Toward Better PAC-Bayes Bounds for Uniformly Stable Algorithms

Zhou, Sijia; Lei, Yunwen; Kaban, Ata

Toward Better PAC-Bayes Bounds for Uniformly Stable Algorithms

Sijia Zhou[1] **Yunwen Lei**[2*] **Ata Kabán**[1]
[1]School of Computer Science, University of Birmingham, Birmingham B15 2TT, United Kingdom
[2] Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong, China
sxz115@student.bham.ac.uk leiyw@hku.hk a.kaban@bham.ac.uk

Abstract

We give sharper bounds for uniformly stable randomized algorithms in a PAC-Bayesian framework, which improve the existing results by up to a factor of $\sqrt{n}$ (ignoring a log factor), where n is the sample size. The key idea is to bound the moment generating function of the generalization gap using concentration of weakly dependent random variables due to Bousquet et al (2020). We introduce an assumption of sub-exponential stability parameter, which allows a general treatment that we instantiate in two applications: stochastic gradient descent and randomized coordinate descent. Our results eliminate the requirement of strong convexity from previous results, and hold for non-smooth convex problems.

1 Introduction

Stochastic optimization methods are the workhorses of many modern machine learning problems. A lot of progress has been made in reducing optimization errors with less training time in stochastic settings [6, 21, 41, 46, 64]. One such method is stochastic gradient descent (SGD), which builds a stochastic gradient estimator iteratively based on a randomly sampled example to approximate the gradient for the next iteration. SGD is appealing for large-scale data analysis due to its cheap computational cost, simplicity and efficiency.

Obtaining models that generalize well is the goal in machine learning and we undertake a theoretical analysis of this problem. The generalization behavior of a model can be quantified by the excess risk, which decomposes into two components: the optimization error and the generalization error (or generalization gap). In this paper, we focus on the analysis of the generalization error of stochastic optimization methods. We combine two useful approaches to bound the generalization error, that is algorithmic stability [7, 16] and PAC-Bayes bounds [9, 26, 34, 53]. Our hope is to retain the benefits of both approaches to derive powerful generalization bounds to better understand the behavior of stochastic optimization methods.

Seminal work by Bousquet et al. [7] provided generalization bounds for uniformly stable algorithms in the deterministic case. Sharper generalization bounds were obtained via a moment bound on the generalization gap and a concentration inequality of weakly dependent random variables [8, 18]. Extending stability-based bounds to randomized algorithms is a challenge. In [16], where the stability-based bounds for the randomized algorithms were considered, the results hold for fixed distributions. To give bounds uniformly for all distributions, we need to turn to PAC-Bayes analysis. A recent work analyzed stable algorithms in the PAC-Bayes framework and derived generalization bounds that hold for any distribution [32]. However, a comparison between [8] and [32] reveals that the error convergence rate in the randomized case [32] is much slower than that for the deterministic case [8]. In more detail, it was shown that β-uniformly stable (β is decreasing w.r.t. n, where n is the sample

*Corresponding author

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

size) and deterministic algorithms would imply generalization bounds of order $\widetilde{O}(1/\sqrt{n}+\beta)$ [8, 18][2], while PAC-Bayes bounds of order $O(1/\sqrt{n}+\sqrt{n}\beta)$ were developed for randomized algorithms [32]. In [32], the randomness comes from the sampling of hyperparameters such as the index of examples chosen in SGD, and the PAC-Bayes bounds hold for any posterior distribution which may depend on the dataset. It is clear that the PAC-Bayes bounds can be slower than those in [8, 18] by a factor of $\sqrt{n}$, which motivates a natural question: can we develop PAC-Bayes bounds for randomized algorithms which match the rate of deterministic algorithms?

This paper explores the above question. We provide sharper PAC-Bayes bounds that hold for uniformly stable randomized algorithms. We adapt a moment bound [8] previously used for stable algorithms in deterministic cases and extend it to the PAC-Bayesian framework, to give bounds that hold for randomized predictors. The PAC-Bayes framework is based on the work of [32]. However, we take a different analysis strategy to control the change in hyperparameters, which is based on an assumption of sub-exponential stability parameter. This general assumption allows us to bound the moment generating function (MGF) of the generalization gap within a high probability domain where our assumption holds (see Appendix A.2). Furthermore, we prove that this assumption holds for both SGD and randomized coordinate descent (RCD). We then illustrate the advantage of our results over existing bounds.

Regarding the convergence rate, our main result improves on the existing PAC-Bayes bounds [32] by a factor of $\sqrt{n}$ (ignoring a log factor). This improvement holds under weaker conditions under which convergence is not guaranteed in [32]. Our primary technical tool is a moment bound, which we extend to randomized learning algorithms in the PAC-Bayes framework. This allows our bounds to hold for all possible posteriors, not just fixed ones [16] or deterministic algorithms [8, 18]. We need to introduce novel techniques to handle the randomness of the hyperparameter, which is a challenge in the PAC-Bayes analysis (details will be given in Section 3.2).

Regarding assumptions, our result holds without the requirement of hyperparameter stability in previous work [32]. Instead, we introduce a new assumption on the sub-exponential behavior of uniform stability by viewing the uniform stability as a function of the random hyperparameter. Interestingly, it suffices to study the sub-exponential behavior of uniform stability under the prior distribution, which makes the sub-exponential assumption easy to check.

We show that the uniform stability of SGD and RCD have a mixture of sub-Gaussian and sub-exponential tails that satisfy this assumption. Thus, we remove the strong convexity assumption in the existing analysis of SGD [32] and extend the result to non-smooth problems. Our result also applies to RCD, where the randomness arises from the selected coordinates.

The remainder of the paper is organized as follows. We discuss the basics on PAC-Bayes and stability analysis in Section 2. We present our main results in Section 3, and apply it to SGD and RCD in Section 4. We survey the related work on stability and PAC-Bayesian analysis in Section 5, and conclude the paper in Section 6.

2 Preliminaries

We first introduce some notations. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output space respectively and let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We are given a set of training examples of size n, $S = \{z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)\}$, drawn independently and identically distributed (i.i.d.) from an unknown distribution $\mathcal{D}$ on $\mathcal{Z}$. We hope to learn a predictor from a class of hypotheses $\mathcal{H}$ to predict unseen new data drawn from $\mathcal{D}$. Let $\mathcal{W}$ denote the weight space, $\mathcal{W} \subseteq \mathbb{R}^d$, and Θ denote a hyperparameter space. A deterministic learning algorithm $A : \mathcal{Z}^n \times \Theta \to \mathcal{H}$ maps the training examples to a hypothesis $h_{\mathbf{w}} \in \mathcal{H}$ determined by $\mathbf{w} \in \mathcal{W}$. The quality of a hypothesis is measured by a loss function, $\ell : \mathcal{H} \times \mathcal{Z} \to [0, M]$.

The goal of any learning algorithm is to produce a predictor that generalizes well. That means that the learned predictor applied on previously unseen input data from the marginal of $\mathcal{D}$ should have a small expected loss. For any $\theta \in \Theta$, the risk of a predictor returned by the algorithm A is a random variable as a function of S, defined as

$$R(A(S;\theta)) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(A(S;\theta), z)]. \tag{2.1}$$

[2]We use $\widetilde{O}$ to hide poly-logarithmic factors.

Since $\mathcal{D}$ is unknown, we have no access to the true risk. Instead, the empirical risk is often used as an approximation

$$R_S(A(S;\theta)) = \frac{1}{n}\sum_{i=1}^{n} \ell(A(S;\theta), z_i). \tag{2.2}$$

We are interested in how well the empirical risk can estimate the risk. The difference between them is the generalization gap $G(S;\theta) \triangleq R(A(S;\theta)) - R_S(A(S;\theta))$. We can upper bound the risk by bounding this difference.

2.1 PAC-Bayes basics

We are interested in randomized algorithms, such as stochastic gradient descent (SGD) and randomized coordinate descent (RCD). In such algorithms, there is an in-built random sampling mechanism that we can think of as a random hyperparameter. In other words, a randomized algorithm may be viewed as a deterministic algorithm with hyperparameters θ that follow a distribution. Therefore, for the analysis that follows, we will explicitly define distributions on Θ. There has been a lot of interest in SGD because it is often used to train a model and its randomness comes from independent sampling of training instances to estimate gradient directions. In this case, the hyperparameters $\theta \in \Theta$ form a random sequence $\theta = (\theta_1, ..., \theta_T)$, where every $\theta_t \in [n]$, for $t \in [T]$, is an i.i.d. index of a training point from S. Here $[n]$ denotes $\{1, \ldots, n\}$.

Any distribution over the hyperparameter space Θ induces a distribution over the predictors (or weights since we often consider parametric models). In the PAC-Bayes methodology, the common approach is to define a prior distribution on the domain of the weights [13, 22, 34, 53]. However, in the context of randomized algorithms of the kind mentioned above it is more natural to exploit the randomness already present through a distribution on the domain of the hyperparameters.

Hence, we will conduct PAC-Bayesian analysis by focusing on the randomness of the hyperparameters. Let $\mathbb{P}$ be any prior distribution on the hyperparameter domain Θ, chosen before seeing the training data. Given a randomized algorithm, we will only need to prove a condition for the prior $\mathbb{P}$ on Θ, and in return our generalization guarantees will hold with high probability for *all* posteriors $\mathbb{Q}$ on Θ. In particular, we can assume a simple uniform distribution as a prior for SGD, i.e. SGD with uniform sampling, and derive bounds for data-dependent non-uniform sampling [49, 56, 66]. For example, SGD with importance sampling [66] shows better results than uniform sampling, and there are efforts to learn good sampling distributions [32] in a line of research towards self-certified learning algorithms [13, 44].

The quality of a predictor learned by an algorithm A from the training sample S, is defined as the expected risk w.r.t the PAC-Bayes posterior $\mathbb{Q}$ as

$$R(S, \mathbb{Q}) = \mathop{\mathbb{E}}_{\theta\sim\mathbb{Q}}[R(A(S;\theta))]. \tag{2.3}$$

Its empirical counterpart, the expected empirical risk w.r.t. $\mathbb{Q}$ is

$$R_S(S, \mathbb{Q}) = \mathop{\mathbb{E}}_{\theta\sim\mathbb{Q}}[R_S(A(S;\theta))]. \tag{2.4}$$

In special cases this expectation can be computed analytically; most often it is approximated either by Monte Carlo sampling or by analytical upper bounds.

PAC-Bayes bounds aim to estimate $R(S, \mathbb{Q})$ in terms of $R_S(S, \mathbb{Q})$ and the divergence between $\mathbb{P}$ and $\mathbb{Q}$. A key ingredient that PAC-Bayes bounds rest upon is a change of measure inequality, also known as variational formula. For completeness this is given in Appendix (Lemma A.3).

2.2 Algorithmic stability basics

Another relatively recent framework for generalization analysis is based on the concept of algorithmic stability [7]. The key concept in this framework quantifies how sensitive a learning algorithm is to small perturbations of the training data. There are several notions of algorithmic stability, and the one we use in our work is the following uniform stability [7]. We denote $S \sim S'$ if they are neighboring datasets, i.e., S and S' differ by at most a single example.

