L_2-Uniform Stability of Randomized Learning Algorithms: Sharper Generalization Bounds and Confidence Boosting

Xiao-Tong Yuan
School of Intelligence Science and Technology
Nanjing University
Suzhou, 215163, China
xtyuan1980@gmail.com

Ping Li
VecML Inc. www.vecml.com
Bellevue, WA 98004, USA
pingli98@gmail.com

Abstract

Exponential generalization bounds with near-optimal rates have recently been established for uniformly stable algorithms (Feldman and Vondrák, 2019; Bousquet et al., 2020). We seek to extend these best known high probability bounds from deterministic learning algorithms to the regime of randomized learning. One simple approach for achieving this goal is to define the stability for the expectation over the algorithm's randomness, which may result in sharper parameter but only leads to guarantees regarding the on-average generalization error. Another natural option is to consider the stability conditioned on the algorithm's randomness, which is way more stringent but may lead to generalization with high probability jointly over the randomness of sample and algorithm. The present paper addresses such a tension between these two alternatives and makes progress towards relaxing it inside a classic framework of confidence-boosting. To this end, we first introduce a novel concept of L_2-uniform stability that holds uniformly over data but in second-moment over the algorithm's randomness. Then as a core contribution of this work, we prove a strong exponential bound on the first-moment of generalization error under the notion of L_2-uniform stability. As an interesting consequence of the bound, we show that a bagging-based meta algorithm leads to near-optimal generalization with high probability jointly over the randomness of data and algorithm. We further substantialize these generic results to stochastic gradient descent (SGD) to derive sharper exponential bounds for convex or non-convex optimization with natural time-decaying learning rates, which have not been possible to prove with the existing stability-based generalization guarantees.

1 Introduction

In many statistical learning problems, we are interested in designing a randomized algorithm $A : \mathcal{Z}^N \times \mathcal{R} \mapsto \mathcal{W}$ that maps a training data sample $S = \{Z_i\}_{i\in[N]} \in \mathcal{Z}^N$ with an algorithm's random parameter $\xi \in \mathcal{R}$ to a model $A(S, \xi) \in \mathcal{W}$. Here $\mathcal{Z}$ and $\mathcal{R}$ are some measurable sets, and $\mathcal{W}$ is a closed subset of an Euclidean space. The ultimate goal is to find a suitable algorithm such that the following population risk evaluated at the model should be as small as possible:

$$R(A(S, \xi)) := \mathbb{E}_Z[\ell(A(S, \xi); Z)],$$

where $Z \in \mathcal{Z}$ and $\ell : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}^+$ is a non-negative bounded loss function whose value $\ell(w; z)$ measures the loss evaluated at z with parameter w. It is generally the case that the underlying data distribution is unknown, and in this case the data points Z_i are usually assumed to be independent.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Then, a natural alternative measurement that mimics the computationally intractable population risk is the empirical risk given by

$$R_S(A(S,\xi)) := \mathbb{E}_{Z\sim\mathtt{Unif}(S)}[\ell(A(S,\xi);Z)] = \frac{1}{N}\sum_{i=1}^{N}\ell(A(S,\xi);Z_i).$$

The bound on the difference between the population and empirical risks is of central interest in understanding the generalization performance of a learning algorithm. In particular, we hope to derive a suitable law of large numbers, i.e., a sample size vanishing rate b_N such that the generalization bound $|R_S(A(S,\xi)) - R(A(S,\xi))| \lesssim b_N$ holds with high probability over the randomness of S and hopefully the randomness of ξ as well. Let $R^* := \min_{w\in\mathcal{W}} R(w)$ be the optimal value of the population risk. Conditioned on S, suppose that $A(S,\xi)$ is an almost minimizer of the empirical risk R_S such that $R_S(A(S,\xi)) - \min_{w\in\mathcal{W}} R_S(w) \le \varepsilon$, then the generalization bound immediately implies an *excess risk* bound $R(A(S,\xi)) - R^* \lesssim b_N + \frac{1}{\sqrt{N}} + \varepsilon$ based on the standard risk decomposition and Hoeffding's inequality. Therefore, generalization guarantees also play a crucial role in understanding the stochastic optimization performance of a learning algorithm.

A powerful proxy for analyzing the generalization bounds is the *stability* of learning algorithms to changes in the training dataset. Since the seminal work of Bousquet and Elisseeff (2002), stability has been extensively demonstrated to beget dimension-independent generalization bounds for deterministic learning algorithms (Mukherjee et al., 2006; Shalev-Shwartz et al., 2010), as well as for randomized learning algorithms such as bagging and SGD (Elisseeff et al., 2005; Hardt et al., 2016). So far, the best known results about generalization bounds are offered by approaches based on the notion of uniform stability (Feldman and Vondrák, 2018, 2019; Bousquet et al., 2020; Klochkov and Zhivotovskiy, 2021) which is independent to the underlying distribution of data. For randomized algorithms, the definition of uniform stability can be extended in two natural ways by respectively considering 1) the stability averaged over the algorithm's randomness (Hardt et al., 2016) and 2) the stability conditioned on the algorithm's randomness (Feldman and Vondrák, 2019). The former is simpler to show but typically leads to on-average generalization bounds, while the latter is relatively more stringent but may yield deviation bounds given that the conditional stability holds with high probability over the algorithm's randomness. Between these two extreme cases, however, the generalization behavior of randomized learning algorithm still remains largely under explored.

To address the above mentioned theoretical gap between the current lines of results, we explore the opportunities of *deriving exponential generalization bounds for randomized learning algorithms beyond the notions of on-average stability and conditional stability*. A concrete working example of our study is the widely used stochastic gradient descent (SGD) algorithm that carries out the following recursion for all $t \ge 1$ with learning rate $\eta_t > 0$:

$$w_t := \Pi_{\mathcal{W}}\left(w_{t-1} - \eta_t \nabla_w \ell(w_{t-1}; Z_{i_t})\right), \tag{1}$$

where $i_t \in [N]$ is a random index of data under with or without replacement sampling, and $\Pi_{\mathcal{W}}$ is the Euclidean projection operator associated with $\mathcal{W}$. The in-expectation generalization of SGD has been studied under on-average stability (Hardt et al., 2016; Zhou et al., 2022; Lei and Ying, 2020), while the exponential bounds have recently been established given that the stability holds with high probability over the sampling path of SGD (Feldman and Vondrák, 2019; Bassily et al., 2020).

1.1 Prior results

Let us start by briefly reviewing some state-of-the-art exponential generalization bounds under the notion of uniform stability and its randomized variants. We denote by $S \doteq \tilde{S}$ if a pair of data sets S and $\tilde{S}$ differ in a single element. A randomized learning algorithm A is said to have *on-average γ_N-uniform stability* (Elisseeff et al., 2005) if it satisfies the following uniform bound:

$$\sup_{S\doteq\tilde{S}, Z\in\mathcal{Z}} \left|\mathbb{E}_\xi\left[\ell(A(S,\xi);Z) - \ell(A(\tilde{S},\xi);Z)\right]\right| \le \gamma_N. \tag{2}$$

This definition is equivalent to the concept of uniform stability defined for the expectation of loss $\mathbb{E}_\xi[\ell(A(S,\xi);Z)]$. Suppose that the loss function is bounded in the interval $[0,M]$. Then essentially it has been shown in Feldman and Vondrák (2019) that for any $\delta \in (0,1)$, with probability at least $1-\delta$ over S, the on-average generalization error is upper bounded by

$$|\mathbb{E}_\xi\left[R(A(S,\xi)) - R_S(A(S,\xi))\right]| \lesssim \gamma_N \log(N)\log\left(\frac{N}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{N}}. \tag{3}$$

Bousquet et al. (2020) later derived a slightly improved exponential bound that implies

$$|\mathbb{E}_\xi [R(A(S,\xi)) - R_S(A(S,\xi))]| \lesssim \gamma_N \log(N) \log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{N}}. \tag{4}$$

These bounds are near-tight (up to logarithmic factors) in the sense of an $\mathcal{O}\big(\gamma_N \log\left(\frac{1}{\delta}\right) + \sqrt{\frac{\log(1/\delta)}{N}}\big)$ lower deviation bound on sum of random functions with γ_N-uniform stability (Bousquet et al., 2020, Proposition 9). Concerning the excess risk bound, Klochkov and Zhivotovskiy (2021) essentially derived the following result using the sample-splitting techniques of Bousquet et al. (2020):

$$\mathbb{E}_\xi [R(A(S,\xi))] - R^* \lesssim \Delta_{\text{opt}} + \mathbb{E}[\Delta_{\text{opt}}] + \gamma_N \log(N)\log\left(\frac{1}{\delta}\right) + \frac{(M+B)\log(1/\delta)}{N}, \tag{5}$$

where $\Delta_{\text{opt}} := \mathbb{E}_\xi [R_S(A(S,\xi))] - \min_{w\in\mathcal{W}} R_S(w)$ represents the in-expectation empirical risk sub-optimality, and B is the constant of the generalized Bernstein condition (Koltchinskii, 2006). While sharp in the dependence on sample size, one common limitation of the above uniform stability implied generalization and risk bounds lies in that these high-probability results only hold *in expectation* with respect to ξ, the internal randomness of algorithm.

Alternatively, consider that A has γ_N-uniform stability with probability at least $1-\delta'$ for some $\delta' \in (0,1)$ over the random draw of ξ, i.e.,

$$\mathbb{P}\left\{\sup_{S \doteq \tilde{S}, Z\in\mathcal{Z}} |\ell(A(S,\xi);Z) - \ell(A(\tilde{S},\xi);Z)| \le \gamma_N\right\} \ge 1-\delta'. \tag{6}$$

Suppose that the randomness of A is independent of the training set S. Then the bound of Bousquet et al. (2020) naturally implies that with probability at least $1-\delta-\delta'$ over S and ξ,

$$|R(A(S,\xi)) - R_S(A(S,\xi))| \lesssim \gamma_N \log(N)\log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{N}}. \tag{7}$$

This is by far the best known generalization bound of randomized stable algorithms that hold with high probability jointly over the randomness of data and algorithm. The result, however, relies heavily on the high-probability uniform stability as expressed in (6). For the SGD recursion (1) with fixed learning rate $\eta_t \equiv \eta$, it is possible to show that $\gamma_N \lesssim \eta\sqrt{T} + \frac{\eta T}{N}$ and $\delta' = N\exp(-\frac{N}{2})$ in (6) (Bassily et al., 2020). For SGD with time decaying learning rates, which has been widely studied in theory (Harvey et al., 2019; Rakhlin et al., 2012) and applied in practice for training popular deep nets such as ResNet and DenseNet (Bengio et al., 2017), it is not clear if the condition in (6) is still valid for γ_N and δ' of interest. Madden et al. (2020) indeed have established a high-probability uniform stability bound for minibatch SGD with learning rates $\eta_t \lesssim \frac{1}{Nt}$. However, such a fairly conservative choice of learning rates tends to impair the empirical minimization performance of SGD and thus is of limited interest from the perspective of risk minimization.

More specially for randomized learning methods such as bagging (Breiman, 1996) and SGD, the randomness of algorithm can be precisely characterized by a vector of i.i.d. parameters $\xi = \{i_1, ..., i_t\}$ which are independent on data S. In such cases, assume additionally that $A(S,\xi)$ has uniform stability with respect to ξ conditioned on S, i.e., $\sup_{\xi \doteq \tilde{\xi}} |\ell(A(S,\xi)) - \ell(A(S,\tilde{\xi}))| \le \rho_T$. Then the following exponential bound has been derived by Elisseeff et al. (2005):

$$|R(A(S)) - R_S(A(S))| \lesssim \gamma_N + \left(\frac{1+N\gamma_N}{\sqrt{N}} + \sqrt{T}\rho_T\right)\sqrt{\log\left(\frac{1}{\delta}\right)}. \tag{8}$$

Provided that $\gamma_N \lesssim \frac{1}{N}$ and $\rho_T \lesssim \frac{1}{T}$, the above bound shows that the generalization bound scales as $\mathcal{O}\big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{T}}\big)$ with high probability. However, the rate of the above bound is sub-optimal and will show no guarantee on convergence if $\gamma_N \gtrsim \frac{1}{\sqrt{N}}$ and/or $\rho_T \gtrsim \frac{1}{\sqrt{T}}$. As an example, for non-convex SGD with learning rate $\eta_t = O\big(\frac{1}{t}\big)$, it can be shown that $\gamma_N \lesssim \frac{\sqrt{T}}{N}$ and ρ_T scales as large as $\mathcal{O}(1)$.

Open problem. So far, it still remains open if the exponential generalization bounds for deterministic uniformly stable algorithms might be extended to randomized learning algorithms under the variants of uniform stability tighter than the on-average version (2) but less restrictive than the high-probability version (6). Particularly, we are interested in the following notion of L_2*-uniform stability* (as formally introduced in Definition 1) with parameter $\gamma_{\mathrm{L}_2,N}$:

$$\sup_{S\doteq\tilde{S}, Z\in\mathcal{Z}} \mathbb{E}_\xi\left[\left(\ell(A(S,\xi);Z)-\ell(A(\tilde{S},\xi);Z)\right)^2\right] \le \gamma^2_{\mathrm{L}_2,N}, \tag{9}$$

which represents a second-moment variant of the uniform stability for randomized learning algorithms. For example, as we will shortly show in Section 4 that SGD with practical time-decaying learning rates has L_2-uniform stability with favorable parameters. The main goal of the present work is to derive sharper exponential generalization bounds for randomized learning algorithms under the notion of L_2-uniform stability.

1.2 Overview of our contribution

The fundamental contribution of this work is a near-optimal first-moment generalization error bound for L_2-uniformly stable algorithms, which is summarized in Theorem 1 and highlighted below:

$$\mathbb{E}_\xi\left[|R(A(S,\xi))-R_S(A(S,\xi))|\right] \lesssim \gamma_{\mathrm{L}_2,N}\log(N)\log\left(\frac{1}{\delta}\right)+M\sqrt{\frac{\log(1/\delta)}{N}}.$$

While our first-moment bound above has an identical convergence rate to that of the on-average bound in (4), the former is stronger in the sense that the expectation is taken outside the generalization gap and thus implies the latter where the expectation is taken inside. The key ingredients of our analysis are a set of fine-grained concentration inequalities for randomized functions (Proposition 1) and sums of randomized functions (Proposition 2), which respectively generalize the classic bounded-difference inequalities and a prior result of Bousquet et al. (2020) under the considered L_2-uniform bounded difference conditions. These generalized concentration inequalities and their proof arguments are novel to our knowledge and should be of independent interests in analyzing randomized functions.

As an important consequence of our main result, we reveal that a bagging-based meta procedure (see Algorithm 1) can be used to boost the confidence of generalization for L_2-uniformly stable algorithms. More specifically, in the presented bagging procedure we independently run a randomized algorithm A multiple K times over a fraction of the training set to obtain K solutions. Then we evaluate the validation error of these candidate solutions over a holdout training subset, and output the solution that has the smallest training-validation gap. Our result in Theorem 2 shows that for any confidence level $\delta\in(0,1)$, setting $K\asymp\log(\frac{1}{\delta})$ yields a near-optimal generalization bound for the selected solution that holds with high probability jointly over the randomness of data and algorithm.

We have substantialized our results to SGD with smooth (Corollary 1) or non-smooth (Corollary 2) convex losses, and smooth non-convex losses (Corollary 3) as well. For an instance, our result in Corollary 1 shows that when invoked to SGD with smooth convex loss and learning rates $\eta_t=\mathcal{O}(\frac{1}{\sqrt{t}})$, the generalization bound of the output of Algorithm 1 may scale as $\mathcal{O}\big(\log(N)\log\left(\frac{1}{\delta}\right)\sqrt{\frac{\log(T)}{N}+\frac{\sqrt{T}}{N}}\big)$. To compare with the $\mathcal{O}\big(\frac{\sqrt{T}}{N}\big)$ in-expectation bound of smooth convex SGD (Hardt et al., 2016), our bound above for the boosted SGD is comparable in convergence rate while it holds with high probability jointly over the randomness of data and sampling path.

2 L_2-Uniform Stability and Generalization

2.1 Notation and definitions

Let us introduce some notation to be used in our analysis. We abbreviate $[N]:=\{1,...,N\}$. Recall that $S=\{Z_i\}_{i\in[N]}$ is a set of i.i.d. training data. Denote by $S'=\{Z'_i\}_{i\in[N]}$ an independent copy of S and we write $S^{(i)}=\{Z_1,...,Z_{i-1},Z'_i,Z_{i+1},...,Z_N\}$. For a real-valued random variable Y, its L_q-norm for $q\ge1$ is given by $\|Y\|_q=(\mathbb{E}[|Y|^q])^{1/q}$. By definition it can be verified that $\forall q\ge2$,

$$\|Y\|_q^2=(\mathbb{E}[|Y|^q])^{2/q}=\left(\mathbb{E}[|Y^2|^{q/2}]\right)^{2/q}=\left\|Y^2\right\|_{q/2}. \tag{10}$$

Let $h:\mathcal{Z}^N\mapsto\mathbb{R}$ be some measurable function and consider the random variable $h(S)=h(Z_1,...,Z_N)$. For $h(S)$ and any index set $I\subseteq[N]$, we define the following abbreviations:

$$h(S_I):=\mathbb{E}\left[h(S)\mid S_I\right],\quad \|h\|_q(S_I):=\left(\mathbb{E}\left[|h(S)|^q\mid S_I\right]\right)^{1/q}.$$

We say a function f to be G-Lipschitz continuous over $\mathcal{W}$ if $|f(w) - f(\tilde{w})| \le G\|w - \tilde{w}\|$ for all $w, \tilde{w} \in \mathcal{W}$, and it is L-smooth if $\|\nabla f(w) - \nabla f(\tilde{w})\| \le L\|w - \tilde{w}\|$. For a pair of functions $f, f' \ge 0$, we use $f \lesssim f'$ (or $f' \gtrsim f$) to denote $f \le cf'$ for some universal constant $c > 0$.

In the following definition, we formally introduce the concept of L_2-uniform stability for randomized learning algorithms to be investigated in this work.

