Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent

Yunwen Lei [1 2] **Yiming Ying** [3]

Abstract

Recently there are a considerable amount of work devoted to the study of the algorithmic stability and generalization for stochastic gradient descent (SGD). However, the existing stability analysis requires to impose restrictive assumptions on the boundedness of gradients, smoothness and convexity of loss functions. In this paper, we provide a fine-grained analysis of stability and generalization for SGD by substantially relaxing these assumptions. Firstly, we establish stability and generalization for SGD by removing the existing bounded gradient assumptions. The key idea is the introduction of a new stability measure called *on-average model stability*, for which we develop novel bounds controlled by the risks of SGD iterates. This yields generalization bounds depending on the behavior of the best model, and leads to the *first-ever-known fast bounds* in the low-noise setting using stability approach. Secondly, the smoothness assumption is relaxed by considering loss functions with Hölder continuous (sub)gradients for which we show that optimal bounds are still achieved by balancing computation and stability. To our best knowledge, this gives the *first-ever-known* stability and generalization bounds for SGD with *non-smooth* loss functions (e.g., hinge loss). Finally, we study learning problems with (strongly) convex objectives but non-convex loss functions.

1. Introduction

Stochastic gradient descent (SGD) has become the workhorse behind many machine learning problems. As an iterative algorithm, SGD updates the model sequentially upon receiving a new datum with a cheap per-iteration cost, making it amenable for big data analysis. There is a plethora of theoretical work on its convergence analysis as an optimization algorithm (e.g. Duchi et al., 2011; Lacoste-Julien et al., 2012; Nemirovski et al., 2009; Rakhlin et al., 2012; Shamir & Zhang, 2013; Zhang, 2004).

Concurrently, there are a considerable amount of work with focus on its generalization analysis (Dieuleveut & Bach, 2016; Hardt et al., 2016; Lin et al., 2016; Rosasco & Villa, 2015; Ying & Zhou, 2016). For instance, using the tool of integral operator the work (Dieuleveut & Bach, 2016; Lin & Rosasco, 2017; Rosasco & Villa, 2015; Ying & Pontil, 2008) studied the excess generalization error of SGD with the least squares loss, i.e. the difference between the true risk of SGD iterates and the best possible risk. An advantage of this approach is its ability to capture the regularity of regression functions and the capacity of hypothesis spaces. The results were further extended in Lei & Tang (2018); Lin et al. (2016) based on tools of empirical processes which are able to deal with general convex functions even without a smoothness assumption. The idea is to bound the complexity of SGD iterates in a controllable manner, and apply concentration inequalities in empirical processes to control the uniform deviation between population risks and empirical risks over a ball to which the SGD iterates belong.

Recently, in the seminal work (Hardt et al., 2016) the authors studied the generalization bounds of SGD via algorithmic stability (Bousquet & Elisseeff, 2002; Elisseeff et al., 2005) for convex, strongly convex and non-convex problems. This motivates several appealing work on some weaker stability measures of SGD that still suffice for guaranteeing generalization (Charles & Papailiopoulos, 2018; Kuzborskij & Lampert, 2018; Zhou et al., 2018). An advantage of this stability approach is that it considers only the particular model produced by the algorithm, and can imply generalization bounds independent of the dimensionality.

However, the existing stability analysis of SGD is established under the strong assumptions on the loss function such as the boundedness of the gradient and strong smoothness. Such assumptions are very restrictive which are not satisfied in many standard contexts. For example, the bounded

[1]Department of Computer Science, University of Kaiserslautern, Germany [2]School of Computer Science, University of Birmingham, United Kingdom [3]Department of Mathematics and Statistics, State University of New York at Albany, USA. Correspondence to: Yiming Ying <yying@albany.edu>.

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

gradient assumption does not hold for the simple least-squares regression, where the model parameter belongs to an unbounded domain. The strong smoothness assumption does not hold for the popular support vector machine. Furthermore, the analysis in the strongly convex case requires strong convexity of each loss function which is not true for many problems such as the important problem of least squares regression.

In this paper, we provide a fine-grained analysis of stability and generalization for SGD. Our new results remove the bounded gradient assumption for differentiable loss functions and remove the smoothness assumption for Lipschitz continuous loss functions, and therefore broaden the impact of the algorithmic stability approach for generalization analysis of SGD. In summary, our main contributions are listed as follows.

I think it's direct from using smoothness to bound gradients i.e.

- Firstly, we study stability and generalization for SGD by removing the existing bounded gradient assumptions. The key is an introduction of a novel stability measure called on-average model stability, whose connection to generalization is established by using the smoothness of loss functions able to capture the low risks of output models for better generalization. An advantage of on-average model stability is that the corresponding bounds involve a weighted sum of empirical risks instead of the uniform Lipschitz constant. The weighted sum of empirical risks can be bounded via tools in analyzing optimization errors, which implies a key message that optimization is beneficial to generalization. Furthermore, our stability analysis allows us to develop generalization bounds depending on the risk of the best model. In particular, we have established fast generalization bounds $O(1/n)$ for the setting of low noises, where n is the sample size. To our best knowledge, this is the first fast generalization bound of SGD based on stability approach in a low-noise setting.

- Secondly, we consider loss functions with their (sub)gradients satisfying the Hölder continuity which is a much weaker condition than the strong smoothness in the literature. Although stability decreases by weakening the smoothness assumption, optimal generalization bounds can be surprisingly achieved by balancing computation and stability. In particular, we show that an optimal generalization bound can be achieved for the hinge loss by running SGD with $O(n^2)$ iterations. To our best knowledge, this is the first generalization analysis of SGD for non-smooth loss functions based on algorithmic stability. Fast learning rates are further derived in the low-noise case.

- Thirdly, we study learning problems with (strongly) convex objectives but non-convex individual loss functions. The nonconvexity of loss functions makes the corresponding gradient update no longer non-expansive, and therefore the arguments in Hardt et al. (2016) do not apply. We bypass this obstacle by developing a novel quadratic inequality of the stability using only the convexity of the objective, which shows that this relaxation affects neither generalization nor computation.

The paper is structured as follows. We discuss the related work in Section 2 and formulate the problem in Section 3. The stability and generalization for learning with convex loss functions is presented in Section 4. In Sections 5 and 6, we consider problems with relaxed convexity and relaxed strong convexity, respectively. We conclude the paper in Section 7.

2. Related Work

In this section, we discuss related work on algorithmic stability, stability of stochastic optimization algorithms and generalization error of SGD.

