An Alternative Analysis of High-Probability Generalization Bound for Uniformly Stable Algorithms

Wei Xiong *Yong Lin †

April 19, 2022

Abstract

Estimation of the generalization gap between the empirical risk and population risk is one of the most important problems in the field of learning theory. The most prominent approach for this problem is based on uniform convergence whose bound is related to some complexity measure of the underlying function space. Since such an analysis largely ignores the way the learning algorithm searches the model space, it can sub-optimal for a variety of learning algorithms. Stability analysis is another classical approach to derive generalization bound for stable learning algorithms due to Bousquet and Elisseeff (2002). In this paper, we study the high-probability generalization bound for γ-uniformly stable learning algorithms and present an alternative analysis to derive a bound that matches the best existing result of Bousquet et al. (2020) without any additional assumption. We further extend our analysis with structure information to obtain faster rates and show that our method can have some potential advantages in these cases.

1 Introduction

A key issue in learning theory is the estimation of the generalization gap between the empirical risk evaluated at the training set $\mathcal{S}_n$ and population risk evaluated in terms of the data distribution $\mathcal{D}$. Among the approaches that have been proposed to this problem, one of the most popular is based on uniform convergence of all models $f \in \mathcal{F}$. This theory relates the generalization gap to the complexity of underlying hypothesis space (see Section 2 for a brief introduction). However, in practical applications, one may not search for the entire model space. For instance, SGD searches the parameter space $\Omega \subset \mathbb{R}^d$ along a line. Therefore, uniform convergence over the whole space can be non-optimal in this case because the way the learning algorithm $\mathcal{A}$ searches the model space is largely ignored by the uniform convergence analysis. In this paper, we consider another tool, referred to as *stability analysis* to estimate the generalization gap of the learned model (Bousquet and Elisseeff, 2002; Feldman and Vondrak, 2018, 2019; Bousquet et al., 2020).

*Department of Mathematics, The Hong Kong University of Science and Technology; email: wxiongae@connect.ust.hk

†Department of Computer Science and Engineering, The Hong Kong University of Science and Technology; email: ylindf@connect.ust.hk

Stability of a learning algorithm refers to the changes in the output of the system when we change the input training dataset $\mathcal{S}_n$. Intuitively, a learning algorithm is said to be stable if the learned model does not change "much" when the training dataset is modified. Here, the word "much" is usually characterized by putting an upper bound for the change. As a motivating example, result from VC-theory is useless for k-Nearest Neighbors algorithm (k-NN) whose VC-dimension is known to be infinite. On the contrary, the k-NN algorithm is very stable because of its "locality" and its stability is employed to derive meaningful generalization bound (Rogers and Wagner, 1978).

Stability analysis has been applied to various learning algorithms, including the empirical risk minimization (ERM) method with strongly convex losses (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010), stochastic gradient descent (SGD) with convex and smooth losses (Hardt et al., 2016), and Gibbs distribution with non-convex losses. It was also conjectured by Hardt et al. (2016) that stability can be used to understand the generalization properties of neural network. Despite a handful of seminal works along the line, the most existing bounds are on the expectation or the second moment of the generalization gap over the random choice of the dataset. The seminal work Feldman and Vondrak (2019) significantly improve the high-probability bound for uniformly-stable learning algorithms based on techniques about range reduction and dataset size reduction.

In this paper, we present an alternative approach to derive the high-probability generalization bound without additional assumptions based on controlling the logarithmic moment generating function. Under the same boundedness assumption, the bound we obtain matches that of Bousquet et al. (2020). Then, we rewrite the analysis of Bousquet et al. (2020) by directly controlling the logarithmic moment generating function to obtain a result that is possibly more flexible to use. Moreover, we extend our analysis by additional structure information to obtain faster rates and show that our method can have some potential advantage in these cases.

2 Preliminaries

In this section, we first define the notion of uniform stability, which is first introduced by Bousquet and Elisseeff (2002), and then formulate the problem. Then, we review the results of uniform convergence and results of stability analysis that are closely related to our work.

2.1 Problem Formulation

Suppose that we are interested in a specific joint probability distribution $\mathcal{D}$ over the input space $\mathcal{X}$ and the ouput space $\mathcal{Y}$. We assume that we are given a training set $\mathcal{S}_n$ consisting of n i.i.d. samples drawn from $\mathcal{D}$. We consider an arbitrary randomized learning algorithm $\mathcal{A}$ that maps the training set $\mathcal{S}_n$ to a model $f \in \mathcal{F}$. Given a loss function $\ell : \mathcal{F} \times \mathcal{Y} \to \mathbb{R}$ that measures the loss of model $f \in \mathcal{F}$ on point $z \in \mathcal{Z}$, with slight abuse of notation, we define the *population risk* of $\mathcal{A}(\mathcal{S}_n)$ as

$$\ell(\mathcal{A}(\mathcal{S}_n), \mathcal{D}) := \mathbb{E}_{Z\sim\mathcal{D}}\ell(\mathcal{A}(\mathcal{S}_n), Z). \tag{2.1}$$

While we cannot compute the population risk directly, we can compute the *empirical risk* over the training set $\mathcal{S}_n$ by:

$$\ell(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) := \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}(\mathcal{S}_n), Z_i). \tag{2.2}$$

Our goal is to estimate the generalization gap between the population risk and empirical risk:

$$\Delta_{\mathcal{D}-\mathcal{S}}(\ell(\mathcal{A})) := \ell(\mathcal{A}(\mathcal{S}_n), \mathcal{D}) - \ell(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n). \tag{2.3}$$

Throughout the rest of this paper, we make the following boundedness assumption.

Assumption 1 (Boundedness)**.** *We assume that* $\ell(f, z) \in [0, 1]$ *for all functions* $f \in \mathcal{F}$ *and point* $z \in \mathcal{Z}$*.*

2.2 Uniform Convergence

Although we consider a general $\mathcal{A}$ here, it is useful to consider $\mathcal{A}$ to be the *Empirical Risk Minimization* (ERM) method for a clearer presentation. Formally, ERM method finds the minimizer of the empirical risk by

$$\hat{f} := \operatorname*{argmin}_{f\in\mathcal{F}} \ell(f, \mathcal{S}_n). \tag{2.4}$$

Let f^* be the minimizer of the population risk. We can derive an upper bound for the *excess risk* as follows:

$$\begin{aligned}
\ell(\hat{f}, \mathcal{D}) - \ell(f^*, \mathcal{D}) &= \underbrace{\left(\ell(\hat{f}, \mathcal{D}) - \ell(\hat{f}, \mathcal{S}_n)\right)}_{A} + \underbrace{\left(\ell(\hat{f}, \mathcal{S}_n) - \ell(f^*, \mathcal{S}_n)\right)}_{B} + \underbrace{\left(\ell(f^*, \mathcal{S}_n) - \ell(f^*, \mathcal{D})\right)}_{C} \\
&\le \left(\ell(\hat{f}, \mathcal{D}) - \ell(\hat{f}, \mathcal{S}_n)\right) + \left(\ell(f^*, \mathcal{S}_n) - \ell(f^*, \mathcal{D})\right) \\
&\le 2\sup_{f\in\mathcal{F}} \left|\ell(f, \mathcal{D}) - \ell(f, \mathcal{S}_n)\right|.
\end{aligned} \tag{2.5}$$

