Generalization performance of multi-pass stochastic gradient descent with convex loss functions

Lei, Yunwen; Hu, Ting; Tang, Ke

Journal of Machine Learning Research 22 (2021) 1-41 Submitted 8/19; Revised 12/20; Published 1/21

Generalization Performance of Multi-pass Stochastic Gradient Descent with Convex Loss Functions

Yunwen Lei* Y.LEI@BHAM.AC.UK
School of Computer Science
University of Birmingham
Birmingham B15 2TT, United Kingdom

Ting Hu TINGHU@WHU.EDU.CN
School of Mathematics and Statistics
Wuhan University
Wuhan 430072, China

Ke Tang TANGK3@SUSTECH.EDU.CN
Research Institute of Trustworthy Autonomous Systems
Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen 518055, China

Editor: Lorenzo Rosasco

Abstract

Stochastic gradient descent (SGD) has become the method of choice to tackle large-scale datasets due to its low computational cost and good practical performance. Learning rate analysis, either capacity-independent or capacity-dependent, provides a unifying viewpoint to study the computational and statistical properties of SGD, as well as the implicit regularization by tuning the number of passes. Existing capacity-independent learning rates require a nontrivial bounded subgradient assumption and a smoothness assumption to be optimal. Furthermore, existing capacity-dependent learning rates are only established for the specific least squares loss with a special structure. In this paper, we provide both optimal capacity-independent and capacity-dependent learning rates for SGD with *general* convex loss functions. Our results require neither bounded subgradient assumptions nor smoothness assumptions, and are stated with high probability. We achieve this improvement by a refined estimate on the norm of SGD iterates based on a careful martingale analysis and concentration inequalities on empirical processes.

Keywords: Stochastic gradient descent, learning theory, generalization bound

1. Introduction

1.1 Background

Stochastic gradient descent (SGD) has found wide applications in machine learning due to its simplicity in implementation, low memory requirement and low computational complexity per iteration, as well as good practical behavior (Zhang, 2004; Bach and Moulines, 2013;

*. Part of the work was done at the Department of Computer Science and Engineering, Southern University of Science and Technology.

Rakhlin et al., 2012; Shamir and Zhang, 2013; Bottou et al., 2018; Orabona, 2019). As an iterative method, SGD minimizes empirical risks by moving iterates along the direction of a negative gradient calculated based on a loss function on either a single training example or a batch of few examples. This strategy of processing few examples per iteration makes SGD particularly suitable for practical applications with very large data sets (Zhang, 2004; Bach and Moulines, 2013), which are becoming ubiquitous in the big data era.

Theoretical analysis of SGD was mainly conducted from an optimization viewpoint to understand how the optimization error would decrease along the iterations (Zhang, 2004; Nemirovski et al., 2009; Rakhlin et al., 2012; Shamir and Zhang, 2013). In a machine learning setting, we are more interested in the population risks of SGD, i.e., how the model output by SGD would generalize to unseen examples (Bousquet and Bottou, 2008; Vapnik, 1998; Pillaud-Vivien et al., 2018a). To this aim, we need also to take into account simultaneously the difference between population risks and empirical risks, which are referred to estimation errors in the statistical learning theory setting. Intuitively, optimization errors would decrease along the SGD iterations, while the complexity of the SGD iterates and therefore the estimation errors would increase meanwhile. This suggests that an implicit regularization can be achieved by tuning the number of passes to balance the optimization and estimation errors. A unifying consideration of optimization and estimation errors for SGD would provide a theoretical principle towards this aim (Lin et al., 2016a; Lin and Rosasco, 2017; Yao et al., 2007).

Existing learning rates of SGD are mainly derived in the setting with only one-pass over the data allowed, i.e., each example can be used at most once (Ying and Pontil, 2008; Orabona, 2014; Ying and Zhou, 2006; Bach and Moulines, 2011). However, in practical applications the strategy of multi-pass SGD is often adopted to produce a model with good generalization performance (Pillaud-Vivien et al., 2018b). The key difference between the learning rate analysis is that the expectation of loss functions over the stochastic algorithm for one-pass SGD is the population risk, while the expectation for multi-pass SGD is the empirical risk. Therefore, the learning rate analysis of multi-pass SGD raises a new challenge to control the estimation errors.

Motivated by the popularity of multi-pass SGD, the generalization properties of multi-pass SGD have received increasing attention recently. Stability of multi-pass SGD were established in Hardt et al. (2016), which in turn yields capacity-independent learning rates ignoring the capacity information on the hypothesis spaces (Lin et al., 2016a; Feldman and Vondrak, 2019). The learning rates in Hardt et al. (2016); Lin et al. (2016a); Feldman and Vondrak (2019) require to impose a bounded subgradient assumption for iterates, and are either stated for smooth loss functions (Hardt et al., 2016; Feldman and Vondrak, 2019) or not optimal for non-smooth and Lipschitz loss functions (Lin et al., 2016a). Capacity-dependent learning rates were also studied recently, where the information on the capacity of hypothesis spaces is exploited to derive better learning rates (Rosasco and Villa, 2015; Dieuleveut and Bach, 2016; Lin and Rosasco, 2017; Pillaud-Vivien et al., 2018b; Mücke et al., 2019). However, to our best knowledge, the existing capacity-dependent learning rates of multi-pass SGD are all stated for the specific least squares loss (Rosasco and Villa, 2015; Dieuleveut and Bach, 2016; Lin and Rosasco, 2017; Pillaud-Vivien et al., 2018b; Mücke et al., 2019). Indeed, a key property in these analysis is the closed-form update of the SGD iterates by integral operator which does not hold for general loss functions.

1.2 Main Contributions

In this paper, we investigate both capacity-independent and capacity-dependent learning rates for multi-pass SGD with general convex loss functions. Our results outperform the existing capacity-independent analysis by removing bounded subgradient assumptions as well as smoothness assumptions on loss functions (Hardt et al., 2016; Lin et al., 2016a; Feldman and Vondrak, 2019), and complements the existing capacity-dependent analysis for the specific least squares loss (Rosasco and Villa, 2015; Dieuleveut and Bach, 2016; Lin and Rosasco, 2017) by considering general convex loss functions. Furthermore, our results are stated with high probability and are optimal in the sense of matching the best available bound for Tikhonov regularization where optimization errors are ignored. Our results show that different number of passes are required for multi-pass SGD with different polynomially decaying stepsizes to achieve optimal learning rates, which show in a clear way how statistical errors and computational resources should be balanced for general convex loss functions. Our novelty in the analysis consists in an exploitation of self-bounding properties of loss functions to remove bounded subgradient assumptions, and a refined estimate on the norm of SGD iterates based on a careful martingale analysis together with concentration inequalities in empirical processes. In particular, we build a novel polynomial inequality to relate the norm of an iterate to the norm of previous iterates.

This paper is extended from our previous conference article published in Advance in Neural Information Processing Systems 2018 (Lei and Tang, 2018), where capacity-independent bounds were established. However, the analysis there requires to impose a non-intuitive assumption on the existence of an empirical risk minimizer with a finite norm, which depends on the sampling of training examples and is hard to check in practice. We successfully remove this assumption by a different error decomposition with respect to (w.r.t.) a regularized risk minimizer together with concentration inequalities to relate empirical risks to population risks in the norm estimation. Furthermore, we also add new capacity-dependent learning rates in this paper, which were not considered in the conference article. Our capacity-dependent learning rates can be as fast as $\widetilde{O}(n^{-1})$[1] under suitable capacity assumption on the hypothesis space and variance-expectation assumption, where n is the sample size.

Organization of the paper. The remainder of this paper is organized as follows. We formulate the problem and present main results in Section 2. We give the comparison of our results with the state of the art in Section 3. We provide preliminary complexity control in Section 4, which is then used to study the capacity-independent and capacity-dependent rates in Section 5 and Section 6, respectively. The conclusion is given in Section 7.

2. Problem Formulation and Main Results

Let ρ be a probability measure defined on a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$ being the input space and $\mathcal{Y} \subseteq [-b, b]$ being the output space, where $\mathcal{X}$ may be any set equipped with a measure. We assume a training sample $\mathbf{z} = \{z_1, \ldots, z_n\}$ of size $n \in \mathbb{N}$ is drawn independently from ρ, and our aim is to learn a hypothesis $h : \mathcal{X} \mapsto \mathbb{R}$ from a hypothesis space $\mathcal{W}$ with good generalization performance. The quality of h at (x, y) is quantified by $\ell(h(x), y)$, where

1. We use $\widetilde{O}$ to hide logarithmic factors.

$\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ is convex w.r.t. the first argument. The population risk and empirical risk of h are defined respectively by $\mathcal{E}(h) = \mathbb{E}_z\big[\ell(h(x), y)\big]$ and $\mathcal{E}_{\mathbf{z}}(h) = \frac{1}{n}\sum_{i=1}^n \ell\big(h(x_i), y_i\big)$. The best model minimizing the population risk then becomes $h_\rho = \arg\min_h \mathcal{E}(h)$. We consider a non-parametric learning setting with $\mathcal{W}$ being a reproducing kernel Hilbert space (RKHS) associated to a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ which is continuous, symmetric and positive semi-definite (Aronszajn, 1950; Schölkopf and Smola, 2001; Cristianini and Shawe-Taylor, 2000; Shawe-Taylor and Cristianini, 2004). Let $\Phi : \mathcal{X} \mapsto \mathcal{W}$ be the associated feature map satisfying $K(x, \tilde{x}) = \langle \Phi(x), \Phi(\tilde{x}) \rangle$ for all $x, \tilde{x} \in \mathcal{X}$. In this learning setting, the candidate models take the form $h_{\mathbf{w}}(x) = \langle \mathbf{w}, \Phi(x) \rangle$ with $\mathbf{w} \in \mathcal{W}$. For brevity, we denote the norm in the RKHS $\mathcal{W}$ by $\|\cdot\|_2$ and introduce the abbreviations $\mathcal{E}(\mathbf{w}) = \mathcal{E}(h_{\mathbf{w}}), \mathcal{E}_{\mathbf{z}}(\mathbf{w}) = \mathcal{E}_{\mathbf{z}}(h_{\mathbf{w}})$ for any $\mathbf{w} \in \mathcal{W}$. Let $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x, x)}$.

Let $\mathbf{w}_1 = 0$ and $\{\eta_t\}_{t \in \mathbb{N}}$ be a positive stepsize sequence. At the t-th iteration, we randomly choose an index j_t from the uniform distribution over $\{1, \ldots, n\}$ and update $\mathbf{w}_{t+1}$ as follows

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \ell'\big(\langle \mathbf{w}_t, \Phi(x_{j_t}) \rangle, y_{j_t}\big)\Phi(x_{j_t}) = \mathbf{w}_t - \eta_t f'(\mathbf{w}_t, z_{j_t}), \tag{2.1}$$

where ℓ' and f' respectively denote subgradients of ℓ and f w.r.t. the first argument, and we introduce the notation $f(\mathbf{w}, z) = \ell(\langle \mathbf{w}, \Phi(x) \rangle, y)$. In this paper, we are interested in the population risk of the iterates produced by (2.1). To this aim, we need to introduce some assumptions.

Our first assumption is the so-called self-bounding property of loss functions meaning that the subgradient can be bounded by function values.

Assumption 1 *We assume the existence of $\widetilde{A}$ and $\widetilde{B} \geq 0$ such that*

$$|\ell'(a, y)|^2 \leq \widetilde{A}\ell(a, y) + \widetilde{B}, \quad \forall a \in \mathbb{R}, y \in \mathcal{Y}. \tag{2.2}$$

Remark 1 *Many popular loss functions satisfy* (2.2)*, including the p-norm hinge loss $\ell(a, y) = \max\{0, 1 - ya\}^p$ $(1 \leq p \leq 2)$ (Steinwart and Christmann, 2008), the logistic loss $\ell(a, y) = \log(1 + \exp(-ya))$ for classification, and the p-th power absolute distance loss $\ell(a, y) = |a - y|^p$ $(1 \leq p \leq 2)$, the ϵ-insensitive loss $\ell(a, y) = \max\{0, |y - a| - \epsilon\}$, the Huber loss $\ell(a, y) = (a - y)^2$ if $|a - y| \leq 1$ and $\ell(a, y) = 2|a - y| - 1$ otherwise for regression (Zhang, 2004). We refer the interested readers to Zhang (2004) for constants $\widetilde{A}, \widetilde{B}$ in* (2.2) *with different loss functions ℓ.*

For loss functions satisfying (2.2), we have the following inequality useful for our learning rate analysis (Part (a) of Lemma 16)

$$\|f'(\mathbf{w}, z)\|_2^2 \leq Af(\mathbf{w}, z) + B, \quad \forall \mathbf{w} \in \mathcal{W} \text{ and } z \in \mathcal{Z}, \tag{2.3}$$

where $A = \widetilde{A}\kappa^2$ and $B = \widetilde{B}\kappa^2$.

Assumption 2 *Let $\lambda > 0$ and $\mathbf{w}_\lambda$ be the minimizer*

$$\mathbf{w}_\lambda := \arg\min_{\mathbf{w} \in \mathcal{W}} \big\{\mathcal{E}(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2\big\}. \tag{2.4}$$

The approximation error is defined by $\mathcal{D}(\lambda) = \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + \lambda\|\mathbf{w}_\lambda\|_2^2$. We assume that $\mathcal{D}(\lambda)$ enjoys a polynomial decay with exponent $0 < \alpha \leq 1$ in the sense $\mathcal{D}(\lambda) \leq c_\alpha \lambda^\alpha, \forall \lambda > 0$

for some $c_\alpha > 0$.

Remark 2 *Assumption 2 is standard in learning theory and satisfied under some mild conditions on the smoothness of the function h_ρ and the representation power of $\mathcal{W}$ (Cucker and Zhou, 2007; Steinwart and Christmann, 2008). If ℓ is smooth, then $\mathcal{D}(\lambda)$ can be controlled by $\widetilde{D}(\lambda) := \inf_{\mathbf{w}\in\mathcal{W}} \|h_{\mathbf{w}} - h_\rho\|^2_{L^2_{\rho_\mathcal{X}}} + \lambda\|\mathbf{w}\|_2^2$, which quantifies the approximation of h_ρ by RKHS in $L^2_{\rho_\mathcal{X}}$ (square-integrable function class with marginal measure $\rho_\mathcal{X}$) and is well studied in approximation theory. $\widetilde{D}(\lambda)$ decays polynomially with $\alpha \in (0,1]$ if $h_\rho \in L_K^{\alpha/2}(L^2_{\rho_\mathcal{X}})$, where $L_K : L^2_{\rho_\mathcal{X}} \mapsto L^2_{\rho_\mathcal{X}}$ is the integral operator associated to K (Cucker and Zhou, 2007, Proposition 8.5). Similar results hold if ℓ is Lipschitz continuous. Assumption 2 also holds if we use Gaussian kernels with flexible variances and distributions with geometric noise conditions (Steinwart and Scovel, 2007).*

Our next assumption is to assume that the projection of a predicted output onto the interval $[-b, b]$ can always improve the prediction accuracy, which is natural since we assume the output belongs to $[-b, b]$. Examples of loss functions satisfying this assumption include the p-norm hinge loss, p-th power absolute distance loss ($1 \le p \le 2$), the Huber loss and the ϵ-insensitive loss (Wu et al., 2007; Steinwart and Scovel, 2007). With this assumption, we can further improve the learning performance by taking a projection (Steinwart and Christmann, 2009; Cucker and Zhou, 2007; Steinwart and Christmann, 2008; Wu et al., 2007). For any $h : \mathcal{X} \mapsto \mathbb{R}$, we define $\hat{h} : \mathcal{X} \mapsto [-b, b]$ by

$$\hat{h}(x) = \min\big\{b, \max\{-b, h(x)\}\big\}.$$

Assumption 3 *We assume $\ell(\hat{h}(x), y) \le \ell(h(x), y)$ for all h, x and y.*

2.1 Capacity-independent Learning Rates

Our first main result is a probabilistic learning rate for an average of iterates produced by (2.1). It is known that any nonnegative and convex loss function satisfying Assumption 1 would satisfy $\ell(a, y) \le c(a^2 + 1), \forall a \in \mathbb{R}$ and $y \in \mathcal{Y}$ for a constant c (Lei and Tang, 2018). Therefore, we can introduce parameters to better quantify the behavior of ℓ. Let $q \in [1, 2]$ and $c_q \ge 0$ be constants such that

$$\ell(a, y) \le c_q(|a|^q + 1), \quad \forall a \in \mathbb{R}, y \in \mathcal{Y}. \tag{2.5}$$

All Lipschitz loss functions satisfy (2.5) with $q = 1$, including the hinge loss function and the logistic loss function. The p-norm hinge loss function and the p-th power absolute loss function satisfy (2.5) with $q = p$ ($p \in [1, 2]$).

Remark 3 *We use Eq. (2.5) to further characterize the growth behavior of loss functions. As we will show in Remark 11 and Remark 15, we recover (almost) exactly the existing learning rates for Tikhonov regularization with Lipschitz loss functions ($q = 1$) and SGD with the least squares loss ($q = 2$). This shows that the above growth assumption seems to capture the learning rates of SGD in different scenario. A similar assumption as $|\ell'(a, y)| \le c_q(|\alpha|^{q-1}+1)$ for $q \ge 1$ was also imposed in the literature (Lin et al., 2016b), and the learning rates there also depend on the parameter q. It would be very interesting to investigate*

whether this assumption is really necessary or can be removed. This growth assumption can be implied by assuming the loss function to be Nemitski when $\mathcal{Y}$ is bounded, which is introduced in Vito et al. (2004) (see also Steinwart and Christmann (2008)). Nemitski condition is satisfied by most loss functions and provides natural variational characterization of loss functions.

The learning rate in the following theorem is capacity-independent since it does not rely on an assumption on the capacity of the associated hypothesis spaces, e.g., covering numbers. The proof of Theorem 4 is given in Section 5.3. The notation $U \asymp V$ means that there are constants $\widetilde{C}_1, \widetilde{C}_2 > 0$ such that $\widetilde{C}_1 U \leq V \leq \widetilde{C}_2 U$, where $\widetilde{C}_1$ and $\widetilde{C}_2$ are two expressions depending on other variables.

Theorem 4 *Let Assumptions 1-3 hold. Assume $\delta \in (0,1)$ and $\eta_1 \leq 1/A$. If $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ is the sequence produced by (2.1) with $\eta_t = \eta_1 t^{-\theta}, \theta > 1/2$. then with probability at least $1-\delta$ we have the following inequality for $T \asymp n^{\frac{1}{(1+\alpha)(1-\theta)}}$*

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(n^{-\frac{\alpha}{1+\alpha}} \log^{\frac{3}{2}} \frac{n}{\delta}),$$

where $\bar{\mathbf{w}}_T = \big(\sum_{t=1}^T \eta_t\big)^{-1} \sum_{t=1}^T \eta_t \mathbf{w}_t$ is an average of the first T iterates.

