High-Probability Generalization Bounds for Pointwise Uniformly Stable Algorithms

Jun Fan[a], Yunwen Lei[b,*]

[a] *Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong, China*
[b] *Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong, China*

Abstract

Algorithmic stability is a fundamental concept in statistical learning theory to understand the generalization behavior of optimization algorithms. Existing high-probability bounds are developed for the generalization gap as measured by function values and require the algorithm to be uniformly stable. In this paper, we introduce a novel stability measure called pointwise uniform stability by considering the sensitivity of the algorithm with respect to the perturbation of each training example. We show this weaker pointwise uniform stability guarantees almost optimal bounds, and gives the first high-probability bound for the generalization gap as measured by gradients. Sharper bounds are given for strongly convex and smooth problems. We further apply our general result to derive improved generalization bounds for stochastic gradient descent. As a byproduct, we develop concentration inequalities for a summation of weakly-dependent vector-valued random variables.

Keywords:
Learning Theory, Algorithmic Stability, Stochastic Gradient Descent, Generalization Analysis

1. Introduction

How to understand the generalization behavior of a learning algorithm is a central problem in statistical learning theory. A popular approach to developing generalization bounds is based on the uniform convergence, which controls the uniform deviation between population risks and empirical risks over a function space [39, 2, 36, 9]. This approach ignores how an algorithm explores over the function space, and leads to generalization bounds depending on the complexity of function spaces such as VC dimension [39], covering numbers [45, 36] and Rademacher complexities [2].

An alternative approach for generalization analysis is based on a fundamental concept of algorithmic stability. Roughly speaking, we say a learning algorithm is algorithmically stable if a change of a single example in the training dataset brings only a small change in the output model, i.e., the algorithm is insensitive with respect to (w.r.t.) the perturbation of training datasets [32, 5]. Algorithmic stability was introduced in 1970s to derive leave-one-out bounds for certain nonparametric local learning algorithms (such as nearest-neighbor rules) [11, 32]. The modern framework of stability analysis

[*] Corresponding author
Email addresses: junfan@hkbu.edu.hk (Jun Fan), leiyw@hku.hk (Yunwen Lei)

was established in [5], where a celebrated concept called the uniform stability has been introduced to study regularization methods.

We need to answer two questions in applying algorithmic stability to get generalization bounds for an algorithm. The first question is how to guarantee the generalization by stability, i.e., whether a stable algorithm can always produce models with good generalization behavior. The second question is how to develop stability bounds for an algorithm in terms of algorithm parameters such as the regularization parameter, the step size and the number of iterations.

The second question is algorithm-dependent, which allows us to exploit the special property of algorithms to get bounds better than algorithm-independent bounds based on complexity measures [2]. The stability of various optimization algorithms has been developed in the literature. For example, the uniform stability has been developed for stochastic gradient descent (SGD) [17], which is one of the most widely used optimization methods to solve large-scale problems in machine learning.

For the first question, quantitative connection either in expectation or with high probability has been established. In particular, with probability at least $1-\delta$ the following generalization bounds were developed for β-uniformly stable algorithms[1] [6, 14]

$$\big|F(A(S)) - F_S(A(S))\big| \lesssim \beta \log n \log(1/\delta) + \log^{\frac{1}{2}}(1/\delta)/n^{\frac{1}{2}}, \tag{1.1}$$

where $A(S)$ denotes the output model by applying an algorithm A to the dataset S, $F(\mathbf{w})$ denotes the population risk of a model $\mathbf{w}$, $F_S(\mathbf{w})$ denotes the empirical risk of $\mathbf{w}$ (definitions are given in Section 3.1) and n is the sample size. Eq. (1.1) is a breakthrough result on the high-probability generalization analysis for uniformly stable algorithms initialized in 2002 [5]. However, some questions on Eq. (1.1) still remain.

- Eq. (1.1) provides generalization bounds in terms of function values. For nonconvex problems, optimization algorithms can only find a local minimizer and therefore we can only get optimization error bounds for $\|\nabla F(A(S))\|$ [15], where ∇ denotes the gradient operator. Therefore, it is interesting to study the generalization behavior of $A(S)$ as measured by $\nabla F(A(S))$, which motivates the question of developing high-probability bounds on the generalization gap as measured by gradients, i.e., $\|\nabla F(A(S)) - \nabla F_S(A(S))\|$.
- Eq. (1.1) requires the algorithm to be uniformly stable, which is arguably the strongest concept of algorithmic stability. Is it possible to relax this uniform stability to a weaker version of uniform stability, and can we develop better bounds on this weaker stability for popular algorithms such as SGD?
- The recent sharper generalization bounds in [18] require the loss function to be simultaneously Lipschitz continuous and λ-strongly convex, which cannot be satisfied globally due to the conflict

[1] We use the notation $\lesssim$ to ignore constant factors in an inequality.

between the Lipschitz continuity and strong convexity. Furthermore, their generalization bounds involve $\hat{\Delta}_\lambda^{\frac{1}{2}}$, where $\hat{\Delta}_\lambda = \lambda^{-1}\big(F_S(A(S)) - \min_{\mathbf{w}} F_S(\mathbf{w})\big)$ denotes a weighted suboptimality of the output model in terms of the empirical risk. This square-root dependency on $\hat{\Delta}_\lambda$ is slow in practice. Can we address the above conflict and improve the dependency on $\hat{\Delta}_\lambda$?

In this paper, we aim to provide affirmative answers to the above questions. Our main contributions are as follows.

- We develop a concentration inequality for a summation of weakly-dependent vector-valued random variables, which generalizes a similar result in Bousquet et al. [6] from real-valued random variables to random variables taking values in a Hilbert space.

- We introduce a new stability measure termed as the pointwise uniform stability. While this stability is weaker than the uniform stability, we show it guarantees high-probability generalization bounds on $F(A(S)) - F_S(A(S))$. We also give the first high-probability bound for $\|\nabla F(A(S)) - \nabla F_S(A(S))\|$ based on stability analysis.

- We improve the high-probability bound in [18] by considering a loss function of a structure, which reconciles the conflict between Lipschitz continuity and strong convexity. Furthermore, we derive a sharper bound involving $\hat{\Delta}_\lambda^{\frac{1+\alpha}{2}}$ to exploit the α-Hölder continuity of gradients. In particular, if $\alpha = 1$, the term $\hat{\Delta}_\lambda^{\frac{1+\alpha}{2}}$ decays quadratically faster than $\hat{\Delta}_\lambda^{\frac{1}{2}}$ in [18].

- We study the pointwise uniform stability of SGD for convex and strongly convex problems, covering smooth and nonsmooth problems. We then apply our connection between stability and generalization to give high-probability generalization bounds.

The paper is organized as follows. We review the related work in Section 2. We present our main results in Section 3, and give applications to SGD in Section 4. We present the proofs on connecting stability and generalization in Section 5, and the proofs on SGD in Section 6. The conclusion is given in Section 7. Some lemmas and proofs are given in the Appendix.

2. Related Work

2.1. Connection on Stability and Generalization

Algorithmic stability can imply generalization bounds in expectation and with high probability. We first consider generalization bounds in expectation. On-average stability can imply generalization under a Lipschitz condition of loss functions [34]. For non-Lipschitz problems, an on-average model stability was proposed to give generalization bounds by exploiting the smoothness of loss functions [22], which can further imply fast rates under a low-noise condition. On-average stability can imply generalization bounds for any learning algorithms to solve gradient-dominated problems [23, 7]. For nonconvex and smooth problems, generalization as measured by gradients can be guaranteed by stability in gradients [21].

We now consider generalization bounds with high probability. In a seminal paper [5], β-uniform stability was introduced to give bounds of order $O((\beta+1/n)\sqrt{n\log(1/\delta)})$, which was extended to randomized learning algorithms [12]. These results were significantly improved to $O(\sqrt{(\beta+1/n)\log(1/\delta)})$ in [13] by techniques in adaptive data analysis. Almost optimal generalization bounds in Eq. (1.1) were further derived by developing concentration inequalities for a summation of weakly-dependent random variables [6, 14]. The above-mentioned high-probability analysis can imply bounds of the order at most $O(1/\sqrt{n})$. Under a Bernstein condition on variances, it was shown that β-uniformly stable algorithms can enjoy high-probability bounds of the order $O((\beta\log n + 1/n)\log(1/\delta))$ [18].

2.2. Stability of Learning Algorithms

Algorithmic stability has been studied for various learning algorithms. Uniform stability bounds of order $O(1/(n\lambda))$ were developed for empirical risk minimization to solve λ-strongly convex problems [5]. In a seminal paper, uniform stability bounds of order $O(G^2\sum_{t=1}^T \eta_t/n)$ were developed for SGD with T iterations and step size sequences $\{\eta_t\}$ for convex, smooth and G-Lipschitz problems [17]. Data-dependent stability bounds reflecting the effect of initialization point were established for SGD [20]. For nonsmooth and convex problems, stability bounds of order $O(\eta\sqrt{T}+\eta T/n)$ were developed for SGD with $\eta_t=\eta$ either in expectation [22] or with high probability [3]. The Lipschitz constant G in the existing stability bounds [17] was replaced by the training error based on on-average model stability, which can imply fast excess risk bounds under a low-noise condition [22]. On-average model stability was also used to understand the benefit of overparameterization for shallow neural networks [30, 37, 24], and the implicit bias of gradient methods for separable data and self-bounding loss functions [33]. Other than the standard SGD/GD, the stability of differentially private SGD [42, 3, 21], gradient-free optimization methods [28], accelerated methods [41] and noisy SGD [46, 25, 38, 27] was studied in the literature. Lower bounds on the stability of gradient methods were also developed [3, 19, 1].

3. Main Results

3.1. Problem Setup

Let ρ be a probability measure defined on a sample space $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}$ is an input space and $\mathcal{Y}$ is an output space. Let $S=(z_1,\ldots,z_n)$ be a training dataset drawn independently from ρ, based on which we aim to find a model $h:\mathcal{X}\mapsto\mathcal{Y}$ for further prediction. We consider a parametric model, i.e., a model can be indexed by a parameter $\mathbf{w}\in\mathcal{W}$, where $\mathcal{W}\subset\mathbb{R}^d$ is the parameter space. The performance of a model $\mathbf{w}$ on an example z can be measured by $f(\mathbf{w};z)$, where $f:\mathcal{W}\times\mathcal{Z}\mapsto\mathbb{R}^+$ is the loss function. The empirical behavior of a model $\mathbf{w}$ can be quantified by the empirical risk $F_S(\mathbf{w})=\frac{1}{n}\sum_{i=1}^n f(\mathbf{w};z_i)$, while the prediction behavior can be quantified by the population risk $F(\mathbf{w})=\mathbb{E}_z[f(\mathbf{w};z)]$, where $\mathbb{E}_z[\cdot]$ denotes the expectation w.r.t. z. We often apply an algorithm A onto S to get a model $A(S)\in\mathcal{W}$ with a small empirical risk. However, this does not necessarily imply a small population risk referred to as the overfitting phenomenon. To this aim, we need to handle an important concept called the generalization gap $F(A(S)) - F_S(A(S))$, i.e., the difference between population risk and empirical risk at the output model $A(S)$. In this paper, we will leverage the celebrated concept called algorithmic stability to develop high-probability bounds on the generalization gap.

3.2. Concentration Inequality

We give a p-norm bound for a summation of weakly-dependent random variables taking values in a Hilbert space, whose proof is given in Section 5.1. It will play a fundamental role in deriving the connection between stability and generalization. The L_p-norm of a real-valued random variable Z is denoted by $\|Z\|_p := \left(\mathbb{E}[|Z|^p]\right)^{\frac{1}{p}}, p \geq 1$. Let $\|\cdot\|$ denote the norm in a Hilbert space $\mathcal{H}$. Then $\|\nabla f(\mathbf{w}; Z)\|$ is a real-valued random variable (as a function of Z). According to our notation, we have

$$\|\|\nabla f(\mathbf{w}; Z)\|\|_p = \Big(\mathbb{E}_Z\big[\|\nabla f(\mathbf{w}; Z)\|^p\big]\Big)^{\frac{1}{p}}, \quad \forall p \geq 1.$$

Theorem 1. *Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ be a sequence of independent random variables taking values in a Hilbert space $\mathcal{H}$. Let $g_1, \ldots, g_n$ be functions $g_i : \mathcal{Z}^n \mapsto \mathcal{H}$ such that the following holds.*

1. *For any $i \in [n]$, almost surely we have $\sup_{z_i} \|\mathbb{E}[g_i(\mathbf{Z})|Z_i = z_i]\| \leq M$.*

2. *For any $i \in [n]$, almost surely we have $\mathbb{E}[g_i(\mathbf{Z})|\mathbf{Z}_{[n]\setminus\{i\}} = (z_j)_{j\neq i}] = 0, \forall z_j \in \mathcal{Z}, j \neq i$.*

3. *For any $i \in [n]$, the following inequality holds*

$$\sup_{z_1,\ldots,z_n,z'_j : j\neq i} \big\|g_i(z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, z_n) - g_i(z_1, \ldots, z_{j-1}, z'_j, z_{j+1}, \ldots, z_n)\big\| \leq \beta_j. \tag{3.1}$$

Then, for any $p \geq 2$ we have

$$\Big\|\Big\|\sum_{i=1}^n g_i\Big\|\Big\|_p \leq 2(\sqrt{2}+1)M\sqrt{np} + 2(\sqrt{2}+1)p\lceil\log_2 n\rceil\Big(n\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}}.$$

Remark 1. If $\mathcal{H} = \mathbb{R}$, a similar bound was established in [6]. That is, let $\tilde{g}_1, \ldots, \tilde{g}_n$ be real-valued functions such that $\|\mathbb{E}[\tilde{g}_i(Z)|Z_i]\| \leq M$, $\mathbb{E}[\tilde{g}_i(Z)|Z_{[n]\setminus\{i\}}] = 0$ and

$$\sup_{z_j,z'_j} \big|\tilde{g}_i(z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, z_n) - \tilde{g}_i(z_1, \ldots, z_{j-1}, z'_j, z_{j+1}, \ldots, z_n)\big| \leq \beta.$$

Then, the following inequality was established for any $p \geq 2$ [6]

$$\Big\|\sum_{i=1}^n \tilde{g}_i(Z)\Big\|_p \leq 4M\sqrt{np} + 12\sqrt{2}pn\beta\lceil\log_2 n\rceil. \tag{3.2}$$

There are two differences between our result and Eq. (3.2). First, we extend the discussion in [6] from real-valued random variables to random variables taking values in a general Hilbert space, and slightly improve the constant factor. Second, the discussions [6] assume the change of j-th example in $\mathbf{z} = (z_1, \ldots, z_n)$ would lead to a change of value uniformly bounded by β. As a comparison, we allow different β_j for different $j \in [n]$. As we will show, this is useful for us to get a new generalization bound based on our pointwise uniform stability.

3.3. Stability and Generalization

Stability measures the sensitivity of an algorithm up to the perturbation of the training dataset by a single example. A very popular stability measure is the uniform stability, which considers the change of any single example of any training dataset by any $z \in \mathcal{Z}$.

Definition 1 (Uniform Stability)**.** Let A be an algorithm and $\beta > 0$.

1. We say A is β-uniformly-stable in *function values* if for all datasets S, S' such that S and S' differ by a single example, we have

$$\sup_z |f(A(S); z) - f(A(S'); z)| \leq \beta. \tag{3.3}$$

2. We say A is β-uniformly-stable in *gradients* if for all datasets S, S' such that S and S' differ by a single example, we have

$$\sup_z \|\nabla f(A(S); z) - \nabla f(A(S'); z)\| \leq \beta. \tag{3.4}$$

In this paper, we introduce a new stability measure which we call the pointwise uniform stability. The basic idea is to give a single stability parameter β_i for perturbing the i-th example of the dataset. It is clear that if A is β-uniformly stable, then it is also $(\beta, \ldots, \beta)$-pointwise uniformly stable. Therefore, pointwise uniform stability is weaker than the uniform stability. In this paper, we will show that this weaker stability can also imply high-probability generalization bounds. We say two datasets S and $S^{(i)}$ differ only by the i-th example if $S = (z_1, \ldots, z_n)$ and $S^{(i)} = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)$ for some $z_i' \in \mathcal{Z}$.

Definition 2 (Pointwise Uniform Stability)**.** Let A be an algorithm and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n), \beta_i > 0$.

1. We say A is $\boldsymbol{\beta}$-pointwise uniformly-stable *in function values* if for all $S, S^{(i)}$ such that S and $S^{(i)}$ differ by the i-th example, we have

$$\sup_z |f(A(S); z) - f(A(S^{(i)}); z)| \leq \beta_i. \tag{3.5}$$

2. We say A is $\boldsymbol{\beta}$-pointwise uniformly-stable *in gradients* if for all $S, S^{(i)}$ such that S and $S^{(i)}$ differ by the i-th example, we have

$$\sup_z \|\nabla f(A(S); z) - \nabla f(A(S^{(i)}); z)\| \leq \beta_i. \tag{3.6}$$

Theorem 2 gives a high-probability bound on the generalization gap $F(A(S)) - F_S(A(S))$ for pointwise uniformly stable algorithms. We omit the proof due to its similarity with Theorem 3.