Algorithmic Stability. The study of stability can be dated back to Rogers & Wagner (1978). A modern framework of quantifying generalization via stability was established in the paper (Bousquet & Elisseeff, 2002), where a concept of uniform stability was introduced and studied for empirical risk minimization (ERM) in the strongly convex setting. This framework was then extended to study randomized learning algorithms (Elisseeff et al., 2005), transfer learning (Kuzborskij & Lampert, 2018) and privacy-preserving learning (Bassily et al., 2019; Dwork & Feldman, 2018), etc. The interplay between various notions of stability, learnability and consistency was further studied (Rakhlin et al., 2005; Shalev-Shwartz et al., 2010). The power of stability analysis is especially reflected by its ability in deriving optimal generalization bounds in expectation (Shalev-Shwartz et al., 2010). Very recently, almost optimal high-probability generalization bounds were established via the stability approach (Bousquet et al., 2019; Feldman & Vondrak, 2018; 2019). In addition to the notion of uniform stability mentioned above, various other notions of stability were recently introduced, including uniform argument stability (Liu et al., 2017) and hypothesis set stability (Foster et al., 2019).

Stability of Stochastic Optimization Algorithms. In the seminal paper (Hardt et al., 2016), the co-coercivity of gradients was used to study the uniform stability of SGD in convex, strongly convex and non-convex problems. The uniform stability was relaxed to a weaker notion of on-average stability (Shalev-Shwartz et al., 2010), for which the corresponding bounds of SGD can capture the impact of the risk at the initial point (Kuzborskij & Lampert, 2018) and the variance of stochastic gradients (Zhou et al., 2018). For non-convex learning problems satisfying either a gradient dominance or a quadratic growth condition, pointwise-hypothesis stabilities were studied for a class of learning algorithms that converge to global optima (Charles & Papailiopoulos, 2018),

which relaxes and extends the uniform stability of ERM under strongly convex objectives (Bousquet & Elisseeff, 2002). A fundamental stability and convergence trade-off of iterative optimization algorithms was recently established, where it was shown that a faster converging algorithm can not be too stable, and vice versa (Chen et al., 2018). This together with some uniform stability bounds for several first-order algorithms established there, immediately implies new convergence lower bounds for the corresponding algorithms. Algorithmic stability was also established for stochastic gradient Langevin dynamics with non-convex objectives (Li et al., 2020; Mou et al., 2018) and SGD implemented in a stagewise manner (Yuan et al., 2019).

Generalization Analysis of SGD. A framework to study the generalization performance of large-scale stochastic optimization algorithms was established in Bousquet & Bottou (2008), where three factors influencing generalization behavior were identified as optimization errors, estimation errors and approximation errors. Uniform stability was used to establish generalization bounds $O(1/\sqrt{n})$ in expectation for SGD for convex and strongly smooth cases (Hardt et al., 2016). For convex and nonsmooth learning problems, generalization bounds $O(n^{-\frac{1}{3}})$ were established based on the uniform convergence principle (Lin et al., 2016). An interesting observation is that an implicit regularization can be achieved without an explicit regularizer by tuning either the number of passes or the step sizes (Lin et al., 2016; Rosasco & Villa, 2015). For the specific least squares loss, optimal excess generalization error bounds (up to a logarithmic factor) were established for SGD based on the integral operator approach (Lin & Rosasco, 2017; Pillaud-Vivien et al., 2018). The above mentioned generalization results are in the form of expectation. High-probability bounds were established based on either an uniform-convergence approach (Lei & Tang, 2018) or an algorithmic stability approach (Feldman & Vondrak, 2019). A novel combination of PAC-Bayes and algorithmic stability was used to study the generalization behavior of SGD, a promising property of which is its applications to all posterior distributions of algorithms' random hyperparameters (London, 2017).

3. Problem Formulation

Let $S = \{z_1, \ldots, z_n\}$ be a set of training examples independently drawn from a probability measure ρ defined over a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is an input space and $\mathcal{Y} \subseteq \mathbb{R}$ is an output space. Our aim is to learn a prediction function parameterized by $\mathbf{w} \in \Omega \subseteq \mathbb{R}^d$ to approximate the relationship between an input variable x and an output variable y. We quantify the loss of a model $\mathbf{w}$ on an example $z = (x, y)$ by $f(\mathbf{w}; z)$. The corresponding empirical and population risks are respectively given by

$$F_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}; z_i) \quad \text{and} \quad F(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}; z)].$$

Here we use $\mathbb{E}_z[\cdot]$ to denote the expectation with respect to (w.r.t.) z. In this paper, we consider stochastic learning algorithms A, and denote by $A(S)$ the model produced by running A over the training examples S.

We are interested in studying the *excess generalization error* $F(A(S)) - F(\mathbf{w}^*)$, where $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\Omega} F(\mathbf{w})$ is the one with the best prediction performance over Ω. It can be decomposed as

$$\begin{aligned}\mathbb{E}_{S,A}\big[F(A(S)) - F(\mathbf{w}^*)\big] = &\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \\ &+ \mathbb{E}_{S,A}\big[F_S(A(S)) - F_S(\mathbf{w}^*)\big]. \quad (3.1)\end{aligned}$$

The first term is called the estimation error due to the approximation of the unknown probability measure ρ based on sampling. The second term is called the optimization error induced by running an optimization algorithm to minimize the empirical objective, which can be addressed by tools in optimization theory. A popular approach to control estimation errors is to consider the stability of the algorithm, for which a widely used stability measure is the uniform stability (Elisseeff et al., 2005; Hardt et al., 2016).

Definition 1 (Uniform Stability)**.** A stochastic algorithm A is ϵ-uniformly stable if for all training datasets $S, \widetilde{S} \in \mathcal{Z}^n$ that differ by at most one example, we have

$$\sup_z \mathbb{E}_A\big[f(A(S); z) - f(A(\widetilde{S}); z)\big] \leq \epsilon. \quad (3.2)$$

The celebrated relationship between generalization and uniform stability is established in the following lemma (Hardt et al., 2016; Shalev-Shwartz et al., 2010).

Lemma 1 (Generalization via uniform stability)**.** *Let A be ϵ-uniformly stable. Then*

$$\big|\mathbb{E}_{S,A}\big[F_S(A(S)) - F(A(S))\big]\big| \leq \epsilon.$$

Throughout the paper, we restrict our interest to a specific algorithm called projected stochastic gradient descent. It is worth mentioning that our main results in Section 4 hold also when $\Omega = \mathbb{R}^d$, i.e., no projections.

Definition 2 (Projected Stochastic Gradient Descent)**.** Let $\Omega \subseteq \mathbb{R}^d$ and Π_Ω denote the projection on Ω. Let $\mathbf{w}_1 = 0 \in \mathbb{R}^d$ be an initial point and $\{\eta_t\}_t$ be a sequence of positive step sizes. Projected SGD updates models by

$$\mathbf{w}_{t+1} = \Pi_\Omega\big(\mathbf{w}_t - \eta_t \partial f(\mathbf{w}_t; z_{i_t})\big), \quad (3.3)$$

where $\partial f(\mathbf{w}_t, z_{i_t})$ denotes a subgradient of f w.r.t. the first argument and i_t is independently drawn from the uniform distribution over $\{1, \ldots, n\}$.