Definition 1 (L_2-Uniform stability of randomized learning algorithms)**.** *We say a randomized learning algorithm $A : \mathcal{Z}^N \times \mathcal{R} \mapsto \mathcal{W}$ to have L_2-uniform stability with parameter $\gamma_{L_2,N} \ge 0$ if*

$$\sup_{S, Z_i', Z} \mathbb{E}_\xi \left[\left(\ell(A(S,\xi); Z) - \ell(A(S^{(i)}, \xi); Z) \right)^2 \right] \le \gamma^2_{L_2,N}.$$

Remark 1. *By definition the L_2-uniform stability has a second-moment dependence on the internal randomness of algorithm conditioned on data, while it is invariant to the data distribution. This justifies the name of such a notion of mixed algorithmic stability.*

Remark 2. *On one hand, by Jensen's inequality the L_2-uniform stability implies the on-average uniform stability defined in (2). On the other hand, the second-order form of L_2-uniform stability is by definition weaker than the high-probability uniform stability in (6). If the algorithm's randomness ξ can be expressed as a set of i.i.d. random bits, then the L_2-uniform stability is also weaker than the conditional uniform stability conditioned on data S (Elisseeff et al., 2005).*

Throughout this paper, we assume for simplicity that the output models $A(S^{(i)}, \xi)$ and $A(S, \xi)$ share the same internal random bit ξ which is invariant to data. With similar analysis techniques, it is indeed possible to extend Definition 1 and our main results to the general setting where the randomness of algorithm is allowed to be dependent on data, such as in posterior sampling for Bayesian learning.

2.2 Concentration inequalities for randomized functions

We begin by establishing in the following result a group of first- and second-order concentration inequalities (in moments) for *randomized* functions of independent random variables.

Proposition 1. *Let $S = \{Z_1, Z_2, ..., Z_N\}$ be a set of independent random variables valued in $\mathcal{Z}$ and ξ be a random variable valued in $\mathcal{R}$. Let $g : \mathcal{Z}^N \times \mathcal{R} \mapsto \mathbb{R}$ be a measurable function that satisfies the following L_2-bounded-difference condition:*

$$\sup_{S, Z_i'} \mathbb{E}_\xi \left[\left(g(S,\xi) - g(S^{(i)}, \xi) \right)^2 \right] \le \beta^2.$$

Then for any $q \ge 2$,

$$\left\| \mathbb{E}_\xi \left[\left| g(S,\xi) - \mathbb{E}_S \left[g(S,\xi) \right] \right| \right] \right\|_q \le 3\beta\sqrt{Nq}, \tag{11}$$

and

$$\left\| \mathbb{E}_\xi \left[\left(g(S,\xi) - \mathbb{E}_S \left[g(S,\xi) \right] \right)^2 \right] \right\|_q \le 68N\beta^2 q. \tag{12}$$

Proof in sketch. Let us consider $h(S) := \mathbb{E}_\xi \left[|g(S,\xi) - \mathbb{E}_S[g(S,\xi)]| \right]$. The given L_2-bounded-difference condition implies that $h(S)$ has bounded-difference property. Then the desired first-order bound in (11) can be obtained by respectively invoking a moment Efron-Stein inequality (Boucheron et al., 2005, Theorem 2) to upper bound $\|h(S) - \mathbb{E}[h(S)]\|_q$ and a slightly modified Efron-Stein inequality to bound the mean $\mathbb{E}[h(S)]$. To prove the second-order concentration bound, we consider the function $h'(S) := \mathbb{E}_\xi \left[(g(S,\xi) - \mathbb{E}_S[g(S,\xi)])^2 \right]$, which can be shown to be *weakly self-bounding* (see Definition 2) under the L_2-bounded-difference condition. Then the desired bound (12) can be derived by applying the upper tail bound of Boucheron et al. (2005, Theorem 6.19) and lower tail bound of Klochkov and Zhivotovskiy (2021, Proposition 3.1) for weakly self-bounding functions. See Appendix A.2 for a detailed proof of this result. □

The moment bound in (11) extends the McDiarmid's (bounded difference) inequality (McDiarmid et al., 1989) to randomized functions with the L_2-bounded-difference property. The second-order concentration bound in (12) is crucial for proving the moment bound of sums in Proposition 2, as it can be used to sharply control some second-order components involved in the arguments. These generic inequalities are expected to be of independent interests for understanding the first-/second-order concentration behavior of randomized functions.

2.3 A moment inequality for sums of randomized functions

As a key intermediate result, we further establish in the following proposition a moment concentration inequality for sums of randomized functions that satisfy the L_2-bounded-difference condition. This result extends the moment bound for sums of functions (Bousquet et al., 2020, Theorem 4) to sums of randomized functions.

Proposition 2. *Let $S = \{Z_1, Z_2, ..., Z_N\}$ be a set of independent random variables valued in $\mathcal{Z}$ and ξ be a random variable valued in $\mathcal{R}$. Let $g_1, ..., g_N$ be a set of measurable functions $g_i : \mathcal{Z}^N \times \mathcal{R} \mapsto \mathbb{R}$ that satisfy the following conditions for any $i \in [N]$:*

- $\mathbb{E}\left[g_i(S, \xi) \mid S \setminus Z_i, \xi\right] = 0$ *and* $|\mathbb{E}[g_i(S, \xi) \mid Z_i, \xi]| \leq M$*, almost surely;*
- $g_i(S, \xi)$ *has the following L_2-bounded-difference property with respect to all variables in S except Z_i, i.e.,* $\forall j \neq i$,

$$\sup_{S, Z_j'} \mathbb{E}_\xi \left[\left(g_i(S, \xi) - g_i(S^{(j)}, \xi)\right)^2\right] \leq \beta^2.$$

Then for all $q \geq 2$,

$$\left\| \mathbb{E}_\xi \left[\left| \sum_{i=1}^N g_i(S, \xi) \right|\right] \right\|_q \leq 3M\sqrt{3Nq} + 38N \lceil \log_2 N \rceil \beta q.$$

Proof in sketch. The main idea is inspired by the sample-splitting arguments of Feldman and Vondrák (2019); Bousquet et al. (2020), with some new ingredients developed to handle the first-moment operator taken over the internal randomness of functions. Here we just highlight a fundamental difference, which arises from using a newly developed moment inequality (Lemma 9) for bounding the sums of *conditionally independent randomized functions* inside each individual data splits. Different from the version of Marcinkiewicz-Zygmund's inequality used in the original analysis of Bousquet et al. (2020), our new bound in Lemma 9 relies on some second-order (over the function's randomness) components which might be tightly bounded by the second-order concentration inequality in Proposition 1. A full proof is provided in Appendix A.3. □

Remark 3. *For sums of deterministic functions, our result in Proposition 2 reduces to the existing moment bound of Bousquet et al. (2020, Theorem 4) which is known to be near-tight up to logarithmic factors. We comment in passing that the tightness analysis of Bousquet et al. (2020, Proposition 9) for deterministic functions can be more or less straightforwardly extended to randomized functions.*

Remark 4. *The bound of Proposition 2 would still be valid when the bounded-loss condition $|\mathbb{E}[g_i(S, \xi) \mid Z_i, \xi]| \leq M$ is relaxed to certain sub-Gaussian or sub-exponential stochastic versions.*

2.4 Main result on generalization bound

Consequently from Proposition 2, we can now establish our main result on the generalization bound of L_2-uniformly stable randomized learning algorithms.

Theorem 1. *Let $A : \mathcal{Z}^N \times \mathcal{R} \mapsto \mathcal{W}$ be a randomized learning algorithm that has L_2-uniform stability with parameter $\gamma_{L_2,N}$. Assume that the loss function ℓ is valued in $[0, M]$. Then for any $\delta \in (0, 1)$, the following bound holds with probability at least $1 - \delta$ over the draw of S:*

$$\mathbb{E}_\xi \left[|R(A(S, \xi)) - R_S(A(S, \xi))|\right] \lesssim \gamma_{L_2,N} \log(N) \log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{N}}.$$

Proof. See Appendix A.4 for a proof of this result. □

Remark 5. *The first-moment bound in Theorem 1 naturally implies the on-average bound in (4) with an identical rate of convergence, though the former is obtained under the relatively stronger notion of L_2-uniform stability. As we will see shortly that the L_2-uniform stability can indeed be fulfilled by the popularly applied SGD algorithm and thus Theorem 1 is of practical importance for showcasing sharper generalization performance of SGD. When A is deterministic, our bound reduces to the near-optimal (up to logarithmic factors on sample size and failure tail) generalization bound for uniformly stable algorithms (Bousquet et al., 2020).*

Algorithm 1: Confidence-Boosting for Randomized Learning Algorithms

Input : Randomized learning algorithm A, data set $S = \{Z_i\}_{i\in[N]}$, $\mu \in (0,1)$ and $K \in \mathbb{Z}^+$.
Output : $A(S, \xi_{k^*})$.
Uniformly divide S into two disjoint subsets S_1 and S_2 with $|S_1| = (1-\mu)N, |S_2| = \mu N$.
for $k = 1, 2, ..., K$ **do**
  Estimate $A(S_1, \xi_k)$ as an output of A over the subset S_1 with random bit ξ_k.
end
Select the random bit k^* according to $k^* = \arg\min_{k\in[K]} |R_{S_2}(A(S_1, \xi_k)) - R_{S_1}(A(S_1, \xi_k))|$.

In view of the standard risk decomposition, the following excess risk tail bound can be readily obtained via applying Theorem 1 and Hoeffding's inequality:

$$\mathbb{E}_\xi \left[R(A(S,\xi)) - R^*\right] \lesssim \Delta_{\text{opt}} + \gamma_{L_2,N} \log(N) \log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{N}}. \tag{13}$$

Here recall that $\Delta_{\text{opt}} := \mathbb{E}_\xi\left[R_S(A(S,\xi))\right] - \min_{w\in\mathcal{W}} R_S(w)$ is the sub-optimality of empirical risk minimization. Since the excess risk is by definition non-negative, the above bound can also be obtained under the weaker notion of on-average uniform stability (2) via applying (4). In this sense, the first-moment generalization error bound in Theorem 1 is substantially more challenging to derive than the excess risk bound. Additionally, under the generalized Bernstein condition (Koltchinskii, 2006), the risk bound (13) can be readily improved to (5) by directly applying the corresponding deviation optimal risk bound of Klochkov and Zhivotovskiy (2021) to the on-average loss function $\mathbb{E}_\xi[\ell(A(S,\xi); Z)]$ under on-average uniform stability condition.

3 Boosting the Confidence of Generalization

The confidence-boosting technique of Schapire (1990) is a classic meta approach that allows one to boost the dependence of a learning algorithm on the failure probability δ from $1/\delta$ to $\log(1/\delta)$, at a certain cost of computational complexity. In this section, we show an implication of our first-moment bound in Theorem 1 for achieving high-probability generalization jointly over the randomness of data and algorithm, inside a natural framework of confidence-boosting.

3.1 Confidence boosting via bagging

Given a randomized learning algorithm A, we propose to study a bagging based confidence-boosting procedure as outlined in Algorithm 1. In this meta procedure, we independently run the algorithm A for K times over S_1, a fraction of the training set, to obtain K different candidate solutions $\{A(S_1, \xi_k)\}_{k\in[K]}$. Then we evaluate the validation error of these candidate solutions over the holdout training subset S_2, and cherry pick $A(S_1, \xi_{k^*})$ that has the smallest gap between the training error and validation error, i.e., $k^* = \arg\min_{k\in[K]} |R_{S_2}(A(S_1, \xi_k)) - R_{S_1}(A(S_1, \xi_k))|$. Particularly, consider that the internal randomness of A arises from random sampling with replacement of data points, such as SGD under with-replacement sampling. Then in this setting, the procedure can be regarded as a version of bagging (Breiman, 1996) with a greedy model ensemble scheme, which is invoked to the deterministic counterpart of A with fixed random bits (e.g., SGD with identity permutation) over the training subset S_1.

3.2 Jointly exponential bounds

The following theorem is our main result about the generalization error bound of the output $A(S_1, \xi_{k^*})$ that holds with high probability over the entire training set S and the random seeds $\{\xi_k\}_{k\in[K]}$.

Theorem 2. *Suppose that a randomized learning algorithm $A : \mathcal{Z}^N \times \mathcal{R} \mapsto \mathcal{W}$ has L_2-uniform stability with parameter $\gamma_{L_2,N}$. Assume that the loss function ℓ is valued in $[0, M]$. Then for any $\delta \in (0,1)$ and $K \geq 2\log(\frac{2}{\delta})$, with probability at least $1-\delta$ over the randomness of S and $\{\xi_k\}_{k\in[K]}$, the output of Algorithm 1 satisfies*

$$|R(A(S_1, \xi_{k^*})) - R_S(A(S_1, \xi_{k^*}))| \lesssim \gamma_{L_2,(1-\mu)N} \log(N) \log\left(\frac{1}{\delta}\right) + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(K/\delta)}{N}}.$$

Algorithm 2: $A_{\texttt{SGD-w}}$: SGD under With-Replacement Sampling

Input :Data set $S = \{Z_i\}_{i\in[N]}$, step-sizes $\{\eta_t\}_{t\ge1}$, #iterations T, initialization w_0.
Output : $\bar{w}_T = \frac{1}{T}\sum_{t\in[T]} w_t$.
for $t = 1, 2, ..., T$ **do**
  Uniformly randomly sample an index $i_t \in [N]$ with replacement;
  Compute $w_t = \Pi_{\mathcal{W}}\left(w_{t-1} - \eta_t \nabla_w \ell(w_{t-1}; Z_{i_t})\right)$.
end

Proof in sketch. Based on Theorem 1, we first prove an intermediate result to show that the minimal generalization error of the K outputs satisfies $\min_{k\in[K]} |R(A(S_1,\xi_k)) - R_{S_1}(A(S_1,\xi_k))| \lesssim \gamma_{\mathrm{L}_2,(1-\mu)N}\log(N)\log\left(\frac{1}{\delta}\right) + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(1/\delta)}{N}}$ provided that $K \gtrsim \log(\frac{1}{\delta})$. Next we show that the used greedy model selection strategy guarantees that the selected $A(S,\xi_{k^*})$ mimics the generalization behavior of that best performer among the K candidates, with a slightly expanded $\log(K/\delta)$ factor representing the overhead of simultaneously bounding the generalization performance of K different candidate solutions over the holdout validation set. Finally the desired bound follows from the union probability argument. See Appendix B.1 for its full proof. □

Remark 6. *The bound in Theorem 2 holds with high probability jointly over the randomness of sample and algorithm. Different from the bound in (7) that requires high probability uniform stability, Theorem 2 is valid under a substantially milder notion of L_2-uniform stability, though at the cost of multiple running of algorithm for confidence boosting. Compared to the bound in (8) that requires certain conditional uniform stability over the random bits of algorithm, our bound has sharper dependence on the uniform stability parameter yet under a weaker notion of stability.*

Remark 7. *Regarding the scale of the factor $1/\sqrt{\mu(1-\mu)}$ in the bound of Theorem 2, if setting $\mu = 0.01$ (i.e., 99% of S are used as S_1 for training), then the factor is around 10.05.*

Concerning the excess risk of Algorithm 1, we consider a slightly modified output $A(S_1,\xi_{k^*})$ such that $k^* = \arg\min_{k\in[K]} R_{S_2}(A(S_1,\xi_k))$. Then based on the in-expectation risk bound (13), we can derive the following excess risk bound under the conditions of Theorem 2 using similar arguments:

$$R(A(S_1,\xi_{k^*})) - R^* \lesssim \Delta_{\text{opt}} + \gamma_{\mathrm{L}_2,(1-\mu)N}\log(N)\log\left(\frac{1}{\delta}\right) + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(K/\delta)}{N}}. \quad (14)$$

Again, the above risk bound is still valid under the weaker notion of on-average uniform stability (2).

4 Implications for SGD

This section is devoted to demonstrating the implications of Theorem 1 and Theorem 2 for the widely used SGD algorithm and its confidence-boosted versions as well. We focus on a variant of SGD under with-replacement sampling as outlined in Algorithm 2, which we call $A_{\texttt{SGD-w}}$. In what follows, we substantialize $\xi = \{i_t\}_{t\in[T]}$ the sample path of $A_{\texttt{SGD-w}}$ over a given data set, and $\{\xi_k\}_{k\in[K]}$ the K independent copies of ξ when implemented with bagging as shown in Algorithm 1. Our results can also be extended to the without-replacement variant of SGD and the corresponding results are provided in Appendix D for the sake of completeness.

4.1 Convex optimization with smooth loss

We first present the following lemma that establishes the L_2-uniform stability of $A_{\texttt{SGD-w}}$ with convex and smooth loss functions, such as logistic loss. See Appendix C.2 for its proof.

Lemma 1. *Suppose that the loss function $\ell(\cdot;\cdot)$ is convex, G-Lipschitz and L-smooth with respect to its first argument. Assume that $\eta_t \le 2/L$ for all $t \ge 1$. Then $A_{\texttt{SGD-w}}$ has L_2-uniform stability with parameter*

$$\gamma_{\mathrm{L}_2,N} = 2G^2\sqrt{10\left(\frac{1}{N}\sum_{t=1}^{T}\eta_t^2 + \frac{1}{N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2\right)}.$$

Given Lemma 1, we can apply Theorem 1 and Theorem 2 to immediately obtain the following generalization result for $A_{\texttt{SGD-w}}$ and its confidence-boosted version with smooth and convex losses.