Theorem 2 (Generalization via Function Values)**.** *Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$. Consider an algorithm A and $\delta \in (0, 1)$. Assume for any S and any z, $|f(A(S); z)| \leq M$. If A is $\boldsymbol{\beta}$-pointwise uniformly-stable in function values, then the following inequality holds with probability at least $1 - \delta$*

$$\big|F(A(S)) - F_S(A(S))\big| \lesssim \frac{M \log^{\frac{1}{2}}(1/\delta)}{\sqrt{n}} + \Big(\frac{1}{n}\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}} \log n \log(1/\delta).$$

Remark 2. If A is β-uniformly stable in function values, the generalization bound in Eq. (1.1) was developed [5, 14]. As a comparison, our bound involves an average of stability parameters over all indices, i.e., the term $\big(\frac{1}{n}\sum_{i=1}^n \beta_i^2\big)^{\frac{1}{2}}$, which is smaller than the uniform stability parameter $\beta = \max_i \beta_i$ considered in [5, 14]. As we will show, for SGD we can establish a bound for $\big(\frac{1}{n}\sum_{i=1}^n \beta_i^2\big)^{\frac{1}{2}}$ which is smaller than that for $\max_i \beta_i$.

Our next result is a high-probability bound on the generalization gap in terms of gradients, which extends the high-probability generalization bound in function values in [5, 14]. We show that the deviation between population gradients and empirical gradients at the output model can be bounded by the stability parameter in gradients. We require f to be differentiable, and do not require a convexity or smoothness assumption in Theorem 3. The proof is given in Section 5.2.

Theorem 3 (Generalization via Gradients)**.** *Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$. Consider an algorithm A and $\delta \in (0,1)$. Assume for any S and any z, $\|\nabla f(A(S);z)\| \le M$. If A is $\boldsymbol{\beta}$-pointwise uniformly-stable in gradients, then the following inequality holds with probability at least $1-\delta$*

$$\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\| \lesssim \frac{M \log^{\frac{1}{2}}(1/\delta)}{\sqrt{n}} + \Big(\frac{1}{n}\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}} \log n \log(1/\delta).$$

Remark 3. If A is β-uniformly stable in gradients with, then it was shown [21]

$$\mathbb{E}\big[\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|\big] \lesssim \beta + \sqrt{\frac{1}{n}\mathbb{E}\big[\mathbb{V}_Z[f(A(S);Z)]\big]},$$

where $\mathbb{V}_Z[f(A(S);Z)]$ is the variance of $\nabla f(A(S);Z)$ as a function of Z. This bound was established in expectation. As a comparison, we develop high-probability bounds on the generalization gap between population and empirical gradients. High-probability bounds of order $\sqrt{d\log(1/\delta)/n}$ were also established for $\sup_{\mathbf{w}} \|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|$ based on complexity measures of function spaces, which, however, depend on the dimensionality d of the problem and are not appealing for high-dimensional learning problems. As a comparison, our stability analysis implies dimension-free generalization bounds.

3.4. Sharper Generalization Bounds

Theorem 2 implies generalization bounds of the order $O(1/\sqrt{n})$. In this section, we improve this dependency to $O(1/n)$ for pointwise uniformly stable algorithms. The following theorem is an extension of the stability analysis in [18]. We consider functions with a composite structure.

Definition 3 (Lipschitzness, Smoothness and Convexity)**.** Let $G, L_\alpha, L > 0, \lambda \ge 0$ and $g : \mathcal{W} \mapsto \mathbb{R}$.

- We say g is G-Lipschitz continuous if $|g(\mathbf{w}) - g(\mathbf{w}')| \le G\|\mathbf{w} - \mathbf{w}'\|, \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$.
- We say g has (α, L_α)-Hölder continuous gradients ($\alpha \in [0,1]$) if

$$\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\| \le L_\alpha \|\mathbf{w} - \mathbf{w}'\|^\alpha, \ \ \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}.$$

We say g is L-smooth if g has $(1, L)$-Hölder continuous gradients.

- We say g is λ-strongly convex if

$$g(\mathbf{w}) \geq g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}') \rangle + \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}'\|^2, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}.$$

  We say g is convex if the above inequality holds with $\lambda = 0$.

Assumption 1. Let $\lambda > 0, \ell : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$ and $r : \mathcal{W} \mapsto \mathbb{R}_+$. Assume $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$ has the following structure

$$f(\mathbf{w}; z) = \ell(\mathbf{w}; z) + r(\mathbf{w}). \tag{3.7}$$

Assume for any z, the function $\mathbf{w} \mapsto \ell(\mathbf{w}; z)$ is nonnegative and has (α, L_α)-Hölder continuous gradients. Assume r is L_r-smooth, and for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is λ-strongly convex.

For non-composite problems, our analysis can still imply faster rates if f is strongly convex, smooth and $\|\nabla f(A(S); z)\| \leq G, \|\nabla f(A_{\mathrm{e}}(S); z)\| \leq G$, where we denote by A_{e} the empirical risk minimization (ERM) algorithm, i.e.,

$$A_{\mathrm{e}}(S) = \arg\min_{\mathbf{w} \in \mathcal{W}} F_S(\mathbf{w}).$$

Let $L_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{w}; z_i)$ and $L(\mathbf{w}) = \mathbb{E}_z[\ell(\mathbf{w}; z)]$. Let $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w})$ be a minimizer of the population risk. The proof is given in Section C.

Theorem 4. *Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ and f take a structure in Eq.* (3.7)*. Assume A is $\boldsymbol{\beta}$-pointwise uniformly-stable in function values (measured by ℓ), i.e., Eq.* (3.5) *holds with f replaced by ℓ. Let $M = \sup_z \big|\mathbb{E}_S[\ell(A(S)); z] - \ell(\mathbf{w}^*; z)\big|$. Then for any $\delta \in (0, 1)$, the following inequality holds with probability at least $1 - \delta$*

$$F(A(S)) - F_S(A(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*) \lesssim \Big(\frac{1}{n} \sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}} \log n \log(1/\delta) + \frac{M \log \frac{1}{\delta}}{n} + \Big(\frac{\sigma_A^2 \log(1/\delta)}{n}\Big)^{\frac{1}{2}},$$

where

$$\sigma_A^2 = \mathbb{E}_Z\Big[\Big(\mathbb{E}_S\big[\ell(A(S); Z)\big] - \ell(\mathbf{w}^*; Z)\Big)^2\Big] - \Big(\mathbb{E}_S\big[L(A(S))\big] - L(\mathbf{w}^*)\Big)^2.$$

Remark 4. A key difference between Theorem 4 and Theorem 2 is that the term $n^{-\frac{1}{2}} M \log^{\frac{1}{2}}(1/\delta)$ in Theorem 2 is replaced by $n^{-1} M \log(1/\delta)$ in Theorem 4, at the cost of introducing $\sigma_A n^{-\frac{1}{2}} \log^{\frac{1}{2}}(1/\delta)$. Then, Theorem 4 can imply fast excess risk bounds if the variance σ_A^2 is small. Similar bounds were derived in [18] under the following Bernstein assumption

$$\mathbb{E}_Z\big[(f(\mathbf{w}; Z) - f(\mathbf{w}^*; Z))^2\big] \leq B(F(\mathbf{w}) - F(\mathbf{w}^*)), \quad \forall \mathbf{w} \in \mathcal{W}. \tag{3.8}$$

The bound in [18] involves the uniform stability. As a comparison, our analysis uses the pointwise uniform stability. Furthermore, we do not impose a Bernstein assumption, and instead include the variance term σ_A^2 in the upper bound. Finally, we consider a problem with a composite structure and our stability assumption is imposed to ℓ instead of f. The underlying reason is that it is possible ℓ is Lipschitz continuous but f not. In this case, if we can derive a bound on $\|A(S) - A(S^{(i)})\|$, we can use the Lipschitz continuity of ℓ to get a bound on $\ell(A(S); z) - \ell(A(S^{(i)}); z)$ but not a bound on $f(A(S); z) - f(A(S^{(i)}); z)$. As a comparison, the analysis in [18] does not consider this composite structure.

To apply Theorem 4, we need to estimate the variance term σ_A^2, which can be related to the excess risk $F(A(S))-F(\mathbf{w}^*)$. In the following theorem to be proved in Section 5.3, we show that the Bernstein condition holds if Assumption 1 holds. Furthermore, we also give generalization bounds in expectation, which involves optimization error $F_S(A(S))-F_S(A_{\rm e}(S))$ and the strong convexity parameter λ. Define

$$c_\alpha = \begin{cases} (1+1/\alpha)^{\frac{\alpha}{1+\alpha}} L_\alpha^{\frac{1}{1+\alpha}}, & \text{if } \alpha \in (0,1], \\ \sup_z \|\nabla \ell(0;z)\| + L_\alpha, & \text{if } \alpha = 0. \end{cases} \tag{3.9}$$

The proof is given in Section D.

Lemma 5. *Let Assumption 1 hold. Then*

$$\sigma_A^2 \le C\lambda^{-1}\mathbb{E}_S[F(A(S)) - F(\mathbf{w}^*)], \tag{3.10}$$

where

$$C = 2c_\alpha^2 \mathbb{E}_{S,Z}\big[\max\big\{\ell^{\frac{2\alpha}{1+\alpha}}(A(S);Z), \ell^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*;Z)\big\}\big]. \tag{3.11}$$

Furthermore, if $\big(F(A_e(S)) - F(\mathbf{w}^*)\big)^{\frac{1-\alpha}{1+\alpha}} \le \widetilde{C}$ *for some* $\widetilde{C} > 0$ *independent of* n *or* λ*, then any algorithm* A *satisfies*

$$\mathbb{E}\big[F(A(S))\big] - F(\mathbf{w}^*) \le \mathfrak{C}\Big(\Delta_\lambda^{\frac{1+\alpha}{2}} + \Delta_\lambda + \nabla_\lambda\Big), \tag{3.12}$$

where $\mathfrak{C}$ *is a constant independent of* λ *or* n *(may depend on* α, L_α, L_r *and is explicitly given in Eq.* (D.5)*) and*

$$\Delta_\lambda = \lambda^{-1}\mathbb{E}\big[F_S(A(S)) - F_S(A_e(S))\big], \qquad \nabla_\lambda = \frac{1}{n\lambda}\mathbb{E}\Big[L_S^{\frac{2\alpha}{1+\alpha}}(A_e(S)) + L^{\frac{2\alpha}{1+\alpha}}(A_e(S))\Big].$$

The assumption $\big(F(A_{\rm e}(S)) - F(\mathbf{w}^*)\big)^{\frac{1-\alpha}{1+\alpha}} \le \widetilde{C}$ is introduced just for simplifying the analysis, and can be removed with more complicated computation. This assumption holds automatically if $\alpha = 1$. We can combine Eq. (3.10) and Eq. (3.12) to derive

$$\sigma_A^2 \le 2c_\alpha^2 \mathfrak{C}(\mathbb{E}[F(A(S))])^{\frac{2\alpha}{1+\alpha}} \Big(\frac{1}{\lambda^{1+\frac{1+\alpha}{2}}}\Big(\mathbb{E}\big[F_S(A(S)) - F_S(A_{\rm e}(S))\big]\Big)^{\frac{1+\alpha}{2}} + \frac{\mathbb{E}\big[F_S(A(S)) - F_S(A_{\rm e}(S))\big]}{\lambda^2} + \frac{2\big(\mathbb{E}\big[L(A_{\rm e}(S))\big]\big)^{\frac{2\alpha}{1+\alpha}}}{n\lambda^2}\Big),$$

where we have used the Jensen's inequality. We can plug the above bound back into Theorem 4, and get the following high-probability bound. We omit the proof for simplicity. For simplicity, we assume $\Delta_\lambda = O(1)$ and absorb all constant factors independent of β_i, n, λ (e.g., α, L_α, L_r) into the $\lesssim$ notation.

Corollary 6. *Let Assumptions in Lemma 5 hold. Let* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ *and assume* A *is* $\boldsymbol{\beta}$*-pointwise uniformly-stable in function values (measured by* ℓ*). Let* $M = \sup_z \big|\mathbb{E}_S[\ell(A(S)); z] - \ell(\mathbf{w}^*; z)\big|$*. Then for any* $\delta \in (0,1)$*, the following inequality holds with probability at least* $1-\delta$

$$F(A(S)) - F_S(A(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*) \lesssim \Big(\frac{1}{n}\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}} \log n \log(1/\delta) + \frac{M\log\frac{1}{\delta}}{n} + \frac{\log^{1/2}(1/\delta)(\mathbb{E}[F(A(S))])^{\frac{\alpha}{1+\alpha}}}{\sqrt{n}}\Big(\frac{1}{\lambda^{\frac{3+\alpha}{4}}}\Big(\mathbb{E}\big[F_S(A(S)) - F_S(A_e(S))\big]\Big)^{\frac{1+\alpha}{4}} + \frac{\big(\mathbb{E}\big[L(A_e(S))\big]\big)^{\frac{\alpha}{1+\alpha}}}{\sqrt{n}\lambda}\Big).$$

If we further assume ℓ is Lipschitz continuous, we can have the following high-probability bounds for any algorithm to solve strongly convex problems. The proof is given in Section 5.3.

Theorem 7. *Let Assumptions in Lemma 5 hold and ℓ be G-Lipschitz continuous. If* $\sup_z \big|\mathbb{E}_S[\ell(A(S)); z] - \ell(\mathbf{w}^*; z)\big| < \infty$, *then for any $\delta \in (0,1)$ with probability at least $1-\delta$ we have*

$$F(A(S)) - F(\mathbf{w}^*) \lesssim (n\lambda)^{-1} \log n \log(1/\delta) + \hat{\Delta}_\lambda^{\frac{1+\alpha}{2}},$$

where $\hat{\Delta}_\lambda = \lambda^{-1}\big(F_S(A(S)) - F_S(A_e(S))\big)$.

Remark 5. The term $F_S(A(S)) - F_S(A_{\mathrm{e}}(S))$ is the optimization error, which measures the suboptimality of $A(S)$ to the minimal empirical risk. The recent work [18] gives the following high probability bound if F_S is λ-strongly convex and f is Lipschitz continuous

$$F(A(S)) - F(\mathbf{w}^*) \lesssim \Big(\frac{1}{n\lambda} + \bar{\Delta}_\lambda^{\frac{1}{2}}\Big) \log n \log(1/\delta), \tag{3.13}$$

where $\bar{\Delta}_\lambda$ is a deterministic number and an upper bound of $\hat{\Delta}_\lambda$. However, a strongly convex function cannot be Lipschitz continuous in the whole region. Therefore, the strong convexity assumption is contradictory to the Lipschitz condition. As a comparison, we consider an objective with a composite structure where ℓ has α-Hölder continuous gradients and is Lipschitz continuous. Our assumption is satisfied by various machine learning problems. For example, for logistic regression we have

$$f(\mathbf{w}; z) = \log(1 + \exp(-y\mathbf{w}^\top x)) + 2^{-1}\lambda\|\mathbf{w}\|^2,$$

which satisfies Assumption 1 with $\alpha = 1$. Moreover, the function $z \mapsto \log(1+\exp(-y\mathbf{w}^\top x))$ is Lipschitz continuous.

Furthermore, we show that the term $\bar{\Delta}_\lambda^{\frac{1}{2}}$ in Eq. (3.13) can be replaced by a faster-decaying term $\hat{\Delta}_\lambda^{\frac{1+\alpha}{2}}$. In particular, if ℓ is smooth, we have $\hat{\Delta}_\lambda^{\frac{1+\alpha}{2}} = \hat{\Delta}_\lambda$, which decays quadratically faster than $\bar{\Delta}_\lambda^{\frac{1}{2}}$ in Eq. (3.13). This shows that we can stop the algorithm earlier if we impose a stronger assumption on the smoothness, and shows the benefit of smoothness in improving the generalization. Indeed, the analysis in [18] first shows that the algorithm A is β-uniformly stable with $\beta = 4G^2/(\lambda n) + \sqrt{8G^2\bar{\Delta}_\lambda}$. Then, they apply the high-probability bound on uniform stability to A and give the bound in Eq. (3.13). Since a smoothness assumption would not affect the uniform stability, the uniform stability parameter there involves $\bar{\Delta}_\lambda^{\frac{1}{2}}$, and the strategy fails to use the smoothness assumption to improve the bound. We take a different strategy. We apply Theorem 4 to the algorithm A_{e} to first give a bound on $F(A_{\mathrm{e}}(S)) - F_S(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*)$, which does not involve $\bar{\Delta}_\lambda$ since A_{e} outputs the ERM model. Then we control $F(A(S)) - A_{\mathrm{e}}(S))$ in terms of $\hat{\Delta}_\lambda$, and use the smoothness assumption to show this bound improves as f is becoming more and more smooth. Finally, Eq. (3.13) requires $\hat{\Delta}_\lambda$ to be upper bounded by a deterministic number $\bar{\Delta}_\lambda$. As a comparison, our result directly involves $\hat{\Delta}_\lambda$.

