=\mathbb{E}_z\big[f(\mathbf{w};z)\big]$, where $\mathbb{E}_z$ denotes the expectation with respect to (w.r.t.) the random variable z drawn from ρ. The population risk is not accessible since the probability measure ρ is unknown. In practice, we often draw a training dataset $S=\{z_1,\ldots,z_n\}$ independently from ρ and construct an empirical approximation of F by $F_S(\mathbf{w})=\frac{1}{n}\sum_{i=1}^n f(\mathbf{w};z_i)$. The empirical risk F_S takes a finite-sum structure, for which an efficient method to use this structure is SGD. Let $\mathbf{w}_1=0$ and $\{\eta_t\}_{t\in\mathbb{N}}$ be a sequence of positive step sizes. At the t-th iteration, we first draw an index j_t from the uniform distribution over $\{1,\ldots,n\}$, and update the model by

$$\mathbf{w}_{t+1}=\mathbf{w}_t-\eta_t\nabla f(\mathbf{w}_t;z_{j_t}),\tag{1}$$

where $\nabla f(\mathbf{w}_t;z_{j_t})$ denotes the gradient of f w.r.t. the first argument. It should be noted that the randomness of the iterate $\mathbf{w}_t$ comes from two sources: one from the sampling of training examples according to ρ and one from the sampling of the indices $\{j_t\}_t$ according to the uniform distribution over $\{1,\ldots,n\}$. In this paper, we are interested in the generalization behavior of $\{\mathbf{w}_t\}$ measured by the gradient of population risks $\nabla F(\mathbf{w}_t)$. We introduce some notations in this paper. Denote $b=\sup_{z\in\mathcal{Z}}f(0;z)$, $\tilde{b}=\sup_{z\in\mathcal{Z}}\|\nabla f(0,z)\|_2$ and $\kappa=\sup_{x\in\mathcal{X}}\|x\|_2$, where $\|\cdot\|_2$ is the Euclidean norm. For any $R>0$, define $B_R=\{\mathbf{w}\in\mathbb{R}^d:\|\mathbf{w}\|_2\le R\}$. Let e be the base of the natural logarithm.

2.1 Assumptions

We need some assumptions. We say a differentiable function $g:\mathcal{W}\mapsto\mathbb{R}$ is L-smooth for a constant $L>0$ if

$$\big\|\nabla g(\mathbf{w})-\nabla g(\tilde{\mathbf{w}})\big\|_2\le L\|\mathbf{w}-\tilde{\mathbf{w}}\|_2,\quad\forall\mathbf{w},\tilde{\mathbf{w}}\in\mathcal{W}.$$

For L-smooth g, we have two useful properties [22]

$$g(\mathbf{w})\le g(\tilde{\mathbf{w}})+\big\langle\mathbf{w}-\tilde{\mathbf{w}},\nabla g(\tilde{\mathbf{w}})\big\rangle+2^{-1}L\|\mathbf{w}-\tilde{\mathbf{w}}\|_2^2,\tag{2}$$

$$(2L)^{-1}\|\nabla g(\mathbf{w})\|_2^2\le g(\mathbf{w})-\inf_{\mathbf{w}}g(\mathbf{w}).\tag{3}$$

Our first assumption is a standard and widely used assumption on the smoothness of loss functions [11–13, 23].

Assumption 1. We assume that for any z, the function $\mathbf{w}\mapsto f(\mathbf{w};z)$ is L-smooth.

In our next assumption, we suppose that a weighted norm of gradient is bounded.

Assumption 2. We assume the existence of $G>0$ such that

$$\sqrt{\eta_t}\|\nabla f(\mathbf{w}_t;z)\|_2\le G,\quad\forall t\in\mathbb{N},z\in\mathcal{Z}.$$

In the literature, a bounded gradient assumption as $\|\nabla f(\mathbf{w}_t;z)\|_2\le G$ (i.e., the Lipschitz continuity of f) is often imposed to study statistical properties of SGD [15–17, 24] as well as the computational properties [11, 25]. As compared to this bounded gradient assumption, Assumption 2 is much milder since the step sizes should diminish to zero for the convergence of the algorithm. Indeed, typical choices of step sizes are $\eta_t=O(1/\sqrt{t})$ and $\eta_t=O(1/t)$ (depending on whether the objective functions satisfy additional conditions such as gradient-dominance condition) [13, 26], in which case, the gradients can respectively grow with the rate $O(t^{\frac{1}{4}})$ and $O(t^{\frac{1}{2}})$ without violating Assumption 2.

Our third assumption is on the boundedness of variances of stochastic gradients, which is widely adopted in the literature to study either computational errors [13, 14] or stability of SGD [16, 27] in the nonconvex setting.

Assumption 3. We assume the existence of $\sigma>0$ such that

$$\mathbb{E}_{j_t}\big[\big\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\big\|_2^2\big]\le\sigma^2,\quad\forall t\in\mathbb{N},$$

where $\mathbb{E}_{j_t}$ denotes the expectation w.r.t. j_t.

2.2 Structure of loss functions

Other than the general smooth loss functions, we also consider in particular a class of loss functions of the structure

$$f(\mathbf{w};z)=\phi(\langle\mathbf{w},x\rangle,y),\tag{4}$$

where $\phi:\mathbb{R}^2\mapsto\mathbb{R}_+$. Functions of this structure can be found in *generalized linear models* and *robust regression*.

1. For generalized linear models in binary classification, we would like to learn a model of the form $\Pr\{Y=1|X=x\}=\ell(\langle\mathbf{w}^*,x\rangle)$ with $\mathbf{w}^*\in\mathbb{R}^d$ being a parameter vector and $\ell:\mathbb{R}\mapsto[0,1]$ being a link function [18]. A commonly used loss is the non-linear squared loss $f(\mathbf{w};z)=\big(y-\ell(\langle\mathbf{w},x\rangle)\big)^2$ which can be written as the form of (4) with $\phi(a,b)=(\ell(a)-b)^2$. Standard choices of ℓ include the logistic link function $\ell(s)=(1+e^{-s})^{-1}$ and the probit link function $\ell(s)=\Phi(s)$ with Φ being the Gaussian cumulative distribution function.

2. For robust regression, we pick a potentially nonconvex function $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ and assume a linear model $y = \langle \mathbf{w}^*, x\rangle + \xi$, where the noise term ξ is i.i.d. with mean zero [18, 20]. The loss function is then $f(\mathbf{w}; z) = \ell(y - \langle \mathbf{w}, x\rangle)$, which takes a form of (4) with $\phi(a, b) = \ell(b - a)$.

For loss functions of the structure (4), we will develop dimensionality-independent learning rates which would be appealing for the high-dimensional learning setting.

3 Uniform Convergence of Gradients

We are interested in the population gradients ∇F instead of empirical gradients ∇F_S. The gap between these two terms can be studied via the uniform convergence on the gradient of empirical risks to the population counterparts [18–20, 28]. In this section, we continue this direction by presenting a new connection between the uniform convergence of gradients and the Rademacher chaos complexities of order two, which is an extension of Rademacher complexities to U-processes [21, 29]. As we will see, this connection allows us to improve the existing uniform convergence by removing a logarithmic factor. It should be noted that the uniform convergence of gradients is not our main contribution.

Definition 1. Let $\mathcal{F} : \mathcal{Z}\times\mathcal{Z} \mapsto \mathbb{R}$ be a function class and $S = \{z_i\}_{i=1}^n \subset \mathcal{Z}$. Let $\{\epsilon_i\}_{i=1}^n$ be independent Rademacher variables with $\Pr\{\epsilon_i = 1\} = \Pr\{\epsilon_i = -1\} = 1/2$. The empirical Rademacher chaos complexity of order two for $\mathcal{F}$ w.r.t. S is defined as $\mathcal{U}_S(\mathcal{F}) = \frac{1}{n}\mathbb{E}_\epsilon\Big[\sup_{f\in\mathcal{F}} \sum_{1\le i<j\le n} \epsilon_i\epsilon_j f(x_i, x_j)\Big]$.

Theorem 1. *Let $\delta \in (0, 1), R > 0$ and $S = \{z_1, \ldots, z_n\}$ be examples drawn independently from ρ. Suppose Assumption 1 holds. Then with probability at least $1 - \delta$ we have*

$$\sup_{\mathbf{w}\in B_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 \le \frac{LR + \tilde{b}}{\sqrt{n}}\big(2 + \sqrt{2\log(1/\delta)}\big) + 2\sqrt{\frac{2\mathcal{U}_S(\mathcal{F}_R)}{n}},$$

where $\mathcal{F}_R = \big\{(z, \tilde{z}) \mapsto \langle\nabla f(\mathbf{w}; z), \nabla f(\mathbf{w}; \tilde{z})\rangle : \mathbf{w} \in B_R\big\}$.

As a corollary, we can derive the following bound by controlling the Rademacher chaos complexities in Theorem 1. The proofs of Theorem 1 and Corollary 2 are given in Section II.1 and Section II.2 (Supplementary Material).

Corollary 2. *Under Assumptions of Theorem 1, the following inequality holds with probability $1 - \delta$*

$$\sup_{\mathbf{w}\in B_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 \le (LR + \tilde{b})n^{-\frac{1}{2}}\times \Big(2 + 2\sqrt{48e\sqrt{2}\big(\log 2 + d\log(3e)\big)} + \sqrt{2\log(1/\delta)}\Big).$$

Remark 1. Based on covering numbers, it was shown [18]

$$\sup_{\mathbf{w}\in B_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 = O\Big(\sqrt{\frac{d\log(n)\log(1/\delta)}{n}}\Big).$$

As a comparison, our argument based on Rademacher chaos complexities yields the bound $O(\sqrt{\frac{d+\log(1/\delta)}{n}})$, which is slightly tighter than the above one by removing the term $\log(n)$. Furthermore, our bound involves $d + \log(1/\delta)$ while the above bound involves $d\log(1/\delta)$. Although the improvement by removing a logarithmic factor is marginal, the obtained bound is generally the best one can get. These improvements are achieved by establishing a new connection between uniform convergence of gradients and Rademacher chaos complexities (Theorem 1), which might be interesting on its own due to the important role of Rademacher-type complexities in learning theory.

The uniform convergence rates in Corollary 2 involves the dimensionality d. Dimensionality-free bounds are possible if we consider loss functions f of the structure (4). Proposition 3 is a direct corollary of the chain rule for vector-valued Rademacher complexities [20]. For completeness, we provide a complete proof in Section II.3 (Supplementary Material) based on Gaussian complexities.

Proposition 3. *Let $\delta \in (0, 1), R > 0$ and $S = \{z_1, \ldots, z_n\}$ be examples drawn independently from ρ. Suppose $f : \mathcal{W}\times\mathcal{Z} \mapsto \mathbb{R}$ takes the form* (4) *with ϕ being L_ϕ-smooth w.r.t. the first argument. Then with probability $1 - \delta$, $\sup_{\mathbf{w}\in B_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2$ is upper bounded by*

$$\frac{2\sqrt{2}\kappa\big(2L_\phi R\kappa + \tilde{b}\big)}{\sqrt{n}} + \sqrt{\frac{2\big(L_\phi\kappa^2 R + \tilde{b}\big)^2\log(1/\delta)}{n}}.$$

4 Learning Rates

In this section, we present our main results on the behavior of gradients for population risks of SGD iterates. We consider general nonconvex objectives in Section 4.1, and then move to a subclass of nonconvex objectives satisfying the gradient-dominance condition in Section 4.2.

4.1 General nonconvex objective functions

Before providing learning rates, we first present a high-probability bound on the gradients of empirical risks to be proved in Section III.1 (Supplementary Material).

Lemma 4. *Suppose Assumptions 1-3 hold. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by* (1) *with $\eta_t \le 1/(2L)$. Then for any $\delta \in (0, 1)$, the following inequality holds with probability $1 - \delta$*

$$\sum_{k=1}^{t} \eta_k\|\nabla F_S(\mathbf{w}_k)\|_2^2 \le C_t + C_1\log(2/\delta), \tag{5}$$

where we introduce $C_t = 4b + 4L\big(\sigma^2 + G^2\big)\sum_{k=1}^t \eta_k^2$ and $C_1 = 8\max\{G^2, \sigma^2/L\} + 32LG^2$.

Remark 2. Under Assumptions 1 and 3, it was shown for SGD with $\eta_t \le 1/(2L)$ that [13] $\sum_{k=1}^t \eta_k\mathbb{E}[\|\nabla F_S(\mathbf{w}_k)\|_2^2] = O(\sum_{k=1}^t \eta_k^2 + 1)$, from which and the Markov's inequality it follows that $\Pr\{\sum_{k=1}^t \eta_k\mathbb{E}[\|\nabla F_S(\mathbf{w}_k)\|_2^2] \ge \delta\} \le \delta^{-1}O(\sum_{k=1}^t \eta_k^2 + 1)$. As a comparison, we develop a high-probability bound where the dependency on the confidence parameter $1/\delta$ is logarithmic instead of linear.