Corollary 1. *Suppose that the loss function $\ell(\cdot;\cdot) \in [0, M]$ is convex, G-Lipschitz and L-smooth with respect to its first argument. Then for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ over the randomness of S that $\mathbb{E}_\xi\left[|R(A_{\texttt{SGD-w}}(S,\xi)) - R_S(A_{\texttt{SGD-w}}(S,\xi))|\right] \lesssim$*

$$G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\frac{1}{N}\sum_{t=1}^{T}\eta_t^2 + \frac{1}{N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2} + M\sqrt{\frac{\log(1/\delta)}{N}}.$$

Moreover, consider Algorithm 1 specified to $A_{\texttt{SGD-w}}$ with learning rate $\eta_t \le 2/L$ and $K \asymp \log(\frac{1}{\delta})$. Then with probability at least $1-\delta$ over the randomness of S and $\{\xi_k\}_{k\in[K]}$, it holds that $|R(A_{\texttt{SGD-w}}(S_1,\xi_{k^*})) - R_S(A_{\texttt{SGD-w}}(S_1,\xi_{k^*}))| \lesssim$

$$G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\frac{1}{(1-\mu)N}\sum_{t=1}^{T}\eta_t^2 + \frac{1}{(1-\mu)^2N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2} + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(1/\delta)}{N}}.$$

Remark 8. *For the conventional choice of $\eta_t = \frac{2}{L\sqrt{t}}$, the high-probability (w.r.t. data) generalization bounds in Corollary 1 for SGD and its confidence boosted version are roughly of scale $\mathcal{O}\big(\log(N)\log\left(\frac{1}{\delta}\right)\sqrt{\frac{\log(T)}{N}+\frac{\sqrt{T}}{N}}\big)$, which matches the corresponding $\mathcal{O}\big(\frac{\sqrt{T}}{N}\big)$ in-expectation bound of SGD with smooth and convex losses (Hardt et al., 2016).*

Combining with the standard in-expectation optimization error bound of convex SGD (see, e.g., Shamir and Zhang, 2013), we can show the following excess risk bound of (modified) Algorithm 1 as a direct consequence of the generic bound (14) to $A_{\texttt{SGD-w}}$ with convex and smooth losses:

$$\begin{aligned} R(A_{\texttt{SGD-w}}(S_1,\xi_{k^*})) - R^* \lesssim & G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\frac{1}{(1-\mu)N}\sum_{t=1}^{T}\eta_t^2 + \frac{1}{(1-\mu)^2N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2} \\ & + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(1/\delta)}{N}} + \frac{D^2(w_0,W^*) + G^2\sum_{t=1}^{T}\eta_t^2}{\sum_{t=1}^{T}\eta_t}, \end{aligned}$$

where $W^* := \mathrm{Argmin}_{w\in\mathcal{W}} R(w)$ and $D(w, W^*) = \min_{w^*\in W^*}\|w - w^*\|$. With learning rate $\eta_t = \frac{2}{L\sqrt{t}}$, the right hand side of the above roughly scales as $\mathcal{O}\big(\sqrt{\log(N)\log\left(\frac{1}{\delta}\right)\frac{\log(T)}{N}} + \frac{\sqrt{T}}{N} + \frac{\log(T)}{\sqrt{T}}\big)$ which matches the prior high-probability excess risk bounds of SGD with convex losses (Harvey et al., 2019, Remark 3.7).

4.2 Convex optimization with non-smooth loss

Now we turn to study the case where the loss is convex but not necessarily smooth, such as the hinge loss and absolute loss. We first establish the following lemma about the L_2-uniform stability parameter of $A_{\texttt{SGD-w}}$ in the considered setting. See Appendix C.3 for its proof.

Lemma 2. *Suppose that the loss function $\ell(\cdot;\cdot)$ is convex and G-Lipschitz with respect to its first argument. Then $A_{\texttt{SGD-w}}$ has L_2-uniform stability with parameter*

$$\gamma_{L_2,N} = G^2\sqrt{40\sum_{t=1}^{T}\eta_t^2 + \frac{32}{N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2}.$$

With Lemma 2 in place, we can readily apply Theorem 1 and Theorem 2 to establish the following corollary about the generalization bounds of $A_{\texttt{SGD-w}}$ and its confidence-boosted version with convex and non-smooth loss functions.

Corollary 2. *Suppose that the loss function $\ell(\cdot;\cdot) \in [0, M]$ is convex and G-Lipschitz with respect to its first argument. Then for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ over the randomness of S that $\mathbb{E}_\xi\left[|R(A_{\texttt{SGD-w}}(S,\xi)) - R_S(A_{\texttt{SGD-w}}(S,\xi))|\right] \lesssim$*

$$G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\sum_{t=1}^{T}\eta_t^2 + \frac{1}{N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2} + M\sqrt{\frac{\log(1/\delta)}{N}}.$$

Moreover, consider Algorithm 1 specified to $A_{\texttt{SGD-w}}$ with $K \asymp \log(\frac{1}{\delta})$. Then with probability at least $1-\delta$ over S and $\{\xi_k\}_{k\in[K]}$, it holds that $|R(A_{\texttt{SGD-w}}(S_1,\xi_{k^*})) - R_S(A_{\texttt{SGD-w}}(S_1,\xi_{k^*}))| \lesssim$

$$G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\sum_{t=1}^{T} \eta_t^2 + \frac{1}{(1-\mu)^2 N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2} + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(1/\delta)}{N}}.$$

Remark 9. *For SGD with decaying learning rates $\eta_t = \frac{1}{\sqrt{Nt}}$, Corollary 2 admits high-probability generalization bounds of scale $\mathcal{O}\big(\log(N)\log\left(\frac{1}{\delta}\right)\sqrt{\frac{\log(T)}{N} + \frac{T}{N^3}} + \sqrt{\frac{\log(1/\delta)}{N}}\big)$. With fixed rates $\eta_t \equiv \eta$, Corollary 2 yields deviation bounds of scale $\mathcal{O}\big(\eta \log(N)\log\left(\frac{1}{\delta}\right)(\sqrt{T} + \frac{T}{N}) + \sqrt{\frac{\log(1/\delta)}{N}}\big)$ which matches the near-optimal rate by Bassily et al. (2020, Theorem 3.3).*

4.3 Non-convex optimization with smooth loss

We further study the performance of Algorithm 1 for $A_{\texttt{SGD-w}}$ with smooth but not necessarily convex loss functions, such as normalized sigmoid loss (Mason et al., 1999). The following lemma estimates the L_2-uniform stability of $A_{\texttt{SGD-w}}$ in the considered setting. See Appendix C.4 for its proof.

Lemma 3. *Suppose that the loss function $\ell(\cdot;\cdot)$ is G-Lipschitz and L-smooth with respect to its first argument. Consider $\eta_t \le 1/L$. Then $A_{\texttt{SGD-w}}$ has L_2-uniform stability with parameter*

$$\gamma_{L_2,N} = 2G^2 \sqrt{\frac{1}{N}\sum_{t=1}^{T} \exp\left(3L \sum_{\tau=t+1}^{T} \eta_\tau\right) u_t},$$

where

$$u_t := \eta_t^2 + 2\eta_t \sum_{\tau=1}^{t-1} \exp\left(L \sum_{i=\tau+1}^{t-1} \eta_i\right)\eta_\tau.$$

Based on Lemma 3, we can invoke Theorem 1 and Theorem 2 to show the following generalization result for $A_{\texttt{SGD-w}}$ and its confidence-boosted version with non-convex and smooth loss functions.

Corollary 3. *Suppose that the loss function $\ell(\cdot;\cdot) \in [0, M]$ is G-Lipschitz and L-smooth with respect to its first argument. Then for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ over the randomness of S that $\mathbb{E}_\xi\left[|R(A_{\texttt{SGD-w}}(S,\xi)) - R_S(A_{\texttt{SGD-w}}(S,\xi))|\right] \lesssim$*

$$G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\frac{1}{N}\sum_{t=1}^{T} \exp\left(L\sum_{\tau=t+1}^{T}\eta_\tau\right) u_t} + M\sqrt{\frac{\log(1/\delta)}{N}},$$

where $u_t := \eta_t^2 + 2\eta_t \sum_{\tau=1}^{t-1}\exp(L\sum_{i=\tau+1}^{t-1}\eta_i)\eta_\tau$ for all $t \ge 1$. Moreover, consider Algorithm 1 specified to $A_{\texttt{SGD-w}}$ with $\eta_t \le \frac{1}{L}$ and $K \asymp \log(\frac{1}{\delta})$. Then with probability at least $1-\delta$ over S and $\{\xi_k\}_{k\in[K]}$, it holds that $|R(A_{\texttt{SGD-w}}(S_1,\xi_{k^*})) - R_S(A_{\texttt{SGD-w}}(S_1,\xi_{k^*}))| \lesssim$

$$G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\frac{1}{(1-\mu)N}\sum_{t=1}^{T} \exp\left(L\sum_{\tau=t+1}^{T}\eta_\tau\right) u_t} + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(1/\delta)}{N}}.$$

Remark 10. *For the decaying learning rates $\eta_t = \frac{1}{L\nu t}$ with arbitrary $\nu \ge 1$, the generalization bounds in Corollary 3 are of scale $\mathcal{O}\big(\log(N)\log\left(\frac{1}{\delta}\right)\sqrt{\frac{T^{1/\nu}\log(T)}{\nu N}} + \sqrt{\frac{\log(1/\delta)}{N}}\big)$. For the constant learning rates $\eta_t \equiv \frac{1}{LT}$, the bounds are of scale $\mathcal{O}\big(\log(N)\log\left(\frac{1}{\delta}\right)\sqrt{\frac{\log(1/\delta)}{N}}\big)$.*

5 Conclusion

In this paper, we have introduced a novel concept of L_2-uniform stability for randomized learning algorithms and proved a strong first-moment generalization bound that holds with high probability over training sample. Equipped with this result, we have further developed a bagging based confidence-boosting procedure and shown that it yields near-optimal generalization bounds with high confidence jointly over the randomness of sample and algorithm. The power of our theory has been demonstrated through an application to SGD with time-decaying learning rates, where sharper generalization bounds have been obtained for both convex and non-convex loss functions.

Acknowledgments and Disclosure of Funding

The authors sincerely thank the anonymous reviewers and area chairs for their insightful comments on this paper. The research was conducted while XTY worked for Nanjing University of Information Science and Technology and both authors worked for Baidu Cognitive Computing Lab. The work of XTY is also funded in part by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, and in part by the Natural Science Foundation of China (NSFC) under Grant No. U21B2049, 61936005.

References

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11279–11288, Vancouver, Canada, 2019.

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

Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514–560, 2005.

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 610–626, Virtual Event [Graz, Austria], 2020.

Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

Peter Bühlmann. Bagging, boosting and ensemble methods. In *Handbook of computational statistics*, pages 985–1022. Springer, 2012.

Zachary Charles and Dimitris S. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 744–753, Stockholmsmässan, Stockholm, Sweden, 2018.

Yuan Shih Chow and Henry Teicher. *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media, 2003.

Luc Devroye and Terry J. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. Inf. Theory*, 25(2):202–207, 1979.

André Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *J. Mach. Learn. Res.*, 6:55–79, 2005.

Vitaly Feldman and Jan Vondrák. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9770–9780, Montréal, Canada, 2018.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 1270–1279, Phoenix, AZ, 2019.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 439–449, Chicago, IL, 2020.

Dylan J. Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6726–6736, Vancouver, Canada, 2019.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 1225–1234, New York City, NY, 2016.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 1579–1613, Phoenix, AZ, 2019.

Matthew Holland. Robustness and scalability under heavy tails, without strong convexity. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 865–873, Virtual Event, 2021.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5065–5076, virtual, 2021.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Ilja Kuzborskij and Christoph H. Lampert. Data-dependent stability of stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2820–2829, Stockholmsmässan, Stockholm, Sweden, 2018.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5809–5819, Virtual Event, 2020.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.

Liam Madden, Emiliano DallAnese, and Stephen Becker. High probability convergence and uniform stability bounds for nonconvex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2020.

Llew Mason, Jonathan Baxter, Peter L. Bartlett, and Marcus R. Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 512–518, Denver, CO, 1999.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in Combinatorics*, 141(1): 148–188, 1989.

Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1085–1093, Fort Lauderdale, FL, 2017.

Sayan Mukherjee, Partha Niyogi, Tomaso A. Poggio, and Ryan M. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.*, 25(1-3):161–193, 2006.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 4703–4711, Long Beach, CA, 2019.

David W. Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.*, 11:169–198, 1999.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, UK, 2012.

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

Robert E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5:197–227, 1990.

Mark Schmidt, Nicolas Le Roux, and Francis R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1458–1466, Granada, Spain, 2011.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, 2010.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems (NIPS)*, pages 46–54, Barcelona, Spain, 2016.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 71–79, Atlanta, GA, 2013.

Giorgio Valentini and Thomas G. Dietterich. Low bias bagged support vector machines. In *Proceedings of the Twentieth International Conference (ICML)*, pages 752–759, Washington, DC, 2003.

V. N. Vapnik and A. Ya. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, 1974.

Xiao-Tong Yuan and Ping Li. Stability and risk bounds of iterative hard thresholding. *IEEE Trans. Inf. Theory*, 68(10):6663–6681, 2022.

Xiao-Tong Yuan and Ping Li. Exponential generalization bounds with near-optimal rates for l_q-stable algorithms. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023.

Tong Zhang. Leave-one-out bounds for kernel methods. *Neural Comput.*, 15(6):1397–1437, 2003.

Yi Zhou, Yingbin Liang, and Huishuai Zhang. Understanding generalization error of SGD in nonconvex optimization. *Mach. Learn.*, 111(1):345–375, 2022.

A Proofs for Section 2

In this section, we provide the technical proofs for the main results stated in Section 2.

A.1 Auxiliary lemmas

Here we collect a set of preliminary lemmas to be used in our analysis. The following lemma is an L_q-norm extension of the celebrated Efron-Stein inequality (see, e.g., Boucheron et al., 2005, Theorem 2).

Lemma 4 (Generalized Efron-Stein inequality)**.** *Let $S = \{Z_1, ..., Z_N\}$ be a set of independent random variables valued in $\mathcal{Z}$ and $h : \mathcal{Z}^N \mapsto \mathbb{R}$ be some measurable function. Then for all $q \geq 2$,*

$$\|h(S) - \mathbb{E}[h(S)]\|_q \leq \sqrt{3q}\sqrt{\left\|\sum_{i=1}^{N}\left(h(S) - h(S^{(i)})\right)^2\right\|_{q/2}}.$$

The following result compares the moments and conditional moments of a random function.

Lemma 5. *Let $S = \{Z_1, ..., Z_N\}$ be a set of independent random variables valued in some measure space $\mathcal{Z}$ and $h : \mathcal{Z}^N \mapsto \mathbb{R}$ be some measurable function. Then for all $I \subseteq [N]$ and $q \geq 1$, we have*

$$\|h(S_I)\|_q \leq \|h(S)\|_q = \|\|h\|_q(S_I)\|_q .$$

Proof. Recall $h(S_I) = \mathbb{E}[h(S) \mid S_I]$. Using Jensen's inequality we can show that

$$\|h(S_I)\|_q = \left(\mathbb{E}\left[|\mathbb{E}[h(S) \mid S_I]|^q\right]\right)^{1/q} \leq \left(\mathbb{E}\left[\mathbb{E}[|h(S)|^q \mid S_I]\right]\right)^{1/q} = \left(\mathbb{E}[|h(S)|^q]\right)^{1/q} = \|h(S)\|_q.$$

By definition we can also express $\|h(S)\|_q = \left(\mathbb{E}\left[\mathbb{E}[|h(S)|^q \mid S_I]\right]\right)^{1/q} = \|\|h(S)\|_q(S_I)\|_q$. The proof is completed. □

We further need to introduce the concept of weakly self-bounding function to be used in the analysis of second-order concentration bounds.

Definition 2 (Weakly self-bounding function)**.** *A non-negative function $h : \mathcal{Z}^N \mapsto \mathbb{R}^+ \cup \{0\}$ is said to be weakly (a, b)-self-bounding with parameters $a, b \geq 0$ if there exist non-negative functions $h_i : \mathcal{Z}^{N-1} \mapsto \mathbb{R}^+ \cup \{0\}$ such that for all $S = \{Z_1, ..., Z_n\} \in \mathcal{Z}^N$,*

$$\sum_{i=1}^{N}\left(h(S) - h_i(S^{\setminus i})\right)^2 \leq ah(S) + b,$$

where $S^{\setminus i} := S \setminus \{Z_i\}$.

The following lemma is a combination of the upper tail bound of Boucheron et al. (2005, Theorem 6.19) and lower tail bound of Klochkov and Zhivotovskiy (2021, Proposition 3.1) for weakly self-bounding functions.

Lemma 6. *Let $S = \{Z_1, Z_2, ..., Z_N\}$ be a set of independent random variables valued in $\mathcal{Z}$ and $h : \mathcal{Z}^N \mapsto \mathbb{R}^+ \cup \{0\}$ be a weakly (a, b)-self-bounding function.*

- *Assume that the relevant h_i satisfy $h_i(S^{\setminus i}) \leq h(S)$ for any $i = 1, ..., N$. Then for all $t > 0$,*

$$\mathbb{P}\left\{h(S) \geq \mathbb{E}[h(S)] + t\right\} \leq \exp\left(-\frac{t^2}{2a\mathbb{E}(h(S)) + 2b + at}\right).$$

- *Assume that the relevant h_i satisfy $h_i(S^{\setminus i}) \geq h(S)$ for any $i = 1, ..., N$. Then for all $t > 0$,*

$$\mathbb{P}\left\{h(S) \leq \mathbb{E}[h(S)] - t\right\} \leq \exp\left(-\frac{t^2}{2a\mathbb{E}(h(S)) + 2b}\right).$$

We also need the following preliminary result about the equivalence between tails and moments (Bousquet et al., 2020).

Lemma 7. *Let Y be a real-valued random variable.*

- *If Y satisfies the following inequality for some $a, b \geq 0$ with probability at least $1-\delta$ for any $\delta \in (0,1)$,*
$$|Y| \leq a \log\left(\frac{e}{\delta}\right) + b\sqrt{\log\left(\frac{e}{\delta}\right)}.$$
Then, for any $q \geq 1$ it holds that
$$\|Y\|_q \leq 3aq + 9b\sqrt{q}.$$
- *If Y satisfies $\|Y\|_q \leq aq + b\sqrt{q}$ for any $q \geq 1$. Then the following holds with probability at least $1-\delta$ for any $\delta \in (0,1)$:*
$$|Y| \leq e\left(a \log\left(\frac{e}{\delta}\right) + b\sqrt{\log\left(\frac{e}{\delta}\right)}\right).$$

A.2 Proof of Proposition 1

The following lemma is key to our proof.