Note if f is differentiable, then ∂f denotes the gradient of f w.r.t. the first argument. We say a function $g : \mathbb{R}^d \mapsto \mathbb{R}$ is σ-strongly convex if

$$g(\mathbf{w}) \geq g(\tilde{\mathbf{w}}) + \langle \mathbf{w} - \tilde{\mathbf{w}}, \partial g(\tilde{\mathbf{w}}) \rangle + \frac{\sigma}{2}\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 \quad (3.4)$$

for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d$, where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\|\mathbf{w}\|_2$ denotes the ℓ_2 norm of $\mathbf{w} = (w_1, \ldots, w_d)$, i.e., $\|\mathbf{w}\|_2 = \big(\sum_{j=1}^d w_j^2\big)^{\frac{1}{2}}$. If (3.4) holds with $\sigma = 0$, then we say g is convex. We denote $B \asymp \widetilde{B}$ if there are absolute constants c_1 and c_2 such that $c_1 B \leq \widetilde{B} \leq c_2 B$.

4. Stability with Convexity

An essential assumption to establish the uniform stability of SGD is the uniform Lipschitz continuity (boundedness of gradients) of loss functions as follows (Bousquet & Elisseeff, 2002; Charles & Papailiopoulos, 2018; Hardt et al., 2016; Kuzborskij & Lampert, 2018; Zhou et al., 2018).

they don't assume this

Assumption 1. We assume $\|\partial f(\mathbf{w}; z)\|_2 \leq G$ for all $\mathbf{w} \in \Omega$ and $z \in \mathcal{Z}$.

→ They just point the weaknes

Unfortunately, the Lipschitz constant G can be very large or even infinite for some learning problems. Consider the simple least squares loss $f(\mathbf{w}; z) = \frac{1}{2}(\langle \mathbf{w}, x \rangle - y)^2$ with the gradient $\partial f(\mathbf{w}; z) = (\langle \mathbf{w}, x \rangle - y)x$. In this case the G-Lipschitzness of f requires to set $G = \sup_{\mathbf{w}\in\Omega} \sup_{z\in\mathcal{Z}} \|(\langle \mathbf{w}, x \rangle - y)x\|_2$, which is infinite if Ω is unbounded. As another example, the Lipschitz constant of deep neural networks can be prohibitively large. In this case, existing stability bounds fail to yield meaningful generalization bounds. Furthermore, another critical assumption in the literature is the L-smoothness on f, i.e. for any z and $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d$

$$\big\|\partial f(\mathbf{w}, z) - \partial f(\tilde{\mathbf{w}}, z)\big\|_2 \leq L\|\mathbf{w} - \tilde{\mathbf{w}}\|_2. \quad (4.1)$$

In this section, we will remove the boundedness assumption on the gradients for differentiable loss functions, and establish stability and generalization only under the assumption where loss functions have Hölder continuous (sub)gradients– a condition much weaker than the strong smoothness (Lei et al., 2018; Nesterov, 2015; Ying & Zhou, 2017). Note that the loss functions can be *non-differentiable* if $\alpha = 0$.

Definition 3. Let $L > 0, \alpha \in [0, 1]$. We say ∂f is (α, L)-Hölder continuous if for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,

$$\big\|\partial f(\mathbf{w}, z) - \partial f(\tilde{\mathbf{w}}, z)\big\|_2 \leq L\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^\alpha. \quad (4.2)$$

If (4.2) holds with $\alpha = 1$, then f is smooth as defined by (4.1). If (4.2) holds with $\alpha = 0$, then this amounts to saying that f is Lipschitz continuous as considered in Assumption 1. Examples of loss functions satisfying Definition 3 include the q-norm hinge loss $f(\mathbf{w}; z) = \big(\max(0, 1 - y\langle \mathbf{w}, x \rangle)\big)^q$ for classification and the q-th power absolute distance loss $f(\mathbf{w}; z) = |y - \langle \mathbf{w}, x \rangle|^q$ for regression (Steinwart & Christmann, 2008), whose (sub)gradients are $(q-1, C)$-Hölder continuous for some $C > 0$ if $q \in [1, 2]$. If $q = 1$, we get the hinge loss and absolute distance loss with wide applications in machine learning and statistics.

4.1. On-average model stability

The key to remove the bounded gradient assumption and the strong smoothness assumption is the introduction of a novel stability measure which we refer to as the on-average model stability. We use the term "on-average model stability" to differentiate it from on-average stability in Kearns & Ron (1999); Shalev-Shwartz et al. (2010) as we measure stability on *model parameters* $\mathbf{w}$ instead of *function values*. Intuitively, on-average model stability measures the on-average sensitivity of models by traversing the perturbation of each single coordinate.

Definition 4 (On-average Model Stability). Let $S = \{z_1, \ldots, z_n\}$ and $\widetilde{S} = \{\tilde{z}_1, \ldots, \tilde{z}_n\}$ be drawn independently from ρ. For any $i = 1, \ldots, n$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, \tilde{z}_i, z_{i+1}, \ldots, z_n\}$ as the set formed from S by replacing the i-th element with $\tilde{z}_i$. We say a randomized algorithm A is ℓ_1 on-average model ϵ-stable if

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n \|A(S) - A(S^{(i)})\|_2\Big] \leq \epsilon,$$

and ℓ_2 on-average model ϵ-stable if

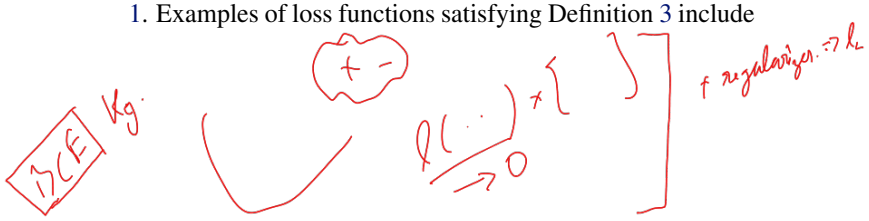
$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n \|A(S) - A(S^{(i)})\|_2^2\Big] \leq \epsilon^2.$$

In the following theorem, we build the connection between generalization in expectation and the on-average model stabilities to be proved in Appendix B. Although the generalization by ℓ_1 on-average model stability requires Assumption 1, it is removed for ℓ_2 on-average model stability. We introduce a free parameter γ to tune according to the property of problems. Note we require a convexity assumption in Part (c) by considering non-smooth loss functions. Let $c_{\alpha,1} = (1 + 1/\alpha)^{\frac{\alpha}{1+\alpha}} L^{\frac{1}{1+\alpha}}$ if $\alpha > 0$ and $c_{\alpha,1} = \sup_z \|\partial f(0; z)\|_2 + L$ if $\alpha = 0$.