Definition 1 (Uniform Stability)**.** For any θ, an algorithm $A : S \mapsto A(S;\theta)$ is β_θ-uniformly stable w.r.t. a loss function ℓ if $\forall S \sim S' \in \mathcal{Z}^n, \forall z \in \mathcal{Z}$,

$$|\ell(A(S;\theta), z) - \ell(A(S';\theta), z)| \leq \beta_\theta. \tag{2.5}$$

Recall that A is a randomized algorithm whose randomness is exclusively due to a random draw of θ. Hence, given a fixed instance of θ, the algorithm A becomes a deterministic algorithm. This simple observation helps us reduce the problem from randomized learning to deterministic learning. Based on this observation, the next section presents sharper generalization bounds for randomized algorithms in comparison to previous results [32].

3 Main Results

First we introduce a sub-exponential assumption on the stability parameter. This will allow us to make some key innovations: 1) We will be able to remove one of the assumptions required by [32, Theorem 2] (hyperparameter stability) and only require uniform stability. 2) we will state our result in more general terms, and instantiate it to specific algorithms such as SGD and RCD.

Assumption 1 (Sub-exponential stability)**.** Let $\mathbb{P}$ be a fixed probability distribution on $\Theta = \prod_{t=1}^{T} \Theta_t$. We say that a randomized algorithm with random hyperparameters $\theta \sim \mathbb{P}$ satisfies sub-exponential stability if, for any fixed instance of θ it satisfies β_θ-uniform stability w.r.t. a loss function ℓ, and there exists $c \in \mathbb{R}$ such that for any $\delta \in (0, 1/n]$, with probability at least $1 - \delta$ over draws of $\theta \sim \mathbb{P}$:

$$\beta_\theta \leq \mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c \log(1/\delta). \tag{3.1}$$

In other words, this assumption says that the deviation of β_θ from its mean roughly has a sub-exponential tail. We will show that c is dominated by $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]$ in typical stochastic optimization algorithms such as SGD and RCD.

Note that in the above assumption, the probability is made w.r.t. $\theta \sim \mathbb{P}$, which is a prior distribution independent of S. We can choose $\mathbb{P}$ to be a uniform distribution. In this case, we will show that SGD satisfies this assumption under mild conditions on the loss functions with very small $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]$ and c.

Recall that for any fixed instance of θ, the trained model of the learning algorithm A on data set S with hyperparameters θ is a deterministic predictor. Our proof strategy for our main result below is to first fix θ, and apply a recent technique of obtaining sharp bounds [8] to the resulting deterministic algorithm. After that, we deal with the randomness of θ.

In Theorem 1 we give our main result. This bound is useful when we have small $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]$ and c. We will see two examples in Section 4. We denote $B \gtrsim B'$ if there exists a universal constant $c_1 > 0$ such that $B \geq c_1 B'$. We use $B \lesssim B'$ if there exists a universal constant $c_2 > 0$ such that $B \leq c_2 B'$. We use $B \asymp B'$ if $B \lesssim B' \lesssim B$. The proof is given in Appendix A.2.

Theorem 1 (Generalization of sub-exponentially stable randomized algorithms)**.** *Consider a learning algorithm $A(S;\theta)$ that satisfies Assumption 1 w.r.t. $\mathbb{P}$ and $c \lesssim \mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]$. Assume $\ell(A(S;\theta), z) \in [0, M]$. Then for any $\delta_1 \in (0, 1)$ the following inequality holds with probability at least $1 - \delta_1$ uniformly for all $\mathbb{Q}$*

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \lesssim \Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta_1)\Big) \max\Big\{\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]\log^2 n, \frac{M}{\sqrt{n}}\Big\},$$

where $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})$ means the KL divergence between $\mathbb{P}$ and $\mathbb{Q}$, i.e., $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{\theta\sim\mathbb{Q}}\Big[\log\frac{\mathbb{Q}(\theta)}{\mathbb{P}(\theta)}\Big]$.

To apply Theorem 1, we only need to check the sub-exponential assumption w.r.t. the prior $\mathbb{P}$. Then, we use the PAC-Bayesian analysis to transfer this stability assumption w.r.t. the specific $\mathbb{P}$ to a bound holding for all posterior distributions.

3.1 Comparison

Next, we compare our result (Theorem 1) with the previous generalization bounds on randomized algorithms due to [32] and observe the advantages that our approach offers.

To blend the stability into the PAC-Bayes framework, where hyperparameters θ follow a distribution, [32] defined the following new hyperparameter stability assumption for learning algorithms (Definition 2) and obtained the stability-based PAC-Bayes bound given in Theorem 2.

Definition 2 (Hyperparameter Stability)**.** A learning algorithm A has uniform hyperparameter stability β_Θ w.r.t. the loss function ℓ, if

$$\sup_{S\in\mathcal{Z}^n}\sup_{z\in\mathcal{Z}}\sup_{\theta,\theta'\in\Theta:D_H(\theta,\theta')=1}|\ell(A(S;\theta),z)-\ell(A(S;\theta'),z)|\leq\beta_\Theta, \tag{3.2}$$

where $D_H(\mathbf{v},\mathbf{v}')\triangleq\sum_{i=1}^{|\mathbf{v}|}\mathbb{I}[v_i\neq v_i']$ is the Hamming distance and $\mathbb{I}[\cdot]$ denotes the indicator function, i.e., $\mathbb{I}[E]=1$ if the event E holds and 0 otherwise.

Observe that, uniform stability (Definition 1) concerns the stability of an algorithm with respect to a change in the training set. In contrast, the above Definition 2 requires stability w.r.t. a change in the hyperparameters. Moreover, the approach in [32] requires stability w.r.t. both the loss function and the hyperparameters to derive the following PAC-Bayes bound.

Theorem 2 (Theorem 2 of [32])**.** *Let A be a randomized learning algorithm and $\ell(A(S;\theta),z)\in[0,M]$. Assume A is β_θ-uniformly stable w.r.t. loss functions, and β_Θ-uniformly stable w.r.t. hyperparamters. Consider the prior $\mathbb{P}$ as a fixed probability distribution defined on $\Theta=\prod_{t=1}^T\Theta_t$. Then for any $n\geq1$, $T\geq1$, and $\delta\in(0,1)$, the following inequality holds with probability at least $1-\delta$ over draws of a data set, $S\sim\mathbb{D}^n$, for every posterior $\mathbb{Q}$ on Θ*

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta))-R_S(A(S;\theta))\big]\lesssim\sqrt{\Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log(1/\delta)\Big)\Big((M+n\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta])^2/n+T\beta_\Theta^2\Big)}.$$

We now compare our bound (Theorem 1) with the bound above (Theorem 2).

In terms of the rate, our bound in Theorem 1 improves the previous result of Theorem 2 by up to a factor of $\sqrt{n}$ (up to poly-logarithmic factors when the divergence is polylogarithmic in n [32]). To see this, consider the case where in both Theorem 1 and Theorem 2, the term that contains the stability parameter dominates over the KL term. Then, in Theorem 1, for β_θ-uniformly stable randomized algorithms, we have

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta))-R_S(A(S;\theta))\big]=\max\Big\{\widetilde{O}(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]),\widetilde{O}(n^{-\frac{1}{2}})\Big\}D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}).$$

In contrast, Theorem 2 in [32] gives

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta))-R_S(A(S;\theta))\big]=\widetilde{O}\big(\sqrt{n}\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]+\sqrt{T}\beta_\Theta\big)D_{\mathrm{KL}}^{\frac{1}{2}}(\mathbb{Q}\|\mathbb{P}). \tag{3.3}$$

It is clear that, in the case when $\widetilde{O}(\sqrt{n}\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta])$ dominates the KL divergence then having replaced it with $\widetilde{O}(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta])$ we improved (3.3) by up to a factor of $\sqrt{n}$ (ignoring a log term). The prior $\mathbb{P}$ can be selected freely if it is independent of S. The posterior $\mathbb{Q}$ can depend on the data. Indeed, a strength of the PAC-Bayesian analysis is that it applies to any $\mathbb{Q}$, and therefore it allows to choose a distribution $\mathbb{Q}$ in a data-dependent manner. The posterior $\mathbb{Q}$ can be optimized to control the divergence between $\mathbb{P}$ and $\mathbb{Q}$. Therefore, we typically have a small KL divergence.

Observe also that, in Theorem 1, when $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]\lesssim O(n^{-\frac{1}{2}})$, our generalization bound is $\widetilde{O}(n^{-\frac{1}{2}})$, while (3.3) implies a vacuous bound $O(1)$ so generalization is not guaranteed. Furthermore, our bounds require $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]\lesssim1/\sqrt{n}$ to get almost optimal rates $\widetilde{O}(1/\sqrt{n})$. As a comparison, the results in Eq. (3.3) require stronger condition $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]\lesssim1/n$ and $\beta_\Theta\leq1/T$ to get the rate $O(1/\sqrt{n})$.

In the unlikely case when the KL divergence is the dominating term, for example, $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})\gtrsim O(n)$, then Theorem 2 achieves better result. But in this circumstance, both Theorems 1 and 2 require $\beta_\theta\lesssim O(n^{-\frac{3}{2}})$ to converge at a rate of $O(n^{-\frac{1}{2}})$, which we believe this to be also rather uncommon.

In terms of assumptions, we eliminate the hyperparameter stability assumption in [32] at the price of an additional sub-exponential tail assumption on the uniform stability parameter. We will prove later in Section 4 that this general assumption holds for more stochastic optimization methods and even for non-smooth problems. In conclusion, we achieve better results under weaker conditions.

3.2 Proof sketch and challenge in the analysis

We begin by noting that ℓ is applied to the output of $A(S;\theta)$, which depends on the training data and the hyperparameters. Therefore, $\ell(A(S;\theta))$ could be sensitive to changes in both the dataset and the hyperparameters. To address this issue, Theorem 2 assumes hyperparameter stability, which requires small changes in $\ell(A(S;\theta))$ when the hyperparameters are perturbed.

In contrast to the previous method, we adopt a different strategy for controlling changes in θ. Our β_θ is a random variable w.r.t. θ. We assume a sub-exponential concentration behavior of β_θ to control its deviation (Assumption 1). This allows us to control β_θ without necessitating hyperparameter stability. Based on this assumption, our proof proceeds by bounding the MGF of the generalization gap. A key challenge is to handle the randomness of β_θ. More precisely, for any temporarily fixed θ it has been shown that $G(S;\theta)$ is a mixture of sub-Gaussian and sub-exponential random variables when only considering the randomness of S [8]. Then a bound of $\mathbb{E}_S[\exp(\lambda G(S;\theta))]$ requires an assumption $\lambda \lesssim 1/\beta_\theta$ (by Eq. (A.5) on sub-exponential random variables). This constraint makes it challenging since we need to choose an appropriate λ and control the associated $\mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{P}}[\exp(\lambda G(S;\theta)]$ (the random constraint makes the selection of λ difficult). Our key idea to address this problem is to control the MGF of another function $H : \mathcal{Z}^n \times \Theta \mapsto \mathbb{R}$ defined as follows

$$H(S;\theta) = \begin{cases} R(A(S;\theta)) - R_S(A(S;\theta)) & \text{if } \theta \in \Omega_\delta, \\ 0 & \text{otherwise,} \end{cases}$$

where Assumption 1 holds everywhere on Ω_δ, which is a subset of Θ with probability measure at least $1-\delta$. This definition of H has two benefits.

• First, $H(S;\theta)$ is equal to $G(S;\theta)$ with high probability. Therefore, a high-probability bound on H would imply a high-probability bound on G.