Lemma 8. *Let $S = \{Z_1, Z_2, ..., Z_N\}$ be a set of independent random variables valued in $\mathcal{Z}$ and ξ be a random variable valued in $\mathcal{R}$. Let $g : \mathcal{Z}^N \times \mathcal{R} \mapsto \mathbb{R}$ be a measurable function. Then it holds that*
$$\mathbb{E}\left[(g(S,\xi) - \mathbb{E}_S[g(S,\xi)])^2\right] \leq \sum_{i=1}^N \mathbb{E}\left[\left(g(S,\xi) - g(S^{(i)},\xi)\right)^2\right]. \tag{15}$$
Moreover, for any $q \geq 2$, the following bound holds:
$$\begin{aligned}&\|\mathbb{E}_\xi\left[|g(S,\xi) - \mathbb{E}_S\left[g(S,\xi)\right]|\right]\|_q \\ \leq&\sqrt{\sum_{i=1}^N \mathbb{E}\left[\left(g(S,\xi) - g(S^{(i)},\xi)\right)^2\right]} + \sqrt{3q}\sqrt{\left\|\sum_{i=1}^N \mathbb{E}_\xi^2\left[\left|g(S,\xi) - g(S^{(i)},\xi)\right|\right]\right\|_{q/2}}.\end{aligned} \tag{16}$$

Proof. To prove the variance bound (15), we consider the following conditional expectation operators, conditioned on the random variables $(Z_1, ..., Z_i)$ and the random bit ξ of algorithm:
$$f_i := \mathbb{E}\left[g(S,\xi) \mid Z_1, ..., Z_i, \xi\right], \quad \forall i = 1, ..., N.$$
We conventionally define $f_0 = \mathbb{E}_S[g(S,\xi)]$. Clearly, the following telescope decomposition holds:
$$g(S,\xi) - \mathbb{E}_S[g(S,\xi)] = f_N - f_0 = \sum_{i=1}^N \left\{\Delta_i := f_i - f_{i-1}\right\}.$$
Then we have
$$\begin{aligned}&\mathbb{E}\left[(g(S,\xi) - \mathbb{E}_S[g(S,\xi)])^2\right] \\ =&\mathbb{E}\left[\left(\sum_{i=1}^N \Delta_i\right)^2\right] = \sum_{i=1}^N \mathbb{E}[\Delta_i^2] + 2\sum_{i<j}\mathbb{E}[\Delta_i\Delta_j] = \sum_{i=1}^N \mathbb{E}[\Delta_i^2],\end{aligned} \tag{17}$$
where in the last equality we have used the fact that for any index pair $i < j$, $\mathbb{E}\left[\Delta_j \mid Z_1, ..., Z_i, \xi\right] = 0$ which implies $\mathbb{E}[\Delta_i\Delta_j] = \mathbb{E}\left[\Delta_i \mathbb{E}\left[\Delta_j \mid Z_1, ..., Z_i, \xi\right]\right] = 0$. Note that,
$$\begin{aligned}\mathbb{E}[\Delta_i^2] =&\mathbb{E}\left[\left(\mathbb{E}\left[g(S,\xi) \mid Z_1, ..., Z_i, \xi\right] - \mathbb{E}\left[g(S,\xi) \mid Z_1, ..., Z_{i-1}, \xi\right]\right)^2\right] \\ \overset{\zeta_1}{=}&\mathbb{E}\left[\left(\mathbb{E}\left[g(S,\xi) - g(S^{(i)},\xi)\right]\right)^2 \mid Z_1, ..., Z_i, \xi\right] \\ \overset{\zeta_2}{\leq}&\mathbb{E}\left[\mathbb{E}\left[\left(g(S,\xi) - g(S^{(i)},\xi)\right)^2\right] \mid Z_1, ..., Z_i, \xi\right] \\ =&\mathbb{E}\left[\left(g(S,\xi) - g(S^{(i)},\xi)\right)^2\right],\end{aligned} \tag{18}$$

where "ζ_1" makes use of the independence of $\xi, Z_1, ..., Z_i, Z_i'$, "ζ_2" is due to Jensen's inequality. Substituting (18) into (17) yields

$$\mathbb{E}\left[(g(S,\xi)-\mathbb{E}_S[g(S,\xi)])^2\right] \le \sum_{i=1}^{N}\mathbb{E}\left[\left(g(S,\xi)-g(S^{(i)},\xi)\right)^2\right],$$

which is the first desired bound.

We now proceed to prove the q-moment bound (16). Let us define $h(S) := \mathbb{E}_\xi\left[|g(S,\xi)-\mathbb{E}_S[g(S,\xi)]|\right]$. Based on Jensen's inequality and triangle inequality we can show that

$$\begin{aligned}\left|h(S)-h(S^{(i)})\right| =& \left|\mathbb{E}_\xi\left[|g(S,\xi)-\mathbb{E}_S[g(S,\xi)]| - \left|g(S^{(i)},\xi)-\mathbb{E}_{S^{(i)}}[g(S^{(i)},\xi)]\right|\right]\right| \\ \le& \mathbb{E}_\xi\left[\left|g(S,\xi)-\mathbb{E}_S[g(S,\xi)] - g(S^{(i)},\xi)+\mathbb{E}_{S^{(i)}}[g(S^{(i)},\xi)]\right|\right] \\ =& \mathbb{E}_\xi\left[\left|g(S,\xi)-g(S^{(i)},\xi)\right|\right].\end{aligned}$$

Then by invoking the generalized Efron-Stein inequality (Lemma 4) to $h(S)$ we get that for all $q\ge 2$,

$$\begin{aligned}\|h(S)-\mathbb{E}_S[h(S)]\|_q \le& \sqrt{3q}\sqrt{\left\|\sum_{i=1}^{N}\left(h(S)-h(S^{(i)})\right)^2\right\|_{q/2}} \\ \le& \sqrt{3q}\sqrt{\left\|\sum_{i=1}^{N}\mathbb{E}_\xi^2\left[|g(S,\xi)-g(S^{(i)},\xi)|\right]\right\|_{q/2}}.\end{aligned}$$

It follows that

$$\begin{aligned}\|h(S)\|_q \le& |\mathbb{E}_S[h(S)]| + \sqrt{3q}\sqrt{\left\|\sum_{i=1}^{N}\mathbb{E}_\xi^2\left[|g(S,\xi)-g(S^{(i)},\xi)|\right]\right\|_{q/2}} \\ =& \mathbb{E}\left[|g(S,\xi)-\mathbb{E}_S[g((S,\xi)]|\right] + \sqrt{3q}\sqrt{\left\|\sum_{i=1}^{N}\mathbb{E}_\xi^2\left[|g(S,\xi)-g(S^{(i)},\xi)|\right]\right\|_{q/2}} \\ \le& \sqrt{\sum_{i=1}^{N}\mathbb{E}\left[\left(g(S,\xi)-g(S^{(i)},\xi)\right)^2\right]} + \sqrt{3q}\sqrt{\left\|\sum_{i=1}^{N}\mathbb{E}_\xi^2\left[|g(S,\xi)-g(S^{(i)},\xi)|\right]\right\|_{q/2}},\end{aligned}$$

where in the last inequality we have used Jensen's inequality and (15). This gives the desired q-moment bound in the second part. □

Remark 11. *The first variance bound in (15) can be regarded as a natural extension of the Efron-Stein inequality to randomized functions.*

Based on Lemma 8, we can prove the main result in Proposition 1.

Proof of Proposition 1. The concentration bound (11) can be implied by (16) and the bounded-difference condition as in the following:

$$\begin{aligned}&\|\mathbb{E}_\xi\left[|g(S,\xi)-\mathbb{E}_S[g(S,\xi)]|\right]\|_q \\ \le& \sqrt{\sum_{i=1}^{N}\mathbb{E}\left[\left(g(S,\xi)-g(S^{(i)},\xi)\right)^2\right]} + \sqrt{3q}\sqrt{\left\|\sum_{i=1}^{N}\mathbb{E}_\xi^2\left[|g(S,\xi)-g(S^{(i)},\xi)|\right]\right\|_{q/2}} \\ \le& \beta\sqrt{N}+\sqrt{3q}\sqrt{N\beta^2} \le 3\beta\sqrt{Nq}.\end{aligned}$$

To prove the second-order concentration bound (12), we first show via the inequality (15) in Lemma 8 and the bounded difference condition that

$$\begin{aligned}&\mathbb{E}_\xi\left[(g(S,\xi)-\mathbb{E}_S[g(S,\xi)])^2\right] \\ \le& \sum_{i=1}^{N}\mathbb{E}\left[\left(g(S,\xi)-g(S^{(i)},\xi)\right)^2\right] = \sum_{i=1}^{N}\mathbb{E}_S\left[\mathbb{E}_\xi\left[\left(g(S,\xi)-g(S^{(i)},\xi)\right)^2\right]\right] \le N\beta^2.\end{aligned} \tag{19}$$

Let us define $h(S) := \mathbb{E}_\xi\left[(g(S,\xi) - \mathbb{E}_S[g(S,\xi)])^2\right]$. Let $h_i^-(S^{\setminus i}) := \inf_{Z_i\in\mathcal{Z}} h(S)$ such that $h_i^- \le h$. It can be shown that

$$\begin{aligned}
&\sum_{i=1}^N \left(h(S) - h_i^-(S^{\setminus i})\right)^2 \\
=&\sum_{i=1}^N \left(\mathbb{E}_\xi\left[(g(S,\xi) - \mathbb{E}_S[g(S,\xi)])^2\right] - \inf_{Z_i\in\mathcal{Z}} \mathbb{E}_\xi\left[(g(S,\xi) - \mathbb{E}_S[g(S,\xi)])^2\right]\right)^2 \\
\overset{\zeta_1}{\le}& N\beta^2\left(\beta + 2\mathbb{E}_\xi\left[|g(S,\xi) - \mathbb{E}_S[g(S,\xi)]|\right]\right)^2 \\
\le& 8N\beta^2 h(S) + 2N\beta^4,
\end{aligned}$$

where in "ζ_1" we have used Jensen's inequality, Cauchy-Schwarz inequality and the bounded difference assumption. Therefore h is a weakly $(8N\beta^2, 2N\beta^4)$-self-bounding function. Then for any $\delta \in (0,1)$, it follows from the first upper tail bound in Lemma 6 that the following bound holds with probability at least $1-\frac{\delta}{2}$:

$$h(S) \le \mathbb{E}[h(S)] + 8N\beta^2\log\left(\frac{2}{\delta}\right) + 2\sqrt{(4N\beta^2\mathbb{E}[h(S)] + N\beta^4)\log\left(\frac{2}{\delta}\right)},$$

Now consider $h_i^+(S^{\setminus i}) := \sup_{Z_i\in\mathcal{Z}} h(S)$ such that $h_i^+ \ge h$. Similar to the previous arguments we can show according to the second lower tail bound in Lemma 6 that with probability at least $1-\frac{\delta}{2}$,

$$h(S) \ge \mathbb{E}[h(S)] - 2\sqrt{(4N\beta^2\mathbb{E}[h(S)] + N\beta^4)\log\left(\frac{2}{\delta}\right)}.$$

Combing the preceding two inequalities yields that the following holds with probability at least $1-\delta$

$$|h(S) - \mathbb{E}[h(S)]| \le 8N\beta^2\log\left(\frac{2}{\delta}\right) + 2\sqrt{(4N\beta^2\mathbb{E}[h(S)] + N\beta^4)\log\left(\frac{2}{\delta}\right)}.$$

In view of Lemma 7 we further have that for any $q \ge 1$,

$$\|h(S) - \mathbb{E}[h(S)]\|_q \le 24N\beta^2 q + 18\sqrt{(4N\beta^2\mathbb{E}[h(S)] + N\beta^4)\,q}.$$

Consequently we have

$$\begin{aligned}
\|h(S)\|_q \le& \mathbb{E}[h(S)] + 24N\beta^2 q + 9\left(\mathbb{E}[h(S)] + \frac{\beta^2}{4} + 4N\beta^2 q\right) \\
\le& 10\mathbb{E}[h(S)] + 63N\beta^2 q \le 10N\beta^2 + 63N\beta^2 q \le 68N\beta^2 q,
\end{aligned}$$

where we have used the fact $2ab \le a^2 + b^2$, (19) and $q \ge 2$. The proof is completed. □

A.3 Proof of Proposition 2

We need the following lemma as another important and useful consequence of Lemma 8 which can be regarded as an extension of the Marcinkiewicz-Zygmund inequality (Chow and Teicher, 2003) to conditionally independent functions.

Lemma 9. *Let $Z_1, Z_2, ..., Z_N$ be a set of independent random variables valued in $\mathcal{Z}$ and ξ be a random variable valued in $\mathcal{R}$. For any $i \in [N]$, let $f_i : \mathcal{Z}\times\mathcal{R} \mapsto \mathbb{R}$ be a measurable function satisfies $\mathbb{E}[f_i(Z_i,\xi) \mid \xi] = 0$. Then for any $q \ge 2$,*

$$\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N f_i(Z_i,\xi)\right|\right]\right\|_q \le \sqrt{2\sum_{i=1}^N \mathbb{E}\left[f_i^2(Z_i,\xi)\right] + 2\sqrt{3q}\sqrt{\left\|\sum_{i=1}^N \mathbb{E}_\xi^2\left[|f_i(Z_i,\xi)|\right]\right\|_{q/2}}}. \quad (20)$$

Proof. Let $S = \{Z_1, Z_2, ..., Z_N\}$ and consider $g(S, \xi) = \sum_{i=1}^{N} f_i(Z_i, \xi)$. Then it can be verified that $\mathbb{E}_S [g(S, \xi)] = 0$,

$$\mathbb{E}\left[\left(g(S,\xi) - g(S^{(i)},\xi)\right)^2\right] = \mathbb{E}\left[(f_i(Z_i,\xi) - f_i(Z_i',\xi))^2\right] = 2\mathbb{E}\left[f_i^2(Z_i,\xi)\right],$$

and

$$\mathbb{E}_\xi^2\left[\left|g(S,\xi) - g(S^{(i)},\xi)\right|\right] = \mathbb{E}_\xi^2\left[|f_i(Z_i,\xi) - f_i(Z_i',\xi)|\right] \le 2\mathbb{E}_\xi^2\left[|f_i(Z_i,\xi)|\right] + 2\mathbb{E}_\xi^2\left[|f_i(Z_i',\xi|\right].$$

Applying Lemma 8 to $g(S, \xi)$ yields

$$\begin{aligned}
\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^{N} f_i(Z_i,\xi)\right|\right]\right\|_q &= \left\|\mathbb{E}_\xi\left[|g(S,\xi) - \mathbb{E}_S[g(S,\xi)]|\right]\right\|_q \\
&\le \sqrt{2\sum_{i=1}^{N}\mathbb{E}\left[f_i^2(Z_i,\xi)\right]} + \sqrt{3q}\sqrt{\left\|2\sum_{i=1}^{N}\left(\mathbb{E}_\xi^2\left[|f_i(Z_i,\xi)|\right] + \mathbb{E}_\xi^2\left[|f_i(Z_i',\xi)|\right]\right)\right\|_{q/2}} \\
&= \sqrt{2\sum_{i=1}^{N}\mathbb{E}\left[f_i^2(Z_i,\xi)\right]} + 2\sqrt{3q}\sqrt{\left\|\sum_{i=1}^{N}\mathbb{E}_\xi^2\left[|f_i(Z_i,\xi)|\right]\right\|_{q/2}}.
\end{aligned}$$

This proves the desired bound. □

We also need the following lemma which indicates that conditional expectation does not expand the L_2-bounded-difference parameter.

Lemma 10. *Let $S = \{Z_1, Z_2, ..., Z_N\}$ be a set of independent random variables valued in $\mathcal{Z}$ and ξ be a random variable valued in $\mathcal{R}$. Let $g : \mathcal{Z}^N \times \mathcal{R} \mapsto \mathbb{R}$ be some measurable function. Let $I \subseteq [N]$ be an index set. Then for all $i \in I$,*

$$\sup_{S_I, S_I^{(i)}} \sqrt{\mathbb{E}_\xi\left[\left(g(S_I,\xi) - g(S_I^{(i)},\xi)\right)^2\right]} \le \sup_{S, S^{(i)}} \sqrt{\mathbb{E}_\xi\left[\left(g(S,\xi) - g(S^{(i)},\xi)\right)^2\right]}$$

Proof. Recall that $g(S_I, \xi) = \mathbb{E}[g(S, \xi) \mid S_I, \xi]$. Based on Jensen's inequality we can show that

$$\begin{aligned}
\mathbb{E}_\xi\left[\left(g(S_I,\xi) - g(S_I^{(i)},\xi)\right)^2\right] &\le \mathbb{E}_\xi\left[\mathbb{E}\left[\left(g(S,\xi) - g(S^{(i)},\xi)\right)^2 \mid S_I, S_I^{(i)}, \xi\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_\xi\left[\left(g(S,\xi) - g(S^{(i)},\xi)\right)^2\right] \mid S_I, S_I^{(i)}\right] \\
&\le \sup_{S,S^{(i)}} \mathbb{E}_\xi\left[\left(g(S,\xi) - g(S^{(i)},\xi)\right)^2\right],
\end{aligned}$$

where in the last inequality we have used the fact that expectation is always no larger than maximum. The desired bound then follows immediately from the above inequality. □

Now we are in the position to prove the main result. The proof is inspired by the sample-splitting arguments of Feldman and Vondrák (2019); Bousquet et al. (2020), with several non-trivial modifications along made to deal with the challenges arisen from the first-moment operation taken over the randomness of algorithm.