In the following theorem, we present our main result on the decay rate of population gradients in general nonconvex cases. As we will see in the proof given in Section III.3 (Supplementary Material), the term $\|\nabla F(\mathbf{w}_t)\|_2^2$ depends on two terms: the computational error $\|\nabla F_S(\mathbf{w}_t)\|_2^2$ and the statistical error $\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|_2^2$. Generally, the computational error decreases as a function of the iteration number, while the statistical error would increase along the learning process ($\mathbf{w}_t$ traverses over a ball with an increasing radius as t increases, and therefore one has to apply Theorem 1 with an increasing R). Therefore, one should trade-off these two errors by selecting an appropriate iteration number. We denote $A \asymp B$ if there exist universal constants $\widetilde{C}_1, \widetilde{C}_2 > 0$ such that $\widetilde{C}_1 A \leq B \leq \widetilde{C}_2 A$.

Theorem 5. *Suppose Assumptions 1-3 hold. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by* (1) *with $\eta_t = \eta_1 t^{-\theta}, \theta \in (0,1)$ and $\eta_1 \leq 1/(2L)$. Then for any $\delta \in (0,1)$, we can choose $T \asymp (nd^{-1})^{\frac{1}{2(1-\theta)}}$ to derive with probability $1-\delta$*

$$\Big(\sum_{t=1}^{T}\eta_t\Big)^{-1}\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 = \begin{cases} O\Big((nd^{-1})^{\frac{\theta}{2(\theta-1)}}\log^3(1/\delta)\Big), & \text{if } \theta < 1/2,\\ O\Big(n^{-\frac{1}{2}}d^{\frac{1}{2}}\log(T/\delta)\log^3(1/\delta)\Big), & \text{if } \theta = 1/2,\\ O\Big(n^{-\frac{1}{2}}d^{\frac{1}{2}}\log^3(1/\delta)\Big), & \text{if } \theta > 1/2. \end{cases} \quad (6)$$

Remark 3. According to Theorem 5, one can select an appropriate early-stopping iteration number to achieve similar learning rates for polynomially decaying step sizes with $\theta \in (1/2, 1)$. However, the corresponding computational cost measured by $T \asymp (nd^{-1})^{\frac{1}{2(1-\theta)}}$ varies for different θ, which achieves its minimum by taking θ close to $1/2$.

Remark 4. We compare our results with the existing stability bounds. For the specific step-size sequence $\eta_t = O(1/t)$, stability bounds $O(T^\alpha/n)$ with $\alpha \in (0,1)$ were established for SGD with T iterations in nonconvex learning (Theorem 3.8 in [15]), from which one can derive $\mathbb{E}[F(\mathbf{w}_T) - F_S(\mathbf{w}_T)] = O(T^\alpha/n)$. As compared to our uniform convergence approach, this stability approach has an appealing property of implying dimensionality-independent learning rates. However, the stability-based approach requires a fast-decaying step sizes as $\eta_t = O(1/t)$, for which the associated computational errors would generally decay as

$$\min_{t=1,\ldots,T}\mathbb{E}[\|\nabla F_S(\mathbf{w}_t)\|_2^2] = O\Big(\big(1+\sum_{t=1}^{T}\eta_t^2\big)/\sum_{t=1}^{T}\eta_t\Big) = O(\log^{-1}(T)),$$

where the first step is due to [13] and the second step is due to the elementary inequalities $\sum_{t=1}^{T} t^{-2} = O(1), \sum_{t=1}^{T} t^{-1} \asymp \log T$. To get a meaningful statistical error bound, one requires $T = o(n^{\frac{1}{\alpha}})$, for which the associated computational error bound decays as $O(\log^{-1}(n))$. As a comparison, our statistical errors are developed for the step sizes $\eta_t = O(t^{-\theta}), \theta \in (0,1)$, for which the associated computational errors decay significantly faster as $O\big(T^{\max\{\theta-1,-\theta\}}\big)$. The ability to balance the statistical and computational errors with a general step-size sequence allows us to develop learning rate of the order $\widetilde{O}((n^{-1}d)^{\frac{1}{2}})$ in the general nonconvex learning setting, where we use the notation $\widetilde{O}$ to hide logarithmic factors. It is now clear that each of these two approaches has its own advantages over the other, depending on whether the dependency on the dimensionality or the iteration number is more interesting. In more details, the stability approach is preferable in a high-dimensional learning setting with $n = O(d)$, while the uniform-convergence approach is better for low-dimensional learning problems with $d = o(n)$.

If the loss function takes a specific structure, we can improve the learning rates in Theorem 5 by removing the dependency on the dimensionality. Theorem 6 can be proved similarly to Theorem 5 by using Proposition 3 instead of Corollary 2, and we omit the proof for brevity.

Theorem 6. *Let assumptions in Theorem 5 hold. If $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}$ takes the form* (4) *with ϕ being L_ϕ-smooth w.r.t. the first argument, we can choose $T \asymp n^{\frac{1}{2(1-\theta)}}$ to derive Eq.* (6) *with the involved d removed.*

Remark 5. We can argue the effectiveness of learning rates in Theorem 6 as follows. If we assume ϕ is convex w.r.t. the first argument, the best known generalization bound of SGD is of the order $F(\bar{\mathbf{w}}_T) - F(\mathbf{w}^*) = \widetilde{O}(n^{-\frac{1}{2}})$, where $\bar{\mathbf{w}}_T$ is an average of the first T iterates with an appropriately chosen T and $\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w})$ [15, 30, 31]. It then follows from (3) that $\|\nabla F(\bar{\mathbf{w}}_T)\|_2^2 = \widetilde{O}(n^{-\frac{1}{2}})$ with high probabilities. This matches our learning rates derived in the nonconvex learning setting in Theorem 6.

4.2 Gradient-dominated objective functions

In this section, we show that a faster learning rate is possible if we impose a gradient-dominance condition on the objective function. Roughly speaking, gradient-dominance condition means that the suboptimality of function values can be bounded by the squared magnitude of gradients. It also means that every stationary point must be a global optimum. The gradient-dominance condition is widely used in the analysis of nonconvex learning algorithms, see, e.g., [11, 17, 20, 26, 32]. Many nonconvex objective functions satisfy the gradient-dominance condition, including one hidden layer neural networks, ResNets with linear activations, phase retrieval and matrix factorization [20, 32].

Assumption 4 (Gradient-dominance condition)**.** We assume for all $S \subset \mathcal{Z}$, there exists an $\mu_S > 0$ such that

$$F_S(\mathbf{w}) - F_S(\mathbf{w}_S) \leq (4\mu_S)^{-1}\|\nabla F_S(\mathbf{w})\|_2^2, \quad \forall \mathbf{w} \in \mathcal{W},$$

where $\mathbf{w}_S = \arg\min_{\mathbf{w}} F_S(\mathbf{w})$.

For gradient-dominated objectives, we can implement SGD with $O(n)$ iterations to achieve the learning rates $\widetilde{O}(n^{-1}d)$ for excess risks instead of gradients. The proof of Theorem 7 is given in Section IV (Supplementary Material).

Theorem 7. *Suppose Assumptions 1-4 hold. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by* (1) *with $\eta_t = 2/(\mu_S(t+t_0))$ for all $t \in \mathbb{N}$ and $t_0 \geq \max\{4L/\mu_S, 1\}$. Let $\delta \in (0,1)$. Then we can choose $T \asymp n$ to derive the following inequality with probability $1-\delta$*

$$F(\mathbf{w}_T) - F(\mathbf{w}^*) = O\Big(n^{-1}(d + \log(1/\delta))\log^2 n \log^2(1/\delta)\Big). \quad (7)$$

Remark 6. We compare our results with the existing stability analysis. For gradient-dominated objectives, computational error bounds of the order $O(1/T)$ were developed for step sizes $\eta_t = \frac{2t+1}{2\mu_S(t+1)^2}$ [26]. For this step-size sequence, uniform stability bounds were shown to decay as $O(n^{-1}T^{\frac{\alpha}{\alpha+1}})$ (Theorem 3.8 of [15]), where $\alpha = L/\mu_S$ behaves as a condition number. Therefore, the stability approach implies $\mathbb{E}[F(\mathbf{w}_T)]-F(\mathbf{w}^*) = O\big(T^{-1}+n^{-1}T^{\frac{\alpha}{\alpha+1}}\big)$, and one can choose $T \asymp n^{\frac{\alpha+1}{2\alpha+1}}$ to derive the learning rate $O(n^{-\frac{\alpha+1}{2\alpha+1}})$. As compared to Theorem 7, it is clear that the stability-based rates have an advantage of being dimensionality-free but meanwhile a worse dependency on n. Indeed, the stability approach is preferable if $n = O(d^{\frac{2\alpha+1}{\alpha}})$, while our learning rate is better if $d = O(n^{\frac{\alpha}{2\alpha+1}})$.

Analogous to the case with general nonconvex objectives, we can remove the dependency of learning rates on the dimensionality for loss functions of the structure (4). We omit the proof of Theorem 8 for brevity.

Theorem 8. *Let assumptions in Theorem 7 hold. If $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}$ takes the form* (4) *with ϕ being L_ϕ-smooth w.r.t. the first argument, then we can choose $T \asymp n$ to derive* (7) *with the involved d removed.*

Remark 7. The effectiveness of learning rates in Theorem 8 can be shown as follows. If we assume ϕ is Lipschitz continuous and strongly convex w.r.t. the first argument, the best known generalization bound of the order $F(\bar{\mathbf{w}}_T)-F(\mathbf{w}^*) = \widetilde{O}(T^{-1})$ was derived in the online learning setting, where $\bar{\mathbf{w}}_T$ is an average of the first T iterates [33]. This online learning setting corresponds to the one-pass SGD since the released training examples there need to be independently drawn from ρ. Therefore the results in [33] correspond to the generalization bound $F(\bar{\mathbf{w}}_T) - F(\mathbf{w}^*) = \widetilde{O}(n^{-1})$ in the stochastic learning setting with T chosen proportionally to the sample size, which matches the bounds in Theorem 8.

4.3 Sketch of the proof

We sketch here our main idea in proving our results in Section 4. Our learning rates are derived by controlling the population gradients in terms of two components as follows

$$\|\nabla F(\mathbf{w}_t)\|_2^2 \le 2\|\nabla F_S(\mathbf{w}_t)\|_2^2 + 2\|\nabla F_S(\mathbf{w}_t) - \nabla F(\mathbf{w}_t)\|_2^2.$$

We refer to the first term as the computational error since it is related to the optimization algorithm to minimize the empirical risk F_S. The second term is called the statistical error since it is related to approximating the true gradient with its empirical counterpart based on random samples.

Computational errors have received much attention in the optimization community [11–13, 23]. However, existing results are mainly studied in expectation, while our focus here is to establish high-probability bounds (Lemma 4). Our idea to this aim is to bound $\sum_{k=1}^t \eta_k\|\nabla F_S(\mathbf{w}_k)\|_2^2$ in terms of two martingale difference sequences where the variance of martingales plays an essential role in deriving the stated high probability bounds. To apply uniform convergence of gradients in Section 3 to control statistical errors, an essential step is to control the norm of iterates in the following lemma to be proved in Section III.2 (Supplementary Material).

Lemma 9. *Suppose Assumptions 1-3 hold. Let $\{\mathbf{w}_t\}_t$ be produced by* (1) *with $\eta_t \le 1/(2L)$. Then for any $\delta \in (0,1)$, with probability $1-\delta$ we have uniformly for all $t = 1,\ldots,T$*

$$\|\mathbf{w}_{t+1}\|_2 \le C_2\Big(\Big(\sum_{k=1}^T \eta_k^2\Big)^{\frac{1}{2}}+1\Big)\Big(\Big(\sum_{k=1}^t \eta_k\Big)^{\frac{1}{2}}+1\Big)\log(4/\delta), \quad (8)$$

where C_2 is independent of T and δ (explicitly given in the proof).

Our idea in proving Lemma 9 is to use (1) to get

$$\|\mathbf{w}_{t+1}\|_2 \le \Big\|\sum_{k=1}^t \eta_k \nabla f(\mathbf{w}_k; z_{j_k})\Big\|_2 \le \Big\|\sum_{k=1}^t \eta_k \nabla F_S(\mathbf{w}_k)\Big\|_2 + \Big\|\sum_{k=1}^t \eta_k\big(\nabla F_S(\mathbf{w}_k) - \nabla f(\mathbf{w}_k; z_{j_k})\big)\Big\|_2.$$

The last second term of the above inequality can be controlled by Lemma 4 together with the Schwartz's inequality. Furthermore, the term $\sum_{k=1}^t \eta_k\big(\nabla F_S(\mathbf{w}_k) - \nabla f(\mathbf{w}_k; z_{j_k})\big)$ is a summation of a martingale difference sequence, which we can control by concentration inequalities of martingales.