• Second, H is convenient to control. For any $\theta \in \Theta \setminus \Omega_\delta$, the MGF satisfies $\mathbb{E}_S[\exp(\lambda H(S;\theta)] \leq 1$. For any $\theta \in \Omega_\delta$, the term $\mathbb{E}_S[\exp(\lambda H(S;\theta))]$ can be controlled under an assumption $\lambda \lesssim 1/\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]$ due to the definition of Ω_δ which relates β_θ with its expectation. A key difference is that the random β_θ is replaced by $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]$ in the constraint of λ, which allows us to choose a deterministic λ uniformly for all θ. A further expectation w.r.t. θ would imply a bound on $\mathbb{E}_{S;\theta}[\exp(\lambda H(S;\theta))]$, which is needed in the PAC-Bayes analysis.

Finally, we manipulate this MGF bound in the PAC-Bayes framework by the variational formula (Lemma A.3), building upon the framework of [32].

3.3 Bounds in Expectation

Our previous analysis gives high-probability bounds. In this subsection, we give PAC-Bayes bounds in expectation. In this case, we no longer need an assumption on the concentration behavior of β_θ around its expectation. The proof is given in Appendix A.3.

Theorem 3. *Consider any β_θ-uniformly stable algorithm A and M-bounded loss ℓ. For any distribution $\mathbb{Q}$, we have*

$$\mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \leq \big(\chi^2(\mathbb{Q}\|\mathbb{P}) + 1\big)^{\frac{1}{2}} \left(\frac{2M^2}{n} + 16\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta^2]\right)^{\frac{1}{2}}. \quad (3.4)$$

Remark 1. Under the same assumption, it was shown in Theorem 1 in [32] that

$$\mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \leq \big(\chi^2(\mathbb{Q}\|\mathbb{P}) + 1\big)^{\frac{1}{2}} \left(\frac{2M^2}{n} + 12M\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]\right)^{\frac{1}{2}}. \quad (3.5)$$

Therefore, our analysis gives a bound of the order $O\big(1/\sqrt{n} + \big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta^2]\big)^{\frac{1}{2}}\big)$, while Eq. (3.5) gives a bound $O\big(1/\sqrt{n} + \big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]\big)^{\frac{1}{2}}\big)$. It is known that $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta^2]$ can be much smaller than $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]$. For example, for SGD with t iterations and $\mathbb{P}$ being the uniform distribution, it was shown $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] = O(\eta t/n)$ and $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta^2] = O(\eta^2 t/n)$ if the loss function is convex and smooth [27]. In the typical setting with $\eta = O(1/\sqrt{t})$ and $t \asymp n$ (in this setting SGD achieves optimal rates), we have $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] = O(1/\sqrt{n})$ and $\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta^2] = O(1/n)$. In this case, our analysis gives PAC-Bayes bounds of the order $O(1/\sqrt{n})$, while Eq. (3.5) gives PAC-Bayes bounds of the order $O(1/n^{\frac{1}{4}})$. Therefore, our analysis implies much better PAC-Bayes bounds than that in [32].

4 Applications

We apply our general results to derive PAC-Bayes bounds for two optimization algorithms: Stochastic Gradient Descent (SGD) and Randomized Coordinate Descent (RCD). To this aim, we introduce some necessary definitions. Let $\|\cdot\|_2$ denote the Euclidean norm.

Definition 3 (Lipschitz continuity)**.** We say a loss function $\ell(\cdot; z)$ is L-Lipschitz if for any $\mathbf{w} \in \mathcal{W}$ and $z \in \mathcal{Z}$, we have $\|\nabla \ell(\mathbf{w}; z)\|_2 \leq L$. This implies that for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$,

$$|\ell(\mathbf{w}_1; z) - \ell(\mathbf{w}_2; z)| \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|_2. \tag{4.1}$$

Definition 4 (Convexity)**.** Let $\kappa \geq 0$. We say a loss function $\ell(\cdot; z)$ is κ-strongly convex if for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ and $z \in \mathcal{Z}$, we have

$$\ell(\mathbf{w}_1; z) \geq \ell(\mathbf{w}_2; z) + \left\langle \nabla \ell(\mathbf{w}_2; z), \mathbf{w}_1 - \mathbf{w}_2 \right\rangle + \frac{\kappa}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2. \tag{4.2}$$

We say the loss function ℓ is convex if the above inequality holds with $\kappa = 0$.

Definition 5 (Smoothness)**.** Let $\alpha \geq 0$. We say a loss function $\ell(\cdot; z)$ is α-smooth if for any $\mathbf{w}_1$, $\mathbf{w}_2$ $\in \mathcal{W}$ and $z \in \mathcal{Z}$, we have

$$\|\nabla \ell(\mathbf{w}_1; z) - \nabla \ell(\mathbf{w}_2; z)\|_2 \leq \alpha\|\mathbf{w}_1 - \mathbf{w}_2\|_2. \tag{4.3}$$

For the applications, we only need to verify the sub-exponential stability of the algorithm w.r.t. the prior sampling $\mathbb{P}$, which is often chosen to be simple such as the uniform distribution.

4.1 Applications to Stochastic Gradient Descent

SGD is one of the most popular algorithms to solve optimization problems in machine learning due to its simplicity and efficiency. The basic idea is to build a stochastic gradient based on a randomly selected example, which is used to update iterates. Here we consider SGD with a general sampling scheme, where the random index follows from a general distribution. This general SGD has already been considered in the literature to improve the efficiency of SGD with uniform sampling, including importance sampling [66] and Markov chain sampling [54, 58, 63].

Definition 6 (SGD with general sampling)**.** Let $\mathbf{w}_1$ be an initial point. Let $\mathbb{P}$ be a probability measure over $[n]^T$ and $S = \{z_1, \ldots, z_n\}$ be a training dataset. Let $(i_1, \ldots, i_T)$ be drawn according to $\mathbb{P}$. At the t-th iteration, SGD with sampling scheme $\mathbb{P}$ updates the model by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t; z_{i_t}), \tag{4.4}$$

where $\{\eta_t\}$ is a positive step-size sequence. If $\mathbb{P}$ is the uniform distribution, then we call it SGD with uniform sampling (SGDU).

Now we apply Theorem 1 to develop PAC-Bayes bounds for SGD applied to convex problems, covering both smooth and non-smooth cases. We will show that SGD enjoy sub-exponential stability.

4.1.1 Smooth case

In the following lemma to be proved in Appendix II.1.1, we give stability bounds for SGDU and show it satisfies Assumption 1. Recall that the indicator function $\mathbb{I}[\cdot]$ is defined in Definition 2.

Lemma 4 (Stability bound)**.** *Let S and S' be neighboring datasets. Suppose for all $z \in \mathcal{Z}$ the loss function is convex, α-smooth and L-Lipschitz. Let $\{\mathbf{w}_t\}, \{\mathbf{w}'_t\}$ be the sequence produced by SGDU on S and S' respectively with $\eta_t \leq 2/\alpha$. Then SGDU with t iterations and the hyperparameter θ is β_θ-uniformly stable with*

$$\beta_\theta = 2L^2 \max_{k \in [n]} \sum_{j=1}^{t} \eta_j \mathbb{I}[i_j = k].$$

If $\eta_t = \eta$, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have

$$\beta_\theta \leq \mathbb{E}_{\theta \sim \mathbb{P}}[\beta_\theta] + 4L^2\eta\big(1 + (t/n)^{\frac{1}{2}}\big)\log(1/\delta).$$

That is, Assumption 1 holds with $c = 4L^2\eta\big(1 + (t/n)^{\frac{1}{2}}\big)$ w.r.t. $\mathbb{P}$.

We can combine the above lemma with Theorem 1 to obtain PAC-Bayes bounds for SGD, whose proof is given in Appendix II.1.1. An interesting property is that generalization bounds for SGD with general sampling can be derived based on the stability analysis for SGD with the uniform sampling. We always let $\mathbb{P}$ denote a uniform prior on Θ.

Corollary 5 (Generalization bound)**.** *Assume ℓ is M-bounded, L-Lipschitz, convex and α-smooth. For uniform distribution $\mathbb{P}$, every $n \in \mathbb{N}^+$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of a data set, $S \sim \mathcal{D}^n$, for all posterior sampling distribution $\mathbb{Q}$ on $[n]^T$, SGD with $\eta \leq 2/\alpha$ satisfies*

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \lesssim \Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta)\Big) \max\Big\{L^2\eta(T/n + (1+(T/n)^{\frac{1}{2}})\log(n))\log n, \frac{M}{\sqrt{n}}\Big\}.$$

Based on [20, 27], which studied the trade-off between optimization and stability, the recommended choices of parameters are $T \asymp n$ and $\eta \asymp 1/\sqrt{T}$ to get a SGD iterate with good generalization behavior. In this setting, the above corollary implies a PAC-Bayes bound $\widetilde{O}(1/\sqrt{n})$.

London [32] gave PAC-Bayes bounds for SGD under strong convexity and smoothness assumptions.

Corollary 6 (Corollary 1 in [32])**.** *Suppose ℓ is M-bounded. Let the objective function be κ-strongly convex, L-Lipschitz and α-smooth. Then, for uniform distribution $\mathbb{P}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of a data set, $S \sim \mathcal{D}^n$, SGD with $\eta_t = (\kappa t + \alpha)^{-1}$ and any posterior sampling distribution $\mathbb{Q}$ on $[n]^T$ satisfies*

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \lesssim \sqrt{\Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta)\Big)\Big(\frac{\big(M + L^2/\kappa\big)^2}{n} + \frac{L^2}{\kappa^2 T}\Big)}.$$

Remark 2. We now compare Corollary 5 and Corollary 6. First, Corollary 6 requires a strong convexity assumption, which is removed in our analysis. Second, our analysis implies PAC-Bayes bounds of the order $\widetilde{O}(1/\sqrt{n})$, while Corollary 6 implies bounds of order $O(1/(\sqrt{n}\kappa))$. The strong convexity parameter κ is often very small in both theoretical and empirical analysis. For example, the existing generalization analysis of regularization schemes suggests $\kappa = O(n^{-\frac{1}{2}})$ to get an optimal bound (Section 3 in [55]), for which Corollary 6 implies a vacuous bound. By contrast, our bound $\widetilde{O}(1/\sqrt{n})$ is optimal up to a log factor.

Remark 3. Our uniform stability (Definition 1) is slightly different from the uniform stability $\beta_{\mathcal{Z}} := \sup_{S\sim S'}\sup_z \big|\mathbb{E}_{\theta\sim\mathbb{P}}[\ell(A(S;\theta),z) - \ell(A(S';\theta),z)]\big|$ in [32] in the sense of taking expectation in different places. The expectation of $\theta \sim \mathbb{P}$ is taken outside $\sup$ in our case and inside $\sup$ in $\beta_{\mathcal{Z}}$ [32]. However, we often have similar upper bounds for $\mathbb{E}[\beta_\theta]$ and $\beta_{\mathcal{Z}}$. Consider SGD for smooth, Lipschitz and convex problems as an example. It is shown in [20] that $\beta_{\mathcal{Z}} \leq 2L^2\sum_{k=1}^t \eta_k/n$, while we can show $\mathbb{E}[\beta_\theta] \lesssim L^2\log n\sum_{k=1}^t \eta_k/n$ (the proof of Lemma 4) . These two upper bounds are the same order up to a logarithmic factor. Furthermore, lower bounds were established in [65] (Theorem 1) where $\beta_{\mathcal{Z}} \geq \frac{L}{2}\sum_{k=1}^t \eta_k/n$, which match the existing upper bounds up to a constant factor. This shows that $\beta_{\mathcal{Z}}$ and $\mathbb{E}[\beta_\theta]$ are of similar order.