Remark 5 *The bound $O\big(n^{-\frac{\alpha}{1+\alpha}} \log^{\frac{3}{2}} \frac{n}{\delta}\big)$ coincides with $O(n^{-\frac{\alpha}{1+\alpha}} \log n)$ (up to a logarithmic factor) established in expectation for convex and smooth loss functions (Lin et al., 2016a), and largely improves the bound $O(n^{-\frac{\alpha}{1+2\alpha}} \log n)$ in expectation for convex and non-smooth loss functions (Lin et al., 2016a). In particular, if $\alpha = 1$ we derive the optimal bound $O(n^{-\frac{1}{2}} \log^{\frac{3}{2}} \frac{n}{\delta})$ in the capacity-independent analysis (up to a logarithmic factor). It is also clear that SGD with different stepsizes can achieve similar learning rates. However, the computational complexity (measured by T) to fulfill this statistical potential can be significantly different.*

Remark 6 *In the conference article (Lei and Tang, 2018), we also establish the learning rate $O\big(n^{-\frac{\alpha}{1+\alpha}} \log^{\frac{3}{2}} \frac{n}{\delta}\big)$ with high probability. However, the discussion there requires to assume the existence of an empirical risk minimizer $\mathbf{w}_\mathbf{z} = \arg\inf_\mathbf{w} \mathcal{E}_\mathbf{z}(\mathbf{w})$ with $\|\mathbf{w}_\mathbf{z}\|_2 < \infty$. Since $\mathbf{w}_\mathbf{z}$ varies with different realizations of training examples, this assumption is not intuitive and hard to check in practice. Furthermore, this assumption leads to a misleading result that over-fitting would not happen for SGD. In Theorem 4, we establish the same learning rate without imposing such an assumption.*

2.2 Capacity-dependent Learning Rates

In this section, we show that learning rates faster than $O(n^{-\frac{1}{2}})$ are possible if we impose assumptions on the capacity of the hypothesis space as well as a relationship between variance and expectation of excess loss functions. We measure the capacity of a function space by empirical covering numbers defined as the number of elements required to approximate the whole function space to some accuracy.

Definition 7 (Covering number) *Let $\widetilde{\mathcal{F}}$ be a class of real-valued functions defined over a space $\widetilde{\mathcal{Z}}$ and $\widetilde{S} := \{\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n\} \subset \widetilde{\mathcal{Z}}$ of cardinality n. For any $\epsilon > 0$, the empirical ℓ_2-norm*

covering number $\mathcal{N}_2(\epsilon, \widetilde{\mathcal{F}}, \widetilde{S})$ *w.r.t.* $\widetilde{S}$ *is defined as the minimal number* m *of a collection of vectors* $\mathbf{v}^1, \ldots, \mathbf{v}^m \in \mathbb{R}^n$ *such that (*$\mathbf{v}_i^j$ *is the* i*-th component of the vector* $\mathbf{v}^j$*)*

$$\sup_{f \in \widetilde{\mathcal{F}}} \min_{j=1,\ldots,m} n^{-\frac{1}{2}} \Big(\sum_{i=1}^n |f(\tilde{\mathbf{z}}_i) - \mathbf{v}_i^j|^2 \Big)^{\frac{1}{2}} \leq \epsilon.$$

In the following assumption, we assume that the logarithm of covering numbers grows polynomially w.r.t. the reciprocal of approximation accuracy. This is a standard assumption in statistical learning theory which is closely related to an assumption on the eigenvalue decay of the integral operator associated with K (Cucker and Zhou, 2007; Steinwart and Christmann, 2008). Denote $\mathcal{H}_R = \{h_{\mathbf{w}} : \|\mathbf{w}\|_2 \leq R\}$.

Assumption 4 *Let* $S_{\mathbf{x}} \subset \mathcal{X}$ *be a set of random examples. We assume the existence of some* $\zeta \in (0, 2)$ *and* $c_\zeta > 0$ *such that* $\mathbb{E}_{S_{\mathbf{x}}}\big[\log \mathcal{N}_2(\epsilon, \mathcal{H}_1, S_{\mathbf{x}})\big] \leq c_\zeta \epsilon^{-\zeta}$ *for all* $\epsilon > 0$*, where* $\mathbb{E}_{S_{\mathbf{x}}}$ *denotes the expectation w.r.t.* $S_{\mathbf{x}}$*.*

We also need to impose a variance-expectation assumption on excess loss functions, which means that the variance of the function $x \mapsto \ell(h(x), y) - \ell(h_\rho(x), y)$ can be bounded by its expectation (Bartlett et al., 2006; Tsybakov, 2004).

Assumption 5 *We assume the existence of* $c_\beta > 0$ *and* $\beta \in (0, 1]$ *such that the following inequality holds for all* $h : \mathcal{X} \mapsto [-b, b]$

$$\mathbb{E}\big[\big(\ell(h(x), y)) - \ell(h_\rho(x), y)\big)^2\big] \leq c_\beta \big(\mathcal{E}(h) - \mathcal{E}(h_\rho)\big)^\beta.$$

The variance-expectation assumption plays an important role in deriving fast learning rates due to the intuitive observation that a good model with low population risk also exhibits a low variance under Assumption 5, and therefore one can apply Bernstein-type concentration inequalities to exploit this variance assumption (Blanchard et al., 2008; Bartlett et al., 2006; Tsybakov, 2004; Blanchard et al., 2003). It was shown that Assumption 5 holds for loss functions ℓ satisfying some strict convexity (Bartlett et al., 2006), which include p-norm absolute distance loss, p-norm hinge loss with $p \in (1, 2]$, truncated least squares loss function, logistic loss function and exponential loss function (Bartlett et al., 2006). Assumption 5 is also related to the property of marginal condition for classification problems. For example, if the conditional property $\rho(y|x)$ satisfies the so-called Tsybakov margin condition (Tsybakov, 2004) as follows

$$\rho_{\mathcal{X}}\big(\big\{x \in \mathcal{X} : |\rho(1|x) - 1/2| \leq \delta\big\}\big) \leq C\delta^s$$

for some C and all $\delta > 0$, then Assumption 5 holds with $\beta = \frac{s}{s+1}$ for the hinge loss.

Under the above assumptions, we now present our second main results on fast learning rates of an average of SGD iterates. The proof of Theorem 8 is given in Section 6.3.

Theorem 8 *Let Assumptions 1-5 hold. Assume* $\delta \in (0, 1)$ *and* $\eta_1 \leq 1/A$*. Let* $\{\mathbf{w}_t\}_{t \in \mathbb{N}}$ *be the sequence produced by* (2.1) *with* $\eta_t = \eta_1 t^{-\theta}, \theta > 1/2$*.*

(a) If $2\alpha + \alpha\beta\zeta + \zeta + \alpha q \geq \alpha\zeta + 2\alpha\beta + q$*, then with* $T \asymp n^{\frac{2}{(\zeta + 4\alpha + \alpha\beta\zeta - \alpha\zeta - 2\alpha\beta)(1-\theta)}}$ *we derive the following inequality with probability at least* $1 - \delta$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(\log^{\frac{3}{2}}(n/\delta)) n^{\frac{2\alpha}{(\alpha-1)\zeta - 4\alpha + 2\alpha\beta - \alpha\beta\zeta}}. \tag{2.6}$$

(b) If $2\alpha + \alpha\beta\zeta + \zeta + \alpha q < \alpha\zeta + 2\alpha\beta + q$, *then with* $T \asymp n^{\frac{2(2\alpha-2\alpha\beta+\alpha\beta\zeta-\alpha\zeta-4+2\beta-\beta\zeta+\zeta+\alpha q-q)}{(4-2\beta+\beta\zeta)(\alpha q-2\alpha-q)(1-\theta)}}$ *we derive the following inequality with probability at least* $1-\delta$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(\log^{\frac{3}{2}}(n/\delta))n^{\frac{2\alpha}{\alpha q-2\alpha-q}}. \tag{2.7}$$

Remark 9 *According to Theorem 8, several parameters affect the generalization behavior of SGD. In particular, we need to early stop SGD by considering two cases. Here, we give an intuitive interpretation. As we will show in Lemma 19, we need to control* $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)$ *in the estimation of* $\|\mathbf{w}_t\|_2$. *Our idea is to relate* $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)$ *by* $T_A = \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(h_\rho)$ *and* $T_B = \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}_k}) - \mathcal{E}(h_\rho) + \mathcal{E}(\hat{h}_{\mathbf{w}_k})$ *(Eq. (6.4)). To apply Bernstein inequality to control* T_A, *we need to control* $\ell(h_{\mathbf{w}_\lambda}(x), y)$, *which is bounded by* $O(\|\mathbf{w}_\lambda\|_2^q) = O(\lambda^{\frac{q(\alpha-1)}{2}})$ *due to Assumption 2 and Assumption* (2.5) *(we actually only use Assumption* (2.5) *to control* $\ell(h_{\mathbf{w}_\lambda}(x), y)$*). We apply concentration inequalities in empirical process to control* T_B *since* $\mathbf{w}_k$ *is a random variable, which depends on the capacity of the hypothesis space as reflected by the parameter* ζ. *It is clear that* T_A *and* T_B *increase with increasing* q *and* ζ, *respectively. We distinguish the two cases in Theorem 8 by considering whether* T_A *or* T_B *is dominant. If* ζ *is large then the term* T_B *would be dominant. This corresponds to the case (a) in Theorem 8 and this explains why the parameter* q *does not affect the rate in* (2.6). *If* q *is large (as compared to* ζ*), then the dominant term between* T_A *and* T_B *is* T_A. *This corresponds to the case (b) and this explains why the parameter* ζ *does not affect the generalization bound in* (2.7). *Notice that the capacity-independent analysis corresponds to the case* $\zeta = 2$, *and in this case* T_A *is dominated by* T_B, *and therefore the associated learning rates do not depend on* q. *Furthermore, the generalization performance always improves if* α *and* β *increase.*

We now present some specific learning rates which follow directly from Theorem 8 with specific instantiations of either q or β. We omit the proof for brevity. If we consider Lipschitz continuous loss functions (e.g., hinge loss and ϵ-insensitive loss), we can derive the following learning rates.

Corollary 10 *Let Assumptions 1-5 hold and* $q = 1$. *Assume* $\delta \in (0,1)$ *and* $\eta_1 \le 1/A$. *Let* $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ *be the sequence produced by* (2.1) *with* $\eta_t = \eta_1 t^{-\theta}, \theta > 1/2$.

(a) If $3\alpha + \alpha\beta\zeta + \zeta \ge \alpha\zeta + 2\alpha\beta + 1$, *then with* $T \asymp n^{\frac{2}{(\zeta+4\alpha+\alpha\beta\zeta-\alpha\zeta-2\alpha\beta)(1-\theta)}}$ *we derive the following inequality with probability at least* $1-\delta$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(\log^{\frac{3}{2}}(n/\delta))n^{\frac{2\alpha}{(\alpha-1)\zeta-4\alpha+2\alpha\beta-\alpha\beta\zeta}}.$$

(b) If $3\alpha + \alpha\beta\zeta + \zeta < \alpha\zeta + 2\alpha\beta + 1$, *then with* $T \asymp n^{\frac{2(3\alpha-2\alpha\beta+\alpha\beta\zeta-\alpha\zeta-4+2\beta-\beta\zeta+\zeta-1)}{(4-2\beta+\beta\zeta)(\alpha q-2\alpha-q)(1-\theta)}}$ *we derive the following inequality with probability at least* $1-\delta$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(\log^{\frac{3}{2}}(n/\delta))n^{-\frac{2\alpha}{\alpha+1}}.$$

Remark 11 *The best learning rate for Tikhonov regularization with Lipschitz loss functions is* $O(\max\{n^{\frac{2\alpha}{(\alpha-1)\zeta-4\alpha+2\alpha\beta-\alpha\beta\zeta}}, n^{-\frac{2\alpha}{\alpha+1}}\})$, *(see, e.g., Steinwart and Christmann, 2008,*

Chapter 7), which, however, does not take into account optimization errors. Corollary 10 shows that this best learning rate can be achieved by SGD (up to a logarithmic factor) where the trade-off among optimization errors, approximation errors and estimation errors is considered.

If Assumption 5 holds with $\beta = 1$, then we derive the following learning rates.

Corollary 12 *Let Assumptions 1-5 hold and $\beta = 1$. Assume $\delta \in (0,1)$ and $\eta_1 \leq 1/A$. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be the sequence produced by (2.1) with $\eta_t = \eta_1 t^{-\theta}, \theta > 1/2$.*

(a) If $\zeta + \alpha q \geq q$, then with $T \asymp n^{\frac{2}{(2\alpha+\zeta)(1-\theta)}}$ we derive the following inequality with probability at least $1-\delta$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(\log^{\frac{3}{2}}(n/\delta))n^{-\frac{2\alpha}{2\alpha+\zeta}}.$$

(b) If $\zeta + \alpha q < q$, then with $T \asymp n^{\frac{2(2+q-\alpha q)}{(2+\zeta)(2\alpha+q-\alpha q)(1-\theta)}}$ we derive the following inequality with probability at least $1-\delta$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(\log^{\frac{3}{2}}(n/\delta))n^{\frac{2\alpha}{\alpha q-2\alpha-q}}.$$

Remark 13 *Fast learning rates of the order $O(n^{-\frac{2\alpha}{2\alpha+\zeta}})$ were established for regularized risk minimization with self-concordant loss functions (Marteau-Ferey et al., 2019), which include generalized linear models. Their discussion makes a source assumption and a capacity assumption in terms of Hessian at optimum, which are different from Assumption 2 and Assumption 4, respectively. Furthermore, the variance-expectation assumption is removed in their discussion due to the concordance of loss functions. A nice property of their analysis is that it applies also to the case $\alpha > 1$, and therefore does not suffer from the saturation effect. As a comparison, our analysis applies only to $\alpha \in (0,1]$. Our analysis differs from theirs by considering different assumptions and a different algorithm where the optimization errors need to be addressed.*

As a direct application of Corollary 12 with $q = 2$, we can derive the optimal learning rates for SGD applied to learning problems with the least squares loss function (Lin and Rosasco, 2017).

Corollary 14 (Least squares) *Let Assumptions 2, 4 hold and the loss function be the least squares. Assume $\delta \in (0,1)$ and $\eta_1 \leq 1/A$. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be the sequence produced by (2.1) with $\eta_t = \eta_1 t^{-\theta}, \theta > 1/2$.*

(a) If $\alpha + \zeta/2 \geq 1$, then with $T \asymp n^{\frac{2}{(2\alpha+\zeta)(1-\theta)}}$, we derive the inequality $\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(\log^{\frac{3}{2}}(n/\delta))n^{-\frac{2\alpha}{2\alpha+\zeta}}$ with probability at least $1-\delta$.

(b) If $\alpha + \zeta/2 < 1$, then with $T \asymp n^{\frac{4-2\alpha}{(\zeta+2)(1-\theta)}}$, we derive the inequality $\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(\log^{\frac{3}{2}}(n/\delta))n^{-\alpha}$ with probability at least $1-\delta$.

Remark 15 *Optimal learning rates for SGD have already been derived for the specific least squares loss. In more details, learning rates of the order $O(n^{-\frac{2\alpha}{2\alpha+\zeta}})$ and $O(n^{\epsilon-\alpha})$ for an arbitrary small $\epsilon > 0$ were derived for the case $\alpha+\zeta/2 \geq 1$ and $\alpha+\zeta/2 < 1$ (up to logarithmic factors), respectively (Lin and Rosasco, 2017). Our learning rates are consistent with theirs.*

3. Discussions

In this section, we compare our learning rates with state-of-the-art results. We first compare our results to capacity-independent and capacity-dependent analysis of SGD, for each of which we consider both one-pass SGD and multi-pass SGD. We also give discussions with related work on stepsize, averaging and gradient descent (GD).

3.1 Related Work on Capacity-independent Analysis

One-pass SGD. One-pass SGD has been well studied in the literature (Ying and Pontil, 2008; Nemirovski et al., 2009; Rakhlin et al., 2012; Shamir and Zhang, 2013; Tarres and Yao, 2014; Orabona, 2014; Ying and Zhou, 2006; Bach and Moulines, 2013; Orabona, 2019). For the specific least squares loss, learning rates of the order $O(n^{-\frac{\alpha}{\alpha+1}}\log n)$ were derived for unregularized SGD (Ying and Pontil, 2008) and regularized SGD (Tarres and Yao, 2014). For general loss functions, learning rates of the order $O(n^{\epsilon-\frac{\alpha}{2(\alpha+1)}}\log n)$ for an arbitrary small $\epsilon > 0$ were derived in Ying and Zhou (2006), which were improved to $O(n^{-\frac{\alpha}{\alpha+1}}\log n)$ in Lin and Zhou (2018). If one does not consider approximation errors ($\alpha = 1$), the minimax optimal learning rates of the order $O(1/\sqrt{n})$ were well-known for one-pass SGD with averaging (Zhang, 2004; Nemirovski et al., 2009; Bach and Moulines, 2011). Some later work showed that one can go pass the rate $O(1/\sqrt{n})$ for an averaged SGD with smooth self-concordant losses, e.g., the least square and logistic loss (Bach and Moulines, 2013). More recently, an averaged SGD with a constant step size was shown to be able to adapt the local strong convexity of the objective function with Lipschitz self-concordant losses, leading to convergence rates of the order $O(1/(n\mu))$ with μ being the lowest eigenvalue of the Hessian at the global optimum (Bach, 2014). A nice property is that the implementation does not require the information of μ. This result applies to logistic loss and generalized linear models.

Multi-pass SGD. As compared to the one-pass SGD, the generalization performance of multi-pass SGD was much less studied. The landmark work in Bousquet and Bottou (2008) developed a framework to analyze the generalization performance of multi-pass stochastic learning algorithms by taking into account the computational complexity of learning algorithms. Under this framework, the interplay among estimation errors, optimization errors and approximation errors can be studied, showing that an implicit regularization can be achieved in the absence of penalization or constraints by tuning either the stepsize or the number of passes (the number of iterations divided by the sample size) (Rosasco and Villa, 2015; Lin and Rosasco, 2017; Lin et al., 2016a; Hardt et al., 2016).

Multi-pass SGD with a fixed ordering at each pass was investigated for the specific least squares loss, where learning rates $O(n^{-\frac{\alpha}{\alpha+2}})$ were established for a constant stepsize (Rosasco and Villa, 2015). In a parametric setting, it was shown that SGD is algorithmically stable (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010) and the stability measure of SGD with T iterations scales as $O(n^{-1}\sum_{t=1}^{T}\eta_t)$ (Hardt et al., 2016), based on which a generalization bound $\mathbb{E}[\mathcal{E}(\bar{\mathbf{w}}_T)] - \inf_{w\in\mathcal{W}}\mathcal{E}(\mathbf{w}) = O(n^{-\frac{1}{2}})$ was established for $\eta_t = O(1/\sqrt{n})$ and $T \asymp n$ without considering approximation errors. The discussion in Hardt et al. (2016) requires to impose a smoothness assumption on loss functions. Generalization analysis was considered separately for smooth and non-smooth loss functions (Lin

et al., 2016a). For smooth loss functions, it was shown $\mathbb{E}[\mathcal{E}(\bar{\mathbf{w}}_T)] - \mathcal{E}(h_\rho) = O(n^{-\frac{\alpha}{1+\alpha}} \log n)$ for $\eta_t = \eta_1/\sqrt{t}$ with $T \asymp n^{\frac{2}{\alpha+1}}$ (Lin et al., 2016a), based on the stability property of SGD established in Hardt et al. (2016). For non-smooth loss functions, it was shown $\mathbb{E}[\mathcal{E}(\bar{\mathbf{w}}_T)] - \mathcal{E}(h_\rho) = O(n^{-\frac{\alpha}{2\alpha+1}} \log n)$ for $\eta_t = \eta_1/\sqrt{t}$ and $T \asymp n^{\frac{2}{2\alpha+1}}$ (Lin et al., 2016a), by controlling estimation errors with Rademacher complexities (Bartlett and Mendelson, 2002). Still, the bounds in Lin et al. (2016a); Hardt et al. (2016) require to impose a boundedness assumption on subgradients and are stated in expectation. Feldman and Vondrak (2019) established a framework to get nearly optimal learning rates for uniformly stable algorithm, and then applied it to projected SGD for Lipschitz and smooth loss functions. For projected SGD with constant stepsize $\eta_t = 1/\sqrt{T}$ and $T \asymp n$, they derived almost optimal learning rates $O(\log(n)\log^2(n/\delta)/\sqrt{n})$. The projected SGD requires a projection of SGD iterates onto a compact domain at each iteration. As a comparison, we consider projection-free algorithm and our purpose is to show that the implicit regularization can be achieved without either a penalty or a constraint. Furthermore, our learning rate analysis extends the analysis in Hardt et al. (2016) to non-smooth loss functions and substantially improve the bound $O(n^{-\frac{\alpha}{2\alpha+1}} \log n)$ (Lin et al., 2016a) in this setting. The generalization error bound $O\big(n^{-\frac{\alpha}{1+\alpha}} \log^{\frac{3}{2}} \frac{n}{\delta}\big)$ in Theorem 4 is optimal in the sense that it matches the best available bound for Tikhonov regularization (up to a logarithmic factor) (Cucker and Zhou, 2007; Steinwart and Christmann, 2008; Lin et al., 2016a). Our learning rates are stated with high probability, which are beneficial to understand the variety of the learned model as compared to the results in expectation (Hardt et al., 2016; Lin et al., 2016a). It should be emphasized that our discussion requires no bounded subgradient assumption imposed in the literature (Hardt et al., 2016; Lin et al., 2016a; Feldman and Vondrak, 2019).