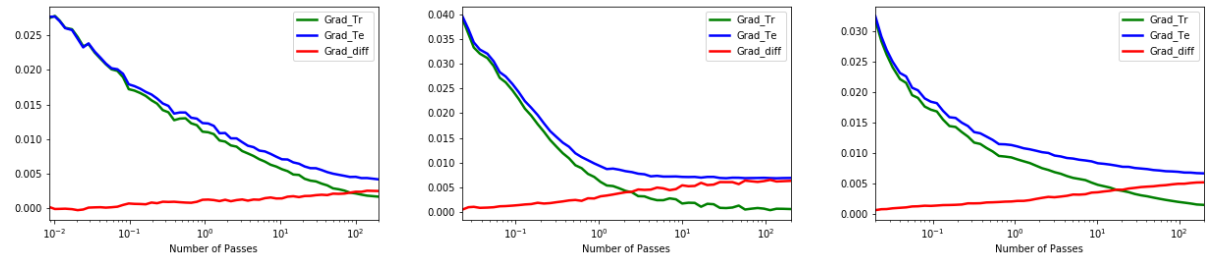
5 Discussions

We discuss here related work on computational errors, uniform convergence of gradients and stability-based analysis.

Computational errors. SGD for nonconvex learning is mainly studied from the perspective of computational errors. Initially, asymptotic properties of SGD were studied [34]. The key property on the smoothness of loss functions was used to establish the first nonasymptotic convergence rates in expectation $\mathbb{E}[\|\nabla F_S(\mathbf{w}_u)\|_2^2] = O(1/\sqrt{T})$ where u is drawn from a probability distribution on $\{1,\ldots,T\}$ [13]. For gradient-dominated and smooth objectives, it was further shown that $\mathbb{E}[F_S(\mathbf{w}_T)] - \inf_{\mathbf{w}} F_S(\mathbf{w}) = O(1/T)$ [26]. These discussions were then extended successfully to other variants of SGD for nonconvex optimization, including distributed SGD [25], stochastic composite optimization [6, 14] and stochastic variance reduction [11, 12]. In particular, near optimal iteration complexities were recently achieved by some interesting variance-reduced algorithms for nonconvex optimization [35–38].

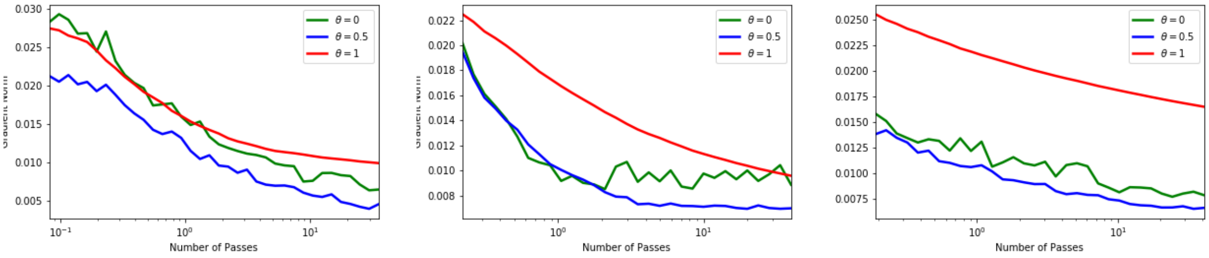
Uniform convergence of gradients. In nonconvex learning, convergence rates are generally stated for the gradients of empirical risks [11–13, 39], which not necessarily means that similar convergence rates can be carried to their population counterparts. Motivated by this, the uniform convergence of gradients is recently drawing increasing attention [18–20]. Based on the tool of covering numbers, convergence rates $O(\sqrt{n^{-1}\log n})$ were established for the uniform deviation between population and empirical gradients [18, 19]. A chain rule for vector-valued Rademacher complexities was established to show the uniform convergence of gradients for nonconvex functions with the structure (4) [20]. The above mentioned uniform convergence results are established for smooth functions. Recently, an interesting graphical convergence was studied for nonconvex and nonsmooth loss functions, for which the convergence is measured by the gradient of the Moreau envelops [40]. Similar to our results in Section 3, the graphic convergence in [40] involves a square-root dependency on the dimension d in general, and is dimensionality-independent for loss functions of the structure (4). These discussions apply to empirical risk minimization, which do not take into account the computational property of learning algorithms.

Stability-based analysis. Other than the uniform convergence approach, another popular approach to investigate the generalization behavior of a learning algorithm is to study its algorithmic stability [41–43]. An advantage of



(a) Segment (b) Fourclass (c) German

Fig. 1. Computational errors, statistical errors and test errors versus the number of passes for some datasets. Grad_Tr, Grad_Te and Grad_diff refer to computational errors, test errors and statistical errors, respectively.



(a) Segment (b) Fourclass (c) German

Fig. 2. Test errors versus the number of passes for step sizes of the form $\eta_t = \eta_1 t^{-\theta}, \theta \in \{0, 1/2, 1\}$ and some datasets.

stability-based approach is that it considers the specific model produced by the learning algorithm, and therefore can imply dimensionality-independent learning rates. As a comparison, the dependence on the dimensionality is generally inevitable for the uniform convergence approach. Initially, stability bounds are especially useful to derive optimal generalization error bounds stated *in expectation* [44]. Recently, almost optimal generalization bounds *with high probability* were developed for uniformly-stable algorithms in a series of breakthrough papers [45–47]. The uniform stability of SGD was established in a seminar paper [15], which motivates some interesting work on data-dependent stability [16] and on-average stability of SGD [17, 44]. However, stability bounds there are mainly derived for the specific step-size sequence $\eta_t = O(1/t)$ in nonconvex learning, for which one generally needs an exponential number of iterations for a moderate decay of the associated computational errors [13]. Under gradient-dominance conditions or quadratic growth conditions, black-box stability results for first-order methods were established [32]. However, the discussions there require to impose a strict assumption on the convergence of a learning algorithm to global optimum which is difficult to hold for nonconvex problems. Stability of stochastic gradient Langevin dynamics is also drawing increasing attention recently, where isotropic Gaussian noises are added to each stochastic gradient steps [24].

The primary focus of stability bounds is to control the generalization gap $F(\mathbf{w}_T) - F_S(\mathbf{w}_T)$ which ignores computational errors, while we aim to study the gradients of population risks where the interplay between computational and statistical errors plays an important role to achieve satisfactory learning rates. Unlike existing bounds stated in expectation [15–17, 24, 32], our analysis is able to develop more challenging high-probability bounds. A key observation is to show that the complexity of SGD iterates grows in a controllable manner w.r.t. the step-size sequence (Lemma 9), which allows us to relate the gradients of empirical risks to their population counterparts via the uniform convergence of gradients. Our analysis therefore provides an alternative approach to study learning rates of SGD in nonconvex learning without considering the stability of an algorithm. As stated in Remark 5 and Remark 7, the effectiveness of our learning rates is justified by matching the existing generalization bounds in the convex learning setting. Furthermore, our statistical errors are developed for step sizes $\eta_t = O(t^{-\theta})$ with $\theta \in (0, 1)$ in the general nonconvex learning setting, for which the associated computational errors decay significantly faster than $O(\log^{-1} T)$ achieved by the step sizes $\eta_t = O(t^{-1})$. Although our learning rates are dimensionality-dependent in general, one can exploit the specific structure of loss functions to derive dimensionality-independent bounds in specific cases.

For convex learning, generalization bounds of SGD were studied via algorithmic stability [15, 16], integral operators [48–50] and tools in empirical process [30, 31].

6 Simulations

In this section, we present some simulation results to validate our theory. According to Section 2.2, we consider a generalized linear model for binary classification where the loss function takes the form $f(\mathbf{w}; z) = \big(\ell(\langle \mathbf{w}, x\rangle) - y\big)^2$ and ℓ is the logistic link function $\ell(s) = (1 + e^{-s})^{-1}$. We consider three datasets available from the LIBSVM dataset:

Segment, Fourclass and German [51]. For each dataset, we take 80 percents as the training dataset, and reserve the remaining 20 percents for testing. Let $F_S(\mathbf{w})$ and $F_{\widetilde{S}}(\mathbf{w})$ be the objective function built on the training dataset S and the testing dataset $\widetilde{S}$, i.e., $F_{\widetilde{S}}(\mathbf{w}) = \frac{1}{|\widetilde{S}|}\sum_{z\in\widetilde{S}} f(\mathbf{w};z)$, where $|\widetilde{S}|$ denotes the cardinality of the set $\widetilde{S}$. It is clear that $F_{\widetilde{S}}$ serves as a good approximation of F. We repeat experiments 100 times and report the average of results.

Our first aim is to illustrate how the computational and statistical errors behave versus the number of passes, which is the iteration number divided by the sample size. We apply SGD with 200 passes and consider the step sizes $\eta_t = 5/\sqrt{t}$. We plot the behavior of test errors $\|\nabla F_{\widetilde{S}}(\mathbf{w}_t)\|_2$, computational errors $\|\nabla F_S(\mathbf{w}_t)\|_2$ and statistical errors versus the number of passes. According to Fig. 1, along the optimization process the computational errors continue to decrease while the statistical errors continue to increase, which is well consistent with our theory. It can be also seen that the model trained by SGD also generalizes well by its resistance to overfitting, i.e., the test error $\|\nabla F_{\widetilde{S}}(\mathbf{w}_t)\|_2$ does not fluctuate even if we apply SGD with 200 passes.

Our second aim is to show how test errors would behave versus different step sizes. To this aim, we consider step sizes of the form $\eta_t = \eta_1 t^{-\theta}$ with $\theta \in \{0, 1/2, 1\}$. For each θ, we validate η_1 over the set $\{1, 2, \ldots, 2^8\}$ based on a validation set, and report the behavior of the associated test errors versus the number of passes in Fig. 2. In our experiments, SGD with $\theta = 1/2$ outperforms those with $\theta = 0$ and $\theta = 1$. This is consistent with Theorem 5, where $\theta = 1/2$ is shown to achieve the best learning rates (up to logarithmic factors) while achieving the minimal computation cost.

7 Conclusions

In this paper, we present learning rates of SGD for nonconvex learning problems from a joint perspective of optimization and statistics, which allows us to see how an optimal number of passes should be taken to trade-off the computational and statistical errors. We consider both general nonconvex and gradient-dominated objectives, and derive learning rates comparable to the corresponding ones in the convex case. For objective functions with a specific structure, we show that the learning rates can be dimensionality-independent. We control the statistical errors by giving high-probability bounds on the complexity of SGD iterates, which provides an alternative explanation on the generalization performance of SGD to train nonconvex models.

There remain some interesting directions to pursue. Firstly, it is interesting to extend the learning rates here to some variants of SGD for nonconvex learning, including SGD with variance reduction [52–54] and SGD with momentum [22, 55]. Secondly, the existing stability bounds for nonconvex learning are developed for the step-size sequence $\eta_t = O(1/t)$. It is very interesting to study the stability of SGD for the general polynomially-decaying step sizes.

Acknowledgments

We are grateful to the editor and referees for the constructive comments, which are very helpful to improve the paper.

References

[1] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *International Conference on Machine Learning*, 2003, pp. 928–936.

[2] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *International Conference on Machine Learning*, 2004, pp. 919–926.

[3] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[4] S. Huang and Z. Zhou, "Fast multi-instance multi-label learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.

[5] O. Bousquet and L. Bottou, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems*, 2008, pp. 161–168.

[6] W. Zhang, L. Zhang, Z. Jin, R. Jin, D. Cai, X. Li, R. Liang, and X. He, "Sparse learning with stochastic composite optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1223–1236, 2017.

[7] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 5947–5956.

[8] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 6240–6249.

[9] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *International Conference on Learning Representations*, 2017.

[10] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer Science & Business Media, 2008.

[11] S. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *International Conference on Machine Learning*, 2016, pp. 314–323.

[12] Z. Allen-Zhu and E. Hazan, "Variance reduction for faster non-convex optimization," in *International Conference on Machine Learning*, 2016, pp. 699–707.

[13] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[14] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Mathematical Programming*, vol. 155, no. 1-2, pp. 267–305, 2016.

[15] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International Conference on Machine Learning*, 2016, pp. 1225–1234.

[16] I. Kuzborskij and C. Lampert, "Data-dependent stability of stochastic gradient descent," in *International Conference on Machine Learning*, 2018, pp. 2820–2829.

[17] Y. Zhou, Y. Liang, and H. Zhang, "Generalization error bounds with probabilistic guarantee for SGD in nonconvex optimization," *arXiv preprint arXiv:1802.06903*, 2018.

[18] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for nonconvex losses," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2747–2774, 2018.

[19] L. Zhang, T. Yang, and R. Jin, "Empirical risk minimization for stochastic convex optimization: O$(1/n)$-and O$(1/n^2)$-type of risk bounds," in *Conference on Learning Theory*, 2017,

pp. 1954–1979.

[20] D. J. Foster, A. Sekhari, and K. Sridharan, "Uniform convergence of gradients for non-convex learning and optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 8759–8770.

[21] V. De la Peña and E. Giné, *Decoupling: From Dependence to Independence*. New York: Springer-Verlag, 1999.

[22] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2013, vol. 87.

[23] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.

[24] W. Mou, L. Wang, X. Zhai, and K. Zheng, "Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints," in *Conference on Learning Theory*, 2018, pp. 605–638.

[25] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Advances in Neural Information Processing Systems*, 2015, pp. 2737–2745.

