Proceedings of Machine Learning Research vol 195:1–37, 2023 36th Annual Conference on Learning Theory

Stability and Generalization of Stochastic Optimization with Nonconvex and Nonsmooth Problems

Yunwen Lei LEIYW@HKU.HK
Department of Mathematics, The University of Hong Kong

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Stochastic optimization has found wide applications in minimizing objective functions in machine learning, which motivates a lot of theoretical studies to understand its practical success. Most of existing studies focus on the convergence of optimization errors, while the generalization analysis of stochastic optimization is much lagging behind. This is especially the case for nonconvex and nonsmooth problems often encountered in practice. In this paper, we initialize a systematic stability and generalization analysis of stochastic optimization on nonconvex and nonsmooth problems. We introduce novel algorithmic stability measures and establish their quantitative connection on the gap between population gradients and empirical gradients, which is then further extended to study the gap between the Moreau envelope of the empirical risk and that of the population risk. To our knowledge, these quantitative connection between stability and generalization in terms of either gradients or Moreau envelopes have not been studied in the literature. We introduce a class of sampling-determined algorithms, for which we develop bounds for three stability measures. Finally, we apply these results to derive error bounds for stochastic gradient descent and its adaptive variant, where we show how to achieve an implicit regularization by tuning the step sizes and the number of iterations.

1. Introduction

Stochastic optimization has become the workhorse behind many successful applications of machine learning (ML) (Zhang, 2004; Bottou et al., 2018). The basic idea is to introduce randomness into the design of optimization algorithms to speed up the learning process by using the sum structure of objective functions in ML. A representative algorithm is the stochastic gradient descent (SGD). As an iterative algorithm, SGD first randomly selects a single example from a training dataset to build a stochastic gradient, and then moves along the negative direction of this stochastic gradient to get the next iterate. Due to its cheap computation cost and simplicity, SGD is especially interesting to solve large-scale and complex learning problems. In the last decade, SGD has been improved in various directions from the viewpoint of Nesterov acceleration (Nesterov, 1983), variance reduction (Johnson and Zhang, 2013; Schmidt et al., 2017; Defazio et al., 2014; Fang et al., 2018) and adaptive learning rates (Duchi et al., 2010; Kingma and Ba, 2015; Zhou et al., 2018).

Motivated by the increasing popularity, researchers have studied the theoretical behavior of stochastic optimization. Depending on the property of objective functions, one can measure the progress of optimization in terms of different performance metrics. For strongly convex problems, one can use the distance between the output model and the best model as the performance measure since there is only a unique minimizer (Bottou et al., 2018; Zhang and Zhou, 2019). For convex problems, one can develop convergence rates in terms of functional suboptimality gap since there may exist several models with the same global function value (Zhang, 2004). For nonconvex and smooth problems, one can measure the performance through the magnitude of gradients since an algorithm is only guaranteed to find a local minimum (Ghadimi and Lan, 2013).

The performance metric becomes more tricky for nonconvex and nonsmooth problems (Davis and Drusvyatskiy, 2019, 2021). For an objective function ψ, neither the functional suboptimality gap $\psi(\mathbf{w}_t) - \inf \psi(\mathbf{w})$, nor the stationarity measure, $\text{dist}(0, \partial\psi(\mathbf{w}_t))$, necessarily decay to zero along the optimization process (Davis and Drusvyatskiy, 2019). Here $\mathbf{w}_t$ denotes an iterate of the algorithm, $\partial\psi(\mathbf{w}_t)$ denotes the subdifferential and dist denotes the Euclidean distance function. Recently, Davis and Drusvyatskiy (2019) proposed to use the Moreau envelope $\psi_\lambda(\mathbf{w}) = \inf_{\mathbf{v}} \left\{\psi(\mathbf{v}) + \frac{1}{2\lambda}\|\mathbf{v} - \mathbf{w}\|_2^2\right\}$ as a useful potential function to study stochastic optimization for weakly convex problems [1]. An intuitive understanding is that a small gradient $\|\nabla\psi_\lambda(\mathbf{w}_t)\|_2$ implies that $\mathbf{w}_t$ is near some point that is nearly stationary for the problem $\min_{\mathbf{w}} \psi(\mathbf{w})$, which motivates the use of the performance measure $\|\nabla\psi_\lambda(\mathbf{w}_t)\|_2$ for weakly convex problems. Weakly convex problems form an importance class of nonconvex and nonsmooth problems, with instantiations in various application domains such as phase retrieval, robust principal component analysis, covariance matrix estimation and sparse dictionary learning (Davis and Drusvyatskiy, 2019).

Most of existing studies focus on the convergence behavior of stochastic optimization algorithms from the perspective of optimization, i.e., how the trained model would behave on training examples. However, in ML we are more interested in the prediction behavior from the perspective of learning (Mohri et al., 2012), i.e., how these models would behave on testing examples, which is much less studied for stochastic optimization. The gap between training and testing is a central topic in statistical learning theory (SLT). There are two major approaches to study the generalization gap: a *uniform convergence* approach based on the complexity analysis of hypothesis spaces (Bartlett and Mendelson, 2002) and an *algorithmic stability* approach based on the sensitivity analysis of algorithms (Bousquet and Elisseeff, 2002) (for simplicity we always mean algorithmic stability when mentioning stability). Uniform convergence analysis applies to nonconvex problems, which, however, often leads to a square-root dependency on the dimensionality and therefore unfavorable for high-dimensional learning problems (Feldman and Vondrak, 2019). Stability analysis can yield dimension-free bounds, which, however, often requires strong assumptions on loss functions such as convexity or smoothness. For example, most of the algorithmic stability analysis of stochastic optimization requires a convexity and a smoothness assumption (Hardt et al., 2016; Kuzborskij and Lampert, 2018). The smoothness assumption is removed in the recent study (Lei and Ying, 2020; Bassily et al., 2020). In particular, the paper (Bassily et al., 2020) develops matching lower bounds for convex and nonsmooth problems. For nonconvex problems, one typically requires a Polyak-Łojasiewicz (PL) condition to get nontrivial error bounds of SGD (Charles and Papailiopoulos, 2018). In the general nonconvex case, the stability analysis of SGD requires very small step sizes to get meaningful stability bounds (Hardt et al., 2016; Kuzborskij and Lampert, 2018), for which one cannot get meaningful optimization error bounds within reasonable computations. The strong assumption restricts the application domain of stability analysis for nonconvex and nonsmooth problems, which are often encountered in practice. To our knowledge, there is no stability analysis of stochastic optimization for problems that are simultaneously nonconvex and nonsmooth without restrictive assumptions such as the PL condition.

In this paper, we initialize the stability and generalization analysis of stochastic optimization for weakly convex problems, where the objective functions are nonconvex and nonsmooth. As a warm up, we first consider convex and nonsmooth problems, then nonconvex and smooth problems, and finally move onto weakly convex problems. As indicated before, we require to use different metrics

1. A function is weakly convex if eigenvalues of Hessian matrices are lower bounded by a negative value.

to measure the generalization performance, which also asks for different stability measures as well as a different connection between stability and generalization. Our contributions are as follows. Comparisons between our results and existing results are given in Table 1 and Table 2.

(a) We introduce a stability measure called uniform stability in gradients, and establish its quantitative relationship to the generalization measured by gradients for smooth problems. In particular, we show that the gap between population and empirical gradients can be bounded by our stability measure plus $O(1/\sqrt{n})$, where n is the sample size.

(b) We consider a specific class of nonconvex and nonsmooth problems called weakly convex problems, for which we measure the performance of trained models by Moreau envelopes. We develop, to our best knowledge, the first connection between argument stability and the generalization gap measured by Moreau envelopes.

(c) We introduce the concept of sampling-determined algorithms, for which we establish stability bounds in terms of either function values, gradients or arguments.

(d) We apply our results to SGD and its adaptive variant. For nonconvex and smooth problems, we develop stability-based risk bounds without the PL condition. We also develop the first risk bounds in terms of Moreau envelops for weakly convex problems.

2. Related Work

In this section, we review the related work on generalization analysis. We will focus on two approach: the algorithmic stability approach and the uniform convergence approach.

Algorithmic Stability. We first review the related work on algorithmic stability. Algorithmic stability is a fundamental concept in SLT to measure the sensitivity of an algorithm up to a perturbation of the training dataset, which is closely related to learnability (Shalev-Shwartz et al., 2010; Rakhlin et al., 2005). There are various algorithmic stability concepts. Some stability concepts measure the sensitivity in terms of function values, e.g., uniform stability (Bousquet and Elisseeff, 2002), hypothesis stability (Bousquet and Elisseeff, 2002; Elisseeff et al., 2005), Bayes stability (Li et al., 2020) and on-average stability (Shalev-Shwartz et al., 2010; Kuzborskij and Lampert, 2018), while others measure the sensitivity in terms of output models, e.g., argument stability (Liu et al., 2017) and on-average argument stability (Lei and Ying, 2020). A most widely used stability concept is the uniform stability (Bousquet and Elisseeff, 2002), which can imply almost optimal generalization bounds with high probability (Feldman and Vondrak, 2019; Bousquet et al., 2020; Klochkov and Zhivotovskiy, 2021). The celebrated connection between stability and generalization motivates the discussion of stability for many specific algorithms, including regularization algorithms (Bousquet and Elisseeff, 2002; Attia and Koren, 2022), stochastic optimization algorithms (Hardt et al., 2016; Chen et al., 2018; Kuzborskij and Lampert, 2018; Charles and Papailiopoulos, 2018; Mou et al., 2018), iterative hard thresholding (Yuan and Li, 2021), structured prediction (London et al., 2016), meta learning (Maurer, 2005) and transfer learning (Kuzborskij and Lampert, 2018). In particular, the influential work gives the first stability analysis of SGD applied to convex and smooth problems (Hardt et al., 2016). The smoothness assumption is removed in the recent study (Lei and Ying, 2020; Bassily et al., 2020), and a tight lower bound on the stability of SGD was developed (Bassily et al., 2020). Stability analysis can be also used to study the convergence of optimization error for multi-epoch SGD (Koren et al.).

Uniform Convergence. Machine learning models may achieve good performance on the training dataset but bad generalization behavior, which motivates the generalization analysis by the uniform convergence approach to study the difference between training and testing over the whole hypothesis space. Initially, the uniform convergence was mainly studied in terms of *function values* (Bartlett and Mendelson, 2002), which, however, is not appropriate to stochastic optimization with nonconvex loss functions. The underlying reason is that an algorithm can only guarantee to find a local minimizer (one cannot get convergence rate of training errors to the that of the best model). Then, the uniform convergence of function values (Lei et al., 2021) fail to show the convergence of testing errors to that of the best model. Instead, one has to turn to other performance measures such as the gradients of risks for smooth problems (Ghadimi and Lan, 2013) and the gradients of Moreau envelope for weakly convex problems (Davis and Drusvyatskiy, 2019). In particular, Ghadimi and Lan (2013) gave the first nonasymptotical convergence rate of the gradient norm. Motivated by this observation, the uniform convergence for gradients have been recently studied (Mei et al., 2018; Foster et al., 2018; Lei and Tang, 2021; Davis and Drusvyatskiy, 2021). The work (Mei et al., 2018) initialized the discussion on the uniform convergence of gradients by characterizing the complexity of function spaces with covering numbers, which was extended to the uniform convergence in terms of Rademacher complexities (Foster et al., 2018). These discussions are devoted to control the uniform deviation between gradients of empirical and population risks under a smoothness condition. For nonsmooth problems, the gradients are not well defined since the functions may not be differentiable. This problem was recently addressed by considering the gradients of Moreau envelope of empirical/population risks (Davis and Drusvyatskiy, 2021), which are appropriate stationary measures for weakly convex problems. Specifically, the uniform deviation of gradients for the Moreau envelope between empirical and population risks was studied based on covering numbers (Davis and Drusvyatskiy, 2021).

Other than the above two approaches, there are also interesting discussions on generalization analysis by using tools in integral operators (Smale and Zhou, 2007; Guo et al., 2017; Mücke et al., 2019; Pillaud-Vivien et al., 2018) and information theory (Russo and Zou, 2016; Xu and Raginsky, 2017; Neu et al., 2021; Neu and Lugosi, 2022).

3. Problem Setup

Let ρ be a probability measure defined on a sample space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, from which a dataset $S = \{z_1, \ldots, z_n\}$ are independently drawn. Based on S, we wish to build a model $h : \mathcal{X} \mapsto \mathcal{Y}$ for prediction. We consider a parametric learning setting where the model is determined by a parameter $\mathbf{w}$ in a parameter space $\mathcal{W} \subset \mathbb{R}^d$. The performance of a model $\mathbf{w}$ on an example z can be quantified by a loss function $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$. The training and testing behavior of $\mathbf{w}$ then can be measured by the empirical risk $F_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n f(\mathbf{w}; z_i)$ and the population risk $F(\mathbf{w}) := \mathbb{E}_Z[f(\mathbf{w}; Z)]$, where $\mathbb{E}_Z$ denotes the expectation w.r.t. Z. Let $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ be the model with the minimal population risk in $\mathcal{W}$. Let A be a randomized learning algorithm and $A(S)$ be the output model when applying A to the dataset S. In this paper, we are interested in the quality of $A(S)$ in prediction under different performance measures. We require necessary definitions on Lipschitz continuity, smoothness and convexity. Let $\|\cdot\|_2$ denote the Euclidean norm and $\nabla g(\mathbf{w})$ denote a subgradient of g at $\mathbf{w}$. If g is differentiable then $\nabla g(\mathbf{w})$ becomes the gradient of g at $\mathbf{w}$.

Definition 1 *Let $g : \mathcal{W} \mapsto \mathbb{R}$. Let $L, \rho, G > 0$.*

(a) We say g is L-smooth if $\left\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\right\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2, \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$.

(b) We say g is convex if $g(\mathbf{w}) \geq g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}') \rangle, \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$. We say g is ρ-weakly-convex if $\mathbf{w} \mapsto g(\mathbf{w}) + \frac{\rho}{2}\|\mathbf{w}\|_2^2$ is convex, and ρ-strongly convex if $\mathbf{w} \mapsto g(\mathbf{w}) - \frac{\rho}{2}\|\mathbf{w}\|_2^2$ is convex.

(c) We say g is G-Lipschitz if $|g(\mathbf{w}) - g(\mathbf{w}')| \leq G\|\mathbf{w} - \mathbf{w}'\|_2, \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$.

Weakly convex functions are widespread in applications with a common source being the composite function class: $g(\mathbf{w}) := h(c(\mathbf{w}))$, where $h : \mathbb{R}^m \mapsto \mathbb{R}$ is convex and G-Lipschitz and $c : \mathbb{R}^d \mapsto \mathbb{R}^m$ has β-Lipschitz continuous Jacobians (Davis and Drusvyatskiy, 2019). Concrete examples include robust phase retrieval, covariance matrix estimation, sparse dictionary learning, robust PCA and conditional value-at-risk. We will use error decomposition to study the generalization behavior of learning models. Depending on the property of learning tasks, we will introduce different error decompositions.

For **convex learning** problems, a learning algorithm can be guaranteed to produce a model with a small empirical error. Therefore, we quantify the behavior of a model by the associated population risk. A standard approach to studying the population risk is to decompose it into two error terms (Bousquet and Bottou, 2008)

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

where we have used $\mathbb{E}_{S,A}[F_S(\mathbf{w}^*)] = F(\mathbf{w}^*)$ since $\mathbf{w}^*$ is independent of A and S. We refer to the term $F(A(S)) - F_S(A(S))$ in (3.1) as the generalization error since it is related to the generalization from the training behavior to testing behavior. The second term $F_S(A(S)) - F_S(\mathbf{w}^*)$ is called the optimization error since it quantifies how well the algorithm minimizes the empirical risk. We will apply stability analysis to study the generalization error, and tools in optimization theory to study the optimization error.