Generalization bounds for multi-pass SGD were also studied from a PAC-Bayesian perspective (London, 2017). However, the high-probability bounds there require to impose Lipschitz continuity, smoothness and strong convexity assumptions on loss functions, and ignore optimization and approximation errors (London, 2017).

3.2 Related Work on Capacity-dependent Analysis

One-pass SGD. Better learning rates are possible if a capacity assumption is imposed (Cucker and Zhou, 2007; Steinwart and Christmann, 2008). For the one-pass SGD with the specific least squares loss, the learning rates $O(n^{-\frac{2\alpha}{2\alpha+\zeta}})$ and $O(n^{-\alpha})$ have been derived for the case $\alpha + \zeta/2 \geq 1$ and $\alpha + \zeta/2 < 1$, respectively, under a source condition different from Assumption 2 (Dieuleveut and Bach, 2016). Their results suffer from a saturation effect, i.e., the learning rate does not improve after α reaching a critical point. It was recently observed that this saturation effect is due to the use of uniform averaging (Mücke et al., 2019), based on which the authors proposed a tail-averaging to successfully address the saturation effect.

Multi-pass SGD. Optimal learning rates for multi-pass SGD with the least squares loss were recently developed (Lin and Rosasco, 2017). Specifically, the learning rates $O(n^{-\frac{2\alpha}{2\alpha+\zeta}})$ were derived in the case $\alpha+\zeta/2 \geq 1$ with $T \asymp n^{\frac{2}{2\alpha+\zeta}+1}$, while the learning rates $O(n^{\epsilon-\alpha})$ with an arbitrarily small $\epsilon > 0$ were derived in the case $\alpha+\zeta/2 < 1$ with $T \asymp n^{2-\epsilon}$. The results there hold with high probability w.r.t. the random training examples and in expectation w.r.t.

the random indices $j_1, \ldots, j_T$. As a comparison, Corollary 14 slightly extends the results in Lin and Rosasco (2017) from $O(n^{\epsilon-\alpha})$ to $O(n^{-\alpha}\log^{\frac{3}{2}}(n))$ in the case $\alpha + \zeta/2 < 1$, and is stated with high probability w.r.t. both the random training examples and the random indices. The results in Lin and Rosasco (2017) also cover the strategy with mini-batches and an attainable case, which were then extended to multi-pass SGD with the least squares loss in a distributed learning setting (Lin and Cevher, 2018). A refined analysis for the case $\alpha + \zeta/2 < 1$ was further conducted by imposing an additional assumption on the regularity of objective functions in the hypothesis space w.r.t. the ℓ_∞-norm (Pillaud-Vivien et al., 2018b). It should be mentioned that, different from Assumption 4, a related capacity assumption measured by the associated integral operator was imposed in these analysis (Lin and Rosasco, 2017; Lin and Cevher, 2018; Dieuleveut and Bach, 2016; Pillaud-Vivien et al., 2018b).

To our best knowledge, the existing capacity-dependent learning rates of multi-pass SGD are all stated for the specific least squares loss function. In this paper, we establish the first capacity-dependent learning rates of multi-pass SGD for a general convex loss function. With specific instantiations, we immediately derive learning rates matching the best one developed for the Tikhonov regularization with Lipschitz loss functions (Steinwart and Christmann, 2008), and the minimax optimal one for the Tikhonov regularization with the least squares loss function (Caponnetto and De Vito, 2007). It should be mentioned that the optimization errors are ignored in the analysis for Tikhonov regularization. In this sense, our analysis sheds new insights on how optimization errors, approximation errors and estimation errors should be balanced to achieve an optimal generalization performance for learning with general convex loss functions.

The optimal learning rates for general convex loss functions are achieved here by establishing a unifying norm estimate of $\mathbf{w}_t$ applicable to both capacity-independent and capacity-dependent case. We can apply different concentration inequalities (Boucheron et al., 2013) in these two cases to control the associated random variables, which in turn lead to explicit norm estimates of SGD iterates able to imply satisfactory learning rates.

3.3 Discussions on Stepsize and Averaging

In this paper, we consider averaged SGD with a decaying stepsize. In the literature, several variants of SGD with different averaging schemes and stepsizes were shown to achieve optimal generalization bounds with less computation (Bach and Moulines, 2013; Dieuleveut and Bach, 2016; Pillaud-Vivien et al., 2018b; Mücke et al., 2019). For example, it was shown that the uniform averaging allows for a large constant stepsize (Dieuleveut and Bach, 2016). Recently, it was shown that tail-averaging and minibatch can further overcome the saturation effect (Mücke et al., 2019). The consideration of a large constant stepsize significantly reduces the number of passes required for getting optimal learning rates. In particular, the optimal bounds in Dieuleveut and Bach (2016); Mücke et al. (2019) were achieved for one-pass SGD. As a comparison, we require multi-pass SGD to get optimal bounds due to the consideration of a decaying stepsize. The reason to consider a decaying stepsize is that we study general convex loss functions, and use a uniform convergence approach instead of the integral operator approach (Dieuleveut and Bach, 2016; Mücke et al., 2019). In our approach, an essential ingredient is to estimate the norm of SGD

iterates. A large constant stepsize makes these norm challenging to control. Actually our estimation error analysis applies to any single SGD iterate and fails to use the advantage of averaging schemes in allowing for a large stepsize. It would be very interesting to investigate how to exploit the advantage of averaging in the uniform convergence approach, and how to achieve optimal learning rates for one-pass SGD with a large constant stepsize for general convex loss functions. The reason to consider averaged SGD is to simplify the optimization error analysis. Optimization errors for the last iterate of SGD were studied in Shamir and Zhang (2013); Harvey et al. (2019). However, the results in Shamir and Zhang (2013) are stated in expectation, while the high-probability analysis in Harvey et al. (2019) imposes assumptions on the boundedness of gradients and is performed for projected SGD. As a comparison, our aim is to show the implicit regularization effect of SGD without any explicit constraint and we also want to remove the bounded gradient assumptions. How to remove these assumptions and get optimal learning rates for the last SGD iterate remains an interesting problem for future study.

3.4 Related Work on Gradient Descent

In this subsection, we compare our results with related work on GD.

We first consider GD with general convex loss functions. The iterative regularization effect of GD was studied in Lin et al. (2016b) for general convex loss functions, where the iterates are updated by $\mathbf{w}_{t+1}^G = \mathbf{w}_t^G - \eta_t \mathcal{E}_{\mathbf{z}}'(\mathbf{w}_t^G)$. If $q = 1$, it was shown that GD with $\eta_t = \eta_1/\sqrt{t}$ and $T \asymp n^{\frac{4}{(2\alpha+1)(2-\beta+\beta\zeta/2)}}$ satisfies $\mathcal{E}(\mathbf{w}_T^G) - \mathcal{E}(h_\rho) = \widetilde{O}\big(n^{-\frac{2\alpha}{(2\alpha+1)(2-\beta+\beta\zeta/2)}}\big)$. It can be checked that $(\alpha-1)\zeta \geq \beta-2-\beta\zeta/2$ and $\alpha+1 \leq (2\alpha+1)(2-\beta+\beta\zeta/2)$, and therefore the learning rates in Lin et al. (2016b) are worse than those in Corollary 10. In particular, if $\alpha = 1$, the discussion in Lin et al. (2016b) implies generalization bounds $\widetilde{O}(n^{\frac{2}{3(\beta-2-\beta\zeta/2)}})$, while Corollary 10 implies better learning rates $\widetilde{O}(n^{\frac{2}{2\beta-4-\beta\zeta}})$. If the loss function is further smooth, then it was shown that GD with a constant stepsize and $T \asymp n^{\frac{2}{(1+2\alpha)(2-\beta+\beta\zeta/2)}}$ is able to show the learning rate $O(n^{-\frac{2\alpha}{(2\alpha+1)(2-\beta+\beta\zeta/2)}})$ (Lin et al., 2016b), which is again slower than our learning rate in Corollary 10.

We now consider GD with the least square loss. Capacity-independent learning rates of the order $O(n^{-\frac{\alpha}{\alpha+1}})$ were developed in Bauer et al. (2007). Optimal capacity-dependent learning rates $O(n^{-\frac{2\alpha}{2\alpha+\zeta}})$ were established in Caponnetto and Yao (2010) by introducing extra unlabeled data. If $\alpha + \zeta/2 \geq 1$, it was shown that GD with a constant stepsize and $T \asymp n^{\frac{1}{\alpha+\zeta/2}}$ achieves the learning rate $\widetilde{O}(n^{-\frac{2\alpha}{2\alpha+\zeta}})$ (Lin and Rosasco, 2017). If $\alpha + \zeta/2 < 1$, then GD with $T \asymp n^{1-\epsilon}$ implies the learning rate $\widetilde{O}(n^{-\alpha(1-\epsilon)})$ (Lin and Rosasco, 2017). It is clear that these learning rates match Corollary 14. If we choose $\theta = 1/2$ in Corollary 14 (our results hold for θ arbitrarily close to $1/2$), then Corollary 14 requires $T \asymp n^{\frac{4}{2\alpha+\zeta}}$ for the case $\alpha + \zeta/2 \geq 1$ and $T \asymp n^{\frac{8-4\alpha}{\zeta+2}}$ for the case $\alpha + \zeta/2 < 1$. If $\alpha + \zeta/2 \geq 1$, then it is clear that $4/(2\alpha+\zeta) \leq 2/(2\alpha+\zeta) + 1$ (an iterate of GD corresponds to n iterations of SGD) and therefore our analysis shows that SGD can achieve optimal learning rates with less computation as compared with the results in Lin and Rosasco (2017). If $\alpha + \zeta/2 < 1$, then our analysis requires more computation but achieves a slightly better learning rate by removing ϵ. It should be mentioned that the discussions in Lin and Rosasco

(2017) considered a source assumption $h_\rho \in L_K^{\alpha/2}(L^\alpha_{\rho_X})$ different from Assumption 2. A nice property of these results is that their analysis also applies to the case $\alpha > 1$ (Lin and Rosasco, 2017), and the generalization improves if the regression function has more regularity. As a comparison, our rates only apply to the case $\alpha \in (0, 1]$.

4. Complexity Control and Basic Idea

4.1 Basic Complexity Control

An essential component to control estimation errors and optimization errors is to control the complexity of SGD iterates measured by the RKHS norm. In this subsection, we first give a lemma to show $\|\mathbf{w}_t\|_2^2$ can be bounded by $O(1)\sum_{k=1}^t \eta_k$ (Lemma 17), which is further refined by showing $\|\mathbf{w}_t\|_2^2 = O(1)\sum_{k=1}^t \eta_k \max\{0, \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\}$ (Lemma 19). Furthermore, Lemma 17 gives a bound on $\sum_{k=1}^t \eta_k^2 f(\mathbf{w}_k, z_{j_k})$, which together with the self-bounding property of loss functions, gives a bound on $\sum_{k=1}^t \eta_k^2 \|f'(\mathbf{w}_k, z_{j_k})\|_2^2$. The existing studies require a bounded gradient assumption to show $\sum_{k=1}^t \eta_k^2 \|f'(\mathbf{w}_k, z_{j_k})\|_2^2 = O\big(\sum_{k=1}^t \eta_k^2\big)$ (Hardt et al., 2016; Lin et al., 2016a), which is removed here. Before introducing Lemma 17, we give a simple lemma to be proved in Section B relating properties of loss functions ℓ to properties of f. Without loss of generality, we always assume $\tilde{c}_q \geq 1$.

Lemma 16 *(a) If the loss function ℓ satisfies* (2.2)*, then* (2.3) *holds.*

(b) If the loss function ℓ satisfies $\ell(a, y) \leq c_q(|a|^q + 1)$ for all a and $y \in \mathcal{Y}$, then for any $\mathbf{w} \in \mathcal{W}$ we have $f(\mathbf{w}, z) \leq \tilde{c}_q\big(\|\mathbf{w}\|_2^q + 1\big)$, where $\tilde{c}_q = c_q \max\{\kappa^q, 1\}$.

Lemma 17 *Let Assumption 1 hold and $C_1 = 2\sup_y \ell(0, y) + A^{-1}B$. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be the sequence produced by* (2.1) *with $\eta_t \leq 1/A$. Then*

$$\|\mathbf{w}_{t+1}\|_2^2 \leq C_1 \sum_{k=1}^t \eta_k \tag{4.1}$$

and

$$\sum_{k=1}^t \eta_k^2 f(\mathbf{w}_k, z_{j_k}) \leq C_1 \sum_{k=1}^t \eta_k^2. \tag{4.2}$$

Proof It follows from (2.1), Part (a) of Lemma 16 and the convexity of f that

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &= \big\|\mathbf{w}_t - \eta_t f'(\mathbf{w}_t, z_{j_t}) - \mathbf{w}\big\|_2^2 \\
&= \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 \|f'(\mathbf{w}_t, z_{j_t})\|_2^2 + 2\big\langle \mathbf{w} - \mathbf{w}_t, \eta_t f'(\mathbf{w}_t, z_{j_t})\big\rangle \\
&\leq \|\mathbf{w} - \mathbf{w}_t\|_2^2 + \eta_t^2\big(Af(\mathbf{w}_t, z_{j_t}) + B\big) + 2\big\langle \mathbf{w} - \mathbf{w}_t, \eta_t f'(\mathbf{w}_t, z_{j_t})\big\rangle \\
&\leq \|\mathbf{w} - \mathbf{w}_t\|_2^2 + \eta_t^2\big(Af(\mathbf{w}_t, z_{j_t}) + B\big) + 2\eta_t\big(f(\mathbf{w}, z_{j_t}) - f(\mathbf{w}_t, z_{j_t})\big).
\end{aligned} \tag{4.3}$$

If we take $\mathbf{w} = 0$ in the above inequality, then

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|_2^2 &\leq \|\mathbf{w}_t\|_2^2 + \eta_t^2\big(Af(\mathbf{w}_t, z_{j_t}) + B\big) + 2\eta_t\big(\ell(0, y_{j_t}) - f(\mathbf{w}_t, z_{j_t})\big) \\
&= \|\mathbf{w}_t\|_2^2 + \eta_t f(\mathbf{w}_t, z_{j_t})(A\eta_t - 2) + \eta_t\big(2\ell(0, y_{j_t}) + \eta_t B\big) \\
&\leq \|\mathbf{w}_t\|_2^2 - \eta_t f(\mathbf{w}_t, z_{j_t}) + \eta_t\big(2\ell(0, y_{j_t}) + A^{-1}B\big),
\end{aligned} \tag{4.4}$$

where we have used $\eta_t \leq 1/A$. Taking a summation of the above inequality then shows

$$\|\mathbf{w}_{t+1}\|_2^2 = \|\mathbf{w}_1\|_2^2 + \sum_{k=1}^{t} \big[\|\mathbf{w}_{k+1}\|_2^2 - \|\mathbf{w}_k\|_2^2\big] \leq C_1 \sum_{k=1}^{t} \eta_k,$$

This establishes (4.1). We now prove (4.2). Multiplying both sides of (4.4) by η_t followed with a reformulation gives

$$\begin{aligned}\eta_t^2 f(\mathbf{w}_t, z_{j_t}) &\leq \eta_t \|\mathbf{w}_t\|_2^2 - \eta_t \|\mathbf{w}_{t+1}\|_2^2 + \eta_t^2 C_1 \\ &\leq \eta_t \|\mathbf{w}_t\|_2^2 - \eta_{t+1} \|\mathbf{w}_{t+1}\|_2^2 + C_1 \eta_t^2,\end{aligned}$$

where we have used $\eta_{t+1} \leq \eta_t$. A summation of the above inequality then implies

$$\sum_{k=1}^{t} \eta_k^2 f(\mathbf{w}_k, z_{j_k}) \leq \sum_{k=1}^{t} \big(\eta_k \|\mathbf{w}_k\|_2^2 - \eta_{k+1} \|\mathbf{w}_{k+1}\|_2^2\big) + C_1 \sum_{k=1}^{t} \eta_k^2 \leq C_1 \sum_{k=1}^{t} \eta_k^2.$$

The proof is complete. ■

To prove Lemma 19, we first introduce a lemma to relate $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda)$ to $\mathcal{E}(\mathbf{w}_\lambda)$ with $\mathbf{w}_\lambda$ defined in (2.4). The proof is given in Section B.

Lemma 18 *Let $\delta \in (0,1), \rho \in (0,1]$ and $\mathbf{w}_\lambda$ be defined in (2.4). Then, the following inequality holds with probability at least $1-\delta$*

$$\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda) \leq \rho \mathcal{E}(0) + \tilde{c}_q (\rho n)^{-1} \big(\|\mathbf{w}_\lambda\|_2^q + 1\big) \log(1/\delta). \tag{4.5}$$

Lemma 19 *Let Assumption 1 hold. Assume $n^{-1}\|\mathbf{w}_\lambda\|_2^q \leq c_\alpha^{\frac{q}{2}}$ and $\delta \in (0, 2/e)$ (e is the base of the natural logarithm). If $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ is the sequence produced by (2.1) with $\eta_t \leq 1/A, \eta_{t+1} \leq \eta_t$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$, then with probability at least $1-\delta$ we have the following inequality*

$$\|\mathbf{w}_{t+1}\|_2^2 \leq 8 \sum_{k=1}^{t} \eta_k \max\{0, \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\} + C_2 \log \frac{2T}{\delta} (\|\mathbf{w}_\lambda\|_2^2 + 1), \quad \forall\, t = 1, \ldots, T, \tag{4.6}$$

where C_2 is a constant independent of T and n (explicitly given in the proof).