[26] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *European Conference on Machine Learning*, 2016, pp. 795–811.

[27] Y. Zhou, Y. Liang, and H. Zhang, "Generalization error bounds with probabilistic guarantee for SGD in nonconvex optimization," *arXiv preprint arXiv:1802.06903*, 2018.

[28] L. Zhang and Z.-H. Zhou, "Stochastic approximation of smooth and strongly convex functions: Beyond the $o(1/t)$ convergence rate," in *Conference on Learning Theory*, 2019, pp. 3160–3179.

[29] Y. Ying and C. Campbell, "Rademacher chaos complexities for learning the kernel problem," *Neural computation*, vol. 22, no. 11, pp. 2858–2886, 2010.

[30] J. Lin, R. Camoriano, and L. Rosasco, "Generalization properties and implicit regularization for multiple passes SGM," in *International Conference on Machine Learning*, 2016, pp. 2340–2348.

[31] Y. Lei and K. Tang, "Stochastic composite mirror descent: Optimal bounds with high probabilities," in *Advance in Neural Information Processing Systems*, 2018, pp. 1524–1534.

[32] Z. Charles and D. Papailiopoulos, "Stability and generalization of learning algorithms that converge to global optima," in *International Conference on Machine Learning*, 2018, pp. 744–753.

[33] S. M. Kakade and A. Tewari, "On the generalization ability of online strongly convex programming algorithms," in *Advances in Neural Information Processing Systems*, 2009, pp. 801–808.

[34] L. Bottou, "On-line learning and stochastic approximations," in *On-line Learning in Neural Networks*, D. Saad, Ed. New York, NY, USA: Cambridge University Press, 1998, pp. 9–42.

[35] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," in *Advances in Neural Information Processing Systems*, 2018, pp. 689–699.

[36] D. Zhou, P. Xu, and Q. Gu, "Stochastic nested variance reduced gradient descent for nonconvex optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 3921–3932.

[37] L. M. Nguyen, M. van Dijk, D. T. Phan, P. H. Nguyen, T.-W. Weng, and J. R. Kalagnanam, "Finite-sum smooth optimization with SARAH," 2019.

[38] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh, "Spiderboost and momentum: Faster variance reduction algorithms," in *Advances in Neural Information Processing Systems*, 2019, pp. 2403–2413.

[39] Z. Allen-Zhu, "Natasha 2: Faster non-convex optimization than sgd," in *Advances in Neural Information Processing Systems*, 2018, pp. 2675–2686.

[40] D. Davis and D. Drusvyatskiy, "Uniform graphical convergence of subgradients in nonconvex optimization and learning," *arXiv preprint arXiv:1810.07590*, 2018.

[41] A. Elisseeff, T. Evgeniou, and M. Pontil, "Stability of randomized learning algorithms," *Journal of Machine Learning Research*, vol. 6, no. Jan, pp. 55–79, 2005.

[42] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, no. Mar, pp. 499–526, 2002.

[43] T. Liu, D. Tao, M. Song, and S. J. Maybank, "Algorithm-dependent generalization bounds for multi-task learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 227–241, 2016.

[44] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2635–2670, 2010.

[45] V. Feldman and J. Vondrak, "Generalization bounds for uniformly stable algorithms," in *Advances in Neural Information Processing Systems*, 2018, pp. 9747–9757.

[46] ——, "High probability generalization bounds for uniformly stable algorithms with nearly optimal rate," in *Conference on Learning Theory*, 2019, pp. 1270–1279.

[47] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy, "Sharper bounds for uniformly stable algorithms," in *Conference on Learning Theory*, 2020, pp. 610–626.

[48] L. Rosasco and S. Villa, "Learning with incremental iterative regularization," in *Advances in Neural Information Processing Systems*, 2015, pp. 1630–1638.

[49] A. Dieuleveut and F. Bach, "Nonparametric stochastic approximation with large step-sizes," *Annals of Statistics*, vol. 44, no. 4, pp. 1363–1399, 2016.

[50] J. Lin and L. Rosasco, "Optimal rates for multi-pass stochastic gradient methods," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3375–3421, 2017.

[51] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[52] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems*, 2013, pp. 315–323.

[53] L. Zhang, M. Mahdavi, and R. Jin, "Linear convergence with condition number independent access of full gradients," in *Advances in Neural Information Processing Systems*, 2013, pp. 980–988.

[54] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1-2, pp. 83–112, 2017.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

Supplemental Material for "Learning Rates for Stochastic Gradient Descent with Nonconvex Objectives"

Yunwen Lei and Ke Tang, *Senior Member, IEEE*

Abstract—In the Supplemental Material, we prove the theoretical results stated in the main text. The Supplemental Material consists of four parts: some useful concentration inequalities on martingales, the proofs on uniform convergence of gradients, the proofs on learning rates for general nonconvex objectives and the proofs of learning rates for gradient-dominated objectives.

I Concentration Inequalities

Our theoretical analysis is based on some concentration inequalities on martingales. In Lemma I.1, we present concentration inequalities for real-valued martingales. Part (a) is the Auzan-Hoeffding inequality for martingales with bounded increments, while Part (b) is a Bernstein-type inequality where the concentration behavior is better quantified in terms of the variance. Lemma I.2 is a Pinelis-Bernstein inequality for martingale difference sequences in a Hilbert space [1].

Lemma I.1. *Let $z_1,\ldots,z_n$ be a sequence of random variables such that z_k may depend on the previous random variables $z_1,\ldots,z_{k-1}$ for all $k=1,\ldots,n$. Consider a sequence of functionals $\xi_k(z_1,\ldots,z_k), k=1,\ldots,n$. Let $\sigma_n^2=\sum_{k=1}^n\mathbb{E}_{z_k}\big[(\xi_k-\mathbb{E}_{z_k}[\xi_k])^2\big]$ be the conditional variance.*

(a) Assume that $|\xi_k-\mathbb{E}_{z_k}[\xi_k]|\le b_k$ for each k. Let $\delta\in(0,1)$. With probability at least $1-\delta$ we have

$$\sum_{k=1}^n\xi_k-\sum_{k=1}^n\mathbb{E}_{z_k}[\xi_k]\le\Big(2\sum_{k=1}^n b_k^2\log\frac{1}{\delta}\Big)^{\frac{1}{2}}. \tag{I.1}$$

(b) Assume that $\xi_k-\mathbb{E}_{z_k}[\xi_k]\le b$ for each k. Let $\rho\in(0,1]$ and $\delta\in(0,1)$. With probability at least $1-\delta$ we have

$$\sum_{k=1}^n\xi_k-\sum_{k=1}^n\mathbb{E}_{z_k}[\xi_k]\le\frac{\rho\sigma_n^2}{b}+\frac{b\log\frac{1}{\delta}}{\rho}. \tag{I.2}$$

Lemma I.2. *Let $\{\xi_k\}_{k\in\mathbb{N}}$ be a martingale difference sequence in $\mathbb{R}^d$. Suppose that almost surely $\|\xi_k\|_2\le B$ and $\sum_{k=1}^t\mathbb{E}[\|\xi_k\|_2^2|\xi_1,\ldots,\xi_{k-1}]\le\sigma_t^2$. Then, for any $0<\delta<1$, the following inequality holds with probability at least $1-\delta$*

$$\max_{1\le j\le t}\big\|\sum_{k=1}^j\xi_k\big\|_2\le2\Big(\frac{B}{3}+\sigma_t\Big)\log(2/\delta).$$

II Proofs on Uniform Convergence

In this section, we present the proofs on uniform convergence for the gradients of empirical risk to that of population gradients. To prove Corollary 2 and Proposition 3, we first relate the uniform deviation to the associated expectation by McDimarid's inequality (Lemma II.1). To this aim, we need to show that the uniform deviation satisfies a bounded increment condition.

Lemma II.1. *Let $c_1,\ldots,c_n\in\mathbb{R}_+$. Let $Z_1,\ldots,Z_n$ be independent random variables taking values in a set $\mathcal{Z}$, and assume that $g:\mathcal{Z}^n\mapsto\mathbb{R}$ satisfies*

$$\sup_{z_1,\ldots,z_n,\bar{z}_i\in\mathcal{Z}}|g(z_1,\cdots,z_n)-g(\cdots,z_{i-1},\bar{z}_i,z_{i+1},\cdots)|\le c_i \tag{II.1}$$

for $i=1,\ldots,n$. Then, for any $0<\delta<1$, with probability at least $1-\delta$ we have

$$g(Z_1,\ldots,Z_n)\le\mathbb{E}\big[g(Z_1,\ldots,Z_n)\big]+\sqrt{\frac{\sum_{i=1}^n c_i^2\log(1/\delta)}{2}}.$$

Lemma II.2. *Let $\delta\in(0,1)$ and $S=\{z_1,\ldots,z_n\}$ be examples drawn independently from ρ. Suppose Assumption 1 holds. Then with probability at least $1-\delta$ we have*

$$\sup_{\mathbf{w}\in B_R}\big\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\big\|_2\le$$
$$\frac{2}{n}\mathbb{E}_S\mathbb{E}_\epsilon\sup_{\mathbf{w}\in B_R}\big\|\sum_{i=1}^n\epsilon_i\nabla f(\mathbf{w};z_i)\big\|_2+\sqrt{\frac{2\big(LR+\tilde{b}\big)^2\log(1/\delta)}{n}}.$$

Proof. By the L-Lipschitz continuity of ∇f, the following inequality holds for all $\mathbf{w}\in B_R$

$$\|\nabla f(\mathbf{w};z_i)\|_2\le\|\nabla f(0;z_i)\|_2+L\|\mathbf{w}\|_2\le LR+\tilde{b}. \tag{II.2}$$

Let $\widetilde{S}=\{\tilde{z}_1,\ldots,\tilde{z}_n\}$ be independent examples drawn independently from ρ and $\widetilde{S}_i=\{z_1,\ldots,z_{i-1},\tilde{z}_i,z_{i+1},\ldots,z_n\}$. Then, we have

$$\begin{aligned}&\Big|\sup_{\mathbf{w}\in B_R}\big\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\big\|_2-\sup_{\mathbf{w}\in B_R}\big\|\nabla F(\mathbf{w})-\nabla F_{\widetilde{S}_i}(\mathbf{w})\big\|_2\Big|\\&\le\sup_{\mathbf{w}\in B_R}\Big|\big\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\big\|_2-\big\|\nabla F(\mathbf{w})-\nabla F_{\widetilde{S}_i}(\mathbf{w})\big\|_2\Big|\\&\le\sup_{\mathbf{w}\in B_R}\big\|\nabla F_S(\mathbf{w})-\nabla F_{\widetilde{S}_i}(\mathbf{w})\big\|_2\\&=n^{-1}\sup_{\mathbf{w}\in B_R}\big\|\nabla f(\mathbf{w};z_i)-\nabla f(\mathbf{w};\tilde{z}_i)\big\|\le2(LR+\tilde{b})/n,\end{aligned}$$

where we have used (II.2) for all $\mathbf{w}\in B_R$. Therefore, (II.1) holds with $g(z_1,\ldots,z_n):=\sup_{\mathbf{w}\in B_R}\big[\|\nabla F(\mathbf{w})-$

$\nabla F_S(\mathbf{w})\|_2]$ and $c_i = 2(LR + \tilde{b})/n$. We can apply Lemma II.1 to derive the following inequality with probability $1 - \delta$

$$\sup_{\mathbf{w}\in B_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 \leq$$

$$\mathbb{E}_S\Big[\sup_{\mathbf{w}\in B_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2\Big] + \sqrt{\frac{2\big(LR + \tilde{b}\big)^2 \log(1/\delta)}{n}}.$$

By the standard symmetric trick, we get

$$\begin{aligned}
&\mathbb{E}_S \sup_{\mathbf{w}\in B_R} \big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \\
&= \mathbb{E}_S \sup_{\mathbf{w}\in B_R} \Big\|\nabla F_S(\mathbf{w}) - \mathbb{E}_{\widetilde{S}}\big[\nabla F_{\widetilde{S}}(\mathbf{w})\big]\Big\|_2 \\
&\leq n^{-1}\mathbb{E}_{S,\widetilde{S}} \sup_{\mathbf{w}\in B_R} \Big\|\sum_{i=1}^n \big(\nabla f(\mathbf{w}; z_i) - \nabla f(\mathbf{w}; \tilde{z}_i)\big)\Big\|_2 \\
&= n^{-1}\mathbb{E}_{S,\widetilde{S}}\mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R} \Big\|\sum_{i=1}^n \epsilon_i\big(\nabla f(\mathbf{w}; z_i) - \nabla f(\mathbf{w}; \tilde{z}_i)\big)\Big\|_2 \\
&\leq \frac{2}{n}\mathbb{E}_S\mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R} \Big\|\sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i)\Big\|_2.
\end{aligned}$$

We can combine the above two inequalities to derive the stated inequality with probability $1 - \delta$. The proof is complete. □

It remains to estimate the vector Rademacher complexities $\mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R} \big\|\sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i)\big\|_2$ [2, 3]. In the following two subsections, we will estimate it by Rademacher Chaos complexities and Gaussian complexities, respectively.