Theorem 2 (Generalization via Model Stability). *Let $S, \widetilde{S}$ and $S^{(i)}$ be constructed as Definition 4. Let $\gamma > 0$.*

(a) Let A be ℓ_1 on-average model ϵ-stable and Assumption 1 hold. Then

$$\big|\mathbb{E}_{S,A}\big[F_S(A(S)) - F(A(S))\big]\big| \leq G\epsilon.$$



So can we workout an example where if loss is of order $O(\frac{1}{n})$ the L_2 is much much smaller ?

(b) *If for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and L-smooth, then*

$$\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq \frac{L}{\gamma}\mathbb{E}_{S,A}\big[F_S(A(S))\big] + \frac{L+\gamma}{2n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},A}\big[\|A(S^{(i)}) - A(S)\|_2^2\big].$$

(c) *If for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and $\mathbf{w} \mapsto \partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1]$, then*

$$\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq \frac{c_{\alpha,1}^2}{2\gamma}\mathbb{E}_{S,A}\Big[F^{\frac{2\alpha}{1+\alpha}}(A(S))\Big] + \frac{\gamma}{2n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},A}\big[\|A(S^{(i)}) - A(S)\|_2^2\big].$$

Remark 1. We explain here the benefit of ℓ_2 on-average model stability. If A is ℓ_2 on-average model ϵ-stable, then we take $\gamma = \sqrt{2L\mathbb{E}\big[F_S(A(S))\big]}/\epsilon$ in Part (b) and derive

$$\mathbb{E}\big[F(A(S)) - F_S(A(S))\big] \leq L\epsilon^2/2 + \sqrt{2L\mathbb{E}\big[F_S(A(S))\big]}\epsilon.$$

In particular, if the output model has a small empirical risk in the sense of $\mathbb{E}\big[F_S(A(S))\big] = O(1/n)$, we derive $\mathbb{E}\big[F(A(S)) - F_S(A(S))\big] = O(\epsilon^2 + \epsilon/\sqrt{n})$. That is, our relationship between the generalization and ℓ_2 on-average stability allows us to exploit small risk of output model to get a generalization bound with an improved dependency on the stability measure ϵ. As a comparison, the discussions based on uniform stability (Lemma 1) and the ℓ_1 on-average model stability (Part (a)) only show $\mathbb{E}\big[F(A(S)) - F_S(A(S))\big] = O(\epsilon)$, which fail to exploit the low-noise condition. We can also take $\gamma = c_{\alpha,1}\big(\mathbb{E}[F(A(S))]\big)^{\frac{\alpha}{1+\alpha}}/\epsilon$ in part (c) to derive

$$\mathbb{E}\big[F(A(S)) - F_S(A(S))\big] = O\Big(\epsilon\big(\mathbb{E}[F(A(S))]\big)^{\frac{\alpha}{1+\alpha}}\Big).$$

The above equation can be written as an inequality of $\mathbb{E}[F(A(S)) - F_S(A(S))]$ (using the sub-additivity of $t \mapsto t^{\frac{\alpha}{1+\alpha}}$), from which we derive

$$\mathbb{E}\big[F(A(S)) - F_S(A(S))\big] = O\Big(\epsilon^{1+\alpha} + \epsilon\big(\mathbb{E}[F_S(A(S))]\big)^{\frac{\alpha}{1+\alpha}}\Big).$$

If $\mathbb{E}[F_S(A(S))]$ is small, this also implies an improved dependency of the generalization bound on ϵ.

4.2. Strongly smooth case

To justify the effectiveness of the on-average model stability, we first consider its application to learning with smooth loss functions. We first study stability and then generalization.

Stability bounds. The following theorem to be proved in Appendix C.1 establishes on-average model stability bounds in the smooth setting. A key difference from the existing stability bounds is that the uniform Lipschitz constant G is replaced by empirical risks. Since we are minimizing empirical risks by SGD, it is expected that these risks would be significantly smaller than the uniform Lipschitz constant. Actually we will control the weighted sum of empirical risks by tools in analyzing optimization errors. In the optimistic case with $F(\mathbf{w}^*) = 0$, we expect $\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)] = O(1/t)$, and in this case the discussion based on on-average model stability would imply significantly better generalization bounds. The idea of introducing a parameter p in (4.4) is to make $(1 + p/n)^t \leq e$ by setting $p = n/t$, where e is the base of the nature logarithm.

Theorem 3 (Stability bounds)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth. Let $S, \widetilde{S}$ and $S^{(i)}$ be constructed as Definition 4. Let $\{\mathbf{w}_t\}$ and $\{\mathbf{w}_t^{(i)}\}$ be produced by (3.3) with $\eta_t \leq 2/L$ based on S and $S^{(i)}$, respectively. Then for any $p > 0$ we have*

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2\Big] \leq \frac{2\sqrt{2L}}{n}\sum_{j=1}^{t}\eta_j\mathbb{E}_{S,A}\Big[\sqrt{F_S(\mathbf{w}_j)}\Big]. \quad (4.3)$$

and

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\Big] \leq \frac{8(1+p^{-1})L}{n}\sum_{j=1}^{t}(1+p/n)^{t-j}\eta_j^2\mathbb{E}_{S,A}\big[F_S(\mathbf{w}_j)\big]. \quad (4.4)$$

Remark 2. Kuzborskij & Lampert (2018) developed an interesting on-average stability bound $O(\frac{\tilde{\sigma}}{n}\sum_{j=1}^{t}\eta_j)$ under the bounded variance assumption $\mathbb{E}_{S,z}\big[\|\partial f(\mathbf{w}_t; z) - \partial F(\mathbf{w}_t; z)\|_2^2\big] \leq \tilde{\sigma}^2$ for all t. Although this bound successfully replaces the uniform Lipschitz constant by the milder uniform variance constant $\tilde{\sigma}$, the corresponding generalization analysis still requires a bounded gradient assumption. A nice property of the stability bound in Kuzborskij & Lampert (2018) is that it depends on the quality of the initialization, i.e., the stability improves if we start with a good model. Our stability bound also enjoys this property. As we can see from Theorem 3, the stability increases if we find good models with small optimization errors in the optimization process. This illustrates a key message that optimization is beneficial to improve the generalization.