We remark that we cannot directly apply concentration inequalities (e.g. Hoeffding's inequality) to $(\ell(\hat{f}, \mathcal{D}) - \ell(\hat{f}, \mathcal{S}_n))$ because $\hat{f}$ depends on the dataset and $\hat{f}(Z_i)$ are no longer independent of each other. Here we are concerning a *uniform convergence* for all $f \in \mathcal{F}$ instead of a fixed one where traditional law of large number applies. The prominent approach is based on a union concentration over an ϵ-cover of $\mathcal{F}$, which typically leads to the following bound:

$$\sup_{f\in\mathcal{F}} \left|\ell(f, \mathcal{D}) - \ell(f, \mathcal{S}_n)\right| \le O\left(\sqrt{\log\frac{N(\mathcal{F}, \epsilon, \rho)}{\delta}\frac{1}{n}}\right),$$

where $N(\mathcal{F}, \epsilon, \rho)$ is the ϵ-covering number of $\mathcal{F}$ with respect to some metric ρ. The bounds obtained through uniform convergence typically depend on the covering number, or more generally, some notion of the complexity of $\mathcal{F}$. However, in practice, algorithms like SGD search a model parameter along a path which do not cover the entire model space $\mathcal{F}$. Since the uniform convergence bounds largely ignore the way where the model is searched, the bounds can be non-optimal for a variety

of learning algorithms. See Section A for examples.

2.3 Uniform Stability

A different approach to derive generalization bound is based on stability analysis. We introduce the notion of algorithmic stability as follows (Bousquet and Elisseeff, 2002; Feldman and Vondrak, 2018, 2019; Bousquet et al., 2020).

Definition 1 (Uniform Stability). *An algorithm $\mathcal{A}$ is γ-uniformly stable if for all $\mathcal{S}_n$ and $\mathcal{S}_n'$ that differ by only one element, it holds that*

$$\sup_{z\in\mathcal{Z}}\left[\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n'\right),z\right)-\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right),z\right)\right]\leq\gamma, \tag{2.6}$$

where $\mathbb{E}_{\mathcal{A}}$ is taken with respect to the randomness of $\mathcal{A}$.

Stability can be used to derive expected generalization bound for a learning algorithm $\mathcal{A}$. In particular, we have the following theorem from Bousquet and Elisseeff (2002). The proof is presented here for completeness.

Theorem 1 (Expected generalization bound by stability). *If a learning algorithm $\mathcal{A}$ is γ-uniformly stable, then we have*

$$\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right),\mathcal{D}\right)\leq\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right),\mathcal{S}_n\right)+\gamma. \tag{2.7}$$

Proof. Consider two independent datasets $\mathcal{S}_n=\{Z_1,\cdots,Z_n\}$ and $\mathcal{S}_n'=\{Z_1',\cdots,Z_n'\}$. Let

$$\mathcal{S}_n^{(i)}=\{Z_1,\cdots,Z_{i-1},Z_{n+1},Z_{i+1},\cdots,Z_n\}$$

where we replace Z_i with Z_{n+1}. Then it holds that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right),\mathcal{D}\right)-\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right),\mathcal{S}_n\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Z_{n+1}}\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right),Z_i\right)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right),Z_i\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Z_{n+1}}\mathbb{E}_{\mathcal{S}_n}\left[\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right),Z_i\right)-\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right),Z_i\right)\right]\leq\gamma
\end{aligned}$$

where the first equality is because Z_i is independent of $\mathcal{S}_n^{(i)}$ so the distribution of $\ell(\mathcal{A}(\mathcal{S}_n^{(i)}),Z_i)$ is the same as that of $\ell(\mathcal{A}(\mathcal{S}_n),Z)$ with $Z\sim\mathcal{D}$. □

This shows that the expected population risk of a γ-uniformly stable algorithm is bounded by the expected empirical risk plus the stability parameter γ where the expectation is for the training set $\mathcal{S}_n$. In this paper, however, we are mainly concerning the high-probability bound of the generalization error that is generally worse than the expected generalization bound. We summarize existing results as follows.

Bousquet and Elisseeff (2002) establishes an upper bound of

$$\Delta_{\mathcal{D}-\mathcal{S}}(\ell(\mathcal{A})) \leq O\left(\gamma\sqrt{n\log(1/\delta)} + \sqrt{\log(1/\delta)/n}\right),$$

which holds with probability at least $1-\delta$. We note that this high-probability bound is larger than the expected generalization error at least by a factor of $\sqrt{n\log(1/\delta)}$. This bound is tight when γ scales as $\frac{1}{n}$ but becomes vacuous when $\gamma \geq 1/\sqrt{n}$. A stronger bound is proved by Feldman and Vondrak (2018) as

$$\Delta_{\mathcal{D}-\mathcal{S}}(\ell(\mathcal{A})) \leq O\left(\sqrt{(\gamma + 1/n)\log(1/\delta)}\right).$$

This bound is non-vacuous for any non-trivial stability parameter $\gamma = o(1)$. Moreover, the overhead of the high-probability bound as compared to the bound in expectation is reduced from $\sqrt{n}$ to $n^{1/4}$. Feldman and Vondrak (2019) further improves the high-probability bound to

$$\Delta_{\mathcal{D}-\mathcal{S}}(\ell(\mathcal{A})) \leq O\left(\gamma\left(\log^2(n) + \log(n)\log(1/\delta)\right) + \sqrt{\log(1/\delta)/n}\right).$$

Remarkably, this bound implies that algorithms with $\gamma = O(1/\sqrt{n})$ enjoy essentially the same generalization error guarantees up to some logarithmic factors as algorithms that output a fixed function, which has uniform stability 0. Moreover, the upper bound is optimal whenever

$$\gamma \leq \frac{\sqrt{\log(1/\delta)}}{\sqrt{n}\log(n/\delta)\log(n)},$$

while in Bousquet and Elisseeff (2002); Feldman and Vondrak (2018), similar guarantees are achieved only when $\gamma = O(1/n)$. Bousquet et al. (2020) further improve the bound to

$$\Delta_{\mathcal{D}-\mathcal{S}}(\ell(\mathcal{A})) \leq O\left(\gamma\log(n)\log(1/\delta) + \sqrt{\log(1/\delta)/n}\right),$$

whose analysis is based on moment bound which implies generalization bound. We remark that the analysis is also much more straightforward as compared to Feldman and Vondrak (2019).

3 Main Results

In this section, we present the main result of this paper.