4.1.2 Non-smooth case

The following lemma shows that SGDU applied to non-smooth problems enjoys the sub-exponential stability. The proof follows the analysis in Section 4.2 in [29] and is given in Appendix II.1.2.

Lemma 7 (Stability bound)**.** *Let S and S' be neighboring datasets. Suppose for all $z \in \mathcal{Z}$ the loss function is convex and L-Lipschitz. Let $\{\mathbf{w}_t\}, \{\mathbf{w}'_t\}$ be the sequence produced by SGDU on S and S' respectively with fixed step sizes. Then SGDU with t iterations and the hyperparameter θ is β_θ-uniformly stable with $\beta_\theta = 2\sqrt{e}L^2\eta\big(\sqrt{t} + \max_{k\in[n]}\sum_{j=1}^t \mathbb{I}[i_j = k]\big)$. For any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\beta_\theta \leq \mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + 4\sqrt{e}L^2\eta\big(1 + (t/n)^{\frac{1}{2}}\big)\log(1/\delta).$$

That is, Assumption 1 holds with $c = 4\sqrt{e}L^2\eta\big(1+(t/n)^{\frac{1}{2}}\big)$ w.r.t. $\mathbb{P}$.

Based on the above lemma, we derive the following corollary for the PAC-Bayes bounds of SGD in non-smooth problems.

Corollary 8 (Generalization bound)**.** *Assume ℓ is M-bounded, L-Lipschitz and convex. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of a data set, $S \sim \mathcal{D}^n$, for all posterior sampling distribution $\mathbb{Q}$ on $[n]^T$, SGD with T iterations and $\eta_t = \eta$ satisfies*

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \lesssim \Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{1}{\delta}\Big)\max\Big\{L^2\eta(\sqrt{T} + T/n + (1+(T/n)^{\frac{1}{2}})\log n)\log n, \frac{M}{\sqrt{n}}\Big\}.$$

Remark 4. If we choose $\eta \asymp T^{-\frac{3}{4}}$ and $T \asymp n^2$, then Corollary 8 gives the PAC-Bayes bounds of order $\widetilde{O}(1/\sqrt{n})$. This choice of parameters was suggested in Theorem 7 in [27]. This gives the $O(1/\sqrt{n})$ optimization bounds to get optimal trade-off between stability and optimization for non-smooth problems. The analysis in [32] cannot imply PAC-Bayes bounds for non-smooth problems.

4.2 Applications to Randomized Coordinate Descent

In this subsection, we consider RCD; this has not been studied in the PAC-Bayesian literature. RCD is an efficient optimization algorithm that randomly chooses a coordinate to update at each iteration [37]. Here we consider RCD with general sampling scheme, i.e. the coordinate to update follows a general distribution. This scheme has been studied before in the optimization context [1, 66].

Definition 7 (RCD with general sampling)**.** Let $\mathbf{w}_1$ be an initial point. Let $\mathbb{P}$ be a probability measure over $[d]^T$ and $S = \{z_1, \ldots, z_n\}$ be a training dataset. Let $(i_1, \ldots, i_T)$ be drawn according to $\mathbb{P}$. At the t-th iteration, RCD with sampling scheme $\mathbb{P}$ updates the model by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla_{i_t} R_S(\mathbf{w}_t)\mathbf{e}_{i_t}, \tag{4.5}$$

where $\{\eta_t\}$ is a step-size sequence, $\mathbf{e}_i$ is the i-th coordinate vector in $\mathbb{R}^d$, and $\nabla_i g$ is the derivative of g w.r.t. the i-th coordinate. If $\mathbb{P}$ is the uniform distribution, then we call it RCD with uniform sampling (RCDU).

Before giving the generalization bound for RCD, we first introduce coordinate-wise smoothness.

Definition 8 ([37])**.** A differentiable function $g : \mathcal{W} \to \mathbb{R}$ has coordinate-wise Lipschitz continuous gradients with parameter $\hat{\alpha} > 0$, if for all $\lambda \in \mathbb{R}$, $\mathbf{w} \in \mathcal{W}$, $i \in [d]$,

$$g(\mathbf{w} + \lambda\mathbf{e}_i) \le g(\mathbf{w}) + \lambda\nabla_i g(\mathbf{w}) + \hat{\alpha}\lambda^2/2.$$

In Lemma 9, to be proved in Appendix B.2, we develop stability bounds for RCDU with convex and smooth loss functions. In particular, we show the stability follows a sub-exponential distribution.

Lemma 9 (Stability bound)**.** *Let S and S' be neighboring datasets. Suppose for all $z \in \mathcal{Z}$ the loss function ℓ is convex, α-smooth, L-Lipschitz and has coordinate-wise Lipschitz continuous gradients with parameter $\hat{\alpha} \ge 0$. Let $\{\mathbf{w}_t\}, \{\mathbf{w}'_t\}$ be the sequence produced by RCDU on S and S' respectively with $\eta_t \le 2/\hat{\alpha}$. Then RCD with t iterations is β_θ-uniformly stable with*

$$\beta_\theta = \frac{L}{n}\max_{k\in[n]}\sum_{j=1}^{t}\eta_j|\nabla_{i_j}\ell(\mathbf{w}_j; z_k) - \nabla_{i_j}\ell(\mathbf{w}_j; z'_k)|, \tag{4.6}$$

where $\|\nabla\ell(\mathbf{w}; z)\|_1 \le L_1$. Furthermore, if $\eta_t = \eta$, then for any $\delta \in (0,1)$ the following inequality holds with probability at least $1-\delta$ over $\theta \sim \mathbb{P}$ (the uniform distribution over $\{(i_1, \ldots, i_t) : i_j \in [d]\}$)

$$\beta_\theta \le \frac{2L_1 L\eta t}{nd} + \frac{\eta L^2\log(1/\delta)}{n}\Big(\frac{8}{3} + \sqrt{\frac{32t}{d}}\Big).$$

Remark 5. Stability bounds of the order $O(\eta t/(nd))$ were developed for RCD in [62] (Theorem 2). Their stability bounds hold in expectation. In contrast, we develop high-probability stability bounds of the order $O\big(\frac{\eta t}{nd} + \frac{\eta}{n}\big)$.

We plug the above bounds into Theorem 1, and derive the following PAC-Bayes bounds for RCD.

Corollary 10 (Generalization bound)**.** *Let the assumptions in Corollary 5 hold. We further assume that the gradient is coordinate-wise Lipschitz continuous. When $\mathbb{P}$ is the uniform distribution, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of $S \sim \mathcal{D}^n$, for all posterior sampling distributions $\mathbb{Q}$ on $[d]^T$, RCD with the hyperparameter θ and fixed step sizes $\eta \le 2/\hat{\alpha}$ satisfies*

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \lesssim \Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta)\Big)\max\Big\{\frac{LL_1\eta T}{nd}\log n, \frac{M}{\sqrt{n}}\Big\}.$$

According to the above corollary, we can derive PAC-Bayes bounds of the order $\widetilde{O}(1/\sqrt{n})$ if $T = O(d\sqrt{n})$ and $\eta = O(1)$. These choices of parameters were suggested in Theorem 7 in [62] to balance optimization and stability for RCD.

5 Related Work

Related work on stability The analysis of the generalization error through algorithmic stability is based on the landmark work of [7]. Generalization bounds via stability are algorithm-specific and have been applied to regularization algorithms, such as SVM regression and classification [7, 11, 16, 52]. Pioneering work on stability analysis of SGD was introduced in [20], which motivated much subsequent stability analyses of randomized iterative algorithms [10, 24, 25, 31, 38, 45]. The smoothness assumption in [20] was removed in more recent stability analyses [3, 27] of SGD. Stability bounds showing the benefit of low training errors on generalization were also developed for convex [27, 39, 50], nonconvex [28] and overparameterized models [12, 47, 59]. In recent works, stronger high-probability bounds via uniform stability have been developed. Some breakthroughs have narrowed down the difference between the risk and the empirical risk, leading to faster convergence rates with high probability [8, 17, 18].

Related work on PAC-Bayes bounds The PAC-Bayes theory of generalization dates back to works by [53] and [34] and further improved by [9, 26] and others. PAC-Bayes bounds are upper bounds on the generalization error of randomized learning algorithms, given in terms of data-dependent quantities that we can compute: the empirical error, and a quantity to measure the divergence between the PAC-Bayesian prior and posterior distributions, such as Kullback-Leibler (KL) divergence or the Rényi divergence [4]. The framework was later extended to allow learning the prior from the data, resulting in tighter bounds [2, 14, 15, 35, 42–44, 48]. The sensitivity of learning algorithms to small perturbations in the weights can be analyzed through the properties of a distribution of predictors, which may lead to regularity and improve the generalization bounds [5]. By evaluating the algorithmic stability on all possible outputs, stability of learning algorithms and PAC-Bayes bounds can be combined [32, 33, 35, 36, 40, 48, 57] and applied to randomized learning algorithms such as SGD and SGLD [30, 32, 35, 36]. For PAC-Bayes bounds, usually, the randomness is induced by a distribution on the parameters of a model. In [32], the authors isolated the randomness to view the randomized learning algorithm as deterministic, with hyperparameters following the distribution instead.

6 Conclusions

Under an assumption of a sub-exponential stability parameter, we derive sharper stability-based PAC-Bayes bounds for randomized learning algorithms by utilizing a moment bound. We show that the sub-exponential stability assumption holds for SGD and RCD, for which we develop PAC-Bayes bounds as corollaries. Our results remove the need for the requirements of strong convexity, hyperparameter stability, and even smoothness in previous results.

Limitations. Future work of interest includes exploring other optimization methods that meet the sub-exponential assumption. It would also be interesting to study the quality of bounds obtainable when placing the PAC-Bayes prior on the parameters of a model (as in the classic approach) versus the hyperparametrs of the optimiser of the model (as in this work). Another interesting research direction is to further improve the bound by including additional assumptions. As noted in [23], for the deterministic case, the bounds can be $\sqrt{n}$-times faster under a Bernstein condition between expectation and variance of loss functions.

Acknowledgements. The authors are grateful to the anonymous reviewers for their thoughtful comments and constructive suggestions. The work of Yunwen Lei is partially supported by the Research Grants Council of Hong Kong [Project No. 22303723]. The work of Sijia Zhou is funded by CSC and UoB scholarship. AK acknowledges funding by EPSRC Fellowship grant EP/P004245/1.

References

[1] Z. Allen-Zhu, Z. Qu, P. Richtárik, and Y. Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pages 1110–1119. PMLR, 2016.

[2] A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter pac-bayes bounds. *Advances in Neural Information Processing Systems*, 19, 2006.

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

[4] L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. Pac-bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444. PMLR, 2016.

[5] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[6] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

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

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

[9] O. Catoni. A pac-bayesian approach to adaptive classification. *preprint*, 840, 2003.

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

[11] J. Chen, H. Chen, X. Jiang, B. Gu, W. Li, T. Gong, and F. Zheng. On the stability and generalization of triplet learning. *arXiv preprint arXiv:2302.09815*, 2023.

[12] P. Deora, R. Ghaderi, H. Taheri, and C. Thrampoulidis. On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*, 2023.

[13] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *33-rd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017.

[14] G. K. Dziugaite and D. M. Roy. Data-dependent pac-bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[15] G. K. Dziugaite, K. Hsu, W. Gharbieh, G. Arpino, and D. Roy. On the role of data in pac-bayes bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 604–612. PMLR, 2021.