Remark 3. The stability bounds in Theorem 3 can be extended to the non-convex case. Specifically, let assumptions of Theorem 3, except the convexity of $\mathbf{w} \mapsto f(\mathbf{w}; z)$, hold.

Then for any $p > 0$ one gets (see Proposition C.3)

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\Big] \leq (1+p/n)(1+\eta_t L)^2\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\Big] + \frac{8(1+p^{-1})L\eta_t^2}{n}\mathbb{E}_{S,A}\big[F_S(\mathbf{w}_t)\big].$$

This result improves the recurrence relationship in Hardt et al. (2016) for uniform stability by replacing the uniform Lipschitz constant with empirical risks.

Generalization bounds. We now establish generalization bounds based on ℓ_2 on-average model stability. This approach not only removes a bounded gradient assumption, but also allows us to fully exploit the smoothness of loss functions to derive bounds depending on the behavior of the best model $\mathbf{w}^*$. As we will see in Corollary 5, Theorem 4 interpolates between $O(1/\sqrt{n})$ bound in the "pessimistic" case ($F(\mathbf{w}^*) > 0$) and the $O(1/n)$ bound in the "low-noise" case ($F(\mathbf{w}^*) = 0$) (Reeve & Kabán, 2020; Srebro et al., 2010), which is becoming more and more interesting in the deep learning era with possibly more parameters than training examples. To our best knowledge, this is the first optimistic bound for SGD based on a stability approach. Eq. (4.6) still holds if $F(\mathbf{w}^*) = O(1/n)$. The proofs are given in Appendix C.2.

Theorem 4 (Generalization bounds)**.** *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth. Let $\{\mathbf{w}_t\}$ be produced by (3.3) with nonincreasing step sizes satisfying $\eta_t \leq 1/(2L)$. If $\gamma \geq 1$, then*

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O\Big(\Big(\frac{1}{\gamma}+\frac{\sum_{t=1}^T\eta_t^2}{\sum_{t=1}^T\eta_t}\Big)F(\mathbf{w}^*) + \frac{1}{\sum_{t=1}^T\eta_t} + \frac{\gamma(1+T/n)}{n}\big(1+\sum_{t=1}^T\eta_t^2F(\mathbf{w}^*)\big)\Big),$$

where $\mathbf{w}_T^{(1)} = \big(\sum_{t=1}^T\eta_t\mathbf{w}_t\big)/\sum_{t=1}^T\eta_t$.

Corollary 5. *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth.*

(a) Let $\{\mathbf{w}_t\}$ be produced by (3.3) with $\eta_t = c/\sqrt{T} \leq 1/(2L)$ for a constant $c > 0$. If $T \asymp n$, then

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O\Big(\frac{F(\mathbf{w}^*)+1}{\sqrt{n}}\Big). \quad (4.5)$$

(b) Let $\{\mathbf{w}_t\}$ be produced by (3.3) with $\eta_t = \eta_1 \leq 1/(2L)$. If $F(\mathbf{w}^*) = 0$ and $T \asymp n$, then

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O(1/n). \quad (4.6)$$

Remark 4. Based on the stability bound in Hardt et al. (2016), we can show $\mathbb{E}_{S,A}\big[F(\mathbf{w}_T^{(1)})\big] - F(\mathbf{w}^*)$ decays as

$$\frac{2G^2\sum_{t=1}^T\eta_t}{n} + O\Big(\frac{\sum_{t=1}^T\eta_t^2F(\mathbf{w}^*)+1}{\sum_{t=1}^T\eta_t}\Big), \quad (4.7)$$

from which one can derive the $O(1/\sqrt{n})$ bound at best even if $F(\mathbf{w}^*) = 0$. The improvement of our bounds over (4.7) is due to the consideration of on-average model stability bounds involving empirical risks (we use the same optimization error bounds in these two approaches). Based on the on-average stability bound in Kuzborskij & Lampert (2018), one can derive a generalization bound similar to (4.7) with G^2 replaced by $G\tilde{\sigma}$ ($\tilde{\sigma}$ is the uniform variance constant in Remark 2), which also could not yield a fast bound $O(1/n)$ if $F(\mathbf{w}^*) = 0$.

Remark 5. We compare here our results with some fast bounds for SGD. Some fast convergence rates of SGD were recently derived for SGD under low noise conditions (Bassily et al., 2018; Ma et al., 2018; Srebro et al., 2010) or growth conditions relating stochastic gradients to full gradients (Vaswani et al., 2019). The discussions there mainly focused on optimization errors, which are measured w.r.t. the iteration number t. As a comparison, our fast rates measured by n are developed for generalization errors of SGD (Part (b) of Corollary 5), for which we need to trade-off optimization errors and estimation errors by stopping at an appropriate iteration number. Fast generalization bounds are also established for the specific least squares based on an integral operator approach (Dieuleveut et al., 2017; Lin & Rosasco, 2017; Mücke et al., 2019; Pillaud-Vivien et al., 2018). However, these discussions heavily depend on the structure of the square loss and require capacity assumptions in terms of the decay rate of eigenvalues for the associated integral operator. As a comparison, we consider general loss functions and do not impose a capacity assumption.

4.3. Non-smooth case

As a further application, we apply our on-average model stability to learning with non-smooth loss functions (e.g., the hinge loss), which have not been studied in the literature.

Stability bounds. We first present stability bounds in Theorem 7. As compared to (4.4), the stability bound below involves an additional term $O(\sum_{j=1}^t\eta_j^{\frac{2}{1-\alpha}})$, which is the cost we pay by relaxing the smoothness condition to a Hölder continuity of (sub)gradients. Indeed, in this case we show that the gradient update operator $\mathbf{w} \mapsto \mathbf{w} - \eta\partial f(\mathbf{w}; z)$ is approximately contractive in the sense of Lemma 6, which plays a key role in establishing the stability bounds in the non-smooth case. The constant hidden in the big O notation is stated explicitly in (D.4) in the appendix. The proofs of Lemma 6 and Theorem 7 are given in Appendix D.1.