Proof of Proposition 2. Consider k such that $2^{k-1} < N \le 2^k$. If $N < 2^k$, we pad the training set S with extra zero-functions so that $N = 2^k$. Consider the partition $\mathcal{I}_0, \mathcal{I}_1, ..., \mathcal{I}_k$ of $[N]$ given by

$$\mathcal{I}_0 = \{\{1\}, ..., \{2^k\}\}, \ \ \mathcal{I}_1 = \{\{1,2\}, \{3,4\}..., \{2^k - 1, 2^k\}\}, \ \ \mathcal{I}_k = \{\{1, ..., 2^k\}\}.$$

For any $i \in [N]$ and $l = 0, ..., k$, we denote by $I^l(i) \in \mathcal{I}_l$ the only set from $\mathcal{I}_l$ that contains i and consider the following random variables:

$$g_i^l = \mathbb{E}\left[g_i \mid Z_i, S_{\overline{I^l(i)}}, \xi\right].$$

In particular, $g_i^0 = g_i$ and $g_i^k = \mathbb{E}[g_i \mid Z_i, \xi]$. Clearly we have the following telescope sum:

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

It follows that

$$\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^{N}\left(g_i - \mathbb{E}[g_i \mid Z_i, \xi]\right)\right|\right]\right\|_q \le \sum_{l=0}^{k-1}\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^{N} g_i^l - g_i^{l+1}\right|\right]\right\|_q. \tag{21}$$

We need to upper bound the r.h.s. of the above inequality. To this end, it can be verified that

$$g_i^{l+1} = \mathbb{E}\left[g_i \mid Z_i, S_{\overline{I^{l+1}(i)}}, \xi\right] = \mathbb{E}\left[g_i^l \mid Z_i, S_{\overline{I^{l+1}(i)}}, \xi\right].$$

Since by assumption g_i has bounded L_2-uniform difference by β with respect to all variables except the i-th variable, it can be verified by Lemma 10 that so is g_i^l for each $l = 0, ..., k$. Conditioned on $Z_i, S_{\overline{I^{l+1}(i)}}$, invoking Theorem 1 to g_i^l yields that for any $q \ge 1$,

$$\begin{aligned}
\left\|\mathbb{E}_\xi\left[|g_i^l - g_i^{l+1}|\right]\right\|_q\left(Z_i, S_{\overline{I^{l+1}(i)}}\right) \le& 3\beta\sqrt{q2^l},\\
\left\|\mathbb{E}_\xi\left[\left(g_i^l - g_i^{l+1}\right)^2\right]\right\|_q\left(Z_i, S_{\overline{I^{l+1}(i)}}\right) \le& 68\beta^2 2^l q,
\end{aligned}$$

as there are 2^l indices in $I^{l+1}(i) \setminus I^l(i)$. Consequently according to Lemma 5 we have that for any $q \ge 1$,

$$\left\|\mathbb{E}_\xi\left[|g_i^l - g_i^{l+1}|\right]\right\|_q = \left\|\left\|\mathbb{E}_\xi\, |g_i^l - g_i^{l+1}|\right\|_q\left(Z_i, S_{\overline{I^{l+1}(i)}}\right)\right\|_q \le 3\beta\sqrt{q2^l}, \tag{22}$$

and

$$\left\|\mathbb{E}_\xi\left[\left(g_i^l - g_i^{l+1}\right)^2\right]\right\|_q = \left\|\left\|\mathbb{E}_\xi\left[\left(g_i^l - g_i^{l+1}\right)^2\right]\right\|_q\left(Z_i, S_{\overline{I^{l+1}(i)}}\right)\right\|_q \le 68\beta^2 2^l q. \tag{23}$$

Now consider any $I^l \in \mathcal{I}_l$. Since for each $i \in I_l$, $g_i^l - g_i^{l+1}$ depends only on $(Z_i, S_{\overline{I^l}}, \xi)$, these random terms are essentially of the form $f(Z_i, \xi)$ conditioned on $(S_{\overline{I^l}})$. Given the assumption $\mathbb{E}\left[g_i(S,\xi) \mid S \setminus Z_i, \xi\right] = 0$, it holds that $\mathbb{E}\left[g_i^l - g_i^{l+1} \mid S_{\overline{I^l}}, \xi\right] = 0$, $\forall i \in I_l$. Therefore, applying Lemma 9 yields

$$\begin{aligned}
&\left\|\mathbb{E}_\xi\left[\left|\sum_{i\in I^l} g_i^l - g_i^{l+1}\right|\right]\right\|_q (S_{\overline{I^l}})\\
\le&\sqrt{2\sum_{i=1}^{N}\mathbb{E}\left[\left(g_i^l - g_i^{l+1}\right)^2 \mid S_{\overline{I^l}}\right]} + 2\sqrt{3q}\sqrt{\left\|\sum_{i=1}^{N}\mathbb{E}_\xi^2\left[|g_i^l - g_i^{l+1}|\right]\right\|_{q/2}(S_{\overline{I^l}})}.
\end{aligned}$$

Next we proceed to bound the moment $\left\|\mathbb{E}_\xi\left[\left|\sum_{i\in I^l} g_i^l - g_i^{l+1}\right|\right]\right\|_q$. By applying Lemma 5 again and using the above bound we can show that

$$
\begin{aligned}
&\left\|\mathbb{E}_\xi\left[\left|\sum_{i\in I^l} g_i^l - g_i^{l+1}\right|\right]\right\|_q = \left\|\left\|\mathbb{E}_\xi\left[\left|\sum_{i\in I^l} g_i^l - g_i^{l+1}\right|\right]\right\|_q (S_{\overline{I^l}})\right\|_q \\
\leq &\left\|\sqrt{2\sum_{i\in I^l}\mathbb{E}\left[\left(g_i^l - g_i^{l+1}\right)^2 \mid S_{\overline{I^l}}\right]} + 2\sqrt{3q}\sqrt{\left\|\sum_{i\in I^l}\mathbb{E}_\xi^2\left[|g_i^l - g_i^{l+1}|\right]\right\|_{q/2}(S_{\overline{I^l}})}\right\|_q \\
\leq &\left\|\sqrt{2\sum_{i\in I^l}\mathbb{E}\left[\left(g_i^l - g_i^{l+1}\right)^2 \mid S_{\overline{I^l}}\right]}\right\|_q + 2\sqrt{3q}\left\|\sqrt{\left\|\sum_{i\in I^l}\mathbb{E}_\xi^2\left[|g_i^l - g_i^{l+1}|\right]\right\|_{q/2}(S_{\overline{I^l}})}\right\|_q \\
\overset{\zeta_1}{=} &\sqrt{2\left\|\sum_{i\in I^l}\mathbb{E}\left[\left(g_i^l - g_i^{l+1}\right)^2 \mid S_{\overline{I^l}}\right]\right\|_{q/2}} + 2\sqrt{3q}\sqrt{\left\|\left\|\sum_{i\in I^l}\mathbb{E}_\xi^2\left[|g_i^l - g_i^{l+1}|\right]\right\|_{q/2}(S_{\overline{I^l}})\right\|_{q/2}} \\
\leq &\sqrt{2\sum_{i\in I^l}\left\|\mathbb{E}\left[\left(g_i^l - g_i^{l+1}\right)^2 \mid S_{\overline{I^l}}\right]\right\|_{q/2}} + 2\sqrt{3q}\sqrt{\sum_{i\in I^l}\left\|\mathbb{E}_\xi^2\left[|g_i^l - g_i^{l+1}|\right]\right\|_{q/2}} \\
\overset{\zeta_2}{\leq} &\sqrt{2\sum_{i\in I^l}\left\|\mathbb{E}_\xi\left[\left(g_i^l - g_i^{l+1}\right)^2\right]\right\|_{q/2}} + 2\sqrt{3q}\sqrt{\sum_{i\in I^l}\left\|\mathbb{E}_\xi\left[|g_i^l - g_i^{l+1}|\right]\right\|_q^2} \\
\overset{\zeta_3}{\leq} &\sqrt{2\times 2^l \times 34\beta^2 2^l q} + 2\sqrt{3q}\sqrt{2^l\times 9\beta^2 2^l q} = 2^l\beta\left(2\sqrt{17q} + 6\sqrt{3q}\right) \leq 2^l \times 19\beta q,
\end{aligned}
$$

where "ζ_1" is due to (10), in "ζ_2" we have used Lemma 5, and in "ζ_3" we have used (22) and (23). Then based on the triangle inequality we get

$$
\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N g_i^l - g_i^{l+1}\right|\right]\right\|_q \leq \sum_{I^l\in\mathcal{I}_l}\left\|\mathbb{E}_\xi\left[\left|\sum_{i\in I^l} g_i^l - g_i^{l+1}\right|\right]\right\|_q \leq 2^{k-l}\times 2^l \times 19\beta q \leq 38N\beta q.
$$

Therefore the r.h.s. of (21) can be bounded as

$$
\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N (g_i - \mathbb{E}[g_i \mid Z_i,\xi])\right|\right]\right\|_q \leq \sum_{l=0}^{k-1}\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N g_i^l - g_i^{l+1}\right|\right]\right\|_q \leq 38N\lceil\log_2 N\rceil\beta q. \tag{24}
$$

Based on (24) and the triangle inequality we have

$$
\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N g_i\right|\right]\right\|_q \leq \left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N \mathbb{E}[g_i \mid Z_i,\xi]\right|\right]\right\|_q + 38N\lceil\log_2 N\rceil\beta q. \tag{25}
$$

Let $f_i(Z_i,\xi) = \mathbb{E}[g_i(S,\xi) \mid Z_i,\xi]$. We must have $|f_i| \leq M$ and $\mathbb{E}[f_i(Z_i,\xi) \mid \xi] = 0$ as $|\mathbb{E}[g_i \mid Z_i,\xi]| \leq M$ and $\mathbb{E}[g_i \mid S\setminus Z_i,\xi] = 0$. Then it follows from Lemma 9 that the first term at the right hand side of (25) can be bounded as

$$
\begin{aligned}
&\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N \mathbb{E}[g_i \mid Z_i,\xi]\right|\right]\right\|_q \\
=&\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N f_i(Z_i,\xi)\right|\right]\right\|_q \leq M\sqrt{2N} + 2M\sqrt{3Nq} \leq 3M\sqrt{3Nq}.
\end{aligned} \tag{26}
$$

Finally, the desired bound is obtained by plugging (26) into (25). □

A.4 Proof of Theorem 1

Proof. Let us consider $g_i(S,\xi) = \mathbb{E}_{Z_i'}\left[R(A(S^{(i)},\xi)) - \ell(A(S^{(i)},\xi);Z_i)\right]$. Then the L_q-norm of the on-average generalization gap can be upper bounded by the triangle inequality as

$$\begin{aligned}
&\left\|\mathbb{E}_\xi\left[|R(A(S,\xi)) - R_S(A(S,\xi))|\right]\right\|_q = \frac{1}{N}\left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N R(A(S,\xi)) - \ell(A(S,\xi);Z_i)\right|\right]\right\|_q \\
\leq &\frac{1}{N}\left(\underbrace{\left\|\mathbb{E}\left[\left|\sum_{i=1}^N g_i(S,\xi)\right|\right]\right\|_q}_{T_1} + \underbrace{\left\|\mathbb{E}\left[\left|\sum_{i=1}^N (R(A(S,\xi)) - \ell(A(S,\xi);Z_i) - g_i(S,\xi))\right|\right]\right\|_q}_{T_2}\right).
\end{aligned} \tag{27}$$

We next respectively upper bound the two terms T_1 and T_2 in (27). To bound the term T_1, by definition it holds that $\mathbb{E}[g_i(S,\xi) \mid S \setminus Z_i, \xi] = 0$. Based on the triangle inequality and Jensen's inequality we have that for any $i \in [N]$,

$$|\mathbb{E}[g_i(S,\xi) \mid Z_i,\xi]| \leq \mathbb{E}[\ell(A(S^{(i)},\xi);Z) \mid \xi] + \mathbb{E}[\ell(A(S^{(i)},\xi);Z_i) \mid Z_i,\xi] \leq 2M.$$

Further we show that g_i satisfies the L_2-uniform bounded difference property with respect to all variables in S except Z_i. Indeed, for each $j \neq i$ and conditioned on S, Z_j' it can be verified that

$$\begin{aligned}
&\left\|g_i(S,\xi) - g_i(S^{(j)},\xi)\right\|_2 (S, Z_j') \\
\leq &\left\|\mathbb{E}_{Z_i'}\left[R(A(S^{(i)},\xi)) - R(A((S^{(i)})^{(j)},\xi))\right]\right\|_2 (S, Z_j') \\
&+ \left\|\mathbb{E}_{Z_i'}\left[\ell(A(S^{(i)},\xi);Z_i) - \ell(A((S^{(i)})^{(j)},\xi);Z_i)\right]\right\|_2 (S, Z_j') \\
= &\left\|\mathbb{E}_{Z_i'}\mathbb{E}_Z\left[\ell(A(S^{(i)},\xi);Z) - \ell(A((S^{(i)})^{(j)},\xi);Z)\right]\right\|_2 (S, Z_j') \\
&+ \left\|\mathbb{E}_{Z_i'}\left[\ell(A(S^{(i)},\xi);Z_i) - \ell(A((S^{(i)})^{(j)},\xi);Z_i)\right]\right\|_2 (S, Z_j') \\
\leq &\sup_{Z_i',Z}\left\|\ell(A(S^{(i)},\xi);Z) - \ell(A((S^{(i)})^{(j)},\xi);Z)\right\|_2 (S, Z_i', Z_j', Z) \\
&+ \sup_{Z_i'}\left\|\ell(A(S^{(i)},\xi);Z_i) - \ell(A((S^{(i)})^{(j)},\xi);Z_i)\right\|_2 (S, Z_i', Z_j') \\
\leq &\sup_{S^{(i)},Z_j',Z}\left\|\ell(A(S^{(i)},\xi);Z) - \ell(A((S^{(i)})^{(j)},\xi);Z)\right\|_2 (S^{(i)}, Z_j', Z) \\
&+ \sup_{S^{(i)},Z_j',Z_i}\left\|\ell(A(S^{(i)},\xi);Z_i) - \ell(A((S^{(i)})^{(j)},\xi);Z_i)\right\|_2 (S^{(i)}, Z_j', Z_i) \\
\leq &2\gamma_{\mathrm{L}_2,N},
\end{aligned}$$

where we have frequently used the fact that expectation is always no larger than supreme, and in the last equality we have used the L_2-uniform stability assumption on the algorithm A. Therefore, $\{g_i\}$ satisfy the conditions of Proposition 2 and thus

$$T_1 = \left\|\mathbb{E}_\xi\left[\left|\sum_{i=1}^N g_i(S,\xi)\right|\right]\right\|_q \leq 6M\sqrt{3Nq} + 76N\lceil\log_2 N\rceil\gamma_{\mathrm{L}_2,N}\, q. \tag{28}$$

Now we proceed to bound the second term T_2. It can be verified that

$$
\begin{aligned}
T_2 \leq & \left\| \mathbb{E}_\xi \left[\left| \sum_{i=1}^N \mathbb{E}_{Z_i'} \left[R(A(S,\xi)) - R(A(S^{(i)},\xi)) \right] \right| \right] \right\|_q \\
& + \left\| \mathbb{E}_\xi \left[\left| \sum_{i=1}^N \mathbb{E}_{Z_i'} \left[\ell(A(S,\xi); Z_i) - \ell(A(S^{(i)},\xi); Z_i) \right] \right| \right] \right\|_q \\
= & \left\| \mathbb{E}_\xi \left[\left| \sum_{i=1}^N \mathbb{E}_{Z_i'} \mathbb{E}_Z \left[\ell(A(S,\xi); Z) - \ell(A(S^{(i)},\xi); Z) \right] \right| \right] \right\|_q \\
& + \left\| \mathbb{E}_\xi \left[\left| \sum_{i=1}^N \mathbb{E}_{Z_i'} \left[\ell(A(S,\xi); Z_i) - \ell(A(S^{(i)},\xi); Z_i) \right] \right| \right] \right\|_q \\
\leq & 2 \sum_{i=1}^N \left(\sup_{S, Z_i', Z} \mathbb{E}_\xi \left[\left| \ell(A(S,\xi); Z) - \ell(A(S^{(i)},\xi); Z) \right| \right] \right) \\
\leq & 2N \gamma_{\mathrm{L}_2,N},
\end{aligned} \tag{29}
$$

where in the last equality we have used the L_2-uniform stability assumption. Plugging bounds (28) and (29) into (27) and preserving leading terms yields

$$
\left\| \mathbb{E}_\xi \left[|R(A(S,\xi)) - R_S(A(S,\xi))| \right] \right\|_q \leq 6M\sqrt{\frac{3q}{N}} + 77\lceil \log_2 N \rceil \gamma_{\mathrm{L}_2,N} q.
$$

According to the equivalence of tails and moments as shown in Lemma 7, the above moment bound immediately implies the desired exponential generalization bound. □

B Proofs for Section 3

In this section, we present the technical proofs for the main results stated in Section 3.

B.1 Proof of Theorem 2

We first establish the following intermediate result that captures the effects of bagging on randomized algorithms: it basically tells that with $K \gtrsim \log(\frac{1}{\delta})$, at least one of the solutions generated by bagging generalizes well with high probability.

Lemma 11. *Suppose that a randomized learning algorithm $A : \mathcal{Z}^N \times \mathcal{R} \mapsto \mathcal{W}$ has L_2-uniform stability with parameter $\gamma_{L_2,N}$. Assume that the loss function ℓ is ranged in $[0, M]$. Then for any $\delta \in (0,1)$ and $K \geq 2\log(2/\delta)$, with probability at least $1-\delta$ over the randomness of S_1 and $\{\xi_k\}_{k\in[K]}$, the sequence $\{A(S_1,\xi_k)\}_{k\in[K]}$ generated by Algorithm 1 satisfies*

$$
\min_{k\in[K]} |R(A(S_1,\xi_k)) - R_{S_1}(A(S_1,\xi_k))| \lesssim \gamma_{L_2,(1-\mu)N} \log(N) \log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{(1-\mu)N}}.
$$

Proof. For any data set S, let us define $h(S) := \mathbb{E}_\xi \left[|R(A(S,\xi)) - R_S(A(S,\xi))| \right]$. From Theorem 1 we have that with probability at least $1-\delta$ over S_1,

$$
h(S_1) \lesssim \gamma_{\mathrm{L}_2,(1-\mu)N} \log((1-\mu)N) \log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{(1-\mu)N}}. \tag{30}
$$

Let us now consider the following defined events:

$$
\mathcal{E}: \ \min_{k\in[K]} |R(A(S_1,\xi_k)) - R_{S_1}(A(S_1,\xi_k))| \lesssim \gamma_{\mathrm{L}_2,(1-\mu)N} \log((1-\mu)N) \log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{(1-\mu)N}},
$$

$$
\mathcal{E}_0: \ h(S_1) \lesssim \gamma_{\mathrm{L}_2,(1-\mu)N} \log(N) \log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{(1-\mu)N}},
$$

$$
\mathcal{E}_k: \ |R(A(S_1,\xi_k)) - R_{S_1}(A(S_1,\xi_k))| \leq 2h(S_1), \ \ k = 1, ..., K.
$$

We first show that $\mathbb{P}\left\{\bigcap_{k\in[K]}\overline{\mathcal{E}_k}\right\} \le \frac{\delta}{2}$. To this end, for each k, let us consider the random indication function $g(S_1, \xi_k) := \mathbf{1}_{\{\overline{\mathcal{E}_k}\}}$ where $\mathbf{1}_{\{C\}}$ is the indicator function of the condition C. Then we can show that

$$\begin{aligned}
\mathbb{P}\left\{\bigcap_{k=1}^{K}\overline{\mathcal{E}_k}\right\} =& \mathbb{E}\left[\prod_{k=1}^{K} g(S_1, \xi_k)\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\prod_{k=1}^{K} g(S_1, \xi_k) \mid S_1\right]\right] \\
=& \mathbb{E}\left[\prod_{k=1}^{K} \mathbb{E}\left[g(S_1, \xi_k) \mid S_1\right]\right] \\
=& \mathbb{E}\left[\prod_{k=1}^{K} \mathbb{P}\left\{|R(A(S_1, \xi_k)) - R_{S_1}(A(S_1, \xi_k))| > 2h(S_1)\right\} \mid S_1\right] \\
\le& \mathbb{E}\left[\left(\frac{1}{2}\right)^K \mid S_1\right] = \left(\frac{1}{2}\right)^K \le \frac{\delta}{2},
\end{aligned}$$

where we have used the independence among $\{\xi_k\}$ and S_1, Markov inequality and the condition on K as well. From the high-probability bound (30) we have $\mathbb{P}\left\{\overline{\mathcal{E}_0}\right\} \le \frac{\delta}{2}$. Combining this bound and the preceding bound yields

$$\mathbb{P}\left\{\mathcal{E}\right\} \ge \mathbb{P}\left\{\left(\bigcup_{k=1}^{K}\mathcal{E}_k\right)\bigcap\mathcal{E}_0\right\} \ge 1 - \mathbb{P}\left\{\bigcap_{k=1}^{K}\overline{\mathcal{E}_k}\right\} - \mathbb{P}\left\{\overline{\mathcal{E}_0}\right\} \ge 1 - \frac{\delta}{2} - \frac{\delta}{2} = 1 - \delta.$$

This proves the desired bound. □

With Lemma 11 in place, we are ready to prove the main result in Theorem 2.