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

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

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

[19] D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

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

[21] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[22] A. Kabán. Fractional norm regularization: Learning with very few relevant features. *IEEE Transactions on Neural Networks and Learning Systems*, 24(6):953–963, 2013.

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

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

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

[26] J. Langford and R. Schapire. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6(3), 2005.

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

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

[29] Y. Lei, M. Liu, and Y. Ying. Generalization guarantee of sgd for pairwise learning. *Advances in Neural Information Processing Systems*, 34:21216–21228, 2021.

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

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

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

[33] B. London, B. Huang, and L. Getoor. Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859, 2016.

[34] D. A. McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

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

[36] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

[37] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

[39] K. E. Nikolakakis, F. Haddadpour, A. Karbasi, and D. S. Kalogerias. Beyond lipschitz: sharp generalization and excess risk bounds for full-batch gd. *arXiv preprint arXiv:2204.12446*, 2022.

[40] L. Oneto, M. Donini, M. Pontil, and J. Shawe-Taylor. Randomized learning and generalization of fair and private classifiers: From pac-bayes to stability and differential privacy. *Neurocomputing*, 416:231–243, 2020.

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

[42] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. Pac-bayes bounds with data dependent priors. *The Journal of Machine Learning Research*, 13(1):3507–3531, 2012.

[43] M. Pérez-Ortiz, O. Rivasplata, B. Guedj, M. Gleeson, J. Zhang, J. Shawe-Taylor, M. Bober, and J. Kittler. Learning pac-bayes priors for probabilistic neural networks. *arXiv preprint arXiv:2109.10304*, 2021.

[44] M. Pérez-Ortiz, O. Rivasplata, J. Shawe-Taylor, and C. Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22, 2021.

[45] A. Raj, L. Zhu, M. Gurbuzbalaban, and U. Simsekli. Algorithmic stability of heavy-tailed sgd with general loss functions. In *International Conference on Machine Learning*, pages 28578–28597, 2023.

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

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

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

[49] F. Salehi, L. E. Celis, and P. Thiran. Stochastic optimization with bandit sampling. *arXiv preprint arXiv:1708.02544*, 2017.

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

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

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

[53] J. Shawe-Taylor and R. C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 2–9, 1997.

[54] S. Smale and D.-X. Zhou. Online learning with markov sampling. *Analysis and Applications*, 7(01): 87–113, 2009.

[55] K. Sridharan, S. Shalev-Shwartz, and N. Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, pages 1545–1552, 2009.

[56] S. U. Stich, A. Raj, and M. Jaggi. Safe adaptive importance sampling. *Advances in Neural Information Processing Systems*, 30, 2017.

[57] S. Sun, M. Yu, J. Shawe-Taylor, and L. Mao. Stability-based pac-bayes analysis for multi-view learning algorithms. *Information Fusion*, 86:76–92, 2022.

[58] T. Sun, Y. Sun, and W. Yin. On markov chain gradient descent. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[60] R. Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

[61] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[62] P. Wang, L. Wu, and Y. Lei. Stability and generalization for randomized coordinate descent. In *International Joint Conference on Artificial Intelligence*, pages 3104–3110, 2021.

[63] P. Wang, Y. Lei, Y. Ying, and D.-X. Zhou. Stability and generalization for markov chain stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 37735–37748, 2022.

[64] B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. Mcmahan, O. Shamir, and N. Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020.

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

[66] P. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *International Conference on Machine Learning*, pages 1–9, 2015.

Appendix for "Toward Better PAC-Bayes Bounds for Uniformly Stable Algorithms"

A Proof of PAC-Bayes Bounds for Uniformly Stable Algorithms

A.1 Lemma

In this subsection, we collect several lemmas to prove Theorem 1.

We build on the following lemma from Theorem 4 in [8], which provides a moment inequality for a summation of weakly dependent functions with zero mean and bounded changes under small perturbations. We denote the L_p-norm of a random variable Z by $\|Z\|_p := \left(\mathbb{E}[|Z|^p]\right)^{\frac{1}{p}}, p \geq 1$. Let the set $\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$ be denoted by $S\backslash\{z_i\}$.

Lemma A.1 (Theorem 4 in [8]). *Let $S = \{z_1, \ldots, z_n\}$ be a set of independent random variables that each takes values in $\mathcal{Z}$ and $M > 0$. Let $g_1, \ldots, g_n$ be some functions $g_i : \mathcal{Z}^n \mapsto \mathbb{R}$ such that the following holds for any $i \in [n]$*

- $\left|\mathbb{E}_{S\backslash\{z_i\}}[g_i(S)]\right| \leq M$ *almost surely (a.s.),*
- $\mathbb{E}_{z_i}\left[g_i(S)\right] = 0$ *a.s.,*
- *a.s. for any $j \in [n]$ with $j \neq i$, and $z_j'' \in \mathcal{Z}$*

$$\left|g_i(S) - g_i(z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\right| \leq \beta. \tag{A.1}$$

Then, we can decompose $\sum_{i=1}^n g_i(S)$ as follows

$$\sum_{i=1}^n g_i(S) = X_1 + X_2,$$

where X_1 and X_2 are two random variables satisfying $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$. Furthermore for any $p \geq 1$ we have

$$\|X_1\|_p \leq 4\sqrt{pn}M$$

and for any $p \geq 2$

$$\|X_2\|_p \leq 12\sqrt{2}pn\beta\lceil\log_2 n\rceil.$$

The moment-generating function of a random variable Z is defined as $\mathbb{E}[\exp(\lambda Z)]$, where $\lambda > 0$ is a parameter. The following lemma gives bounds on the MGF of a random variable X once we are given its bounds on the p-norm. The first part considers a sub-gaussian random variable, while the second part considers a sub-exponential random variable.

Lemma A.2 (Prop. 2.5.2 and Prop. 2.7.1 in [61]). *Let X be a random variable and $\mathbb{E}[X] = 0$. Then if the moments of X satisfy*

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq \sqrt{p} \tag{A.2}$$

for all $p \geq 2$, then there exists $K_1 \geq 0$ such that

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(K_1\lambda^2), \tag{A.3}$$

for all $\lambda \in \mathbb{R}$. If the moments of X satisfy

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq p \tag{A.4}$$

for all $p \geq 2$, then

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(2e^2\lambda^2), \quad \forall|\lambda| \leq 1/(2e). \tag{A.5}$$

The following lemma gives a variational formula for the logarithm of moment generating function of $h(\theta)$, which is a foundation for our PAC-Bayesian analysis

Lemma A.3 (Gibbs variational principle [60]). *For any measurable function $h : \Theta \mapsto \mathbb{R}$ we have*

$$\log \mathbb{E}_{\theta\sim\mathbb{P}}[\exp(h(\theta))] = \sup_{\mathbb{Q}} \Big[\mathbb{E}_{\theta\sim\mathbb{Q}}[h(\theta)] - D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})\Big].$$

The following lemma provides an inequality on the change of measure.

Lemma A.4 (Rényi change of measure [4], Theorem 8). *Let X be a random variable and let ψ be a measurable function. Then for any $\alpha > 1$, and any two distributions $\mathbb{P}$ and $\mathbb{Q}$*

$$\log \mathbb{E}_{X\sim\mathbb{Q}}[|\psi(X)|] \le \frac{1}{\alpha}\log \mathbb{E}_{X\sim\mathbb{P}}\Big[\Big(\frac{\mathbb{Q}(X)}{\mathbb{P}(X)}\Big)^{\alpha}\Big] + \frac{\alpha-1}{\alpha}\log \mathbb{E}_{X\sim\mathbb{P}}\big[|\psi(X)|^{\frac{\alpha}{\alpha-1}}\big].$$

In particular, if $\psi(X) = \mathbb{I}_{[X\in A]}$ for a set A, then Lemma A.4 implies that the following inequality for any $\alpha > 1$

$$\log \mathbb{Q}(A) \le \frac{1}{\alpha}\log \mathbb{E}_{X\sim\mathbb{P}}\Big[\Big(\frac{\mathbb{Q}(X)}{\mathbb{P}(X)}\Big)^{\alpha}\Big] + \frac{\alpha-1}{\alpha}\log \mathbb{P}(A),$$

which further shows

$$\mathbb{Q}(A) \le \inf_{\alpha>1}\Big(\mathbb{E}_{X\sim\mathbb{P}}\Big[\Big(\frac{\mathbb{Q}(X)}{\mathbb{P}(X)}\Big)^{\alpha}\Big]\Big)^{\frac{1}{\alpha}} \cdot \mathbb{P}^{\frac{\alpha-1}{\alpha}}(A). \tag{A.6}$$

A.2 Proof of Theorem 1

We are now in a position to prove Theorem 1. Let $e = 2.718...$ be the Euler's number.

Proof of Theorem 1. Let $\delta = 1/n^{\gamma}$, with $\gamma \ge 1$. According to Assumption 1, there exists a set Ω_δ with probability measure $\mathbb{P}(\Omega_\delta) \ge 1-\delta$ such that

$$\beta_\theta \le \mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta), \quad \forall\theta\in\Omega_\delta.$$

We define $H : \mathcal{Z}^n \times \Theta \mapsto \mathbb{R}$ as follows

$$H(S;\theta) = \begin{cases} R(A(S;\theta)) - R_S(A(S;\theta)), & \text{if } \theta\in\Omega_\delta \\ 0, & \text{otherwise.}\end{cases}$$

Since $\ell(A(S;\theta),z)\in[0,M]$, the law of total expectation shows

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\mathbb{Q}}[R(A(S;\theta)) - R_S(A(S;\theta))] \\
&= \mathbb{E}_{\theta\sim\mathbb{Q}}[R(A(S;\theta)) - R_S(A(S;\theta))|\theta\in\Omega_\delta]\mathbb{Q}(\Omega_\delta) + \mathbb{E}_{\theta\sim\mathbb{Q}}[R(A(S;\theta)) - R_S(A(S;\theta))|\theta\in\Omega_\delta^c]\mathbb{Q}(\Omega_\delta^c) \\
&= \mathbb{E}_{\theta\sim\mathbb{Q}}[H(S;\theta)] + \mathbb{E}_{\theta\sim\mathbb{Q}}[R(A(S;\theta)) - R_S(A(S;\theta))|\theta\in\Omega_\delta^c]\mathbb{Q}(\Omega_\delta^c),
\end{aligned} \tag{A.7}$$

where Ω_δ^c denotes the complement of Ω_δ. By Eq. (A.6), we know

$$\mathbb{Q}(\Omega_\delta^c) \le \inf_{\alpha>1}\Big(\mathbb{E}_{\theta\sim\mathbb{P}}\Big[\Big(\frac{\mathbb{Q}(\theta)}{\mathbb{P}(\theta)}\Big)^{\alpha}\Big]\Big)^{\frac{1}{\alpha}} \cdot \mathbb{P}^{\frac{\alpha-1}{\alpha}}(\Omega_\delta^c) \le \inf_{\alpha>1}\delta^{\frac{\alpha-1}{\alpha}}\Big(\mathbb{E}_{\theta\sim\mathbb{P}}\Big[\Big(\frac{\mathbb{Q}(\theta)}{\mathbb{P}(\theta)}\Big)^{\alpha}\Big]\Big)^{\frac{1}{\alpha}}.$$

It then follows that

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[\big|R(A(S;\theta)) - R_S(A(S;\theta))\big|\big] \le \mathbb{E}_{\theta\sim\mathbb{Q}}[H(S;\theta)] + M\inf_{\alpha>1}\delta^{\frac{\alpha-1}{\alpha}}\Big(\mathbb{E}_{\theta\sim\mathbb{P}}\Big[\Big(\frac{\mathbb{Q}(\theta)}{\mathbb{P}(\theta)}\Big)^{\alpha}\Big]\Big)^{\frac{1}{\alpha}}. \tag{A.8}$$