3.1 Main Result

Theorem 2. *Assume algorithm $\mathcal{A}$ is γ-uniformly stable. Let $\mathcal{S}_n$ be a dataset of n i.i.d. samples drawn from $\mathcal{D}$. Then, with probability at least $1-\delta$, it holds that*

$$\Delta_{\mathcal{D}-\mathcal{S}}(\ell(\mathcal{A})) \leq O\left(\gamma\log(n)\log(1/\delta) + \epsilon_n(\delta)\right), \tag{3.1}$$

where $\epsilon_n(\delta)$ comes from the concentration of

$$\frac{1}{n}\sum_{i=1}^{n}[\mathbb{E}_{S_n^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right),Z_i\right)-\mathbb{E}_{S_n^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right),\mathcal{D}\right)]\leq\epsilon_n(\delta)$$

In particular, due to the boundedness assumption, the Hoeffding's inequality implies that

$$\Delta_{\mathcal{D}-\mathcal{S}}(\ell(\mathcal{A}))\leq O\left(\gamma\log(n)\log(1/\delta)+\sqrt{\log(1/\delta)/n}\right). \tag{3.2}$$

Note that under Assumption 1, the obtained bound matches that of Bousquet et al. (2020). When additional structural assumption is available, our result may be easier to apply because the analysis of $\epsilon_n(\delta)$ is separate and can be controlled by any method.

4 Proof of Theorem 2: The first method via m.g.f.

We prove Theorem 2 in this subsection.

Lemma 1. *Assume that $g(\mathcal{S}_n,z)$ is zero-mean with respect to z for all $\mathcal{S}_n$, i.e.,*

$$\mathbb{E}_{Z\sim\mathcal{D}}g(S_n,Z)=0.$$

Assume also that $g(\mathcal{S}_n,z)$ is an γ-uniformly stable function, i.e., for all $z\in\mathcal{Z}$ and S_n' that differ from $\mathcal{S}_n$ by one element, we have

$$\left|g\left(\mathcal{S}_n;z\right)-g\left(\mathcal{S}_n';z\right)\right|\leq\gamma.$$

Let

$$\bar{g}\left(\mathcal{S}_{n+1}\right)=\frac{1}{n}\sum_{i=1}^{n}\left[g\left(\mathcal{S}_n^{(i)};Z_i\right)-\mathbb{E}_{\mathcal{S}_{n+1}^{(i)}}g\left(\mathcal{S}_n^{(i)};Z_i\right)\right]$$

where $S_n^{(i)}:=\{Z_1,\cdots,Z_{i-1},Z_{n+1},Z_{i+1},\cdots,Z_n\}$. Then, if the samples of $\mathcal{S}_{n+1}$ are i.i.d. drawn from $\mathcal{D}$, it holds that

$$\ln\mathbb{E}_{\mathcal{S}_{n+1}}\exp\left((\lambda/L)\bar{g}\left(\mathcal{S}_{n+1}\right)\right)\leq 0.3\lambda^2\gamma^2+6\lambda^4\gamma^4$$

where $L=\lceil\log_2 n\rceil$. This implies that with probability. at least $1-\delta$, we have

$$\bar{g}\left(\mathcal{S}_{n+1}\right)\leq L\gamma+2.5L\gamma\ln(1/\delta)$$

We now invoke the above lemma to prove the main theorem.

Proof of Theorem 2. We define

$$g\left(\mathcal{S}_n;z\right)=\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right),\mathcal{D}\right)-\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right),z\right),$$

which is mean-zero w.r.t. z. For all $z \in \mathcal{Z}$, we have

$$\begin{aligned}|g(\mathcal{S}_n, z) - g(\mathcal{S}_n', z)| &\leq |\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right), \mathcal{D}\right) - \mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n'\right), \mathcal{D}\right)| + |\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right), z\right) - \mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n'\right), z\right)| \\ &\leq 2\gamma,\end{aligned}$$

where we use $\mathcal{A}$ is γ-uniformly stable. Therefore, $g(\mathcal{S}_n, z)$ is 2γ-uniformly stable and it holds that

$$\bar{g}\left(\mathcal{S}_{n+1}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[g\left(\mathcal{S}_n^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_{n+1}^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right] \leq 2\lceil\log_2 n\rceil\gamma + 5\lceil\log_2 n\rceil\gamma\ln(2/\delta), \tag{4.1}$$

with probability at least $1 - \delta/2$ by Lemma 1. Specifically, we can write $\bar{g}(\mathcal{S}_{n+1})$ as

$$\begin{aligned}\bar{g}(\mathcal{S}_{n+1}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}[\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), \mathcal{D}\right) - \mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), Z_i\right)}_{(i)} \\ \underbrace{- \mathbb{E}_{S_{n+1}^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), \mathcal{D}\right) + \mathbb{E}_{S_{n+1}^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), Z_i\right)]}_{(ii)}.\end{aligned}$$

We note that the analysis of (ii) is separate and since $\mathbb{E}_{S_n^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), Z_i\right)$ are independent of each other, it is simply a concentration problem with bound $\epsilon(\delta)$. In particular, due to the boundedness assumption, we can apply Hoeffding's inequality to obtain that

$$\frac{1}{n}\sum_{i=1}^{n}[\mathbb{E}_{S_n^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), Z_i\right) - \mathbb{E}_{S_n^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), \mathcal{D}\right)] \leq \sqrt{\frac{\log(2/\delta)}{2n}}, \tag{4.2}$$

with probability at least $1 - \delta/2$. By a union bound, we conclude that with probability at least $1 - \delta$, it holds that

$$\begin{aligned}&\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right), \mathcal{D}\right), \\ \leq&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Z\sim\mathcal{D}}\left[\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right); Z\right)\right] + \gamma, \\ \leq&\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right); Z_i\right)\right] + 2\lceil\log_2 n\rceil\gamma(1 + 2.5\ln(2/\delta)) + \sqrt{\log(2/\delta)/(2n)} + \gamma, \\ \leq&\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n\right); Z_i\right)\right] + 2\lceil\log_2 n\rceil\gamma(1 + 2.5\ln(2/\delta)) + \sqrt{\log(2/\delta)/(2n)} + 2\gamma,\end{aligned}$$

where the first inequality and the third inequality use $\mathcal{A}$ is γ-uniformly stable, and the second inequality is because Eqn. (4.1) and (4.2). □

5 Deriving generalization bound via moment

We present the analysis from Bousquet et al. (2020) in this section with modification for a clear presentation. We define $\|Y\|_p = (\mathbb{E}|Y|^p)^{1/p}$ and $\|Y\|_p(X) = (\mathbb{E}[|Y|^p|X])^{1/p}$ We start with presenting the key steps in their proof.

Lemma 2 (Equivalence of tails and moments)**.** *Suppose that*

$$Y \leq a\sqrt{\log(e/\delta)} + b\log(e/\delta),$$

with probability at least $1-\delta$ *for any* $\delta \in (0,1)$ *for some* $a, b > 0$*. Then, for any* $p \geq 1$*, we have*

$$\|Y\|_p \leq 3\sqrt{p}a + 9pb.$$

And vice versa, if $\|Y\|_p \leq 3\sqrt{p}a + 9pb$ *for all* $p \geq 1$*, then for any* $\delta \in (0,1)$*, we have with probability at least* $1-\delta$*,*

$$|Y| \leq e\left(a\sqrt{\log\left(\frac{e}{\delta}\right)} + b\log\left(\frac{e}{\delta}\right)\right).$$

We also need the following Concentration inequality for the function with bounded difference property.