Proof of Theorem 2. The key idea is to show that the proposed greedy model selection strategy guarantees that the selected $A(S, \xi_{k^*})$ mimics the generalization behavior of the best performer among the K candidates. To do so, let us consider the following three events:

$$\begin{aligned}
&\mathcal{E}: |R(A(S_1, \xi_{k^*})) - R_S(A(S_1, \xi_{k^*}))| \lesssim \gamma_{\mathrm{L}_2,(1-\mu)N}\log(N)\log\left(\frac{1}{\delta}\right) + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(K/\delta)}{N}}, \\
&\mathcal{E}_1: \max_{k\in[K]} |R(A(S_1, \xi_k)) - R_{S_2}(A(S_1), \xi_k)| \le M\sqrt{\frac{\log(2K/\delta)}{2\mu N}}, \\
&\mathcal{E}_2: \min_{k\in[K]} |R(A(S_1, \xi_k)) - R_{S_1}(A(S_1, \xi_k))| \lesssim \gamma_{\mathrm{L}_2,(1-\mu)N}\log(N)\log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{(1-\mu)N}}.
\end{aligned}$$

We first show that $\mathcal{E} \supseteq \mathcal{E}_1 \cap \mathcal{E}_2$. Indeed, suppose that $\mathcal{E}_1$ and $\mathcal{E}_2$ simultaneously occur. Consequently the following inequality can be verified:

$$
\begin{aligned}
& |R(A(S_1,\xi_{k^*})) - R_S(A(S_1,\xi_{k^*}))| \\
=& |R(A(S_1,\xi_{k^*})) - (1-\mu)R_{S_1}(A(S_1,\xi_{k^*})) - \mu R_{S_2}(A(S_1,\xi_{k^*}))| \\
\le& (1-\mu)\,|R(A(S_1,\xi_{k^*})) - R_{S_1}(A(S_1,\xi_{k^*}))| + \mu\,|R(A(S_1,\xi_{k^*})) - R_{S_2}(A(S_1,\xi_{k^*}))| \\
\le& (1-\mu)\,|R_{S_2}(A(S_1,\xi_{k^*})) - R_{S_1}(A(S_1,\xi_{k^*}))| + |R(A(S_1,\xi_{k^*})) - R_{S_2}(A(S_1,\xi_{k^*}))| \\
\overset{\zeta_1}{=}& (1-\mu)\min_{k\in[K]} |R_{S_2}(A(S_1,\xi_k)) - R_{S_1}(A(S_1,\xi_k))| + |R(A(S_1,\xi_{k^*})) - R_{S_2}(A(S_1,\xi_{k^*}))| \\
=& (1-\mu)\min_{k\in[K]} |R_{S_2}(A(S_1,\xi_k)) - R(A(S_1,\xi_k)) + R(A(S_1,\xi_k)) - R_{S_1}(A(S_1,\xi_k))| \\
& + |R(A(S_1,\xi_{k^*})) - R_{S_2}(A(S_1,\xi_{k^*}))| \\
\le& (1-\mu)\min_{k\in[K]} |R(A(S_1,\xi_k)) - R_{S_1}(A(S_1,\xi_k))| + (1-\mu)\max_{k\in[K]} |R_{S_2}(A(S_1,\xi_k)) - R(A(S_1,\xi_k))| \\
& + |R(A(S_1,\xi_{k^*})) - R_{S_2}(A(S_1,\xi_{k^*}))| \\
\le& \min_{k\in[K]} |R(A(S_1,\xi_k)) - R_{S_1}(A(S_1,\xi_k))| + 2\max_{k\in[K]} |R_{S_2}(A(S_1,\xi_k)) - R(A(S_1,\xi_k))| \\
\overset{\zeta_2}{\lesssim}& \gamma_{\mathrm{L}_2,(1-\mu)N}\log(N)\log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{(1-\mu)N}} + M\sqrt{\frac{\log(K/\delta)}{\mu N}},
\end{aligned}
$$

where in "ζ_1" we have used the definition of k^*, and "ζ_2" follows from $\mathcal{E}_1, \mathcal{E}_2$. After some algebraic manipulation with leading terms preserved in the above we can see that $\mathcal{E}$ occurs.

Next we aim to show that $\mathbb{P}\{\overline{\mathcal{E}_1}\} \le \frac{\delta}{2}$. To this end, let us consider the random indication function $g(S,\{\xi_k\}) := \mathbf{1}_{\{\overline{\mathcal{E}_1}\}}$ associated with $\overline{\mathcal{E}_1}$. Then we have that

$$
\begin{aligned}
\mathbb{P}\left\{\overline{\mathcal{E}_1}\right\} =& \mathbb{E}\left[g(S,\{\xi_k\})\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[g(S,\{\xi_k\}) \mid S_1, \{\xi_k\}\right]\right] \\
=& \mathbb{E}\left[\mathbb{P}\left\{\max_{k\in[K]} |R(A(S_1,\xi_k)) - R_{S_2}(A(S_1,\xi_k))| \ge M\sqrt{\frac{\log(2K/\delta)}{2\mu N}}\right\} \mid S_1, \{\xi_k\}\right] \\
\overset{\zeta_1}{\le}& \mathbb{E}\left[\frac{\delta}{2} \mid S_1, \{\xi_k\}\right] = \frac{\delta}{2},
\end{aligned}
$$

where in "ζ_1" we have used Hoeffding's inequality and union bound, keeping in mind the independence among $\{A_k\}$, S_1 and S_2. Further, from Lemma 11 we have $\mathbb{P}\left\{\overline{\mathcal{E}_2}\right\} \le \frac{\delta}{2}$. Combining this and the preceding bound yields

$$
\mathbb{P}\{\mathcal{E}\} \ge \mathbb{P}_{S,\{A_k\}}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \ge 1 - \mathbb{P}\left\{\overline{\mathcal{E}_1}\right\} - \mathbb{P}\left\{\overline{\mathcal{E}_2}\right\} \ge 1 - \frac{\delta}{2} - \frac{\delta}{2} = 1-\delta.
$$

The proof is concluded. □

C Proofs for Section 4

In this section, we present the technical proofs for the main results stated in Section 4.

C.1 Auxiliary lemmas

We need the following lemma from Hardt et al. (2016) which shows that SGD iteration is non-expansive for convex and smooth losses.

Lemma 12 (Hardt et al. (2016))**.** *Let $f : \mathcal{W} \mapsto \mathbb{R}$ be a convex and L-smooth function. Then for any $w, \tilde{w} \in \mathcal{W}$ and $\alpha \le 2/L$, we have the following bound holds*

$$
\|w - \alpha\nabla f(w) - (\tilde{w} - \alpha\nabla f(\tilde{w}))\| \le \|w - \tilde{w}\|.
$$

The following lemma, which can be proved by induction (see, e.g., Schmidt et al., 2011), will be used to prove the main results in Section 4.

Lemma 13. *Assume that the nonnegative sequence $\{u_\tau\}_{\tau\ge 1}$ satisfies the following recursion for all $t \ge 1$:*

$$u_t^2 \le S_t + \sum_{\tau=1}^{t} \alpha_\tau u_\tau,$$

with $\{S_\tau\}_{\tau\ge1}$ an increasing sequence, $S_0 \ge u_0^2$ and $\alpha_\tau \ge 0$ for all τ. Then, the following inequality holds for all $t \ge 1$:

$$u_t \le \sqrt{S_t} + \sum_{\tau=1}^{t} \alpha_\tau.$$

For analyzing SGD with convex and non-smooth loss functions, we need the following lemma by Bassily et al. (2020, Lemma 3.1) that quantifies the deviation between the online gradient descent trajectories.

Lemma 14 (Bassily et al. (2020))**.** *Consider the two sequences $\{w_t\}_{t\ge0}$ and $\{\tilde{w}_t\}_{t\ge0}$ generated according to the following recursions respectively over the convex and G-Lipschitz objectives $\{f_t\}_{t\ge0}$ and $\{\tilde{f}_t\}_{t\ge0}$ from $w_0 = \tilde{w}_0$:*

$$\begin{aligned} w_t =& \Pi_{\mathcal{W}}\left(w_{t-1} - \eta_t \nabla f_{t-1}(w_{t-1})\right) \\ \tilde{w}_t =& \Pi_{\mathcal{W}}\left(\tilde{w}_{t-1} - \eta_t \nabla \tilde{f}_{t-1}(\tilde{w}_{t-1})\right). \end{aligned}$$

Let $t_0 := \inf\{t : f_t \neq f'_t\}$ and $\beta_t := \mathbf{1}_{\{f_t \neq f'_t\}}$. Then for any $T \ge 1$,

$$\|w_T - \tilde{w}_T\| \le 2G\sqrt{\sum_{t=t_0}^{T-1} \eta_t^2} + 4G \sum_{t=t_0+1}^{T-1} \eta_t \beta_t.$$

C.2 Proof of Lemma 1

Here we prove Lemma 1 that establishes the L_2-uniform stability of $A_{\texttt{SGD-w}}$ with convex and smooth loss functions.

Proof. Given any pair of data sets $S, \tilde{S}$ that differ in a single element, let us define the sequences $\{w_t\}_{t\in[T]}$ and $\{\tilde{w}_t\}_{t\in[T]}$ that are respectively generated over S and $\tilde{S}$ via $A_{\texttt{SGD-w}}$ via sample path $\xi = \{i_t\}_{t\in[T]}$. Note by assumption that $w_0 = \tilde{w}_0$. We distinguish the following two complementary cases.

Case I: $Z_{i_t} = \tilde{Z}_{i_t}$. In this case, by invoking Lemma 12 we get

$$\begin{aligned} \|w_t - \tilde{w}_t\|^2 =& \|\Pi_{\mathcal{W}}(w_{t-1} - \eta_t \nabla_w \ell(w_{t-1}; Z_{i_t})) - \Pi_{\mathcal{W}}(\tilde{w}_{t-1} - \eta_t \nabla_w \ell(\tilde{w}_{t-1}; \tilde{Z}_{i_t}))\|^2 \\ \le& \|w_{t-1} - \eta_t \nabla_w \ell(w_{t-1}; Z_{i_t}) - (\tilde{w}_{t-1} - \eta_t \nabla_w \ell(\tilde{w}_{t-1}; Z_{i_t}))\|^2 \\ \le& \|w_{t-1} - \tilde{w}_{t-1}\|^2. \end{aligned} \tag{31}$$

Case II: $Z_{i_t} \neq \tilde{Z}_{i_t}$. In this case, we have

$$\begin{aligned} \|w_t - \tilde{w}_t\|^2 =& \|\Pi_{\mathcal{W}}(w_{t-1} - \eta_t \nabla_w \ell(w_{t-1}; Z_{i_t})) - \Pi_{\mathcal{W}}(\tilde{w}_{t-1} - \eta_t \nabla_w \ell(\tilde{w}_{t-1}; \tilde{Z}_{i_t}))\|^2 \\ \le& \|w_{t-1} - \eta_t \nabla_w \ell(w_{t-1}; Z_{i_t}) - (\tilde{w}_{t-1} - \eta_t \nabla_w \ell(\tilde{w}_{t-1}; \tilde{Z}_{i_t}))\|^2 \\ \le& \Big(\|w_{t-1} - \tilde{w}_{t-1}\| + \eta_t(\|\nabla_w \ell(w_{t-1}; Z_{i_t})\| + \|\nabla_w \ell(\tilde{w}_{t-1}; \tilde{Z}_{i_t})\|)\Big)^2 \\ \le& \left(\|w_{t-1} - \tilde{w}_{t-1}\| + 2G\eta_t\right)^2 \\ =& \|w_{t-1} - \tilde{w}_{t-1}\|^2 + 4G\eta_t\|w_{t-1} - \tilde{w}_{t-1}\| + 4G^2\eta_t^2, \end{aligned} \tag{32}$$

where in the last but inequality we have used $\ell(\cdot;\cdot)$ is G-Lipschitz with respect to its first argument.

Let $\beta_t = \beta_t(S, \tilde{S}, i_t) := \mathbf{1}_{\{Z_{i_t} \neq \tilde{Z}_{i_t}\}}$ be the random indication function associated with the event $Z_{i_t} \neq \tilde{Z}_{i_t}$. Based on the recursion forms (31) and (32) and the condition $w_0 = \tilde{w}_0$ we can show that for all $t \geq 1$,

$$\|w_t - \tilde{w}_t\|^2 \leq \sum_{\tau=1}^{t} 4G\beta_\tau \eta_\tau \|w_{\tau-1} - \tilde{w}_{\tau-1}\| + \sum_{\tau=1}^{t} 4G^2 \beta_\tau \eta_\tau^2.$$

Then applying Lemma 13 with simple algebraic manipulation yields

$$\|w_t - \tilde{w}_t\|^2 \leq 8G^2 \left(\sum_{\tau=1}^{t} \beta_\tau \eta_\tau^2 + 4 \left(\sum_{\tau=1}^{t} \beta_\tau \eta_\tau \right)^2 \right).$$

Since by assumption S and $\tilde{S}$ differ only in a single element, under the scheme of uniform sampling without replacement, we can see that $\beta_t \sim \texttt{Bernoulli}(1/N)$ and $\{\beta_t\}_{t\in[T]}$ is an i.i.d. sequence of Bernoulli random variables. It follows that

$$\begin{aligned}
&\mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right] \\
\leq& 8G^2 \left(\sum_{\tau=1}^{t} \mathbb{E}\left[\beta_\tau\right] \eta_\tau^2 + 4\mathbb{E}\left[\left(\sum_{\tau=1}^{t} \beta_\tau \eta_\tau \right)^2 \right] \right) \\
=& 8G^2 \left(\sum_{\tau=1}^{t} \mathbb{E}\left[\beta_\tau + 4\beta_\tau^2\right] \eta_\tau^2 + 4 \sum_{\tau \neq \tau'} 1\mathbb{E}\left[\beta_\tau \beta_{\tau'}\right] \eta_\tau \eta_{\tau'} \right) \\
=& 8G^2 \left(\frac{5}{N} \sum_{\tau=1}^{t} \eta_\tau^2 + \frac{4}{N^2} \left(\sum_{\tau=1}^{t} \eta_\tau \right)^2 \right) \leq 40G^2 \left(\frac{1}{N} \sum_{\tau=1}^{T} \eta_\tau^2 + \frac{1}{N^2} \left(\sum_{\tau=1}^{T} \eta_\tau \right)^2 \right),
\end{aligned}$$

where we have used $\mathbb{E}[\beta_t] = \mathbb{E}[\beta_t^2] = \frac{1}{N}$. The convexity of squared Euclidean norm leads to

$$\mathbb{E}\left[\|\bar{w}_T - \tilde{\bar{w}}_T\|^2\right] \leq \frac{\sum_{t=1}^{T} \mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right]}{T} \leq 40G^2 \left(\frac{1}{N} \sum_{t=1}^{T} \eta_t^2 + \frac{1}{N^2} \left(\sum_{t=1}^{T} \eta_t \right)^2 \right).$$

For each $i \in [N]$, let $\left\{w_t^{(i)}\right\}_{t\in[T]}$ be the sequence generated over $S^{(i)}$ by $A_{\texttt{SGD-w}}$. Since the above holds for any $S \doteq \tilde{S}$, we must have

$$\sup_{S, Z_i'} \mathbb{E}_\xi \left[\left\| \bar{w}_T - \bar{w}_T^{(i)} \right\|^2 \right] \leq 40G^2 \left(\frac{1}{N} \sum_{t=1}^{T} \eta_t^2 + \frac{1}{N^2} \left(\sum_{t=1}^{T} \eta_t \right)^2 \right).$$

Finally, since the loss is G-Lipschitz, it follows from the above that for all $i \in [N]$,

$$\sup_{S, Z_i', Z} \mathbb{E}_\xi \left[\left(\ell(\bar{w}_T; Z) - \ell(\bar{w}_T^{(i)}; Z) \right)^2 \right] \leq 40G^4 \left(\frac{1}{N} \sum_{t=1}^{T} \eta_t^2 + \frac{1}{N^2} \left(\sum_{t=1}^{T} \eta_t \right)^2 \right).$$

This proves the desired L_2-uniform stability of algorithm. □

C.3 Proof of Lemma 2

In this subsection we prove Lemma 2 that establishes the L_2-uniform stability of $A_{\texttt{SGD-w}}$ with convex and non-smooth loss functions.