Proof Taking $\mathbf{w} = \mathbf{w}_\lambda$ in (4.3), we know

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_\lambda\|_2^2 \leq{}& \|\mathbf{w}_t - \mathbf{w}_\lambda\|_2^2 + \eta_t^2 \big(A f(\mathbf{w}_t, z_{j_t}) + B\big) + 2\eta_t \big\langle \mathbf{w}_\lambda - \mathbf{w}_t, f'(\mathbf{w}_t, z_{j_t}) \\ &- \mathbb{E}_{j_t}\big[f'(\mathbf{w}_t, z_{j_t})\big]\big\rangle + 2\eta_t \big\langle \mathbf{w}_\lambda - \mathbf{w}_t, \mathbb{E}_{j_t}\big[f'(\mathbf{w}_t, z_{j_t})\big]\big\rangle \\ \leq{}& \|\mathbf{w}_t - \mathbf{w}_\lambda\|_2^2 + \eta_t^2 \big(A f(\mathbf{w}_t, z_{j_t}) + B\big) + 2\eta_t \big\langle \mathbf{w}_\lambda - \mathbf{w}_t, f'(\mathbf{w}_t, z_{j_t}) \\ &- \mathbb{E}_{j_t}\big[f'(\mathbf{w}_t, z_{j_t})\big]\big\rangle + 2\eta_t \big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_t)\big),\end{aligned}$$

where the last inequality is due to the convexity of f. Taking a summation of the above inequality shows

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\lambda\|_2^2 \leq \|\mathbf{w}_1 - \mathbf{w}_\lambda\|_2^2 + (AC_1 + B) \sum_{k=1}^{t} \eta_k^2 + 2 \sum_{k=1}^{t} \xi_k + 2 \sum_{k=1}^{t} \eta_k \big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\big), \tag{4.7}$$

where we have used Lemma 17 and introduced

$$\xi_k = \eta_k \big\langle \mathbf{w}_\lambda - \mathbf{w}_k, f'(\mathbf{w}_k, z_{j_k}) - \mathbb{E}_{j_k}\big[f'(\mathbf{w}_k, z_{j_k})\big]\big\rangle. \tag{4.8}$$

It is clear that $\mathbb{E}_{j_k}[\xi_k] = 0$ and therefore $\{\xi_k\}_{k\in\mathbb{N}}$ is a martingale difference sequence. The variance of ξ_k can be bounded by

$$\begin{aligned}
\mathbb{E}_{j_k}[(\xi_k - \mathbb{E}_{j_k}[\xi_k])^2] = \mathbb{E}_{j_k}[\xi_k^2] &\le \eta_k^2 \mathbb{E}_{j_k}\Big[\big\langle \mathbf{w}_\lambda - \mathbf{w}_k, f'(\mathbf{w}_k, z_{j_k})\big\rangle^2\Big] \\
&\le \eta_k^2 \|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2 \mathbb{E}_{j_k}\big[\big\|f'(\mathbf{w}_k, z_{j_k})\big\|_2^2\big] \\
&\le \eta_k^2 \|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2 \mathbb{E}_{j_k}\big[Af(\mathbf{w}_k, z_{j_k}) + B\big] \\
&= \eta_k^2 \|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2 \big(A\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) + B\big),
\end{aligned}$$

where in the first step we have used $\mathbb{E}[(\xi - \mathbb{E}[\xi])^2] \le \mathbb{E}[\xi^2]$ for any real-valued random variable ξ and in the last second step we have used Part (a) of Lemma 16. A summation of the above inequality then shows

$$\begin{aligned}
&\sum_{k=1}^t \mathbb{E}_{j_k}\big[\big(\xi_k - \mathbb{E}_{j_k}[\xi_k]\big)^2\big] \\
&\le A\sum_{k=1}^t \eta_k^2 \|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2 \max\{\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda), 0\} + \big(A\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) + B\big)\sum_{k=1}^t \eta_k^2 \|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2.
\end{aligned}$$

By the elementary inequality

$$(a+b)^2 \le 2(a^2+b^2), \quad \forall a, b \in \mathbb{R}, \tag{4.9}$$

we know

$$\eta_k^2 \|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2 \le 2\eta_k^2\big(\|\mathbf{w}_\lambda\|_2^2 + \|\mathbf{w}_k\|_2^2\big) \le 2\eta_k\big(\|\mathbf{w}_\lambda\|_2^2\eta_k + \eta_k C_1 \sum_{j=1}^k \eta_j\big) \le 2\eta_k\big(\|\mathbf{w}_\lambda\|_2^2\eta_k + C_1C_3\big),$$

where we have used Lemma 17 and introduced $C_3 := \sup_k \eta_k \sum_{j=1}^k \eta_j \le \sum_{j=1}^k \eta_j^2 < \infty$. It then follows that

$$\begin{aligned}
\sum_{k=1}^t \mathbb{E}_{j_k}\big[\big(\xi_k - \mathbb{E}_{j_k}[\xi_k]\big)^2\big] \le 2A\sum_{k=1}^t \eta_k\big(\|\mathbf{w}_\lambda\|_2^2\eta_k + C_1C_3\big) \max\{\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda), 0\} \\
+ \big(A\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) + B\big)\sum_{k=1}^t \eta_k^2 \|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2.
\end{aligned}$$

By (4.9), we also have

$$\begin{aligned}
|\xi_k| &\le \frac{\eta_k}{2}\Big[\|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2 + \|f'(\mathbf{w}_k, z_{j_k}) - \mathbb{E}_{j_k}[f'(\mathbf{w}_k, z_{j_k})]\|_2^2\Big] \\
&\le \eta_k\Big[\|\mathbf{w}_\lambda\|_2^2 + \|\mathbf{w}_k\|_2^2 + \|f'(\mathbf{w}_k, z_{j_k})\|_2^2 + \|\mathbb{E}_{j_k}[f'(\mathbf{w}_k, z_{j_k})]\|_2^2\Big].
\end{aligned} \tag{4.10}$$

It follows from Lemma 16 that

$$\|f'(\mathbf{w}_k, z_{j_k})\|_2^2 \leq Af(\mathbf{w}_k, z_{j_k}) + B \leq A\tilde{c}_q\big(\|\mathbf{w}_k\|_2^q + 1\big) + B \leq A\tilde{c}_q\|\mathbf{w}_k\|_2^2 + 2A\tilde{c}_q + B, \tag{4.11}$$

where we have used the inequality $\|\mathbf{w}_k\|_2^q \leq \|\mathbf{w}_k\|_2^2 + 1$ for $q \in [1, 2]$. Combining the above two inequalities together, we derive the following inequality on the magnitude of $\xi_k - \mathbb{E}_{j_k}[\xi_k]$

$$\xi_k - \mathbb{E}_{j_k}[\xi_k] \leq \eta_1\big(\|\mathbf{w}_\lambda\|_2^2 + 4A\tilde{c}_q + 2B\big) + \big(2A\tilde{c}_q + 1\big)C_1\eta_k\sum_{j=1}^{k}\eta_j \leq C_4\big(\|\mathbf{w}_\lambda\|_2^2 + 1\big),$$

where we have used (4.1), the definition of C_3 and introduced $C_4 = \max\{\eta_1(4A\tilde{c}_q + 2B) + (2A\tilde{c}_q + 1)C_1C_3, \eta_1\}$.

Plugging the above two bounds on magnitudes and variances of random variables into Part (b) of Lemma A.1, we derive the following inequality with probability at least $1 - \delta/2$

$$\sum_{k=1}^{t}\xi_k \leq \frac{\rho}{C_4(\|\mathbf{w}_\lambda\|_2^2 + 1)}\bigg(2A\sum_{k=1}^{t}\eta_k\big(\|\mathbf{w}_\lambda\|_2^2\eta_k + C_1C_3\big)\max\{\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda), 0\} + \big(A\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) + B\big)\sum_{k=1}^{t}\eta_k^2\|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2\bigg) + \frac{C_4(\|\mathbf{w}_\lambda\|_2^2 + 1)\log\frac{2}{\delta}}{\rho},$$

where $\rho = \min\{1, \frac{C_4(\|\mathbf{w}_\lambda\|_2^2+1)}{2A(\eta_1\|\mathbf{w}_\lambda\|_2^2+C_1C_3)}\}$. For any t, combining the above inequality and (4.7) together, we derive the following inequality with probability at least $1 - \delta/2$

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\lambda\|_2^2 \leq \|\mathbf{w}_\lambda\|_2^2 + (AC_1 + B)\sum_{k=1}^{t}\eta_k^2 + 2\sum_{k=1}^{t}\eta_k\max\{\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda), 0\} \\
&+ \frac{A\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) + B}{AC_1C_3}\sum_{k=1}^{t}\eta_k^2\|\mathbf{w}_k - \mathbf{w}_\lambda\|_2^2 + \frac{2C_4(\|\mathbf{w}_\lambda\|_2^2 + 1)\log\frac{2}{\delta}}{\rho} + 2\sum_{k=1}^{t}\eta_k\big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\big) \\
&= \|\mathbf{w}_\lambda\|_2^2 + (AC_1 + B)\sum_{k=1}^{t}\eta_k^2 + 2\sum_{k=1}^{t}\eta_k\max\{0, \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\} + \\
&\frac{A\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) + B}{AC_1C_3}\Big(\sum_{k=1}^{t_1}\eta_k^2\|\mathbf{w}_k - \mathbf{w}_\lambda\|_2^2 + \sum_{k=t_1+1}^{t}\eta_k^2\|\mathbf{w}_k - \mathbf{w}_\lambda\|_2^2\Big) + \frac{2C_4(\|\mathbf{w}_\lambda\|_2^2 + 1)\log\frac{2}{\delta}}{\rho},
\end{aligned}$$

where t_1 is an integer to be fixed later and we have used $\rho \leq \frac{C_4(\|\mathbf{w}_\lambda\|_2^2+1)}{2A(\eta_1\|\mathbf{w}_\lambda\|_2^2+C_1C_3)}$ in the first step. An union bound on probabilities of events then gives the following inequality with probability at least $1 - \delta/2$ for all $t = 1, \ldots, T$

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\lambda\|_2^2 \leq \|\mathbf{w}_\lambda\|_2^2 + (AC_1 + B)\sum_{k=1}^{t}\eta_k^2 + 2\sum_{k=1}^{t}\eta_k\max\{0, \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\} \\
&+ \frac{A\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) + B}{AC_1C_3}\Big(\sum_{k=1}^{t_1}\eta_k^2\|\mathbf{w}_k - \mathbf{w}_\lambda\|_2^2 + \sup_{1\leq\tilde{k}\leq t}\|\mathbf{w}_{\tilde{k}} - \mathbf{w}_\lambda\|_2^2\sum_{k=t_1+1}^{t}\eta_k^2\Big) + \frac{C_4(\|\mathbf{w}_\lambda\|_2^2 + 1)\log\frac{2T}{\delta}}{\rho}.
\end{aligned}$$

On the other hand, Lemma 18 with $\rho = 1$ and $n^{-1}\|\mathbf{w}_\lambda\|_2^q \le c_\alpha^{\frac{q}{2}}$ implies the following inequality with probability at least $1-\delta/2$

$$\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda) \le \mathcal{E}(0) + \tilde{c}_q(c_\alpha^{\frac{q}{2}}+1)\log(2/\delta).$$

Combining the above two inequalities together with an union bound on probabilities of events then gives the following inequality with probability at least $1-\delta$ for all $t = 1, \ldots, T$ (note $\mathcal{E}(\mathbf{w}_\lambda) \le \mathcal{E}(0)$)

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\lambda\|_2^2 \le \|\mathbf{w}_\lambda\|_2^2 + (AC_1+B)\sum_{k=1}^t \eta_k^2 + 2\sum_{k=1}^t \eta_k \max\{0, \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\} +$$
$$\frac{A\mathcal{E}(0)+B+C_5\log\frac{2}{\delta}}{AC_1C_3}\Big(\sum_{k=1}^{t_1}\eta_k^2\|\mathbf{w}_k-\mathbf{w}_\lambda\|_2^2 + \sup_{1\le\tilde{k}\le t}\|\mathbf{w}_{\tilde{k}}-\mathbf{w}_\lambda\|_2^2\sum_{k=t_1+1}^{t}\eta_k^2\Big) + \frac{2C_4(\|\mathbf{w}_\lambda\|_2^2+1)\log\frac{2T}{\delta}}{\rho},$$

where $C_5 = A\big(\mathcal{E}(0)+\tilde{c}_q(c_\alpha^{\frac{q}{2}}+1)\big)$. The assumption $\sum_{t=1}^\infty \eta_t^2 < \infty$ guarantees the existence of t_1 such that $\sum_{k=t_1+1}^\infty \eta_k^2 < \frac{2^{-1}AC_1C_3}{A\mathcal{E}(0)+B+C_5\log\frac{2}{\delta}}$ and therefore we derive the following inequality with probability $1-\delta$ for all $t = 1, \ldots, T$ (the right-hand side is an increasing function of t)

$$\sup_{\tilde{t}=1,\ldots,t+1}\|\mathbf{w}_{\tilde{t}} - \mathbf{w}_\lambda\|_2^2 \le \|\mathbf{w}_\lambda\|_2^2 + (AC_1+B)\sum_{k=1}^t \eta_k^2 + 2\sum_{k=1}^t \eta_k \max\{0, \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\}$$
$$+\frac{A\mathcal{E}(0)+B+C_5\log\frac{2}{\delta}}{AC_1C_3}\sum_{k=1}^{t_1}\eta_k^2\|\mathbf{w}_k-\mathbf{w}_\lambda\|_2^2 + 2^{-1}\sup_{1\le\tilde{k}\le t+1}\|\mathbf{w}_{\tilde{k}}-\mathbf{w}_\lambda\|_2^2 + \frac{2C_4(\|\mathbf{w}_\lambda\|_2^2+1)\log\frac{2T}{\delta}}{\rho},$$

from which we derive the following inequality with probability at least $1-\delta$

$$\sup_{\tilde{t}=1,\ldots,t+1}\|\mathbf{w}_{\tilde{t}} - \mathbf{w}_\lambda\|_2^2 \le 2\|\mathbf{w}_\lambda\|_2^2 + 2(AC_1+B)\sum_{k=1}^t \eta_k^2 + 4\sum_{k=1}^t \eta_k \max\{0, \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\}$$
$$+\frac{4A\mathcal{E}(0)+4B+4C_5\log\frac{2}{\delta}}{AC_1C_3}\Big(C_1C_3\sum_{k=1}^{t_1}\eta_k + \|\mathbf{w}_\lambda\|_2^2\sum_{k=1}^{t_1}\eta_k^2\Big) + \frac{4C_4(\|\mathbf{w}_\lambda\|_2^2+1)\log\frac{2T}{\delta}}{\rho}.$$

Here we have used the following inequality due to (4.9), Lemma 17 and the definition of C_3

$$\sum_{k=1}^{t_1}\eta_k^2\|\mathbf{w}_k - \mathbf{w}_\lambda\|_2^2 \le 2C_1\sum_{k=1}^{t_1}\eta_k^2\sum_{j=1}^{k}\eta_j + 2\|\mathbf{w}_\lambda\|_2^2\sum_{k=1}^{t_1}\eta_k^2.$$

This together with Eq. (4.9) and the definition of ρ gives the stated result with C_2 as

$$C_2 = \max\Bigg\{6 + 8\max\big\{C_4, 2A\eta_1\big\} + \frac{8A\mathcal{E}(0)+8B+8C_5}{AC_1C_3}\sum_{k=1}^{t_1}\eta_k^2,$$
$$8\max\{C_4, 2AC_1C_3\} + 4(AC_1+B)\sum_{k=1}^t\eta_k^2 + 8\big(\mathcal{E}(0)+A^{-1}B+A^{-1}C_5\big)\sum_{k=1}^{t_1}\eta_k\Bigg\}.$$

The proof is complete. ■

4.2 Challenge and Novelty of the Analysis

In this subsection, we present the challenge and novelty of our analysis. We describe the challenge and novelty for both capacity-independent analysis as compared to the NeurIPS article (Lei and Tang, 2018) and the capacity-dependent analysis as compared to the capacity-independent analysis. Note that our novelty consists not only in Lemma 17 and Lemma 19 we have established but also in the analysis we will present in Sections 5 and Section 6. In particular, a challenge in Sections 5 and Section 6 is how to use Lemma 19 to give satisfactory bounds of $\|\mathbf{w}_t\|_2$ for optimal learning rates, which are presented in Lemma 22 (capacity-independent case) and Lemma 27 (capacity-dependent case).

In our NeurIPS article (Lei and Tang, 2018), we assume the finite-norm of $\mathbf{w}_{\mathbf{z}} = \arg\inf_{\mathbf{w}} \mathcal{E}_{\mathbf{z}}(\mathbf{w})$. Under this assumption, we show

$$\|\mathbf{w}_t - \mathbf{w}_{\mathbf{z}}\|_2^2 \leq \sum_{k=1}^t \eta_k \big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\big) + O\big(\sum_{k=1}^t \eta_k^2\big) + 2\sum_{k=1}^t \xi_k^{(n)}, \tag{4.12}$$

where $\xi_k^{(n)} = \eta_k \langle \mathbf{w}_{\mathbf{z}} - \mathbf{w}_k, f'(\mathbf{w}_k; z_{i_k}) - \mathbb{E}_{i_k}[f'(\mathbf{w}_k; z_{i_k})]\rangle$. By Schwartz's inequality and (2.3), we control the variance of $\xi_k^{(n)}$ by

$$\mathbb{E}_{i_k}[(\xi_k^{(n)})^2] \leq A\eta_k^2 \|\mathbf{w}_{\mathbf{z}} - \mathbf{w}_k\|_2^2 \big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_{\mathbf{z}})\big) + \eta_k^2 \|\mathbf{w}_{\mathbf{z}} - \mathbf{w}_k\|_2^2 \big(A\mathcal{E}_{\mathbf{z}}(\mathbf{w}_{\mathbf{z}}) + B\big). \tag{4.13}$$

Since $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) \geq \mathcal{E}_{\mathbf{z}}(\mathbf{w}_{\mathbf{z}})$, we can use an upper bound of $\|\mathbf{w}_{\mathbf{z}} - \mathbf{w}_k\|_2^2$ to control the variance of $\xi_k^{(n)}$. A notable observation is that the term $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_{\mathbf{z}})$ in (4.13) can be offset by the term $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)$ in (4.12). This leads to the result $\max_{t=1,\ldots,T} \|\mathbf{w}_t\|_2^2 = O(\log T)$ (Theorem 3 in Lei and Tang (2018)). This result shows that $\mathbf{w}_t$ always belongs to a small ball and therefore the overfitting phenomenon will never happen for SGD. The underlying reason is the strong assumption on the finite norm of $\mathbf{w}_{\mathbf{z}}$, which depends on training examples and is not intuitive.

In this paper, we do not impose a bounded-norm assumption. Therefore, we consider the term $\|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2$ instead of $\|\mathbf{w}_{\mathbf{z}} - \mathbf{w}_k\|_2^2$ as $\|\mathbf{w}_{\mathbf{z}}\|_2$ may be infinite. The corresponding analysis involves the estimation of the martingale difference sequence $\xi_k = \eta_k \big\langle \mathbf{w}_\lambda - \mathbf{w}_k, f'(\mathbf{w}_k, z_{j_k}) - \mathbb{E}_{j_k}\big[f'(\mathbf{w}_k, z_{j_k})\big]\big\rangle$, whose variance can be bounded by

$$\mathbb{E}_{j_k}\big[\xi_k^2\big] \leq A\eta_k^2 \|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2 \big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda)\big) + \big(A\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) + B\big)\eta_k^2 \|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2. \tag{4.14}$$

A notable difference from (4.13) is that $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda)$ is no longer non-negative, and therefore we can not plug an upper bound of $\|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2$ into (4.14) to give an upper bound of $\mathbb{E}_{j_k}\big[\xi_k^2\big]$. We have to replace $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda)$ by $\max\{\mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda), 0\}$ to ensure the nonnegativity. Then we can use an upper bound of $\|\mathbf{w}_\lambda - \mathbf{w}_k\|_2^2$ to control (4.14), which, combined with (4.12), leads to an upper bound of $\|\mathbf{w}_{t+1}\|_2^2$ as follows

$$\|\mathbf{w}_{t+1}\|_2^2 = O\Big(\sum_{k=1}^t \eta_k \max\{0, \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\} + \|\mathbf{w}_\lambda\|_2^2 \log T\Big). \tag{4.15}$$

As compared to the bound $\|\mathbf{w}_t\|_2^2 = O(\log T)$ (Lei and Tang, 2018), the above bound involves an additional term $\sum_{k=1}^t \eta_k \max\{0, \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)\}$ due to the removal of the

bounded-norm assumption, which is the dominant term and much more challenging to address.