For **nonconvex and smooth** learning problems, a learning algorithm can only be guaranteed to produce an approximate stationary point, i.e., a point $\mathbf{w}$ with a small $\|\nabla F_S(\mathbf{w})\|_2$. In this case, the population risk is not a reasonable quality measure since there may be many local minimizers with different risks. As an alternative, we use the population gradient norm as the performance measure. We use the following error decomposition

$$\mathbb{E}_{S,A}\big[\|\nabla F(\mathbf{w})\|_2\big] \leq \mathbb{E}_{S,A}\big[\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|_2\big] + \mathbb{E}_{S,A}\big[\|\nabla F_S(\mathbf{w})\|_2\big]. \quad (3.2)$$

We call the first term $\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|_2$ the generalization error for smooth problems, and $\|\nabla F_S(\mathbf{w})\|_2$ the optimization error (empirical gradient norm). We will introduce a stability concept as well as its connection to generalization to study the generalization error for nonconvex problems. The optimization error is well studied in the literature (Ghadimi and Lan, 2013).

For **weakly convex** learning problems, we cannot measure the quality of a model by gradients since the function may not be differentiable. An elegant performance measure is in terms of the Moreau envelope. Intuitively, Moreau envelope of f is a smoothed approximation of f. An illustration of the Moreau envelope was given in Fig. 1 of Davis and Drusvyatskiy (2019).

Definition 2 (Moreau envelope) *For any $\lambda > 0$ and $\psi : \mathcal{W} \mapsto \mathbb{R}$, we define the Moreau envelope (with parameter λ) $\psi_\lambda : \mathcal{W} \mapsto \mathbb{R}$ by*

$$\psi_\lambda(\mathbf{w}) = \min_{\mathbf{v} \in \mathbb{R}^d} \big\{\psi(\mathbf{v}) + 1/(2\lambda)\|\mathbf{w} - \mathbf{v}\|_2^2\big\}$$

and the proximal operator $Prox_{\lambda\psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ by

$$Prox_{\lambda\psi}(\mathbf{w}) = \arg\min_{\mathbf{v} \in \mathbb{R}^d} \big\{\psi(\mathbf{v}) + 1/(2\lambda)\|\mathbf{w} - \mathbf{v}\|_2^2\big\}.$$

Standard results show that as long as ψ is ρ-weakly-convex and $\lambda < 1/\rho$, the envelope ψ_λ is strongly smooth with the gradient given by $\nabla\psi_\lambda(\mathbf{w}) = \lambda^{-1}\big(\mathbf{w} - \text{Prox}_{\lambda\psi}(\mathbf{w})\big)$, where $\nabla\psi_\lambda(\mathbf{w})$ denotes $\nabla(\psi_\lambda)(\mathbf{w})$. For smooth ψ, the norm of $\nabla\psi_\lambda(\mathbf{w})$ is proportional to the magnitude of the true gradient $\nabla\psi$. For nonsmooth ψ, it was shown that $\|\nabla\psi_\lambda(\mathbf{w})\|_2$ has an intuitive interpretation in terms of near-stationarity of the target problem $\min_{\mathbf{w}} \psi(\mathbf{w})$ (Davis and Drusvyatskiy, 2019). Therefore, we use $\|\nabla F_{1/(2\rho)}(\mathbf{w})\|_2$ to quantify the generalization behavior of $\mathbf{w}$ for ρ-weakly-convex F ($F_{1/(2\rho)}$ means the Moreau envelope of F with the parameter $1/(2\rho)$). We need the following error decomposition in this case

$$\mathbb{E}_{S,A}\big[\|\nabla F_{1/(2\rho)}(\mathbf{w})\|_2\big] \leq \mathbb{E}_{S,A}\big[\|\nabla F_{1/(2\rho)}(\mathbf{w}) - \nabla F_{S,1/(2\rho)}(\mathbf{w})\|_2\big] + \mathbb{E}_{S,A}\big[\|\nabla F_{S,1/(2\rho)}(\mathbf{w})\|_2\big], \tag{3.3}$$

where we denote $F_{S,1/(2\rho)} := (F_S)_{1/(2\rho)}$. We call the first term $\|\nabla F_{1/(2\rho)}(\mathbf{w}) - \nabla F_{S,1/(2\rho)}(\mathbf{w})\|_2$ the generalization error for weakly-convex (possibly nonsmooth) problems, and $\|\nabla F_{S,1/(2\rho)}(\mathbf{w})\|_2$ the optimization error. We will introduce a novel connection between argument stability and generalization to study the generalization error for weakly-convex problems. The optimization error on $\|\nabla F_{S,1/(2\rho)}(\mathbf{w})\|_2$ is well studied in the literature (Davis and Drusvyatskiy, 2019).

We summarize our results and give comparisons with existing results in Table 1 and Table 2. We consider two classes of problems: smooth & nonconvex problems, and weakly convex & nonsmooth problems. Table 1 considers the generalization gap, while Table 2 considers the error bounds for SGD.

Problems	Reference	Bounds
smooth & nonconvex	Mei et al (2018)	$O(\sqrt{d(\log n)/n})$
	Thm. 6 (our work)	$O(\epsilon + n^{-\frac{1}{2}})$
weakly convex & nonsmooth	Davis and Drusvyatskiy (2021)	$O(\sqrt{d/n})$
	Thm. 8 (our work)	$O(\sqrt{\epsilon} + n^{-\frac{1}{2}})$

Table 1: Generalization bounds. For smooth and nonconvex problems, the generalization bounds are derived for $\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2$. For weakly convex and nonsmooth problems, the generalization bounds are derived for $\big\|\nabla F_{S,1/(2\rho)}(A(S)) - \nabla F_{1/(2\rho)}(A(S))\big\|_2$. The existing generalization bounds are based on uniform convergence approach, and admit a square-root dependency on the dimension. Our generalization bounds depend on the stability parameter ϵ.

4. Stability and Generalization

4.1. Connecting Stability and Generalization

Algorithmic stability measures the insensitiveness on an algorithm under a perturbation of a training dataset by a single example. The uniform stability and uniform argument stability were discussed in the literature (Bousquet and Elisseeff, 2002). To tackle the performance measure in terms of gradient norms for nonconvex learning problems, we introduce a *uniform stability in gradients*. We say S, S' are neighboring datasets if they differ by at most a single example.

Problems	Reference	Bounds
smooth & nonconvex	Ghadimi and Lan (2013)	$\|\nabla F_S(\mathbf{w}_r))\| = O(T^{-\frac{1}{4}})$
	Prop. 21 (our work)	$\|\nabla F(\mathbf{w}_r)\|_2 = O(n^{-\frac{1}{6}})$
weakly convex & nonsmooth	Davis and Drusvyatskiy (2019)	$\|\nabla F_{S,1/(2\rho)}(\mathbf{w}_r)\|_2 = O(T^{-\frac{1}{4}})$
	Prop. 25 (our work)	$\|\nabla F_{1/(2\rho)}(\mathbf{w}_r)\|_2 = O(n^{-\frac{1}{6}})$

Table 2: Error bounds for SGD. The existing analysis considers the performance of SGD on the empirical risk F_S, while our results consider the performance of SGD on the population risk F. Here T is the number of iterations and $\mathbf{w}_r$ is a randomly selected SGD iterate.

Definition 3 (Uniform Stability) *Let A be a randomized algorithm. We say A is ϵ-uniformly-stable in* function values *if for all neighboring datasets S, S', we have*

$$\sup_z \mathbb{E}_A\big[f(A(S); z) - f(A(S'); z)\big] \leq \epsilon. \tag{4.1}$$

We say A is ϵ-uniformly-argument-stable if for all neighboring datasets S, S', we have

$$\mathbb{E}_A\big[\|A(S) - A(S')\|_2\big] \leq \epsilon. \tag{4.2}$$

We say A is ϵ-uniformly-stable in gradients *if for all neighboring datasets S, S', we have*

$$\sup_z \mathbb{E}_A\big[\|\nabla f(A(S); z) - \nabla f(A(S'); z)\|_2^2\big] \leq \epsilon^2. \tag{4.3}$$

Remark 4 The motivation of introducing the gradient-based stability is to use it to study the generalization performance for nonconvex problems. For nonconvex problems, an optimization algorithm generally only finds a local minimizer, and therefore one cannot use the function value to measure the convergence (the local minimizer the algorithm finds may be far away from the global minimizer and therefore the convergence in function values do not make much sense). In this case, one often studies the convergence of ∇F_S in the optimization community (Ghadimi and Lan, 2013). To use this convergence to study the behavior of $A(S)$ in prediction, we need to address $\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2$, which, as we will see, can be achieved by stability in gradients. In summary, the stability on gradients allows us to incorporate the existing optimization error bounds to study the prediction performance as measured by $\|\nabla F(A(S))\|_2$.

The connection between uniform stability in function values and generalization is given in the following lemma (Shalev-Shwartz et al., 2010; Hardt et al., 2016).

Lemma 5 (Generalization via Stability in Function Values) *Let A be ϵ-uniformly stable in function values. Then $\big|\mathbb{E}_{S,A}\big[F_S(A(S)) - F(A(S))\big]\big| \leq \epsilon$.*

Our first result is a connection between generalization and stability in gradients. This result cannot be derived by using the standard arguments in the literature (Shalev-Shwartz et al., 2010; Hardt et al., 2016) since one can not exchange the summation operator and norm. We will give more explanations in the proof, which is given in Section A.1.

Theorem 6 (Generalization via Stability in Gradients) *Let A be ϵ-uniformly-stable in gradients. Assume for any z, the function $f(\mathbf{w}; z)$ is differentiable. Then*

$$\mathbb{E}_{S,A}\big[\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2\big] \leq 4\epsilon + \sqrt{n^{-1}\mathbb{E}_S\Big[\mathbb{V}_Z(\nabla f(A(S); Z))\Big]}, \tag{4.4}$$

where $\mathbb{V}_Z(\nabla f(A(S); Z)) = \mathbb{E}_Z\big[\|\nabla f(A(S); Z) - \mathbb{E}_Z[\nabla f(A(S); Z)]\|_2^2\big]$ is the variance of $\nabla f(A(S); Z)$ as a function of the random variable Z.

Remark 7 Note the left-hand side of Eq. (4.4) can be addressed by the uniform convergence of gradients $\sup_{\mathbf{w}\in\mathcal{W}} \|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|_2$, which was established in terms of covering numbers (Mei et al., 2018) and Rademacher complexities (Foster et al., 2018). These bounds generally involve a square-root dependency on the dimension of $\mathcal{W}$. As a comparison, Theorem 6 considers the convergence of empirical gradients to population gradients at the output model $A(S)$. Therefore, it implies dimension-free bounds which would be effective for high-dimensional learning problems.

Our second result is a connection between the uniform argument-stability and generalization measured by the Moreau envelope for weakly convex problems. Theorem 8 shows that the difference between empirical and population gradients of the Moreau envelope at $A(S)$ can be bounded by the uniform argument stability of A. With this result, we can transfer the existing bound on $\|\nabla F_{S,1/(2\rho)}\|_2$ to $\|\nabla F_{1/(2\rho)}\|_2$ on the performance of models for prediction. The proof of Theorem 8 is totally different from that of Theorem 6. The proof is given in Section A.2.

Theorem 8 (Generalization via Uniform Argument Stability) *Let A be ϵ-argument stable. Assume for any z, the function $f(\mathbf{w}; z)$ is G-Lipschitz continuous. Assume for any S, the function F_S is ρ-weakly-convex and F is ρ-weakly-convex. Then*

$$\mathbb{E}\big[\big\|\nabla F_{S,1/(2\rho)}(A(S)) - \nabla F_{1/(2\rho)}(A(S))\big\|_2\big] \leq \frac{4G}{\sqrt{n}} + \sqrt{32G\epsilon\rho}. \tag{4.5}$$

Remark 9 For ρ-weakly convex f, the uniform convergence $\sup_{\mathbf{w}\in\mathcal{W}} \|\nabla F_{S,1/(2\rho)}(\mathbf{w}) - \nabla F_{1/(2\rho)}(\mathbf{w})\|_2$ was studied in terms of the covering number of $\mathcal{W}$ (Davis and Drusvyatskiy, 2021), which generally involves a square-root dependency on the dimensionality. For example, if $\mathcal{W}$ is a ball in $\mathbb{R}^d$, then the following result was established (Davis and Drusvyatskiy, 2021)

$$\sup_{\mathbf{w}\in\mathcal{W}} \|\nabla F_{S,1/(2\rho)}(\mathbf{w}) - \nabla F_{1/(2\rho)}(\mathbf{w})\|_2 = O\big(G\sqrt{d/n}\big). \tag{4.6}$$

The underlying reason to consider a uniform convergence is noting the dependency of $A(S)$ in Eq. (4.5) on S. We address this dependency by giving a bound in terms of the argument stability of A. Theorem 8 yields dimension-free bounds since it only considers the convergence of $\nabla F_{S,1/(2\rho)}$ to $\nabla F_{1/(2\rho)}$ at the particular output model $A(S)$.

Finally, we give a high-probability bound on the generalization gap measured by the Moreau envelope. The proof is given in Section A.2.

Theorem 10 (High-probability Bound via Uniform Argument Stability) *Let A be ϵ-argument stable almost surely, i.e., $\|A(S) - A(S')\|_2 \leq \epsilon$ for any neighboring S, S'. Assume for any z, the function $f(\mathbf{w}; z)$ is G-Lipschitz continuous and $f(A(S); z) = O(1)$ almost surely. Assume for any*

S, the function F_S is ρ-weakly-convex and F is ρ-weakly-convex. For any $\delta \in (0,1)$, the following inequality holds with probability at least $1-\delta$

$$\left\|\nabla F_{S,1/(2\rho)}(A(S)) - \nabla F_{1/(2\rho)}(A(S))\right\|_2 = O\Big(\big(Gn^{-\frac{1}{2}} + \sqrt{G\epsilon\rho}\big)\sqrt{\log(n)\log(1/\delta)} + \big(n^{-1}\rho^2\log(1/\delta)\big)^{\frac{1}{4}}\Big).$$

4.2. Stability Bounds

We now consider a class of randomized algorithms called sampling-determined algorithms for our stability analysis. We say a randomized algorithm A is symmetric if its output is independent on the order of the elements in the training set.

Definition 11 (Sampling-determined Algorithm) *Let A be a randomized algorithm which randomly chooses an index sequence $I(A) = \{i_t\}$ to build stochastic gradients. We say a symmetric algorithm A is sampling-determined if the output model is determined by $\{z_j : j \in I(A)\}$. To be precise, $A(S)$ is independent of z_j if $j \notin I$.*

An important property of sampling-determined algorithms is that these algorithms will produce the same model when applied to two neighboring datasets if the differing example is not selected in the algorithm. For example, if two neighboring datasets differ by the first example and the index 1 is not selected by the algorithm, then the algorithm would produce the same model when applied to these two neighboring datasets. This property is critical for us to study the stability. The class of sampling-determined algorithms include several famous randomized algorithms. Below, we give some representative algorithms. The first algorithm is the SGD, which is a most simple and most popular stochastic optimization algorithm. Let $\Pi_{\mathcal{W}}(\mathbf{w})$ denote the projection of $\mathbf{w}$ onto $\mathcal{W}$. Note $\mathcal{W}$ can be $\mathbb{R}^d$ and in this case there is no projection.

Definition 12 (Stochastic Gradient Descent) *Let $\mathbf{w}_1 = 0 \in \mathbb{R}^d$ be an initial point and $\{\eta_t\}_t$ be a sequence of positive step sizes. SGD updates models by $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\big(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{i_t})\big)$, where $\nabla f(\mathbf{w}_t, z_{i_t})$ denotes a subgradient of f w.r.t. the first argument and i_t is independently drawn from the uniform distribution over $[n] := \{1, 2, \ldots, n\}$.*

The second algorithm is an adaptive variant of SGD, which introduces a sequence $\{b_t^2\}$ to store the accumulated gradient norm square (Duchi et al., 2010; Li and Orabona, 2019; Ward et al., 2020). We then set the step size as the reciprocal of b_t multiplied by a parameter η (Ward et al., 2020). This algorithm has a nice advantage of being able to adapt the level of stochastic noise of the problem, and can achieve robust convergence without the need to fine-tune stepsize schedule.