4. Applications to Stochastic Gradient Descent

In this section, we apply our connection between stability and generalization to derive generalization bounds for SGD.

Definition 4 (SGD)**.** Let $\mathbf{w}_1 \in \mathcal{W}$ and $\{\eta_t\}$ be a sequence of positive step sizes. At each iteration, we first randomly select an index j_t according to the uniform distribution over $[n]$ and update the model as follows

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

4.1. Stability Bounds

We first develop the pointwise uniform stability bounds for SGD. We consider three classes of problems: convex and smooth problems, convex and nonsmooth problems, and strongly convex and smooth problems. Let $\mathbb{I}_{[E]}$ be the indicator function, i.e., $\mathbb{I}_{[E]} = 1$ if the event E happens, and 0 otherwise. The proofs are given in Section 6.

Theorem 8 (Stability of SGD: Smooth Case)**.** *Assume $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}^+$ is convex, L-smooth and G-Lipschitz. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be produced by SGD with $\eta_t = \eta \leq 2/L$. Then SGD with T iterations is $\boldsymbol{\beta}$-pointwise uniformly stable in function values, where $\frac{1}{n}\sum_{i=1}^n \beta_i = \frac{2G^2T\eta}{n}$ and*

$$\frac{1}{n}\sum_{i=1}^n \beta_i^2 = \frac{4G^4\eta^2}{n}\sum_{i=1}^n \Big(\sum_{t=1}^T \mathbb{I}_{[j_t=i]}\Big)^2. \tag{4.1}$$

Remark 6. Under the same condition, one can show that SGD is β_{unif}-uniformly stable in function values with (implicitly shown in the proof of Theorem 8)

$$\beta_{\text{unif}} = 2G^2\eta \max_{i\in[n]} \sum_{t=1}^T \mathbb{I}_{[j_t=i]}. \tag{4.2}$$

To see the comparison between the uniform stability bound in Eq. (4.2) and the pointwise stability bound in Eq. (4.1), we introduce

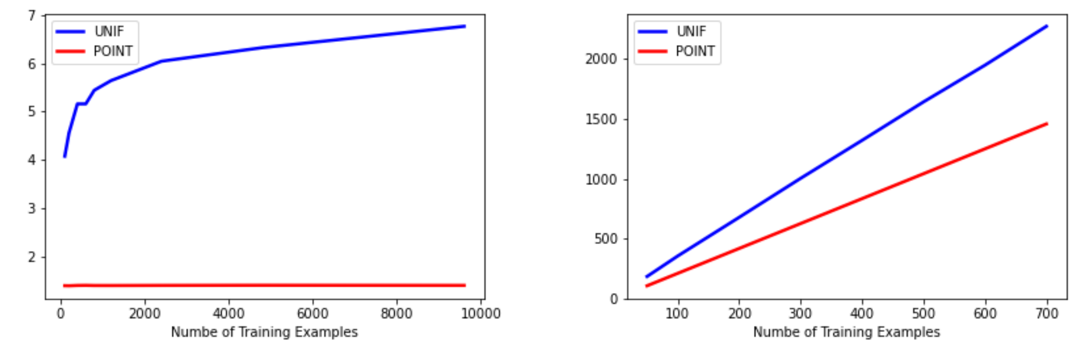
$$\tilde{\beta}_{\text{unif}} = \max_{i\in[n]} \sum_{t=1}^T \mathbb{I}_{[j_t=i]}, \quad \tilde{\beta}_{\text{point}} = \Big(\frac{1}{n}\sum_{i=1}^n \Big(\sum_{t=1}^T \mathbb{I}_{[j_t=i]}\Big)^2\Big)^{\frac{1}{2}}. \tag{4.3}$$

It is clear that $\tilde{\beta}_{\text{unif}}, \tilde{\beta}_{\text{point}}$ differ from the above uniform/pointwise stability bounds by a factor of $2G^2\eta$. For simplicity, we set $T = n$ as this implies the optimal excess risk bounds [17]. Then, we have

$$\tilde{\beta}_{\text{point}} \leq \Big(\frac{1}{n}\sum_{i=1}^n \Big(\sum_{t=1}^n \mathbb{I}_{[j_t=i]}\Big) \max_{i\in[n]}\Big(\sum_{t=1}^n \mathbb{I}_{[j_t=i]}\Big)\Big)^{\frac{1}{2}} = \Big(\frac{1}{n}\sum_{t=1}^n\sum_{i=1}^n \mathbb{I}_{[j_t=i]}\Big)^{\frac{1}{2}} \tilde{\beta}_{\text{unif}}^{\frac{1}{2}} = \tilde{\beta}_{\text{unif}}^{\frac{1}{2}}, \tag{4.4}$$

where we have used the identity $\sum_{i=1}^n \mathbb{I}_{[j_t=i]} = 1$ for any t. The term $\tilde{\beta}_{\text{unif}}$ is related to the balls and bins problem [29]. It was shown that with probability at least $1 - 1/n$, $\tilde{\beta}_{\text{unif}} = \Theta\big(\frac{\log n}{\log\log n}\big)$ [29]. Then, by Eq. (4.4), with probability at least $1 - 1/n$ we have $\tilde{\beta}_{\text{point}} = O\big(\frac{\log^{\frac{1}{2}} n}{(\log\log n)^{\frac{1}{2}}}\big)$. Note Eq. (4.4) is not tight, and we expect that $\tilde{\beta}_{\text{point}}$ has a tighter upper bound. For example, we can show that the second moment of $\tilde{\beta}_{\text{point}}$ is bounded by a constant independent of n:

$$\begin{aligned}\mathbb{E}[\tilde{\beta}_{\text{point}}^2] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\big(\sum_{t=1}^n \mathbb{I}_{[j_t=i]}\big)^2\big] = \frac{1}{n}\sum_{i=1}^n\sum_{t=1}^n \mathbb{E}[\mathbb{I}^2_{[j_t=i]}] + \frac{1}{n}\sum_{i=1}^n \sum_{t\neq t'\in[n]} \mathbb{E}[\mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_{t'}=i]}] \\ &= \frac{1}{n}\sum_{t=1}^n\sum_{i=1}^n \mathbb{E}[\mathbb{I}_{[j_t=i]}] + \frac{1}{n}\sum_{i=1}^n \sum_{t\neq t'\in[n]} \mathbb{E}[\mathbb{I}_{[j_t=i]}]\mathbb{E}[\mathbb{I}_{[j_{t'}=i]}] = 1 + \frac{n^2-n}{n^2} \leq 2.\end{aligned}$$



(a) Smooth case

(b) Nonsmooth case

Figure 1: $\tilde{\beta}_{\text{unif}}$ (blue curve) and $\tilde{\beta}_{\text{point}}$ (red curve) as a function of n for SGD applied to convex and Lipschitz problems. Left panel considers the smooth case with $T = n$, where $\tilde{\beta}_{\text{unif}}$ and $\tilde{\beta}_{\text{point}}$ are defined in Eq. (4.3). Right panel considers the nonsmooth case with $T = n^2$, where $\tilde{\beta}_{\text{unif}}$ and $\tilde{\beta}_{\text{point}}$ are defined in Eq. (4.6).

As a comparison, $\mathbb{E}[\tilde{\beta}_{\text{unif}}] = \Theta\big(\frac{\log n}{\log\log n}\big)$ [29], which grows as n increases. We perform a simulation to compare $\tilde{\beta}_{\text{unif}}$ and $\tilde{\beta}_{\text{point}}$. We set $T = n$, and get a sequence of indices $\{j_t\}_{t\in[T]}$ by drawing j_t from the uniform distribution over $[n]$. Then, we compute $\tilde{\beta}_{\text{unif}}$ and $\tilde{\beta}_{\text{point}}$ according to Eq. (4.3). We repeat the experiments 25 times, and report the average of the experimental results. In Figure 1 (left panel), we plot $\tilde{\beta}_{\text{unif}}$ and $\tilde{\beta}_{\text{point}}$ as functions of n. The plot shows that $\tilde{\beta}_{\text{unif}}$ is substantially larger than $\tilde{\beta}_{\text{point}}$, and the difference grows as n increases. This shows the benefit of using pointwise uniform stability to study generalization.

Remark 7. Recently, fast excess risk bounds were derived for SGD based on the on-average model stability in the realizable (low-noise) setting [22]. Their bounds are stated in expectation, and their key idea is to incorporate the empirical risk in the stability bounds by using the expectation over S. For example, for SGD in a convex and smooth case, we can build the following inequality for two datasets $S, S^{(i)}$ differing by the i-th example

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\| \leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\| + \eta_t \mathbb{I}_{[j_t=i]}\Big(\|\nabla f(\mathbf{w}_t; z_i)\| + \|\nabla f(\mathbf{w}_t^{(i)}; z_i')\|\Big), \tag{4.5}$$

where z_i and z_i' are respectively the i-th example in S and $S^{(i)}$, and $\{\mathbf{w}_t\}, \{\mathbf{w}_t^{(i)}\}$ are SGD iterates on S and $S^{(i)}$, respectively. Then, the self-bounding property of smooth functions and the symmetry between z_i and z_i' imply

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|] \leq \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|] + \frac{\sqrt{2L}\eta_t}{n}\mathbb{E}\big[f^{\frac{1}{2}}(\mathbf{w}_t; z_i) + f^{\frac{1}{2}}(\mathbf{w}_t^{(i)}; z_i')\big] = \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|] + \frac{2\sqrt{2L}\eta_t}{n}\mathbb{E}\big[f^{\frac{1}{2}}(\mathbf{w}_t; z_i)\big].$$

An average over $i \in [n]$ further includes the empirical risks in the stability bounds

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|] + \frac{2\sqrt{2L}\eta_t}{n}\mathbb{E}[F_S^{\frac{1}{2}}(\mathbf{w}_t)],$$

which implies fast rates if $F_S(\mathbf{w}_t)$ are small.

As a comparison, the pointwise uniform stability takes a supremum over all neighboring datasets, and this supremum comes from the bounded increment condition in Eq. (A.1), which takes the supremum over all z_j. Then, we need to take supremum over S on both sides of Eq. (4.5) to get

$$\sup_{S,z_i'}\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|\leq\sup_{S,z_i'}\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|+\eta_t\mathbb{I}_{[j_t=i]}\sup_{S,z_i'}\Big(\|\nabla f(\mathbf{w}_t;z_i)\|+\|\nabla f(\mathbf{w}_t^{(i)};z_i')\|\Big),$$

from which we fail to incorporate empirical risks in the stability bounds for a fast rate.

We now consider the convex and nonsmooth case. The following theorem shows that the stability of SGD in the nonsmooth case is worse than that in the smooth case.

Theorem 9 (Stability of SGD: Nonsmooth Case)**.** *Assume $f:\mathcal{W}\times\mathcal{Z}\mapsto\mathbb{R}^+$ is convex and G-Lipschitz. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be produced by SGD with $\eta_t=\eta$. Then SGD with T iterations is $\boldsymbol{\beta}$-pointwise uniformly-stable in function values, where*

$$\frac{1}{n}\sum_{i=1}^n\beta_i^2\leq\frac{4G^4\eta^2}{n}\Big(Tn+4(T+1)^{\frac{3}{2}}/3+4\sum_{i=1}^n\sum_{t=1}^T\sum_{k=1}^t\mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_k=i]}\Big).$$

Remark 8. As we will show in the proof of Theorem 9, we can show that SGD is β_{unif}-uniformly stable with

$$\beta_{\mathrm{unif}}\leq 2G^2\eta\sqrt{T}+4G^2\eta\max_{i\in[n]}\sum_{t=1}^T\mathbb{I}_{[j_t=i]}.$$

Analogous to Remark 6, we introduce

$$\tilde{\beta}_{\mathrm{unif}}=\sqrt{T}+2\max_{i\in[n]}\sum_{t=1}^T\mathbb{I}_{[j_t=i]},\quad\tilde{\beta}_{\mathrm{point}}=\Big(T+\frac{4(T+1)^{\frac{3}{2}}}{3n}+\frac{4}{n}\sum_{i=1}^n\sum_{t=1}^T\sum_{k=1}^t\mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_k=i]}\Big)^{\frac{1}{2}}.\tag{4.6}$$

It is clear that $\tilde{\beta}_{\mathrm{unif}},\tilde{\beta}_{\mathrm{point}}$ differ from the above uniform/pointwise stability bounds by a factor of $2G^2\eta$. Since $\sum_{t=1}^T\sum_{k=1}^{t-1}\mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_k=i]}=\sum_{k=1}^T\sum_{t=k+1}^T\mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_k=i]}=\sum_{t=1}^T\sum_{k=t+1}^T\mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_k=i]}$, we know

$$\begin{aligned}2\sum_{i=1}^n\sum_{t=1}^T\sum_{k=1}^t\mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_k=i]}&=\sum_{i=1}^n\sum_{t=1}^T\mathbb{I}^2_{[j_t=i]}+\sum_{i=1}^n\sum_{t=1}^T\sum_{k=1}^T\mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_k=i]}=T+\sum_{i=1}^n\Big(\sum_{t=1}^T\mathbb{I}_{[j_t=i]}\Big)^2\\&\leq T+\sum_{i=1}^n\Big(\sum_{t=1}^T\mathbb{I}_{[j_t=i]}\Big)\max_{i\in[n]}\sum_{t=1}^T\mathbb{I}_{[j_t=i]}=T+T\max_{i\in[n]}\sum_{t=1}^T\mathbb{I}_{[j_t=i]}.\end{aligned}$$

It then follows that

$$\tilde{\beta}_{\mathrm{point}}\leq\Big(T+\frac{4(T+1)^{\frac{3}{2}}}{3n}+\frac{2T}{n}+\frac{2T}{n}\max_{i\in[n]}\sum_{t=1}^T\mathbb{I}_{[j_t=i]}\Big)^{\frac{1}{2}}\leq\tilde{\beta}_{\mathrm{unif}}.\tag{4.7}$$

For the nonsmooth case, $\tilde{\beta}_{\mathrm{point}}$ and $\tilde{\beta}_{\mathrm{unif}}$ are of similar order. Indeed, if $T=O(n^2)$, the dominating term in both $\tilde{\beta}_{\mathrm{point}}$ and $\tilde{\beta}_{\mathrm{unif}}$ is $\sqrt{T}$. Furthermore, if $T=\Omega(n^2)$, then $\max_{i\in[n]}\sum_{t=1}^T\mathbb{I}_{[j_t=i]}=\Theta(T/n)$ with high probability [29], which implies that $\tilde{\beta}_{\mathrm{point}}=\Theta(T/n)$ and $\tilde{\beta}_{\mathrm{unif}}=\Theta(T/n)$ in this case. In Figure 1 (right panel), we also plot $\tilde{\beta}_{\mathrm{point}}$ and $\tilde{\beta}_{\mathrm{unif}}$ as a function of n. We set $T=n^2$, and get a sequence of indices $\{j_t\}_{t\in[T]}$ by drawing j_t from the uniform distribution over $[n]$. Then, we compute

$\tilde{\beta}_{\text{unif}}$ and $\tilde{\beta}_{\text{point}}$ according to Eq. (4.6). We repeat the experiments 25 times, and report the average of the experimental results. The experimental results show that $\tilde{\beta}_{\text{point}}$ and $\tilde{\beta}_{\text{unif}}$ behave as linear functions of n in the nonsmooth case (note $T/n = n$ in our experiments), which is consistent with our theoretical analysis.

Finally, we consider SGD for strongly convex and smooth problems.

Theorem 10 (Stability of SGD: Strongly Convex Case)**.** *Let Assumption 1 hold with $\alpha = 1$ and ℓ be G-Lipschitz. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be produced by SGD with $\eta_t \leq 1/L$, where $L := L_\alpha + L_r$. Then SGD with T iterations is $\boldsymbol{\beta}$-pointwise uniformly-stable in function values (measured by ℓ), where $\frac{1}{n}\sum_{i=1}^n \beta_i \leq \frac{4G^2}{n\lambda}$ and*

$$\frac{1}{n}\sum_{i=1}^n \beta_i^2 = \frac{4G^4}{n}\sum_{i=1}^n \Big(\sum_{t=1}^T \eta_t \mathbb{I}_{[j_t=i]} \prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2)\Big)^2.$$

Remark 9. Under the same condition, one can show that SGD is β_{unif}-uniformly stable in function values with (implicitly shown in the proof of Theorem 10) $\beta_{\text{unif}} = 2G^2 \max_{i\in[n]} \sum_{t=1}^T \eta_t \mathbb{I}_{[j_t=i]} \prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2)$. It is clear that $\beta_{\text{unif}}^2 \geq \frac{1}{n}\sum_{i=1}^n \beta_i^2$ for β_i in Theorem 10.