We introduce new decompositions to control this term. For the capacity-independent analysis, we use the decomposition (Eq. (5.2))

$$\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) \leq \underbrace{\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(\hat{h}_{\mathbf{w}_k}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}_k})}_{:=\mathfrak{A}} + \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(\hat{h}_{\mathbf{w}_k}).$$

We then use concentration inequalities to show that

$$\mathfrak{A} = O\Big(n^{-\frac{\alpha}{\alpha+1}} + n^{-\frac{1}{1+\alpha}}\|\mathbf{w}_\lambda\|_2^q + R_T n^{-\frac{1}{2}}\Big),$$

where R_T is an upper bound of $\max_{k=1,\ldots,T}\|\mathbf{w}_k\|_2$. We plug this inequality into (4.15) and derive

$$\|\mathbf{w}_t\|_2 = \widetilde{O}\Big(n^{-\frac{1}{2}}\sum_{k=1}^{T}\eta_k + \|\mathbf{w}_\lambda\|_2 + \Big(\sum_{k=1}^{T}\eta_k\Big)^{\frac{1}{2}}\Big(n^{-\frac{1}{2(1+\alpha)}}\|\mathbf{w}_\lambda\|_2^{\frac{q}{2}} + n^{-\frac{\alpha}{2(1+\alpha)}}\Big)\Big). \tag{4.16}$$

A notable difference here is that this upper bound of $\|\mathbf{w}_t\|_2$ goes to infinity as we run more and more iterations. Therefore, whether imposing a bounded-norm assumption makes the analysis and results essentially different.

The bound (4.16) cannot imply optimal learning rates in the capacity-dependent analysis. To see this, if $\alpha = 1, \zeta = 0$, then Corollary 14 requires $T \asymp n^2$ iterations with $\theta = 1/2$. In this case, the bound (4.16) is larger than $n^{-\frac{1}{2}}\sum_{k=1}^{n^2} k^{-\frac{1}{2}} \asymp \sqrt{n}$ from which one can only get very crude learning rates. The underlying reason is that we require much more iterations to get rates better than $n^{-1/2}$ as compared to capacity-independent case (in the capacity-independent case we only need to run $T \asymp n$ iterations to get almost optimal learning rate $\widetilde{O}(n^{-1/2})$, and then the norm estimation in (4.16) is sufficient). To address this problem, we use the following different error decomposition to fully use the variance-expectation assumption and the capacity assumption (Eq. (6.4))

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) \leq{}& \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(\hat{h}_{\mathbf{w}_k}) + \\ &\underbrace{\big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(h_\rho)\big) + \big(\mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}_k}) - \mathcal{E}(h_\rho) + \mathcal{E}(\hat{h}_{\mathbf{w}_k})\big)}_{:=\mathfrak{B}}.\end{aligned} \tag{4.17}$$

We then use Bernstein-type inequalities to address the term $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(h_\rho)$ (Lemma A.2) and the term $\mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}_k}) - \mathcal{E}(h_\rho) + \mathcal{E}(\hat{h}_{\mathbf{w}_k})$ (Lemma 24), respectively. In turn, we derive a novel polynomial inequality on an upper bound $\widetilde{R}_T$ of $\max_{t=1,\ldots,T}\|\mathbf{w}_t\|_2$ as follows

$$\widetilde{R}_T^2 \leq \tilde{c}_1\widetilde{R}_T^{\frac{2\zeta}{4-2\beta+\beta\zeta}} + \tilde{c}_2\widetilde{R}_T^{\frac{2\zeta}{2+\zeta}} + \tilde{c}_3,$$

where $\tilde{c}_1, \tilde{c}_2, \tilde{c}_3$ are quantities independent of $\widetilde{R}_T$. We solve this inequality by Lemma 21 and derive a refined estimation of $\|\mathbf{w}_t\|_2$ as follows

$$\widetilde{R}_T = \widetilde{O}\Big(n^{-\frac{1}{4-2\beta+\beta\zeta-\zeta}}\Big(\sum_{k=1}^T \eta_k\Big)^{\frac{4-2\beta+\beta\zeta}{2(4-2\beta+\beta\zeta-\zeta)}}\Big) + n^{-\frac{1}{2}}\Big(\sum_{k=1}^T \eta_k\Big)^{\frac{2+\zeta}{4}} + \|\mathbf{w}_\lambda\|_2 + \big(\sum_{k=1}^T \eta_k\big)^{\frac{1}{2}}\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + n^{\frac{1}{\beta-2}} + n^{-1}\|\mathbf{w}_\lambda\|_2^q\big)^{\frac{1}{2}}\Big).$$

For the least square loss, if we choose $\alpha = 1, \zeta = 0, T \asymp n^2$ (ζ can be arbitrarily close to 0) and $\theta = 1/2$, then the above upper bound is of the order of $\|\mathbf{w}_\lambda\|_2 + \big(\sum_{k=1}^{n^2}\eta_k\big)^{\frac{1}{2}}\big(n^{-\frac{1}{2}}\|\mathbf{w}_\lambda\|_2 + \lambda^{\frac{\alpha}{2}}\big) \asymp \|\mathbf{w}_\lambda\|_2$ (we choose $\lambda \asymp n^{-1}$ in this case and note $q = 2, \beta = 1$). That is, even we need much more iterations in the capacity-dependent case, our analysis is still able to imply a good estimation of $\|\mathbf{w}_t\|_2$, which is the key to get an almost optimal capacity-dependent bound.

In summary, the estimation of $\|\mathbf{w}_t\|_2$ is much more challenging than the NeurIPS version due to the removal of the bounded norm assumption. We need to introduce new decomposition and concentration inequalities to address some additional terms. Furthermore, it is more challenging to develop satisfactory bounds of $\|\mathbf{w}_t\|_2$ in the capacity-dependent case since we need much more iterations (can be n^2 versus n) to achieve optimal learning rates (the bound of $\|\mathbf{w}_t\|_2$ in the capacity-independent case can be as large as $O(\sqrt{n})$). We notice that very nice capacity-dependent bounds for least squares regression are very well developed in the literature (Dieuleveut and Bach, 2016; Pillaud-Vivien et al., 2018b; Mücke et al., 2019; Lin and Rosasco, 2017). However, the analysis there uses integral operators which can not be applied here. We have to develop totally different techniques for learning with general convex loss functions.

4.3 Error Decomposition

Before proceeding the capacity-independent and capacity-dependent analysis, we introduce the following error decomposition as a foundation of the subsequent analysis

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = \big(\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) + \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda)\big) + \big(\mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda)\big) + \big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big). \quad (4.18)$$

The term $\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) + \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda)$ is called an estimation error, which comes from approximating the probability measure ρ by its empirical counterpart. The term $\mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda)$ is called an optimization error since it is related to the optimization algorithm for minimizing the empirical risk. The last term is an approximation error which is related to the regularity of the model $\mathbf{w}_\lambda$ and the approximation power of hypothesis space. We will use powerful tools in statistical learning theory, optimization theory and approximation theory to control estimation errors, optimization errors and approximation errors, respectively.

5. Capacity-independent Analysis

In this section, we perform a detailed capacity-independent analysis by estimating norms of $\mathbf{w}_t$, optimization errors and learning rates.

5.1 Estimation of Norm

Our estimation of norm requires two additional lemmas. The following lemma controls the uniform deviation between empirical risks and population risks over a RKHS ball. The proof is standard and put to Section B.

Lemma 20 *Let $R \geq 1$ and define $B_R = \{\mathbf{w} \in \mathcal{W} : \|\mathbf{w}\|_2 \leq R\}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\sup_{\mathbf{w}\in B_R} \left[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\right] \leq C_7 R n^{-\frac{1}{2}} \log^{\frac{1}{2}}(1/\delta), \tag{5.1}$$

where C_7 is a constant independent of R and n (explicitly given in the proof).

We will use an induction strategy to estimate the norm of SGD iterates. Suppose $\|\mathbf{w}_{\tilde{i}}\|_2 \leq R_T$ for $\tilde{i} = 1, \ldots, t$ with R_T defined in Lemma 22. We can essentially derive the result

$$\|\mathbf{w}_{t+1}\|_2^2 = O(\log(T/\delta)) R_T \sum_{k=1}^{T} \eta_k n^{-\frac{1}{2}} + \mathfrak{C},$$

where $\mathfrak{C}$ is independent of R_T. This upper bound is a linear function of R_T, which can be upper bounded by R_T^2 according to the following inequality on univariate polynomials.

Lemma 21 (Cucker and Zhou 2007) *Let $s \in \mathbb{N}, c_1, \ldots, c_s > 0$ and $2 > q_1 > q_2 > \cdots > q_{s-1} > 0$. Then any*

$$x \geq \max\{(sc_1)^{\frac{1}{2-q_1}}, (sc_2)^{\frac{1}{2-q_2}}, \ldots, (sc_s)^{\frac{1}{2}}\}$$

satisfies $c_1 x^{q_1} + c_2 x^{q_2} + \cdots + c_{s-1} x^{q_{s-1}} + c_s \leq x^2$.

Based on the above two lemmas and Lemma 19, we can derive the following lemma on the norm estimation of SGD iterates.

Lemma 22 *Let Assumptions 1, 3 hold, $n^{-1}\|\mathbf{w}_\lambda\|_2^q \leq c_\alpha^q$ and $\delta \in (0,1)$. If $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ is the sequence produced by (2.1) with $\eta_{t+1} \leq \eta_t \leq 1/A$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$, then with probability at least $1-\delta$ we have $\|\mathbf{w}_t\|_2 \leq R_T$ uniformly for all $t = 1, \ldots, T$, where $R_T \geq 1$ is defined by*

$$R_T = C_6 \log^{\frac{1}{2}}(4T/\delta) \max\left\{ \sum_{k=1}^{T} \eta_k n^{-\frac{1}{2}}, \right.$$
$$\left. \left(\|\mathbf{w}_\lambda\|_2^2 + \sum_{k=1}^{T} \eta_k n^{-\frac{1}{1+\alpha}} \|\mathbf{w}_\lambda\|_2^q + \sum_{k=1}^{T} \eta_k \left(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + n^{-\frac{\alpha}{\alpha+1}}\right) + 1 \right)^{\frac{1}{2}} \right\},$$

and C_6 is a constant independent of T, λ, n and δ (explicitly given in the proof).

Proof By $\mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}}) \leq \mathcal{E}_{\mathbf{z}}(\mathbf{w})$, Lemma 18 with $\rho = n^{-\frac{\alpha}{1+\alpha}}$ and Lemma 20, we have the following inequality with probability at least $1-\delta$ uniformly for all $\mathbf{w}$ with $\|\mathbf{w}\|_2 \leq R_T$

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}) &\leq \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}}) = \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(\hat{h}_{\mathbf{w}}) + \mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}}) \\ &\leq \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + \mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}}) \\ &\leq C_8 \log\frac{2}{\delta}\big(n^{-\frac{\alpha}{\alpha+1}} + n^{-\frac{1}{1+\alpha}}\|\mathbf{w}_\lambda\|_2^q\big) + \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + C_7 R_T n^{-\frac{1}{2}} \log^{\frac{1}{2}}\frac{2}{\delta}, \end{aligned} \tag{5.2}$$

where $C_8 = \mathcal{E}(0) + \tilde{c}_q$ ($\tilde{c}_q \geq 1$). According to Lemma 19, Eq. (4.6) holds with probability at least $1-\delta$ simultaneously for all $t = 1, \ldots, T$. In the remainder of the proof we always assume (4.6) and (5.2) hold, which happen with probability at least $1-2\delta$. We now use the induction principle to show that under (4.6) and (5.2) we have $\|\mathbf{w}_{\tilde{i}}\|_2 \leq R_T$ for all $\tilde{i} = 1, \ldots, T$. The case $\tilde{i} = 1$ is clear from the definition of R_T and $\mathbf{w}_1 = 0$. Suppose $\|\mathbf{w}_{\tilde{i}}\|_2 \leq R_T$ for $\tilde{i} = 1, \ldots, t$. We now need to show $\|\mathbf{w}_{t+1}\|_2 \leq R_T$. Plugging (5.2) with $\mathbf{w} = \mathbf{w}_1, \ldots, \mathbf{w}_t$ back into (4.6) and using the induction assumption $\|\mathbf{w}_{\tilde{i}}\|_2 \leq R_T$ for $\tilde{i} = 1, \ldots, t$, we derive

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2^2 \leq C_2 \log\frac{2T}{\delta}\big(\|\mathbf{w}_\lambda\|_2^2 + 1\big) + 8C_7 \log^{\frac{1}{2}}\frac{2}{\delta} R_T \sum_{k=1}^T \eta_k n^{-\frac{1}{2}} \\ + 8\sum_{k=1}^T \eta_k \log\frac{2}{\delta}\Big(C_8\big(n^{-\frac{\alpha}{\alpha+1}} + n^{-\frac{1}{1+\alpha}}\|\mathbf{w}_\lambda\|_2^q\big) + \big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big)\Big).\end{aligned}$$

The right-hand side is a linear function of R_T, which is no larger than R_T^2 by Lemma 21 if R_T satisfies

$$\begin{aligned}R_T \geq \max\Bigg\{16C_7 \log^{\frac{1}{2}}\frac{2}{\delta}\sum_{k=1}^T \eta_k n^{-\frac{1}{2}}, \sqrt{2\log(2T/\delta)}\Big(C_2 + C_2\|\mathbf{w}_\lambda\|_2^2 + 8C_8\sum_{k=1}^T \eta_k n^{-\frac{1}{1+\alpha}}\|\mathbf{w}_\lambda\|_2^q \\ + 8\sum_{k=1}^T \eta_k\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + C_8 n^{-\frac{\alpha}{\alpha+1}}\big)\Big)^{\frac{1}{2}}\Bigg\},\end{aligned}$$

which is further satisfied by the definition of R_T with C_6 defined by $C_6 = \max\big\{16C_7, \sqrt{2C_2}, 4\sqrt{C_8}\big\}$. This shows the stated inequality for $\tilde{i} = t+1$ and finishes the proof. ■

Note that the upper bound established in Lemma 22 goes to infinity as we run more SGD iterations, which is totally different from the almost finiteness of $\mathbf{w}_t$ based on a norm-bounded assumption $\|\mathbf{w}_{\mathbf{z}}\|_2 = O(1)$ (Lei and Tang, 2018). Our analysis here therefore highlights the importance of early-stoping to achieve satisfactory learning rates.

5.2 Estimation of Optimization Errors

Based on the bound of $\|\mathbf{w}_t\|_2$ in Lemma 22, we can provide optimization error bounds stated in Theorem 23. It should be noted that the existing probabilistic bound on optimization errors of SGD requires to assume that either $\mathbf{w}_t$ is bounded or the existence of $\mathbf{w}_{\mathbf{z}}$ with a

finite norm (Lei and Tang, 2018). As a comparison, our optimization error bounds do not require these assumptions. Our idea is to construct a different martingale sequence $\{\xi_t'\}$ conditioned on the probabilistic bound of $\mathbf{w}_t$ established in Lemma 22.

Theorem 23 *Let Assumptions 1, 3 hold, $n^{-1}\|\mathbf{w}_\lambda\|_2^q \le c_\alpha^{\frac{q}{2}}$ and $\delta \in (0,1)$. If $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ is the sequence produced by (2.1) with $\eta_{t+1} \le \eta_t \le 1/A$ and $\sum_{t=1}^\infty \eta_t^2 < \infty$, then with probability at least $1-\delta$*

$$\mathcal{E}_{\mathbf{z}}(\bar{\mathbf{w}}_T) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) = O(1)\log^{\frac{3}{2}}\frac{8T}{\delta}\Big(n^{-1}\sum_{t=1}^T \eta_t + \big(\sum_{t=1}^T \eta_t\big)^{-1}\|\mathbf{w}_\lambda\|_2^2 + n^{-\frac{1}{1+\alpha}}\|\mathbf{w}_\lambda\|_2^q + \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + n^{-\frac{\alpha}{\alpha+1}}\Big).$$

Proof Eq. (4.7) can be reformulated as

$$\sum_{t=1}^T \eta_t\big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_t) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda)\big) \le 2^{-1}\|\mathbf{w}_\lambda\|_2^2 + 2^{-1}(AC_1+B)\sum_{t=1}^T \eta_t^2 + \sum_{t=1}^T \xi_t, \tag{5.3}$$

where $\{\xi_t\}$ is defined in (4.8). Let

$$\xi_t' = \eta_t\big\langle \mathbf{w}_\lambda - \mathbf{w}_t, f'(\mathbf{w}_t, z_{j_t}) - \mathbb{E}_{j_t}\big[f'(\mathbf{w}_t, z_{j_t})\big]\big\rangle \mathbb{I}_{\{\|\mathbf{w}_t\|_2 \le R_T\}},$$

where R_T is defined in Lemma 22 with δ replaced by $\delta/2$ and $\mathbb{I}_{\mathcal{A}}$ denotes the indicator function of an event $\mathcal{A}$, i.e., $\mathbb{I}_{\mathcal{A}} = 1$ if $\mathcal{A}$ happens and 0 otherwise. Analyzing analogously to (4.10) and (4.11), we know

$$\begin{aligned}|\xi_t'| &\le \eta_t\Big[\|\mathbf{w}_\lambda\|_2^2 + (2A\tilde{c}_q+1)\|\mathbf{w}_t\|_2^2 + 2(2A\tilde{c}_q+B)\Big]\mathbb{I}_{\{\|\mathbf{w}_t\|_2\le R_T\}}\\ &\le \eta_t\Big(\|\mathbf{w}_\lambda\|_2^2 + 2(2A\tilde{c}_q+B) + (2A\tilde{c}_q+1)R_T^2\Big).\end{aligned}$$

It is clear that $\mathbb{E}_{j_t}[\xi_t'] = 0$ and ξ_t' depends only on $j_1,\ldots,j_t$. According to Part (a) of Lemma A.1, we can find an event Ω_T with $\Pr\{\Omega_t\} \ge 1 - \frac{\delta}{2}$ such that under the event Ω_T the following inequality holds

$$\sum_{t=1}^T \xi_t' \le \Big(\|\mathbf{w}_\lambda\|_2^2 + 2(2A\tilde{c}_q+B) + (2A\tilde{c}_q+1)R_T^2\Big)\Big(2\sum_{t=1}^T \eta_t^2 \log\frac{2}{\delta}\Big)^{\frac{1}{2}}.$$

Furthermore, according to Lemma 22, there exists an event Ω_T' with $\Pr\{\Omega_t'\} \ge 1 - \frac{\delta}{2}$ such that under the event Ω_T' the inequality $\max_{1\le t\le T}\|\mathbf{w}_t\|_2^2 \le R_T^2$ holds. Under the intersection of these two events, we have $\xi_t = \xi_t'$ and therefore

$$\sum_{t=1}^T \xi_t = \sum_{t=1}^T \xi_t' \le \Big(\|\mathbf{w}_\lambda\|_2^2 + 2(2A\tilde{c}_q+B) + (2A\tilde{c}_q+1)R_T^2\Big)\Big(2\sum_{t=1}^T \eta_t^2 \log\frac{2}{\delta}\Big)^{\frac{1}{2}}, \tag{5.4}$$

which, together with $\Pr\{\Omega_T \cap \Omega_T'\} \geq 1-\delta$ and (5.3), shows the following inequality with probability at least $1-\delta$

$$\begin{aligned}
&\sum_{t=1}^T \eta_t\big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_t) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda)\big) \\
&\leq 2^{-1}\|\mathbf{w}_\lambda\|_2^2 + 2^{-1}\big(AC_1+B\big)\sum_{t=1}^T \eta_t^2 + \Big(\|\mathbf{w}_\lambda\|_2^2 + 2(2A\tilde{c}_q+B) + (2A\tilde{c}_q+1)R_T^2\Big)\Big(2\sum_{t=1}^T \eta_t^2 \log\frac{2}{\delta}\Big)^{\frac{1}{2}} \\
&= O(1)\log^{\frac{3}{2}}\frac{8T}{\delta}\Big(\big(\sum_{t=1}^T \eta_t\big)^2 n^{-1} + \|\mathbf{w}_\lambda\|_2^2 + \sum_{t=1}^T \eta_t n^{-\frac{1}{1+\alpha}}\|\mathbf{w}_\lambda\|_2^q + \sum_{t=1}^T \eta_t\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + n^{-\frac{\alpha}{\alpha+1}}\big)\Big).
\end{aligned}$$

In the above deduction we have used the definition of R_T in Lemma 22 with δ replaced by $\delta/2$. The stated inequality then follows from the convexity of the empirical risk. ■

5.3 Estimation of Learning Rates

We are now ready to prove Theorem 4. As sketched in Section 4.3, our basic idea is to decompose $\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho)$ into estimation errors, optimization errors and approximation errors. We will apply Lemma 20 together with the norm estimate in Lemma 22 to control estimation errors, apply Theorem 23 to control optimization errors and apply Assumption 2 to control approximation errors.