II.1 Proof of Theorem 1

We can prove Theorem 1 by showing that vector Rademacher complexities in Lemma II.2 can be controlled by Rademacher Chaos complexities.

Proof of Theorem 1. According to Jensen's inequality, we know

$$\begin{aligned}
&\Big(\mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R} \Big[\Big\|\sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i)\Big\|_2\Big]\Big)^2 \\
&\leq \mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in B_R} \Big\|\sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i)\Big\|_2^2\Big] \\
&= \mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in B_R} \Big\langle \sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i), \sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i)\Big\rangle\Big] \\
&\leq \sup_{\mathbf{w}\in B_R} \sum_{i=1}^n \langle \nabla f(\mathbf{w}; z_i), \nabla f(\mathbf{w}; z_i)\rangle \\
&+ 2\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in B_R} \sum_{1\leq i<j\leq n} \epsilon_i\epsilon_j\big\langle \nabla f(\mathbf{w}; z_i), \nabla f(\mathbf{w}; z_j)\big\rangle\Big] \\
&\leq n(LR + \tilde{b})^2 + 2n\mathcal{U}_S(\mathcal{F}_R), \qquad \text{(II.3)}
\end{aligned}$$

where we have used (II.2) and the definition of Rademacher Chaos complexities. It then follows that

$$\mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R} \Big[\Big\|\sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i)\Big\|_2\Big] \leq \sqrt{n}(LR + \tilde{b}) + \sqrt{2n\mathcal{U}_S(\mathcal{F}_R)}.$$

We can plug the above bound into Lemma II.2 to derive the stated bound with high probabilities. □

II.2 Proof of Corollary 2

To prove Corollary 2, it remains to estimate the involved Rademacher Chaos complexity, which can be further controlled by the following entropy integral in terms of covering numbers.

Definition 1. Let $(\mathcal{G}, d)$ be a metric space and set $\mathcal{F} \subseteq \mathcal{G}$. For any $\epsilon > 0$, a set $\mathcal{F}^\triangle \subset \mathcal{F}$ is called an ϵ-cover of $\mathcal{F}$ if for every $f \in \mathcal{F}$ we can find an element $g \in \mathcal{F}^\triangle$ satisfying $d(f, g) \leq \epsilon$. The covering number $\mathcal{N}(\epsilon, \mathcal{F}, d)$ is the cardinality of the minimal ϵ-cover of $\mathcal{F}$:

$$\mathcal{N}(\epsilon, \mathcal{F}, d) := \min\Big\{|\mathcal{F}^\triangle| : \mathcal{F}^\triangle \text{ is an } \epsilon\text{-cover of } \mathcal{F}\Big\}.$$

Lemma II.3 ([4]). *Let $\mathcal{F} : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$ be a function class with $\sup_{f\in\mathcal{F}} d_S(f, 0) \leq D$ and $S = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$, where d_S is a pseudometric on $\mathcal{F}$ defined as follows*

$$d_S(f, g) := \Big(\frac{1}{n^2}\sum_{1\leq i<j\leq n} |f(x_i, x_j) - g(x_i, x_j)|^2\Big)^{1/2}. \qquad \text{(II.4)}$$

Then

$$\mathcal{U}_S(\mathcal{F}) \leq 24e\int_0^D \log\big(\mathcal{N}(r, \mathcal{F}, d_S) + 1\big)\mathrm{d}r.$$

Proof of Corollary 2. We define a metric d_S over $\mathcal{F}_R$ by

$$\begin{aligned}
d_S(\mathbf{w}, \tilde{\mathbf{w}}) = \Big(\frac{1}{n^2}\sum_{1\leq i<j\leq n} \big|&\langle \nabla f(\mathbf{w}; z_i), \nabla f(\mathbf{w}; z_j)\rangle \\
&- \langle \nabla f(\tilde{\mathbf{w}}; z_i), \nabla f(\tilde{\mathbf{w}}; z_j)\rangle\big|^2\Big)^{1/2}.
\end{aligned}$$

For any $\mathbf{w}$ and $\tilde{\mathbf{w}}$ in B_R, there holds

$$\begin{aligned}
&n^2 d_S^2(\mathbf{w}, \tilde{\mathbf{w}}) \\
&= \sum_{1\leq i<j\leq n} \big|\langle \nabla f(\mathbf{w}; z_i), \nabla f(\mathbf{w}; z_j)\rangle - \langle \nabla f(\tilde{\mathbf{w}}; z_i), \nabla f(\tilde{\mathbf{w}}; z_j)\rangle\big|^2 \\
&\leq 2\sum_{1\leq i<j\leq n} \big\langle \nabla f(\mathbf{w}; z_i) - \nabla f(\tilde{\mathbf{w}}; z_i), \nabla f(\mathbf{w}; z_j)\big\rangle^2 \\
&+ 2\sum_{1\leq i<j\leq n} \big\langle \nabla f(\tilde{\mathbf{w}}; z_i), \nabla f(\mathbf{w}; z_j) - \nabla f(\tilde{\mathbf{w}}; z_j)\big\rangle^2 \\
&\leq 2\sum_{1\leq i<j\leq n} \big\|\nabla f(\mathbf{w}; z_i) - \nabla f(\tilde{\mathbf{w}}; z_i)\big\|_2^2 \big\|\nabla f(\mathbf{w}; z_j)\big\|_2^2 \\
&+ 2\sum_{1\leq i<j\leq n} \|\nabla f(\tilde{\mathbf{w}}; z_i)\|_2^2 \big\|\nabla f(\mathbf{w}; z_j) - \nabla f(\tilde{\mathbf{w}}; z_j)\big\|_2^2 \\
&\leq 2L^2\sum_{1\leq i<j\leq n} \Big[\big\|\nabla f(\mathbf{w}; z_j)\big\|_2^2 + \|\nabla f(\tilde{\mathbf{w}}; z_i)\|_2^2\Big]\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 \\
&\leq 2L^2(LR + \tilde{b})^2 n(n-1)\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2, \qquad \text{(II.5)}
\end{aligned}$$

where we have used (III.3) and the decomposition

$$\begin{aligned}
\langle \nabla f(\mathbf{w}; z_i), \nabla f(\mathbf{w}; z_j)\rangle - \langle \nabla f(\tilde{\mathbf{w}}; z_i), \nabla f(\tilde{\mathbf{w}}; z_j)\rangle = \\
\langle \nabla f(\mathbf{w}; z_i) - \nabla f(\tilde{\mathbf{w}}; z_i), \nabla f(\mathbf{w}; z_j)\rangle \\
+ \langle \nabla f(\tilde{\mathbf{w}}; z_i), \nabla f(\mathbf{w}; z_j) - \nabla f(\tilde{\mathbf{w}}; z_j)\rangle
\end{aligned}$$

in the first inequality, the L-smoothness of f in the third inequality and (II.2) in the last inequality. It then follows that

$$\begin{aligned}
\log\mathcal{N}(r, \mathcal{F}_R, d_S) &\leq \log\mathcal{N}\Big(r/\big(\sqrt{2}L(LR + \tilde{b})\big), B_R, d_2\Big) \\
&\leq d\log\Big(3\sqrt{2}LR(LR + \tilde{b})r^{-1}\Big),
\end{aligned}$$

where we have used the classical result $\log \mathcal{N}(r, B_R, d_2) \leq d \log(3R/r)$ [5] and d_2 is the metric over B_R defined by $d_2(\mathbf{w}, \tilde{\mathbf{w}}) = \|\mathbf{w} - \tilde{\mathbf{w}}\|_2$. Furthermore, (II.5) also implies $d_S(\mathbf{w}, 0) \leq \sqrt{2}LR(LR + \tilde{b})$ for $\mathbf{w} \in B_R$. We can now apply Lemma II.3 to show that

$$
\begin{aligned}
\mathcal{U}_S(\mathcal{F}_R) &\leq 24e \int_0^{(LR+\tilde{b})\sqrt{2}LR} \log\big(1 + \mathcal{N}(r, \mathcal{F}_R, d_S)\big) dr \\
&\leq 24e \int_0^{(LR+\tilde{b})\sqrt{2}LR} \Big(\log 2 + d \log\Big(3\sqrt{2}LR(LR+\tilde{b})r^{-1}\Big)\Big) \mathrm{d}r \\
&\leq 24\sqrt{2}e(LR + \tilde{b})LR\big(\log 2 + d\log(3e)\big),
\end{aligned}
$$

where we have used

$$
\begin{aligned}
&\int_0^{(LR+\tilde{b})\sqrt{2}LR} \log\Big(3\sqrt{2}LR(LR + \tilde{b})r^{-1}\Big)\mathrm{d}r \\
&= \sqrt{2}LR(LR + \tilde{b}) \int_0^1 \log(3/r)\mathrm{d}r = \sqrt{2}LR(LR + \tilde{b}) \log(3e).
\end{aligned}
$$

The stated bound then follows by plugging the above bound on Rademacher Chaos complexities into Theorem 1. The proof is complete. □

II.3 Proof of Proposition 3

Before proving Proposition 3, we first apply a comparison (Slepian's lemma, Lemma II.4 below) on the suprema of Gaussian processes to estimate the vector Rademacher complexity in Lemma II.2.

Lemma II.4. *Let $\{\mathfrak{X}_\theta : \theta \in \Theta\}$ and $\{\mathfrak{Y}_\theta : \theta \in \Theta\}$ be two mean-zero separable Gaussian processes indexed by the same set Θ and suppose that*

$$
\mathbb{E}[(\mathfrak{X}_\theta - \mathfrak{X}_{\bar{\theta}})^2] \leq \mathbb{E}[(\mathfrak{Y}_\theta - \mathfrak{Y}_{\bar{\theta}})^2], \quad \forall \theta, \bar{\theta} \in \Theta. \tag{II.6}
$$

Then $\mathbb{E}[\sup_{\theta\in\Theta} \mathfrak{X}_\theta] \leq \mathbb{E}[\sup_{\theta\in\Theta} \mathfrak{Y}_\theta]$.

Lemma II.5. *Suppose $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}$ takes the form $f(\mathbf{w}; z) = \phi(\langle \mathbf{w}, x\rangle, y)$ with ϕ being L_ϕ-smooth with respect to the first argument. Then*

$$
\mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R} \Big\|\sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i)\Big\|_2 \leq \big(2L_\phi R\kappa + \tilde{b}\big)\sqrt{2}\Big(\sum_{i=1}^n \|x_i\|_2^2\Big)^{\frac{1}{2}}.
$$

Proof. By the structure of f, we know

$$
\begin{aligned}
&\mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R} \Big\|\sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i)\Big\|_2 \\
&= \mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R} \Big\|\sum_{i=1}^n \epsilon_i \phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big) x_i\Big\|_2 \\
&= \mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R, \mathbf{v}\in B_1} \Big\langle \sum_{i=1}^n \epsilon_i \phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big) x_i, \mathbf{v}\Big\rangle \\
&\leq \mathbb{E}_g \sup_{\mathbf{w}\in B_R, \mathbf{v}\in B_1} \sum_{i=1}^n g_i \phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big)\langle x_i, \mathbf{v}\rangle,
\end{aligned} \tag{II.7}
$$

where ϕ' denotes the derivative of ϕ with respect to the first argument, $g_1, \ldots, g_n$ are i.i.d. $N(0, 1)$ random variables and we have used the following inequality on Rademacher and Gaussian complexities

$$
\mathbb{E}_\epsilon \sup_f \sum_{i=1}^n \epsilon_i f(z_i) \leq \mathbb{E}_g \sup_f \sum_{i=1}^n g_i f(z_i).
$$