Furthermore, there holds

$$\begin{aligned}
&\mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{P}}[\exp(\lambda H(S;\theta))] \\
&= \mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{P}}[\exp(\lambda H(S;\theta))|\theta\in\Omega_\delta]\mathbb{P}(\Omega_\delta) + \mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{P}}[\exp(\lambda H(S;\theta))|\theta\in\Omega_\delta^c]\mathbb{P}(\Omega_\delta^c) \\
&= \mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{P}}\big[\exp\big(\lambda\big(R(A(S;\theta)) - R_S(A(S;\theta))\big)|\theta\in\Omega_\delta\big)\big]\mathbb{P}(\Omega_\delta) + \exp(0)\mathbb{P}(\Omega_\delta^c) \\
&\le \mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{P}}\big[\exp\big(\lambda\big(R(A(S;\theta)) - R_S(A(S;\theta))\big)|\theta\in\Omega_\delta\big)\big] + \delta.
\end{aligned} \tag{A.9}$$

We now fix any $\theta \in \Omega_\delta$. We denote that $S_i = \{z_1, ..., z_{i-1}, z_i', z_{i+1}, ..., z_n\}$, and $z_i' = (x_i', y_i')$ is an independent copy of (x_i, y_i). It was shown in [8]

$$R(A(S;\theta)) - R_S(A(S;\theta)) \leq 2\beta_\theta + \frac{1}{n}\sum_{i=1}^n g_i, \tag{A.10}$$

where

$$g_i = \mathbb{E}_{z_i'}\big[\mathbb{E}_z[\ell(A(S_i;\theta), z)] - \ell(A(S_i;\theta), z_i)\big].$$

As shown in [8], g_i satisfies all the conditions in Lemma A.1 and therefore one can apply Lemma A.1 to show the existence of two random variables X_1 and X_2 such that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$

$$\frac{1}{n}\sum_{i=1}^n g_i = X_1 + X_2 \tag{A.11}$$

and

$$\|X_1\|_p \leq 4\sqrt{p}Mn^{-\frac{1}{2}}, \quad \forall p \geq 1, \tag{A.12}$$

$$\|X_2\|_p \leq 12\sqrt{2}p\beta_\theta\lceil\log_2 n\rceil \leq 12\sqrt{2}p\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)\big)\lceil\log_2 n\rceil, \quad \forall p \geq 2. \tag{A.13}$$

According to Lemma A.2, we know (we apply Lemma A.2 with $X = X_1/4Mn^{-\frac{1}{2}}$)

$$\mathbb{E}_S[\exp(\lambda X_1)] \leq \exp(16M^2n^{-1}K_1\lambda^2) \tag{A.14}$$

and (we apply Lemma A.2 with $X = X_2/12\sqrt{2}\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)\big)\lceil\log_2 n\rceil$

$$\mathbb{E}_S[\exp(\lambda X_2)] \leq \exp(576e^2\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)\big)^2\lceil\log_2 n\rceil^2\lambda^2), \quad \forall|\lambda| \leq \frac{1}{24e\sqrt{2}\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)\big)\lceil\log_2 n\rceil}. \tag{A.15}$$

By Jensen's inequality

$$\exp(\lambda X_1 + \lambda X_2) = \exp(\lambda X_1)\exp(\lambda X_2) \leq \exp(2\lambda X_1) + \exp(2\lambda X_2), \tag{A.16}$$

we know

$$\begin{aligned}\mathbb{E}_S\big[\exp\big(\lambda\big(R(A(S;\theta)) - R_S(A(S;\theta))\big)\big)\big] &\leq \mathbb{E}_S\big[\exp\big(2\lambda\beta_\theta + \lambda X_1 + \lambda X_2\big)\big] \\ &\leq \exp(2\lambda\beta_\theta)\Big(\mathbb{E}_S[\exp(2\lambda X_1)] + \mathbb{E}_S[\exp(2\lambda X_2)]\Big).\end{aligned} \tag{A.17}$$

This inequality together with Eq. (A.14) and Eq. (A.15) implies the following inequality for all $0 < \lambda \leq \frac{1}{48e\sqrt{2}\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]+c\log(1/\delta)\big)\lceil\log_2 n\rceil}$

$$\begin{aligned}\mathbb{E}_S\big[\exp\big(\lambda\big(R(A(S;\theta)) - R_S(A(S;\theta))\big)\big)\big] \leq \exp(2\lambda(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)))\times \\ \Big(\exp(64M^2n^{-1}K_1\lambda^2) + \exp(576(2e)^2\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)\big)^2\lceil\log_2 n\rceil^2\lambda^2)\Big).\end{aligned}$$

We combine the above inequality with Eq. (A.9) to get

$$\begin{aligned}\mathbb{E}_{\theta\sim\mathbb{P}}\mathbb{E}_S[\exp(\lambda H(S;\theta))] \leq \exp(2\lambda(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)))\times \\ \Big(\exp(64M^2n^{-1}K_1\lambda^2) + \exp(576(2e)^2\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)\big)^2\lceil\log_2 n\rceil^2\lambda^2)\Big) + \delta.\end{aligned} \tag{A.18}$$

For any $u, v, w > 0$ and $\delta \in (0, 1)$, we have

$$\exp(u)(\exp(v) + \exp(w)) + \delta \leq \exp(u + 1/2)(\exp(v) + \exp(w)).$$

Applying this with $u = 2\lambda(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]+c\log(1/\delta))$, $v = 64M^2n^{-1}K_1\lambda^2$, $w = 576(2e)^2\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta]+c\log(1/\delta)\big)^2\lceil\log_2 n\rceil^2\lambda^2$, Eq. (A.18) further implies

$$\begin{aligned}\mathbb{E}_{\theta\sim\mathbb{P}}\mathbb{E}_S[\exp(\lambda H(S;\theta))] \leq \exp\big(2\lambda(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)) + 1/2\big)\times \\ \Big(\exp(64M^2n^{-1}K_1\lambda^2) + \exp(576(2e)^2\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)\big)^2\lceil\log_2 n\rceil^2\lambda^2)\Big).\end{aligned} \tag{A.19}$$

Now, we choose

$$\lambda = \min\Big\{\frac{1}{48e\sqrt{2}\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)\big)\lceil\log_2 n\rceil}, \frac{\sqrt{n}}{8\sqrt{K_1}M}\Big\}. \tag{A.20}$$

Then we have

$$\begin{gathered}
2\lambda(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)) + 1/2 \leq 1,\\
64M^2 n^{-1}K_1\lambda^2 \leq 1,\\
576(2e)^2\big(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(1/\delta)\big)^2\lceil\log_2 n\rceil^2\lambda^2 \leq 1.
\end{gathered}$$

We can combine the above inequalities and Eq. (A.19) to derive

$$\mathbb{E}_{\theta\sim\mathbb{P}}\mathbb{E}_S[\exp(\lambda H(S;\theta))] \leq e(e+e) \leq e^3.$$

Switching to the PAC-Bayes framework, with the above choice of λ, we have by the variational formula (Lemma A.3) that

$$\begin{aligned}
\mathbb{E}_S\big[\exp\big(\sup_{\mathbb{Q}}\big(\lambda\mathbb{E}_{\theta\sim\mathbb{Q}}H(S;\theta) - D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})\big)\big)\big] &\leq \mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{P}}[\exp(\lambda H(S;\theta))]\\
&= \mathbb{E}_{\theta\sim\mathbb{P}}\mathbb{E}_S[\exp(\lambda H(S;\theta))] \leq e^3.
\end{aligned}$$

By Markov's inequality, we further get

$$\begin{aligned}
&\Pr\{\sup_{\mathbb{Q}}[\mathbb{E}_{\theta\sim\mathbb{Q}}\big[\lambda H(S;\theta)\big] - D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})] - 3 \geq \epsilon\}\\
&= \Pr\{\exp\big(\sup_{\mathbb{Q}}[\mathbb{E}_{\theta\sim\mathbb{Q}}\big[\lambda H(S;\theta)\big] - D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})] - 3\big) \geq \exp(\epsilon)\}\\
&\leq \mathbb{E}_S\big[\exp\big(\sup_{\mathbb{Q}}\big(\lambda\mathbb{E}_{\theta\sim\mathbb{Q}}\big[H(S;\theta)\big] - D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})\big) - 3\big)\big]\exp(-\epsilon)\\
&\leq \exp(3-3)\exp(-\epsilon) = \exp(-\epsilon).
\end{aligned}$$

We choose $\epsilon = \log(1/\delta_1)$ and derive the following inequality with probability at least $1-\delta_1$

$$\sup_{\mathbb{Q}}\big[\mathbb{E}_{\theta\sim\mathbb{Q}}\big[\lambda H(S;\theta)\big] - D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) - 3\big] \leq \log(1/\delta_1).$$

Therefore, with probability at least $1-\delta_1$ the following inequality holds uniformly for all $\mathbb{Q}$

$$\mathbb{E}_{\theta\sim\mathbb{Q}}[H(S;\theta)] \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta_1) + 3}{\lambda}.$$

This together with Eq. (A.7) with $\delta = 1/n^\gamma$ gives with probability at least $1-\delta_1$ the following inequality for all $\mathbb{Q}$

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta_1) + 3}{\lambda} + M\inf_{\alpha>1} n^{\frac{(1-\alpha)\gamma}{\alpha}}\Big(\mathbb{E}_{\theta\sim\mathbb{P}}\Big[\Big(\frac{\mathbb{Q}(\theta)}{\mathbb{P}(\theta)}\Big)^\alpha\Big]\Big)^{\frac{1}{\alpha}}\\
&\leq \Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta_1) + 3\Big)\max\Big\{48e\sqrt{2}(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + \gamma c\log(n))\lceil\log_2 n\rceil, \frac{8\sqrt{K_1}M}{\sqrt{n}}\Big\}\\
&+ M\inf_{\alpha>1} n^{\frac{(1-\alpha)\gamma}{\alpha}}\Big(\mathbb{E}_{\theta\sim\mathbb{P}}\Big[\Big(\frac{\mathbb{Q}(\theta)}{\mathbb{P}(\theta)}\Big)^\alpha\Big]\Big)^{\frac{1}{\alpha}}.
\end{aligned}$$

If we take $\gamma = 4$, we get

$$\begin{aligned}
\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] &\lesssim M\inf_{\alpha>1} n^{\frac{4(1-\alpha)}{\alpha}}\Big(\mathbb{E}_{\theta\sim\mathbb{P}}\Big[\Big(\frac{\mathbb{Q}(\theta)}{\mathbb{P}(\theta)}\Big)^\alpha\Big]\Big)^{\frac{1}{\alpha}}\\
&+ \Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta_1)\Big)\max\Big\{(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(n))\lceil\log_2 n\rceil, \frac{M}{\sqrt{n}}\Big\}.
\end{aligned}$$

Recall that the χ^2 divergence between $\mathbb{P}$ and $\mathbb{Q}$ is defined as follows [4]

$$\chi^2(\mathbb{Q}\|\mathbb{P}) \triangleq \mathbb{E}_{\theta\sim\mathbb{P}}\left[\left(\frac{\mathbb{Q}(\theta)}{\mathbb{P}(\theta)}\right)^2 - 1\right].$$

Therefore, if we choose $\alpha = 2$ in the above inequality, we get

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \lesssim Mn^{-2}\Big(\chi^2(\mathbb{Q}\|\mathbb{P}) + 1\Big)^{\frac{1}{2}} + \Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta_1)\Big)\max\Big\{(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(n))\lceil\log_2 n\rceil, \frac{M}{\sqrt{n}}\Big\}.$$

It is reasonable to assume $Mn^{-2}\big(\chi^2(\mathbb{Q}\|\mathbb{P}) + 1\big)^{\frac{1}{2}}$ is negligible (actually one can replace this term as $Mn^{-\gamma/2}\big(\chi^2(\mathbb{Q}\|\mathbb{P}) + 1\big)^{\frac{1}{2}}$ for any $\gamma > 1$) as compared to the second term in the above inequality, and in this case our analysis shows

$$\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \lesssim \Big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta_1)\Big)\max\Big\{(\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + c\log(n))\lceil\log_2 n\rceil, \frac{M}{\sqrt{n}}\Big\}.$$

The proof is completed. □

A.3 Proof of Theorem 3

We now present the proof of Theorem 3. To this aim, we need to introduce two lemmas. The following lemma [32] bounds the generalization gap with the second moment of the generalization error. Recall that $G(S;\theta) \triangleq R(A(S;\theta)) - R_S(A(S;\theta))$.