Lemma 6. *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is convex, and $\mathbf{w} \mapsto \partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1)$. Then for all $\mathbf{w} \in \mathbb{R}^d$ and $\eta > 0$ there holds*

$$\|\mathbf{w}-\eta\partial f(\mathbf{w};z)-\tilde{\mathbf{w}}+\eta\partial f(\tilde{\mathbf{w}};z)\|_2^2 = \|\mathbf{w}-\tilde{\mathbf{w}}\|_2^2+O(\eta^{\frac{2}{1-\alpha}}).$$

Theorem 7 (Stability bounds)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and $\partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1)$. Let $S, \widetilde{S}$ and $S^{(i)}$ be constructed in Definition 4. Let $\{\mathbf{w}_t\}$ and $\{\mathbf{w}_t^{(i)}\}$ be produced by (3.3) based on S and $S^{(i)}$, respectively. Then*

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n \|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\Big] = O\Big(\sum_{j=1}^t \eta_j^{\frac{2}{1-\alpha}}\Big)+ O\Big(n^{-1}(1+t/n)\sum_{j=1}^t \eta_j^2 \mathbb{E}_{S,A}\Big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j)\Big]\Big).$$

Note if $\alpha = 0$, the above stability bounds is simplified as

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n \|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\Big] = O\Big((1+t/n^2)\sum_{j=1}^t \eta_j^2\Big).$$

Generalization bounds. We now present generalization bounds for learning by loss functions with Hölder continuous (sub)gradients, which are specific instantiations of a general result (Proposition D.3) stated and proved in Appendix D.2.

Theorem 8 (Generalization bounds)**.** *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex, and $\partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1)$. Let $\{\mathbf{w}_t\}_t$ be given by (3.3) with $\eta_t = cT^{-\theta}, \theta \in [0, 1], c > 0$.*

(a) If $\alpha \geq 1/2$, we can take $\theta = 1/2$ and $T \asymp n$ to derive $\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}})$.

(b) If $\alpha < 1/2$, we can take $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$ and $\theta = \frac{3-3\alpha}{2(2-\alpha)}$ to derive $\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}})$.

(c) If $F(\mathbf{w}^*) = 0$, we take $T \asymp n^{\frac{2}{1+\alpha}}$ and $\theta = \frac{3-\alpha^2-2\alpha}{4}$ to derive $\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O(n^{-\frac{1+\alpha}{2}})$.

Remark 6. Although relaxing smoothness affects stability by introducing $O(\sum_{j=1}^t \eta_j^{\frac{2}{1-\alpha}})$ in the stability bound, we achieve a generalization bound similar to the smooth case with a similar computation cost if $\alpha \geq 1/2$. For $\alpha < 1/2$, a minimax optimal generalization bound $O(n^{-\frac{1}{2}})$ (Agarwal et al., 2012) can be also achieved with more computation cost as $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$. In particular, if $\alpha = 0$ we develop the optimal generalization bounds $O(n^{-\frac{1}{2}})$ for SGD with $T \asymp n^2$ iterations. To our best knowledge, this gives the first generalization bounds for SGD with non-smooth loss functions (e.g., hinge loss) based on stability analysis. Analogous to the smooth case, we can derive generalization bounds better than $O(n^{-\frac{1}{2}})$ in the case with low noises. To our best knowledge, this is the first optimistic generalization bound for SGD with non-smooth loss functions.

Remark 7. We can extend our discussion to ERM. If F_S is σ-strongly convex and $\partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous, we can apply the on-average model stability to show (see Proposition D.5)

$$\mathbb{E}_S\big[F(A(S))-F_S(A(S))\big] = O(\mathbb{E}_S\big[F^{\frac{2\alpha}{1+\alpha}}(A(S))\big]/(n\sigma)),$$

where $A(S) = \arg\min_{\mathbf{w}\in\mathbb{R}^d} F_S(\mathbf{w})$. This extends the error bounds developed for ERM with strongly-smooth loss functions (Shalev-Shwartz & Ben-David, 2014; Srebro et al., 2010) to the non-smooth case, and removes the G-admissibility assumption in Bousquet & Elisseeff (2002). In a low-noise case with a small $\mathbb{E}_S\big[F(A(S))\big]$, the discussion based on an on-average stability can imply optimistic generalization bounds for ERM.

5. Stability with Relaxed Convexity

We now turn to stability and generalization of SGD for learning problems where the empirical objective F_S is convex but each loss function $f(\mathbf{w}; z)$ may be non-convex. For simplicity, we impose Assumption 1 here and use the arguments based on the uniform stability. The proofs of Theorem 9 and Theorem 10 are given in Appendix E.1.

Theorem 9. *Let Assumption 1 hold. Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is L-smooth. Let $S = \{z_1, \ldots, z_n\}$ and $\widetilde{S} = \{\tilde{z}_1, \ldots, \tilde{z}_n\}$ be two sets of training examples that differ by a single example. Let $\{\mathbf{w}_t\}_t$ and $\{\tilde{\mathbf{w}}_t\}_t$ be produced by (3.3) based on S and $\widetilde{S}$, respectively. If for all S, F_S is convex, then*

$$\Big(\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\tilde{\mathbf{w}}_{t+1}\|_2^2\big]\Big)^{\frac{1}{2}} \leq 4GC_t\sum_{j=1}^t \frac{\eta_j}{n} + 2G\Big(C_t\sum_{j=1}^t \frac{\eta_j^2}{n}\Big)^{\frac{1}{2}},$$

where we introduce $C_t = \prod_{j=1}^t \big(1 + L^2\eta_j^2\big)$.

Remark 8. The derivation of uniform stability bounds in Hardt et al. (2016) is based on the non-expansiveness of the operator $\mathbf{w} \mapsto \mathbf{w} - \partial f(\mathbf{w}; z)$, which requires the convexity of $\mathbf{w} \mapsto f(\mathbf{w}; z)$ for all z. Theorem 9 relaxes this convexity condition to a milder convexity condition on F_S. If $\sum_{j=1}^\infty \eta_j^2 < \infty$, the stability bounds in Theorem 9 become $O(n^{-1}\sum_{j=1}^t \eta_j + n^{-\frac{1}{2}})$ since $C_t < \infty$.

As shown below, minimax optimal generalization bounds can be achieved for step sizes $\eta_t = \eta_1 t^{-\theta}$ for all $\theta \in (1/2, 1)$ as well as the step sizes $\eta_t \asymp 1/\sqrt{T}$ with $T \asymp n$.

Theorem 10. *Let Assumption 1 hold. Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is L-smooth. Let $\{\mathbf{w}_t\}_t$ be produced by (3.3). Suppose for all S, F_S is convex.*

(a) If $\eta_t = \eta_1 t^{-\theta}, \theta \in (1/2, 1)$, then

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O\Big(n^{-1}T^{1-\theta} + n^{-\frac{1}{2}} + T^{\theta-1}\Big).$$

If $T \asymp n^{\frac{1}{2-2\theta}}$, then $\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}})$.

(b) If $\eta_t = c/\sqrt{T}$ for some $c > 0$ and $T \asymp n$, then $\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}})$.