Introduce two mean-zero separable Gaussian processes indexed by $B_R \times B_1$

$$
\mathfrak{X}_{\mathbf{w},\mathbf{v}} = \sum_{i=1}^n g_i \phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big)\langle x_i, \mathbf{v}\rangle
$$

$$
\mathfrak{Y}_{\mathbf{w},\mathbf{v}} = \sqrt{2}\kappa \sum_{i=1}^n g_i \phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big) + \sqrt{2}\big(\tilde{b} + L_\phi R\kappa\big) \sum_{i=1}^n \tilde{g}_i \langle x_i, \mathbf{v}\rangle,
$$

where $\tilde{g}_1, \ldots, \tilde{g}_n$ are independent $N(0, 1)$ random variables. For any $\mathbf{w}, \tilde{\mathbf{w}} \in B_R$ and $\mathbf{v}, \tilde{\mathbf{v}} \in B_1$, the independence among g_i and $\mathbb{E}g_i^2 = 1, \forall i = 1, \ldots, n$ imply that

$$
\begin{aligned}
&\mathbb{E}_g\big[\big(\mathfrak{X}_{\mathbf{w},\mathbf{v}} - \mathfrak{X}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}}\big)^2\big] \\
&= \sum_{i=1}^n \Big(\phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big)\langle x_i, \mathbf{v}\rangle - \phi'\big(\langle \tilde{\mathbf{w}}, x_i\rangle, y_i\big)\langle x_i, \tilde{\mathbf{v}}\rangle\Big)^2 \\
&\leq 2\sum_{i=1}^n \Big(\phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big) - \phi'\big(\langle \tilde{\mathbf{w}}, x_i\rangle, y_i\big)\Big)^2 \big(\langle x_i, \mathbf{v}\rangle\big)^2 \\
&\quad + 2\sum_{i=1}^n \Big(\phi'\big(\langle \tilde{\mathbf{w}}, x_i\rangle, y_i\big)\Big)^2 \big(\langle x_i, \mathbf{v}\rangle - \langle x_i, \tilde{\mathbf{v}}\rangle\big)^2 \\
&\leq 2\kappa^2 \sum_{i=1}^n \Big(\phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big) - \phi'\big(\langle \tilde{\mathbf{w}}, x_i\rangle, y_i\big)\Big)^2 \\
&\quad + 2\big(\tilde{b} + L_\phi R\kappa\big)^2 \sum_{i=1}^n \big(\langle x_i, \mathbf{v}\rangle - \langle x_i, \tilde{\mathbf{v}}\rangle\big)^2 \\
&= \mathbb{E}_g\big[\big(\mathfrak{Y}_{\mathbf{w},\mathbf{v}} - \mathfrak{Y}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}}\big)^2\big],
\end{aligned}
$$

where we have used the elementary inequality (III.3), the decomposition

$$
\begin{aligned}
&\phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big)\langle x_i, \mathbf{v}\rangle - \phi'\big(\langle \tilde{\mathbf{w}}, x_i\rangle, y_i\big)\langle x_i, \tilde{\mathbf{v}}\rangle = \\
&\qquad \Big(\phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big) - \phi'\big(\langle \tilde{\mathbf{w}}, x_i\rangle, y_i\big)\Big)\langle x_i, \mathbf{v}\rangle \\
&\qquad\qquad + \phi'\big(\langle \tilde{\mathbf{w}}, x_i\rangle, y_i\big)\big(\langle x_i, \mathbf{v}\rangle - \langle x_i, \tilde{\mathbf{v}}\rangle\big)
\end{aligned}
$$

and the following inequality due to the L_ϕ-smoothness of ϕ

$$
\big|\phi'\big(\langle \tilde{\mathbf{w}}, x_i\rangle, y_i\big)\big| \leq \tilde{b} + L_\phi|\langle \tilde{\mathbf{w}}, x_i\rangle - 0| \leq \tilde{b} + L_\phi R\kappa.
$$

Therefore, we can apply Lemma II.4 to show

$$
\begin{aligned}
&\mathbb{E}_g \sup_{\mathbf{w}\in B_R, \mathbf{v}\in B_1} \mathfrak{X}_{\mathbf{w},\mathbf{v}} \leq \mathbb{E}_g \sup_{\mathbf{w}\in B_R, \mathbf{v}\in B_1} \mathfrak{Y}_{\mathbf{w},\mathbf{v}} \\
&\leq \sqrt{2}\kappa \mathbb{E}_g \sup_{\mathbf{w}\in B_R} \sum_{i=1}^n g_i \phi'\big(\langle \mathbf{w}, x_i\rangle, y_i\big) \\
&\quad + \sqrt{2}\big(\tilde{b} + L_\phi R\kappa\big)\mathbb{E}_g \sup_{\mathbf{v}\in B_1} \sum_{i=1}^n g_i \langle x_i, \mathbf{v}\rangle \\
&\leq \sqrt{2}L_\phi \kappa \mathbb{E}_g \sup_{\mathbf{w}\in B_R} \sum_{i=1}^n g_i \langle \mathbf{w}, x_i\rangle \\
&\quad + \sqrt{2}\big(\tilde{b} + L_\phi R\kappa\big)\mathbb{E}_g \sup_{\mathbf{v}\in B_1} \sum_{i=1}^n g_i \langle x_i, \mathbf{v}\rangle,
\end{aligned}
$$

where we have used the L_ϕ-Lipschitz continuity of ϕ' and the contraction lemma of Gaussian complexities in the last step. Furthermore, it follows from the Jensen's inequality that

$$\begin{aligned}&\mathbb{E}_g \sup_{\mathbf{w}\in B_R} \sum_{i=1}^n g_i \langle \mathbf{w}, x_i\rangle = \mathbb{E}_g \sup_{\mathbf{w}\in B_R} \Big\langle \mathbf{w}, \sum_{i=1}^n g_i x_i \Big\rangle \\ &\leq R\mathbb{E}_g \Big\| \sum_{i=1}^n g_i x_i \Big\|_2 \leq R\sqrt{\mathbb{E}_g\Big[\Big\langle \sum_{i=1}^n g_i x_i, \sum_{i=1}^n g_i x_i\Big\rangle\Big]} \\ &= R\Big(\sum_{i=1}^n \|x_i\|_2^2\Big)^{\frac{1}{2}}.\end{aligned}$$

In a similar way, we can show $\mathbb{E}_g \sup_{\mathbf{v}\in B_1} \sum_{i=1}^n g_i\langle x_i, \mathbf{v}\rangle \leq \Big(\sum_{i=1}^n \|x_i\|_2^2\Big)^{\frac{1}{2}}$. Therefore,

$$\mathbb{E}_g \sup_{\mathbf{w}\in B_R, \mathbf{v}\in B_1} \mathfrak{X}_{\mathbf{w},\mathbf{v}} \leq \big(2L_\phi R\kappa + \tilde{b}\big)\sqrt{2}\Big(\sum_{i=1}^n \|x_i\|_2^2\Big)^{\frac{1}{2}}.$$

Plugging the above inequality into (II.7) then gives the stated bound. The proof is complete. □

Proof of Proposition 3. By the L_ϕ-smoothness of ϕ, we know that f is $(L_\phi\kappa^2)$-smooth

$$\begin{aligned}\big\|\nabla f(\mathbf{w};z) - \nabla f(\tilde{\mathbf{w}};z)\big\|_2 &= \big|\phi'(\langle \mathbf{w}, x\rangle, y) - \phi'(\langle\tilde{\mathbf{w}}, x\rangle, y)\big|\|x\|_2 \\ &\leq L_\phi |\langle \mathbf{w} - \tilde{\mathbf{w}}, x\rangle|\|x\|_2 \leq L_\phi\kappa^2\|\mathbf{w} - \tilde{\mathbf{w}}\|_2.\end{aligned}$$

Therefore, Lemma II.2 holds with $L = L_\phi\kappa^2$. By Lemma II.5 and the definition of κ, we know

$$\mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R} \Big\| \sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w};z_i)\Big\|_2 \leq \big(2L_\phi R\kappa + \tilde{b}\big)\sqrt{2n}\kappa,$$

which, together with Lemma II.2, gives the stated bound. The proof is complete. □

III Proofs on Learning Rates for General Nonconvex Objectives

In this section, we prove learning rates for general nonconvex objectives. We first prove Lemma 4 related to computational errors. Then, we move on to the proof of Lemma 9 on the norm of SGD iterates, based on which we finally prove the learning rates in Theorem 5.

III.1 Proof of Lemma 4

In this section, we present the proof of Lemma 4. Our idea is to use the L-smoothness of f to derive

$$\sum_{k=1}^t \eta_k \|\nabla F_S(\mathbf{w}_k)\|_2^2 = O(1)\big(\sum_{k=1}^t \xi_k + \sum_{k=1}^t \xi_k' + \sum_{k=1}^t \eta_k^2\big),$$

where $\{\xi_k\}_k$ and $\{\xi_k'\}$ are two martingale difference sequences. We can apply Lemma I.1 to establish high probability bounds for $\sum_{k=1}^t \xi_k$ and $\sum_{k=1}^t \xi_k'$, which then yield the stated bound.

Proof of Lemma 4. According to Assumption 1, we know F_S is also L-smooth, from which and 2′ we derive

$$\begin{aligned}F_S(\mathbf{w}_{t+1}) &\leq F_S(\mathbf{w}_t) + \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ &= F_S(\mathbf{w}_t) + \eta_t\langle \nabla F_S(\mathbf{w}_t) - \nabla f(\mathbf{w}_t; z_{j_t}), \nabla F_S(\mathbf{w}_t)\rangle \\ &\quad - \eta_t\|\nabla F_S(\mathbf{w}_t)\|_2^2 + 2^{-1}L\eta_t^2\|\nabla f(\mathbf{w}_t;z_{j_t})\|_2^2 \\ &= F_S(\mathbf{w}_t) + \xi_t - \eta_t\|\nabla F_S(\mathbf{w}_t)\|_2^2 + 2^{-1}L\eta_t^2\|\nabla f(\mathbf{w}_t;z_{j_t})\|_2^2,\end{aligned} \tag{III.1}$$

where we have introduced a martingale difference sequence

$$\xi_t = \eta_t\langle \nabla F_S(\mathbf{w}_t) - \nabla f(\mathbf{w}_t;z_{j_t}), \nabla F_S(\mathbf{w}_t)\rangle, \quad t\in\mathbb{N}. \tag{III.2}$$

By the elementary inequality

$$(a_1 + a_2)^2 \leq 2(a_1^2 + a_2^2), \quad \forall a_1, a_2 \in \mathbb{R}, \tag{III.3}$$

we know

$$\begin{aligned}\|\nabla f(\mathbf{w}_t;z_{j_t})\|_2^2 &= \big\|\nabla f(\mathbf{w}_t;z_{j_t}) - \nabla F_S(\mathbf{w}_t) + \nabla F_S(\mathbf{w}_t)\big\|_2^2 \\ &\leq 2\|\nabla f(\mathbf{w}_t;z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|_2^2 + 2\|\nabla F_S(\mathbf{w}_t)\|_2^2 \\ &= 2\xi_t' + 2\mathbb{E}_{j_t}\big[\|\nabla f(\mathbf{w}_t;z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|_2^2\big] + 2\|\nabla F_S(\mathbf{w}_t)\|_2^2,\end{aligned} \tag{III.4}$$

where we have introduced another martingale difference sequence

$$\begin{aligned}\xi_t' &= \|\nabla f(\mathbf{w}_t;z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|_2^2 \\ &\quad - \mathbb{E}_{j_t}\big[\|\nabla f(\mathbf{w}_t;z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|_2^2\big],\ t\in\mathbb{N}.\end{aligned} \tag{III.5}$$

Combining (III.1), (III.4) together and using Assumption 3, we know that $F_S(\mathbf{w}_{t+1})$ is upper bounded by

$$\begin{aligned}&F_S(\mathbf{w}_t) + \xi_t - \eta_t\|\nabla F_S(\mathbf{w}_t)\|_2^2 + L\eta_t^2\big(\sigma^2 + \xi_t' + \|\nabla F_S(\mathbf{w}_t)\|_2^2\big) \\ &\leq F_S(\mathbf{w}_t) + \xi_t - 2^{-1}\eta_t\|\nabla F_S(\mathbf{w}_t)\|_2^2 + L\eta_t^2\sigma^2 + L\eta_t^2\xi_t',\end{aligned} \tag{III.6}$$

where we have used the inequality $L\eta_t^2 - \eta_t \leq -\eta_t/2$ due to the assumption $\eta_t \leq 1/(2L)$.