Definition 13 (AdaGrad-Norm) *Let $\mathbf{w}_1 = 0 \in \mathbb{R}^d$, $b_0 > 0$ and $\eta > 0$. At each iteration, we first draw i_t from the uniform distribution over $[n]$ and update $\{b_t\}, \{\mathbf{w}_t\}$ by*

$$b_t^2 = b_{t-1}^2 + \|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2, \qquad \mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\Big(\mathbf{w}_t - \frac{\eta}{b_t}\nabla f(\mathbf{w}_t; z_{i_t})\Big). \tag{4.7}$$

Remark 14 Let A be either SGD or AdaGrad-Norm with T iterations. Note $A(S)$ does not depend on z_j if $j \in [n]$ is not selected in the implementation of A. Therefore, both SGD and AdaGrad-Norm are sampling-determined algorithms and $I(A) = \{i_1, \ldots, i_T\}$. It is also clear from the definition that Adam is a sampling-determined algorithm.

Remark 15 There are also some randomized algorithms that are not sampling-determined. A notable example is the stochastic variance reduction gradient (SVRG) (Johnson and Zhang, 2013). Note that SVRG is implemented in epochs, for each of which we need to compute the full gradient at a reference point. Therefore, SVRG will produce different models when applied to neighboring datasets even if the differing example is not selected to compute a stochastic gradient. One can also check that other variance reduction algorithms are not sampling-determined, including stochastic average gradient (Schmidt et al., 2017) and SAGA (Defazio et al., 2014).

The following theorem to be proved in Section C (supplementary material) establishes the uniform stability bounds for sampling-determined algorithms. It shows that the uniform stability of a sampling-determined algorithm A can be bounded by the probability of an index not selected in $I(A)$. We consider stability in function values (Part (a)), stability in gradients (Part (b)) and stability in arguments (Part (c)). The proof is motivated by the arguments in Hardt et al. (2016).

Theorem 16 *Let A be a sampling-determined algorithm and S, S' be neighboring datasets.*

(a) If $\sup_z \mathbb{E}_A[f(A(S); z)|n \in I(A)] \leq B$ for any S, then

$$\sup_z \mathbb{E}_A\big[f(A(S); z) - f(A(S'); z)\big] \leq 2B \cdot \Pr\{n \in I(A)\}.$$

(b) If $\sup_z \mathbb{E}_A[\|\nabla f(A(S); z)\|_2^2|n \in I(A)] \leq G^2$ for any S, then

$$\sup_z \mathbb{E}_A\big[\|\nabla f(A(S); z) - \nabla f(A(S'); z)\|_2^2\big] \leq 4G^2 \cdot \Pr\{n \in I(A)\}.$$

(c) If $\mathbb{E}_A[\|A(S)\|_2|n \in I(A)] \leq R$ for any S, then $\mathbb{E}_A\big[\|A(S) - A(S')\|_2\big] \leq 2R \cdot \Pr\{n \in I(A)\}$.

We derive Corollary 17 by computing $\Pr\{n \in I(A)\}$. The proof is given in Section C.

Corollary 17 *Let A be SGD or AdaGrad-Norm with T iterations and S, S' be neighboring datasets.*

(a) If $\sup_z \mathbb{E}_A[f(A(S); z)|n \in I(A)] \leq B, \forall S$, then $\sup_z \mathbb{E}_A\big[f(A(S); z) - f(A(S'); z)\big] \leq \frac{2BT}{n}$.

(b) If $\sup_z \mathbb{E}_A[\|\nabla f(A(S); z)\|_2^2|n \in I(A)] \leq G^2$ for any S, then $\sup_z \mathbb{E}_A\big[\|\nabla f(A(S); z) - \nabla f(A(S'); z)\|_2^2\big] \leq \frac{4G^2T}{n}$.

(c) If $\mathbb{E}_A[\|A(S)\|_2|n \in I(A)] \leq R$ for any S, then $\mathbb{E}_A\big[\|A(S) - A(S')\|_2\big] \leq \frac{2RT}{n}$.

Remark 18 Since we consider symmetric algorithms, the condition $n \in I(A)$ can be replaced by $i \in I(A)$ for any $i \in [n]$. Both Theorem 16 and Corollary 17 require boundedness assumptions on either function values, gradients and arguments, which hold immediately if we impose a projection operator on $A(S)$. Note we do not require a projection for each iterate. A projection for the final output $A(S)$ suffices for our analysis.

5. Applications to Stochastic Gradient Descent

We now apply our stability results to SGD. We denote $B \asymp \widetilde{B}$ if there exist constants $c_1, c_2 > 0$ such that $c_1\widetilde{B} < B \leq c_2\widetilde{B}$. Recall n is the sample size and T is the iteration number. We will consider different problem settings: convex and smooth cases, nonconvex and smooth cases, and weakly convex cases. All the proofs in this subsection can be found in Section D. We will give applications to adaptive gradient descent in Section E, and differentially private SGD in Section F.

Convex and Nonsmooth Problems. In Proposition 19, we show SGD applied to convex and nonsmooth problems can imply the excess population risk bounds $O(n^{-\frac{1}{3}})$ with $O(n^{\frac{2}{3}})$ iterations. The algorithm is computationally efficient in the sense that SGD with $T \asymp n^{\frac{2}{3}}$ iterations can at most imply optimization error bounds $O(1/\sqrt{T}) = O(n^{-\frac{1}{3}})$. Therefore, our analysis implies excess risk bounds of the same order of optimization error bounds with the same computation complexity. There is no additional cost by going from optimization to generalization if we run $O(n^{\frac{2}{3}})$ iterations. This proposition is not a main result since our focus is on nonconvex case. We present it just as a byproduct. Recall $\mathbf{w}^*$ is a minimizer of the population risk F and we assume $\|\mathbf{w}^*\|_2$ is finite.

Proposition 19 (Convex and Nonsmooth Case) *Let $\{\mathbf{w}_t\}_t$ be the sequence produced by SGD and $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2$ for all $t \in [T]$. Let A output $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$. If F_S is convex, $\eta_t = \eta$ and $\sup_z \mathbb{E}_A[f(A(S); z)|n \in I(A)] \leq B$, then*

$$\mathbb{E}_{S,A}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O\Big(\frac{T\eta^2G^2 + \|\mathbf{w}^*\|_2^2}{T\eta}\Big) + O\big(BT/n\big). \tag{5.1}$$

If $\eta \asymp n^{-\frac{1}{3}}\|\mathbf{w}^*\|_2/G, T \asymp n^{\frac{2}{3}}G\|\mathbf{w}^*\|_2/B$ we have $\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O\big(\big(G\|\mathbf{w}^*\|_2 + B\big)n^{-\frac{1}{3}}\big)$.

Remark 20 We compare Proposition 19 with existing results. The following excess risk bounds of SGD without smoothness assumptions were established (Lei and Ying, 2020; Bassily et al., 2020)

$$\mathbb{E}_{S,A}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O\big(G^2\sqrt{T}\eta + T\eta G^2/n + \|\mathbf{w}^*\|_2^2/(T\eta)\big). \tag{5.2}$$

By setting $T \asymp n^2$ and $\eta \asymp T^{-\frac{3}{4}}\|\mathbf{w}^*\|_2/G$, the above bound implies the excess risk bounds $\mathbb{E}_{S,A}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(G\|\mathbf{w}^*\|_2 n^{-\frac{1}{2}})$. As a comparison, our analysis implies the bounds $O((G\|\mathbf{w}^*\|_2 + B)n^{-\frac{1}{3}})$. However, the bound (5.2) requires $O(n^2)$ iterations to achieve this optimal risk bounds, which is computationally expensive. As a comparison, our analysis requires $O(n^{\frac{2}{3}}G\|\mathbf{w}^*\|_2/B)$ iterations to achieve the bound $O((G\|\mathbf{w}^*\|_2 + B)n^{-\frac{1}{3}})$. To achieve the bound $O(G\|\mathbf{w}^*\|_2 n^{-\frac{1}{3}})$, the existing analysis (Lei and Ying, 2020; Bassily et al., 2020) requires to run SGD with $O(n^{\frac{4}{3}})$ iterations. Indeed, the right-hand-side of (5.2) is at least of the order of $O\big(G^2\sqrt{T}\eta + \|\mathbf{w}^*\|_2^2/(T\eta)\big) \geq O(G\|\mathbf{w}^*\|_2 T^{-\frac{1}{4}})$. Setting $T^{-\frac{1}{4}} = n^{-\frac{1}{3}}$ gives the complexity requirement $T = n^{\frac{4}{3}}$, which is larger than the iteration complexity $n^{\frac{2}{3}}G\|\mathbf{w}^*\|_2/B$ in Proposition 19. Note we require an assumption $\sup_z \mathbb{E}_A[f(A(S); z)|n \in I(A)] \leq B$ in Proposition 19, which is not required in Lei and Ying (2020); Bassily et al. (2020). The discussion in Bassily et al. (2020) requires a Lipschitz assumption and imply high-probability bounds, while we require the assumption $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2$ and derive bounds in expectation. Furthermore, a tight lower bound on the stability is developed in Bassily et al. (2020).

Excess risk bounds of the order $O(G\|\mathbf{w}^*\|_2 n^{-\frac{1}{3}}\log n)$ were also established for SGD based on the uniform convergence approach (Lin et al., 2016). Their discussions consider kernel methods and

would imply dimension-dependent bounds if applied to general nonlinear models. As a comparison, our stability analysis always yields dimension-free bounds.

Nonconvex and Smooth Problems. We now consider the performance of SGD for nonconvex and smooth problems. In the remainder, we always let r be randomly selected from the uniform distribution over $[T]$. We show SGD with $O(n^{\frac{2}{3}})$ iterations achieves the population gradient bound $\mathbb{E}_{S,A,r}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O(n^{-\frac{1}{6}})$. Again, this result shows considering generalization does not bring additional computation cost since SGD with $T \asymp n^{\frac{2}{3}}$ iterations is only guaranteed to achieve empirical gradient bounds $\mathbb{E}_{S,A,r}\big[\|\nabla F_S(\mathbf{w}_r)\|_2\big] = O(n^{-\frac{1}{6}})$ (Ghadimi and Lan, 2013). That is, with $n^{\frac{2}{3}}$ iterations, our population gradient bounds match the existing empirical gradient bounds.

Proposition 21 (Nonconvex and Smooth Case) *Let $\{\mathbf{w}_t\}_t$ be produced by SGD with $\eta_t = \eta$ and $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \le G^2$ for all $t \in [T]$. If $A(S) = \mathbf{w}_r$, F_S is L-smooth and*

$$\sup_z \mathbb{E}_A[\|\nabla f(A(S); z)\|_2^2 | n \in I(A)] \le G^2,$$

then

$$\mathbb{E}_{S,A,r}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O\Big(\frac{G\sqrt{T}}{\sqrt{n}} + \frac{G\sqrt{T}\eta + 1}{\sqrt{T\eta}}\Big).$$

If $T \asymp n^{\frac{2}{3}}/G^{\frac{2}{3}}, \eta \asymp 1/(G\sqrt{T})$, we get $\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O(G^{\frac{2}{3}} n^{-\frac{1}{6}})$.

Remark 22 We compare our bounds with existing results. For nonconvex, smooth and Lipschitz loss functions, the uniform stability bound of order $O(n^{-1}T^{\frac{Lc}{Lc+1}})$ was established for SGD with $\eta_t \le c/t$ (Hardt et al., 2016). While this analysis gives nontrivial bounds on the generalization gap, the proposed step size is small to enjoy a good decay of optimization errors. Indeed, with this step size one can only derive optimization error bounds $\mathbb{E}_{S,A,r}\big[\|\nabla F_S(\mathbf{w}_r)\|_2\big] = O(1/\log T)$. One cannot trade-off the generalization bounds $O(n^{-1}T^{\frac{Lc}{Lc+1}})$ and optimization error bounds $O(1/\log T)$ for a non-vacuous population gradient bound. Indeed, to get a non-vacuous bound, one requires $T = O(n^{\frac{Lc+1}{Lc}})$. However, in this case the optimization error bounds become $O(1/\log n)$, which are very slow. As a comparison, our discussion suggests a step size $\eta_t \asymp n^{-\frac{1}{3}}$ for a significantly better population risk bound $O(n^{-\frac{1}{6}})$. We should mention that the discussion in Hardt et al. (2016) considers the stability in function values, while we consider stability in gradients. High probability bounds on a weighted average of $\|\nabla F(\mathbf{w}_t)\|_2^2$ were developed in Lei and Tang (2021). Their discussions use a uniform convergence approach and therefore admits a square-root dependency on the dimensionality. As a comparison, Proposition 21 yields dimension-free bounds.

We can improve population gradient bounds under a strong growth condition (SGC), which connects the rates at which the stochastic gradients shrink to the full gradient (Vaswani et al., 2019).

Definition 23 *We say SGC holds if $\frac{1}{n}\sum_{i=1}^n[\|\nabla f(\mathbf{w}; z_i)\|_2^2] \le \rho\|\nabla F_S(\mathbf{w})\|_2^2$.*

Proposition 24 shows that the learning performance improves under the SGC condition.

Proposition 24 (Nonconvex, Smooth and SGC Case) *Assume for all z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is L-smooth and SGC holds. Let $\{\mathbf{w}_t\}_t$ be produced by SGD with $\eta_t = 1/(\rho L)$ and suppose $\mathbb{E}_{S,A}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2$ for all $t \in [T]$. If $A(S) = \mathbf{w}_r$, $T \asymp \sqrt{L\rho n}/G$ and*

$$\sup_z \mathbb{E}_A[\|\nabla f(A(S); z)\|_2^2 | n \in I(A)] \leq G^2,$$

then $\mathbb{E}_{S,A}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O((L\rho G^2/n)^{\frac{1}{4}})$.

Weakly Convex Problems. Finally, we consider weakly convex problems. Note we impose a bounded subgradient assumption $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2$ as in Davis and Drusvyatskiy (2019). In the appendix G, we will relax this assumption as $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq B_1 \mathbb{E}[f(\mathbf{w}_t; z_{i_t})] + B_2$ for some $B_1, B_2 > 0$ and derive the corresponding convergence rates of SGD. To our knowledge, this convergence analysis under the relaxed condition is new for SGD with weakly convex problems.

Proposition 25 (Weakly-convex Case) *Let $\{\mathbf{w}_t\}_t$ be given by SGD with $\eta_t = \eta$ and $A(S) = \mathbf{w}_r$. Assume $\mathbb{E}_{S,A}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2, \mathbb{E}_A[\|A(S)\|_2 | n \in I(A)] \leq R$. If F_S is ρ-weakly convex, then*

$$\mathbb{E}_{S,A,r}\big[\|\nabla F_{1/(2\rho)}(\mathbf{w}_r)\|_2\big] = O\Big(G\sqrt{\rho\eta} + \sqrt{GR\rho T/n} + 1/\sqrt{T\eta}\Big).$$

If $T \asymp n^{\frac{2}{3}}/(R^{\frac{2}{3}}\rho^{\frac{1}{3}})$ and $\eta \asymp 1/(G\sqrt{\rho T})$, we get $\mathbb{E}\big[\|\nabla F_{1/(2\rho)}(\mathbf{w}_r)\|_2\big] = O(\sqrt{G}\rho^{\frac{1}{3}}R^{\frac{1}{6}}/n^{\frac{1}{6}})$.