Proof of Theorem 4 Let $\lambda = n^{-\frac{1}{1+\alpha}}$. It follows from Assumption 2 that

$$\lambda\|\mathbf{w}_\lambda\|_2^2 \leq \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + \lambda\|\mathbf{w}_\lambda\|_2^2 \leq c_\alpha \lambda^\alpha,$$

from which we derive

$$\|\mathbf{w}_\lambda\|_2^q \leq \big(c_\alpha \lambda^{\alpha-1}\big)^{\frac{q}{2}} = c_\alpha^{\frac{q}{2}} n^{\frac{(1-\alpha)q}{2+2\alpha}} \leq c_\alpha^{\frac{q}{2}} n. \tag{5.5}$$

Therefore, assumptions in Lemma 22 and Theorem 23 hold. We also have the following elementary inequality

$$(1-\theta)^{-1}(T^{1-\theta}-1) \leq \sum_{t=1}^T t^{-\theta} \leq (1-\theta)^{-1}T^{1-\theta}, \quad \theta \in (0,1). \tag{5.6}$$

We use the following error decomposition w.r.t. $\mathbf{w}_\lambda$ to study the excess population risk

$$\begin{aligned}
\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = \big(\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T})\big) + \big(\mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda)\big) \\
+ \big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda)\big) + \big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big).
\end{aligned} \tag{5.7}$$

According to Theorem 23 and (5.6) with $T \asymp n^{\frac{1}{(1+\alpha)(1-\theta)}}$, we derive the following inequality with probability at least $1-\delta/4$ (notice $\|\mathbf{w}\|_2^q \leq 1 + \|\mathbf{w}\|_2^2$)

$$\mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) = O(1)\log^{\frac{3}{2}}\frac{32T}{\delta}\Big(n^{-\frac{1}{1+\alpha}}\|\mathbf{w}_\lambda\|_2^2 + \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + n^{-\frac{\alpha}{\alpha+1}}\Big). \tag{5.8}$$

We can apply Lemma 18 with $\rho = n^{-\frac{\alpha}{1+\alpha}}$ to derive the following inequality with probability at least $1-\delta/4$

$$\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda) \leq C_8 \log \frac{4}{\delta}\big(n^{-\frac{\alpha}{\alpha+1}} + n^{-\frac{1}{1+\alpha}}\|\mathbf{w}_\lambda\|_2^q\big). \tag{5.9}$$

According to Lemma 22, with probability at least $1-\delta/4$ we derive $\max_{1\leq t\leq T}\|\mathbf{w}_t\|_2 \leq R_T$ with R_T defined in Lemma 22 (δ replaced by $\delta/4$), from which and the convexity of norm we derive with probability at least $1-\delta/4$ that $\|\bar{\mathbf{w}}_T\|_2 \leq R_T$. According to Eq. (5.6) with $T \asymp n^{\frac{1}{(1+\alpha)(1-\theta)}}$, we know

$$R_T = O(\log^{\frac{1}{2}}(16T/\delta)) \max\left\{n^{\frac{1}{1+\alpha}-\frac{1}{2}}, \Big(\|\mathbf{w}_\lambda\|_2^2 + \|\mathbf{w}_\lambda\|_2^q + n^{\frac{1}{1+\alpha}}\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big)\Big)^{\frac{1}{2}}\right\}.$$

This, together with an application of Lemma 20 with R being R_T and union bounds of probability of events, implies the following inequality with probability at least $1-\delta/2$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\bar{\mathbf{w}}_T) = O(\log(16T/\delta)) \max\left\{n^{-\frac{\alpha}{1+\alpha}}, n^{-\frac{\alpha}{2(1+\alpha)}}\Big(n^{-\frac{1}{1+\alpha}}\|\mathbf{w}_\lambda\|_2^2 + \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\Big)^{\frac{1}{2}}\right\}. \tag{5.10}$$

Plugging (5.8), (5.9) and (5.10) into (5.7) and choosing $\lambda = n^{-\frac{1}{1+\alpha}}$, we derive the following inequality with probability at least $1-\delta$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) = O(1)\log^{\frac{3}{2}}\frac{32T}{\delta}\Big(n^{-\frac{\alpha}{\alpha+1}} + \mathcal{D}(n^{-\frac{1}{1+\alpha}}) + n^{-\frac{\alpha}{2(1+\alpha)}}\sqrt{\mathcal{D}(n^{-\frac{1}{\alpha+1}})}\Big).$$

This together with Assumption 2 establishes the stated inequality with probability at least $1-\delta$. The proof is complete. ■

6. Capacity-dependent Analysis

We now develop capacity-dependent learning rates. We will first estimate the norm of $\mathbf{w}_t$, which is then used to study optimization errors and learning rates.

6.1 Estimation of Norm

As we state in Section 4.2, the norm estimation in Lemma 22 can only imply very crude bounds in the capacity-dependent case since we require much more iterations to achieve optimal learning rates here. We therefore need to resort the capacity assumption and the variance-expectation assumption to develop refined estimates of $\|\mathbf{w}_t\|_2$. To this aim, we introduce a different error decomposition (4.17) to tackle $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)$, where the involved term $\mathfrak{B}$ can be shown to decay with a faster rate by applying Bernstein-type concentration inequalities. Lemma 24 is a Bernstein-type inequality to control the uniform deviation between empirical means from expectations for a class of random variables, where the information on variances is included to investigate the concentration behavior.

Lemma 24 (Wu et al. 2007) *Let $\widetilde{\mathcal{F}}$ be a set of measurable functions defined on a space $\widetilde{\mathcal{Z}}$ and $S = \{z_1, \ldots, z_n\}$ be n examples drawn randomly from $\widetilde{\mathcal{Z}}$. Let $\tau \in [0,1], M, c_\tau$ be constants such that each function $f \in \widetilde{\mathcal{F}}$ satisfies $\sup_{z\in\widetilde{\mathcal{Z}}} |f(z)| \le M$ and $\mathbb{E}[f^2(Z)] \le c_\tau(\mathbb{E}[f(Z)])^\tau$. If for some $c_\zeta \ge M^\zeta$ and $\zeta \in (0,2)$, $\mathbb{E}_S\big[\log \mathcal{N}_2(\epsilon, \widetilde{\mathcal{F}}, S)\big] \le c_\zeta \epsilon^{-\zeta}$ for all $\epsilon > 0$, then there exists a positive c'_ζ depending only on ζ such that for any $\delta \in (0,1)$, with probability at least $1-\delta$ the following inequality holds uniformly for all $f \in \widetilde{\mathcal{F}}$*

$$\mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^n f(z_i) \le \frac{1}{2}\eta^{1-\tau}\big(\mathbb{E}[f(Z)]\big)^\tau + c'_\zeta\eta + 2(c_\tau n^{-1}\log(1/\delta))^{\frac{1}{2-\tau}} + 18Mn^{-1}\log(1/\delta),$$

where

$$\eta := \max\Big\{c_\tau^{\frac{2-\zeta}{4-2\tau+\zeta\tau}}(c_\zeta n^{-1})^{\frac{2}{4-2\tau+\zeta\tau}}, M^{\frac{2-\zeta}{2+\zeta}}(c_\zeta n^{-1})^{\frac{2}{2+\zeta}}\Big\}.$$

Based on Lemma 24, we can derive a probabilistic bound on $\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}}) + \mathcal{E}_{\mathbf{z}}(h_\rho)$ uniformly for all $\mathbf{w} \in B_R$, which is an essential component in Eq. (4.17).

Lemma 25 *Let Assumptions 1, 3-5 hold. Let $\delta \in (0,1)$ and $R \ge 1$, then the following inequality holds with probability at least $1-\delta$ uniformly for all $\mathbf{w} \in B_R$*

$$\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}}) + \mathcal{E}_{\mathbf{z}}(h_\rho) \le 2^{-1}\big(\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho)\big) + \eta R^{\frac{2\zeta}{4-2\beta+\beta\zeta}} + \tilde{\eta}R^{\frac{2\zeta}{2+\zeta}} + 2(c_\beta n^{-1}\log(1/\delta))^{\frac{1}{2-\beta}} + 18n^{-1}\tilde{b}\log(1/\delta),$$

where $\tilde{b} = \sup_{|y|,|\tilde{y}|\le b}\ell(\tilde{y}, y), C_9 = \sup_{|y|,|\tilde{y}|\le b}|\ell'(\tilde{y}, y)|$, c'_ζ is a constant depending only on ζ and

$$\eta = (2^{-1} + c'_\zeta)c_\beta^{\frac{2-\zeta}{4-2\beta+\beta\zeta}}\big(n^{-1}c_\zeta C_9^\zeta\big)^{\frac{2}{4-2\beta+\beta\zeta}}, \quad \tilde{\eta} = (2^{-1} + c'_\zeta)\tilde{b}^{\frac{2-\zeta}{2+\zeta}}\big(n^{-1}c_\zeta C_9^\zeta\big)^{\frac{2}{2+\zeta}}. \tag{6.1}$$

Proof Introduce the class of loss functions and excess loss functions as

$$\mathcal{F}_R = \big\{\ell(\hat{h}_{\mathbf{w}}(x), y) : \mathbf{w} \in B_R\big\} \quad \text{and} \quad \mathcal{F}_R^* = \big\{\ell(\hat{h}_{\mathbf{w}}(x), y) - \ell(h_\rho(x), y) : \mathbf{w} \in B_R\big\}.$$

For any $\mathbf{w} \in B_R$, we know

$$\sup_{z, \|\mathbf{w}\|_2 \le R}\big|\ell(\hat{h}_{\mathbf{w}}(x), y) - \ell(h_\rho(x), y)\big| \le \sup_{|y|,|\tilde{y}|\le b}\ell(\tilde{y}, y). \tag{6.2}$$

and $\big|\ell'(\hat{h}_{\mathbf{w}}(x), y)\big| \le \sup_{|y|,|\tilde{y}|\le b}|\ell'(\tilde{y}, y)| = C_9$. It then follows from the standard structural result on covering numbers that

$$\mathcal{N}_2(\epsilon, \mathcal{F}_R^*, S) = \mathcal{N}_2(\epsilon, \mathcal{F}_R, S) \le \mathcal{N}_2(\epsilon/C_9, \mathcal{H}_R, S_{\mathbf{x}}) = \mathcal{N}_2(\epsilon/(C_9R), \mathcal{H}_1, S_{\mathbf{x}}),$$

where $S_{\mathbf{x}} = \{x_1, \ldots, x_n\}$. This together with Assumption 4 shows that

$$\mathbb{E}_S\big[\log \mathcal{N}_2(\epsilon, \mathcal{F}_R^*, S)\big] \le c_\zeta(C_9R)^\zeta\epsilon^{-\zeta}, \quad \forall \epsilon > 0. \tag{6.3}$$

An application of Lemma 24 with $\widetilde{\mathcal{F}} = \mathcal{F}_R^*$ together with (6.2) and (6.3) establishes the following inequality with probability at least $1-\delta$ for all $\mathbf{w} \in B_R$

$$\begin{aligned}
&\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}}) + \mathcal{E}_{\mathbf{z}}(h_\rho) \\
&\le 2^{-1}(\eta')^{1-\beta}\big(\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho)\big)^{\beta} + c'_\zeta \eta' + 2(c_\beta n^{-1}\log(1/\delta))^{\frac{1}{2-\beta}} + 18n^{-1}\tilde{b}\log(1/\delta) \\
&\le 2^{-1}\big(\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho)\big) + (2^{-1} + c'_\zeta)\eta' + 2(c_\beta n^{-1}\log(1/\delta))^{\frac{1}{2-\beta}} + 18n^{-1}\tilde{b}\log(1/\delta),
\end{aligned}$$

where we have used $(\eta')^{1-\beta}\big(\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho)\big)^{\beta} \le \eta' + \mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho)$, and introduced

$$\eta' = \max\Big\{c_\beta^{\frac{2-\zeta}{4-2\beta+\beta\zeta}}\big(n^{-1}c_\zeta C_9^\zeta R^\zeta\big)^{\frac{2}{4-2\beta+\beta\zeta}}, \tilde{b}^{\frac{2-\zeta}{2+\zeta}}\big(n^{-1}c_\zeta C_9^\zeta R^\zeta\big)^{\frac{2}{2+\zeta}}\Big\}.$$

The stated result then follows from the definition of η and $\tilde{\eta}$. The proof is complete. ■

Similarly, we can apply Bernstein inequality (Lemma A.2) to derive a probabilistic bound for $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(h_\rho)$. The proof of Lemma 26 can be found in the appendix.

Lemma 26 *Let $\mathbf{w}_\lambda$ be defined by (2.4), Assumption 5 hold and $\delta \in (0,1)$. The following inequality holds with probability at least $1-\delta$*

$$\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(h_\rho) \le \frac{\beta}{2}\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big) + C_{10}n^{\frac{1}{\beta-2}}\log\frac{2}{\delta} + C_{10}n^{-1}\big(\|\mathbf{w}_\lambda\|_2^q + 1\big)\log\frac{2}{\delta},$$

where

$$C_{10} = \max\Big\{2\tilde{b}/3 + \big(1-2^{-1}\beta\big)(2c_\beta)^{\frac{1}{2-\beta}}, \big(2/3 + 1/\beta\big)\tilde{c}_q\Big\}.$$

We are now able to present our bound of $\|\mathbf{w}_t\|_2$. Plugging Lemma 25 and Lemma 26 into (4.17), we can derive a probabilistic bound on $\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)$. Based on this we can derive a bound on $\|\mathbf{w}_t\|_2$ with an induction strategy. Specifically, by an application of Lemma 19 with the induction assumption $\|\mathbf{w}_{\tilde{i}}\|_2 \le \widetilde{R}_T$ for $\tilde{i} = 1, \ldots, t$, we can show with high probability $\|\mathbf{w}_{t+1}\|_2^2 = O(1)\big(\widetilde{R}_T^{\frac{2\zeta}{4-2\beta+\beta\zeta}} + \widetilde{R}_T^{\frac{2\zeta}{2+\zeta}}\big) + \mathfrak{D}$, where $\widetilde{R}_T$ is defined in Lemma 27 and $\mathfrak{D}$ is independent of $\widetilde{R}_T$. We can apply the inequality on univariate polynomials (Lemma 21) to show that $\|\mathbf{w}_{t+1}\|_2 \le \widetilde{R}_T$ with high probability.

Lemma 27 *Let Assumptions 1, 3-5 hold, $n^{-1}\|\mathbf{w}_\lambda\|_2^q \le c_\alpha^{\frac{q}{2}}$ and $\delta \in (0,1)$. If $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ is the sequence produced by (2.1) with $\eta_{t+1} \le \eta_t \le 1/A$ and $\sum_{t=1}^\infty \eta_t^2 < \infty$, then with probability at least $1-\delta$ we have $\|\mathbf{w}_t\|_2 \le \widetilde{R}_T$ uniformly for all $t = 1, \ldots, T$, where $\widetilde{R}_T \ge 1$ is defined by*

$$\begin{aligned}
\widetilde{R}_T = \max\Bigg\{&\Big(24\eta\sum_{k=1}^T \eta_k\Big)^{\frac{4-2\beta+\beta\zeta}{8-4\beta+2\beta\zeta-2\zeta}}, \Big(24\tilde{\eta}\sum_{k=1}^T \eta_k\Big)^{\frac{2+\zeta}{4}}, \log^{\frac{1}{2}}\frac{6T}{\delta}\Bigg(C_{11}\big(\|\mathbf{w}_\lambda\|_2^2 + \\
&n^{-1}\sum_{k=1}^T \eta_k\|\mathbf{w}_\lambda\|_2^q\big) + 12(\beta+2)\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big)\sum_{k=1}^T \eta_k + C_{12}\big(1 + n^{\frac{1}{\beta-2}}\sum_{k=1}^T \eta_k\big)\Bigg)^{\frac{1}{2}}\Bigg\},
\end{aligned}$$

$\eta, \tilde{\eta}$ are given in (6.1) and $C_{11} \ge 1, C_{12}$ are two constants independent of n, T and λ (explicitly given in the proof).

Proof According to Lemma 19, Eq. (4.6) holds with probability at least $1-\delta$ simultaneously for all $t = 1, \ldots, T$. In the remainder of the proof we always assume (4.6), Lemma 25 and Lemma 26 hold, which happen with probability at least $1 - 3\delta$. We now use the induction principle to show that under (4.6), Lemma 25 and Lemma 26 we have $\|\mathbf{w}_{\tilde{i}}\|_2 \leq \widetilde{R}_T$ for all $\tilde{i} = 1, \ldots, T$. The case $\tilde{i} = 1$ is clear from the definition of $\widetilde{R}_T$. Suppose $\|\mathbf{w}_{\tilde{i}}\|_2 \leq \widetilde{R}_T$ for $\tilde{i} = 1, \ldots, t$. We now need to show $\|\mathbf{w}_{t+1}\|_2 \leq \widetilde{R}_T$. Since $\mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}_k}) \leq \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k)$, we know

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) &\leq \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}_k}) \\ &= \Big(\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(h_\rho)\Big) \\ &\quad + \Big(\mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}_k}) - \mathcal{E}(h_\rho) + \mathcal{E}(\hat{h}_{\mathbf{w}_k})\Big) + \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(\hat{h}_{\mathbf{w}_k}). \qquad (6.4)\end{aligned}$$

Since $\|\mathbf{w}_k\|_2 \leq \widetilde{R}_T$ for $k = 1, 2, \ldots, t$, we can combine Lemma 25 with $\mathbf{w} = \mathbf{w}_k$, Lemma 26 and (6.4) together to derive the following inequality for all $k = 1, \ldots, t$

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_k) \leq \big(1 + 2^{-1}\beta\big)\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big) + \eta\widetilde{R}_T^{\frac{2\zeta}{4-2\beta+\beta\zeta}} + \tilde{\eta}\widetilde{R}_T^{\frac{2\zeta}{2+\zeta}} + \\ n^{\frac{1}{\beta-2}}(2c_\beta^{\frac{1}{2-\beta}} + C_{10})\log(2/\delta) + n^{-1}\big(C_{10}\|\mathbf{w}_\lambda\|_2^q + C_{10} + 18\tilde{b}\big)\log(2/\delta), \quad (6.5)\end{aligned}$$

where we used $2^{-1}\big(\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho)\big) \leq \mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}(h_\rho)$ and the definition of $\eta, \tilde{\eta}$ given in (6.1).