Remark 10 (Lower bounds)**.** Recently, lower bounds on the uniform stability were also developed for Lipschitz problems [3, 44, 19, 1]. A lower bound of order $\Omega\big(\min\{1, t/n\}\eta\sqrt{t} + \eta t/n\big)$ was established for SGD with convex and nonsmooth problems [3], a lower bound of order $\Omega\big(\eta t/n\big)$ was established for convex and smooth problems [44], and a lower bound of order $\Omega(\eta^2 n)$ was established for nonconvex problems [19]. These bounds are developed for uniform stability and are stated in expectation. As a comparison, this paper considers pointwise uniform stability. It is interesting to develop lower bounds on pointwise uniform stability with high probability.

4.2. Generalization Bounds

We now apply the above stability bounds to get high-probability generalization bounds of SGD. To our knowledge, Corollary 11 gives the first high-probability bounds on $\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|$ based on algorithmic stability. The bounds can be directly derived by plugging the stability bounds in Section 4.1 to Theorem 2 (Theorem 3). We omit the proofs for simplicity.

Corollary 11 (Generalization of SGD: Smooth Case)**.** *Assume $f : \mathcal{W}\times\mathcal{Z} \mapsto \mathbb{R}^+$ is convex, L-smooth and G-Lipschitz. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be produced by SGD with $\eta_t = \eta$. Let $\delta \in (0,1)$. Then with probability at least $1-\delta$ we have*

$$\big|F(A(S)) - F_S(A(S))\big| \lesssim \mathfrak{T}_1 \quad \textit{and} \quad \big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\| \lesssim \mathfrak{T}_1,$$

where

$$\mathfrak{T}_1 = \frac{\log^{\frac{1}{2}}(1/\delta)}{\sqrt{n}} + \frac{\eta \log n \log(1/\delta)}{\sqrt{n}}\Big(\sum_{i=1}^n\Big(\sum_{k=1}^T \mathbb{I}_{[j_k=i]}\Big)^2\Big)^{\frac{1}{2}}.$$

We now turn to high-probability bounds for SGD applied to nonsmooth problems.

Corollary 12 (Generalization of SGD: Nonsmooth Case)**.** *Assume $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}^+$ is convex and G-Lipschitz. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be produced by SGD with $\eta_t = \eta$ and $\delta \in (0,1)$. With probability at least $1-\delta$, we have*

$$\big|F(A(S)) - F_S(A(S))\big| \lesssim \frac{\log^{\frac{1}{2}}(1/\delta)}{\sqrt{n}} + \frac{\eta \log n \log(1/\delta)}{\sqrt{n}}\Big(Tn + T^{\frac{3}{2}} + \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{t}\mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_k=i]}\Big)^{\frac{1}{2}}.$$

Finally, we can directly plug Theorem 10 to Corollary 6 to get high-probability bounds for SGD applied to strongly convex problems.

Corollary 13 (Generalization of SGD: Strongly Convex Case)**.** *Let Assumptions in Lemma 5 hold with $\alpha = 1$ and ℓ be G-Lipschitz continuous. Let A be SGD with T iterations and $\eta_t \leq 1/L$. If $\sup_z \big|\mathbb{E}_S[\ell(A(S)); z] - \ell(\mathbf{w}^*; z)\big| < \infty$, then for any $\delta \in (0,1)$ with probability at least $1-\delta$ we have*

$$F(A(S)) - F_S(A(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*) \lesssim \mathfrak{T}_2 \quad \textit{and} \quad \|\nabla F(A(S)) - \nabla F_S(A(S)) + \nabla F_S(\mathbf{w}^*)\| \lesssim \mathfrak{T}_2,$$

where

$$\mathfrak{T}_2 = \Big(\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{t=1}^{T}\eta_t\mathbb{I}_{[j_t=i]}\prod_{t'=t+1}^{T}(1-\eta_{t'}\lambda/2)\Big)^2\Big)^{\frac{1}{2}}\log n\log(1/\delta)$$
$$+ \frac{\log^{1/2}(1/\delta)(\mathbb{E}[F(A(S))])^{\frac{1}{2}}}{\lambda\sqrt{n}}\Big(\Big(\mathbb{E}\big[F_S(A(S)) - F_S(A_{\mathbf{e}}(S))\big]\Big)^{\frac{1}{2}} + \frac{\mathbb{E}\big[L(A_{\mathbf{e}}(S))\big]}{\sqrt{n}}\Big).$$

5. Proofs on Connecting Stability and Generalization

5.1. Proof of Theorem 1

To prove Theorem 1, we need the following Marcinkiewicz-Zygmund's inequality for random variables taking values in a Hilbert space. It shows that the p-norm of a summation of independent random variables can be bounded by the summation of the p-norm of random variables.

Lemma 14. *Let $X_1, \ldots, X_n$ be independent random variables taking values in a Hilbert space with $\mathbb{E}[X_i] = 0$ for all $i \in [n]$. Then for any $p \geq 2$ we have*

$$\Big\|\Big\|\sum_{i=1}^{n}X_i\Big\|\Big\|_p \leq 2\sqrt{np}\Big(\frac{1}{n}\sum_{i=1}^{n}\big\|\|X_i\|\big\|_p^p\Big)^{\frac{1}{p}}.$$

The Marcinkiewicz-Zygmund's inequality can be proved by using its connection to Khintchine-Kahane's inequality [4, page 441], where the Marcinkiewicz-Zygmund's inequality was established for real-valued random variables. To get Marcinkiewicz-Zygmund's inequality for vector-valued random variables, we need to use the following Khintchine-Kahane's inequality [10, Theorem 1.3.1]

$$\mathbb{E}_\epsilon\|\sum_{i=1}^{n}\epsilon_iX_i\|^p \leq \max((p-1)^{\frac{p}{2}}, 1)\Big(\sum_{i=1}^{n}\|X_i\|^2\Big)^{\frac{p}{2}} \quad p \geq 2,$$

where $X_1, \ldots, X_n$ are elements in a Hilbert space, and $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher variables (i.e., taking values in $\{1, -1\}$ with the same probability). For brevity, we omit the proof of Lemma 14.

We now give the proof of Theorem 1, which is motivated by the analysis in [6]. For $f(Z_1, \ldots, Z_n)$ and $A \subset [n]$, we write $\|f\|_p(Z_A) = \big(\mathbb{E}[|f|^p|Z_A]\big)^{\frac{1}{p}}$.

Proof of Theorem 1. For simplicity, we assume $n = 2^k$. Define a sequence of partitions $\mathcal{B}_0, \ldots, \mathcal{B}_k$ with $\mathcal{B}_k = \{1, 2, \ldots, 2^k\}$, where $\mathcal{B}_l$ is derived from $\mathcal{B}_{l+1}$ by splitting each subset in $\mathcal{B}_{l+1}$ into two equal parts. Then, there holds

$$\mathcal{B}_0 = \{\{1\}, \{2\}, \ldots, \{2^k\}\},\ \mathcal{B}_1 = \{\{1, 2\}, \{3, 4\}, \ldots, \{2^k - 1, 2^k\}\}, \ldots, \mathcal{B}_k = \{[n]\}.$$

For each $i \in [n]$ and $l = 0, 1, \ldots, k$, denote by $B^l(i) \in \mathcal{B}_l$ the only set from $\mathcal{B}_l$ containing i. According to this definition, we know $B^0(i) = \{i\}$ and $B^k(i) = [n]$.

For each $i \in [n]$ and each $l = 0, 1, \ldots, k$, we introduce random vectors as follows

$$g_i^l := g_i^l(Z_i, Z_{[n]\setminus B^l(i)}) = \mathbb{E}[g_i | Z_i, Z_{[n]\setminus B^l(i)}].$$

That is, we condition on Z_i and all the variables that are not in the same set as Z_i in $\mathcal{B}_l$. This definition shows that $g_i^0 = g_i$ and $g_i^k = \mathbb{E}[g_i|Z_i]$. For each $i \in [n]$, we can decompose g_i as follows

$$g_i = \mathbb{E}[g_i|Z_i] + \sum_{l=0}^{k-1}(g_i^l - g_i^{l+1}).$$

It then follows from the triangle inequality that

$$\begin{aligned}\Big\|\Big\|\sum_{i=1}^n g_i\Big\|\Big\|_p &= \Big\|\Big\|\sum_{i=1}^n \Big(\mathbb{E}[g_i|Z_i] + \sum_{l=0}^{k-1}(g_i^l - g_i^{l+1})\Big)\Big\|\Big\|_p \\ &\le \Big\|\Big\|\sum_{i=1}^n \mathbb{E}[g_i|Z_i]\Big\|\Big\|_p + \sum_{l=0}^{k-1}\Big\|\Big\|\sum_{i=1}^n(g_i^l - g_i^{l+1})\Big\|\Big\|_p. \end{aligned} \tag{5.1}$$

Since $\|\mathbb{E}[g_i|Z_i]\| \le M$, one can check that $f(Z_1, \ldots, Z_n) = \sum_{i=1}^n \mathbb{E}[g_i|Z_i]$ satisfies Eq. (A.1) with $\beta_i = 2M$. Furthermore, we have $\mathbb{E}[\mathbb{E}[g_i|Z_i]] = 0$. Now we can apply Lemma A.3 with $\beta_i = 2M$ to derive the following inequality

$$\Big\|\Big\|\sum_{i=1}^n \mathbb{E}[g_i|Z_i]\Big\|\Big\|_p \le 2(\sqrt{2}+1)\sqrt{np}M. \tag{5.2}$$

The definition of g_i^l implies that

$$\mathbb{E}_{Z_{B^{l+1}(i)\setminus B^l(i)}}[g_i^l] = g_i^{l+1}.$$

We view g_i^l as a function of $Z_j, j \in B^{l+1}(i)\setminus B^l(i)$. Changing any Z_j would change g_i^l by β_j. Therefore, one can apply Lemma A.3 with $f = g_i^l$ to derive the following inequality with (there are 2^l random variables)

$$\Big\|\|g_i^l - g_i^{l+1}\|\Big\|_p(Z_i, Z_{[n]\setminus \mathcal{B}^{l+1}(i)}) \le (\sqrt{2}+1)\Big(p\sum_{j\in B^{l+1}(i)\setminus B^l(i)}\beta_j^2\Big)^{\frac{1}{2}}. \tag{5.3}$$

We now turn to the sum $\sum_{i\in B}(g_i^l - g_i^{l+1})$ for any $B \in \mathcal{B}_l$. Consider any $i \in B \in \mathcal{B}_l$. Note $Z_i' := g_i^l - g_i^{l+1}$ is a function of $Z_i, Z_{[n]\setminus B}$. We now condition on $Z_{[n]\setminus B}$ and then Z_i' is a function of Z_i, which are independent. We can apply Lemma 14 to derive the following inequality

$$\Big\|\Big\|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big\|\Big\|_p^p(Z_{[n]\setminus B}) \le \frac{(2\sqrt{p|B|})^p}{|B|}\sum_{i\in B}\|\|g_i^l - g_i^{l+1}\|\|_p^p(Z_{[n]\setminus B}).$$

Taking integration over $Z_{[n]\backslash B}$ gives

$$\Big|\Big|\Big|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big|\Big|\Big|_p^p \le \frac{(2\sqrt{p|B|})^p}{|B|}\sum_{i\in B}\||g_i^l - g_i^{l+1}\||_p^p,$$

which implies

$$\Big|\Big|\Big|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big|\Big|\Big|_p \le 2\sqrt{p|B|}\Big(\frac{1}{|B|}\sum_{i\in B}\||g_i^l - g_i^{l+1}\||_p^p\Big)^{\frac{1}{p}}.$$

This together with Eq. (5.3) implies that

$$\Big|\Big|\Big|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big|\Big|\Big|_p \le 2p(\sqrt{2}+1)|B|^{\frac{1}{2}}\Big(\frac{1}{|B|}\sum_{i\in B}\Big(\sum_{j\in B^{l+1}(i)\backslash B^l(i)}\beta_j^2\Big)^{\frac{p}{2}}\Big)^{\frac{1}{p}}. \tag{5.4}$$

The rest of the proof is where we depart from the analysis of [6]. Note for any $i, i' \in B \in \mathcal{B}^l$, we have

$$B^{l+1}(i)\backslash B^l(i) = B^{l+1}(i')\backslash B^l(i').$$

Therefore, for any $B \in \mathcal{B}^l$ we have the following well-defined notation

$$\widetilde{B} := \big\{j : j \in B^{l+1}(i)\backslash B^l(i)\big\}, \quad \forall i \in B,$$

which implies

$$\sum_{j\in B^{l+1}(i)\backslash B^l(i)}\beta_j^2 = \sum_{j\in\widetilde{B}}\beta_j^2, \quad \text{if } i \in B \in \mathcal{B}^l. \tag{5.5}$$

One can interpret $\widetilde{B}$ as a sibling set of B in $\mathcal{B}^l$ (they have the same parent set in $\mathcal{B}^{l+1}$). For example, if $B = \{5,6\}$, then $\widetilde{B} = \{7,8\}$. If $B = \{7,8\}$, then $\widetilde{B} = \{5,6\}$. The parent set in $\mathcal{B}^2$ is $\{5,6,7,8\}$. It then follows from Eq. (5.4) and Eq. (5.5) the following inequality for any $B \in \mathcal{B}^l$ ($|B| = 2^l$)

$$\Big|\Big|\Big|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big|\Big|\Big|_p \le 2p(\sqrt{2}+1)|B|^{\frac{1}{2}}\Big(\Big(\sum_{j\in\widetilde{B}}\beta_j^2\Big)^{\frac{p}{2}}\Big)^{\frac{1}{p}} = 2p(\sqrt{2}+1)2^{\frac{l}{2}}\Big(\sum_{j\in\widetilde{B}}\beta_j^2\Big)^{\frac{1}{2}}.$$

It then follows from the fact $|\mathcal{B}_l| = 2^{k-l}$ and the Cauchy-Schwartz inequality that

$$\begin{aligned}\Big|\Big|\Big|\sum_{i\in[n]}(g_i^l - g_i^{l+1})\Big|\Big|\Big|_p &\le \sum_{B\in\mathcal{B}_l}\Big|\Big|\Big|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big|\Big|\Big|_p \le 2^{\frac{k-l}{2}}\Big(\sum_{B\in\mathcal{B}_l}\Big|\Big|\Big|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big|\Big|\Big|_p^2\Big)^{\frac{1}{2}}\\ &\le 2p(\sqrt{2}+1)2^{\frac{k}{2}}\Big(\sum_{B\in\mathcal{B}_l}\sum_{j\in\widetilde{B}}\beta_j^2\Big)^{\frac{1}{2}}.\end{aligned}$$

According to our definition of $\widetilde{B}$, one can check $\sum_{B\in\mathcal{B}_l}\sum_{j\in\widetilde{B}}\beta_j^2 = \sum_{i=1}^n\beta_i^2$ and therefore

$$\Big|\Big|\Big|\sum_{i\in[n]}(g_i^l - g_i^{l+1})\Big|\Big|\Big|_p \le 2p(\sqrt{2}+1)\sqrt{n}\Big(\sum_{i=1}^n\beta_i^2\Big)^{\frac{1}{2}}.$$

This further gives

$$\sum_{l=0}^{k-1}\Big|\Big|\Big|\sum_{i=1}^n(g_i^l - g_i^{l+1})\||\Big|\Big|_p \le 2p(\sqrt{2}+1)k\Big(n\sum_{i=1}^n\beta_i^2\Big)^{\frac{1}{2}}.$$

We can plug Eq. (5.2) and the above inequality back into Eq. (5.1) to derive

$$\Big|\Big|\Big|\sum_{i=1}^n g_i\Big|\Big|\Big|_p \le 2(\sqrt{2}+1)\sqrt{np}M + 2p(\sqrt{2}+1)k\Big(n\sum_{i=1}^n\beta_i^2\Big)^{\frac{1}{2}}.$$

The proof is completed. $\square$

Remark 11. We highlight the difference between our proof and the analysis in [6]. We adopt the analysis in [6] to derive Eq. (5.4), excepting considering vector-valued random variables here. If the algorithm is β_{unif}-uniformly stable, then the analysis in [6] gives the following inequality similar to Eq. (5.4)

$$\Big\|\Big\|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big\|\Big\|_p \le 2p(\sqrt{2}+1)2^l\beta_{\mathrm{unif}}, \quad \forall B\in\mathcal{B}^l.$$

Then one immediately gets

$$\Big\|\Big\|\sum_{i\in[n]}(g_i^l - g_i^{l+1})\Big\|\Big\|_p \le \sum_{B\in\mathcal{B}_l}\Big\|\Big\|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big\|\Big\|_p \le 2^{k-l}2^l 2p(\sqrt{2}+1)\beta_{\mathrm{unif}} = 2np(\sqrt{2}+1)\beta_{\mathrm{unif}}. \quad (5.6)$$

As a comparison, we get Eq. (5.4) to control $\|\|\sum_{i\in B}(g_i^l - g_i^{l+1})\|\|_p$ in terms of $\sum_{i\in B}\big(\sum_{j\in B^{l+1}(i)\setminus B^l(i)}\beta_j^2\big)^{\frac{p}{2}}$. Our observation is that $\sum_{j\in B^{l+1}(i)\setminus B^l(i)}\beta_j^2$ is the same for any $i\in B\in\mathcal{B}^l$, based on which we show

$$\Big\|\Big\|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big\|\Big\|_p^2 \le 4p^2(\sqrt{2}+1)^2 2^l\sum_{j\in\widetilde{B}}\beta_j^2, \quad (5.7)$$

where $\widetilde{B}$ is a sibling of B. We then apply the Cauchy-Schwartz inequality to get

$$\Big\|\Big\|\sum_{i\in[n]}(g_i^l - g_i^{l+1})\Big\|\Big\|_p \le \sum_{B\in\mathcal{B}_l}\Big\|\Big\|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big\|\Big\|_p \le 2^{\frac{k-l}{2}}\Big(\sum_{B\in\mathcal{B}_l}\Big\|\Big\|\sum_{i\in B}(g_i^l - g_i^{l+1})\Big\|\Big\|_p^2\Big)^{\frac{1}{2}}.$$

Finally, we can apply Eq. (5.7) to derive a bound similar to Eq. (5.6).