Remark 26 For weakly convex problems, the convergence rate

$$\mathbb{E}\big[\|\nabla F_{S,1/(2\rho)}(\mathbf{w}_r)\|_2\big] = O(G^{\frac{1}{2}}\rho^{\frac{1}{4}}T^{-\frac{1}{4}})$$

was established for SGD with T iterations (Davis and Drusvyatskiy, 2019). This result is impressive since neither the Moreau envelope nor the proximal map of F_S explicitly appear in the implementation of SGD. This result shows the behavior of SGD on training examples, which we extend to the generalization behavior of SGD on testing examples. Note our analysis requires to set $T \asymp n^{\frac{2}{3}}$ and therefore can only imply the bound of the order $O((G\rho)^{\frac{1}{3}}n^{-\frac{1}{6}})$. It would be interesting to further improve the risk bound here.

Population risk bounds of gradient descent were recently studied for weakly convex problems (Richards and Rabbat, 2021; Richards and Kuzborskij, 2021). Their discussions require the weak convexity parameter to be sufficiently small for meaningful generalization. As a comparison, our discussion does not require this assumption. Furthermore, their discussions consider smooth problems with Lipshictz continuous Hessians and focus on gradient descent (Richards and Rabbat, 2021), while our discussions apply to SGD with nonsmooth problems.

6. Conclusions

We provide a systematic study on the stability and generalization analysis of stochastic optimization for problems that can be either nonconvex or nonsmooth. We consider three stability measures: the stability by function values, the stability by gradients and the stability by arguments, which are used to study convex and nonsmooth problems, nonconvex and smooth problems, and weakly convex problems, respectively. We develop connection between stability and generalization gap measured by gradients for either the population risks or the Moreau envelopes. We then develop bounds for

these stability measures for a class of sampling-determined algorithms. As a combination of these stability bounds and the connection between stability and generalization, we develop error bounds for SGD and AdaGrad-Norm, with the performance measured by either functional suboptimality, stationarity by gradients or stationarity by Moreau envelopes. It is interesting to investigate whether our error bounds can be further improved. It is also very interesting to develop lower bounds for learning with weakly convex problems.

Acknowledgments

We thank Prof. Yiming Ying for interesting discussions. We are grateful to the anonymous reviewers and the area chair for their constructive comments and suggestions.

References

Amit Attia and Tomer Koren. Uniform stability for first-order empirical risk minimization. In *Conference on Learning Theory*, pages 3313–3332. PMLR, 2022.

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

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

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

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

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

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

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

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis and Dmitriy Drusvyatskiy. Graphical convergence of subgradients in nonconvex optimization and learning. *Mathematics of Operations Research*, 2021.

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

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Conference on Learning Theory*, page 257, 2010.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

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

C Fang, CJ Li, Z Lin, and T Zhang. Near-optimal non-convex optimization via stochastic path integrated differential estimator. *Advances in Neural Information Processing Systems*, 31:689, 2018.

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

Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pages 8759–8770, 2018.

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

Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017.

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

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Tomer Koren, Roi Livni, Yishay Mansour, and Uri Sherman. Benign underfitting of stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 19605–19617.

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

Yunwen Lei and Ke Tang. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4505–4511, 2021.

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

Yunwen Lei, Ting Hu, and Ke Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *Journal of Machine Learning Research*, 22:1–41, 2021.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2019.

Zhicong Liang, Bao Wang, Quanquan Gu, Stanley Osher, and Yuan Yao. Exploring private federated learning with laplacian smoothing. *arXiv preprint arXiv:2005.00218*, 2020.

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

Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167, 2017.

Ben London, Bert Huang, and Lise Getoor. Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859, 2016.

Andreas Maurer. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6(Jun):967–994, 2005.

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

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2012.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638, 2018.

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

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o $(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Gergely Neu and Gábor Lugosi. Generalization bounds via convex analysis. In *Conference on Learning Theory*, pages 3524–3546, 2022.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

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

Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

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

Dominic Richards and Mike Rabbat. Learning with gradient descent and weakly convex losses. In *International Conference on Artificial Intelligence and Statistics*, pages 1990–1998. PMLR, 2021.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240. PMLR, 2016.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

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

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pages 1195–1204, 2019.

Rachel Ward, Xiaoxia Wu, and Léon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21:1–30, 2020.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 2017:2525–2534, 2017.

Xiaotong Yuan and Ping Li. Stability and risk bounds of iterative hard thresholding. In *International Conference on Artificial Intelligence and Statistics*, pages 1702–1710. PMLR, 2021.

Lijun Zhang and Zhi-Hua Zhou. Stochastic approximation of smooth and strongly convex functions: Beyond the $o(1/t)$ convergence rate. In *Conference on Learning Theory*, pages 3160–3179, 2019.

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

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

Yingxue Zhou, Belhal Karimi, Jinxing Yu, Zhiqiang Xu, and Ping Li. Towards better generalization of adaptive gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020.

Appendix A. Proofs on Stability and Generalization

A.1. Proof of Theorem 6

In this section, we prove the connection between generalization and uniform stability measured by gradients. For brevity, we use $\mathbb{E}[\cdot]$ to denote $\mathbb{E}_{S,A}[\cdot]$. Before proving Theorem 6, we first present the proof of Lemma 5. This result is known in the literature (Shalev-Shwartz et al., 2010; Hardt et al., 2016; Kuzborskij and Lampert, 2018). We give the proof for completeness and for showing that these arguments cannot be used to prove Theorem 6.

Proof of Lemma 5 Let $S' = \{z_1', \ldots, z_n'\}$ be drawn independently from ρ. For any $i \in [n]$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$. According to the symmetry between z_i and z_i' we have

$$\begin{aligned}\mathbb{E}[F_S(A(S)) - F(A(S))] &= \mathbb{E}\big[F_S(A(S)) - \frac{1}{n}\sum_{i=1}^n F(A(S^{(i)}))\big] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(A(S); z_i) - f(A(S^{(i)}); z_i)\big],\end{aligned}$$

where the last identity holds since $A(S^{(i)})$ is independent of z_i. It then follows that

$$\big|\mathbb{E}[F_S(A(S)) - F(A(S))]\big| \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[|f(A(S); z_i) - f(A(S^{(i)}); z_i)|\big] \leq \epsilon.$$

The proof is completed. ■

An essential argument in proving Lemma 5 is to use the identity

$$\mathbb{E}_{S,A}[F(A(S))] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S,A}[f(A(S^{(i)}); z_i)].$$

However, if we consider gradients of population risks we can only get

$$\mathbb{E}_{S,A}\big[\|\nabla F(A(S))\|_2\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S,A}\big[\|\mathbb{E}_{z_i}[\nabla f(A(S^{(i)}); z_i)]\|_2\big],$$

where the summation is outside of $\|\cdot\|_2$. As a comparison, if we consider gradients of empirical risks we get $\|\nabla F_S(A(S))\|_2 = \big\|\frac{1}{n}\sum_{i=1}^n \nabla f(A(S); z_i)\big\|_2$, where the summation is inside the norm. Since we cannot exchange the norm and the summation, we cannot use the argument in the proof of Lemma 5 to prove Theorem 6.

Intuition. We use an error decomposition in Bousquet et al. (2020) to handle this. Our intuitive *idea* is to show that

$$\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|_2 \leq 2\epsilon + \frac{1}{n}\big\|\sum_{i=1}^n \xi_i\big\|_2,$$

where ξ_i is a sequence of mean-zero variables satisfying $\mathbb{E}[\langle \xi_i, \xi_j \rangle] \leq 4\epsilon^2$ for any $i \neq j$. Then one can show that

$$\frac{1}{n^2}\mathbb{E}[\|\sum_{i=1}^n \xi_i\|_2^2] \leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[\|\xi_i\|_2^2] + 4\epsilon^2 = O(1/n + \epsilon^2).$$

Proof of Theorem 6 Let S, S' and $S^{(i)}$ be defined as in the proof of Lemma 5. We have the following error decomposition

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

where we have used $\mathbb{E}_Z[\nabla f(A(S);Z)] = \nabla F(A(S))$. It then follows that

$$n\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|_2 \leq \sum_{i=1}^n \mathbb{E}_{Z,z_i'}\Big[\|\nabla f(A(S);Z) - \nabla f(A(S^{(i)});Z)\|_2\Big]$$
$$+ \Big\|\sum_{i=1}^n \xi_i(S)\Big\|_2 + \sum_{i=1}^n \mathbb{E}_{z_i'}\Big[\Big\|\nabla f(A(S^{(i)});z_i) - \nabla f(A(S);z_i)\Big\|_2\Big],$$

where we introduce ξ_i as a function of S as follows

$$\xi_i(S) = \mathbb{E}_{z_i'}\Big[\mathbb{E}_Z[\nabla f(A(S^{(i)});Z)] - \nabla f(A(S^{(i)});z_i)\Big], \forall i \in [n].$$

Note S and $S^{(i)}$ differ by a single example. By the assumption on stability, we further get

$$n\mathbb{E}\big[\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|_2\big] \leq 2n\epsilon + \mathbb{E}\Big[\|\sum_{i=1}^n \xi_i(S)\|_2\Big]. \tag{A.1}$$

Due to the symmetry between Z and z_i, one can see that

$$\mathbb{E}_{z_i}[\xi_i(S)] = 0, \quad \forall i \in [n]. \tag{A.2}$$

Introduce $S'' = \{z_1'', \ldots, z_n''\}$ which are drawn independently from ρ. For each $i, j \in [n]$ with $i \neq j$, introduce

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

That is, S_j is formed by replacing the j-th element of S with z_j'', while $S_j^{(i)}$ is formed by replacing the j-th element of $S^{(i)}$ with z_j''. If $i \neq j$, then

$$\mathbb{E}\big[\langle \xi_i(S_j), \xi_j(S)\rangle\big] = \mathbb{E}\mathbb{E}_{z_j}\big[\langle \xi_i(S_j), \xi_j(S)\rangle\big] = \mathbb{E}\Big[\langle \xi_i(S_j), \mathbb{E}_{z_j}[\xi_j(S)]\rangle\Big] = 0,$$

where the second identity holds since $\xi_i(S_j)$ is independent of z_j and the last identity follows from $\mathbb{E}_{z_j}\big[\xi_j(S)\big] = 0$ due to (A.2). In a similar way, one can show the following inequalities for $i \neq j$

$$\mathbb{E}\big[\langle \xi_i(S), \xi_j(S_i)\rangle\big] = \mathbb{E}\mathbb{E}_{z_i}\big[\langle \xi_i(S), \xi_j(S_i)\rangle\big] = \mathbb{E}\Big[\langle \xi_j(S_i), \mathbb{E}_{z_i}[\xi_i(S)]\rangle\Big] = 0$$

and

$$\mathbb{E}\big[\langle \xi_i(S_j), \xi_j(S_i)\rangle\big] = \mathbb{E}\mathbb{E}_{z_i}\big[\langle \xi_i(S_j), \xi_j(S_i)\rangle\big] = \mathbb{E}\big[\langle \xi_j(S_i), \mathbb{E}_{z_i}[\xi_i(S_j)]\rangle\big] = 0.$$

As a combination of the above identities we have ($i \neq j$)

$$\begin{aligned}
\mathbb{E}\big[\langle \xi_i(S), \xi_j(S)\rangle\big] &= \mathbb{E}\Big[\big\langle \xi_i(S) - \xi_i(S_j), \xi_j(S) - \xi_j(S_i)\big\rangle\Big] \\
&\leq \mathbb{E}\Big[\big\|\xi_i(S) - \xi_i(S_j)\big\|_2\big\|\xi_j(S) - \xi_j(S_i)\big\|_2\Big] \\
&\leq \frac{1}{2}\mathbb{E}\Big[\big\|\xi_i(S) - \xi_i(S_j)\big\|_2^2\Big] + \frac{1}{2}\mathbb{E}\Big[\big\|\xi_j(S) - \xi_j(S_i)\big\|_2^2\Big], \qquad \text{(A.3)}
\end{aligned}$$

where we have used $ab \leq \frac{1}{2}(a^2 + b^2)$. According to the definition of $\xi_i(S)$ and $\xi_i(S_j)$ we know the following identity for $i \neq j$

$$\begin{aligned}
\mathbb{E}\big[\big\|\xi_i(S) - \xi_i(S_j)\big\|_2^2\big] = \mathbb{E}\Big[\Big\|&\mathbb{E}_{z_i'}\mathbb{E}_Z\big[\nabla f(A(S^{(i)}); Z) - \nabla f(A(S_j^{(i)}); Z)\big] \\
&+ \mathbb{E}_{z_i'}\big[\nabla f(A(S_j^{(i)}); z_i) - \nabla f(A(S^{(i)}); z_i)\big]\Big\|_2^2\Big].
\end{aligned}$$

It then follows from the elementary inequality $(a+b)^2 \leq 2(a^2+b^2)$ and the Jensen's inequality that

$$\begin{aligned}
\mathbb{E}\big[\big\|\xi_i(S) - \xi_i(S_j)\big\|_2^2\big] \leq\ & 2\mathbb{E}\Big[\Big\|\nabla f(A(S^{(i)}); Z) - \nabla f(A(S_j^{(i)}); Z)\Big\|_2^2\Big] \\
&+ 2\mathbb{E}\Big[\Big\|\nabla f(A(S_j^{(i)}); z_i) - \nabla f(A(S^{(i)}); z_i)\Big\|_2^2\Big].
\end{aligned}$$

Since $S^{(i)}$ and $S_j^{(i)}$ differ by one example, it follows from the definition of stability that

$$\mathbb{E}\big[\big\|\xi_i(S) - \xi_i(S_j)\big\|_2^2\big] \leq 4\epsilon^2, \qquad \forall i \neq j.$$

In a similar way, one can show that

$$\mathbb{E}\Big[\big\|\xi_j(S) - \xi_j(S_i)\big\|_2^2\Big] \leq 4\epsilon^2, \qquad \forall i \neq j.$$

We can plug the above two inequalities back into (A.3) and derive the following inequality if $i \neq j$

$$\mathbb{E}\big[\langle \xi_i(S), \xi_j(S)\rangle\big] \leq 4\epsilon^2.$$

Furthermore, according to the definition of $\xi_i(S)$ and Jensen inequality we know

$$\begin{aligned}
\mathbb{E}\big[\|\xi_i(S)\|_2^2\big] &= \mathbb{E}\Big[\Big\|\mathbb{E}_{z_i'}\Big[\mathbb{E}_Z[\nabla f(A(S^{(i)}); Z)] - \nabla f(A(S^{(i)}); z_i)\Big]\Big\|_2^2\Big] \\
&\leq \mathbb{E}\Big[\Big\|\mathbb{E}_Z[\nabla f(A(S^{(i)}); Z)] - \nabla f(A(S^{(i)}); z_i)\Big\|_2^2\Big] \\
&= \mathbb{E}\Big[\Big\|\mathbb{E}_Z[\nabla f(A(S); Z)] - \nabla f(A(S); z_i')\Big\|_2^2\Big] \\
&= \mathbb{E}_S\Big[\mathbb{V}_Z(\nabla f(A(S); Z))\Big],
\end{aligned}$$

where we have used the symmetry between z_i and z_i' (z_i' has the same distribution of Z). It then follows that

$$\mathbb{E}\big[\big\|\sum_{i=1}^n \xi_i(S)\big\|_2^2\big] = \mathbb{E}\big[\sum_{i=1}^n \|\xi_i(S)\|_2^2\big] + \sum_{i,j\in[n]:i\neq j} \mathbb{E}[\langle \xi_i(S), \xi_j(S)\rangle]$$
$$\leq n\mathbb{E}_S\Big[\mathbb{V}_Z(\nabla f(A(S);Z))\Big] + 4n(n-1)\epsilon^2.$$

We can plug the above inequality back into (A.1) and get

$$n\mathbb{E}\big[\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|_2\big] \leq 2n\epsilon + \sqrt{n\mathbb{E}_S\Big[\mathbb{V}_Z(\nabla f(A(S);Z))\Big]} + 2n\epsilon.$$

The proof is completed. ■

A.2. Proofs of Theorem 8 and Theorem 10

In this section, we provide the proof of Theorem 8 and Theorem 10.