Lemma A.5 (Appendix B.2 in [32]). *For any two distributions, $\mathbb{P}$ and $\mathbb{Q}$, we have*

$$\mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{Q}}\big[R(A(S;\theta)) - R_S(A(S;\theta))\big] \leq \sqrt{(\chi^2(\mathbb{Q}\|\mathbb{P}) + 1)\,\mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{P}}\left[G(S,\theta)^2\right]}. \tag{A.21}$$

The following lemma shows that the second moment of the generalization error can be controlled by the uniform stability.

Lemma A.6 (Theorem 1.2 in [17]). *For any β_θ-uniformly stable algorithm and M-bounded loss ℓ, we have*

$$\mathbb{E}_S\left[G(S,\theta)^2\right] \leq \frac{2M^2}{n} + 16\beta_\theta^2. \tag{A.22}$$

Proof of Theorem 3. According to the above lemma and linearity of expectation, we take the expectation of both sides of Eq. (A.22)

$$\mathbb{E}_S\mathbb{E}_{\theta\sim\mathbb{P}}\left[G(S,\theta)^2\right] \leq \frac{2M^2}{n} + 16\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta^2]. \tag{A.23}$$

Plugging Eq. (A.23) into Eq. (A.21), we get Eq. (3.4). □

B Proofs on Applications

Here, we apply Theorem 1 to different stochastic optimization methods such as SGD and RCD.

B.1 Stochastic gradient descent

First, we present the proofs on the generalization bounds for SGD with smooth and non-smooth convex loss functions. Before that, we need to prove that SGD meets Assumption 1.

The following concentration inequality is useful to bound the summation of i.i.d. events [51].

Lemma B.1 (Chernoff's Bound). *Let $X_1, \ldots, X_t$ be independent random variables taking values in $\{0,1\}$. Let $X = \sum_{k=1}^t X_k$ and $\mu = \mathbb{E}[X]$. Then for any $\epsilon > 0$ with probability at least $1 - \exp(-\mu\epsilon^2/(2+\epsilon))$ we have $X \leq (1+\epsilon)\mu$. Furthermore, for any $\delta \in (0,1)$ with probability at least $1-\delta$ we have*

$$X \leq \mu + \log(1/\delta) + \sqrt{2\mu\log(1/\delta)}.$$

To prove Lemma 4, we introduce a useful lemma due to Hardt et al. [20].

Lemma B.2 (Lemma 3.7 in [20]). *Suppose the function ℓ is κ-strongly convex and α-smooth. Then for all $\mathbf{w}_1$, $\mathbf{w}_2 \in \mathcal{W}$ and $\eta \leq 2/\alpha$, we have*

$$\|\mathbf{w}_1 - \eta\nabla\ell(\mathbf{w}_1) - \mathbf{w}_2 + \eta\nabla\ell(\mathbf{w}_2)\|_2 \leq \Big(1 - \frac{\alpha\eta\kappa}{\alpha+\kappa}\Big)\|\mathbf{w}_1 - \mathbf{w}_2\|_2.$$

When the loss function ℓ is convex, the above inequality holds with $\kappa = 0$, i.e.,

$$\|\mathbf{w}_1 - \eta\nabla\ell(\mathbf{w}_1) - \mathbf{w}_2 + \eta\nabla\ell(\mathbf{w}_2)\|_2 \leq \|\mathbf{w}_1 - \mathbf{w}_2\|_2.$$

II.1.1 Smooth case

Based on Lemma B.2, we then present the proof of stability bounds.

Proof of Lemma 4. We first assume S and S' differ by the last example. Let us consider two cases. We first consider the case $i_t \in [n]$ and $i_t \neq n$. In this case, according to the SGD update and Lemma B.2, we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 = \|\mathbf{w}_t - \eta_t\nabla\ell(\mathbf{w}_t; z_{i_t}) - \mathbf{w}'_t + \eta_t\nabla\ell(\mathbf{w}'_t; z_{i_t})\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2.$$

We now consider the case when $i_t = n$. In this case, we know

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t - \eta_t\nabla\ell(\mathbf{w}_t; z_{i_t}) + \eta_t\nabla\ell(\mathbf{w}'_t; z'_{i_t})\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2\eta_t L,$$

where we use $\|\nabla\ell(\mathbf{w}_t; z_{i_t})\|_2 \leq L$ due to the L-Lipschitzness. As a combination of the above two cases, we derive

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2\eta_t L\mathbb{I}[i_t = n].$$

Taking a summation of the above inequality gives ($\mathbf{w}_1 = \mathbf{w}'_1$)

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq 2L\sum_{k=1}^{t}\eta_k\mathbb{I}[i_k = n].$$

Based on the above inequality, we know that SGD is β_θ-uniformly stable with

$$\beta_\theta = 2L^2\max_{k\in[n]}\sum_{j=1}^{t}\eta_j\mathbb{I}[i_j = k]. \tag{B.1}$$

For simplicity, let $\beta_{\theta,k} = 2L^2\sum_{j=1}^{t}\eta_j\mathbb{I}[i_j = k]$ for any $k \in [n]$. Taking the expectation over both sides of above inequality, we derive

$$\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] \geq \mathbb{E}_{\theta\sim\mathbb{P}}[\beta_{\theta,k}] = \frac{2L^2}{n}\sum_{j=1}^{t}\eta_j, \tag{B.2}$$

where $\mathbb{E}[\mathbb{I}[i_j = k]] = 1/n$. Based on the above stability bounds, it remains to show that the stability parameter of SGD meet Assumption 1. According to Lemma B.1 with $X_j = \mathbb{I}[i_j = k]$, we get the following inequality with probability at least $1 - \delta/n$ ($\eta_j = \eta$)

$$\beta_{\theta,k} \leq 2L^2\eta(t/n + \log(n/\delta) + \sqrt{2t/n\log(n/\delta)}). \tag{B.3}$$

By the union of probability, with probability at least $1 - \delta$, Eq. (B.3) holds for all $k \in [n]$. Therefore, with probability at least $1 - \delta$

$$\begin{aligned}\beta_\theta &\leq 2L^2\eta(t/n + \log(n/\delta) + \sqrt{2t/n\log(n/\delta)}) \leq 2L^2\eta(t/n + 2\log(1/\delta) + \sqrt{4t/n\log(1/\delta)}) \\ &\leq 2L^2\eta t/n + 4L^2\eta(1 + \sqrt{t/n})\log(1/\delta) \leq \mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] + 4L^2\eta(1 + \sqrt{t/n})\log(1/\delta),\end{aligned}$$

where we have used $\delta \in (0, 1/n)$ in the second inequality, and Eq. (B.2) in the last inequality. Therefore, Assumption 1 holds with $c = 4L^2\eta(1 + \sqrt{t/n})$.

The proof is completed. □

Proof of Corollary 5. With $A(S;\theta) = \mathbf{w}_T$, it follows from Lemma 4 that SGD with convex and smooth loss functions satisfies Assumption 1. Applying the upper bound on β_θ to Theorem 1, we derive the result. □

II.1.2 Non-smooth case

Proof of Lemma 7. Without loss of generality, we first assume S and S' differ by the last example. We consider two cases. We first consider $i_t \in [n]$ and $i_t \neq n$. In this case, according to the SGD update rule and Section 4.2 in [29], we get

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 &= \|\mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t; z_{i_t}) - \mathbf{w}'_t + \eta_t \nabla \ell(\mathbf{w}'_t; z'_{i_t})\|_2^2 \\ &= \|\mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t; z_{i_t}) - \mathbf{w}'_t + \eta_t \nabla \ell(\mathbf{w}'_t; z_{i_t})\|_2^2 \\ &\leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}'_t, \nabla \ell(\mathbf{w}_t; z_{i_t}) - \nabla \ell(\mathbf{w}'_t; z_{i_t})\rangle + 4\eta_t^2 L^2 \\ &\leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4\eta_t^2 L^2,\end{aligned}$$

where we use the L-Lipschitzness and $\langle \mathbf{w}_t - \mathbf{w}'_t, \nabla \ell(\mathbf{w}_t; z_{i_t}) - \nabla \ell(\mathbf{w}'_t; z_{i_t})\rangle \geq 0$ because of the convexity of ℓ. We now consider the case when $i_t = n$. In this case, we know

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 &\leq \|\mathbf{w}_t - \mathbf{w}'_t - \eta_t \nabla \ell(\mathbf{w}_t; z_{i_t}) + \nabla \ell(\mathbf{w}'_t; z'_{i_t})\|_2^2 \\ &\leq (1+p)\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4(1+p^{-1})\eta_t^2 L^2,\end{aligned}$$

where we use the L-Lipschitzness and the second inequality is due to the standard inequality $(a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2$. As a combination of the above two cases, we derive

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 &\leq (1+p)\mathbb{I}[i_t = n]\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4\mathbb{I}[i_t = n](1+p^{-1})\eta_t^2 L^2 \\ &\quad + (1 - \mathbb{I}[i_t = n])\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4(1 - \mathbb{I}[i_t = n])\eta_t^2 L^2 \\ &\leq (1 + p\mathbb{I}[i_t = n])\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4(1 + p^{-1}\mathbb{I}[i_t = n])\eta_t^2 L^2 \\ &= (1+p)^{\mathbb{I}[i_t = n]}\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4(1 + p^{-1}\mathbb{I}[i_t = n])\eta_t^2 L^2.\end{aligned}$$

It then follows that

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 &\leq 4L^2 \sum_{j=1}^t \eta_j^2 (1 + p^{-1}\mathbb{I}[i_j = n]) \prod_{j'=j+1}^t (1+p)^{\mathbb{I}[i'_j = n]} \\ &\leq 4L^2\eta^2 \prod_{j=1}^t (1+p)^{\mathbb{I}[i_j = n]} \sum_{j=1}^t (1 + p^{-1}\mathbb{I}[i_j = n]) \\ &= 4L^2\eta^2 (1+p)^{\sum_{j=1}^t \mathbb{I}[i_j = n]} \Big(t + p^{-1}\sum_{j=1}^t \mathbb{I}[i_j = n]\Big).\end{aligned}$$

We set $p = 1/\sum_{j=1}^t \mathbb{I}[i_j = n]$ and use the inequality $(1 + 1/x)^x \leq e$ to derive