Plugging (6.5) with $\mathbf{w} = \mathbf{w}_1, \ldots, \mathbf{w}_t$ back into (4.6) together with the induction assumption, we derive

$$\begin{aligned}&\|\mathbf{w}_{t+1}\|_2^2 \leq C_2 \log\frac{2T}{\delta}\big(\|\mathbf{w}_\lambda\|_2^2 + 1\big) + 8\sum_{k=1}^t \eta_k\big(\eta\widetilde{R}_T^{\frac{2\zeta}{4-2\beta+\beta\zeta}} + \tilde{\eta}\widetilde{R}_T^{\frac{2\zeta}{2+\zeta}}\big) + \\ &8\sum_{k=1}^t \eta_k\Big(\big(1+2^{-1}\beta\big)\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big) + n^{\frac{1}{\beta-2}}(2c_\beta^{\frac{1}{2-\beta}} + C_{10})\log(2/\delta) + n^{-1}\big(C_{10}\|\mathbf{w}_\lambda\|_2^q + C_{10} + 18\tilde{b}\big)\log(2/\delta)\Big).\end{aligned}$$

According to Lemma 21, the right-hand side of the above inequality is less than $\widetilde{R}_T^2$ if $\widetilde{R}_T$ satisfies

$$\begin{aligned}\widetilde{R}_T \geq \max\Bigg\{&\Big(24\eta\sum_{k=1}^T \eta_k\Big)^{\frac{4-2\beta+\beta\zeta}{8-4\beta+2\beta\zeta-2\zeta}}, \Big(24\tilde{\eta}\sum_{k=1}^T \eta_k\Big)^{\frac{2+\zeta}{4}}, \\ &\Bigg(3C_2\log\frac{2T}{\delta}\big(\|\mathbf{w}_\lambda\|_2^2 + 1\big) + 24\sum_{k=1}^T \eta_k\Big(\big(1+2^{-1}\beta\big)\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big) + \\ &n^{\frac{1}{\beta-2}}(2c_\beta^{\frac{1}{2-\beta}} + C_{10})\log(2/\delta) + n^{-1}\big(C_{10}\|\mathbf{w}_\lambda\|_2^q + C_{10} + 18\tilde{b}\big)\log(2/\delta)\Big)\Bigg)^{\frac{1}{2}}\Bigg\},\end{aligned}$$

which is satisfied by $\widetilde{R}_T$ of the stated form with δ replaced by $\delta/3$ if we introduce

$$C_{11} = 3\max\{C_2, 8C_{10}\},\ C_{12} = 3\max\Big\{C_2, 16(c_\beta^{\frac{1}{2-\beta}} + C_{10} + 9\tilde{b})\Big\}.$$

This establishes the induction assumption with $\tilde{i} = t + 1$ and finishes the proof. ■

6.2 Estimation of Optimization Errors

Analogous to Theorem 23 but with the norm estimate in Lemma 27, we can immediately derive the following theorem on optimization errors.

Theorem 28 *Let Assumptions 1, 3-5 hold,* $n^{-1}\|\mathbf{w}_\lambda\|_2^q \leq c_\alpha^{\frac{q}{2}}$ *and* $\delta \in (0,1)$. *If* $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ *is given by* (2.1) *with* $\eta_{t+1} \leq \eta_t \leq 1/A$ *and* $\sum_{t=1}^\infty \eta_t^2 < \infty$, *then with probability at least* $1-\delta$

$$\mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) = O(1)\log^{\frac{3}{2}}\frac{12T}{\delta}\max\Bigg\{n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}}\Big(\sum_{k=1}^T\eta_k\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta-\zeta}}, n^{-1}\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{2}}, \\ \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + n^{\frac{1}{\beta-2}} + \big(\sum_{t=1}^T\eta_t\big)^{-1}\|\mathbf{w}_\lambda\|_2^2 + n^{-1}\|\mathbf{w}_\lambda\|_2^q\Bigg\}. \quad (6.6)$$

Proof With probability at least $1-\delta/2$, Lemma 27 implies that $\|\mathbf{w}_t\|_2 \leq \widetilde{R}_T$ with $\widetilde{R}_T$ defined in Lemma 27 but δ replaced by $\delta/2$. According to the definition of η and $\tilde{\eta}$ given by (6.1), it can be directly checked that

$$\widetilde{R}_T^2 = O(1)\log\frac{12T}{\delta}\max\Bigg\{n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}}\Big(\sum_{k=1}^T\eta_k\Big)^{\frac{4-2\beta+\beta\zeta}{4-2\beta+\beta\zeta-\zeta}}, n^{-1}\Big(\sum_{k=1}^T\eta_k\Big)^{\frac{2+\zeta}{2}}, \\ \big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + n^{\frac{1}{\beta-2}}\big)\sum_{k=1}^T\eta_k + \|\mathbf{w}_\lambda\|_2^2 + n^{-1}\sum_{k=1}^T\eta_k\|\mathbf{w}_\lambda\|_2^q\Bigg\}. \quad (6.7)$$

Analogous to the proof of Theorem 23, it can be shown that (5.4) with R_T replaced by $\widetilde{R}_T$ holds with probability at least $1-\delta$. This together with (5.3) and the above bound of $\widetilde{R}_T$ gives the stated inequality with probability at least $1-\delta$. The proof is complete. ■

6.3 Estimation of Learning Rates

Combining the norm estimate in Lemma 27 and optimization errors in Theorem 28 together, we can finally derive capacity-dependent learning rates for SGD.

Proof of Theorem 8 We will choose appropriate λ such that $n^{-1}\|\mathbf{w}_\lambda\|_2^q \leq c_\alpha^{\frac{q}{2}}$ (verified at the end of the proof). We use the following error decomposition here

$$\begin{aligned}\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) &= \mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) + \mathcal{E}_{\mathbf{z}}(h_\rho) \\ &\quad + \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(h_\rho) \\ &\quad + \mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) + \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho). \quad (6.8)\end{aligned}$$

Let η and $\tilde{\eta}$ be defined in (6.1), and $\widetilde{R}_T$ be defined in Lemma 27 with δ replaced by $\delta/4$. By the Young's inequality (B.5) we know

$$\begin{aligned} n^{-\frac{2}{4-2\beta+\beta\zeta}}\widetilde{R}_T^{\frac{2\zeta}{4-2\beta+\beta\zeta}} &= \Big((\sum_{t=1}^T \eta_t)^{-1}\widetilde{R}_T^2\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta}}\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta}} n^{-\frac{2}{4-2\beta+\beta\zeta}} \\ &\leq \frac{\zeta}{4-2\beta+\beta\zeta}\Big(\sum_{t=1}^T\eta_t\Big)^{-1}\widetilde{R}_T^2 + \frac{4-2\beta+\beta\zeta-\zeta}{4-2\beta+\beta\zeta}\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta-\zeta}} n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}} \end{aligned}$$

and

$$n^{-\frac{2}{2+\zeta}}\widetilde{R}_T^{\frac{2\zeta}{2+\zeta}} = \Big((\sum_{t=1}^T\eta_t)^{-1}\widetilde{R}_T^2\Big)^{\frac{\zeta}{2+\zeta}}\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{2+\zeta}} n^{-\frac{2}{2+\zeta}} \leq \frac{\zeta}{2+\zeta}\Big(\sum_{t=1}^T\eta_t\Big)^{-1}\widetilde{R}_T^2 + \frac{2}{2+\zeta}\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{2}} n^{-1}.$$

Therefore,

$$\eta\widetilde{R}_T^{\frac{2\zeta}{4-2\beta+\beta\zeta}} + \tilde{\eta}\widetilde{R}_T^{\frac{2\zeta}{2+\zeta}} = O(1)\Big(\Big(\sum_{t=1}^T\eta_t\Big)^{-1}\widetilde{R}_T^2 + \Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta-\zeta}} n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}} + \Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{2}} n^{-1}\Big).$$

Since $\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta-\zeta}} n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}} = \Big(\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{2}} n^{-1}\Big)^{\frac{2}{4-2\beta+\beta\zeta-\zeta}}$ and $2 \leq 4-2\beta+\beta\zeta-\zeta$, we can choose appropriate T, λ such that (indeed this is the case of interest)

$$\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{2}} n^{-1} = O(1) \tag{6.9}$$

and therefore

$$\eta\widetilde{R}_T^{\frac{2\zeta}{4-2\beta+\beta\zeta}} + \tilde{\eta}\widetilde{R}_T^{\frac{2\zeta}{2+\zeta}} = O(1)\Big(\Big(\sum_{t=1}^T\eta_t\Big)^{-1}\widetilde{R}_T^2 + \Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta-\zeta}} n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}}\Big).$$

According to Lemma 27, with probability at least $1-\delta/4$ we have $\max_{1\leq t\leq T}\|\mathbf{w}_t\|_2 \leq \widetilde{R}_T$, from which and the convexity of norm we know with probability at least $1-\delta/4$ that $\|\bar{\mathbf{w}}_T\|_2 \leq \widetilde{R}_T$. This together with an application of Lemma 25 with δ replaced by $\delta/4$ plus an union bound of probability of events shows the following inequality with probability at least $1-\delta/2$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) + \mathcal{E}_{\mathbf{z}}(h_\rho) \leq \frac{1}{2}\big(\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho)\big) + \eta\widetilde{R}_T^{\frac{2\zeta}{4-2\beta+\beta\zeta}} + \tilde{\eta}\widetilde{R}_T^{\frac{2\zeta}{2+\zeta}} + O(1)n^{\frac{1}{\beta-2}}\log\frac{4}{\delta}.$$

Combining the above two inequalities together, we derive the following inequality with probability at least $1-\delta/2$

$$\begin{aligned} \mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T}) + \mathcal{E}_{\mathbf{z}}(h_\rho) &\leq 2^{-1}\big(\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T}) - \mathcal{E}(h_\rho)\big) + O(1)n^{\frac{1}{\beta-2}}\log(4/\delta) \\ &\quad + O(1)\Big(\Big(\sum_{t=1}^T\eta_t\Big)^{-1}\widetilde{R}_T^2 + \Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta-\zeta}} n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}}\Big). \end{aligned}$$

According to Theorem 28, with probability at least $1-\delta/4$ we know that (6.6) and (6.7) with δ replaced by $\delta/4$ hold. This further shows the following inequality with probability at least $1-3\delta/4$ (note (6.9))

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T})-\mathcal{E}(h_\rho)-\mathcal{E}_{\mathbf{z}}(\hat{h}_{\bar{\mathbf{w}}_T})+\mathcal{E}_{\mathbf{z}}(h_\rho)=2^{-1}\big(\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T})-\mathcal{E}(h_\rho)\big)+O(\log^{\frac{3}{2}}(48T/\delta))$$
$$\times\bigg(\mathcal{E}(\mathbf{w}_\lambda)-\mathcal{E}(h_\rho)+n^{\frac{1}{\beta-2}}+\Big(\sum_{k=1}^T\eta_k\Big)^{-1}\|\mathbf{w}_\lambda\|_2^2+n^{-1}\|\mathbf{w}_\lambda\|_2^q+\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta-\zeta}}n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}}\bigg).$$

We can plug the inequality in Lemma 26, (6.6) with δ replaced by $\delta/4$ and the above inequality into (6.8) to derive the following inequality with probability at least $1-\delta$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T})-\mathcal{E}(h_\rho)\le 2^{-1}\big(\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T})-\mathcal{E}(h_\rho)\big)+O(\log^{\frac{3}{2}}(48T/\delta))$$
$$\times\bigg(\mathcal{E}(\mathbf{w}_\lambda)-\mathcal{E}(h_\rho)+n^{\frac{1}{\beta-2}}+\Big(\sum_{k=1}^T\eta_k\Big)^{-1}\|\mathbf{w}_\lambda\|_2^2+n^{-1}\|\mathbf{w}_\lambda\|_2^q+\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta-\zeta}}n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}}\bigg).$$

By (5.5) and Assumption 2, we know $\|\mathbf{w}_\lambda\|_2^q=O(1)\big(\lambda^{\frac{q(\alpha-1)}{2}}\big)$ and $\mathcal{E}(\mathbf{w}_\lambda)-\mathcal{E}(h_\rho)=O(\lambda^\alpha)$, which, together with the above inequality, gives the following inequality with probability at least $1-\delta$

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T})-\mathcal{E}(h_\rho)=O(\log^{\frac{3}{2}}(48T/\delta))\bigg(n^{\frac{1}{\beta-2}}+\lambda^\alpha+\Big(\sum_{k=1}^T\eta_k\Big)^{-1}\lambda^{\alpha-1}+n^{-1}\lambda^{\frac{(\alpha-1)q}{2}}$$
$$+\Big(\sum_{t=1}^T\eta_t\Big)^{\frac{\zeta}{4-2\beta+\beta\zeta-\zeta}}n^{-\frac{2}{4-2\beta+\beta\zeta-\zeta}}\bigg).$$

If we choose an appropriate T such that

$$\sum_{t=1}^T\eta_t\asymp n^{\frac{2}{4-2\beta+\beta\zeta}}\lambda^{\frac{(\alpha-1)(4-2\beta+\beta\zeta-\zeta)}{4-2\beta+\beta\zeta}},\tag{6.10}$$

then with probability at least $1-\delta$ there holds

$$\mathcal{E}(\hat{h}_{\bar{\mathbf{w}}_T})-\mathcal{E}(h_\rho)=O(\log^{\frac{3}{2}}(48T/\delta))\bigg(\lambda^\alpha+n^{-\frac{2}{4-2\beta+\beta\zeta}}\lambda^{\frac{(\alpha-1)\zeta}{4-2\beta+\beta\zeta}}+n^{-1}\lambda^{\frac{(\alpha-1)q}{2}}+n^{\frac{1}{\beta-2}}\bigg).\tag{6.11}$$

We first consider the case $2\alpha+\alpha\beta\zeta+\zeta+\alpha q\ge\alpha\zeta+2\alpha\beta+q$. In this case, we choose $\lambda=n^{\frac{2}{(\alpha-1)\zeta-4\alpha+2\alpha\beta-\alpha\beta\zeta}}$, with which we know

$$n^{-1}\lambda^{\frac{(\alpha-1)q}{2}}=n^{-1}n^{\frac{(\alpha-1)q}{(\alpha-1)\zeta-4\alpha+2\alpha\beta-\alpha\beta\zeta}}\le n^{\frac{2\alpha}{(\alpha-1)\zeta-4\alpha+2\alpha\beta-\alpha\beta\zeta}}=\lambda^\alpha=n^{-\frac{2}{4-2\beta+\beta\zeta}}\lambda^{\frac{(\alpha-1)\zeta}{4-2\beta+\beta\zeta}},$$

where the inequality is due to $2\alpha+\alpha\beta\zeta+\zeta+\alpha q\ge\alpha\zeta+2\alpha\beta+q$ and the two identities are due to the choice of λ. Plugging the above inequality back into (6.11) and using $n^{\frac{1}{\beta-2}}\le$

$n^{\frac{2\alpha}{(\alpha-1)\zeta-4\alpha+2\alpha\beta-\alpha\beta\zeta}}$ implies the inequality (2.6) with probability at least $1-\delta$. Furthermore, it is clear from (5.6) and $T \asymp n^{\frac{2}{(\zeta+4\alpha+\alpha\beta\zeta-\alpha\zeta-2\alpha\beta)(1-\theta)}}$ that

$$\sum_{t=1}^{T} \eta_t \asymp n^{-\frac{2}{\alpha\zeta-\zeta-4\alpha+2\alpha\beta-\alpha\beta\zeta}} = n^{\frac{2(\alpha-1)(4-2\beta+\beta\zeta-\zeta)}{(4-2\beta+\beta\zeta)(\alpha\zeta-\zeta-4\alpha+2\alpha\beta-\alpha\beta\zeta)}+\frac{2}{4-2\beta+\beta\zeta}},$$

which shows that (6.9) and (6.10) hold with the choice of T. This proves Part (a).

We now consider the case $2\alpha+\alpha\beta\zeta+\zeta+\alpha q < \alpha\zeta+2\alpha\beta+q$. In this case, we can choose $\lambda = n^{\frac{2}{\alpha q-2\alpha-q}}$, with which we know

$$\begin{aligned} n^{-\frac{2}{4-2\beta+\beta\zeta}}\lambda^{\frac{(\alpha-1)\zeta}{4-2\beta+\beta\zeta}} &= n^{-\frac{2}{4-2\beta+\beta\zeta}} n^{\frac{2(\alpha-1)\zeta}{(4-2\beta+\beta\zeta)(\alpha q-2\alpha-q)}} \\ &= n^{\frac{4\alpha+2q-2\alpha q+2\alpha\zeta-2\zeta}{(4-2\beta+\beta\zeta)(\alpha q-2\alpha-q)}} \le n^{\frac{2\alpha}{\alpha q-2\alpha-q}} = \lambda^{\alpha} = n^{-1}\lambda^{\frac{(\alpha-1)q}{2}}, \end{aligned}$$

where the inequality is due to $2\alpha+\alpha\beta\zeta+\zeta+\alpha q < \alpha\zeta+2\alpha\beta+q$ and the two identities are due to the choice of λ. Plugging the above inequality back into (6.11) and using $n^{\frac{1}{\beta-2}} \le n^{\frac{2\alpha}{\alpha q-2\alpha-q}}$ due to $2\alpha+\alpha\beta\zeta+\zeta+\alpha q < \alpha\zeta+2\alpha\beta+q$ give the inequality (2.7) with probability at least $1-\delta$. Furthermore, it is clear from (5.6) and $T \asymp n^{\frac{2(2\alpha-2\alpha\beta+\alpha\beta\zeta-\alpha\zeta-4+2\beta-\beta\zeta+\zeta+\alpha q-q)}{(4-2\beta+\beta\zeta)(\alpha q-2\alpha-q)(1-\theta)}}$ that

$$\sum_{t=1}^{T} \eta_t \asymp n^{\frac{2(2\alpha-2\alpha\beta+\alpha\beta\zeta-\alpha\zeta-4+2\beta-\beta\zeta+\zeta+\alpha q-q)}{(4-2\beta+\beta\zeta)(\alpha q-2\alpha-q)}} = n^{\frac{2(\alpha-1)(4-2\beta+\beta\zeta-\zeta)}{(4-2\beta+\beta\zeta)(\alpha q-2\alpha-q)}+\frac{2}{4-2\beta+\beta\zeta}},$$

which shows that (6.9) and (6.10) hold with the choice of T in this case. This proves Part (b).

According to (5.5), it is clear that $n^{-1}\|\mathbf{w}_\lambda\|_2^q \le c_\alpha^{\frac{q}{2}} n^{-1}\lambda^{\frac{(\alpha-1)q}{2}} \le c_\alpha^{\frac{q}{2}}$ holds in both two cases with our choice of λ. The proof is complete. ■

7. Conclusions

This paper presents a learning rate analysis of SGD with convex loss functions. We develop both capacity-independent and capacity-dependent learning rates with high probability. Our capacity-independent learning rates remove the bounded subgradient assumption (Hardt et al., 2016), the smoothness assumption (Hardt et al., 2016) and the assumption on the existence of an empirical risk minimizer with a finite norm (Lei and Tang, 2018). Our capacity-dependent rates extend the existing discussion from the least squares loss (Lin and Rosasco, 2017) to general convex loss functions. It would be interesting to extend our analysis to other learning setting, e.g., distributed learning (Lin and Cevher, 2018; Mücke and Blanchard, 2018) and learning with random features (Carratino et al., 2018).