Intuition. Before giving the detailed proof, we first sketch the intuition. For any S, define

$$\mathbf{w}_S = \arg\min_{\mathbf{v}\in\mathbb{R}^d} \big\{F_S(\mathbf{v}) + \rho\|\mathbf{v} - A(S)\|_2^2\big\}, \tag{A.4}$$

$$\tilde{\mathbf{w}}_S = \arg\min_{\mathbf{v}\in\mathbb{R}^d} \big\{F(\mathbf{v}) + \rho\|\mathbf{v} - A(S)\|_2^2\big\}. \tag{A.5}$$

According to the definition of $F_{S,1/(2\rho)}$ and $F_{1/(2\rho)}$, we know

$$\nabla F_{S,1/(2\rho)}(A(S)) = 2\rho\big(A(S) - \mathbf{w}_S\big),$$
$$\nabla F_{1/(2\rho)}(A(S)) = 2\rho\big(A(S) - \tilde{\mathbf{w}}_S\big).$$

Then we know

$$\nabla F_{S,1/(2\rho)}(A(S)) - \nabla F_{1/(2\rho)}(A(S)) = 2\rho(\tilde{\mathbf{w}}_S - \mathbf{w}_S). \tag{A.6}$$

It remains to control $\|\tilde{\mathbf{w}}_S - \mathbf{w}_S\|_2$. According to the definition of $\mathbf{w}_S$, we know

$$F_S(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2 \leq F_S(\tilde{\mathbf{w}}_S) + \rho\|\tilde{\mathbf{w}}_S - A(S)\|_2^2. \tag{A.7}$$

Let A be ϵ-uniformly argument stable. We then show that the algorithm defined in Eq. (A.4) is $O(\epsilon + 1/(n\rho))$-uniformly stable, and the algorithm defined in Eq. (A.5) is $O(\epsilon)$-uniformly stable. It then follows from the connection between generalization and stability that

$$\mathbb{E}[F(\mathbf{w}_S) - F_S(\mathbf{w}_S)] = O(\epsilon + 1/(n\rho)),$$
$$\mathbb{E}[F_S(\tilde{\mathbf{w}}_S) - F(\tilde{\mathbf{w}}_S)] = O(\epsilon).$$

We then can replace F_S in Eq. (A.7) by F to get

$$\big(\mathbb{E}[F(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2]\big) - \big(\mathbb{E}[F(\tilde{\mathbf{w}}_S) + \rho\|\tilde{\mathbf{w}}_S - A(S)\|_2^2]\big) = O(\epsilon + 1/(n\rho)).$$

Furthermore, the weak-convexity and the optimality of $\tilde{\mathbf{w}}_S$ show that the left-hand side of the above inequality is larger than $\frac{\rho}{2}\mathbb{E}[\|\mathbf{w}_S - \tilde{\mathbf{w}}_S\|_2^2]$. We then get the desired bound $\mathbb{E}[\|\mathbf{w}_S - \tilde{\mathbf{w}}_S\|_2] = O(\sqrt{\epsilon/\rho} + 1/(\sqrt{n}\rho))$.

We now give the detailed proof. We first introduce two lemmas. Lemma A.1 shows the argument stability of the algorithm $S \mapsto \text{Prox}_{F_S/(2\rho)}(A(S))$ via the argument stability of A. For any $g : \mathcal{W} \mapsto \mathbb{R}$, let $\partial g(\mathbf{w})$ denote the subdifferential of g at $\mathbf{w}$.

Lemma A.1 *Let A be an algorithm. Assume for any S, the function F_S is ρ-weakly-convex. For any S, let $\mathbf{w}_S$ be defined in Eq. (A.4) and assume $\sup_z \|\nabla f(\mathbf{w}_S; z)\|_2 \leq G$. Let S and S' be neighboring datasets. Then*

$$\|\mathbf{w}_S - \mathbf{w}_{S'}\|_2 \leq \frac{2G}{\rho n} + 2\|A(S) - A(S')\|_2.$$

Proof Without loss of generality, we assume S and S' differ by the last element, i.e., $S = \{z_1, \ldots, z_n\}$ and $S = \{z_1, \ldots, z_{n-1}, z_n'\}$. Since F_S is ρ-weakly convex, we know

$$\langle \mathbf{w}_S - \mathbf{w}_{S'}, \partial F_S(\mathbf{w}_S) - \partial F_S(\mathbf{w}_{S'})\rangle \geq -\rho\|\mathbf{w}_S - \mathbf{w}_{S'}\|_2^2. \tag{A.8}$$

According to the first-order optimality condition we know

$$-2\rho\big(\mathbf{w}_S - A(S)\big) \in \partial F_S(\mathbf{w}_S)$$

and

$$-2\rho\big(\mathbf{w}_{S'} - A(S')\big) \in \partial F_{S'}(\mathbf{w}_{S'}) = \partial F_S(\mathbf{w}_{S'}) + \frac{1}{n}\partial f(\mathbf{w}_{S'}; z_n') - \frac{1}{n}\partial f(\mathbf{w}_{S'}; z_n),$$

where we have used the addition property of subdifferential and the definition of F_S, F_S'. We can plug the above two expressions into Eq (A.8) and get

$$\Big\langle \mathbf{w}_S - \mathbf{w}_{S'}, -2\rho(\mathbf{w}_S - A(S)) + 2\rho(\mathbf{w}_{S'} - A(S')) + \frac{1}{n}\partial f(\mathbf{w}_{S'}; z_n') - \frac{1}{n}\partial f(\mathbf{w}_{S'}; z_n)\Big\rangle \geq -\rho\|\mathbf{w}_S - \mathbf{w}_{S'}\|_2^2.$$

It then follows from the Lipschitz continuity that

$$\begin{aligned}
\rho\|\mathbf{w}_S - \mathbf{w}_{S'}\|_2^2 &\leq \Big\langle \mathbf{w}_S - \mathbf{w}_{S'}, 2\rho(A(S) - A(S')) + \frac{\partial f(\mathbf{w}_{S'}; z_n') - \partial f(\mathbf{w}_{S'}; z_n)}{n}\Big\rangle \\
&\leq \|\mathbf{w}_S - \mathbf{w}_{S'}\|_2 \Big\| 2\rho(A(S) - A(S')) + \frac{\partial f(\mathbf{w}_{S'}; z_n') - \partial f(\mathbf{w}_{S'}; z_n)}{n}\Big\|_2 \\
&\leq \|\mathbf{w}_S - \mathbf{w}_{S'}\|_2 \Big(2\rho\|A(S) - A(S')\|_2 + \frac{2G}{n}\Big).
\end{aligned}$$

The stated bound then follows. The proof is completed. ■

The following lemma connects the argument stability of the algorithm $S \mapsto \text{Prox}_{F/(2\rho)}(A(S))$ via that of A.

Lemma A.2 *Let A be an algorithm and F be ρ-weakly-convex. For any S, let $\tilde{\mathbf{w}}_S$ be defined in Eq. (A.5). Let S and S' be neighboring datasets. Then*

$$\|\tilde{\mathbf{w}}_S - \tilde{\mathbf{w}}_{S'}\|_2 \leq 2\|A(S) - A(S')\|_2.$$

Proof By the weak convexity of F we know

$$\langle \tilde{\mathbf{w}}_S - \tilde{\mathbf{w}}_{S'}, \partial F(\tilde{\mathbf{w}}_S) - \partial F(\tilde{\mathbf{w}}_{S'})\rangle \geq -\rho\|\tilde{\mathbf{w}}_S - \tilde{\mathbf{w}}_{S'}\|_2^2.$$

According to the first-order optimality condition we know

$$\begin{aligned} -2\rho\big(\tilde{\mathbf{w}}_S - A(S)\big) &\in \partial F(\tilde{\mathbf{w}}_S), \\ -2\rho\big(\tilde{\mathbf{w}}_{S'} - A(S')\big) &\in \partial F(\tilde{\mathbf{w}}_{S'}). \end{aligned}$$

As a combination of the above three inequalities, we get

$$\Big\langle \tilde{\mathbf{w}}_S - \tilde{\mathbf{w}}_{S'}, 2\rho\big(\tilde{\mathbf{w}}_{S'} - A(S')\big) - 2\rho\big(\tilde{\mathbf{w}}_S - A(S)\big)\Big\rangle \geq -\rho\|\tilde{\mathbf{w}}_S - \tilde{\mathbf{w}}_{S'}\|_2^2.$$

It then follows from the Lipschitz continuity that

$$\begin{aligned} \rho\|\tilde{\mathbf{w}}_S - \tilde{\mathbf{w}}_{S'}\|_2^2 &\leq \Big\langle \tilde{\mathbf{w}}_S - \tilde{\mathbf{w}}_{S'}, 2\rho A(S) - 2\rho A(S')\Big\rangle \\ &\leq 2\rho\|\tilde{\mathbf{w}}_S - \tilde{\mathbf{w}}_{S'}\|_2\|A(S) - A(S')\|_2. \end{aligned}$$

The stated inequality then follows directly. ■

Proof of Theorem 8 For any S, define $\mathbf{w}_S$ and $\tilde{\mathbf{w}}_S$ according to Eq. (A.4) and Eq. (A.5), respectively. According to Lemma A.2 and the Lipschitz continuity assumption (A is ϵ-argument stable), we know that the algorithm defined by (A.5) is $2G\epsilon$-uniformly stable in function values. It then follows from Lemma 5 that

$$\mathbb{E}\big[F_S(\tilde{\mathbf{w}}_S) - F(\tilde{\mathbf{w}}_S)\big] \leq 2G\epsilon. \tag{A.9}$$

According to Lemma A.1, we know that the algorithm defined by (A.4) is $(\frac{2G^2}{n\rho} + 2G\epsilon)$-uniformly stable. It then follows from Lemma 5 that

$$\mathbb{E}[F(\mathbf{w}_S) - F_S(\mathbf{w}_S)] \leq \frac{2G^2}{n\rho} + 2G\epsilon.$$

It then follows that

$$\begin{aligned} &\mathbb{E}[F(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2] - \mathbb{E}[F_S(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2] \\ &\qquad = \mathbb{E}[F(\mathbf{w}_S) - F_S(\mathbf{w}_S)] \leq \frac{2G^2}{n\rho} + 2G\epsilon. \end{aligned} \tag{A.10}$$

Furthermore, according to the definition of $\mathbf{w}_S$ we know

$$F_S(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2 \leq F_S(\tilde{\mathbf{w}}_S) + \rho\|\tilde{\mathbf{w}}_S - A(S)\|_2^2$$

and therefore it follows from (A.9) that

$$\begin{aligned}\mathbb{E}[F_S(\mathbf{w}_S)+\rho\|\mathbf{w}_S-A(S)\|_2^2] &\leq \mathbb{E}[F_S(\tilde{\mathbf{w}}_S)+\rho\|\tilde{\mathbf{w}}_S-A(S)\|_2^2] \\ &\leq \mathbb{E}[F(\tilde{\mathbf{w}}_S)+\rho\|\tilde{\mathbf{w}}_S-A(S)\|_2^2]+2G\epsilon.\end{aligned}$$

We can combine (A.10) and the above inequality together, and derive

$$\mathbb{E}[F(\mathbf{w}_S)+\rho\|\mathbf{w}_S-A(S)\|_2^2]-\mathbb{E}[F(\tilde{\mathbf{w}}_S)+\rho\|\tilde{\mathbf{w}}_S-A(S)\|_2^2] \leq \frac{2G^2}{n\rho}+4G\epsilon.$$

According to the ρ-strong convexity of $\mathbf{v}\mapsto F(\mathbf{v})+\rho\|\mathbf{v}-A(S)\|_2^2$ (this strong convexity follows from the weak convexity of F) and the definition of $\tilde{\mathbf{w}}_S$ as a minimizer, we know

$$\mathbb{E}[F(\mathbf{w}_S)+\rho\|\mathbf{w}_S-A(S)\|_2^2]-\mathbb{E}[F(\tilde{\mathbf{w}}_S)+\rho\|\tilde{\mathbf{w}}_S-A(S)\|_2^2] \geq \frac{\rho}{2}\mathbb{E}[\|\mathbf{w}_S-\tilde{\mathbf{w}}_S\|_2^2].$$

We can combine the above two inequalities together and derive

$$\frac{\rho}{2}\mathbb{E}[\|\mathbf{w}_S-\tilde{\mathbf{w}}_S\|_2^2] \leq \frac{2G^2}{n\rho}+4G\epsilon.$$

It then follows that

$$\mathbb{E}[\|\mathbf{w}_S-\tilde{\mathbf{w}}_S\|_2] \leq \frac{2G}{\sqrt{n}\rho}+\sqrt{8G\epsilon/\rho}. \tag{A.11}$$

It then follows from Eq. (A.6) that

$$\mathbb{E}\big[\big\|\nabla F_{S,1/(2\rho)}(A(S))-\nabla F_{1/(2\rho)}(A(S))\big\|_2\big]=2\rho\mathbb{E}\big[\big\|\mathbf{w}_S-\tilde{\mathbf{w}}_S\big\|_2\big] \leq \frac{4G}{\sqrt{n}}+\sqrt{32G\epsilon\rho}.$$

The proof is completed. ■

Appendix B. Proof of Theorem 10

In this section, we prove the high probability bounds. To this aim, we first introduce a useful lemma.

Lemma B.1 (Bousquet et al. 2020) *Let A be an ϵ-uniformly stable algorithm. Assume $f(A(S);z)\leq R$ almost surely. Then for any $\delta\in(0,1)$ with probability at least $1-\delta$ we have*

$$\big|F_S(A(S))-F(A(S))\big|=O\Big(\epsilon\log(n)\log(1/\delta)+R\sqrt{n^{-1}\log(1/\delta)}\Big).$$

Proof of Theorem 10 For any S, define $\mathbf{w}_S$ and $\tilde{\mathbf{w}}_S$ according to Eq. (A.4) and Eq. (A.5), respectively. According to Lemma A.2 and the Lipschitz continuity assumption, we know that the algorithm defined by (A.5) is $2G\epsilon$-uniformly stable in function values. The following inequality then follows from Lemma B.1 with probability at least $1-\delta/2$

$$F_S(\tilde{\mathbf{w}}_S)-F(\tilde{\mathbf{w}}_S)=O\Big(\epsilon\log(n)\log(1/\delta)+\sqrt{n^{-1}\log(1/\delta)}\Big). \tag{B.1}$$

According to Lemma A.1, we know that the algorithm defined by (A.4) is $\left(\frac{2G^2}{n\rho} + 2G\epsilon\right)$-uniformly stable. The following inequality then follows from Lemma B.1 with probability at least $1 - \delta/2$

$$F(\mathbf{w}_S) - F_S(\mathbf{w}_S) = O\Big(\big(G^2(n\rho)^{-1} + G\epsilon\big)\log(n)\log(1/\delta) + \sqrt{n^{-1}\log(1/\delta)}\Big).$$

It then follows that

$$\begin{aligned}\big(F(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2\big) - \big(F_S(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2\big) \\ = O\Big(\big(G^2(n\rho)^{-1} + G\epsilon\big)\log(n)\log(1/\delta) + \sqrt{n^{-1}\log(1/\delta)}\Big). \quad \text{(B.2)}\end{aligned}$$

Furthermore, according to the definition of $\mathbf{w}_S$ and (B.1) we know

$$\begin{aligned}F_S(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2 &\le F_S(\tilde{\mathbf{w}}_S) + \rho\|\tilde{\mathbf{w}}_S - A(S)\|_2^2 \\ &\le F(\tilde{\mathbf{w}}_S) + \rho\|\tilde{\mathbf{w}}_S - A(S)\|_2^2 + O\Big(\epsilon\log(n)\log(1/\delta) + \sqrt{n^{-1}\log(1/\delta)}\Big).\end{aligned}$$

We can combine Eq. (B.2) and the above inequality together, and derive the following inequality with probability at least $1 - \delta$

$$\begin{aligned}\big(F(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2\big) - \big(F(\tilde{\mathbf{w}}_S) + \rho\|\tilde{\mathbf{w}}_S - A(S)\|_2^2\big) \\ = O\Big(\big(G^2(n\rho)^{-1} + G\epsilon\big)\log(n)\log(1/\delta) + \sqrt{n^{-1}\log(1/\delta)}\Big).\end{aligned}$$

According to the ρ-strong convexity of $\mathbf{v} \mapsto F(\mathbf{v}) + \rho\|\mathbf{v} - A(S)\|_2^2$ and the definition of $\tilde{\mathbf{w}}_S$, we know the following inequality

$$\big(F(\mathbf{w}_S) + \rho\|\mathbf{w}_S - A(S)\|_2^2\big) - \big(F(\tilde{\mathbf{w}}_S) + \rho\|\tilde{\mathbf{w}}_S - A(S)\|_2^2\big) \ge \frac{\rho}{2}\|\mathbf{w}_S - \tilde{\mathbf{w}}_S\|_2^2.$$

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

$$\frac{\rho}{2}\|\mathbf{w}_S - \tilde{\mathbf{w}}_S\|_2^2 = O\Big(\big(G^2(n\rho)^{-1} + G\epsilon\big)\log(n)\log(1/\delta) + \sqrt{n^{-1}\log(1/\delta)}\Big),$$

from which we derive

$$\|\mathbf{w}_S - \tilde{\mathbf{w}}_S\|_2 = O\Big(\big(Gn^{-\frac{1}{2}}\rho^{-1} + \sqrt{G\epsilon/\rho}\big)\sqrt{\log(n)\log(1/\delta)} + \big(n^{-1}\rho^{-2}\log(1/\delta)\big)^{\frac{1}{4}}\Big).$$

The stated bound then follows from Eq. (A.6). The proof is completed. ■

Appendix C. Proofs on Uniform Stability Bounds

In this section, we present the proofs on the uniform stability bounds of sampling-determined algorithms. Our proof follows the idea in Hardt et al. (2016).