Proof. The proof arguments follow closely those of Lemma 1. Here we reproduce the proof for the sake of completeness. Let us define the sequences $\{w_t\}_{t\in[T]}$ and $\{\tilde{w}_t\}_{t\in[T]}$ that are respectively generated over S and $\tilde{S}$ by $A_{\texttt{SGD-w}}$ via sample path $\xi = \{i_t\}_{t\in[T]}$. Suppose that $S \doteq \tilde{S}$ and consider a hitting time variable $t_0 = \inf\{t : Z_{i_t} \neq \tilde{Z}_{i_t}\}$. Let $\beta_t = \beta_t(S, \tilde{S}, i_t) := \mathbf{1}_{\{Z_{i_t} \neq \tilde{Z}_{i_t}\}}$ be the random indication function associated with event $Z_{i_t} \neq \tilde{Z}_{i_t}$. Then $\{\beta_t\}_{t\in[T]}$ is an i.i.d. sequence of

$\texttt{Bernoulli}(1/N)$ random variables. Conditioned on t_0, it has been shown by Bassily et al. (2020) (see Lemma 14) that

$$\|w_t - \tilde{w}_t\| \le 2G\sqrt{\sum_{\tau=t_0}^{t}\eta_\tau^2} + 4G\sum_{\tau=t_0+1}^{t}\beta_\tau\eta_\tau \le 2G\sqrt{\sum_{\tau=1}^{t}\eta_\tau^2} + 4G\sum_{\tau=1}^{t}\beta_\tau\eta_\tau. \tag{33}$$

Given S and $\tilde{S}$, based on the square of the bound (33) we can show that

$$\begin{aligned}
\mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right] \le& \mathbb{E}\left[8G^2\sum_{\tau=1}^{t}\eta_\tau^2 + 32G^2\left(\sum_{\tau=1}^{t}\beta_\tau\eta_\tau\right)^2\right] \\
=& 8G^2\sum_{\tau=1}^{t}\eta_\tau^2 + 32G^2\mathbb{E}\left[\sum_{\tau=1}^{t}\beta_\tau^2\eta_\tau^2 + \sum_{\tau\neq\tau'}\beta_\tau\beta_{\tau'}\eta_\tau\eta_{\tau'}\right] \\
=& 8G^2\sum_{\tau=1}^{t}\eta_\tau^2 + 32G^2\left(\frac{1}{N}\sum_{\tau=1}^{t}\eta_\tau^2 + \frac{1}{N^2}\sum_{\tau\neq\tau'}\eta_\tau\eta_{\tau'}\right) \\
\le& 40G^2\sum_{\tau=1}^{t}\eta_\tau^2 + \frac{32G^2}{N^2}\left(\sum_{\tau=1}^{t}\eta_\tau\right)^2,
\end{aligned}$$

where we have used $\mathbb{E}[\beta_t] = \mathbb{E}[\beta_t^2] = \frac{1}{N}$. It follows directly from the convexity of squared loss that

$$\mathbb{E}\left[\|\bar{w}_T - \tilde{\bar{w}}_T\|^2\right] \le 40G^2\sum_{t=1}^{T}\eta_t^2 + \frac{32G^2}{N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2.$$

Since the above holds for any pair of $S \doteq \tilde{S}$, we have that for all $i \in [N]$,

$$\sup_{S,Z_i'}\mathbb{E}_\xi\left[\left\|\bar{w}_T - \bar{w}_T^{(i)}\right\|^2\right] \le 40G^2\sum_{t=1}^{T}\eta_t^2 + \frac{32G^2}{N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2,$$

where $\{w_t^{(i)}\}_{t\in[T]}$ is generated over $S^{(i)}$ by $A_{\texttt{SGD-w}}$. Finally, since the loss is G-Lipschitz, it follows from the above bound that for all $i \in [N]$,

$$\sup_{S,Z_i',Z}\mathbb{E}_\xi\left[\left(\ell(\bar{w}_T;Z) - \ell(\bar{w}_T^{(i)};Z)\right)^2\right] \le 40G^4\sum_{t=1}^{T}\eta_t^2 + \frac{32G^4}{N^2}\left(\sum_{t=1}^{T}\eta_t\right)^2.$$

This proves the desired L_2-uniform stability of algorithm. □

C.4 Proof of Lemma 3

In this subsection we prove Lemma 3 which establishes the L_2-uniform stability of $A_{\texttt{SGD-w}}$ with non-convex and smooth loss functions.

Proof. Let us define the sequences $\{w_t\}_{t\in[T]}$ and $\{\tilde{w}_t\}_{t\in[T]}$ that are respectively generated over S and $\tilde{S}$ by $A_{\texttt{SGD-w}}$ via sample path $\xi = \{i_t\}_{t\in[T]}$. Suppose that $S \doteq \tilde{S}$. Let us consider $\Delta_t := \mathbb{E}\left[\|w_t - \tilde{w}_t\|\right]$. Then based on the arguments of Hardt et al. (2016, Theorem 3.8) we know that with probability $1 - \frac{1}{N}$ over i_t, $\|w_t - \tilde{w}_t\| \le (1 + \eta_t L)\|w_{t-1} - \tilde{w}_{t-1}\|$, and $\|w_t - \tilde{w}_t\| \le$

$\|w_{t-1} - \tilde{w}_{t-1}\| + 2G\eta_t$ with probability $\frac{1}{N}$. Therefore we have

$$\begin{aligned}
\Delta_t &\leq \left(1 - \frac{1}{N}\right)(1+\eta_t L)\Delta_{t-1} + \frac{1}{N}\left(\Delta_{t-1} + 2G\eta_t\right) \\
&= \left(\left(1 - \frac{1}{N}\right)(1+\eta_t L) + \frac{1}{N}\right)\Delta_{t-1} + \frac{2G\eta_t}{N} \\
&= \left(1 + \left(1 - \frac{1}{N}\right)\eta_t L\right)\Delta_{t-1} + \frac{2G\eta_t}{N} \\
&\leq \exp\left(\left(1 - \frac{1}{N}\right)\eta_t L\right)\Delta_{t-1} + \frac{2G\eta_t}{N} \\
&\leq \exp\left(\eta_t L\right)\Delta_{t-1} + \frac{2G\eta_t}{N},
\end{aligned}$$

where we have used $1 + x \leq \exp(x)$. Then we can unwind the above recursion form to obtain that for all $t \geq 1$,

$$\Delta_t \leq \sum_{\tau=1}^{t} \prod_{i=\tau+1}^{t} \exp\left(\eta_i L\right) \frac{2G\eta_\tau}{N} = \frac{2G}{N} \sum_{\tau=1}^{t} \exp\left(L \sum_{i=\tau+1}^{t} \eta_i\right)\eta_\tau, \tag{34}$$

where we have used $\Delta_0 = 0$. Now we consider $\Gamma_t := \mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right]$. Then we can verify that with probability $1 - \frac{1}{N}$ over i_t, $\|w_t - \tilde{w}_t\|^2 \leq (1+\eta_t L)^2 \|w_{t-1} - \tilde{w}_{t-1}\|^2$, and with probability $\frac{1}{N}$,

$$\|w_t - \tilde{w}_t\|^2 \leq (\|w_{t-1} - \tilde{w}_{t-1}\| + 2G\eta_t)^2 = \|w_{t-1} - \tilde{w}_{t-1}\|^2 + 4G\eta_t\|w_{t-1} - \tilde{w}_{t-1}\| + 4G^2\eta_t^2.$$

Therefore we have

$$\begin{aligned}
\Gamma_t &\leq \left(1 - \frac{1}{N}\right)(1+\eta_t L)^2\Gamma_{t-1} + \frac{1}{N}\left(\Gamma_{t-1} + 4G\eta_t\Delta_{t-1} + 4G^2\eta_t^2\right) \\
&\leq \left(\left(1 - \frac{1}{N}\right)(1+\eta_t L)^2 + \frac{1}{N}\right)\Gamma_{t-1} + \frac{4G^2}{N}\left(\underbrace{\eta_t^2 + 2\eta_t \sum_{\tau=1}^{t-1} \exp\left(L \sum_{i=\tau+1}^{t-1} \eta_i\right)\eta_\tau}_{u_t}\right) \\
&= \left(1 + \left(1 - \frac{1}{N}\right)(2\eta_t L + \eta_t^2 L^2)\right)\Gamma_{t-1} + \frac{4G^2 u_t}{N} \\
&\leq \exp\left(\left(1 - \frac{1}{N}\right)(2\eta_t L + \eta_t^2 L^2)\right)\Gamma_{t-1} + \frac{4G^2 u_t}{N} \\
&\leq \exp\left(2\eta_t L + \eta_t^2 L^2\right)\Gamma_{t-1} + \frac{4G^2 u_t}{N},
\end{aligned}$$

where in the second inequality we have used the bound (34) on Δ_t. Recall that $\Gamma_0 = 0$. Then we can unwind the above recursion form to obtain

$$\Gamma_t \leq \frac{4G^2}{N} \sum_{\tau=1}^{t} \left\{\prod_{i=\tau+1}^{t} \exp\left(2\eta_i L + \eta_i^2 L^2\right)\right\} u_\tau \leq \frac{4G^2}{N} \sum_{\tau=1}^{t} \exp\left(3L \sum_{i=\tau+1}^{t} \eta_i\right) u_\tau,$$

where we have used $\eta_t \leq 1/L$. It follows immediately from the convexity that

$$\mathbb{E}\left[\|\bar{w}_T - \bar{\tilde{w}}_T\|^2\right] \leq \frac{\sum_{t=1}^{T} \mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right]}{T} \leq \frac{4G^2}{N} \sum_{t=1}^{T} \exp\left(3L \sum_{\tau=t+1}^{T} \eta_\tau\right) u_t.$$

Since the above holds for any $S \doteq \tilde{S}$, we have that for all $i \in [N]$,

$$\sup_{S, Z_i'} \mathbb{E}_\xi\left[\left\|\bar{w}_T - \bar{w}_T^{(i)}\right\|^2\right] \leq \frac{4G^2}{N} \sum_{t=1}^{T} \exp\left(3L \sum_{\tau=t+1}^{T} \eta_\tau\right) u_t,$$

Algorithm 3: $A_{\texttt{SGD-w/o}}$: SGD under Without-Replacement Sampling

Input :Data set $S = \{Z_i\}_{i\in[N]}$, step-sizes $\{\eta_t\}_{t\geq 1}$, #iterations T, initialization w_0.
Output : $\bar{w}_T = \frac{1}{T}\sum_{t\in[T]} w_t$.
for $t = 1, 2, ..., T$ **do**
  Uniformly randomly sample an index $i_t \in [N]$ *without* replacement;
  Compute $w_t = \Pi_{\mathcal{W}}\left(w_{t-1} - \eta_t \nabla_w \ell(w_{t-1}; Z_{i_t})\right)$.
end

where $\{w_t^{(i)}\}_{t\in[T]}$ is generated over $S^{(i)}$ by $A_{\texttt{SGD-w}}$. Finally, since the loss is G-Lipschitz, it follows from the above that for all $i \in [N]$,

$$\sup_{S, Z_i', Z} \mathbb{E}_\xi\left[\left(\ell(\bar{w}_T; Z) - \ell(\bar{w}_T^{(i)}; Z)\right)^2\right] \leq \frac{4G^4}{N}\sum_{t=1}^{T}\exp\left(3L\sum_{\tau=t+1}^{T}\eta_\tau\right)u_t.$$

This proves the desired L_2-uniform stability of algorithm. □

D Augmented Results for SGD under Without-Replacement Sampling

In this section, we further consider applying our main results in Theorem 2 to the variant of SGD under without-replacement sampling ($A_{\texttt{SGD-w/o}}$), as is outlined in Algorithm 3. For the sake of simplicity and readability, we only consider single-epoch processing with $T \leq N$. The extensions of our analysis to multi-epoch processing, i.e., $T \leq rN$ for some integer $r \geq 1$ are more or less straightforward and thus the details are omitted.

D.1 Results for convex and smooth loss

We start by considering the regime where the loss function is convex and smooth. We need the following lemma on the L_2-uniform stability of $A_{\texttt{SGD-w/o}}$ which can be proved based on the result from Bassily et al. (2020, Lemma 3.1).

Lemma 15. *Suppose that the loss function $\ell(\cdot;\cdot)$ is convex, G-Lipschitz and L-smooth with respect to its first argument. Assume that $\eta_t \leq 2/L$ for all $t \geq 1$. Consider $T \leq N$. Then $A_{\texttt{SGD-w/o}}$ has L_2-uniform stability with parameter*

$$\gamma_{L_2,N} = 2G^2\sqrt{\frac{1}{N}\sum_{t=1}^{T}\eta_t^2}.$$

Proof. For any fixed pair of data sets $S, \tilde{S}$ that differ in a single element, let us define the sequences $\{w_t\}_{t\in[T]}$ and $\{\tilde{w}_t\}_{t\in[T]}$ that are respectively generated over S and $\tilde{S}$ by $A_{\texttt{SGD-w/o}}$ via sample path $\xi = \{i_t\}_{t\in[T]}$. Recall that $T \leq N$. Let us define a stopping time variable t_0 such that $Z_{\xi_{t_0}} \neq \tilde{Z}_{\xi_{t_0}}$. Since $S \doteq \tilde{S}$, the uniform randomness of i_t implies that

$$\mathbb{P}(t_0 = j) = \frac{1}{N}, \quad j \in [N].$$

In the proof of Corollary 1 we have already shown that $\|w_t - \tilde{w}_t\|^2 \leq \|w_{t-1} - \tilde{w}_{t-1}\|^2$ if $Z_{i_t} = \tilde{Z}_{i_t}$ and $\|w_t - \tilde{w}_t\|^2 \leq \|w_{t-1} - \tilde{w}_{t-1}\|^2 + 4G\eta_t\|w_{t-1} - \tilde{w}_{t-1}\| + 4G^2\eta_t^2$ otherwise. Therefore, the without-replacement sampling implies that the following bound holds for any given $t_0 \leq t \leq T$:

$$\|w_t - \tilde{w}_t\|^2 \leq 4G^2\eta_{t_0}^2,$$

and $\|w_t - \tilde{w}_t\|^2 = 0$ for $0 \leq t < t_0$. Then based on the law of total expectation we can show that

$$\mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right] \leq \frac{4G^2}{N}\sum_{t_0=1}^{t}\eta_{t_0}^2 \leq \frac{4G^2}{N}\sum_{t=1}^{T}\eta_t^2.$$

The convexity of squared Euclidean norm leads to

$$\mathbb{E}\left[\|\bar{w}_T - \tilde{\bar{w}}_T\|^2\right] \le \frac{\sum_{t=1}^{T} \mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right]}{T} \le \frac{4G^2}{N} \sum_{t=1}^{T} \eta_t^2.$$

Since the above holds for any $S \doteq \tilde{S}$, we have that for all $i \in [N]$,

$$\sup_{S, Z_i'} \mathbb{E}_\xi \left[\left\|\bar{w}_T - \bar{w}_T^{(i)}\right\|^2\right] \le \frac{4G^2}{N} \sum_{t=1}^{T} \eta_t^2,$$

where $\{w_t^{(i)}\}_{t\in[T]}$ is generated over $S^{(i)}$ by $A_{\texttt{SGD-w/o}}$. Finally, since the loss is G-Lipschitz, it follows from the above bound that for all $i \in [N]$,

$$\sup_{S, Z_i', Z} \mathbb{E}_\xi \left[\left(\ell(\bar{w}_T; Z) - \ell(\bar{w}_T^{(i)}; Z)\right)^2\right] \le \frac{4G^4}{N} \sum_{t=1}^{T} \eta_t^2.$$

This proves the desired L_2-uniform stability of algorithm. □

The following result is a direct consequence of Theorem 2 when invoking Algorithm 1 to $A_{\texttt{SGD-w/o}}$ with convex and smooth loss.

Corollary 4. *Suppose that the loss function $\ell(\cdot;\cdot) \in [0, M]$ is convex, G-Lipschitz and L-smooth with respect to its first argument. Consider Algorithm 1 specified to $A_{\texttt{SGD-w/o}}$ with $T = N$ and learning rate $\eta_t \le 2/L$ for all $t \ge 1$. Then for any $\delta \in (0, 1)$ and $K \ge 2\log(\frac{2}{\delta})$, with probability at least $1 - \delta$ over the randomness of S and $\{\xi\}_{k\in[K]}$, the generalization bound of Algorithm 1 is upper bounded as*

$$\begin{aligned}&|R(A_{\texttt{SGD-w/o}}(S_1, \xi_{k^*})) - R_S(A_{\texttt{SGD-w/o}}(S_1, \xi_{k^*}))| \\ \lesssim& G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\frac{1}{(1-\mu)N} \sum_{t=1}^{N} \eta_t^2} + \frac{M}{\sqrt{\mu(1-\mu)}} \sqrt{\frac{\log(K/\delta)}{N}}.\end{aligned}$$

Proof. For the considered convex and smooth losses, from Lemma 15 we know that $A_{\texttt{SGD-w/o}}$ with $T = N$ iterations has L_2-uniform stability with parameter $\gamma_{\mathrm{L}_2, N} = 2G^2 \sqrt{\frac{1}{N}\sum_{t=1}^{N} \eta_t^2}$. Then the generalization error bound follows immediately from Theorem 2. □

Remark 12. *Specially for constant learning rates $\eta_t \equiv \eta \asymp \frac{1}{\sqrt{N}}$ and $K \asymp \log\left(\frac{1}{\delta}\right)$, Corollary 4 admits a high-probability generalization bound of order $\mathcal{O}\big(\sqrt{\frac{\log(1/\delta)}{N}} + \frac{\log(N)\log(1/\delta)}{\sqrt{N}}\big)$. For time varying learning rates $\eta_t \asymp \frac{1}{\sqrt{t}}$, the generalization bound scales as $\mathcal{O}\big(\frac{\log(N)\log(1/\delta)}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\big)$.*

Further assume that $\mathcal{W}$ is bounded with diameter D. Consider the constant learning rate $\eta_t \equiv \min\{\frac{2}{L}, \frac{D}{G\sqrt{N}}\}$. Then the following in-expectation optimization error bound of $A_{\texttt{SGD-w/o}}$ with convex and smooth loss functions is known (Nagaraj et al., 2019, Theorem 3):

$$\mathbb{E}_\xi \left[R_S(\bar{w}_T) - \min_{w \in \mathcal{W}} R_S(w)\right] \lesssim \frac{D^2 L}{N} + \frac{GD}{\sqrt{N}}.$$

Invoking generic bound (14) combined with Lemma 15 and the above sub-optimality bound yields the following excess risk bound of (modified) Algorithm 1:

$$\begin{aligned}R(A_{\texttt{SGD-w/o}}(S_1, \xi_{k^*})) - R^* \lesssim& G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\frac{1}{(1-\mu)N} \sum_{t=1}^{N} \eta_t^2} + \frac{M}{\sqrt{\mu(1-\mu)}} \sqrt{\frac{\log(K/\delta)}{N}} \\ &+ \frac{GDK}{\sqrt{(1-\mu)N}} + \frac{D^2 L}{(1-\mu)N}.\end{aligned}$$

D.2 Results for convex and non-smooth loss

We now turn to study the case of convex and non-smooth losses. The following lemma is about the L_2-uniform stability of $A_{\texttt{SGD-w/o}}$ in this case.