5.2. Proof of Theorem 3

In this section, we give the proof of Theorem 3.

Proof of Theorem 3. Let $S' = \{z_1', \dots, z_n'\}$ be drawn independently from ρ. For any $i\in[n]$, define

$$S^{(i)} = \{z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n\}. \quad (5.8)$$

Since $\mathbb{E}_Z[\nabla f(A(S);Z)] = \nabla F(A(S))$, we can decompose $\nabla F(A(S)) - \nabla F_S(A(S))$ as follows

$$\begin{aligned} n\big(\nabla F(A(S)) - \nabla F_S(A(S))\big) &= \sum_{i=1}^n \mathbb{E}_{Z,z_i'}\Big[\nabla f(A(S);Z) - \nabla f(A(S^{(i)});Z)\Big] \\ &+ \sum_{i=1}^n \mathbb{E}_{z_i'}\Big[\mathbb{E}_Z[\nabla f(A(S^{(i)});Z)] - \nabla f(A(S^{(i)});z_i)\Big] + \sum_{i=1}^n \mathbb{E}_{z_i'}\Big[\nabla f(A(S^{(i)});z_i) - \nabla f(A(S);z_i)\Big]. \end{aligned}$$

Since A is $\boldsymbol{\beta}$-pointwise uniformly stable in gradients, we know

$$n\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\| \le 2\sum_{i=1}^n\beta_i + \Big\|\sum_{i=1}^n g_i\Big\|, \quad (5.9)$$

where $g_i = \mathbb{E}_{z_i'}\big[\mathbb{E}_Z[\nabla f(A(S^{(i)});Z)] - \nabla f(A(S^{(i)});z_i)\big]$. According to our assumption, we know $\|g_i\| \le 2M$ and

$$\begin{aligned} \mathbb{E}_{z_i}[g_i] &= \mathbb{E}_{z_i}\mathbb{E}_{z_i'}\big[\mathbb{E}_Z[\nabla f(A(S^{(i)});Z)] - \nabla f(A(S^{(i)});z_i)\big] \\ &= \mathbb{E}_{z_i'}\big[\mathbb{E}_Z[\nabla f(A(S^{(i)});Z)] - \mathbb{E}_{z_i}[\nabla f(A(S^{(i)});z_i)]\big] = 0, \end{aligned}$$

where we have used the fact that z_i and Z follow from the same distribution. For any $i \in [n]$, any $j \neq i$ and any z_j'', we have

$$\begin{aligned}
&\big\|g_i(z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, z_n) - g_i(z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\big\| \\
&= \big\|\mathbb{E}_{z_i'}\big[\mathbb{E}_Z[\nabla f(A(S^{(i)}); Z)] - \nabla f(A(S^{(i)}); z_i)\big] - \mathbb{E}_{z_i'}\big[\mathbb{E}_Z[\nabla f(A(S_j^{(i)}); Z)] - \nabla f(A(S_j^{(i)}); z_i)\big]\big\| \\
&\leq \big\|\mathbb{E}_{z_i'}\big[\mathbb{E}_Z[\nabla f(A(S^{(i)}); Z) - \nabla f(A(S_j^{(i)}); Z)]\big\| + \big\|\mathbb{E}_{z_i'}[\nabla f(A(S^{(i)}); z_i) - \nabla f(A(S_j^{(i)}); z_i)]\big\| \leq 2\beta_j,
\end{aligned}$$

where

$$S_j^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n\}. \tag{5.10}$$

Therefore, all the assumptions in Theorem 1 hold (with M replaced by $2M$ and β_j replaced by $2\beta_j$) and we can apply Theorem 1 to derive

$$\Big\|\Big\|\sum_{i=1}^n g_i\Big\|\Big\|_p \leq 4(\sqrt{2}+1)\sqrt{np}M + 4p(\sqrt{2}+1)\lceil \log_2 n\rceil \Big(n\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}}.$$

We can combine the above inequality and Eq. (5.9) to derive the following inequality

$$n\Big\|\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|\Big\|_p \leq 2\sum_{i=1}^n \beta_i + 4(\sqrt{2}+1)\sqrt{np}M + 4p(\sqrt{2}+1)\lceil \log_2 n\rceil \Big(n\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}}.$$

By Lemma A.5, the following inequality holds with probability at least $1-\delta$

$$n\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\| \leq 2\sum_{i=1}^n \beta_i + 4e(\sqrt{2}+1)\sqrt{n\log(1/\delta)}M + 4e(\sqrt{2}+1)\lceil \log_2 n\rceil \log(1/\delta)\Big(n\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}}$$

and therefore

$$\begin{aligned}
\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\| \leq{}& \frac{2\sum_{i=1}^n \beta_i}{n} + \\
& 4e(\sqrt{2}+1)M\log^{\frac{1}{2}}(1/\delta)n^{-\frac{1}{2}} + 4e(\sqrt{2}+1)\lceil \log_2 n\rceil \log(1/\delta)\Big(\frac{1}{n}\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}}.
\end{aligned}$$

The proof is completed. □

5.3. Proof of Theorem 7

To prove Theorem 7, we require the following lemma on the uniform stability of ERM for strongly convex problems. It is a direct extension of a similar result in [5] to functions with a structure in Assumption 1. Since the proof is identical to the classical stability analysis, we omit the proof for brevity.

Lemma 15. *Let Assumption 1 hold and ℓ be G-Lipschitz continuous. Then*

$$\max_{i\in[n]} \sup_{S, S^{(i)}} \sup_z \big[\ell\big(A_e(S); z\big) - \ell\big(A_e(S^{(i)}); z\big)\big] \leq 4G^2/(n\lambda),$$

where $S^{(i)}$ is defined in Eq. (5.8).

Proof of Theorem 7. According to Lemma 15, we know that A_{e} is $4G^2/(n\lambda)$-uniformly stable in function values (measured by ℓ). According to Theorem 4, the following inequality holds with probability at least $1-\delta$

$$F(A_{\mathrm{e}}(S)) - F_S(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*) \lesssim (n\lambda)^{-1}\log n \log(1/\delta) + \Big(\frac{\sigma_A^2 \log(1/\delta)}{n}\Big)^{\frac{1}{2}},$$

where (by Lemma 5)

$$\sigma_A^2 \leq C\lambda^{-1}\mathbb{E}_S[F(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*)].$$

We know

$$\begin{aligned}\mathbb{E}_S[F(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*)] &= \mathbb{E}_S[F(A_{\mathrm{e}}(S)) - F_S(A_{\mathrm{e}}(S))] + \mathbb{E}_S[F_S(A_{\mathrm{e}}(S)) - F_S(\mathbf{w}^*)] + \mathbb{E}_S[F_S(\mathbf{w}^*) - F(\mathbf{w}^*)]\\ &\leq \mathbb{E}_S[F(A_{\mathrm{e}}(S)) - F_S(A_{\mathrm{e}}(S))] \leq \frac{4G^2}{n\lambda}.\end{aligned}$$

We can combine the above inequalities to derive the following inequality with probability at least $1-\delta$

$$F(A_{\mathrm{e}}(S)) - F_S(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*) \lesssim (n\lambda)^{-1}\log n \log(1/\delta). \tag{5.11}$$

According to Lemma B.1, we know

$$\begin{aligned}&\langle A(S) - A_{\mathrm{e}}(S), \nabla F(A_{\mathrm{e}}(S))\rangle \leq \|A(S) - A_{\mathrm{e}}(S)\|\|\nabla F(A_{\mathrm{e}}(S))\|\\ &\leq \|A(S) - A_{\mathrm{e}}(S)\|\Big(\frac{L_r(1+\alpha)^{\frac{1}{1+\alpha}}}{2L_\alpha^{\frac{1}{1+\alpha}}}\Big(F(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*)\Big)^{\frac{1}{1+\alpha}} + 2\Big(\frac{L_\alpha}{1+\alpha}\Big)^{\frac{1}{1+\alpha}}\Big(F(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*)\Big)^{\frac{\alpha}{1+\alpha}}\Big)\\ &\leq \|A(S) - A_{\mathrm{e}}(S)\|\Big(F(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*)\Big)^{\frac{\alpha}{1+\alpha}}\Big(\frac{\widetilde{C}L_r(1+\alpha)^{\frac{1}{1+\alpha}}}{2L_\alpha^{\frac{1}{1+\alpha}}} + 2\Big(\frac{L_\alpha}{1+\alpha}\Big)^{\frac{1}{1+\alpha}}\Big)\\ &:= C_1\|A(S) - A_{\mathrm{e}}(S)\|\Big(F(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*)\Big)^{\frac{\alpha}{1+\alpha}},\end{aligned} \tag{5.12}$$

where we have used the assumption $\Big(F(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*)\Big)^{\frac{1-\alpha}{1+\alpha}} \leq \widetilde{C}$ and introduced C_1 in the last step. Since ℓ has (α, L_α)-Hölder continuous gradients and r is L_r-smooth, Eq. (B.1) implies

$$\begin{aligned}F(A(S)) - F(A_{\mathrm{e}}(S)) &\leq \langle A(S) - A_{\mathrm{e}}(S), \nabla F(A_{\mathrm{e}}(S))\rangle + \frac{L_\alpha\|A(S) - A_{\mathrm{e}}(S)\|^{1+\alpha}}{1+\alpha} + \frac{L_r\|A(S) - A_{\mathrm{e}}(S)\|^2}{2}\\ &\leq C_1\|A(S) - A_{\mathrm{e}}(S)\|\Big(F(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*)\Big)^{\frac{\alpha}{1+\alpha}} + \frac{L_\alpha\|A(S) - A_{\mathrm{e}}(S)\|^{1+\alpha}}{1+\alpha} + \frac{L_r\|A(S) - A_{\mathrm{e}}(S)\|^2}{2}.\end{aligned} \tag{5.13}$$

Since $F_S(A_{\mathrm{e}}(S)) \leq F_S(\mathbf{w}^*)$, we can plug Eq. (5.11) to the above inequality and derive the following inequality with probability at least $1-\delta$

$$F(A(S)) - F(A_{\mathrm{e}}(S)) \lesssim \|A(S) - A_{\mathrm{e}}(S)\|\Big((n\lambda)^{-1}\log n\log(1/\delta)\Big)^{\frac{\alpha}{1+\alpha}} + \|A(S) - A_{\mathrm{e}}(S)\|^{1+\alpha} + \|A(S) - A_{\mathrm{e}}(S)\|^2.$$

By the following inequality due to the strong convexity of F_S,

$$F_S(A(S)) - F_S(A_{\mathrm{e}}(S)) \geq \frac{\lambda}{2}\|A(S) - A_{\mathrm{e}}(S)\|^2, \tag{5.14}$$

we get the following inequality with probability at least $1-\delta$

$$F(A(S)) - F(A_{\mathrm{e}}(S)) \lesssim \hat{\Delta}_\lambda^{\frac{1}{2}}\Big((n\lambda)^{-1}\log n\log(1/\delta)\Big)^{\frac{\alpha}{1+\alpha}} + \hat{\Delta}_\lambda^{\frac{1+\alpha}{2}}.$$

We can combine the above inequality and Eq. (5.11) to derive the following inequality with probability at least $1-\delta$

$$F(A(S))-F_S(A_{\mathrm{e}}(S))-F(\mathbf{w}^*)+F_S(\mathbf{w}^*)\lesssim (n\lambda)^{-1}\log n\log(1/\delta)+\hat{\Delta}_\lambda^{\frac{1}{2}}\Big((n\lambda)^{-1}\log n\log(1/\delta)\Big)^{\frac{\alpha}{1+\alpha}}+\hat{\Delta}_\lambda^{\frac{1+\alpha}{2}}.$$

By the following Young's inequality

$$\hat{\Delta}_\lambda^{\frac{1}{2}}\Big((n\lambda)^{-1}\log n\log(1/\delta)\Big)^{\frac{\alpha}{1+\alpha}}\leq\frac{\alpha}{1+\alpha}\Big((n\lambda)^{-1}\log n\log(1/\delta)\Big)^{\frac{\alpha}{1+\alpha}\frac{1+\alpha}{\alpha}}+\frac{1}{1+\alpha}\hat{\Delta}_\lambda^{\frac{1+\alpha}{2}},$$

the following inequality holds with probability at least $1-\delta$

$$F(A(S))-F_S(A_{\mathrm{e}}(S))-F(\mathbf{w}^*)+F_S(\mathbf{w}^*)\lesssim (n\lambda)^{-1}\log n\log(1/\delta)+\hat{\Delta}_\lambda^{\frac{1+\alpha}{2}}.$$

The stated bound then follows by noting $F_S(\mathbf{w}^*)\geq F_S(A_{\mathrm{e}}(S))$. The proof is completed. □

6. Proofs on Stochastic Gradient Descent

In this section, we present the proof on the stability bounds of SGD. Our analysis is based on the following lemma in [17], which shows that the gradient update $\mathbf{w}\mapsto\mathbf{w}-\eta\nabla f(\mathbf{w};z)$ is nonexpansive if f is convex and smooth.

Lemma 16 ([17]). *Suppose the function $\mathbf{w}\mapsto f(\mathbf{w};z)$ is convex and L-smooth. If $\eta\leq 2/L$, then*

$$\big\|\big(\mathbf{w}-\eta\nabla f(\mathbf{w};z)\big)-\big(\mathbf{w}'-\eta\nabla f(\mathbf{w}';z)\big)\big\|\leq\|\mathbf{w}-\mathbf{w}'\|.$$

Furthermore, if $\mathbf{w}\mapsto f(\mathbf{w};z)$ is λ-strongly convex and $\eta\leq 1/L$, then

$$\big\|\big(\mathbf{w}-\eta\nabla f(\mathbf{w};z)\big)-\big(\mathbf{w}'-\eta\nabla f(\mathbf{w}';z)\big)\big\|^2\leq(1-\eta\lambda)\|\mathbf{w}-\mathbf{w}'\|^2.$$

Let $S^{(i)}$ be defined by Eq. (5.8). Let $\{\mathbf{w}_t^{(i)}\}$ be produced by SGD w.r.t. $S^{(i)}$.