Lemma 3 (McDiarmid's inequality)**.** *Let* X_i *be independent random variables and* $f : \mathcal{X}^n \to \mathbb{R}$ *satisfy the bounded difference property:*

$$\left|f\left(x_1, \ldots, x_n\right) - f\left(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n\right)\right| \leq \gamma.$$

Then, it holds that

$$\left\|f\left(X_1, \ldots, X_n\right) - \mathbb{E}f\left(X_1, \ldots, X_n\right)\right\|_p \leq 2\sqrt{np}\gamma.$$

In particular, if $X_i \in [0,1]$ *and* $\mathbb{E}X_i = 0$*, we have*

$$\left\|\sum_{i=1}^n X_i\right\|_p \leq 2\sqrt{np}.$$

5.1 Decomposition of the Generalization Error

According to Lemma 2, we aim to derive an upper bound for the p-th moment of the random variables. To be specific, we define

$$g_i = g_i(\mathcal{S}_n) = \mathbb{E}_{Z_i'}\left(\ell(\mathcal{A}(\mathcal{S}^{(i)}), \mathcal{D}) - \ell(\mathcal{A}(\mathcal{S}^{(i)}), Z_i)\right).$$

Then, we have the following results.

Theorem 3 (Bousquet et al. (2020)). *We aim to prove that for all $p \geq 2$, we have*

$$\left\| \sum_{i=1}^{n} g_i(Z) \right\|_p \leq 12\sqrt{2}pn\beta \lceil \log_2 n \rceil + 4M\sqrt{pn}.$$

Under the boundedness assumption, it implies that

$$\ell(\mathcal{A}(\mathcal{S}_n), \mathcal{D}) \leq \ell(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) \leq O\left(\gamma \log(n) \log(1/\delta) + \sqrt{n \log(1/\delta)}\right).$$

Sketch proof of Theorem 3. We start with a similar decomposition as in the proof of Theorem 1:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}} \ell\left(\mathcal{A}\left(\mathcal{S}_n\right), \mathcal{D}\right) - \mathbb{E}_{\mathcal{A}} \ell(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) \\
\leq & \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}_{Z_i'} \mathbb{E}_{\mathcal{A}} \ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right); \mathcal{D}\right)\right] - \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}_{Z_i'} \mathbb{E}_{\mathcal{A}} \ell(\mathcal{A}(\mathcal{S}_n^{(i)}); Z_i)\right] + 2\gamma, \\
= & \frac{1}{n} \sum_{i=1}^{n} g_i + 2\gamma.
\end{aligned}$$

It remains to derive an high-probability upper bound for $\sum_{i=1}^{n} g_i$. By Lemma 2, it suffices to control the p-th moments of $\sum_{i=1}^{n} g_i$. W.L.O.G., we assume that $n = 2^k$. Otherwise, we can add extra null samples, increasing the number of terms by at most two times. We construct a sequence of partitions of the dataset $\mathcal{S}_n$: $\mathcal{B}_0, \mathcal{B}_1, \cdots, \mathcal{B}_k$ as follows:

$$\mathcal{B}_0 = \{\{1\}, \{2\}, \cdots, \{2^k\}\}, \qquad \mathcal{B}_1 = \{\{1,2\}, \{3,4\}, \cdots, \{2^k - 1, 2^k\}\}, \qquad \mathcal{B}_k = \{\{1, 2, \cdots, 2^k\}\}$$

In other words, to get $\mathcal{B}_l$ from $\mathcal{B}_{l+1}$, we split each subset of $\mathcal{B}_{l+1}$ into two equal parts. By construction, we have $\mathcal{B}_k = 1$, $|\mathcal{B}_0| = 2^k$, and $|\mathcal{B}_l| = 2^{k-l}$. For each $l \in [k]$ (index of partitions), and each $i \in [n]$ (index of sample), we define $B^l(i)$ as the subset that contains i. For instance, $B^0(i) = \{i\}$ and $B^k(i) = \{1, \cdots, n\}$. We also define

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

where we take expectation with all the samples in the subset containing i at partition $\mathcal{B}_l$, except for i. We have

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

Therefore, by triangle inequality, we have

$$\left\| \sum_{i=1}^{n} g_i \right\|_p \leq \underbrace{\left\| \sum_{i=1}^{n} \mathbb{E}\left[g_i \mid Z_i\right] \right\|_p}_{(i)} + \underbrace{\sum_{l=0}^{k-1} \left\| \sum_{i=1}^{n} g_i^l - g_i^{l+1} \right\|_p}_{(ii)}.$$

Then, we control (i) and (ii) to fit the framework of Lemma 2.

Bounding (i)**.** We remark that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[g_i|Z_i] = \frac{1}{n}\sum_{i=1}^{n}[\mathbb{E}_{S_n^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), Z_i\right) - \mathbb{E}_{S_n^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), \mathcal{D}\right)]$$

is exactly the term we are concerning for $\epsilon(\delta)$ in Theorem 2. By $|\mathbb{E}[g_i|Z_i]| \leq 1$ and $\mathbb{E}(\mathbb{E}[g_i|Z_i]) = 0$, by Lemma 3, we have

$$\left\|\sum_{i=1}^{n}\mathbb{E}\left[g_i \mid Z_i\right]\right\|_p \leq 4\sqrt{pn}.$$

Bounding (ii)**.** The analysis of (ii) is more involved. The key idea is the following decomposition:

$$\sum_{l=0}^{k-1}\left\|\sum_{i=1}^{n} g_i^l - g_i^{l+1}\right\|_p \leq \sum_{l=0}^{k-1}\sum_{j=1}^{2^{k-j}}\left\|\sum_{i\in B_j^l} g_i^l - g_i^{l+1}\right\|_p.$$

We note that $g_i^l - g_i^{l+1}$ only depends on $Z_i, Z_{[n]/B^l}$. In particular, it only depends on B^l through Z_i. Conditioned on the samples outside B^l, i.e., $Z_{[n]/B^l}$, they are independent and mean-zero. Therefore, we can apply the following lemma:

Lemma 4. *Let X_i be independent random variables with finite p-th moment for $p \geq 2$. Then,*

$$\left\|\sum_{i=1}^{n} X_i\right\|_p \leq 3\sqrt{2np}\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|_p^p\right)^{\frac{1}{p}}$$

to obtain

$$\left\|\sum_{i\in B^l} g_i^l - g_i^{l+1}\right\|_p^p\left(Z_{[n]\setminus B^l}\right) \leq \left(3\sqrt{2p2^l}\right)^p\frac{1}{2^l}\sum_{i\in B^l}\left\|g_i^l - g_i^{l+1}\right\|_p^p\left(Z_{[n]\setminus B^l}\right), \tag{5.1}$$

where we take p-power on both sides. It is also not hard to see g_i^l preserves the bounded difference property. We have

$$\left\|g_i^l - g_i^{l+1}\right\|_p\left(B^{l+1}(i)/B^l(i)\right) \leq 2\sqrt{p2^l}\gamma, \quad \forall p \geq 2$$

where we use Lemma 3 with $n = 2^l$ because $|B^{l+1}(i)/B^l(i)| = 2^l$. It further holds that

$$\left\|g_i^l - g_i^{l+1}\right\|_p = (\mathbb{E}\mathbb{E}[|g_i^l - g_i^{l+1}|^p|B^{l+1}(i)/B^l(i)])^{1/p} \leq 2\sqrt{p2^l}\gamma.$$