Example: AUC Maximization. We now consider a specific example of AUC (Area under ROC curve) maximization where the objective function is convex but each loss function may be non-convex. As a widely used method in imbalanced classification ($\mathcal{Y} = \{+1, -1\}$), AUC maximization was often formulated as a pairwise learning problem where the corresponding loss function involves a pair of training examples (Gao et al., 2013; Zhao et al., 2011). Recently, AUC maximization algorithms updating models with a single example per iteration were developed (Liu et al., 2018; Natole et al., 2018; Ying et al., 2016). Specifically, AUC maximization with the square loss can be formulated as the minimization of the following objective function

$$F(\mathbf{w}) := p(1-p)\mathbb{E}\big[\big(1 - \mathbf{w}^\top(x - \tilde{x})\big)^2 | y = 1, \tilde{y} = -1\big], \tag{5.1}$$

where $p = \Pr\{Y = 1\}$ is the probability of an example being positive. Let $x_+ = \mathbb{E}[X|Y = 1]$ and $x_- = \mathbb{E}[X|Y = -1]$ be the conditional expectation of X given $Y = 1$ and $Y = -1$, respectively. It was shown that $\mathbb{E}_{i_t}\big[f(\mathbf{w}; z_{i_t})\big] = F(\mathbf{w})$ for all $\mathbf{w} \in \mathbb{R}^d$ (Natole et al., 2018, Theorem 1), where

$$\begin{aligned} f(\mathbf{w}; z) =& (1-p)\big(\mathbf{w}^\top(x - x_+)\big)^2 \mathbb{I}_{[y=1]} + p(1-p) \\ &+ 2\big(1 + \mathbf{w}^\top(x_- - x_+)\big)\mathbf{w}^\top x\big(p\mathbb{I}_{[y=-1]} - (1-p)\mathbb{I}_{[y=1]}\big) \\ &+ p\big(\mathbf{w}^\top(x - x_-)\big)^2 \mathbb{I}_{[y=-1]} - p(1-p)\big(\mathbf{w}^\top(x_- - x_+)\big)^2. \end{aligned} \tag{5.2}$$

An interesting property is that (5.2) involves only a single example z. This observation allows Natole et al. (2018) to develop a stochastic algorithm as (3.3) to solve (5.1). However, for each z, the function $z \mapsto f(\mathbf{w}; z)$ is non-convex since the associated Hessian matrix may not be positively definite. It is clear that its expectation F is convex.

6. Stability with Relaxed Strong Convexity

6.1. Stability and generalization errors

Finally, we consider learning problems with strongly convex empirical objectives but possibly non-convex loss functions. Theorem 11 provides stability bounds, while the minimax optimal generalization bounds $O(1/(\sigma n))$ are presented in Theorem 12. The proofs are given in Appendix F.

Theorem 11. *Let Assumptions in Theorem 9 hold. Suppose for all $S \subset \mathcal{Z}$, F_S is σ_S-strongly convex. Then, there exists a constant t_0 such that for SGD with $\eta_t = 2/((t + t_0)\sigma_S)$ we have*

$$\Big(\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2^2\big]\Big)^{\frac{1}{2}} \leq \frac{4G}{\sigma_S}\Big(\frac{1}{\sqrt{n(t + t_0)}} + \frac{1}{n}\Big).$$

Remark 9. Under the assumption $\mathbf{w} \mapsto f(\mathbf{w}, z)$ is σ-strongly convex and smooth for all z, it was shown that $\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2\big] = O(1/(n\sigma))$ for $\eta_t = O(1/(\sigma t))$ (Hardt et al., 2016). Indeed, this strong convexity condition is used to show that the operator $\mathbf{w} \mapsto \mathbf{w} - \partial f(\mathbf{w}; z)$ is contractive. We relax the strong convexity of $f(\mathbf{w}; z)$ to the strong convexity of F_S. Our stability bound holds even if $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is non-convex. If $t \asymp n$, then our stability bound coincides with the one in Hardt et al. (2016) up to a constant factor.

Theorem 12. *Let Assumption 1 hold. Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is L-smooth. Suppose for all $S \subset \mathcal{Z}$, F_S is σ_S-strongly convex. Then, there exists some t_0 such that for SGD with $\eta_t = 2/((t + t_0)\sigma_S)$ and $T \asymp n$ we have*

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(2)})] - F(\mathbf{w}^*) = O(\mathbb{E}_S\big[1/(n\sigma_S)\big]),$$

where $\mathbf{w}_T^{(2)} = \big(\sum_{t=1}^T (t + t_0 - 1)\mathbf{w}_t\big)/\sum_{t=1}^T (t + t_0 - 1)$.

Example: Least Squares Regression. We now consider an application to learning with the least squares loss, where $f(\mathbf{w}; z) = \frac{1}{2}\big(\langle \mathbf{w}, x\rangle - y\big)^2$. Let $\Omega = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq R\}$. In this case, (3.3) becomes

$$\mathbf{w}_{t+1} = \Pi_\Omega\big(\mathbf{w}_t - \eta_t\big(\langle \mathbf{w}_t, x_t\rangle - y_t\big)x_t\big), \tag{6.1}$$

where $\Pi_\Omega(\mathbf{w}) = \min\{R/\|\mathbf{w}\|_2, 1\}\mathbf{w}$. Note that each individual loss function $f(\mathbf{w}_t; z_t)$ is non-strongly convex. However, as we will show below, the empirical objective satisfies a strong convexity on a subspace containing the iterates $\{\mathbf{w}_t\}$. For any $S = \{z_1, \ldots, z_n\}$ let $C_S = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ be the empirical covariance matrix and σ_S' be the minimal positive eigenvalue of C_S. Then it is clear from (6.1) that $\{\mathbf{w}_t\}_t$ belongs to the range of C_S. [1] Let $\widetilde{S} \subset \mathcal{Z}^n$ differ from S by a single example. For simplicity, we assume S and $\widetilde{S}$ differ by the first example and denote $\widetilde{S} = \{\tilde{z}_1, z_2, \ldots, z_n\}$. We construct a set $\bar{S} = \{0, z_2, \ldots, z_n\}$. Let $\{\mathbf{w}_t\}, \{\tilde{\mathbf{w}}_t\}$ and $\{\bar{\mathbf{w}}_t\}$ be the sequence by (6.1) based on $S, \widetilde{S}$ and $\bar{S}$, respectively. Then our previous discussion implies that $\mathbf{w}_t - \bar{\mathbf{w}}_t \in \mathrm{Range}(C_S), \tilde{\mathbf{w}}_t - \bar{\mathbf{w}}_t \in \mathrm{Range}(C_{\widetilde{S}})$ for all $t \in \mathbb{N}$ ($\mathrm{Range}(C_{\bar{S}}) \subseteq \mathrm{Range}(C_S), \mathrm{Range}(C_{\bar{S}}) \subseteq \mathrm{Range}(C_{\widetilde{S}})$), where we denote by $\mathrm{Range}(C)$ the range of a matrix C. It follows that $\mathbf{w}_t - \bar{\mathbf{w}}_t$ and $\tilde{\mathbf{w}}_t - \bar{\mathbf{w}}_t$ are orthogonal to the kernel of C_S and $C_{\widetilde{S}}$, respectively. Therefore,

$$\langle \mathbf{w}_t - \bar{\mathbf{w}}_t, C_S(\mathbf{w}_t - \bar{\mathbf{w}}_t)\rangle \geq \sigma_S'\|\mathbf{w}_t - \bar{\mathbf{w}}_t\|_2^2,$$