Proof of Theorem 16 Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z_1', \ldots, z_n'\}$. Without loss of generality, we assume S and S' differ only by the last example, i.e., $z_n \neq z_n'$. Let $I(A) = \{i_1, \ldots, i_T\}$ be the set

of indices selected in the implementation of A. We first prove Part (a). According to the property of conditional expectation, we know

$$\begin{aligned}\mathbb{E}_A\big[f(A(S);z)-f(A(S');z)\big] &= \mathbb{E}_A\big[f(A(S);z)-f(A(S');z)|n\notin I(A)\big]\Pr\{n\notin I(A)\}\\ &\quad+\mathbb{E}_A\big[f(A(S);z)-f(A(S');z)|n\in I(A)\big]\Pr\{n\in I(A)\}.\end{aligned}$$

Since A is a sampling-determined algorithm, $A(S)$ is independent of z_n under the condition $n\notin I(A)$. Therefore, under the condition $n\notin I(A)$ we have $A(S)=A(S')$. Therefore,

$$\begin{aligned}\mathbb{E}_A\big[f(A(S);z)-f(A(S');z)\big] &= \mathbb{E}_A\big[f(A(S);z)-f(A(S');z)|n\in I(A)\big]\Pr\{n\in I(A)\}\\ &\le 2B\Pr\{n\in I(A)\},\end{aligned}$$

where we have used the assumption $\mathbb{E}_A\big[f(A(S);z)|n\in I(A)\big]\le B$ for any S.

We now turn to Part (b). It is clear

$$\begin{aligned}\mathbb{E}_A\big[\|\nabla f(A(S);z)-\nabla f(A(S');z)\|_2^2\big] &= \mathbb{E}_A\big[\|\nabla f(A(S);z)-\nabla f(A(S');z)\|_2^2|n\notin I(A)\big]\Pr\{n\notin I(A)\}\\ &\quad+\mathbb{E}_A\big[\|\nabla f(A(S);z)-\nabla f(A(S');z)\|_2^2|n\in I(A)\big]\Pr\{n\in I(A)\}.\end{aligned}$$

It then follows that

$$\begin{aligned}\mathbb{E}_A\big[\|\nabla f(A(S);z)-\nabla f(A(S');z)\|_2^2\big] &= \mathbb{E}_A\big[\|\nabla f(A(S);z)-\nabla f(A(S');z)\|_2^2|n\in I(A)\big]\Pr\{n\in I(A)\}\\ &\le 4G^2\Pr\{n\in I(A)\},\end{aligned}$$

where we have used the assumption $\mathbb{E}_A[\|\nabla f(A(S);z)|n\in I(A)\|_2^2]\le G^2$ for any S.

Finally, we consider Part (c). It is clear

$$\begin{aligned}&\mathbb{E}_A\big[\|A(S)-A(S')\|_2\big]\\ &=\mathbb{E}_A\big[\|A(S)-A(S')\|_2|n\notin I(A)\big]\Pr\{n\notin I(A)\}+\mathbb{E}_A\big[\|A(S)-A(S')\|_2|n\in I(A)\big]\Pr\{n\in I(A)\}\\ &=\mathbb{E}_A\big[\|A(S)-A(S')\|_2|n\in I(A)\big]\Pr\{n\in I(A)\}\le 2R\Pr\{n\in I(A)\},\end{aligned}$$

where we have used the assumption $\mathbb{E}_A[\|A(S)\|_2|n\in I(A)]\le R$ for any S. The proof is completed. ■

Proof of Corollary 17 We consider only SGD (the arguments of AdaGrad-Norm are the same). It is clear that

$$\Pr\{n\in I(A)\}\le\sum_{t=1}^T\Pr\{i_t=n\}\le\frac{T}{n}.$$

It is clear that the algorithm A is sampling-determined, and therefore one can apply Theorem 16 to derive the stated bounds. The proof is completed. ■

Appendix D. Proofs on Stochastic Gradient Descent

The following lemma establishes the optimization error bounds of SGD. Part (a) is a standard result in optimization. Part (b) is due to Ghadimi and Lan (2013), Part (c) is due to Vaswani et al. (2019) and Part (d) is due to Davis and Drusvyatskiy (2019).

Lemma D.1 (Optimization Error Bound for SGD) *Let $\{\mathbf{w}_t\}_t$ be produced by SGD and*

$$\mathbb{E}_A[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2, \quad \forall t \in [T].$$

(a) If F_S is convex, then for all $t \in \mathbb{N}$ and $\mathbf{w}$

$$\mathbb{E}_A\Big[F_S\Big(\frac{\sum_{t=1}^T \eta_t \mathbf{w}_t}{\sum_{t=1}^T \eta_t}\Big)\Big] - F_S(\mathbf{w}) \leq \frac{G^2 \sum_{t=1}^T \eta_t^2 + \|\mathbf{w}\|_2^2}{2\sum_{t=1}^T \eta_t}.$$

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

$$\sum_{t=1}^T \eta_t \mathbb{E}_A\big[\|\nabla F_S(\mathbf{w}_t)\|_2^2\big] \leq F_S(\mathbf{w}_1) + \frac{LG^2}{2}\sum_{t=1}^T \eta_t^2.$$

(c) Assume for all z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is L-smooth and SGC holds with the parameter ρ. If $\eta_t = 1/(\rho L)$, then

$$\sum_{t=1}^T \mathbb{E}_A\big[\|\nabla F_S(\mathbf{w}_t)\|_2^2\big] \leq 2\rho L f(\mathbf{w}_1).$$

(d) If F_S is ρ-weakly convex, then

$$\sum_{t=1}^T \eta_t \mathbb{E}_A\big[\|\nabla F_{S,1/(2\rho)}(\mathbf{w}_t)\|_2^2\big] = O\Big(1 + G^2\rho\sum_{t=1}^T \eta_t^2\Big).$$

Proof of Proposition 19 According to Lemma D.1, Part (a), we have the following optimization error bounds

$$\mathbb{E}_A[F_S(\bar{\mathbf{w}}_T)] - F_S(\mathbf{w}^*) = O\Big(\frac{T\eta^2 G^2 + \|\mathbf{w}^*\|_2^2}{T\eta}\Big).$$

Furthermore, by Corollary 17, Part (a), we have the following stability bounds

$$\sup_z \mathbb{E}\big[f(\bar{\mathbf{w}}_T; z) - f(\bar{\mathbf{w}}'_T; z)\big] \leq \frac{2BT}{n},$$

where $\{\mathbf{w}'_t\}$ is a sequence of iterates produced by SGD based on a neighboring dataset S'. This together with Lemma 5 on the connection between uniform stability and generalization further implies

$$\mathbb{E}\big[F(\bar{\mathbf{w}}_T) - F_S(\bar{\mathbf{w}}_T)\big] = O\big(BT/n\big).$$

We can plug the above generalization error and optimization error bounds into (3.1), and derive (5.1).

If $\eta \asymp \frac{\|\mathbf{w}^*\|_2}{Gn^{\frac{1}{3}}}$ and $T \asymp \frac{n^{\frac{2}{3}}G\|\mathbf{w}^*\|_2}{B}$, we have

$$G^2\eta \asymp \frac{G\|\mathbf{w}^*\|_2}{n^{\frac{1}{3}}}, \quad T\eta \asymp \frac{n^{\frac{2}{3}}G\|\mathbf{w}^*\|_2}{B}\frac{\|\mathbf{w}^*\|_2}{Gn^{\frac{1}{3}}} = \frac{n^{\frac{1}{3}}\|\mathbf{w}^*\|_2^2}{B},$$
$$\frac{BT}{n} \asymp \frac{n^{\frac{2}{3}}G\|\mathbf{w}^*\|_2}{n} = \frac{G\|\mathbf{w}^*\|_2}{n^{\frac{1}{3}}}.$$

The bound $\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O((B + G\|\mathbf{w}^*\|_2)n^{-\frac{1}{3}})$ follows directly from the choice of η and T. The proof is completed. ■

Proof of Proposition 21 According to Lemma D.1, Part (b), we have the following optimization error bounds

$$\mathbb{E}_A\big[\|\nabla F_S(\mathbf{w}_r)\|_2^2\big] = O\Big(\frac{T\eta^2G^2 + 1}{T\eta}\Big)$$

and therefore

$$\mathbb{E}_A\big[\|\nabla F_S(\mathbf{w}_r)\|_2\big] = O\big(G\sqrt{\eta} + 1/\sqrt{T\eta}\big). \tag{D.1}$$

It is clear that A is sampling-determined and one can apply Corollary 17, Part (b) to show the following uniform stability bounds

$$\sup_z \mathbb{E}\big[\|\nabla f(\mathbf{w}_r; z) - \nabla f(\mathbf{w}_r'; z)\|_2^2\big] \leq \frac{4G^2T}{n},$$

where $\{\mathbf{w}_t'\}$ is a sequence of iterates produced by SGD based on a neighboring dataset S'. This together with (3.2) and the connection between uniform stability and generalization established in Theorem 6 gives

$$\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] \leq 8G\sqrt{T/n} + G/\sqrt{n} + \mathbb{E}\big[\|\nabla F_S(\mathbf{w}_r)\|_2\big].$$

We can plug the optimization error bounds (D.2) into the above bound, and get

$$\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O\Big(\frac{G\sqrt{T}}{\sqrt{n}} + \frac{G\sqrt{T}\eta + 1}{\sqrt{T\eta}}\Big).$$

If we choose $\eta \asymp 1/(G\sqrt{T})$, we get

$$\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O\Big(\frac{G\sqrt{T}}{\sqrt{n}} + \frac{\sqrt{G}}{T^{\frac{1}{4}}}\Big).$$

We can choose $T \asymp n^{\frac{2}{3}}/G^{\frac{2}{3}}$ to derive the stated bound $\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O(G^{\frac{2}{3}}n^{-\frac{1}{6}})$. ■

Proof of Proposition 24 Analogous to the proof of Proposition 21, we have

$$\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] \leq 8G\sqrt{T/n} + G/\sqrt{n} + \mathbb{E}\big[\|\nabla F_S(\mathbf{w}_r)\|_2\big].$$

Furthermore, Lemma D.1, Part (c) implies

$$\mathbb{E}\big[\|\nabla F_S(\mathbf{w}_r)\|_2\big] = O(\sqrt{L\rho}/\sqrt{T}).$$

We can combine the above two bounds together and get

$$\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O\Big(G\sqrt{T/n} + \sqrt{L\rho}/\sqrt{T}\Big).$$

Therefore, we can choose $T \asymp \sqrt{L\rho n}/G$ and get $\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O((L\rho G^2/n)^{\frac{1}{4}})$. The proof is completed. ■

Proof of Proposition 25 According to Lemma D.1, Part (d), we have the following optimization error bounds

$$\mathbb{E}_A\big[\|\nabla F_{S,1/(2\rho)}(\mathbf{w}_r)\|_2^2\big] = O\Big(\frac{\rho T G^2\eta^2 + 1}{T\eta}\Big)$$

and therefore

$$\mathbb{E}_A\big[\|\nabla F_{S,1/(2\rho)}(\mathbf{w}_r)\|_2\big] = O\big(G\sqrt{\rho\eta} + 1/\sqrt{T\eta}\big). \tag{D.2}$$

We can apply Corollary 17, Part (c) to show the following argument stability bounds

$$\mathbb{E}_A\big[\|A(S) - A(S')\|_2\big] \leq \frac{2RT}{n}.$$

This together with (3.3) and the connection between argument stability and generalization established in Theorem 8 gives

$$\mathbb{E}\big[\big\|\nabla F_{1/(2\rho)}(A(S))\big\|_2\big] \leq \mathbb{E}\big[\big\|\nabla F_{S,1/(2\rho)}(A(S))\big\|_2\big] + \frac{4G}{\sqrt{n}} + \sqrt{64GRT\rho n^{-1}}. \tag{D.3}$$

We can plug the optimization error bounds (D.2) into the above bound, and get

$$\mathbb{E}\big[\big\|\nabla F_{1/(2\rho)}(A(S))\big\|_2\big] = O\Big(G\sqrt{\rho\eta} + \sqrt{GR\rho T/n} + 1/\sqrt{T\eta} + G/\sqrt{n}\Big).$$

If we choose $\eta \asymp 1/(G\sqrt{\rho T})$, we get

$$\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O\Big(\sqrt{G}(\rho/T)^{\frac{1}{4}} + \sqrt{GR\rho T/n} + G/\sqrt{n}\Big).$$

We can choose $T \asymp n^{\frac{2}{3}}/(R^{\frac{2}{3}}\rho^{\frac{1}{3}})$ to derive the stated bound $\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O(\sqrt{G}\rho^{\frac{1}{3}}R^{\frac{1}{6}}n^{-\frac{1}{6}})$. The proof is completed. ■

Appendix E. AdaGrad-Norm

E.1. Generalization Bounds of AdaGrad-Norm

We now turn to the generalization analysis of AdaGrad-Norm. Proposition E.1 presents the risk bounds in terms of function values for convex and nonsmooth problems, while Proposition E.2 presents the risk bounds in terms of gradients for nonconvex and smooth problems. Note that these bounds match the corresponding results for SGD (w.r.t. n) in Section 5 up to a logarithmic factor. All the proofs are given in Section E.2.