Taking a summation of the inequality (III.6) gives

$$\begin{aligned}F_S(\mathbf{w}_{t+1}) &= F_S(\mathbf{w}_1) + \sum_{k=1}^t \big(F_S(\mathbf{w}_{k+1}) - F_S(\mathbf{w}_k)\big) \\ &\leq F_S(\mathbf{w}_1) + \sum_{k=1}^t \xi_k - 2^{-1}\sum_{k=1}^t \eta_k\|\nabla F_S(\mathbf{w}_k)\|_2^2 \\ &\quad + L\sigma^2\sum_{k=1}^t \eta_k^2 + L\sum_{k=1}^t \eta_k^2\xi_k'.\end{aligned} \tag{III.7}$$

It is clear that $\mathbb{E}_{j_k}[\xi_k] = 0$ and therefore $\{\xi_k\}$ is a martingale difference sequence. According to Assumption 2, the magnitude of ξ_t can be controlled by (note Assumption 2 implies $\sqrt{\eta_t}\|\nabla F_S(\mathbf{w}_t)\|_2 \leq G$)

$$|\xi_k| \leq \eta_k\big(\|\nabla F_S(\mathbf{w}_k)\|_2 + \|\nabla f(\mathbf{w}_k;z_{j_k})\|_2\big)\|\nabla F_S(\mathbf{w}_k)\|_2 \leq 2G^2. \tag{III.8}$$

According to Assumption 3 and the inequality $\mathbb{E}_{j_k}\big[\big(\xi_k - \mathbb{E}_{j_k}[\xi_k]\big)^2\big] \leq \mathbb{E}_{j_k}[\xi_k^2]$, we have the following inequality on conditional variances

$$\begin{aligned}&\sum_{k=1}^t \mathbb{E}_{j_k}\big[\big(\xi_k - \mathbb{E}_{j_k}[\xi_k]\big)^2\big] \\ &\leq \sum_{k=1}^t \eta_k^2\mathbb{E}_{j_k}\big[\|\nabla F_S(\mathbf{w}_k) - \nabla f(\mathbf{w}_k;z_{j_k})\|_2^2\big]\|\nabla F_S(\mathbf{w}_k)\|_2^2 \\ &\leq \sigma^2\sum_{k=1}^t \eta_k^2\|\nabla F_S(\mathbf{w}_k)\|_2^2 \leq \frac{\sigma^2}{2L}\sum_{k=1}^t \eta_k\|\nabla F_S(\mathbf{w}_k)\|_2^2.\end{aligned} \tag{III.9}$$

Plugging (III.8), (III.9) back into Part (b) in Lemma I.1 with $\rho = \min\{1, LG^2/\sigma^2\}$, we derive the following inequality with probability $1-\delta/2$

$$\sum_{k=1}^t \xi_k \le \frac{\rho\sigma^2\sum_{k=1}^t \eta_k\|\nabla F_S(\mathbf{w}_k)\|_2^2}{4LG^2} + \frac{2G^2\log(2/\delta)}{\rho}$$
$$\le 4^{-1}\sum_{k=1}^t \eta_k \|\nabla F_S(\mathbf{w}_k)\|_2^2 + 2\log(2/\delta)\max\{G^2,\sigma^2/L\}.$$

According to Assumption 2, we know

$$\xi_k' \le 2\big(\|\nabla f(\mathbf{w}_k; z_{j_k})\|_2^2 + \|\nabla F_S(\mathbf{w}_k)\|_2^2\big) \le 4\eta_k^{-1}G^2.$$

In a similar way, one can also show that $\xi_k' \ge -4\eta_k^{-1}G^2$. Therefore, one can apply Part (a) of Lemma I.1 to derive the following inequality with probability at least $1-\delta/2$

$$\sum_{k=1}^t \eta_k^2\xi_k' \le 4G^2\Big(2\sum_{k=1}^t\eta_k^2\log(2/\delta)\Big)^{\frac12} \le 8G^2\log(2/\delta) + G^2\sum_{k=1}^t \eta_k^2,$$

where we have used the Schwartz's inequality in the last step. Plugging the above inequalities back into (III.7) shows that with probability $1-\delta$

$$F_S(\mathbf{w}_{t+1}) \le F_S(0)+2\log(2/\delta)\max\{G^2,\sigma^2/L\}+LG^2\sum_{k=1}^t\eta_k^2$$
$$- 4^{-1}\sum_{k=1}^t\eta_k\|\nabla F_S(\mathbf{w}_k)\|_2^2 + L\sigma^2\sum_{k=1}^t\eta_k^2 + 8LG^2\log(2/\delta),$$

which can be written as (5). The proof is complete. □

III.2 Proof of Lemma 9

We are now ready to prove Lemma 9. The idea is to relate $\mathbf{w}_t$ to a summation of martingale difference sequences plus a summation of weighted empirical gradients, which can be respectively controlled by concentration inequalities and Lemma 4.

Proof of Lemma 9. According to (1), we know

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\big(\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\big) - \eta_t\nabla F_S(\mathbf{w}_t).$$

Taking a summation of the above inequality and using $\mathbf{w}_1 = 0$ then give

$$\mathbf{w}_{t+1} = \sum_{k=1}^t \xi_k - \sum_{k=1}^t\eta_k\nabla F_S(\mathbf{w}_k),$$

and therefore

$$\|\mathbf{w}_{t+1}\|_2 \le \Big\|\sum_{k=1}^t \xi_k\Big\|_2 + \Big\|\sum_{k=1}^t\eta_k\nabla F_S(\mathbf{w}_k)\Big\|_2, \qquad \text{(III.10)}$$

where we introduce the martingale difference sequence

$$\xi_k = \eta_k\big(\nabla F_S(\mathbf{w}_k) - \nabla f(\mathbf{w}_k; z_{j_k})\big),\quad k\in\mathbb{N}.$$

By Assumptions 2, 3 and $\eta_k\le 1/(2L)$, we know

$$\|\xi_k\|_2 = \eta_k\big\|\nabla f(\mathbf{w}_k;z_{j_k}) - \nabla F_S(\mathbf{w}_k)\big\|_2 \le 2G\sqrt{\eta_k} \le G\sqrt{2/L}$$

and

$$\sum_{k=1}^T \mathbb{E}_{j_k}[\|\xi_k\|_2^2] \le \sum_{k=1}^T \eta_k^2\sigma^2.$$

Then, we can apply Lemma I.2 to derive the following inequality with probability $1-\delta/2$

$$\max_{1\le t\le T}\Big\|\sum_{k=1}^t\xi_k\Big\|_2 \le 2\Big(\frac{G\sqrt{2/L}}{3} + \sigma\Big(\sum_{k=1}^T\eta_k^2\Big)^{\frac12}\Big)\log(4/\delta). \qquad \text{(III.11)}$$

Furthermore, according to Lemma 4 and the Schwartz's inequality, we derive the following inequality with probability $1-\delta/2$ for all $t=1,\ldots,T$

$$\Big\|\sum_{k=1}^t\eta_k\nabla F_S(\mathbf{w}_k)\Big\|_2^2 \le \Big(\sum_{k=1}^t \eta_k\|\nabla F_S(\mathbf{w}_k)\|_2\Big)^2 \le \Big(\sum_{k=1}^t\eta_k\Big)$$
$$\times \Big(\sum_{k=1}^t\eta_k\|\nabla F_S(\mathbf{w}_k)\|_2^2\Big) \le \Big(\sum_{k=1}^t\eta_k\Big)(C_T + C_1\log(4/\delta)).$$

It then follows the following inequality with probability $1-\delta/2$ uniformly for all $t=1,\ldots,T$

$$\Big\|\sum_{k=1}^t\eta_k\nabla F_S(\mathbf{w}_k)\Big\|_2 \le \Big((C_T + C_1\log(4/\delta))\sum_{k=1}^t\eta_k\Big)^{\frac12}.$$

Plugging the above inequality and (III.11) into (III.10) then gives the following inequality with probability $1-\delta$

$$\|\mathbf{w}_{t+1}\|_2 \le 2\Big(\frac{G\sqrt{2/L}}{3} + \sigma\Big(\sum_{k=1}^T\eta_k^2\Big)^{\frac12}\Big)\log(4/\delta)$$
$$+ \Big((C_T+C_1\log(4/\delta))\sum_{k=1}^t\eta_k\Big)^{\frac12},$$

which can be written as the stated form with

$$C_2 = \max\Big\{2G\sqrt{2/(9L)}, \sqrt{C_1+4b}, 2\sigma, \sqrt{4L(\sigma^2+G^2)}\Big\}.$$

The proof is complete. □

III.3 Proof of Theorem 5

Theorem 5 is proved by decomposing the gradients of population risks into computational and statistical errors. We can apply Lemma 4 to control computational errors, and uniform convergence of gradients together with the norm estimate of iterates in Lemma 9 to control stastistical errors.

Lemma III.1. *We have the following elementary inequalities.*
(a) If $\theta\in(0,1)$, then $\sum_{k=1}^t k^{-\theta} \le t^{1-\theta}/(1-\theta)$;
(b) If $\theta = 1$, then $\sum_{k=1}^t k^{-\theta} \le \log(et)$;
(c) If $\theta > 1$, then $\sum_{k=1}^t k^{-\theta} \le \frac{\theta}{\theta-1}$.

Proof of Theorem 5. By the elementary inequality (III.3) and Lemma 4, we derive the following inequality with probability $1-\delta/3$

$$\sum_{t=1}^T\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2$$
$$= \sum_{t=1}^T\eta_t\big\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t) + \nabla F_S(\mathbf{w}_t)\big\|_2^2$$
$$\le 2\sum_{t=1}^T\eta_t\big\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\big\|_2^2 + 2\sum_{t=1}^T\eta_t\big\|\nabla F_S(\mathbf{w}_t)\big\|_2^2$$
$$\le 2\sum_{t=1}^T\eta_t\max_{t=1,\ldots,T}\big\|\nabla F(\mathbf{w}_t)-\nabla F_S(\mathbf{w}_t)\big\|_2^2 + O\Big(\sum_{t=1}^T\eta_t^2 + \log\frac1\delta\Big),$$

from which we derive

$$\frac{\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2}{\sum_{t=1}^T \eta_t} \leq 2 \max_{t=1,\ldots,T} \big\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\big\|_2^2 \\ + O(1)\Big(\sum_{t=1}^T \eta_t\Big)^{-1}\big(\sum_{t=1}^T \eta_t^2 + \log(1/\delta)\big). \quad \text{(III.12)}$$

According to Lemma 9 and Lemma III.1, with probability $1-\delta/3$ we have the following inequality uniformly for all $t = 1, \ldots, T$

$$\|\mathbf{w}_t\|_2 \leq R_T := \begin{cases} O(\log(1/\delta))T^{1-\frac{3\theta}{2}}, & \text{if } \theta < 1/2 \\ O(\log(1/\delta))T^{\frac{1}{4}}\log^{\frac{1}{2}} T, & \text{if } \theta = 1/2 \\ O(\log(1/\delta))T^{\frac{1-\theta}{2}}, & \text{if } \theta > 1/2. \end{cases} \quad \text{(III.13)}$$

According to Corollary 2, the following inequality holds with probability $1-\delta/3$ (we assume $R_T \geq 1$)

$$\sup_{\mathbf{w}\in B_{R_T}} \|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|_2 \leq C_3 R_T n^{-\frac{1}{2}}, \quad \text{(III.14)}$$

where

$$C_3 = (L+\tilde{b})\big(2+2\sqrt{48e\sqrt{2}\big(\log 2 + d\log(3e)\big)} + \sqrt{2\log(3/\delta)}\big). \quad \text{(III.15)}$$

Combining (III.12), (III.13) and (III.14) together, with probability $1-\delta$ we derive the following inequality

$$\Big(\sum_{t=1}^T \eta_t\Big)^{-1}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 \leq 2\sup_{\mathbf{w}\in B_{R_T}} \|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|_2^2 \\ + O(1)\Big(\sum_{t=1}^T \eta_t\Big)^{-1}\Big(\sum_{t=1}^T \eta_t^2 + \log(1/\delta)\Big) \\ \leq 2C_3^2 R_T^2 n^{-1} + O(1)\Big(\sum_{t=1}^T \eta_t\Big)^{-1}\Big(\sum_{t=1}^T \eta_t^2 + \log(1/\delta)\Big).$$

This, together with Lemma III.1, gives the following inequality with probability $1-\delta$

$$\Big(\sum_{t=1}^T \eta_t\Big)^{-1}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 \\ = \begin{cases} O(\Lambda_{n,d,\delta})T^{2-3\theta} + O(T^{-\theta}\log(1/\delta)), & \text{if } \theta < 1/2 \\ O(\Lambda_{n,d,\delta})T^{\frac{1}{2}}\log T + O(T^{-\frac{1}{2}}\log(T/\delta)), & \text{if } \theta = 1/2 \\ O(\Lambda_{n,d,\delta})T^{1-\theta} + O(T^{\theta-1}\log(1/\delta)), & \text{if } \theta > 1/2, \end{cases}$$

where $\Lambda_{n,d,\delta} := n^{-1}(d + \log(1/\delta))\log^2(1/\delta)$.

If $\theta < 1/2$, we can choose $T \asymp (nd^{-1})^{\frac{1}{2(1-\theta)}}$ to derive the following inequality with probability $1-\delta$

$$\Big(\sum_{t=1}^T \eta_t\Big)^{-1}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 = O\Big((nd^{-1})^{\frac{\theta}{2(\theta-1)}}\log^3(1/\delta)\Big).$$

If $\theta = 1/2$, we can choose $T \asymp nd^{-1}$ to derive the following inequality with probability $1-\delta$

$$\Big(\sum_{t=1}^T \eta_t\Big)^{-1}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 = O\Big(n^{-\frac{1}{2}}d^{\frac{1}{2}}\log(T/\delta)\log^3(1/\delta)\Big).$$

If $\theta > 1/2$, we can choose $T \asymp (nd^{-1})^{\frac{1}{2(1-\theta)}}$ to derive the following inequality with probability $1-\delta$

$$\Big(\sum_{t=1}^T \eta_t\Big)^{-1}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 = O\Big(n^{-\frac{1}{2}}d^{\frac{1}{2}}\log^3(1/\delta)\Big).$$

The stated bound then follows. The proof is complete. □

IV Proofs of Learning Rates for Gradient-dominated Objectives

In this section, we present the proof of Theorem 7 on fast learning rates under gradient-dominance conditions. Our idea is to still control computational and statistical errors, separately. However, in this case we have a refined bound on computational errors as $\sum_{t=1}^T (t+t_0-1)\|\nabla F_S(\mathbf{w}_t)\|_2^2 = \widetilde{O}(T)$ and a refined estimate on the norm of SGD iterates for the associated step sizes. For brevity, we assume $\mu_S \leq 2$ in the proof.