Proof of Theorem 8. We build a recurrent formula on estimating $\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|$. Consider two cases at the t-th iteration. If $j_t\neq i$, then Lemma 16 implies that

$$\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|=\big\|\big(\mathbf{w}_t-\eta_t\nabla f(\mathbf{w}_t;z_{j_t})\big)-\big(\mathbf{w}_t^{(i)}-\eta_t\nabla f(\mathbf{w}_t^{(i)};z_{j_t})\big)\big\|\leq\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|. \tag{6.1}$$

If $j_t=i$, then $\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|\leq\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|+2G\eta_t$. We can combine the above two inequalities to get

$$\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|\leq\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|+2G\eta_t\mathbb{I}_{[j_t=i]}.$$

We apply the above inequality recursively and get

$$\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|\leq 2G\eta\sum_{k=1}^{t}\mathbb{I}_{[j_k=i]}.$$

By the Lipschitz continuity, we know that SGD with T iterations is $\boldsymbol{\beta}$-pointwise uniformly stable, where $\beta_i=2G^2\eta\sum_{k=1}^T\mathbb{I}_{[j_k=i]}$. It then follows that

$$\frac{1}{n}\sum_{i=1}^n\beta_i^2=\frac{4G^4\eta^2}{n}\sum_{i=1}^n\Big(\sum_{k=1}^T\mathbb{I}_{[j_k=i]}\Big)^2,$$

$$\frac{1}{n}\sum_{i=1}^n\beta_i=\frac{2G^2\eta}{n}\sum_{i=1}^n\sum_{k=1}^T\mathbb{I}_{[j_k=i]}=\frac{2G^2\eta}{n}\sum_{k=1}^T\sum_{i=1}^n\mathbb{I}_{[j_k=i]}=\frac{2G^2T\eta}{n},$$

where we have used $\sum_{i=1}^n\mathbb{I}_{[j_k=i]}=1$ for any k. The proof is completed. □

Proof of Theorem 9. Consider two cases at the t-th iteration. In the first case, assume $j_t \neq i$. Then the Lipschitz continuity of f implies that

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|^2 = \big\|\big(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{j_t})\big) - \big(\mathbf{w}_t^{(i)} - \eta_t \nabla f(\mathbf{w}_t^{(i)}; z_{j_t})\big)\big\|^2 \\
&= \big\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\big\|^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}_t^{(i)}, \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla f(\mathbf{w}_t^{(i)}; z_{j_t})\rangle + \eta_t^2 \|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla f(\mathbf{w}_t^{(i)}; z_{j_t})\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|^2 + 4G^2\eta_t^2,
\end{aligned}$$

where we have used the inequality $\langle \mathbf{w}_t - \mathbf{w}_t^{(i)}, \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla f(\mathbf{w}_t^{(i)}; z_{j_t})\rangle \geq 0$ due to the convexity of f. For the second case, assume $j_t = i$. Then

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|^2 = \big\|\big(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{j_t})\big) - \big(\mathbf{w}_t^{(i)} - \eta_t \nabla f(\mathbf{w}_t^{(i)}; z'_{j_t})\big)\big\|^2 \\
&= \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|^2 + \eta_t^2 \|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla f(\mathbf{w}_t^{(i)}; z'_{j_t})\|^2 - 2\eta_t \big\langle \mathbf{w}_t - \mathbf{w}_t^{(i)}, \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla f(\mathbf{w}_t^{(i)}; z'_{j_t})\big\rangle \\
&\leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|^2 + 4G^2\eta_t^2 + 4G\eta_t \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|,
\end{aligned}$$

where we have used the Lipschitz continuity of f. We can combine the above two cases to get

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|^2 + 4G^2\eta_t^2 + 4G\eta_t \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\| \mathbb{I}_{[j_t = i]}.$$

We apply the above inequality recursively and get

$$\|\mathbf{w}_{T+1} - \mathbf{w}_{T+1}^{(i)}\|^2 \leq 4G^2\eta^2 T + 4G\eta \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\| \mathbb{I}_{[j_t = i]}$$

and therefore

$$\sup_{S, S^{(i)}} \|\mathbf{w}_{T+1} - \mathbf{w}_{T+1}^{(i)}\|^2 \leq 4G^2\eta^2 T + 4G\eta \sum_{t=1}^{T} \sup_{S, S^{(i)}} \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\| \mathbb{I}_{[j_t = i]},$$

where the supremum is taken over two neighboring datasets differing by the i-th example. Let

$$\beta_{T+1,i} = G \sup_{S, S^{(i)}} \|\mathbf{w}_{T+1} - \mathbf{w}_{T+1}^{(i)}\|. \tag{6.2}$$

Then we know SGD with T iterations is $\boldsymbol{\beta}_{T+1}$-pointwise uniformly stable in function values, where $\beta_{T+1,i}$ satisfies the following inequality

$$\beta_{T+1,i}^2 \leq 4G^4\eta^2 T + 4G^2\eta \sum_{t=1}^{T} \beta_{t,i} \mathbb{I}_{[j_t = i]}. \tag{6.3}$$

Let $\Delta_{t,i} = \max_{k \leq t} \beta_{k,i}$. Then the above inequality implies that $\Delta_{T,i}^2 \leq 4G^4\eta^2 T + 4G^2\eta \Delta_{T,i} \sum_{t=1}^{T} \mathbb{I}_{[j_t = i]}$. Solving this quadratic inequality of $\Delta_{T,i}$ implies that

$$\Delta_{T,i} \leq 2G^2\eta\sqrt{T} + 4G^2\eta \sum_{t=1}^{T} \mathbb{I}_{[j_t = i]}. \tag{6.4}$$

We can plug the above bound back into Eq. (6.3), and get

$$\begin{aligned}
\beta_{T+1,i}^2 &\leq 4G^4\eta^2 T + 4G^2\eta \sum_{t=1}^{T} \mathbb{I}_{[j_t = i]} \Big(2G^2\eta\sqrt{t} + 4G^2\eta \sum_{k=1}^{t} \mathbb{I}_{[j_k = i]}\Big) \\
&\leq 4G^4\eta^2 T + 8G^4\eta^2 \sum_{t=1}^{T} \sqrt{t}\, \mathbb{I}_{[j_t = i]} + 16G^4\eta^2 \sum_{t=1}^{T} \sum_{k=1}^{t} \mathbb{I}_{[j_t = i]} \mathbb{I}_{[j_k = i]}.
\end{aligned}$$

It then follows that ($\sum_{i=1}^n \mathbb{I}_{[j_t=i]} = 1$)

$$\sum_{i=1}^n \beta_{T+1,i}^2 \leq 4G^4\eta^2 Tn + \frac{16}{3}G^4\eta^2(T+1)^{\frac{3}{2}} + 16G^4\eta^2 \sum_{i=1}^n \sum_{t=1}^T \sum_{k=1}^t \mathbb{I}_{[j_t=i]}\mathbb{I}_{[j_k=i]}.$$

The proof is completed. □

Proof of Theorem 10. It is clear that for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is L-smooth. We now build a recurrent formula on estimating $\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|$. Consider two cases at the t-th iteration. In the first case, assume $j_t \neq i$. Then Lemma 16 implies that

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\| = \big\|\big(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{j_t})\big) - \big(\mathbf{w}_t^{(i)} - \eta_t \nabla f(\mathbf{w}_t^{(i)}; z_{j_t})\big)\big\| \leq (1-\eta_t\lambda/2)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|. \tag{6.5}$$

For the second case, assume $j_t = i$. Then

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\| = \big\|\big(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{j_t})\big) - \big(\mathbf{w}_t^{(i)} - \eta_t \nabla f(\mathbf{w}_t^{(i)}; z'_{j_t})\big)\big\| \\
&\leq \big\|\big(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{j_t})\big) - \big(\mathbf{w}_t^{(i)} - \eta_t \nabla f(\mathbf{w}_t^{(i)}; z_{j_t})\big)\big\| + \eta_t \big\|\nabla f(\mathbf{w}_t^{(i)}; z_{j_t}) - \nabla f(\mathbf{w}_t^{(i)}; z'_{j_t})\big\| \\
&= \big\|\big(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{j_t})\big) - \big(\mathbf{w}_t^{(i)} - \eta_t \nabla f(\mathbf{w}_t^{(i)}; z_{j_t})\big)\big\| + \eta_t \big\|\nabla \ell(\mathbf{w}_t^{(i)}; z_{j_t}) - \nabla \ell(\mathbf{w}_t^{(i)}; z'_{j_t})\big\| \\
&\leq (1-\eta_t\lambda/2)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\| + 2G\eta_t,
\end{aligned} \tag{6.6}$$

where we have used the Lipschitz continuity of ℓ. We can combine Eq. (6.5) and Eq. (6.6) to get

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\| \leq (1-\eta_t\lambda/2)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\| + 2G\eta_t \mathbb{I}_{[j_t=i]}. \tag{6.7}$$

We apply the above inequality recursively, and derive

$$\|\mathbf{w}_{T+1} - \mathbf{w}_{T+1}^{(i)}\| \leq 2G \sum_{t=1}^T \eta_t \mathbb{I}_{[j_t=i]} \prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2). \tag{6.8}$$

Therefore, SGD is $\boldsymbol{\beta}$-pointwise uniformly stable in function values, where

$$\beta_i = 2G^2 \sum_{t=1}^T \eta_t \mathbb{I}_{[j_t=i]} \prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2). \tag{6.9}$$

It then follows that

$$\frac{1}{n}\sum_{i=1}^n \beta_i^2 = \frac{4G^4}{n} \sum_{i=1}^n \Big(\sum_{t=1}^T \eta_t \mathbb{I}_{[j_t=i]} \prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2)\Big)^2$$

and

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \beta_i = \frac{2G^2}{n} \sum_{i=1}^n \sum_{t=1}^T \eta_t \mathbb{I}_{[j_t=i]} \prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2) = \frac{2G^2}{n} \sum_{t=1}^T \eta_t \sum_{i=1}^n \mathbb{I}_{[j_t=i]} \prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2) \\
&= \frac{2G^2}{n} \sum_{t=1}^T \eta_t \prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2) = \frac{4G^2}{n\lambda} \sum_{t=1}^T \big(1-(1-\eta_t\lambda/2)\big) \prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2\big) \\
&= \frac{4G^2}{n\lambda} \sum_{t=1}^T \Big(\prod_{t'=t+1}^T (1-\eta_{t'}\lambda/2) - \prod_{t'=t}^T (1-\eta_{t'}\lambda/2)\Big) = \frac{4G^2}{n\lambda}\Big(1 - \prod_{t=1}^T (1-\eta_t\lambda/2)\Big),
\end{aligned}$$

where we have used $\sum_{i=1}^n \mathbb{I}_{[j_t=i]} = 1$ for any $t \in [T]$. The proof is completed. □

7. Conclusion

In this paper, we introduce the pointwise uniform stability to develop high-probability generalization bounds. The pointwise uniform stability considers the effect of changing each example in the dataset, which is weaker than the uniform stability. We first develop a moment bound for a summation of weakly-dependent vector-valued random variables, and apply it to develop bounds for the generalization gap as measured by either function values or gradients. We improve the recently fast high-probability rates in [18] by relaxing the requirement on strong convexity and Lipschitz continuity, and improving the dependency on optimization errors. Finally, we apply our results to develop improved generalization bounds for SGD.

Our generalization bounds involve a factor of $\log(n)$ in front of $\left(\frac{1}{n}\sum_{i=1}^{n}\beta_i^2\right)^{\frac{1}{2}}$. A very interesting question is to see whether this logarithmic factor can be removed. Indeed, if we can remove this logarithmic factor, the resulting generalization bound would be optimal up to a constant factor. It is also interesting to apply the stability analysis to study SGD with functional data [8, 16].

Acknowledgements

The authors are grateful to the associate editor and the anonymous reviewers for their thoughtful comments and constructive suggestions. The work by Jun Fan is partially supported by the Research Grants Council of Hong Kong [Project No. HKBU 12302819] and [Project No. HKBU 12303220], and Hong Kong Baptist University [Project No. RC-FNRA-IG/22-23/SCI/02]. The work by Yunwen Lei is partially supported by the Research Grants Council of Hong Kong [Project No. 22303723] and URC Seed Fund for Basic Research for New Staff 2023-24.

Appendix

A. Useful Inequalities in Probability

A.1. McDiarmid's Inequality

We first consider Mcdiarmid's inequality for real-valued functions of random variables, which follows from the standard tail-bound of McDiarmid's inequality and Proposition 2.5.2 in [40].

Lemma A.1 (McDiarmid's Inequality for Real-Valued Functions)**.** *Let $Z_1,\ldots,Z_n$ be independent random variables, and $f:\mathcal{Z}^n\mapsto\mathbb{R}$ such that the following inequality holds for any i and $z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_n$*

$$\sup_{z_i,z_i'}\left|f(z_1,\ldots,z_{i-1},z_i,z_{i+1},\ldots,z_n)-f(z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n)\right|\leq\beta_i.$$

Then for any $p\geq1$ we have

$$\left\|f(Z_1,\ldots,Z_n)-\mathbb{E}[f(Z_1,\ldots,Z_n)]\right\|_p\leq\Big(2p\sum_{i=1}^{n}\beta_i^2\Big)^{\frac{1}{2}}.$$

Now we consider **vector-valued** functions of independent random variables. The following lemma gives the expected distance between $f(Z_1, \ldots, Z_n)$ and its expectation.

Lemma A.2 ([31]). *Let $Z_1, \ldots, Z_n$ be independent random variables, and $f : \mathcal{Z}^n \mapsto \mathcal{H}$ a function into a Hilbert space $\mathcal{H}$ such that the following inequality holds for any i and $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n$*

$$\sup_{z_i, z_i'} \big\| f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n) \big\| \leq \beta_i. \tag{A.1}$$

Then

$$\mathbb{E}\big[\big\| f(Z_1, \ldots, Z_n) - \mathbb{E}[f(Z_1, \ldots, Z_n)] \big\|\big] \leq \Big(\sum_{i=1}^n \beta_i^2 \Big)^{\frac{1}{2}}.$$

The following lemma controls the p-norm for the vector-valued random variable $f(Z_1, \ldots, Z_n) - \mathbb{E}[f(Z_1, \ldots, Z_n)]$.

Lemma A.3 (McDiarmid's Inequality for Vector-Valued Functions). *Let assumptions in Lemma A.2 hold. Then for any $p \geq 1$ we have*

$$\Big\| \big\| f(Z_1, \ldots, Z_n) - \mathbb{E}[f(Z_1, \ldots, Z_n)] \big\| \Big\|_p \leq (\sqrt{2} + 1) \Big(p \sum_{i=1}^n \beta_i^2 \Big)^{\frac{1}{2}}. \tag{A.2}$$

Proof. We define a real-valued function $g : \mathcal{Z}^n \mapsto \mathbb{R}$ as

$$g(z_1, \ldots, z_n) = \big\| f(z_1, \ldots, z_n) - \mathbb{E}[f(Z_1, \ldots, Z_n)] \big\|.$$

We first show this function satisfies the increment condition. Indeed, for any i and $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n$ we have

$$\begin{aligned}
&\sup_{z_i, z_i'} \big| g(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n) - g(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n) \big| \\
&= \sup_{z_i, z_i'} \Big| \big\| f(z_1, \ldots, z_n) - \mathbb{E}[f(Z_1, \ldots, Z_n)] \big\| - \big\| f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n) - \mathbb{E}[f(Z_1, \ldots, Z_n)] \big\| \Big| \\
&\leq \sup_{z_i, z_i'} \Big\| f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n) \Big\| \leq \beta_i.
\end{aligned}$$

Therefore, we can apply Lemma A.1 to the real-valued function g and derive the following inequality

$$\Big\| g(Z_1, \ldots, Z_n) - \mathbb{E}[g(Z_1, \ldots, Z_n)] \Big\|_p \leq \Big(2p \sum_{i=1}^n \beta_i^2 \Big)^{\frac{1}{2}}.$$

According to Lemma A.2, we know the following inequality $\mathbb{E}\big[g(Z_1, \ldots, Z_n)\big] \leq \big(\sum_{i=1}^n \beta_i^2 \big)^{\frac{1}{2}}$. We can combine the above two inequalities together and derive the stated inequality. □

A.2. Bernstein Inequality and Tails

The following lemma gives a Bernstein inequality to incorporate the variance information in bounding a summation of independent random variables [9].

Lemma A.4 (Bernstein inequality)**.** *Let $\{\xi(z_i)\}_{i=1}^n$ be a sequence of independent and identically distributed 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 probability at least $1 - \delta$ there holds*

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

The following lemma shows the relationship between tails and moments [6].

Lemma A.5. *Let Y be a random variable. If $\|Y\|_p \le \sqrt{p}a$ for any $p \ge 2$, then for any $\delta \in (0,1)$ with probability at least $1-\delta$: $|Y| \le ea\sqrt{\log(e/\delta)}$.*

B. Self-Bounding Property

We present some useful self-bounding properties for functions of a composite structure in Assumption 1. The self-bounding property will be very important for our stability and generalization analysis.