Proposition E.1 (Convex and Nonsmooth Case) *Let $\{\mathbf{w}_t\}_t$ be produced by (4.7), $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2$ for all $t \in [T]$ and $\sup_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}\|_2 \leq R$. Let A output $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$. If F_S is convex we have*

$$\mathbb{E}_{S,A}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O\big(GRT/n\big) + O(G(R + \|\mathbf{w}^*\|_2)/\sqrt{T}).$$

If $T \asymp n^{\frac{2}{3}}$ we have

$$\mathbb{E}_{S,A}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(G(R + \|\mathbf{w}^*\|_2)n^{-\frac{1}{3}}).$$

Proposition E.2 (Nonconvex and Smooth Case) *Let $\{\mathbf{w}_t\}_t$ be produced by (4.7) and $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2$ for all $t \in [T]$. If $A(S) = \mathbf{w}_r$ and F_S is L-smooth, then*

$$\mathbb{E}_{S,A,r}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O\Big(G\sqrt{T/n} + GT^{-\frac{1}{4}}\log^{\frac{1}{2}} T\Big).$$

If $T \asymp n^{\frac{2}{3}}$, one gets

$$\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O(Gn^{-\frac{1}{6}}\log^{\frac{1}{2}} n).$$

Remark E.3 Generalization behavior of adaptive gradient descent was recently studied by Zhou et al. (2020). They considered minibatch adaptive algorithms with a sufficiently large batch size, while the algorithms we consider here use only a single example to compute a stochastic gradient and is therefore more computationally efficient. Their analysis is based on a connection between generalization and differential privacy, and requires to add noise to achieve differential privacy. This in turn leads to a dimension-dependent bound. As a comparison, we do not require to introduce noise in algorithms and our bounds are dimension-free.

E.2. Proofs on AdaGrad-Norm

The following lemma establishes the convergence rates of AdaGrad-Norm. Part (a) is for convex and nonsmooth problems, while Part (b) is for nonconvex and smooth problems. We give a simple proof of Part (a), while the proof of Part (b) can be found in Ward et al. (2020).

Lemma E.4 (Optimization Error Bound for AdaGrad-Norm) *Let $\{\mathbf{w}_t\}$ be the sequence produced by AdaGrad-Norm.*

(a) Let F_S be convex. Assume $\|\mathbf{w}\|_2 \leq R$ for all $\mathbf{w} \in \mathcal{W}$. Then the following bound holds for all $\mathbf{w} \in \mathcal{W}$

$$\mathbb{E}_A\big[F_S(\bar{\mathbf{w}}_T)\big] - F_S(\mathbf{w}) = O(G(R + \|\mathbf{w}\|_2)/\sqrt{T}).$$

(b) Assume F_S is L-smooth, $\mathbb{E}\big[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2\big] \leq G^2$ for all $\mathbf{w}_t$. Then

$$\mathbb{E}_{A,r}\big[\|\nabla F_S(\mathbf{w}_r)\|_2\big] = O\Big(GT^{-\frac{1}{4}}\log^{\frac{1}{2}} T\Big),$$

where r follows from the uniform distribution over $[T]$.

Proof Denote $\eta_t = \eta/b_t$, then (4.7) can be written as $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{i_t}))$. It then follows that

$$\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 - 2\eta_t\langle \mathbf{w}_t - \mathbf{w}, \nabla f(\mathbf{w}_t; z_{i_t})\rangle + \eta_t^2\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2.$$

Re-arranging the above inequality gives

$$\langle \mathbf{w}_t - \mathbf{w}, \nabla f(\mathbf{w}_t; z_{i_t}) \rangle \leq \frac{1}{2\eta_t}\Big(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\Big) + \frac{\eta_t}{2}\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2.$$

We take conditional expectation w.r.t. z_{i_t} over both sides and get

$$\langle \mathbf{w}_t - \mathbf{w}, \nabla F_S(\mathbf{w}_t) \rangle \leq \mathbb{E}_{i_t}\Big[\frac{1}{2\eta_t}\Big(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\Big)\Big] + \mathbb{E}_{i_t}\Big[\frac{\eta_t}{2}\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2\Big].$$

It then follows from the convexity of F_S that

$$F_S(\mathbf{w}_t) - F_S(\mathbf{w}) \leq \mathbb{E}_{i_t}\Big[\frac{1}{2\eta_t}\Big(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\Big)\Big] + \mathbb{E}_{i_t}\Big[\frac{\eta_t}{2}\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2\Big].$$

We can take an expectation followed with a summation of the above inequality from $t = 1$ to $t = T$, and get

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}_A\big[F_S(\mathbf{w}_t) - F_S(\mathbf{w})\big] - \mathbb{E}_A\big[\frac{1}{2\eta_1}\|\mathbf{w}_1 - \mathbf{w}\|_2^2\big] \\
&\leq \frac{1}{2}\sum_{t=2}^T \mathbb{E}_A\Big[\|\mathbf{w}_t - \mathbf{w}\|_2^2\Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\Big)\Big] + \frac{1}{2}\sum_{t=1}^T \mathbb{E}_A\Big[\eta_t\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2\Big] \\
&\leq (R^2 + \|\mathbf{w}\|_2^2)\sum_{t=2}^T \mathbb{E}_A\Big[\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\Big] + \frac{\eta}{2}\sum_{t=1}^T \mathbb{E}_A\Big[\frac{\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2}{\sqrt{\sum_{\tau=1}^t \|\nabla f(\mathbf{w}_\tau; z_{i_\tau})\|_2^2}}\Big] \\
&\leq (R^2 + \|\mathbf{w}\|_2^2)\eta^{-1}\mathbb{E}_A\Big[\Big(\sum_{t=1}^T \|\nabla f(\mathbf{w}_\tau; z_{i_t})\|_2^2\Big)^{\frac{1}{2}}\Big] + \eta\mathbb{E}_A\Big[\Big(\sum_{t=1}^T \|\nabla f(\mathbf{w}_\tau; z_{i_t})\|_2^2\Big)^{\frac{1}{2}}\Big],
\end{aligned}$$

where we have used the following inequality in the last step

$$\sum_{t=1}^T \frac{a_t}{\sqrt{\sum_{j=1}^t a_j}} \leq \sum_{t=1}^T \int_{\sum_{j=1}^{t-1} a_j}^{\sum_{j=1}^t a_j} \frac{1}{\sqrt{x}} dx = \int_0^{\sum_{j=1}^T a_j} \frac{1}{\sqrt{x}} dx = 2\sqrt{\sum_{t=1}^T a_t}.$$

It then follows from the convexity of F_S that

$$\mathbb{E}_A\Big[F_S(\bar{\mathbf{w}}_T) - F_S(\mathbf{w})\Big] = O\Big(\frac{1}{T}\mathbb{E}_A\Big[\Big(\sum_{t=1}^T \|\nabla f(\mathbf{w}_\tau; z_{i_t})\|_2^2\Big)^{\frac{1}{2}}\Big]\Big((R^2 + \|\mathbf{w}\|_2^2)\eta^{-1} + \eta\Big)\Big).$$

The stated bound then follows. ■

Proof of Proposition E.1 Analogous to the proof of Proposition 19, we have the following generalization error bound

$$\mathbb{E}\big[F(\bar{\mathbf{w}}_T) - F_S(\bar{\mathbf{w}}_T)\big] = O\big(GRT/n\big).$$

Lemma E.4, Part (a), implies the following optimization error bound

$$\mathbb{E}_A\big[F_S(\bar{\mathbf{w}}_T)\big] - F_S(\mathbf{w}^*) = O(G(R + \|\mathbf{w}^*\|_2)/\sqrt{T}).$$

We can plug the above two inequalities back into (3.1) and get

$$\mathbb{E}\big[F(\bar{\mathbf{w}}_T)\big] - F(\mathbf{w}^*) = O\big(GRT/n\big) + O(G(R + \|\mathbf{w}^*\|_2)/\sqrt{T}).$$

One can derive the stated bound by setting $T \asymp n^{\frac{2}{3}}$. The proof is completed. ■

Proof of Proposition E.2 Analogous to the Proof of Proposition 21, we have the following generalization error bound

$$\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] \leq 8G\sqrt{T/n} + G/\sqrt{n} + \mathbb{E}\big[\|\nabla F_S(\mathbf{w}_r)\|_2\big].$$

Lemma E.4, Part (b), implies the following optimization error bound

$$\mathbb{E}_{A,r}\big[\|\nabla F_S(\mathbf{w}_r)\|_2\big] = O\Big(GT^{-\frac{1}{4}}\log^{\frac{1}{2}} T\Big).$$

We can combine the above two inequalities together and get

$$\mathbb{E}\big[\|\nabla F(\mathbf{w}_r)\|_2\big] = O\Big(G\sqrt{T/n} + GT^{-\frac{1}{4}}\log^{\frac{1}{2}} T\Big).$$

One can choose $T \asymp n^{\frac{2}{3}}$ to get the stated bound. The proof is completed. ■

Appendix F. Differentially Private SGD

F.1. Utility and Privacy Guarantee

In this section, we use our stability analysis to develop a differentially private SGD with generalization guarantee for weakly-convex problems, which is useful to handle data with sensitive information (Dwork, 2008). We first introduce the definition of *differential privacy*, which is a well-accepted mathematical definition of privacy.

Definition F.1 (Differential Privacy) *Let $\epsilon > 0$ and $\delta \in (0, 1)$. A randomized mechanism $\mathcal{A}$ provides (ϵ, δ)-differential privacy (DP) if for any two neighboring datasets S and S' and any set E in the range of $\mathcal{A}$ there holds*

$$\mathbb{P}(\mathcal{A}(S) \in E) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(S') \in E) + \delta.$$

Our basic idea to develop differentially private algorithms is to inject noise in the learning process to mask the influence of any single datapoint. In particular, at the t-th iteration we randomly sample a noise b_t from a Gaussian distribution with a variance $\sigma^2\mathbb{I}_d$ and build a new stochastic gradient as $\nabla f(\mathbf{w}_t; z_{i_t}) + b_t$. Then we move along the negative direction of this stochastic gradient as follows

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\big(\mathbf{w}_t - \eta_t(\nabla f(\mathbf{w}_t; z_{i_t}) + b_t)\big), \tag{F.1}$$

where β is a parameter and

$$\sigma^2 = \frac{14G^2T}{\beta n^2\epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big). \tag{F.2}$$

We refer to our algorithm as DP-SGD and summarize the implementation in Algorithm 1. Proposition F.2 shows that Algorithm 1 achieves the (ϵ, δ)-privacy guarantee, while Proposition F.3 gets the utility guarantee as measured by $\|\nabla F(\mathbf{w}_r)\|_2$. The proofs are given in Section F.2.

Algorithm 1: Differentially Private SGD

Input: $\mathbf{w}_1 = 0$, learning rates $\{\eta_t\}_t$, parameter $\beta, \epsilon, \delta > 0$ and dataset $S = \{z_1, \ldots, z_n\}$
for $t = 1, 2, \ldots, T$ **do**
  compute σ by Eq. (F.2)
  draw i_t uniformly from $[n]$ and $b_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$
  update $\mathbf{w}_{t+1}$ according to Eq. (F.1)
end
Output: $\mathbf{w}_r$ where $r \sim \text{unif}[T]$

Proposition F.2 (Privacy guarantee) *Let $\epsilon > 0$ and $\delta \in (0, 1)$. Assume for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is G-Lipschitz. If*

$$\epsilon \geq \frac{14T}{3n^2} \quad \text{and} \quad \frac{\log(1/\delta)}{\epsilon} \leq \frac{\sqrt{n}}{3\sqrt{3}} - \frac{5}{3}, \tag{F.3}$$

then we can choose $\beta = \frac{7T}{3n^2\epsilon}$ and Algorithm 1 satisfies (ϵ, δ)-DP.

Proposition F.3 (Utility guarantee) *Let $\epsilon > 0$ and $\delta \in (0, 1)$. Assume for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is G-Lipschitz and F_S is ρ-weakly convex. Let $\{\mathbf{w}_t\}_t$ be produced by Algorithm 1 with $\eta_t = \eta$ and $A(S) = \mathbf{w}_r$. Assume $\mathbb{E}_{S,A}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2$ and $\mathbb{E}_A[\|A(S)\|_2 | n \in I(A)] \leq R$. Let Eq. (F.3) hold and $\beta = \frac{7T}{3n^2\epsilon}$. If we choose $\eta \asymp 1/((G + \sqrt{d}\sigma)\sqrt{\rho T})$ and $T \asymp (1 + d^{\frac{1}{3}} G^{\frac{2}{3}} \log^{\frac{2}{3}}(1/\delta)\epsilon^{-\frac{2}{3}}) n^{\frac{2}{3}} / (R^{\frac{2}{3}} \rho^{\frac{1}{3}})$, then*

$$\mathbb{E}\big[\|\nabla F_{1/(2\rho)}(\mathbf{w}_r)\|_2\big] = O(\sqrt{G} R^{\frac{1}{6}} \rho^{\frac{1}{3}} (1 + d^{\frac{1}{6}} G^{\frac{1}{3}} \log^{\frac{1}{3}}(1/\delta)\epsilon^{-\frac{1}{3}}) n^{-\frac{1}{6}}). \tag{F.4}$$

F.2. Proofs on Differentially Private SGD

In this section, we prove privacy and utility guarantee for DP-SGD. To this aim, we first study the Rényi differential privacy (Mironov, 2017), and then transform it to (ϵ, δ)-DP.

Definition F.4 *For $\lambda > 1, \rho > 0$, a randomized mechanism $\mathcal{A}$ satisfies (λ, ρ)-Rényi differential privacy (RDP) if for all neighboring datasets S and S' we have*

$$D_\lambda(\mathcal{A}(S) \| \mathcal{A}(S')) := \frac{1}{\lambda - 1} \log \int \Big(\frac{P_{\mathcal{A}(S)}(\mathbf{w})}{P_{\mathcal{A}(S')}(\mathbf{w})} \Big)^\lambda dP_{\mathcal{A}(S')}(\mathbf{w}) \leq \rho,$$

where $P_{\mathcal{A}(S)}(\mathbf{w})$ and $P_{\mathcal{A}(S')}(\mathbf{w})$ are the density of $\mathcal{A}(S)$ and $\mathcal{A}(S')$, respectively.

We first introduce some necessary lemmas. The following lemma establishes the RDP of a Gaussian mechanism together with subsampling (Liang et al., 2020).

Lemma F.5 (Liang et al. 2020) *Consider a mechanism $\mathcal{M} : \mathcal{Z}^m \mapsto \mathbb{R}^d$ and let Δ be its ℓ_2-sensitivity, i.e., $\Delta = \sup_{S \sim S'} \|\mathcal{M}(S) - \mathcal{M}(S')\|_2$. The Gaussian mechanism $\mathcal{A} = \mathcal{M} + \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ applied to a subset of samples that are drawn uniformly without replacement with subsampling rate p satisfies $(\lambda, 3.5p^2\lambda\Delta^2/\sigma^2)$-RDP if*

$$\sigma^2 \geq 0.67\Delta^2 \quad \text{and} \quad \lambda - 1 \leq \frac{2\sigma^2}{3\Delta^2} \log\Big(\frac{1}{\lambda p (1 + \sigma^2/\Delta^2)} \Big).$$

The following lemma shows the RDP of an adaptive composition of several mechanisms.

Lemma F.6 (Mironov 2017) *If $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are randomized algorithms satisfying, respectively, (α, ϵ_1)-RDP,$\ldots$,(α, ϵ_k)-RDP, then their composition defined as $(\mathcal{A}_1(S), \ldots, \mathcal{A}_k(S))$ is $(\alpha, \epsilon_1 + \ldots, +\epsilon_k)$-RDP. Moreover, the ith algorithm can be chosen on the basis of the outputs of $\mathcal{A}_1, \ldots, \mathcal{A}_{i-1}$.*

The following lemma shows the connection between DP and RDP.

Lemma F.7 (Mironov 2017) *If a randomized mechanism $\mathcal{A}$ satisfies (λ, ρ)-RDP, then $\mathcal{A}$ satisfies $(\rho + \log(1/\delta)/(\lambda - 1), \delta)$-DP for all $\delta \in (0, 1)$.*

We are now ready to prove the privacy and utility guarantee.