Combining this with Eqn.(5.1), and integrating w.r.t. $Z_{[n]/B^l}$, we have

$$\left\|\sum_{i\in B^l} g_i^l - g_i^{l+1}\right\|_p \leq 3\sqrt{2p2^l}\times 2\sqrt{p2^l}\gamma = 6\sqrt{2}p2^l\gamma.$$

Therefore, by triangle inequality, we have

$$\left\|\sum_{i\in[n]} g_i^l - g_i^{l+1}\right\|_p \le \sum_{B^l\in\mathcal{B}_l}\left\|\sum_{i\in B^l} g_i^l - g_i^{l+1}\right\|_p \le 2^{k-l}\times 6\sqrt{2}p2^l\gamma = 6\sqrt{2}p2^k\gamma \le 12\sqrt{2}pn\gamma.$$

It follows that

$$\left\|\sum_{i=1}^n g_i\right\|_p \le \left\|\sum_{i=1}^n \mathbb{E}\left[g_i \mid Z_i\right]\right\|_p + \sum_{l=0}^{k-1}\left\|\sum_{i=1}^n g_i^l - g_i^{l+1}\right\|_p \le 4\sqrt{pn} + 12\sqrt{2}pn\gamma\lceil\log_2 n\rceil.$$

This implies the desired generalization bound. □

6 Proof of Theorem 2 (m.g.f. counterpart of Bousquet et al. (2020))

In this section, we rewrite the proof provided in Section 5 where we directly control the moment generating function instead of the moment. We use the same notation as 5. Recall that

$$\sum_{i=1}^n g_i = \underbrace{\sum_{i=1}^n \mathbb{E}\left[g_i \mid Z_i\right]}_{(i)} + \underbrace{\sum_{l=0}^{k-1}\sum_{i=1}^n g_i^l - g_i^{l+1}}_{(ii)}.$$

We know that $(i) \le n\epsilon_n(\delta)$ with high probability so it suffices to bound (ii).

Alternative proof of Theorem 2. Observe that

$$g_i^{l+1}(Z_i, Z_{[n]/B_i^{l+1}}) = \mathbb{E}[g_i^l(Z_i, Z_{[n]/B^l(i)})|Z_i, Z_{[n]/B^{l+1}(i)}],$$

that is, the expectation is taken w.r.t. the variables $Z_j, j \in B^{l+1}(i)/B^l(i)$. We obtain the uniform bound

$$\ln \mathbb{E}\exp(\lambda \sum_{i\in\in B_l^j} g_i^l - g_l^{l+1}) = 2^l \ln \mathbb{E}\exp(\lambda(g_i^l - g_i^{l+1})) \le \lambda^2 2^{2l}\gamma^2/8,$$

where the first inequality is due to $g_i^l - g_i^{l+1}$ for $i \in B^l$ depends only on $Z_i, Z_{[n]/B^{l+1}(i)}$, the terms are independent and zero mean conditioned on $Z_{[n]/B^l}$. The first inequality is due to the McDiarmid's inequality conditioned on $Z_i, Z_{[n]/B^{l+1}(i)}$ because g_i^l preserves the bounded differences property (by

γ) as the function g_i. We then have

$$\begin{aligned}\ln \mathbb{E} \exp(\lambda \sum_{l=0}^{k-1} \sum_{j=1}^{2^{k-l}} \sum_{i \in B_l^j} g_i^l - g_i^{l+1}) \leq & \frac{1}{k} \sum_{k=0}^{k-1} \ln \mathbb{E} \exp(\lambda k \sum_{j=1}^{2^{k-l}} \sum_{i \in B_l^j} g_i^l - g_i^{l+1}) \\ \leq & \frac{1}{k} \sum_{k=0}^{k-1} \frac{1}{2^{k-l}} \sum_{j=1}^{2^{k-l}} \ln \mathbb{E} \exp(\lambda k 2^{k-l} \sum_{i \in B_l^j} g_i^l - g_i^{l+1}) \\ \leq & \lambda^2 k^2 2^{2k-2l} 2^{2l} \gamma^2 / 8 \\ = & n^2 (\log_2 n)^2 \lambda^2 \gamma^2 / 8,\end{aligned}$$

where in the first and second inequalities we use Jesen's Inequality. The last inequality is due to $k = \log_2 n$. It follows that

$$\ln P \left(\sum_{l=0}^{k-1} \sum_{i=1}^{n} g_i^l - g_i^{l+1} \geq n \log_2 n (1+\epsilon')\gamma \right) \leq n^2 (\log_2 n)^2 \lambda^2 \gamma^2 / 8 - \lambda n \log_2 n (1+\epsilon')\gamma.$$

Taking $\lambda = \frac{c}{n \log n \gamma}$ with $0 < c < 1$, we obtain that

$$\ln P \left(\sum_{l=0}^{k-1} \sum_{i=1}^{n} g_i^l - g_i^{l+1} \geq n \log_2 n (1+\epsilon')\gamma \right) \leq c^2/8 - c - c\epsilon' \leq -c\epsilon'.$$

Setting $\delta/2 = \exp(-c\epsilon')$, we obtain that

$$\sum_{l=0}^{k-1} \sum_{i=1}^{n} g_i^l - g_i^{l+1} < n \log_2 n (1 + \frac{1}{c} \ln \frac{2}{\delta})\gamma$$

with probability at least $1 - \delta/2$.
Putting (i) and (ii) together, we have

$$\begin{aligned}\sum_{i=1}^{n} g_i \leq & \sum_{i=1}^{n} \mathbb{E}\left[g_i \mid Z_i\right] + \sum_{l=0}^{k-1} \sum_{i=1}^{n} g_i^l - g_i^{l+1} \\ \leq & n\epsilon_n(\delta) + n \log_2 n (1 + \frac{1}{c} \ln \frac{2}{\delta})\gamma\end{aligned}$$

with probability at least $1 - \delta$. This implies the desired theorem. □

7 Extension with Realizability

In this subsection, we show that our analysis can be easily generalized with additional structural assumption to obtain faster rates.