Lemma B.1. *Assume $F(\mathbf{w}) = L(\mathbf{w}) + r(\mathbf{w})$, where L has (α, L_α)-Hölder continuous gradients and r is L_r-smooth. Then we have*

$$\|\nabla F(\mathbf{w})\| \le \frac{L_r(1+\alpha)^{\frac{1}{1+\alpha}}}{2L_\alpha^{\frac{1}{1+\alpha}}}\Big(F(\mathbf{w}) - F(\mathbf{w}^*)\Big)^{\frac{1}{1+\alpha}} + 2\Big(\frac{L_\alpha}{1+\alpha}\Big)^{\frac{1}{1+\alpha}}\Big(F(\mathbf{w}) - F(\mathbf{w}^*)\Big)^{\frac{\alpha}{1+\alpha}}.$$

Lemma B.2. *Assume $F(\mathbf{w}) = L(\mathbf{w}) + r(\mathbf{w})$, where L has (α, L_α)-Hölder continuous gradients and r is L_r-smooth. If F is nonnegative, then we have*

$$\|\nabla F(\mathbf{w})\| \le \frac{L_r(1+\alpha)^{\frac{1}{1+\alpha}}}{2L_\alpha^{\frac{1}{1+\alpha}}} F^{\frac{1}{1+\alpha}}(\mathbf{w}) + 2\Big(\frac{L_\alpha}{1+\alpha}\Big)^{\frac{1}{1+\alpha}} F^{\frac{\alpha}{1+\alpha}}(\mathbf{w}).$$

Proof. If $\nabla F(\mathbf{w}) = 0$, the inequality holds immediately. Now we only consider the case that $\nabla F(\mathbf{w}) \neq 0$. Since L has (α, L_α)-Hölder continuous gradients, we know $L(\mathbf{w}') \le L(\mathbf{w}) + \langle \mathbf{w}' - \mathbf{w}, \nabla L(\mathbf{w})\rangle + \frac{L_\alpha}{1+\alpha}\|\mathbf{w} - \mathbf{w}'\|^{1+\alpha}$ [43]. Since r is L_r-smooth, we know $r(\mathbf{w}') \le r(\mathbf{w}) + \langle \mathbf{w}' - \mathbf{w}, \nabla r(\mathbf{w})\rangle + \frac{L_r}{2}\|\mathbf{w} - \mathbf{w}'\|^2$ [26]. It then follows that

$$F(\mathbf{w}') \le F(\mathbf{w}) + \langle \mathbf{w}' - \mathbf{w}, \nabla F(\mathbf{w})\rangle + \frac{L_\alpha}{1+\alpha}\|\mathbf{w} - \mathbf{w}'\|^{1+\alpha} + \frac{L_r}{2}\|\mathbf{w} - \mathbf{w}'\|^2. \tag{B.1}$$

We choose

$$\mathbf{w}' = \mathbf{w} - A\|\nabla F(\mathbf{w})\|^{-1}\nabla F(\mathbf{w}), \quad A := \Big(\frac{(1+\alpha)F(\mathbf{w})}{L_\alpha}\Big)^{\frac{1}{1+\alpha}}.$$

It then follows that

$$0 \le F(\mathbf{w}') \le F(\mathbf{w}) - A\|\nabla F(\mathbf{w})\| + \frac{L_r A^2}{2} + \frac{L_\alpha A^{1+\alpha}}{1+\alpha}.$$

That is,

$$\|\nabla F(\mathbf{w})\| \le \frac{L_r A}{2} + \frac{L_\alpha A^\alpha}{1+\alpha} + \frac{F(\mathbf{w})}{A}.$$

According to our construction of A, we further have

$$\begin{aligned}\|\nabla F(\mathbf{w})\| &\leq \frac{L_r}{2}\Big(\frac{(1+\alpha)F(\mathbf{w})}{L_\alpha}\Big)^{\frac{1}{1+\alpha}} + \frac{L_\alpha}{1+\alpha}\Big(\frac{(1+\alpha)F(\mathbf{w})}{L_\alpha}\Big)^{\frac{\alpha}{1+\alpha}} + F(\mathbf{w})\Big(\frac{L_\alpha}{(1+\alpha)F(\mathbf{w})}\Big)^{\frac{1}{1+\alpha}} \\ &= \frac{L_r}{2}\Big(\frac{(1+\alpha)F(\mathbf{w})}{L_\alpha}\Big)^{\frac{1}{1+\alpha}} + 2\Big(\frac{L_\alpha}{1+\alpha}\Big)^{\frac{1}{1+\alpha}} F^{\frac{\alpha}{1+\alpha}}(\mathbf{w}).\end{aligned}$$

The proof is completed. □

We now prove Lemma B.1 as a direct corollary of Lemma B.2.

Proof of Lemma B.1. Define $\widetilde{F} : \mathcal{W} \mapsto \mathbb{R}$ as $\widetilde{F}(\mathbf{w}) = F(\mathbf{w}) - F(\mathbf{w}^*)$. It is clear that $\widetilde{F}(\mathbf{w}) \geq 0$ and $\widetilde{F}(\mathbf{w}) = \widetilde{L}(\mathbf{w}) + r(\mathbf{w})$, where $\widetilde{L}(\mathbf{w}) = L(\mathbf{w}) - F(\mathbf{w}^*)$ and $\widetilde{L}(\mathbf{w})$ has (α, L_α)-Hölder continuous gradients. Therefore, we can apply Lemma B.2 to $\widetilde{F}$ and derive

$$\|\nabla \widetilde{F}(\mathbf{w})\| \leq \frac{L_r(1+\alpha)^{\frac{1}{1+\alpha}}}{2L_\alpha^{\frac{1}{1+\alpha}}}\widetilde{F}^{\frac{1}{1+\alpha}}(\mathbf{w}) + 2\Big(\frac{L_\alpha}{1+\alpha}\Big)^{\frac{1}{1+\alpha}}\widetilde{F}^{\frac{\alpha}{1+\alpha}}(\mathbf{w}).$$

The stated bound then follows directly. The proof is completed. □

The following lemma gives the self-bounding property for a nonnegative function with Hölder continuous gradients [35, 43].

Lemma B.3. *Assume the map $\mathbf{w} \mapsto g(\mathbf{w})$ is nonnegative, and $\mathbf{w} \mapsto \nabla g(\mathbf{w})$ is (α, L_α)-Hölder continuous with $\alpha \in [0,1]$. Let c_α be defined in Eq.* (3.9)*. Then*

$$\|\nabla g(\mathbf{w})\|_2 \leq c_\alpha g^{\frac{\alpha}{1+\alpha}}(\mathbf{w}), \quad \forall \mathbf{w} \in \mathcal{W}. \tag{B.2}$$

C. Proof of Theorem 4

Proof of Theorem 4. Let $S = \{z_1, \ldots, z_n\}, S' = \{z'_1, \ldots, z'_n\}, S'' = \{z''_1, \ldots, z''_n\}$ be drawn independently from ρ. For any $i \in [n] := \{1, \ldots, n\}$, define

$$g_i(S) = \mathbb{E}_{z'_i}\Big[\mathbb{E}_Z\big[\ell(A(S^{(i)}); Z)\big] - \ell(A(S^{(i)}); z_i)\Big],$$

where $S^{(i)}$ is defined in Eq. (5.8). Due to the symmetry between S and S', we have

$$\begin{aligned}\mathbb{E}_{S\backslash z_i}[g_i(S)] &= \mathbb{E}_{S\backslash z_i}\mathbb{E}_{z'_i}\Big[\mathbb{E}_Z\big[\ell(A(S^{(i)}); Z)\big] - \ell(A(S^{(i)}); z_i)\Big] \\ &= \mathbb{E}_{S^{(i)}}[L(A(S^{(i)}))] - \mathbb{E}_{S^{(i)}}[\ell(A(S^{(i)}); z_i)] = \mathbb{E}_S[L(A(S))] - \mathbb{E}_{S^{(i)}}[\ell(A(S^{(i)}); z_i)].\end{aligned} \tag{C.1}$$

According to the definition of pointwise uniform stability, we know

$$\begin{aligned}&\Big|\mathbb{E}_Z\big[\ell(A(S); Z)\big] - \frac{1}{n}\sum_{i=1}^n \ell(A(S); z_i) - \frac{1}{n}\sum_{i=1}^n g_i(S)\Big| \\ &\leq \frac{1}{n}\sum_{i=1}^n \Big|\mathbb{E}_Z\big[\ell(A(S); Z)\big] - \mathbb{E}_{z'_i, Z}\big[\ell(A(S^{(i)}); Z)\big]\Big| + \frac{1}{n}\sum_{i=1}^n \Big|\ell(A(S); z_i) - \mathbb{E}_{z'_i}\big[\ell(A(S^{(i)}); z_i)\big]\Big| \\ &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{z'_i, Z}\Big[\Big|\ell(A(S); Z) - \ell(A(S^{(i)}); Z)\Big|\Big] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{z'_i}\Big[\Big|\ell(A(S); z_i) - \ell(A(S^{(i)}); z_i)\Big|\Big] \leq \frac{2}{n}\sum_{i=1}^n \beta_i.\end{aligned}$$

It then follows that

$$\begin{aligned}
&\Big|L(A(S)) - L_S(A(S)) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S\backslash z_i}[g_i(S)]\Big| \\
&= \Big|L(A(S)) - L_S(A(S)) - \frac{1}{n}\sum_{i=1}^n \Big(\mathbb{E}_{S\backslash z_i}[g_i(S)] - g_i(S) + g_i(S)\Big)\Big| \\
&\le \Big|L(A(S)) - L_S(A(S)) - \frac{1}{n}\sum_{i=1}^n g_i(S)\Big| + \Big|\frac{1}{n}\sum_{i=1}^n \Big(\mathbb{E}_{S\backslash z_i}[g_i(S)] - g_i(S)\Big)\Big| \\
&\le \frac{2}{n}\sum_{i=1}^n \beta_i + \frac{1}{n}\Big|\sum_{i=1}^n h_i(S)\Big|, \qquad \text{(C.2)}
\end{aligned}$$

where we introduce

$$h_i(S) = g_i(S) - \mathbb{E}_{S\backslash z_i}[g_i(S)], \quad \forall i \in [n].$$

We now show that the above h_i satisfies the conditions in Theorem 1. According to the definition of h_i, we know that

$$\mathbb{E}_{S\backslash z_i}[h_i(S)] = \mathbb{E}_{S\backslash z_i}[g_i(S)] - \mathbb{E}_{S\backslash z_i}[g_i(S)] = 0. \qquad \text{(C.3)}$$

It is clear that

$$\mathbb{E}_{z_i}[g_i(S)] = \mathbb{E}_{z_i}\Big[\mathbb{E}_{z_i'}\Big[\mathbb{E}_Z\big[\ell(A(S^{(i)}); Z)\big] - \ell(A(S^{(i)}); z_i)\Big]\Big] = \mathbb{E}_{z_i'}\Big[\mathbb{E}_Z\big[\ell(A(S^{(i)}); Z)\big] - \mathbb{E}_{z_i}[\ell(A(S^{(i)}); z_i)]\Big] = 0.$$

It then follows that

$$\mathbb{E}_{z_i}[h_i(S)] = \mathbb{E}_{z_i}[g_i(S)] - \mathbb{E}_{S\backslash z_i}\mathbb{E}_{z_i}[g_i(S)] = 0. \qquad \text{(C.4)}$$

Finally, for any $j \in [n]$ with $j \ne i$, we have

$$\begin{aligned}
&\big|h_i(S) - h_i(z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\big| \qquad \text{(C.5)} \\
&= \Big|\Big(g_i(S) - \mathbb{E}_{S\backslash z_i}[g_i(S)]\Big) - \Big(g_i(S_j'') - \mathbb{E}_{S_j''\backslash z_i}[g_i(S_j'')]\Big)\Big| \\
&\le \big|g_i(S) - g_i(S_j'')\big| + \Big|\mathbb{E}_{S\backslash z_i}[g_i(S)] - \mathbb{E}_{S_j''\backslash z_i}[g_i(S_j'')]\Big| \\
&\le \big|g_i(S) - g_i(S_j'')\big| + \mathbb{E}_{S\backslash z_i}\mathbb{E}_{S_j''\backslash z_i}\Big|g_i(S) - g_i(S_j'')\Big|, \qquad \text{(C.6)}
\end{aligned}$$

where

$$S_j'' = \big\{z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n\big\}, \quad \forall j \in [n].$$

Note that

$$\begin{aligned}
&\big|g_i(S) - g_i(S_j'')\big| \\
&= \Big|\Big(\mathbb{E}_{z_i'}\Big[\mathbb{E}_Z\big[\ell(A(S^{(i)}); Z)\big] - \ell(A(S^{(i)}); z_i)\Big]\Big) - \mathbb{E}_{z_i'}\Big[\mathbb{E}_Z\big[\ell(A(S_j^{(i)}); Z)\big] - \ell(A(S_j^{(i)}); z_i)\Big]\Big| \\
&\le \Big|\mathbb{E}_{z_i'}\mathbb{E}_Z\big[\ell(A(S^{(i)}); Z)\big] - \mathbb{E}_{z_i'}\mathbb{E}_Z\big[\ell(A(S_j^{(i)}); Z)\big]\Big| + \Big|\mathbb{E}_{z_i'}[\ell(A(S^{(i)}); z_i)] - \mathbb{E}_{z_i'}[\ell(A(S_j^{(i)}); z_i)]\Big| \le 2\beta_j,
\end{aligned}$$

where $S_j^{(i)}$ is defined in Eq. (5.10). We combine the above inequality and Eq. (C.6) together to get

$$\big|h_i(S) - h_i(z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\big| \le 4\beta_j.$$

According to Eq. (C.3), (C.4) and the above inequality, the conditions of Theorem 1 hold with $M = 0$. Therefore, we can apply Theorem 1 to derive the following inequality

$$\Big|\sum_{i=1}^n h_i(S)\Big| \lesssim p \log_2 n\Big(n\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}}.$$

It then follows the following inequality with probability at least $1-\delta/2$

$$\Big|\sum_{i=1}^n h_i(S)\Big| \lesssim \Big(n\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}} \log n \log(1/\delta).$$

The above inequality together with Eq. (C.1), (C.2) gives the following inequality with probability at least $1-\delta/2$

$$\begin{aligned}&\Big|L(A(S)) - L_S(A(S)) - \mathbb{E}_S[L(A(S))] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S^{(i)}}[\ell(A(S^{(i)}); z_i)]\Big|\\ &= \Big|L(A(S)) - L_S(A(S)) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S\backslash z_i}[g_i(S)]\Big| \lesssim \frac{1}{n}\sum_{i=1}^n \beta_i + \Big(\frac{1}{n}\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}} \log n \log(1/\delta). \qquad \text{(C.7)}\end{aligned}$$

We have the following identity

$$\begin{aligned}L(A(S)) - L_S(A(S)) - L(\mathbf{w}^*) + L_S(\mathbf{w}^*) &= \Big(L(A(S)) - L_S(A(S)) - \mathbb{E}_S[L(A(S))] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S'}\big[\ell(A(S'); z_i)\big]\Big)\\ &+ \Big(\mathbb{E}_S[L(A(S))] - L(\mathbf{w}^*) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S'}\big[\ell(A(S'); z_i)\big] + L_S(\mathbf{w}^*)\Big). \qquad \text{(C.8)}\end{aligned}$$

The first term can be controlled by Eq. (C.7) and the identity $\mathbb{E}_{S^{(i)}}[\ell(A(S^{(i)}); z_i)] = \mathbb{E}_{S'}[\ell(A(S'); z_i)]$. We now control the second term by Bernstein's inequality. To this aim, we introduce $\xi(z) = \mathbb{E}_{S'}\big[\ell(A(S'); z)\big] - \ell(\mathbf{w}^*; z)$. Due to the symmetry between S and S', we further get

$$\begin{aligned}&\mathbb{E}_S[L(A(S))] - L(\mathbf{w}^*) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S'}\big[\ell(A(S'); z_i)\big] + L_S(\mathbf{w}^*)\\ &= \mathbb{E}_{S'}[L(A(S'))] - L(\mathbf{w}^*) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S'}\big[\ell(A(S'); z_i)\big] + L_S(\mathbf{w}^*) = \mathbb{E}_Z[\xi(Z)] - \frac{1}{n}\sum_{i=1}^n \xi(z_i).\end{aligned}$$

We can control the variance of ξ as follows

$$\begin{aligned}\mathrm{Var}(\xi(Z)) &= \mathbb{E}_Z\Big[\Big(\mathbb{E}_{S'}\big[\ell(A(S'); Z)\big] - \ell(\mathbf{w}^*; Z)\Big)^2\Big] - \Big(\mathbb{E}_Z\Big[\mathbb{E}_{S'}\big[\ell(A(S'); Z)\big] - \ell(\mathbf{w}^*; Z)\Big]\Big)^2\\ &= \mathbb{E}_Z\Big[\Big(\mathbb{E}_S\big[\ell(A(S); Z)\big] - \ell(\mathbf{w}^*; Z)\Big)^2\Big] - \Big(\mathbb{E}_S\big[L(A(S))\big] - L(\mathbf{w}^*)\Big)^2,\end{aligned}$$

where we have used the symmetry between S and S'. According to Bernstein's inequality (Lemma A.4), the following inequality holds with probability at least $1-\delta/2$

$$\mathbb{E}_Z[\xi(Z)] - \frac{1}{n}\sum_{i=1}^n \xi(z_i) \le \frac{2M\log\frac{2}{\delta}}{3n} + \Big(\frac{2\sigma_A^2 \log(2/\delta)}{n}\Big)^{\frac{1}{2}}.$$

We can plug the above inequality and Eq. (C.7) into Eq. (C.8), and derive the following inequality with probability at least $1-\delta$

$$L(A(S)) - L_S(A(S)) - L(\mathbf{w}^*) + L_S(\mathbf{w}^*) \lesssim \Big(\frac{1}{n}\sum_{i=1}^n \beta_i^2\Big)^{\frac{1}{2}} \log n \log(1/\delta) + \frac{M\log\frac{1}{\delta}}{n} + \Big(\frac{\sigma_A^2\log(1/\delta)}{n}\Big)^{\frac{1}{2}}.$$

The proof is completed by noting the structure of f. $\square$

D. Proof of Lemma 5

Proof of Lemma 5. We first prove Eq. (3.10). Since F is λ-strongly convex, we know

$$F(A(S)) - F(\mathbf{w}^*) \geq \lambda\|A(S) - \mathbf{w}^*\|^2/2. \tag{D.1}$$

According to the definition of σ_A^2, we know $\sigma_A^2 \leq \mathbb{E}_Z\big[\big(\mathbb{E}_S\big[\ell(A(S);Z)\big] - \ell(\mathbf{w}^*;Z)\big)^2\big]$. Since ℓ is convex, we know

$$\big|\ell(A(S);Z) - \ell(\mathbf{w}^*;Z)\big| \leq \|A(S) - \mathbf{w}^*\| \max\big\{\|\nabla\ell(A(S);Z)\|, \|\nabla\ell(\mathbf{w}^*;Z)\|\big\}.$$

It then follows from Eq. (B.2) that

$$\begin{aligned}\Big(\mathbb{E}_S\big[\ell(A(S);Z)\big] - \ell(\mathbf{w}^*;Z)\Big)^2 &\leq \mathbb{E}_S[\|A(S) - \mathbf{w}^*\|^2]\mathbb{E}_S\big[\max\big\{\|\nabla\ell(A(S);Z)\|^2, \|\nabla\ell(\mathbf{w}^*;Z)\|^2\big\}\big]\\ &\leq c_\alpha^2\mathbb{E}_S[\|A(S) - \mathbf{w}^*\|^2]\mathbb{E}_S\big[\max\big\{\ell^{\frac{2\alpha}{1+\alpha}}(A(S);Z), \ell^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*;Z)\big\}\big].\end{aligned}$$

Therefore, we have

$$\sigma_A^2 \leq c_\alpha^2\mathbb{E}_S[\|A(S) - \mathbf{w}^*\|^2]\mathbb{E}_{S,Z}\big[\max\big\{\ell^{\frac{2\alpha}{1+\alpha}}(A(S);Z), \ell^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*;Z)\big\}\big].$$

Eq. (3.10) then follows by combining the above inequality and Eq. (D.1) together.