Lemma 16. *Suppose that the loss function $\ell(\cdot;\cdot)$ is convex and G-Lipschitz with respect to its first argument. Consider $T \le N$. Then $A_{\texttt{SGD-w/o}}$ has L_2-uniform stability with parameter*

$$\gamma_{L_2,N} = 2G^2\sqrt{\frac{1}{N}\sum_{t_0=1}^{T}\sum_{t=t_0}^{T}\eta_t^2}.$$

Proof. The proof arguments are similar to those of Lemma 15. Here we reproduce the proof for the sake of completeness. For any fixed pair of data sets $S, \tilde{S}$ that differ in a single element, let us define the sequences $\{w_t\}_{t\in[T]}$ and $\{\tilde{w}_t\}_{t\in[T]}$ that are respectively generated over S and $\tilde{S}$ by $A_{\texttt{SGD-w/o}}$ via sample path $\xi = \{i_t\}_{t\in[T]}$. Recall that $T \le N$. Let us define a stopping time variable t_0 such that $Z_{\xi_{t_0}} \neq \tilde{z}_{\xi_{t_0}}$. Since $S \doteq \tilde{S}$, the uniform randomness of i_t and the without-replacement sampling strategy yield

$$\mathbb{P}(t_0 = j) = \frac{1}{N}, \quad j \in [N].$$

For any $t_0 \le t \le T$, under without-replacement sampling, it follows from Lemma 14 that

$$\|w_t - \tilde{w}_t\|^2 \le 4G^2\sum_{\tau=t_0}^{t}\eta_\tau^2.$$

We use the convention $\sum_{\tau=t_0}^{t}\eta_\tau^2 = 0$ for $0 \le t < t_0$. Then according to the law of total expectation we must have

$$\mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right] \le \frac{4G^2}{N}\sum_{t_0=1}^{t}\sum_{\tau=t_0}^{t}\eta_\tau^2 \le \frac{4G^2}{N}\sum_{t_0=1}^{T}\sum_{\tau=t_0}^{T}\eta_\tau^2.$$

The convexity of squared Euclidean norm leads to

$$\mathbb{E}\left[\|\bar{w}_T - \tilde{\bar{w}}_T\|^2\right] \le \frac{\sum_{t=1}^{T}\mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right]}{T} \le \frac{4G^2}{N}\sum_{t_0=1}^{T}\sum_{t=t_0}^{T}\eta_t^2.$$

Since the above holds for any $S \doteq \tilde{S}$, the following bound holds for all $i \in [N]$:

$$\sup_{S,Z_i'}\mathbb{E}_\xi\left[\left\|\bar{w}_T - \bar{w}_T^{(i)}\right\|^2\right] \le \frac{4G^2}{N}\sum_{t_0=1}^{T}\sum_{t=t_0}^{T}\eta_t^2,$$

where $\{w_t^{(i)}\}_{t\in[T]}$ is generated over $S^{(i)}$ by $A_{\texttt{SGD-w/o}}$. Finally, since the loss is G-Lipschitz, it follows from the above bound that for all $i \in [N]$,

$$\sup_{S,Z_i',Z}\mathbb{E}_\xi\left[\left(\ell(\bar{w}_T;Z) - \ell(\bar{w}_T^{(i)};Z)\right)^2\right] \le \frac{4G^4}{N}\sum_{t_0=1}^{T}\sum_{t=t_0}^{T}\eta_t^2.$$

This proves the desired L_2-uniform stability of algorithm. □

With the above lemma in hand, we can establish the following result as a direct consequence of Theorem 2 when invoking Algorithm 1 to $A_{\texttt{SGD-w/o}}$ with convex and non-smooth losses.

Corollary 5. *Suppose that the loss function $\ell(\cdot;\cdot)$ is convex and G-Lipschitz with respect to its first argument, and it is bounded in the interval $[0, M]$. Consider Algorithm 1 specified to $A_{\texttt{SGD-w/o}}$ with $T = N$. Then for any $\delta \in (0,1)$ and $K \ge 2\log(\frac{2}{\delta})$, with probability at least $1-\delta$ over the randomness of S and $\{\xi_k\}_{k\in[K]}$, the output of Algorithm 1 satisfies*

$$\begin{aligned}&|R(A_{\texttt{SGD-w/o}}(S_1,\xi_{k^*})) - R_S(A_{\texttt{SGD-w/o}}(S_1,\xi_{k^*}))|\\ \lesssim & G^2\log(N)\log\left(\frac{1}{\delta}\right)\sqrt{\frac{1}{(1-\mu)N}\sum_{t_0=1}^{N}\sum_{t=t_0}^{N}\eta_t^2} + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(K/\delta)}{N}}.\end{aligned}$$

Proof. For the considered convex and Lipschitz loss functions, from Lemma 16 we know that $A_{\texttt{SGD-w/o}}$ with $T = N$ iterations has L_2-uniform stability by $\gamma_{L_2,N} = 2G^2\sqrt{\frac{1}{N}\sum_{t_0=1}^{N}\sum_{t=t_0}^{N}\eta_t^2}$. The results then follow immediately via invoking Theorem 2 to the considered setting. □

Remark 13. *Specially for constant learning rates $\eta_t \equiv \eta$ and setting $K \asymp \log\left(\frac{1}{\delta}\right)$, Corollary 5 admits a high-probability generalization bound of scale $\mathcal{O}\big(\eta\sqrt{N} + \sqrt{\frac{\log(1/\delta)}{N}}\big)$. For time decaying learning rates $\eta_t \asymp \frac{1}{t}$, the generalization bound scales as $\mathcal{O}\big(\sqrt{\frac{\log(N)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\big)$.*

Regarding the excess risk bound, under the conditions of Corollary 5 and $K \asymp \log\left(\frac{1}{\delta}\right)$, the risk bound (14) combined with Lemma 16 yields the following exponential risk bound of (modified) Algorithm 1:

$$
\begin{aligned}
&R(A_{\texttt{SGD-w/o}}(S_1, \xi_{k^*})) - R^* \\
\lesssim&\Delta_{\text{opt}} + G^2\log(N)\log\left(\frac{1}{\delta}\right)\sqrt{\frac{1}{(1-\mu)N}\sum_{t_0=1}^{N}\sum_{t=t_0}^{N}\eta_t^2} + \frac{M}{\sqrt{\mu(1-\mu)}}\sqrt{\frac{\log(1/\delta)}{N}}.
\end{aligned}
$$

In the special case of bounded-norm generalized linear models, Shamir (2016) established an in-expectation empirical risk sub-optimality bound $\Delta_{\text{opt}} \lesssim \frac{1}{\sqrt{N}}$ under suitable learning rates. For generic convex and non-smooth losses, however, it still remains unclear to us if similar sub-optimality bounds are available for SGD under without-replacement sampling.

D.3 Results for non-convex and smooth loss

Finally, we study the performance of Algorithm 1 for $A_{\texttt{SGD-w/o}}$ on smooth but not necessarily convex loss functions. We first establish the following lemma on the L_2-uniform stability of $A_{\texttt{SGD-w/o}}$ in the considered non-convex problem setting.

Lemma 17. *Suppose that the loss function $\ell(\cdot;\cdot)$ is G-Lipschitz and L-smooth with respect to its first argument. Consider $T \le N$. Then $A_{\texttt{SGD-w/o}}$ has L_2-uniform stability with parameter*

$$
\gamma_{L_2,N} = 2G^2\sqrt{\frac{1}{N}\sum_{t=1}^{T}\exp\left(2L\sum_{\tau=t+1}^{T}\eta_\tau\right)\eta_t^2}.
$$

Proof. Let us define the sequences $\{w_t\}_{t\in[T]}$ and $\{\tilde{w}_t\}_{t\in[T]}$ that are respectively generated over S and $\tilde{S}$ by $A_{\texttt{SGD-w/o}}$ via sample path $\xi = \{i_t\}_{t\in[T]}$. Suppose that $S \doteq \tilde{S}$. Recall that $T \le N$. Let us define a stopping time variable t_0 such that $Z_{\xi_{t_0}} \neq \tilde{Z}_{\xi_{t_0}}$. Since $S \doteq \tilde{S}$, the without-replacement sampling implies that

$$
\mathbb{P}(t_0 = j) = \frac{1}{N}, \quad j \in [N].
$$

In view of Lemma 12 we know that $\|w_t - \tilde{w}_t\| \le (1+\eta_t L)\|w_{t-1} - \tilde{w}_{t-1}\|$ if $Z_{i_t} = Z'_{i_t}$, and $\|w_{t_0} - \tilde{w}_{t_0}\| \le 2G\eta_{t_0}$ otherwise due to the assumption that the loss is G-Lipschitz. Therefore, it can be directly verified that the following holds for any $t_0 \le t \le T$:

$$
\|w_t - \tilde{w}_t\|^2 \le 4G^2\prod_{\tau=t_0+1}^{t}(1+\eta_\tau L)^2\eta_{t_0}^2,
$$

where we have used $\prod_{\tau=t_0+1}^{t}(1+\eta_\tau L)^2 = 1$ for $t = t_0$. For $t < t_0$, it is trivial that $\|w_t - \tilde{w}_t\| = 0$. Therefore the law of total expectation yields

$$
\begin{aligned}
\mathbb{E}\left[\|w_t - \tilde{w}_t\|^2\right] \le& \frac{4G^2}{N}\sum_{t_0=1}^{t}\prod_{\tau=t_0+1}^{t}(1+\eta_\tau L)^2\eta_{t_0}^2 \\
\le& \frac{4G^2}{N}\sum_{t_0=1}^{t}\prod_{\tau=t_0+1}^{t}\exp(2\eta_\tau L)\,\eta_{t_0}^2 \\
\le& \frac{4G^2}{N}\sum_{t_0=1}^{T}\exp\left(2L\sum_{\tau=t_0+1}^{T}\eta_\tau\right)\eta_{t_0}^2,
\end{aligned}
$$

where we have used $1 + x \le \exp(x)$. The convexity of Euclidean norm leads to

$$\mathbb{E}\left[\|\bar{w}_T - \tilde{\bar{w}}_T\|\right] \le \frac{4G^2}{N} \sum_{t_0=1}^{T} \exp\left(2L \sum_{\tau=t_0+1}^{T} \eta_\tau\right) \eta_{t_0}^2.$$

Since the above holds for any $S \doteq \tilde{S}$, the following bound holds for all $i \in [N]$:

$$\sup_{S, Z_i'} \mathbb{E}_\xi \left[\left\|\bar{w}_T - \bar{w}_T^{(i)}\right\|^2\right] \le \frac{4G^2}{N} \sum_{t_0=1}^{T} \exp\left(2L \sum_{\tau=t_0+1}^{T} \eta_\tau\right) \eta_{t_0}^2,$$

where $\{w_t^{(i)}\}_{t\in[T]}$ is generated over $S^{(i)}$ by $A_{\texttt{SGD-w/o}}$. Finally, since the loss is G-Lipschitz, it follows from the above bound that for all $i \in [N]$,

$$\sup_{S, Z_i', Z} \mathbb{E}_\xi \left[\left(\ell(\bar{w}_T; Z) - \ell(\bar{w}_T^{(i)}; Z)\right)^2\right] \le \frac{4G^4}{N} \sum_{t_0=1}^{T} \exp\left(2L \sum_{\tau=t_0+1}^{T} \eta_\tau\right) \eta_{t_0}^2.$$

This proves the desired L_2-uniform stability of algorithm. □

With Lemma 17 in place, we can readily derive the following result as a direct application of Theorem 2 to $A_{\texttt{SGD-w/o}}$ with Lipschitz and smooth losses.

Corollary 6. *Suppose that the loss function $\ell(\cdot;\cdot)$ is G-Lipschitz and L-smooth with respect to its first argument, and it is bounded in the interval $[0, M]$. Consider Algorithm 1 specified to $A_{\texttt{SGD-w/o}}$ with $T = N$. Then for any $\delta \in (0, 1)$ and $K \ge 2\log(\frac{2}{\delta})$, with probability at least $1 - \delta$ over the randomness of S and $\{\xi_k\}_{k\in[K]}$, the output of Algorithm 1 satisfies*

$$\begin{aligned} &|R(A_{\texttt{SGD-w/o}}(S_1, \xi_{k^*})) - R_S(A_{\texttt{SGD-w/o}}(S_1, \xi_{k^*}))| \\ \lesssim & G^2 \log(N) \log\left(\frac{1}{\delta}\right) \sqrt{\frac{1}{(1-\mu)N} \sum_{t=1}^{N} \exp\left(L \sum_{\tau=t+1}^{N} \eta_\tau\right) \eta_t^2} + \frac{M}{\sqrt{\mu(1-\mu)}} \sqrt{\frac{\log(K/\delta)}{N}}. \end{aligned}$$

Proof. For the considered smooth and Lipschitz loss functions, from Lemma 17 we know that $A_{\texttt{SGD-w/o}}$ with $T = N$ rounds of iteration has L_2-uniform stability with parameter $\gamma_{\mathrm{L}_2,N} = 2G^2\sqrt{\frac{1}{N}\sum_{t=1}^{T} \exp\left(2L\sum_{\tau=t+1}^{T} \eta_\tau\right) \eta_t^2}$. The desired results then follow immediately via invoking Theorem 2 to the considered problem regime. □

Remark 14. *For $K \asymp \log\left(\frac{1}{\delta}\right)$ and the choice of constant learning rates $\eta_t \equiv \frac{1}{LN}$, Corollary 6 admits high-probability generalization bounds of scale $\mathcal{O}\big(\frac{\log(N)\log(1/\delta)}{N} + \sqrt{\frac{\log(1/\delta)}{N}}\big)$. For the choice of time decaying learning rates $\eta_t = \frac{1}{L\nu t}$ with arbitrary $\nu > 2$, it can be verified that the corresponding generalization bound is of scale $\mathcal{O}\big(\frac{\log(N)\log(1/\delta)}{\nu N^{1/2-1/\nu}} + \sqrt{\frac{\log(1/\delta)}{N}}\big)$.*

E Some Additional Related Work

The idea of using stability of a learning algorithm, namely the sensitivity of estimated model to the changes in training data, for generalization performance analysis dates back to the seventies (Vapnik and Chervonenkis, 1974; Rogers and Wagner, 1978; Devroye and Wagner, 1979). For deterministic learning algorithms, algorithmic stability has been extensively studied with a bunch of applications to establishing strong generalization and excess risk bounds for stable learning models like k-NN and regularized ERMs (Bousquet and Elisseeff, 2002; Zhang, 2003; Klochkov and Zhivotovskiy, 2021; Yuan and Li, 2023). The stability theory for randomized learning algorithms was formally introduced and investigated by Elisseeff et al. (2005). In the celebrated work of Hardt et al. (2016), it was shown in that the solution obtained via stochastic gradient descent is expected to be stable and generalize well for smooth convex and non-convex loss functions. For non-smooth convex losses, the stability induced generalization bounds of SGD have been established in expectation (Lei and Ying, 2020) or deviation (Bassily et al., 2020). In the work of Kuzborskij and Lampert (2018), a set of data-dependent generalization bounds for SGD were derived based on the stability of algorithm. More broadly, generalization bounds for stable learning algorithms that converge to global minima were established in Charles and Papailiopoulos (2018); Lei and Ying (2021). For non-convex sparse learning, algorithmic stability theory has been applied to derive the generalization bounds of the popularly used iterative hard thresholding (IHT) algorithm (Yuan and Li, 2022). The uniform stability bounds on SGD have also been extensively used for designing differential privacy stochastic optimization algorithms (Bassily et al., 2019; Feldman et al., 2020).

The confidence-boosting technique has long been applied for obtaining sharp high-probability excess risk bounds from the corresponding in-expectation bounds (Shalev-Shwartz et al., 2010; Mehta, 2017; Holland, 2021). For generic statistical learning problems, confidence-boosting has been used to convert any low-confidence learning algorithm with linear dependence on $1/\delta$ to a high-confidence algorithm with logarithmic factor $\log(1/\delta)$. For learning with exp-concave losses, a relevant ERM estimator with in-expectation fast rate of convergence was converted to a high-confidence learning algorithm with an almost identical fast rate of convergence up to a logarithmic factor on $1/\delta$ (Mehta, 2017). While sharing a similar spirit, our generalization analysis is substantially more challenging than the existing excess risk analysis in the sense that deriving a favorable first-moment generalization bound for L_2-uniformly stable randomized algorithms is highly non-trivial in itself. Bagging (or bootstrap aggregating) is one of the earliest yet most popular ensemble methods that has been widely applied to reduce the variance for unstable learning algorithms such as decision tree and neural networks (Breiman, 1996; Opitz and Maclin, 1999), and sometimes stable algorithms such as SVMs (Valentini and Dietterich, 2003). As an important variant of bagging, subbagging has been proposed to reduce the computational cost of bagging via training base models under without-replacement sampling (Bühlmann, 2012). The stability and generalization bounds of bagging have been analyzed for both uniform (Elisseeff et al., 2005) and non-uniform (Foster et al., 2019) averaging schemes. Unlike these prior results for bagging with averaging aggregation, our confidence-boosting bounds are obtained based on a greedy aggregation scheme which turns out to yield sharper dependence on the stability parameters.