Corollary 4. *Suppose that the learning algorithm $\mathcal{A}$ is $\gamma = \frac{1}{n}$-uniformly stable and the problem is*

realizable in the sense that for any $\mathcal{S}_n \in \mathcal{Z}^n$, we have

$$\mathbb{E}_{\mathcal{A}}\ell(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) \leq \frac{\log^2(n)\log(2/\delta)}{n},$$

Then, with probability at least $1-\delta$, we have

$$\Delta_{\mathcal{D}-\mathcal{S}}(\ell(\mathcal{A})) \leq O\left(\frac{\log(n)\log(1/\delta)}{n}\right). \tag{7.1}$$

Proof. We show that under the realizability condition, we can obtain a faster rate of $\epsilon_n(\delta)$. We use the short-hand notations: $\mu = \mathbb{E}_{S_n^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), \mathcal{D}\right)$, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_n^{(i)}}\mathbb{E}_{\mathcal{A}}\ell\left(\mathcal{A}\left(\mathcal{S}_n^{(i)}\right), Z_i\right)$. We will use the multiplicative form of Chernoff bound:

$$\bar{X}_n < \mu + \sqrt{\frac{2\mu\log(2/\delta)}{n}} + \frac{\log(2/\delta)}{3n}, \tag{7.2}$$

which holds with probability at least $1-\delta/2$. Therefore, it suffices to bound μ. By the expected generalization bound in Eqn. (2.7), we have

$$\mu \leq \mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\ell(\mathcal{A}(\mathcal{S}_n), \mathcal{S}_n) + \gamma \leq \frac{\log^2(n)\log(2/\delta)}{n} + \frac{1}{n} \leq \frac{2\log^2(n)\log(4/\delta)}{n},$$

where we use the realizability condition in the second inequality. Therefore, we have

$$\bar{X}_n - \mu \leq \frac{3\log(n)\log(4/\delta)}{n}.$$

According to Theorem 2, we know that with probability at least $1-\delta$, it holds that

$$\Delta_{\mathcal{D}-\mathcal{S}}(\ell(\mathcal{A})) \leq O(\frac{\log(n)\log(1/\delta)}{n}).$$

□

8 Conclusion

In this paper, an alternative analysis for the uniformly stable algorithms is presented. The obtained bound matches the best existing high-probability generalization bound of Bousquet et al. (2020). Furthermore, we extend our analysis with realizability condition to obtain a sharper bound. Our analysis can have advantages in some cases with additional structure information.

References

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

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

Feldman, V. and Vondrak, J. (2018). Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31.

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

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR.

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

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

A Examples

In this section, we apply our bounds to several learning algorithms that are known to stable. The detailed proofs are deferred to the Appendix due to space limit.

Regularized Empirical Risk Minimization. (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010) We consider a general ERM method on a closed convex parameter space Ω for convex objectives. The empirical risk now is evaluated with $\bar{\ell}(w,z) = \ell(w,z) + h(w)$ where $h(w) \geq 0$ serves as a regularizer. We assume that $\ell(w,z)$ is $G(z)$-Lipschitz in w and the empirical loss function $\bar{\ell}(w,z)$ is λ-strongly convex. Then, the regularized ERM defined as

$$\mathcal{A}(\mathcal{S}_n) := \operatorname*{argmin}_{w\in\Omega} \bar{\ell}(w, \mathcal{S}_n), \tag{A.1}$$

is uniformly stable with parameter

$$\gamma = \frac{2(\sup_{z\in\mathcal{Z}} G(z))^2}{\lambda n}.$$

Stochastic Gradient Descent (SGD). In many applications, one may run SGD for finite iterations without converging to the minimum solution of ERM. In this case, it is usually much easier to derive generalization bound by stability analysis (Hardt et al., 2016). We make the same assumptions for $\bar{\ell}(w, \mathcal{S}_n)$ and $\ell(w,z)$ as in the regularized ERM example. We define $b_0 = 0$, and

$$b_t = (1 - \eta_t\lambda) b_{t-1} + \frac{2\eta_t}{n} (\sup_{z\in\mathcal{Z}} G(z))^2, \forall t \geq 1.$$

Then, after t iterations, SGD is $\gamma = b_t$-uniformly stable.

Gibbs Distribution. We consider a posterior sampling algorithm, namely, Gibbs Distribution for non-convex objective function. The algorithms will sample a $w \in \Omega$ from the following posterior distribution:

$$p(w|\mathcal{S}_n) \propto p_0(w) \exp(-\beta \sum_{z\in\mathcal{S}_n} \ell(w, Z)), \tag{A.2}$$

where $\beta > 0$ is a tuning parameter. If we assume that $\sup_{w\in\Omega} \ell(w,z) - \inf_{w\in\Omega} \ell(w,z) \leq M$ for all $z \in \mathcal{Z}$, then Gibbs Distribution is $\gamma = (e^{2\beta M} - 1)M$-uniformly stable.

B Proof of Lemma 1

We need the following lemmas.

Lemma 5. *Consider any functions $\tilde{g}(Z)$ where $S = [Z_1, \cdots, Z_m]$ contains m i.i.d. samples from $\mathcal{D}$. Let $\mathcal{S}^{(i)} = [Z_1, \cdots, Z_{i-1}, Z_i', Z_{i+1}, \cdots, Z_n]$ where $\mathcal{S}'$ contains m i.i.d. samples from $\mathcal{D}$ that are independent of Z. Define*

$$V_+(\mathcal{S}) = \mathbb{E}_{\mathcal{S}'} \left[\sum_{i=1}^{n} \left(\tilde{g}(\mathcal{S}) - \tilde{g}\left(\mathcal{S}^{(i)}\right) \right)^2 I\left(\tilde{g}(\mathcal{S}) > \tilde{g}\left(\mathcal{S}^{(i)}\right) \mid \mathcal{S} \right] \right.$$

Assume that there exists positive constants a and b such that

$$V_+(\mathcal{S}) \leq a\tilde{g}(\mathcal{S}) + b,$$

then for $\lambda \in (0, 1/a)$:

$$\log \mathbb{E}_{\mathcal{S}} \exp(\lambda \tilde{g}(\mathcal{S})) \leq \lambda \mathbb{E}_{\mathcal{S}} \tilde{g}(\mathcal{S}) + \frac{\lambda^2}{1 - a\lambda}(a\mathbb{E}\tilde{g}(\mathcal{S}) + b).$$

Lemma 6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\{X_i, X_i'\}$ be $2n$ i.i.d. random variables, then*

$$\mathrm{Var}[f(X)] \leq \frac{1}{2}\mathbb{E}\sum_{i=1}^{n}\left(f(X) - f\left(X^{(i)}\right)^2\right)$$

145 *where $X = [X_1, \cdots, X_n]$ and $X^{(i)} = [X_1, \cdots, X_{i-1}, X_i', X_{i+1}, \cdots, X_n]$.*

Proof of xx. For notation simplicity, we assume $g(\mathcal{S}_n, z)$ is invariant to the order of elements in $\mathcal{S}_n$. The analysis itself holds without this assumption. Consider a fixed $\mathcal{S}_{n+1} = \{Z_1, \cdots, Z_{n+1}\}$. With slight abuse of notaiton, we define for $i \leq m \leq n$,

$$\mathcal{S}_m^{(i)} = \{Z_1, \cdots, Z_{i-1}, Z_{n+1}, z_{i+1}, \cdots, Z_m\},$$

and we define $\mathcal{S}_n^{(i)} = \mathcal{S}_{n+1}^{(i)}$. Now for all $1 \leq m' \leq m \leq n$, we define

$$\bar{g}_{m',m}\left(\mathcal{S}_{n+1}\right) = \frac{1}{m'}\sum_{i=1}^{m'}\left[g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_m^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right].$$