We now turn to Eq. (3.12) on generalization bounds in expectation. We first study the generalization error for the algorithm A_{e}. By the definition $F_S, S, S^{(i)}$, we get

$$f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i) = nF_S(A_{\mathrm{e}}(S^{(i)})) - nF_{S^{(i)}}(A_{\mathrm{e}}(S^{(i)})) + f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i').$$

By symmetry on $\mathbf{z}_i$ and $\mathbf{z}_i'$, we get $\mathbb{E}[f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i')] = \mathbb{E}[f(A_{\mathrm{e}}(S);\mathbf{z}_i)], \mathbb{E}[F_{S^{(i)}}(A_{\mathrm{e}}(S^{(i)}))] = \mathbb{E}[F_S(A_{\mathrm{e}}(S))]$ and

$$\mathbb{E}\big[f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i) - f(A_{\mathrm{e}}(S);\mathbf{z}_i)\big] = n\mathbb{E}\big[F_S(A_{\mathrm{e}}(S^{(i)})) - F_{S^{(i)}}(A_{\mathrm{e}}(S^{(i)}))\big] = n\mathbb{E}\Big[F_S(A_{\mathrm{e}}(S^{(i)})) - F_S(A_{\mathrm{e}}(S))\Big].$$

Since F_S is λ-strongly convex, we further know $F_S(A_{\mathrm{e}}(S^{(i)})) - F_S(A_{\mathrm{e}}(S)) \leq \|\nabla F_S(A_{\mathrm{e}}(S^{(i)}))\|^2/(2\lambda)$ [26]. We can combine the above two inequalities to get

$$\mathbb{E}\big[f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i) - f(A_{\mathrm{e}}(S);\mathbf{z}_i)\big] \leq \frac{n}{2\lambda}\mathbb{E}\big[\|\nabla F_S(A_{\mathrm{e}}(S^{(i)}))\|^2\big]. \tag{D.2}$$

The definition of $A_{\mathrm{e}}(S^{(i)})$ implies $\nabla F_{S^{(i)}}(A_{\mathrm{e}}(S^{(i)})) = 0$, and

$$\begin{aligned}\mathbb{E}\Big[\|\nabla F_S(A_{\mathrm{e}}(S^{(i)}))\|^2\Big] &= \mathbb{E}\Big[\Big\|\nabla F_{S^{(i)}}(A_{\mathrm{e}}(S^{(i)})) - \frac{1}{n}\nabla f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i') + \frac{1}{n}\nabla f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i)\Big\|^2\Big]\\ &= \frac{1}{n^2}\mathbb{E}\Big[\Big\|\nabla f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i') - \nabla f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i)\Big\|^2\Big] = \frac{1}{n^2}\mathbb{E}\Big[\Big\|\nabla f(A_{\mathrm{e}}(S);\mathbf{z}_i) - \nabla f(A_{\mathrm{e}}(S);\mathbf{z}_i')\Big\|^2\Big],\end{aligned}$$

where the last step is due to the symmetry between $\mathbf{z}_i$ and $\mathbf{z}_i'$. We can combine the above inequality and Eq (D.2) together to derive

$$\begin{aligned}\mathbb{E}\big[F(A_{\mathrm{e}}(S)) - F_S(A_{\mathrm{e}}(S))\big] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(A_{\mathrm{e}}(S^{(i)});\mathbf{z}_i) - f(A_{\mathrm{e}}(S);\mathbf{z}_i)\big]\\ &\leq \frac{1}{2n^2\lambda}\sum_{i=1}^n \mathbb{E}\Big[\Big\|\nabla f(A_{\mathrm{e}}(S);\mathbf{z}_i) - \nabla f(A_{\mathrm{e}}(S);\mathbf{z}_i')\Big\|^2\Big],\end{aligned} \tag{D.3}$$

where we have used $\mathbb{E}\big[F(A_{\mathrm{e}}(S))\big] = \mathbb{E}[F(A_{\mathrm{e}}(S^{(i)}))] = \mathbb{E}[f(A_{\mathrm{e}}(S^{(i)}); \mathbf{z}_i)]$. By the structure of f, we get

$$\begin{aligned}\Big\|\nabla f(A_{\mathrm{e}}(S); \mathbf{z}_i) - \nabla f(A_{\mathrm{e}}(S); \mathbf{z}'_i)\Big\|^2 &= \Big\|\nabla \ell(A_{\mathrm{e}}(S); \mathbf{z}_i) - \nabla \ell(A_{\mathrm{e}}(S); \mathbf{z}'_i)\Big\|^2 \\ &\le 2\|\nabla \ell(A_{\mathrm{e}}(S); \mathbf{z}_i)\|^2 + 2\|\nabla \ell(A_{\mathrm{e}}(S); \mathbf{z}'_i)\|^2 \le 2c_\alpha^2 \ell^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S); \mathbf{z}_i) + 2c_\alpha^2 \ell^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S); \mathbf{z}'_i).\end{aligned}$$

It then follows that

$$\frac{1}{n}\sum_{i=1}^n \Big\|\nabla f(A_{\mathrm{e}}(S); \mathbf{z}_i) - \nabla f(A_{\mathrm{e}}(S); \mathbf{z}'_i)\Big\|^2 \le \frac{2c_\alpha^2}{n}\sum_{i=1}^n \ell^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S); \mathbf{z}_i) + \frac{2c_\alpha^2}{n}\sum_{i=1}^n \ell^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S); \mathbf{z}'_i).$$

We can use the above inequality, the concavity of the function $x \mapsto x^{\frac{2\alpha}{1+\alpha}}$ and Eq. (D.3) to derive

$$\mathbb{E}\big[F(A_{\mathrm{e}}(S)) - F_S(A_{\mathrm{e}}(S))\big] \le \frac{c_\alpha^2}{n\lambda}\mathbb{E}\Big[L_S^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S)) + L_{S'}^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S))\Big] \le \frac{c_\alpha^2}{n\lambda}\mathbb{E}\Big[L_S^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S)) + L^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S))\Big]. \tag{D.4}$$

By Eq. (5.12) and the Cauchy-Schwartz's inequality, we know

$$\mathbb{E}\big[\langle A(S) - A_{\mathrm{e}}(S), \nabla F(A_{\mathrm{e}}(S))\big] \le C_1\Big(\mathbb{E}[\|A(S) - A_{\mathrm{e}}(S)\|^2]\Big)^{\frac{1}{2}}\Big(\mathbb{E}\Big[\Big(F(A_{\mathrm{e}}(S)) - F(\mathbf{w}^*)\Big)^{\frac{2\alpha}{1+\alpha}}\Big]\Big)^{\frac{1}{2}}.$$

Since $\mathbb{E}[F_S(A_{\mathrm{e}}(S))] \le F(\mathbf{w}^*)$, we can plug the above inequality back into Eq. (5.13), and derive

$$\begin{aligned}\mathbb{E}\big[F(A(S)) - F(A_{\mathrm{e}}(S))\big] &\le C_1\Big(\mathbb{E}[\|A(S) - A_{\mathrm{e}}(S)\|^2]\Big)^{\frac{1}{2}}\Big(\Big(\mathbb{E}\Big[F(A_{\mathrm{e}}(S)) - F_S(A_{\mathrm{e}}(S))\Big]\Big)^{\frac{2\alpha}{1+\alpha}}\Big)^{\frac{1}{2}} \\ &\quad + \frac{L_\alpha \mathbb{E}[\|A(S) - A_{\mathrm{e}}(S)\|^{1+\alpha}]}{1+\alpha} + \frac{L_r \mathbb{E}[\|A(S) - A_{\mathrm{e}}(S)\|^2]}{2},\end{aligned}$$

where we have used the concavity of $x \mapsto x^{\frac{2\alpha}{1+\alpha}}$ and the Jensen's inequality. We can plug Eq. (D.4) and Eq. (5.14) into the above inequality to show

$$\begin{aligned}\mathbb{E}\big[F(A(S)) - F(A_{\mathrm{e}}(S))\big] &\le \frac{L_\alpha \mathbb{E}[\big(2\lambda^{-1}(F_S(A(S)) - F_S(A_{\mathrm{e}}(S)))\big)^{\frac{1+\alpha}{2}}]}{1+\alpha} + \frac{L_r \mathbb{E}[F_S(A(S)) - F_S(A_{\mathrm{e}}(S))]}{\lambda} \\ &\quad + C_1\Big(2\lambda^{-1}\mathbb{E}[F_S(A(S)) - F_S(A_{\mathrm{e}}(S))]\Big)^{\frac{1}{2}}\Big(\frac{c_\alpha^2}{n\lambda}\mathbb{E}\Big[L_S^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S)) + L^{\frac{2\alpha}{1+\alpha}}(A_{\mathrm{e}}(S))\Big]\Big)^{\frac{\alpha}{1+\alpha}}.\end{aligned}$$

According to Eq. (D.4), the concavity of the function $x \mapsto x^{\frac{1+\alpha}{2}}$ and the decomposition

$$\mathbb{E}\big[F(A(S)) - F_S(A_{\mathrm{e}}(S))\big] = \mathbb{E}\big[F(A(S)) - F(A_{\mathrm{e}}(S))\big] + \mathbb{E}\big[F(A_{\mathrm{e}}(S)) - F_S(A_{\mathrm{e}}(S))\big],$$

we further get

$$\begin{aligned}\mathbb{E}\big[F(A(S)) - F_S(A_{\mathrm{e}}(S))\big] &\le \frac{2^{\frac{1+\alpha}{2}}L_\alpha}{1+\alpha}\Delta_\lambda^{\frac{1+\alpha}{2}} + L_r\Delta_\lambda + \sqrt{2}C_1 c_\alpha^{\frac{2\alpha}{1+\alpha}}\Delta_\lambda^{\frac{1}{2}}\nabla_\lambda^{\frac{\alpha}{1+\alpha}} + c_\alpha^2\nabla_\lambda \\ &\le \Big(\frac{2^{\frac{1+\alpha}{2}}L_\alpha}{1+\alpha} + \frac{\sqrt{2}C_1 c_\alpha^{\frac{2\alpha}{1+\alpha}}}{1+\alpha}\Big)\Delta_\lambda^{\frac{1+\alpha}{2}} + L_r\Delta_\lambda + \Big(\frac{\sqrt{2}C_1 c_\alpha^{\frac{2\alpha}{1+\alpha}}\alpha}{1+\alpha} + c_\alpha^2\Big)\nabla_\lambda,\end{aligned}$$

where we have used the Young's inequality $\Delta_\lambda^{\frac{1}{2}}\nabla_\lambda^{\frac{\alpha}{1+\alpha}} \le \frac{\alpha}{1+\alpha}\nabla_\lambda^{\frac{\alpha}{1+\alpha}\frac{1+\alpha}{\alpha}} + \frac{1}{1+\alpha}\Delta_\lambda^{\frac{\alpha+1}{2}}$. The proof is completed by noting $\mathbb{E}[F_S(A_{\mathrm{e}}(S))] \le \mathbb{E}[F_S(\mathbf{w}^*)] = F(\mathbf{w}^*)$ and

$$\mathfrak{C} = \max\Big\{\frac{2^{\frac{1+\alpha}{2}}L_\alpha}{1+\alpha} + \frac{\sqrt{2}C_1 c_\alpha^{\frac{2\alpha}{1+\alpha}}}{1+\alpha}, L_r, \frac{\sqrt{2}C_1 c_\alpha^{\frac{2\alpha}{1+\alpha}}\alpha}{1+\alpha} + c_\alpha^2\Big\}. \tag{D.5}$$

The proof is completed. $\square$

References

[1] I. Amir, Y. Carmon, T. Koren, and R. Livni. Never go full batch (in stochastic convex optimization). *Advances in Neural Information Processing Systems*, 34:25033–25043, 2021.

[2] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[3] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

[4] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

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

[6] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

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

[8] X. Chen, B. Tang, J. Fan, and X. Guo. Online gradient descent algorithms for functional data learning. *Journal of Complexity*, 70:101635, 2022.

[9] F. Cucker and D.-X. Zhou. *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press, 2007.

[10] V. De la Pena and E. Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

[11] L. Devroye and T. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

[12] A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(Jan):55–79, 2005.

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

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

[15] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[16] X. Guo, Z.-C. Guo, and L. Shi. Capacity dependent analysis for functional online learning algorithms. *Applied and Computational Harmonic Analysis*, 67:101567, 2023.

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

[18] Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] T. Koren, R. Livni, Y. Mansour, and U. Sherman. Benign underfitting of stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2022.

[20] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2820–2829, 2018.

[21] Y. Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *Annual Conference on Learning Theory*, pages 191–227, 2023.

[22] Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819, 2020.

[23] Y. Lei and Y. Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021.

[24] Y. Lei, R. Jin, and Y. Ying. Stability and generalization analysis of gradient methods for shallow neural networks. *Advances in Neural Information Processing Systems*, 35:38557–38570, 2022.

[25] W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638, 2018.

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

[27] G. Neu, G. K. Dziugaite, M. Haghifam, and D. M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

[28] K. Nikolakakis, F. Haddadpour, D. Kalogerias, and A. Karbasi. Black-box generalization: Stability of zeroth-order learning. *Advances in Neural Information Processing Systems*, 35:31525–31541, 2022.

[29] M. Raab and A. Steger. "balls into bins"—a simple and tight analysis. In *International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 159–170. Springer, 1998.

[30] D. Richards and I. Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. *Advances in Neural Information Processing Systems*, 34, 2021.

[31] O. Rivasplata, E. Parrado-Hernández, J. S. Shawe-Taylor, S. Sun, and C. Szepesvári. Pac-bayes bounds for stable algorithms with instance-dependent priors. In *Advances in Neural Information Processing Systems*, pages 9214–9224, 2018.

[32] W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

[33] M. Schliserman and T. Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory*, pages 3380–3394, 2022.

[34] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[35] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

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

[37] H. Taheri and C. Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *arXiv preprint arXiv:2302.09235*, 2023.

[38] E. Ullah, T. Mai, A. Rao, R. A. Rossi, and R. Arora. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pages 4126–4142. PMLR, 2021.

[39] V. Vapnik. *The nature of statistical learning theory*. Springer, 2013.

[40] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47.

Cambridge university press, 2018.

[41] S. Villa, S. Matet, B. C. Vũ, and L. Rosasco. Implicit regularization with strongly convex bias: Stability and acceleration. *Analysis and Applications*, pages 1–27, 2022.

[42] P. Wang, Y. Lei, Y. Ying, and H. Zhang. Differentially private sgd with non-smooth losses. *Applied and Computational Harmonic Analysis*, 56:306–336, 2022.

[43] Y. Ying and D.-X. Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017.

[44] Y. Zhang, W. Zhang, S. Bald, V. Pingali, C. Chen, and M. Goswami. Stability of sgd: Tightness analysis and improved bounds. In *Uncertainty in artificial intelligence*, pages 2364–2373. PMLR, 2022.

[45] D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

[46] L. Zhu, M. Gurbuzbalaban, A. Raj, and U. Simsekli. Uniform-in-time wasserstein stability bounds for (noisy) stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2023.