Lemma IV.1. *For the step size $\eta_t = 2/(\mu_S(t+t_0)), t \in \mathbb{N}$ and $t_0 \geq 1$, we have*

$$\sum_{t=1}^T \eta_t \leq \frac{2}{\mu_S}\sum_{t=1}^T \frac{1}{t+t_0} \leq \frac{2}{\mu_S}\log(T+1).$$

Proof of Theorem 7. Since $t_0 \geq 4L/\mu_S$, we know $\eta_t \leq 1/(2L)$ and therefore Lemma 9 holds. According to Lemma 9 and (III.3), we know the existence of an event $\Omega_T^{(1)}$ with $\Pr\{\Omega_T^{(1)}\} \geq 1-\delta/2$ conditioned on which we know $\|\mathbf{w}_{t+1}\|_2$ can be upper bounded by for all $t = 1, \ldots, T$

$$C_2\Big((2L)^{-\frac{1}{2}}\Big(\sum_{t=1}^T \eta_t\Big)^{\frac{1}{2}} + 1\Big)\Big(\Big(\sum_{t=1}^T \eta_t\Big)^{\frac{1}{2}} + 1\Big)\log(8/\delta) \\ \leq 2C_2 \max\big\{(2L)^{-\frac{1}{2}}, 1\big\}\Big(\sum_{t=1}^T \eta_t + 1\Big)\log(8/\delta) \\ \leq 2C_2 \max\big\{(2L)^{-\frac{1}{2}}, 1\big\}\Big(2\mu_S^{-1}\log(T+1)+1\Big)\log(8/\delta) \leq C_{T,\delta},$$

where we have used Lemma IV.1 and have introduced $(2\mu_S^{-1} \geq 1)$

$$C_{T,\delta} := 2C_2 \max\{(2/L)^{\frac{1}{2}}, 2\}\mu_S^{-1}\log\big(e(T+1)\big)\log(8/\delta).$$

By (III.6) and Assumption 4, we know $F_S(\mathbf{w}_{t+1})$ can be upper bounded by

$$F_S(\mathbf{w}_t) + \xi_t - 2^{-1}\eta_t\|\nabla F_S(\mathbf{w}_t)\|_2^2 + L\eta_t^2\sigma^2 + L\eta_t^2\xi_t' \\ \leq F_S(\mathbf{w}_t) + \xi_t - 4^{-1}\eta_t\|\nabla F_S(\mathbf{w}_t)\|_2^2 \\ + \eta_t\mu_S\big(F_S(\mathbf{w}_S) - F_S(\mathbf{w}_t)\big) + L\eta_t^2\sigma^2 + L\eta_t^2\xi_t',$$

where $\{\xi_t\}_t$ and $\{\xi_t'\}_t$ are defined in (III.2) and (III.5), respectively. It then follows that

$$\frac{1}{2\mu_S(t+t_0)}\|\nabla F_S(\mathbf{w}_t)\|_2^2 + F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_S) \\ \leq (1-\eta_t\mu_S)\big(F_S(\mathbf{w}_t) - F_S(\mathbf{w}_S)\big) + \xi_t + L\eta_t^2\sigma^2 + L\eta_t^2\xi_t' \\ = \frac{t+t_0-2}{t+t_0}\big(F_S(\mathbf{w}_t) - F_S(\mathbf{w}_S)\big) + \xi_t + \frac{4L\sigma^2}{\mu_S^2(t+t_0)^2} + \frac{4L\xi_t'}{\mu_S^2(t+t_0)^2}.$$

Multiplying both sides by $(t+t_0)(t+t_0-1)$, we derive

$$(2\mu_S)^{-1}(t+t_0-1)\|\nabla F_S(\mathbf{w}_t)\|_2^2 + (t+t_0-1)(t+t_0)\times \\ \big(F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_S)\big) \leq (t+t_0-2)(t+t_0-1)\big(F_S(\mathbf{w}_t) - F_S(\mathbf{w}_S)\big) \\ + (t+t_0-1)(t+t_0)\xi_t + 4L\sigma^2\mu_S^{-2} + 4L\mu_S^{-2}\xi_t'.$$

Taking a summation of the above inequality from $t = 1$ to $t = T$, we derive

$$
\begin{aligned}
&(2\mu_S)^{-1}\sum_{t=1}^{T}(t+t_0-1)\|\nabla F_S(\mathbf{w}_t)\|_2^2 \\
&\qquad + (T+t_0-1)(T+t_0)\big(F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}_S)\big) \\
&\leq (t_0-1)t_0\big(F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S)\big) + \sum_{t=1}^{T}(t+t_0-1)(t+t_0)\xi_t \\
&\qquad + 4L\sigma^2\mu_S^{-2}T + 4L\mu_S^{-2}\sum_{t=1}^{T}\xi'_t. \qquad \text{(IV.1)}
\end{aligned}
$$

Introduce two sequences of martingale difference sequences

$$
\tilde{\xi}_t = \eta_t\langle\nabla F_S(\mathbf{w}_t) - \nabla f(\mathbf{w}_t; z_{j_t}), \nabla F_S(\mathbf{w}_t)\rangle\mathbb{I}_{\{\|\mathbf{w}_t\|_2 \leq C_{T,\delta}\}}
$$

$$
\begin{aligned}
\tilde{\xi}'_t = \Big(&\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|_2^2 - \\
&\mathbb{E}_{j_t}\big[\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|_2^2\big]\Big)\mathbb{I}_{\{\|\mathbf{w}_t\|_2 \leq C_{T,\delta}\}}.
\end{aligned}
$$

The following inequality holds for all $t = 1, \ldots, T$

$$
\begin{aligned}
&(t+t_0-1)(t+t_0)|\tilde{\xi}_t| \leq 2\mu_S^{-1}(t+t_0-1)\times \\
&\|\nabla F_S(\mathbf{w}_t) - \nabla f(\mathbf{w}_t; z_{j_t})\|_2\|\nabla F_S(\mathbf{w}_t)\|_2\mathbb{I}_{\{\|\mathbf{w}_t\|_2 \leq C_{T,\delta}\}} \\
&\leq 4\mu_S^{-1}(T+t_0-1)\big(LC_{T,\delta} + \tilde{b}\big)^2,
\end{aligned}
$$

where we have used (II.2). Furthermore, the conditional variances can be bounded by

$$
\begin{aligned}
&\mathbb{E}_{j_t}\big[(t+t_0-1)^2(t+t_0)^2\tilde{\xi}_t^2\big] \\
&\leq 4\mu_S^{-2}(t+t_0-1)^2\|\nabla F_S(\mathbf{w}_t)\|_2^2\mathbb{E}_{j_t}\|\nabla F_S(\mathbf{w}_t) - \nabla f(\mathbf{w}_t; z_{j_t})\|_2^2 \\
&\leq 4\mu_S^{-2}\sigma^2(t+t_0-1)^2\|\nabla F_S(\mathbf{w}_t)\|_2^2.
\end{aligned}
$$

Applying Part (b) of Lemma I.1 together with the above bounds on magnitudes and variances, we know the existence of $\Omega_T^{(2)}$ with $\Pr\{\Omega_T^{(2)}\} \geq 1 - \delta/8$ conditioned on which the following inequality holds

$$
\begin{aligned}
&\sum_{t=1}^{T}(t+t_0-1)(t+t_0)\tilde{\xi}_t \\
&\leq \frac{4\rho\sigma^2\sum_{t=1}^{T}(t+t_0-1)^2\|\nabla F_S(\mathbf{w}_t)\|_2^2}{4\mu_S^2\mu_S^{-1}(T+t_0-1)\widetilde{C}_{T,\delta}^2} + \frac{4(T+t_0-1)\widetilde{C}_{T,\delta}^2\log\frac{8}{\delta}}{\rho\mu_S} \\
&\leq \frac{\rho\sigma^2\sum_{t=1}^{T}(t+t_0-1)\|\nabla F_S(\mathbf{w}_t)\|_2^2}{\mu_S\widetilde{C}_{T,\delta}^2} + \frac{4(T+t_0-1)\widetilde{C}_{T,\delta}^2\log\frac{8}{\delta}}{\rho\mu_S} \\
&\leq (4\mu_S)^{-1}\sum_{t=1}^{T}(t+t_0-1)\|\nabla F_S(\mathbf{w}_t)\|_2^2 \\
&+ 4\mu_S^{-1}(T+t_0-1)\log(8/\delta)\max\big\{4\sigma^2, (LC_{T,\delta} + \tilde{b})^2\big\}, \qquad \text{(IV.2)}
\end{aligned}
$$

where we introduce $\widetilde{C}_{T,\delta} := LC_{T,\delta} + \tilde{b}$ and set $\rho = \min\{(4\sigma^2)^{-1}(LC_{T,\delta} + \tilde{b})^2, 1\}$. According to the definition of $\tilde{\xi}'_t$ and (II.2), we know

$$
\tilde{\xi}'_t \leq 2\big(\|\nabla f(\mathbf{w}_t; z_{j_t})\|_2^2 + \|\nabla F_S(\mathbf{w}_t)\|_2^2\big)\mathbb{I}_{\{\|\mathbf{w}_t\|_2 \leq C_{T,\delta}\}} \leq 4\widetilde{C}_{T,\delta}^2.
$$

In a similar way, we can also show $\tilde{\xi}'_t \geq -4\widetilde{C}_{T,\delta}^2$. We now can apply Part (a) of Lemma I.1 to show the existence of $\Omega_T^{(3)}$ with $\Pr\{\Omega_T^{(3)}\} \geq 1 - \delta/8$ conditioned on which the following inequality holds

$$
\sum_{t=1}^{T}\tilde{\xi}'_t \leq 4(LC_{T,\delta} + \tilde{b})^2(2T\log(8/\delta))^{\frac{1}{2}}.
$$

Under the event $\Omega_T^{(1)}$, we know $\tilde{\xi}_t = \xi_t$ and $\tilde{\xi}'_t = \xi'_t$. Plugging the above inequality and (IV.2) back into (IV.1), we derive the following inequality under the event $\Omega_T^{(1)} \cap \Omega_T^{(2)} \cap \Omega_T^{(3)}$

$$
\begin{aligned}
&(4\mu_S)^{-1}\sum_{t=1}^{T}(t+t_0-1)\|\nabla F_S(\mathbf{w}_t)\|_2^2 \\
&+ (T+t_0-1)(T+t_0)\big(F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}_S)\big) \\
&\leq (t_0-1)t_0\big(F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S)\big) \\
&+ 4\mu_S^{-1}(T+t_0+1)\log(8/\delta)\max\{4\sigma^2, (LC_{T,\delta} + \tilde{b})^2\} \\
&+ 4L\sigma^2\mu_S^{-2}T + 16L\mu_S^{-2}(LC_{T,\delta} + \tilde{b})^2(2T\log(8/\delta))^{\frac{1}{2}}.
\end{aligned}
$$

This together with (3) shows that (notice $C_{T,\delta} = O(\log T\log(1/\delta))$)

$$
\|\nabla F_S(\mathbf{w}_{T+1})\|_2^2 = O\big(T^{-1}\log^2 T\log^3(1/\delta)\big). \qquad \text{(IV.3)}
$$

According to Corollary 2, we know the existence of $\Omega_T^{(4)}$ with $\Pr\{\Omega_T^{(4)}\} \geq 1 - \delta/4$ such that

$$
\sup_{\|\mathbf{w}\|_2 \leq C_{T,\delta}}\big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 \leq C_3 C_{T,\delta}n^{-\frac{1}{2}},
$$

where C_3 is defined in (III.15). Under the event of $\Omega_T^{(1)} \cap \Omega_T^{(2)} \cap \Omega_T^{(3)} \cap \Omega_T^{(4)}$, we can combine (IV.3) and the above inequality to derive the following inequality (notice $C_{T,\delta} = O(\log T\log(1/\delta))$)

$$
\begin{aligned}
&\|\nabla F(\mathbf{w}_{T+1})\|_2^2 \\
&\leq 2\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|_2^2 + 2\|\nabla F_S(\mathbf{w}_{T+1})\|_2^2 \\
&\leq 2\sup_{\|\mathbf{w}\|_2 \leq C_{T,\delta}}\big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2^2 + O\big(T^{-1}\log^2 T\log^3(1/\delta)\big) \\
&= O\Big(\Big(n^{-1}(d + \log(1/\delta))\Big) + T^{-1}\log(1/\delta)\Big)\log^2 T\log^2(1/\delta)\Big).
\end{aligned}
$$

Using the gradient-dominance condition and choosing $T \asymp n$, we derive the stated inequality with probability $1 - \delta$. The proof is complete. □

References

[1] P. Tarres and Y. Yao, "Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence," *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 5716–5735, 2014.

[2] D. J. Foster, A. Sekhari, and K. Sridharan, "Uniform convergence of gradients for non-convex learning and optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 8759–8770.

[3] A. Maurer, "A vector-contraction inequality for rademacher complexities," in *International Conference on Algorithmic Learning Theory*, 2016, pp. 3–17.

[4] Y. Ying and C. Campbell, "Rademacher chaos complexities for learning the kernel problem," *Neural computation*, vol. 22, no. 11, pp. 2858–2886, 2010.

[5] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, 1999, vol. 94.