Proof of Proposition F.2 Consider the mechanism $\mathcal{A}_t = \mathcal{M}_t + b_t$, where $\mathcal{M}_t(z) = \nabla f(\mathbf{w}_t; z)$. Since f is G-Lipschitz continuous, we know

$$\sup_{z,z'} \|\nabla f(\mathbf{w}_t; z) - \nabla f(\mathbf{w}_t; z')\|_2 \leq \Delta := 2G$$

and therefore the ℓ_2 sensitivity of $\mathcal{M}_t$ is $2G$. Note

$$\frac{\sigma^2}{\Delta^2} = \frac{14G^2T}{\beta n^2 \epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big)\frac{1}{4G^2} = \frac{7T}{2\beta n^2 \epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big).$$

According to Lemma F.5, we know $\mathcal{M}_t$ satisfies $\Big(\lambda, \frac{\lambda\beta\epsilon}{T\left(\frac{\log(1/\delta)}{(\beta-1)\epsilon}+1\right)}\Big)$-RDP if

$$\frac{7T}{2\beta n^2 \epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big) \geq 0.67 \tag{F.5}$$

and

$$\lambda - 1 \leq \frac{7T}{3\beta n^2 \epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big)\log\Big(\frac{n}{\lambda\big(1 + \frac{7T}{2\beta n^2\epsilon}\big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\big)\big)}\Big).$$

Let $\lambda = \frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1$. Then the above inequality becomes

$$\frac{\log(1/\delta)}{(1-\beta)\epsilon} \leq \frac{7T}{3\beta n^2 \epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big)\log\Big(\frac{n}{\big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\big)\big(1 + \frac{7T}{2\beta n^2\epsilon}\big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\big)\big)}\Big). \tag{F.6}$$

We first suppose Eq. (F.5), (F.6) hold and prove the stated bound under these conditions. With our definition of λ, we know $\mathcal{M}_t$ satisfies $\big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1, \frac{\beta\epsilon}{T}\big)$-RDP for any $t \in [T]$. By the adaptive composition (Lemma F.6), we know Algorithm 1 satisfies $\big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1, \beta\epsilon\big)$-RDP. It then follows from Lemma F.7 that Algorithm 1 satisfies (ϵ, δ)-DP. We now show that Eq. (F.5) and Eq. (F.6) hold. Since $\epsilon \geq 14T/(3n^2)$ we know $\beta \leq 1/2$. It is clear

$$\frac{7T}{3\beta n^2 \epsilon} = \frac{7T \cdot 3n^2\epsilon}{21Tn^2\epsilon} = 1 \geq 0.67. \tag{F.7}$$

Therefore, Eq. (F.5) holds. Furthermore, the assumption $\frac{\log(1/\delta)}{\epsilon} \leq \frac{\sqrt{n}}{3\sqrt{3}} - \frac{5}{3}$ implies $1 + \frac{3}{2}\big(\frac{2\log(1/\delta)}{\epsilon} + 1\big) \leq \frac{\sqrt{n}}{\sqrt{3}}$. It then follows that

$$\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big)\Big(1 + \frac{7T}{2\beta n^2\epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big)\Big) \leq \Big(\frac{2\log(1/\delta)}{\epsilon} + 1\Big)\Big(1 + \frac{7T}{2\beta n^2\epsilon}\Big(\frac{2\log(1/\delta)}{\epsilon} + 1\Big)\Big)$$
$$= \Big(\frac{2\log(1/\delta)}{\epsilon} + 1\Big)\Big(1 + \frac{3}{2}\Big(\frac{2\log(1/\delta)}{\epsilon} + 1\Big)\Big) \leq \Big(1 + \frac{3}{2}\Big(\frac{2\log(1/\delta)}{\epsilon} + 1\Big)\Big)^2 \leq \frac{n}{3}.$$

We can combine the above inequality and Eq. (F.7) to show Eq. (F.6). The proof is completed. ■

Proof of Proposition F.3 Analogous to Lemma D.1, Part (d), we have the following optimization error bounds for Algorithm 1

$$\mathbb{E}_A\big[\|\nabla F_{S,1/(2\rho)}(\mathbf{w}_r)\|_2^2\big] = O\Big(\frac{\rho T(G^2 + \sigma^2 d)\eta^2 + 1}{T\eta}\Big)$$

and therefore

$$\mathbb{E}_A\big[\|\nabla F_{S,1/(2\rho)}(\mathbf{w}_r)\|_2\big] = O\big((G + \sigma\sqrt{d})\sqrt{\rho\eta} + 1/\sqrt{T\eta}\big).$$

Adding noise does not affect the stability analysis (Bassily et al., 2020), we then use Eq. (D.3) to get

$$\mathbb{E}\big[\big\|\nabla F_{1/(2\rho)}(\mathbf{w}_r)\big\|_2\big] = O\Big((G + \sigma\sqrt{d})\sqrt{\rho\eta} + \sqrt{GR\rho T/n} + 1/\sqrt{T\eta} + G/\sqrt{n}\Big).$$

For our choice of β, we have

$$\sigma^2 = \frac{14G^2T \cdot 3n^2\epsilon}{7Tn^2\epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big) \leq 6G^2\Big(\frac{2\log(1/\delta)}{\epsilon} + 1\Big), \tag{F.8}$$

where we have used $\beta \leq 1/2$ established in the proof of Proposition F.2. If we choose $\eta \asymp 1/((G + \sqrt{d}\sigma)\sqrt{\rho T})$ and use Eq. (F.8), we get

$$\mathbb{E}\big[\|\nabla F_{1/(2\rho)}(\mathbf{w}_r)\|_2\big] = O\Big(\sqrt{G + \sqrt{d}\sigma}(\rho/T)^{\frac{1}{4}} + \sqrt{GR\rho T/n} + G/\sqrt{n}\Big).$$

We can choose $T \asymp (1 + d^{\frac{1}{3}}G^{\frac{2}{3}}\log^{\frac{2}{3}}(1/\delta)\epsilon^{-\frac{2}{3}})n^{\frac{2}{3}}/(R^{\frac{2}{3}}\rho^{\frac{1}{3}})$ to get Eq. (F.4). The proof is completed. ■

Appendix G. Convergence Rates with Relaxed Bounded Gradient Assumptions

In this section, we study the convergence rates of SGD for solving weakly convex problems. The existing convergence analysis requires a bounded subgradient assumption as $\mathbb{E}_{i_t}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq G^2$ for some $G > 0$ (Davis and Drusvyatskiy, 2019). We aim to relax this assumption to a more general assumption as

$$\mathbb{E}_{i_t}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq B_1\mathbb{E}_{i_t}[f(\mathbf{w}_t; z_{i_t})] + B_2, \tag{G.1}$$

where $B_1, B_2 \geq 0$ are two constants. This assumption implies that the gradients can be bounded in terms of function values, which has been considered in the literature (Zhang, 2004).

Theorem G.1 *Let $\mathcal{W} = \mathbb{R}^d$ and $\sum_{t=1}^T \eta_t^2 = O(1)$. Let $\{\mathbf{w}_t\}_t$ be produced by the algorithm A defined by SGD and Eq. (G.1) holds for all $t \in \mathbb{N}$. If F_S is ρ-weakly convex, then*

$$\sum_{t=1}^T \eta_t \mathbb{E}_A[\|\nabla F_{S,1/2\rho}(\mathbf{w}_t)\|_2^2] = O\Big(1 + \sum_{t=1}^T \eta_t^2\Big). \tag{G.2}$$

Proof For any $t \in \mathbb{N}$, denote $\hat{\mathbf{w}}_t = \text{Prox}_{F_S,1/(2\rho)}(\mathbf{w}_t)$. According to the definition of Moreau envelope and the definition of $\hat{\mathbf{w}}_t$, we know

$$\begin{aligned}
&\mathbb{E}_{i_t}[F_{S,1/2\rho}(\mathbf{w}_{t+1})] \le \mathbb{E}_{i_t}\big[F_S(\hat{\mathbf{w}}_t) + \rho\|\hat{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2^2\big] \\
&= F_S(\hat{\mathbf{w}}_t) + \rho\mathbb{E}_{i_t}\big[\|\hat{\mathbf{w}}_t - \mathbf{w}_t + \eta_t \nabla f(\mathbf{w}_t; z_{i_t})\|_2^2\big] \\
&= F_S(\hat{\mathbf{w}}_t) + \rho\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 + 2\rho\eta_t \mathbb{E}_{i_t}\big[\langle \hat{\mathbf{w}}_t - \mathbf{w}_t, \nabla f(\mathbf{w}_t; z_{i_t})\rangle\big] + \rho\eta_t^2 \mathbb{E}_{i_t}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \\
&\le F_{S,1/(2\rho)}(\mathbf{w}_t) + 2\rho\eta_t\langle \hat{\mathbf{w}}_t - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle + \rho\eta_t^2 \mathbb{E}_{i_t}\big[B_1 f(\mathbf{w}_t; z_{i_t}) + B_2\big] \\
&\le F_{S,1/(2\rho)}(\mathbf{w}_t) + 2\rho\eta_t\big(F_S(\hat{\mathbf{w}}_t) - F_S(\mathbf{w}_t) + \frac{\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2\big) + \rho\eta_t^2\big[B_1 F_S(\mathbf{w}_t) + B_2\big],
\end{aligned} \tag{G.3}$$

where in the last second step we have used Eq. (G.1) and in the last inequality we have used the weak convexity of F_S. By the weak convexity of F_S, we know the function $\mathbf{w} \mapsto F_S(\mathbf{w}) + \rho\|\mathbf{w} - \mathbf{v}\|_2^2$ is ρ-strongly convex. This together with the definition of $\hat{\mathbf{w}}_t$ implies

$$\begin{aligned}
&F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}_t) - \frac{\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2 \\
&= \big(F_S(\mathbf{w}_t) + \rho\|\mathbf{w}_t - \mathbf{w}_t\|_2^2\big) - \big(F_S(\hat{\mathbf{w}}_t) + \rho\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2\big) + \frac{\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2 \\
&\ge \rho\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2.
\end{aligned}$$

It then follows that

$$F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}_t) \ge \frac{3\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2 \ge 0. \tag{G.4}$$

This together with the assumption $\eta_t \le 1/B_1$ implies

$$\begin{aligned}
B_1\eta_t^2 F_S(\mathbf{w}_t) &= B_1\eta_t^2\big(F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}_t)\big) + B_1\eta_t^2 F_S(\hat{\mathbf{w}}_t) \\
&\le \eta_t\big(F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}_t)\big) + B_1\eta_t^2 F_S(\hat{\mathbf{w}}_t).
\end{aligned}$$

We can plug the above inequality back into Eq. (G.3) and derive

$$\begin{aligned}
&\mathbb{E}_{i_t}[F_{S,1/2\rho}(\mathbf{w}_{t+1})] \\
&\le F_{S,1/(2\rho)}(\mathbf{w}_t) + \rho^2\eta_t\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2 + \big(2\rho\eta_t - \rho\eta_t\big)\big(F_S(\hat{\mathbf{w}}_t) - F_S(\mathbf{w}_t)\big) + \rho\eta_t^2\big(B_1 F_S(\hat{\mathbf{w}}_t) + B_2\big) \\
&= F_{S,1/(2\rho)}(\mathbf{w}_t) + \rho^2\eta_t\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2 + \rho\eta_t\big(F_S(\hat{\mathbf{w}}_t) - F_S(\mathbf{w}_t)\big) + \rho\eta_t^2\big(B_1 F_S(\hat{\mathbf{w}}_t) + B_2\big) \\
&\le F_{S,1/(2\rho)}(\mathbf{w}_t) - \frac{\rho^2\eta_t\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2}{2} + \rho\eta_t^2\big(B_1 F_{S,1/(2\rho)}(\mathbf{w}_t) + B_2\big),
\end{aligned} \tag{G.5}$$

where we have used Eq. (G.4) and the following inequality in the last step

$$F_{S,1/(2\rho)}(\mathbf{w}_t) = \inf_{\mathbf{v}}\big\{F_S(\mathbf{v}) + \rho\|\mathbf{v} - \mathbf{w}_t\|_2^2\big\} = F_S(\hat{\mathbf{w}}_t) + \rho\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 \ge F_S(\hat{\mathbf{w}}_t).$$

It then follows from Eq. (G.5) that

$$\mathbb{E}_A\big[F_{S,1/2\rho}(\mathbf{w}_{t+1})\big] \leq \big(1+\rho B_1\eta_t^2\big)\mathbb{E}_A\big[F_{S,1/2\rho}(\mathbf{w}_t)\big] + \rho\eta_t^2 B_2$$

and therefore ($1+a \leq \exp(a)$)

$$\begin{aligned}
\mathbb{E}_A\big[F_{S,1/2\rho}(\mathbf{w}_{t+1})\big] &\leq \prod_{k=1}^{t}\big(1+\rho B_1\eta_k^2\big)F_{S,1/2\rho}(\mathbf{w}_1) + \rho B_2\sum_{k=1}^{t}\eta_k^2\prod_{\tilde{k}=k+1}^{t}\big(1+\rho B_1\eta_{\tilde{k}}^2\big) \\
&\leq \prod_{k=1}^{t}\exp(\rho B_1\eta_k^2)F_{S,1/2\rho}(\mathbf{w}_1) + \rho B_2\sum_{k=1}^{t}\eta_k^2\prod_{\tilde{k}=k+1}^{t}\exp(\rho B_1\eta_{\tilde{k}}^2) \\
&= \exp\Big(\rho B_1\sum_{k=1}^{t}\eta_k^2\Big)F_{S,1/2\rho}(\mathbf{w}_1) + \rho B_2\sum_{k=1}^{t}\eta_k^2\exp\Big(\rho B_1\sum_{\tilde{k}=k+1}^{t}\eta_{\tilde{k}}^2\Big).
\end{aligned}$$

Since $\sum_{t=1}^{T}\eta_t^2 = O(1)$, we further get

$$\mathbb{E}_A\big[F_{S,1/2\rho}(\mathbf{w}_t)\big] = O(1), \qquad \forall t\in[T]. \tag{G.6}$$

We can plug the above inequality back into Eq. (G.5) and get

$$\mathbb{E}_A[F_{S,1/2\rho}(\mathbf{w}_{t+1})] = \mathbb{E}_A[F_{S,1/2\rho}(\mathbf{w}_t)] - \frac{\rho^2\eta_t\mathbb{E}_A[\|\mathbf{w}_t-\hat{\mathbf{w}}_t\|_2^2]}{2} + O(\eta_t^2).$$

The above inequality can be reformulated as

$$\frac{\rho^2\eta_t\mathbb{E}_A[\|\mathbf{w}_t-\hat{\mathbf{w}}_t\|_2^2]}{2} = \mathbb{E}_A[F_{S,1/2\rho}(\mathbf{w}_t)] - \mathbb{E}_A[F_{S,1/2\rho}(\mathbf{w}_{t+1})] + O(\eta_t^2).$$

We can take a summation of the above inequality from $t=1$ to $t=T$ and get

$$\frac{\rho^2}{2}\sum_{t=1}^{T}\eta_t\mathbb{E}_A[\|\mathbf{w}_t-\hat{\mathbf{w}}_t\|_2^2] = O\Big(1+\sum_{t=1}^{T}\eta_t^2\Big).$$

According to the definition of $\hat{\mathbf{w}}_t$, we know $\nabla F_{S,1/2\rho}(\mathbf{w}_t) = 2\rho(\mathbf{w}_t-\hat{\mathbf{w}}_t)$. It then follows that

$$\sum_{t=1}^{T}\eta_t\mathbb{E}_A[\|\nabla F_{S,1/2\rho}(\mathbf{w}_t)\|_2^2] = O\Big(1+\sum_{t=1}^{T}\eta_t^2\Big).$$

This gives the bound (G.2). The proof is completed. ■