Then, we have $\bar{g}(\mathcal{S}_{n+1}) = \bar{g}_{n,n}(\mathcal{S}_{n+1})$. Given $m' \leq m \leq n$, we denote by $\tilde{\mathcal{S}}_{m'}$ a uniformly selected subset of $\mathcal{S}_m$ of size m'. By symmetry, we have

$$\begin{aligned}
\bar{g}_{m,m}\left(\mathcal{S}_{n+1}\right) &= \frac{1}{m}\sum_{i=1}^{m}\left[g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_m^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right] \\
&= \mathbb{E}_{\tilde{\mathcal{S}}_{m'}} \frac{1}{m'} \sum_{Z_i \in \tilde{\mathcal{S}}_{m'}}\left[g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_m^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right]
\end{aligned}$$

This is because for each i, we have

$$\frac{C_{m-1}^{m'-1}}{C_m^{m'}}\frac{1}{m'} = \frac{1}{m}.$$

Then, we have

$$\begin{aligned}
&\ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\lambda' \bar{g}_{m,m}\left(\mathcal{S}_{n+1}\right)\right) \\
\leq &\ln \mathbb{E}_{\mathcal{S}_{n+1}} \mathbb{E}_{\tilde{\mathcal{S}}_{m'}} \exp\left(\frac{\lambda'}{m'} \sum_{Z_i \in \tilde{\mathcal{S}}_{m'}} \left[g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_m^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right]\right) \\
\leq &\ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\lambda' \bar{g}_{m',m}\left(\mathcal{S}_{n+1}\right)\right),
\end{aligned} \tag{B.1}$$

where the first inequality uses Jensen's inequality for $\exp(\cdot)$ and the second inequality is because all $\tilde{\mathcal{S}}_{m'}$ have identical distribution. It can be verified by definition (add and minus $\frac{1}{m'} \sum_{i=1}^{m'} \mathbb{E}_{\mathcal{S}_{m'}^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)$) that

$$\bar{g}_{m',m}\left(\mathcal{S}_{n+1}\right) = \bar{g}_{m',m'}\left(\mathcal{S}_{n+1}\right) + \frac{1}{m'} \sum_{i=1}^{m'} \left[\mathbb{E}_{\mathcal{S}_{m'}^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_m^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right].$$

We then apply the Jensen's inequality applied to $\ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp(g(\mathcal{S}_{n+1}))$ w.r.t. $g(\cdot)$ to obtain for all λ' and $\ell > 1$ that

$$\begin{aligned}
&\ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\lambda' \bar{g}_{m',m}\left(\mathcal{S}_{n+1}\right)\right) \\
\leq &\frac{\ell - 1}{\ell} \ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\ell/(\ell-1)\lambda' \bar{g}_{m',m'}\left(\mathcal{S}_{n+1}\right)\right) \\
&+ \frac{1}{\ell} \ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\frac{\ell \lambda'}{m'} \sum_{i=1}^{m'} \left[\mathbb{E}_{\mathcal{S}_{m'}^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_m^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right]\right).
\end{aligned} \tag{B.2}$$

Now we fix $\{Z_{m+1}, \cdots, Z_n\}$ and consider a function of $\mathcal{S}_{m'+1,\cdots,m}\}$ defined as follows:

$$g'\left(\mathcal{S}_{m'+1,m}\right) = \ln \mathbb{E}_{\mathcal{S}_{m'}} \exp\left(\frac{\lambda}{m'} \sum_{i=1}^{m'} \left[\mathbb{E}_{\mathcal{S}_{m'}^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_m^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right]\right),$$

where we note that is does not depend on Z_{n+1} due to the expectation w.r.t. Z_{n+1} over $\mathcal{S}_{m'}^{(i)}$ and $\mathcal{S}_m^{(i)}$. (Note that we replace Z_i with Z_{n+1} in them!) Conditioned on $\mathcal{S}_{m'+1,m}$, Z_i are independent for $i \leq m'$, so we have

$$g'\left(\mathcal{S}_{m'+1,m}\right) = m' \ln \mathbb{E}_Z \exp\left(\frac{\lambda}{m'} \left[\mathbb{E}_{\mathcal{S}_{m'}} g\left(\mathcal{S}_n; Z\right) - \mathbb{E}_{\mathcal{S}_m} g\left(\mathcal{S}_n; Z\right)\right]\right).$$

See written note for details. Then, by uniform stability and telescope sum, we have

$$\left[\mathbb{E}_{\mathcal{S}_{m'}} g\left(\mathcal{S}_n; Z\right) - \mathbb{E}_{\mathcal{S}_m} g\left(\mathcal{S}_n; Z\right)\right] \leq \left(m - m'\right) \epsilon.$$

We further assume that $m \leq 2m'$ so that $(m-m')/m' \leq 1$. It follows that

$$\begin{aligned} g'\left(\mathcal{S}_{m'+1,m}\right) =& m' \ln \mathbb{E}_Z \exp\left(\frac{\lambda}{m'}\left[\mathbb{E}_{\mathcal{S}_{m'}} g\left(\mathcal{S}_n; Z\right) - \mathbb{E}_{\mathcal{S}_m} g\left(\mathcal{S}_n; Z\right)\right]\right) \\ \leq& m'\left[\mathbb{E}_Z \exp\left(\frac{\lambda}{m'}\left[\mathbb{E}_{\mathcal{S}_{m'}} g\left(\mathcal{S}_n; Z\right) - \mathbb{E}_{\mathcal{S}_m} g\left(\mathcal{S}_n; Z\right)\right]\right) - \right. \\ & \left. - \frac{\lambda}{m'} \mathbb{E}_Z\left[\mathbb{E}_{\mathcal{S}_{m'}} g\left(\mathcal{S}_n; Z\right) - \mathbb{E}_{\mathcal{S}_m} g\left(\mathcal{S}_n; Z\right)\right]\right] \\ \leq& \frac{(\lambda)^2}{m'} \psi(0.4) \underbrace{\mathbb{E}_Z\left(\mathbb{E}_{\mathcal{S}_{m'}} g\left(\mathcal{S}_n; Z\right) - \mathbb{E}_{\mathcal{S}_m} g\left(\mathcal{S}_n; Z\right)\right)^2}_{g''\left(\mathcal{S}_{m'+1,m}\right)}, \end{aligned} \tag{B.3}$$

where we use $\log z \leq z-1$ and $\mathbb{E}_Z g(\mathcal{S}_n; Z) = 0$ for all $\mathcal{S}_n$ in the first inequality and use $\psi(z) = (e^z - z - 1)/z^2$ is increasing in z and