Acknowledgments

We thank the anonymous reviewers and the editor for their constructive comments. This work is supported partially by the National Natural Science Foundation of China (Grant

Nos. 61806091, 12071356), and MOE University Scientific-Technological Innovation Plan Program. Yunwen Lei also acknowledges support by the Alexander von Humboldt Foundation. Parts of the results in this paper were presented at Advances in Neural Information Processing Systems 31 (2018), 1526–1536. The corresponding author is Ting Hu.

Appendix A. Concentration Inequalities

In this section, we collect some concentration inequalities useful in our theoretical analysis.

We first introduce powerful concentration inequalities on martingales. Part (a) is the Azuma-Hoeffding inequality for martingales with bounded increments (Boucheron et al., 2013), and part (b) is a conditional Bernstein inequality using the conditional variance to quantify better the concentration behavior of martingales (Zhang, 2005).

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

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

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

(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

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

We then introduce a Bernstein inequality which controls the deviation between empirical means and expectation for random variables via information on variances.

Lemma A.2 (Bernstein inequality) *Let $\{\xi(z_i)\}_{i=1}^m$ be a sequence of real-valued random variables and $\widetilde{M}$ be a constant such that $|\xi| \le \widetilde{M}$ and the variance $Var(\xi) < \infty$, then for any $0 < \delta < 1$ with confidence at least $1-\delta$ there holds*

$$\frac{1}{n}\sum_{i=1}^n \xi(z_i) - \mathbb{E}[\xi] \le \frac{2\widetilde{M}\log\frac{1}{\delta}}{3n} + \sqrt{\frac{2\,Var(\xi)\log\frac{1}{\delta}}{n}}.$$

Finally, we introduce the McDiarmid's inequality for real-valued functions of independent random variables that satisfy a bounded increment condition (McDiarmid, 1989).

Lemma A.3 *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 $f : \mathcal{Z}^n \to \mathbb{R}$ satisfies*

$$\sup_{z_1,\ldots,z_n,\bar{z}_k\in\mathcal{Z}} |f(z_1,\cdots,z_n) - f(z_1,\cdots,z_{k-1},\bar{z}_k,z_{k+1},\cdots,z_n)| \le c_k \tag{A.3}$$

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

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

Appendix B. Proofs of Some Lemmas

We collect here the proofs of some preliminary lemmas used in our learning rate analysis.

Proof of Lemma 16 We first prove Part (a). For any $\mathbf{w} \in \mathcal{W}, z \in \mathcal{Z}$, Eq. (2.2) implies

$$\|f'(\mathbf{w}, z)\|_2^2 = \big\|\ell'(\langle \mathbf{w}, \Phi(x)\rangle, y)\Phi(x)\big\|_2^2 \leq \kappa^2\Big(\widetilde{A}\ell(\langle \mathbf{w}, \Phi(x)\rangle, y) + \widetilde{B}\Big) = \kappa^2\big(\widetilde{A}f(\mathbf{w}, z) + \widetilde{B}\big).$$

We now turn to Part (b). For any $\mathbf{w} \in \mathcal{W}$ and $z \in \mathcal{Z}$, the definition of c_q implies

$$\begin{aligned} f(\mathbf{w}, z) = \ell(\langle \mathbf{w}, \Phi(x)\rangle, y) &\leq c_q\big(|\langle \mathbf{w}, \Phi(x)\rangle|^q + 1\big) \\ &\leq c_q\big(\|\mathbf{w}\|_2^q \kappa^q + 1\big) \leq \tilde{c}_q\big(\|\mathbf{w}\|_2^q + 1\big). \end{aligned}$$

The proof is complete. ■

Proof of Lemma 18 Let $\xi_i = f(\mathbf{w}_\lambda, z_i), i = 1, \ldots, n$. According to the definition of $\mathbf{w}_\lambda$, we know $\mathcal{E}(\mathbf{w}_\lambda) + \lambda\|\mathbf{w}_\lambda\|_2^2 \leq \mathcal{E}(0)$. It then follows that $\xi_i - \mathbb{E}[\xi_i] \leq \sup_z f(\mathbf{w}_\lambda, z)$ (non-negativity of ξ_i) and

$$\mathbb{E}\big[\big(\xi_i - \mathbb{E}[\xi_i]\big)^2\big] \leq \mathbb{E}[f^2(\mathbf{w}_\lambda, z_i)] \leq \sup_z f(\mathbf{w}_\lambda, z)\mathbb{E}[f(\mathbf{w}_\lambda, z)] \leq \sup_z f(\mathbf{w}_\lambda, z)\mathcal{E}(0).$$

Applying Part (b) of Lemma A.1 with $\xi_i = f(\mathbf{w}_\lambda, z_i)$ and the above bounds on variances and magnitudes, we derive the following inequality with probability at least $1 - \delta$

$$\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}(\mathbf{w}_\lambda) = \frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}[\xi] \leq \frac{\rho n \sup_z f(\mathbf{w}_\lambda, z)\mathcal{E}(0)}{n \sup_z f(\mathbf{w}_\lambda, z)} + \frac{\sup_z f(\mathbf{w}_\lambda, z)\log\frac{1}{\delta}}{\rho n}.$$

The stated inequality then follows directly from Part (b) of Lemma 16. ■

Proof of Lemma 20 We prove this lemma by McDiarmid's inequality (Lemma A.3). To this aim, we first show that the function $\mathbf{z} \mapsto \sup_{\mathbf{w}\in B_R}\big[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big]$ satisfies a bounded difference property. Indeed, for any $\mathbf{z} = \{z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n\}$ and $\bar{\mathbf{z}} = \{z_1, \ldots, z_{i-1}, \bar{z}_i, z_{i+1}, \ldots, z_n\}$, we have

$$\begin{aligned} \Big|\sup_{\mathbf{w}\in B_R}\big[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big] - \sup_{\mathbf{w}\in B_R}\big[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\bar{\mathbf{z}}}(\hat{h}_{\mathbf{w}})\big]\Big| &\leq \sup_{\mathbf{w}\in B_R}\big|\mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\bar{\mathbf{z}}}(\hat{h}_{\mathbf{w}})\big| \\ &\leq \frac{1}{n}\sup_{\mathbf{w}\in B_R}\big|\ell(\hat{h}_{\mathbf{w}}(x_i), y_i) - \ell(\hat{h}_{\mathbf{w}}(\bar{x}_i), \bar{y}_i)\big| \leq \frac{1}{n}\sup_{|y|,|\tilde{y}|\leq b}\ell(y, \tilde{y}). \end{aligned}$$

Applying McDiarmid's inequality with increments bounded above, we derive the following inequality with probability at least $1 - \delta$

$$\sup_{\mathbf{w}\in B_R}\big[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big] \leq \mathbb{E}_{\mathbf{z}}\Big[\sup_{\mathbf{w}\in B_R}\big[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big]\Big] + \sqrt{\frac{\log 1/\delta}{2n}}\sup_{|y|,|\tilde{y}|\leq b}\ell(y, \tilde{y}). \quad \text{(B.1)}$$

We now control the term $\mathbb{E}_{\mathbf{z}}\Big[\sup_{\mathbf{w}\in B_R}\big[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big]\Big]$. Let $\tilde{\mathbf{z}} = \{\tilde{z}_1, \ldots, \tilde{z}_n\}$ be training examples independently drawn from ρ and independent of $\mathbf{z}$. Let $\sigma_1, \ldots, \sigma_n$ be a

sequence of independent Rademacher variables with $\Pr\{\sigma_i = 1\} = \Pr\{\sigma_i = -1\} = \frac{1}{2}$. By Jensen's inequality and the standard symmetrization technique, we get

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\Big[\sup_{\mathbf{w}\in B_R}\big[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big]\Big] = \mathbb{E}_{\mathbf{z}}\Big[\sup_{\mathbf{w}\in B_R}\big[\mathbb{E}_{\tilde{\mathbf{z}}}[\mathcal{E}_{\tilde{\mathbf{z}}}(\hat{h}_{\mathbf{w}})] - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big]\Big] \\
&\le \mathbb{E}_{\mathbf{z},\tilde{\mathbf{z}}}\Big[\sup_{\mathbf{w}\in B_R}\big[\mathcal{E}_{\tilde{\mathbf{z}}}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big]\Big] = \frac{1}{n}\mathbb{E}_{\mathbf{z},\tilde{\mathbf{z}}}\Big[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^n\Big(\ell(\hat{h}_{\mathbf{w}}(\tilde{x}_i),\tilde{y}_i) - \ell(\hat{h}_{\mathbf{w}}(x_i),y_i)\Big)\Big] \\
&= \frac{1}{n}\mathbb{E}_{\mathbf{z},\tilde{\mathbf{z}},\sigma}\Big[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^n\sigma_i\Big(\ell(\hat{h}_{\mathbf{w}}(\tilde{x}_i),\tilde{y}_i) - \ell(\hat{h}_{\mathbf{w}}(x_i),y_i)\Big)\Big] \le \frac{2}{n}\mathbb{E}_{\mathbf{z},\sigma}\Big[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^n\sigma_i\ell(\hat{h}_{\mathbf{w}}(x_i),y_i)\Big].
\end{aligned} \tag{B.2}$$

Since $\big|\ell'(\hat{h}_{\mathbf{w}}(x),y)\big| \le C_9$, for any $\mathbf{w},\mathbf{w}' \in \mathcal{W}$ we know

$$\big|\ell(\hat{h}_{\mathbf{w}}(x),y) - \ell(\hat{h}_{\mathbf{w}'}(x),y)\big| \le C_9|\hat{h}_{\mathbf{w}}(x) - \hat{h}_{\mathbf{w}'}(x)| \le C_9\big|\langle \mathbf{w}-\mathbf{w}', \Phi(x)\rangle\big|.$$

Therefore, we can apply Talagrand's contraction lemma (Ledoux and Talagrand, 1991) to the last term of (B.2) to derive

$$\mathbb{E}_{\mathbf{z}}\Big[\sup_{\mathbf{w}\in B_R}\big[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big]\Big] \le \frac{2C_9}{n}\mathbb{E}_{\mathbf{z},\sigma}\Big[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^n\sigma_i\langle\mathbf{w},\Phi(x_i)\rangle\Big]. \tag{B.3}$$

According to the Schwarz's inequality and Jensen's inequality, we get

$$\begin{aligned}
\mathbb{E}_\sigma\Big[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^n\sigma_i\langle\mathbf{w},\Phi(x_i)\rangle\Big] &= \mathbb{E}_\sigma\Big[\sup_{\mathbf{w}\in B_R}\big\langle\mathbf{w},\sum_{i=1}^n\sigma_i\Phi(x_i)\big\rangle\Big] \le \mathbb{E}_\sigma\Big[\sup_{\mathbf{w}\in B_R}\|\mathbf{w}\|_2\sqrt{\Big\|\sum_{i=1}^n\sigma_i\Phi(x_i)\Big\|_2^2}\Big] \\
&\le R\sqrt{\mathbb{E}_\sigma\Big\langle\sum_{i=1}^n\sigma_i\Phi(x_i),\sum_{i=1}^n\sigma_i\Phi(x_i)\Big\rangle} = R\sqrt{\sum_{i=1}^n\|\Phi(x_i)\|_2^2} \le R\kappa\sqrt{n}.
\end{aligned}$$

Combining the above inequality, (B.1) and (B.3), we derive the following inequality with probability at least $1-\delta$

$$\sup_{\mathbf{w}\in B_R}\big[\mathcal{E}(\hat{h}_{\mathbf{w}}) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}})\big] \le \frac{2R\kappa C_9}{\sqrt{n}} + \sqrt{\frac{\log 1/\delta}{2n}}\sup_{|y|,|\tilde{y}|\le b}\ell(y,\tilde{y}),$$

which can be written as (5.1) with the C_7 defined below

$$C_7 = 2\kappa C_9 + 2^{-\frac{1}{2}}\sup_{|y|,|\tilde{y}|\le b}\ell(y,\tilde{y}).$$

The proof is complete. ■

Proof of Lemma 26 Introduce two sequences of random variables as follows

$$\xi_i = \ell(\hat{h}_{\mathbf{w}_\lambda}(x_i),y_i) - \ell(h_\rho(x_i),y_i), \quad \tilde{\xi}_i = \ell(h_{\mathbf{w}_\lambda}(x_i),y_i) - \ell(\hat{h}_{\mathbf{w}_\lambda}(x_i),y_i),$$

$i = 1, 2, \ldots, n$. It is clear that $|\xi_i| \leq \tilde{b}$. An application of Lemma A.2 together with $\mathbb{E}[\xi^2] \leq c_\beta \big(\mathbb{E}[\xi]\big)^\beta$ due to Assumption 5 then implies the following inequality with probability at least $1 - \delta$

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i - \mathbb{E}[\xi] \leq \frac{2\tilde{b}\log(1/\delta)}{3n} + \sqrt{\frac{2c_\beta(\mathcal{E}(\hat{h}_{\mathbf{w}_\lambda}) - \mathcal{E}(h_\rho))^\beta \log\frac{1}{\delta}}{n}}. \tag{B.4}$$

Furthermore, it follows from the Young's inequality for all $\mu, v \in \mathbb{R}, p^{-1} + q^{-1} = 1, p \geq 0$

$$\mu v \leq p^{-1}|\mu|^p + q^{-1}|v|^q \tag{B.5}$$

that

$$\big(\mathcal{E}(\hat{h}_{\mathbf{w}_\lambda}) - \mathcal{E}(h_\rho)\big)^{\frac{\beta}{2}} \Big(\frac{2c_\beta \log\frac{1}{\delta}}{n}\Big)^{\frac{1}{2}} \leq \frac{\beta}{2}\big(\mathcal{E}(\hat{h}_{\mathbf{w}_\lambda}) - \mathcal{E}(h_\rho)\big)^{\frac{\beta}{2}\frac{2}{\beta}} + \big(1 - \frac{\beta}{2}\big)\Big(\frac{2c_\beta \log\frac{1}{\delta}}{n}\Big)^{\frac{1}{2-\beta}}.$$

Plugging the above inequality into (B.4) establishes the following inequality with probability at least $1 - \delta$

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}_\lambda}) - \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}(\hat{h}_{\mathbf{w}_\lambda}) + \mathcal{E}(h_\rho) &= \frac{1}{n}\sum_{i=1}^{n} \xi_i - \mathbb{E}[\xi] \\ &\leq \frac{2\tilde{b}\log(1/\delta)}{3n} + \frac{\beta}{2}\big(\mathcal{E}(\hat{h}_{\mathbf{w}_\lambda}) - \mathcal{E}(h_\rho)\big) + \big(1 - \frac{\beta}{2}\big)\Big(\frac{2c_\beta \log\frac{1}{\delta}}{n}\Big)^{\frac{1}{2-\beta}}. \end{aligned} \tag{B.6}$$

For any $\mathbf{w} \in B_R$, it follows from Assumption 3 and Part (b) of Lemma 16 that

$$0 \leq \tilde{\xi}_i = \ell(h_{\mathbf{w}_\lambda}(x_i), y_i) - \ell(\hat{h}_{\mathbf{w}_\lambda}(x_i), y_i) \leq \ell(h_{\mathbf{w}_\lambda}(x_i), y_i) \leq \tilde{c}_q\big(\|\mathbf{w}_\lambda\|_2^q + 1\big).$$

Since $\tilde{\xi}_i, i = 1, \ldots, n$ are non-negative random variables, we know $\mathbb{E}[\tilde{\xi}^2] \leq \sup_z \tilde{\xi}(z)\mathbb{E}[\tilde{\xi}]$. An application of Lemma A.2 together with the above bound on $\tilde{\xi}_i$ then implies the following inequality with probability at least $1 - \delta$

$$\begin{aligned}&\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(\hat{h}_{\mathbf{w}_\lambda}) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(\hat{h}_{\mathbf{w}_\lambda}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\xi}_i - \mathbb{E}[\tilde{\xi}] \\ &\leq \frac{2\tilde{c}_q\big(\|\mathbf{w}_\lambda\|_2^q + 1\big)\log\frac{1}{\delta}}{3n} + \sqrt{\frac{2\tilde{c}_q\big(\|\mathbf{w}_\lambda\|_2^q + 1\big)\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(\hat{h}_{\mathbf{w}_\lambda})\big)\log\frac{1}{\delta}}{n}} \\ &\leq \frac{2\tilde{c}_q\big(\|\mathbf{w}_\lambda\|_2^q + 1\big)\log\frac{1}{\delta}}{3n} + \frac{\beta}{2}\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(\hat{h}_{\mathbf{w}_\lambda})\big) + \frac{\tilde{c}_q\big(\|\mathbf{w}_\lambda\|_2^q + 1\big)\log\frac{1}{\delta}}{n\beta}. \end{aligned} \tag{B.7}$$

Combining (B.6) and (B.7) together establishes the following inequality with probability at least $1 - \delta$

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\mathbf{w}_\lambda) - \mathcal{E}_{\mathbf{z}}(h_\rho) - \mathcal{E}(\mathbf{w}_\lambda) + \mathcal{E}(h_\rho) \leq{}& \frac{2\tilde{b}\log\frac{2}{\delta}}{3n} + \frac{\beta}{2}\big(\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho)\big) + \big(1 - \frac{\beta}{2}\big)\Big(\frac{2c_\beta \log\frac{2}{\delta}}{n}\Big)^{\frac{1}{2-\beta}} \\ &+ \frac{2\tilde{c}_q\big(\|\mathbf{w}_\lambda\|_2^q + 1\big)\log\frac{2}{\delta}}{3n} + \frac{\tilde{c}_q\big(\|\mathbf{w}_\lambda\|_2^q + 1\big)\log\frac{2}{\delta}}{n\beta}, \end{aligned}$$

which further implies the stated inequality with probability at least $1 - \delta$. ■

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

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

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

Peter Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Gilles Blanchard, Gábor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4(Oct):861–894, 2003.

Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36(2):489–531, 2008.

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

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(02):161–183, 2010.

Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with sgd and random features. In *Advances in Neural Information Processing Systems*, pages 10213–10224, 2018.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge university press, 2000.

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press, 2007.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279, 2019.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory*, pages 1579–1613, 2019.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer, Berlin, 1991.

Yunwen Lei and Ke Tang. Stochastic composite mirror descent: Optimal bounds with high probabilities. In *Advance in Neural Information Processing Systems*, pages 1524–1534, 2018.

Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *International Conference on Machine Learning*, pages 3098–3107, 2018.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(1):3375–3421, 2017.

Junhong Lin and Ding-Xuan Zhou. Online learning algorithms can converge comparably fast as batch learning. *IEEE Transactions on Neural Networks and Learning Systems*, 29 (6):2367–2378, 2018.

Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning*, pages 2340–2348, 2016a.

Junhong Lin, Lorenzo Rosasco, and Ding-Xuan Zhou. Iterative regularization for learning with convex loss functions. *Journal of Machine Learning Research*, 17(77):1–38, 2016b.

Ben London. A PAC-bayesian analysis of randomized learning with application to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2931–2940, 2017.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on Learning Theory*, pages 2294–2340, 2019.

Colin McDiarmid. On the method of bounded differences. In J. Siemous, editor, *Surveys in combinatorics*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

Nicole Mücke and Gilles Blanchard. Parallelizing spectrally regularized kernel algorithms. *The Journal of Machine Learning Research*, 19(1):1069–1097, 2018.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pages 12568–12577, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. In *Conference on Learning Theory*, pages 250–296, 2018a.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018b.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pages 449–456, 2012.

Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support vector machines, Regularization, Optimization, and Beyond*. MIT press, 2001.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Ingo Steinwart and Andreas Christmann. Fast learning from non-iid observations. In *Advances in Neural Information Processing Systems*, pages 1768–1776, 2009.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

Alexander Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5(Oct):1363–1390, 2004.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

Yiming Ying and Ding-Xuan Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788, 2006.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning*, pages 919–926, 2004.

Tong Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Conference on Learning Theory*, pages 173–187, 2005.