[1] The range of C_S is the linear span of $x_1, \ldots, x_n$. Details are given in Proposition F.1 in Appendix F.

$$\langle \tilde{\mathbf{w}}_t - \bar{\mathbf{w}}_t, C_{\widetilde{S}}(\tilde{\mathbf{w}}_t - \bar{\mathbf{w}}_t)\rangle \geq \sigma'_{\widetilde{S}} \|\tilde{\mathbf{w}}_t - \bar{\mathbf{w}}_t\|_2^2.$$

As we will see in the proof, Theorem 11 holds if only the following local strong convexity holds, i.e.,

$$\langle \mathbf{w}_t - \tilde{\mathbf{w}}_t, \partial F_S(\mathbf{w}_t) - \partial F_S(\tilde{\mathbf{w}}_t)\rangle \geq \sigma_S \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2, \ \forall t \in \mathbb{N}.$$

Therefore, we can apply Theorem 11 with $\widetilde{S} = \bar{S}$ and $\sigma_S = \sigma'_S$ to derive (note $\partial F_S(\mathbf{w}) = C_S\mathbf{w} - \frac{1}{n}\sum_{i=1}^n y_i x_i$)

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \bar{\mathbf{w}}_{t+1}\|_2\big] \leq \frac{4G}{\sigma'_S}\Big(\frac{1}{\sqrt{n(t+t_0)}} + \frac{1}{n}\Big).$$

A similar inequality also holds for $\mathbb{E}_A\big[\|\tilde{\mathbf{w}}_{t+1} - \bar{\mathbf{w}}_{t+1}\|_2\big]$, which together with the subadditivity of $\|\cdot\|_2$ immediately gives stability bounds on $\mathbb{E}_A[\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2]$.

7. Conclusions

In this paper, we study stability and generalization of SGD by removing the bounded gradient assumptions, and relaxing the smoothness assumption and the convexity requirement of each loss function in the existing analysis. We introduce a novel on-average model stability able to capture the risks of SGD iterates, which implies fast generalization bounds in the low-noise case and stability bounds for learning with even non-smooth loss functions. For all considered problems, we show that our stability bounds can imply minimax optimal generalization bounds by balancing optimization and estimation errors. We apply our results to practical learning problems to justify the superiority of our approach over the existing stability analysis. Our results can be extended to stochastic proximal gradient descent, high-probability bounds and SGD without replacement (details are given in Appendix G). In the future, it would be interesting to study stability bounds for other stochastic optimization algorithms, e.g., Nesterov's accelerated variants of SGD (Nesterov, 2013).

Acknowledgement

We are grateful to Marius Kloft for interesting discussions. We are also grateful to the anonymous reviewers and area chairs for their insightful and constructive comments. The work of Y. Lei is supported by the National Natural Science Foundation of China (Grant Nos. 61806091, 11771012) and the Alexander von Humboldt Foundation. The work of Y. Ying is supported by the National Science Foundation (NSF) under Grant No. #1816227.

References

Agarwal, A., Bartlett, P., Ravikumar, P., and Wainwright, M. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 5(58):3235–3249, 2012.

Bassily, R., Belkin, M., and Ma, S. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. G. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pp. 11279–11288, 2019.

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

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

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. *arXiv preprint arXiv:1910.07833*, 2019.

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

Chen, Y., Jin, C., and Yu, B. Stability and convergence tradeoff of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

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

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

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Dwork, C. and Feldman, V. Privacy-preserving prediction. In *Conference on Learning Theory*, pp. 1693–1702, 2018.

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

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

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

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems*, pp. 6726–6736. 2019.

Gao, W., Jin, R., Zhu, S., and Zhou, Z.-H. One-pass AUC optimization. In *International Conference on Machine*

Learning, pp. 906–914, 2013.

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

Kearns, M. and Ron, D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, 11(6):1427–1453, 1999.

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

Lacoste-Julien, S., Schmidt, M., and Bach, F. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

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

Lei, Y., Shi, L., and Guo, Z.-C. Convergence of unregularized online learning algorithms. *Journal of Machine Learning Research*, 18(171):1–33, 2018.

Li, J., Luo, X., and Qiao, M. On generalization error bounds of noisy gradient methods for non-convex learning. *International Conference on Learning Representations*, 2020.

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

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

Liu, M., Zhang, X., Chen, Z., Wang, X., and Yang, T. Fast stochastic AUC maximization with O(1/n)-convergence rate. In *International Conference on Machine Learning*, pp. 3195–3203, 2018.

Liu, T., Lugosi, G., Neu, G., and Tao, D. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pp. 2159–2167, 2017.

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

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334, 2018.

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

Mücke, N., Neu, G., and Rosasco, L. Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pp. 12568–12577, 2019.

Natole, M., Ying, Y., and Lyu, S. Stochastic proximal algorithms for AUC maximization. In *International Conference on Machine Learning*, pp. 3707–3716, 2018.

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

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

Nesterov, Y. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152 (1-2):381–404, 2015.

Pillaud-Vivien, L., Rudi, A., and Bach, F. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pp. 8114–8124, 2018.

Rakhlin, A., Mukherjee, S., and Poggio, T. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pp. 449–456, 2012.

Reeve, H. W. J. and Kabán, A. Optimistic bounds for multi-output prediction. *CoRR*, abs/2002.09769, 2020.

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

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

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

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

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

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

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

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster

convergence of sgd for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204, 2019.

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

Ying, Y. and Zhou, D.-X. Online pairwise learning algorithms. *Neural computation*, 28(4):743–777, 2016.

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

Ying, Y., Wen, L., and Lyu, S. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2016.

Yuan, Z., Yan, Y., Jin, R., and Yang, T. Stagewise training accelerates convergence of testing error over sgd. In *Advances in Neural Information Processing Systems*, pp. 2604–2614, 2019.

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

Zhao, P., Hoi, S. C., Jin, R., and Yang, T. Online AUC maximization. In *International Conference on Machine Learning*, pp. 233–240. Omnipress, 2011.

Zhou, Y., Liang, Y., and Zhang, H. Generalization error bounds with probabilistic guarantee for SGD in non-convex optimization. *arXiv preprint arXiv:1802.06903*, 2018.