$$z = \frac{\lambda}{m'}\left[\mathbb{E}_{\mathcal{S}_{m'}} g\left(\mathcal{S}_n; Z\right) - \mathbb{E}_{\mathcal{S}_m} g\left(\mathcal{S}_n; Z\right)\right] \leq \frac{\lambda(m-m')}{m'}\epsilon \leq 0.4,$$

if $\lambda\epsilon \leq 0.4$. By Lemma 6, we have

$$\mathbb{E}_{\mathcal{S}_{m'+1,m}} g''\left(\mathcal{S}_{m'+1,m}\right) \leq 0.5\left(m-m'\right)\epsilon^2,$$

where the left-hand side is the variance of $f(S_{m'+1,m})$ and within each summand of right-hand side, two terms only differ by one elements. Now we consider two sets $\mathcal{S}_{m'+1,m}$ and $\mathcal{S}'_{m'+1,m}$ that differ by only one element, we have

$$\begin{aligned} & \left(g''\left(\mathcal{S}_{m'+1,m}\right) - g''\left(\mathcal{S}'_{m'+1,m}\right)\right)^2 \\ \leq& \left(2\epsilon \mathbb{E}_Z\left|\left[\mathbb{E}_{\mathcal{S}_{m'}} g\left(\mathcal{S}_n; Z\right) - \mathbb{E}_{\mathcal{S}_m} g\left(\mathcal{S}_n; Z\right)\right]\right| + \epsilon^2\right)^2 \\ \leq& 5\epsilon^2 g''\left(\mathcal{S}_{m'+1,m}\right) + 5\epsilon^4. \end{aligned}$$

This corresponds to Lemma (5) with $a = 5(m-m')\epsilon^2$ and $b = (m-m')\epsilon^4$ (note we need to sum from $m'+1$ to m, leading to $(m-m')$). By Eqn. B.3, we have

$$\begin{aligned} & \ln \mathbb{E}_{\mathcal{S}_{m'+1,m}} \exp\left(g'\left(\mathcal{S}_{m'+1,m}\right)\right) \\ \leq& \ln \mathbb{E}_{\mathcal{S}_{m'+1,m}} \exp\left(\frac{(\lambda)^2}{m'}\psi(0.4) g''\left(\mathcal{S}_{m'+1,m}\right)\right) \\ \leq& \frac{(\lambda)^2}{m'}\psi(0.4)\mathbb{E}_{\mathcal{S}_{m'+1,m}} g''\left(\mathcal{S}_{m'+1,m}\right) \\ & + 2\frac{(\lambda)^4}{(m')^2}\psi(0.4)^2\left(5\left(m-m'\right)\epsilon^2 \mathbb{E}_{\mathcal{S}_{m'+1,m}} g''\left(\mathcal{S}_{m'+1,m}\right) + 5\left(m-m'\right)\epsilon^4\right) \end{aligned}$$

with $\lambda_{lemma} = \frac{\lambda^2}{m'}\psi(0.4)$ where the second inequality uses Lemma 5 and

$$1 - a\frac{\lambda^2}{m'}\psi(0.4) \geq 1 - 5(0.4)^2\psi(0.4) \geq 0.5$$

since $\lambda\epsilon \leq 0.5$. Combining this with $\mathbb{E}_{\mathcal{S}_{m'+1,m}} g''\left(\mathcal{S}_{m'+1,m}\right) \leq 0.5\left(m - m'\right)\epsilon^2$, we have

$$\begin{aligned}
&\ln \mathbb{E}_{\mathcal{S}_{m'+1,m}} \exp\left(g'\left(\mathcal{S}_{m'+1,m}\right)\right) \\
=&\ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\frac{\lambda}{m'}\sum_{i=1}^{m'}\left[\mathbb{E}_{\mathcal{S}_{m'}^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_{m}^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right]\right) \\
\leq& 0.3\lambda^2\epsilon^2 + 6\lambda^4\epsilon^4
\end{aligned}$$

where we use $\psi(0.4) \leq 0.58, (m - m')/m' \leq 1$ so

$$\psi(0.4)0.5 \leq 3 \text{ and } 2\psi(0.4)^2 \times [5 \times 0.5 + \frac{1}{m'}5] \leq 6.$$

Now we consider a sequence $1 = m_0 < m_1 < \cdots < m_L = n$ where $m_\ell = \min(2^\ell, n)$. Let $\lambda_\ell = \lambda/\ell$ for $\ell > 0$ and $\lambda_0 = \lambda$ (so $(m - m')/m' \leq 1$). Then, we have

$$\begin{aligned}
&\ell \ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\lambda_\ell \bar{g}_{m_\ell, m_\ell}\left(\mathcal{S}_{n+1}\right)\right) \\
&\leq \ell \ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\lambda_\ell \bar{g}_{m_{\ell-1}, m_\ell}\left(\mathcal{S}_{n+1}\right)\right) \\
&\leq (\ell - 1) \ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\lambda_{\ell-1} \bar{g}_{m_{\ell-1}, m_{\ell-1}}\left(\mathcal{S}_{n+1}\right)\right) \\
&\quad + \ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\frac{\lambda}{m_{\ell-1}}\sum_{i=1}^{m_{\ell-1}}\left[\mathbb{E}_{\mathcal{S}_{m_{\ell-1}}^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right) - \mathbb{E}_{\mathcal{S}_{m_\ell}^{(i)}} g\left(\mathcal{S}_{n+1}^{(i)}; Z_i\right)\right]\right) \\
&\leq (\ell - 1) \ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\lambda_{\ell-1} \bar{g}_{m_{\ell-1}, m_{\ell-1}}\left(\mathcal{S}_{n+1}\right)\right) + 0.3\lambda^2\epsilon^2 + 6\lambda^4\epsilon^4
\end{aligned}$$

where the first inequality is due to Eqn. (B.1); the second inequality comes from Eqn. (B.2) and the last inequality uses the above result. Summing over $\ell = 1, \cdots, L$, we obtain

$$L \ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\lambda_L \bar{g}\left(\mathcal{S}_{n+1}\right)\right) = \ln \mathbb{E}_{\mathcal{S}_{n+1}} \exp\left(\lambda_L \bar{g}_{m_L, m_L}\left(\mathcal{S}_{n+1}\right)\right) \leq L0.3\lambda^2\epsilon^2 + L6\lambda^4\epsilon^4.$$

This implies the first desired bound. The second inequality follows from the Markov's inequality as follows. Considering $\epsilon' > 0$ and taking $\lambda = 0.4/\epsilon$, we have

$$\ln \Pr\left[\bar{g}\left(\mathcal{S}_{n+1}\right) \geq L\left(1 + \epsilon'\right)\epsilon\right] \leq \left[0.3\lambda^2\epsilon^2 + 6\lambda^4\epsilon^4 - (\lambda/L)L\left(1 + \epsilon'\right)\epsilon\right] \leq -0.4\epsilon'.$$

Setting $\exp(-0.4\epsilon') = \delta$ implies that w.p. at least $1 - \delta$, we have

$$\bar{g}(\mathcal{S}_{n+1}) \leq L\epsilon + 2.5L\epsilon \ln(1/\delta).$$

146 □