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 \leq 4eL^2\eta^2 \Big(t + \Big(\sum_{j=1}^t \mathbb{I}[i_j = n]\Big)^2\Big).$$

It then follows that

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq 2\sqrt{e}L\eta \Big(\sqrt{t} + \sum_{j=1}^t \mathbb{I}[i_j = n]\Big).$$

According to the Lipschitz continuity, we further know that SGD is β_θ-uniformly stable with

$$\beta_\theta = 2\sqrt{e}L^2\eta \Big(\sqrt{t} + \max_{k \in [n]} \sum_{j=1}^t \mathbb{I}[i_j = k]\Big). \tag{B.4}$$

For simplicity, let $\beta_{\theta,k} = 2\sqrt{e}L^2\eta\big(\sqrt{t} + \sum_{j=1}^t \mathbb{I}[i_j = k]\big)$. Take the expectation of the above inequality, then gives the following bound

$$\mathbb{E}_{\theta \sim \mathbb{P}}[\beta_{\theta,k}] = 2\sqrt{e}L^2\eta\big(\sqrt{t} + t/n\big).$$

Applying Lemma B.1 to Eq. (B.4), with probability at least $1 - \delta/n$, we have

$$\beta_{\theta,k} \leq 2\sqrt{e}L^2\eta(\sqrt{t} + t/n + \log(n/\delta) + \sqrt{2t/n \log(n/\delta)}).$$

Therefore, with probability at least $1-\delta$, the following inequality holds simultaneously for all $k \in [n]$

$$\beta_{\theta,k} \leq 2\sqrt{e}L^2\eta(\sqrt{t} + t/n + \log(n/\delta) + \sqrt{2t/n\log(n/\delta)}),$$

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

$$\beta_{\theta} \leq 2\sqrt{e}L^2\eta(\sqrt{t} + t/n + 2\log(1/\delta) + \sqrt{4t/n\log(1/\delta)}),$$

where we have used $\delta \in (0, 1/n)$ in the above inequality. Combining the stability bounds above, then we can prove that SGDU with the hyperparameter θ meets the Assumption 1 with

$$\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] \geq 2\sqrt{e}L^2\eta(\sqrt{t} + t/n), \quad c = 4\sqrt{e}L^2\eta(1 + \sqrt{t/n}).$$

The proof is completed. □

Based on the above lemma, we are ready to develop generalization bounds in Corollary 8 for SGD with non-smooth loss functions.

Proof of Corollary 8. With $A(S;\theta) = \mathbf{w}_T$, it is then clear that SGD with convex non-smooth loss functions also satisfies Assumption 1 based on Lemma 7. Therefore, applying the upper bound on β_θ to Theorem 1 derives the result. □

B.2 Randomized coordinate descent

To show that the stability of RCD satisfies Assumption 1, we first introduce a lemma on concentration inequalities of martingales.

Lemma B.3 (Bernstein Inequality for Martingales [19])**.** *Let $Y_1, \ldots, Y_t$ be a sequence of random variables. Consider a sequence of functionals $X_k(Y_1, \ldots, Y_k)$ with X_k taking values in $[0, b]$. Suppose for some $\sigma_t \geq 0$ we have*

$$\sum_{k=1}^{t} \mathbb{V}_{Y_k}[X_k] \leq \sigma_t^2,$$

where $\mathbb{V}_{Y_k}[X_k]$ denotes the variance of X_k w.r.t. Y_k. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have

$$\sum_{k=1}^{t} X_k - \sum_{k=1}^{t} \mathbb{E}_{Y_k}[X_k] \leq \frac{2b\log(1/\delta)}{3} + \sigma_t\sqrt{2\log(1/\delta)}.$$

Lemma B.4. *Suppose loss function ℓ is convex, α-smooth, and L-Lipschitz. We define the martingale difference sequence*

$$X_t = |\nabla_{i_t}\ell(\mathbf{w}_t; z_n) - \nabla_{i_t}\ell(\mathbf{w}_t; z_n')| - \mathbb{E}_{i_t}[|\nabla_{i_t}\ell(\mathbf{w}_t; z_n) - \nabla_{i_t}\ell(\mathbf{w}_t; z_n')|]. \tag{B.5}$$

For any $\delta \in (0,1)$, with probability at least $1-\delta$ there holds

$$\sum_{k=1}^{t} X_k \leq \frac{4L\log(1/\delta)}{3} + L\sqrt{\frac{8t\log(1/\delta)}{d}}.$$

Proof. It is easy to see that $\mathbb{E}_{i_t}[X_t] = 0$ and we have

$$|X_t| = ||\nabla_{i_t}\ell(\mathbf{w}_t; z_n) - \nabla_{i_t}\ell(\mathbf{w}_t; z_n')| - \mathbb{E}_{i_t}[|\nabla_{i_t}\ell(\mathbf{w}_t; z_n) - \nabla_{i_t}\ell(\mathbf{w}_t; z_n')|]| \leq 2L.$$

Because i_t is drawn from the uniform distribution on $[d]$, there holds

$$\begin{aligned}\mathbb{E}_{i_t}[X_t^2] &\leq \frac{2}{d}\sum_{j=1}^{d}\left(|\nabla_j\ell(\mathbf{w}_t; z_n)|^2 + |\nabla_j\ell(\mathbf{w}_t; z_n')|^2\right) \\ &= \frac{2}{d}(\|\nabla\ell(\mathbf{w}_t; z_n)\|_2^2 + \|\nabla\ell(\mathbf{w}_t; z_n')\|_2^2) \leq \frac{4L^2}{d}.\end{aligned}$$

Applying Lemma B.3 with $b = 2L$ and $\sigma_t = 2L\sqrt{t/d}$ gives the following inequality with probability at least $1-\delta$

$$\sum_{k=1}^{t} X_k \leq \frac{4L\log(1/\delta)}{3} + L\sqrt{\frac{8t\log(1/\delta)}{d}}.$$

The proof is completed. □

Proof of Lemma 9. Without loss of generality, we assume S and S' differ by the last example. According to Theorem 2 in [62], we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + \eta_t \|\nabla_{i_t} R_{S'}(\mathbf{w}_t)\mathbf{e}_{i_t} - \nabla_{i_t} R_S(\mathbf{w}_t)\mathbf{e}_{i_t}\|_2. \tag{B.6}$$

Note that

$$\begin{aligned} |\nabla_{i_t} R_{S'}(\mathbf{w}_t) - \nabla_{i_t} R_S(\mathbf{w}_t)| &= \frac{1}{n}|\nabla_{i_t}\ell(\mathbf{w}_t; z_n) - \nabla_{i_t}\ell(\mathbf{w}_t; z'_n)| \\ &\leq \frac{1}{n}\big(|\nabla_{i_t}\ell(\mathbf{w}_t; z_n)| + |\nabla_{i_t}\ell(\mathbf{w}_t; z'_n)|\big) \end{aligned} \tag{B.7}$$

and

$$\begin{aligned} \mathbb{E}_{i_t}[|\nabla_{i_t}\ell(\mathbf{w}_t; z_n) - \nabla_{i_t}\ell(\mathbf{w}_t; z'_n)|] &\leq \frac{1}{d}\sum_{j=1}^{d}\big(|\nabla_j\ell(\mathbf{w}_t; z_n)| + |\nabla_j\ell(\mathbf{w}_t; z'_n)|\big) \\ &= \frac{1}{d}\big(\|\nabla\ell(\mathbf{w}_t; z_n)\|_1 + \|\nabla\ell(\mathbf{w}_t; z'_n)\|_1\big) \leq \frac{2L_1}{d}. \end{aligned} \tag{B.8}$$

Based on Eq. (B.6) and Eq. (B.7), we know

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 &\leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + \frac{\eta_t}{n}|\nabla_{i_t}\ell(\mathbf{w}_t; z_n) - \nabla_{i_t}\ell(\mathbf{w}_t; z'_n)| \\ &\leq \sum_{k=1}^{t}\frac{\eta_k}{n}|\nabla_{i_k}\ell(\mathbf{w}_k; z_n) - \nabla_{i_k}\ell(\mathbf{w}_k; z'_n)|. \end{aligned} \tag{B.9}$$

For L-Lipschitz continuity, we prove that RCD is β_θ-uniformly stable with

$$\beta_\theta = \frac{L}{n}\max_{k\in[n]}\sum_{j=1}^{t}\eta_j|\nabla_{i_j}\ell(\mathbf{w}_j; z_k) - \nabla_{i_j}\ell(\mathbf{w}_j; z'_k)|. \tag{B.10}$$

For simplicity, let $\beta_{\theta,k} = \frac{L}{n}\sum_{j=1}^{t}\eta_j|\nabla_{i_j}\ell(\mathbf{w}_j; z_k) - \nabla_{i_j}\ell(\mathbf{w}_j; z'_k)|$. Taking the expectation of the above inequality, we get

$$\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_{\theta,k}] \leq \frac{2LL_1}{nd}\sum_{j=1}^{t}\eta_j, \tag{B.11}$$

where we use Eq. (B.8). Next, we show that RCD satisfies the Assumption 1. According to Eq. (B.8) and Eq. (B.9), we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + \frac{\eta_t}{n}X_t + \frac{2L_1\eta_t}{nd}, \tag{B.12}$$

where we introduce $\{X_t\}$ in Eq. (B.5). Taking a summation of the above inequality, we get

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \frac{\eta}{n}\sum_{k=1}^{t}X_k + \frac{2L_1\eta t}{nd},$$

for $\eta = \eta_t$ for all t. According to Lemma B.4, with probability at least $1 - \delta/n$ the following inequality holds

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \frac{2L_1\eta t}{nd} + \frac{\eta L\log(n/\delta)}{n}\Big(\frac{4}{3} + \sqrt{\frac{8t}{d}}\Big).$$

The following inequality holds by the Lipschitz continuity

$$\beta_{\theta,k} \leq \frac{2LL_1\eta t}{nd} + \frac{\eta L^2\log(n/\delta)}{n}\Big(\frac{4}{3} + \sqrt{\frac{8t}{d}}\Big).$$

Therefore, with probability at least $1-\delta$, the following inequality holds simultaneously for all $k \in [n]$

$$\beta_{\theta,k} \leq \frac{2LL_1\eta t}{nd} + \frac{\eta L^2\log(n/\delta)}{n}\Big(\frac{4}{3} + \sqrt{\frac{8t}{d}}\Big),$$

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

$$\beta_\theta \leq \frac{2LL_1\eta t}{nd} + \frac{\eta L^2 \log(n/\delta)}{n}\Big(\frac{4}{3} + \sqrt{\frac{8t}{d}}\Big) \leq \frac{2LL_1\eta t}{nd} + \frac{\eta L^2 \log(1/\delta)}{n}\Big(\frac{8}{3} + \sqrt{\frac{32t}{d}}\Big).$$

Combining the stability bounds above, then we can prove that RCDU with the hyperparameter θ meets the Assumption 1 with

$$\mathbb{E}_{\theta\sim\mathbb{P}}[\beta_\theta] \geq \frac{2LL_1\eta t}{nd}, \quad c = \frac{\eta L^2}{n}\Big(\frac{8}{3} + \sqrt{\frac{32t}{d}}\Big).$$

The proof is completed. □

We now apply the above result to prove Corollary 10.

Proof of Corollary 10. With $A(S;\theta) = \mathbf{w}_T$ and Lemma 9, Assumption 1 holds for RCD with convex and smooth loss functions. The proof is completed by combining the upper bound on β_θ with Theorem 1. □