High probability generalization bounds for uniformly stable algorithms with nearly optimal rate

Vitaly Feldman[*]
Google Brain

Jan Vondrák
Stanford University

Abstract

Algorithmic stability is a classical approach to understanding and analysis of the generalization error of learning algorithms. A notable weakness of most stability-based generalization bounds is that they hold only in expectation. Generalization with high probability has been established in a landmark paper of Bousquet and Elisseeff (2002) albeit at the expense of an additional $\sqrt{n}$ factor in the bound. Specifically, their bound on the estimation error of any γ-uniformly stable learning algorithm on n samples and range in $[0,1]$ is $O(\gamma\sqrt{n\log(1/\delta)} + \sqrt{\log(1/\delta)/n})$ with probability $\geq 1-\delta$. The $\sqrt{n}$ overhead makes the bound vacuous in the common settings where $\gamma \geq 1/\sqrt{n}$. A stronger bound was recently proved by the authors (Feldman and Vondrak, 2018) that reduces the overhead to at most $O(n^{1/4})$. Still, both of these results give optimal generalization bounds only when $\gamma = O(1/n)$.

We prove a nearly tight bound of $O(\gamma\log(n)\log(n/\delta) + \sqrt{\log(1/\delta)/n})$ on the estimation error of any γ-uniformly stable algorithm. It implies that algorithms that are uniformly stable with $\gamma = O(1/\sqrt{n})$ have essentially the same estimation error as algorithms that output a fixed function. Our result leads to the first high-probability generalization bounds for multi-pass stochastic gradient descent and regularized ERM for stochastic convex problems with nearly optimal rate — resolving open problems in prior work. Our proof technique is new and we introduce several analysis tools that might find additional applications.

1 Introduction

We consider the following problem. Let $\bar{s} = (s_1,\ldots,s_n) \in Z^n$ be a dataset over an arbitrary domain and $M\colon Z^n \to [0,1]^Z$ be an arbitrary algorithm (or mapping) from datasets to functions over Z with range in $[0,1]$. M is said to be γ-uniformly stable if for all datasets $\bar{s}$ and $\bar{s}'$ that differ in a single element $\|M(\bar{s}) - M(\bar{s}')\|_\infty \leq \gamma$. Equivalently, for every $z \in Z$, $|M(\bar{s},z) - M(\bar{s}',z)| \leq \gamma$ (where $M(\bar{s},z)$ refers to the value of the function $M(\bar{s})$ on z). Assume that $\bar{s}$ consists of samples drawn i.i.d. from some distribution $\mathcal{P}$ over Z. We address the question of how well the true expectation of $M(\bar{s})$ on $\mathcal{P}$, that is $\mathbf{E}_{\mathcal{P}}[M(\bar{s})] = \mathbf{E}_{z\sim\mathcal{P}}[M(\bar{s},z)]$ is approximated by the empirical mean of $M(\bar{s})$ on $\bar{s}$, that is $\mathcal{E}_{\bar{s}}[M(\bar{s})] = \frac{1}{n}\sum_{i\in[n]} M(\bar{s},s_i)$. The value

$$\Delta_{\bar{s}}(M) \doteq |\mathop{\mathbf{E}}_{\mathcal{P}}[M(\bar{s})] - \mathcal{E}_{\bar{s}}[M(\bar{s})]|$$

is referred to as the *estimation error* of M at $\bar{s}$.

The primary motivation and the origin of this question is understanding of the generalization error of learning algorithms that are uniformly stable. In this context, $Z = X \times Y$ is labeled points and the goal

[*] Part of this work was done while the author was visiting the Simons Institute for the Theory of Computing.

is to analyze a learning algorithm A that given $\bar{s}$ outputs a model $f_{\bar{s}}\colon X \to Y$. The output of the learning algorithm is evaluated via some loss function $\ell_Y\colon Y \times Y \to \mathbb{R}_+$, with true loss being defined as $\mathbf{E}_{(x,y)\sim\mathcal{P}}[\ell_Y(f_{\bar{s}}(x), y)]$. By defining $M(\bar{s}, (x,y)) = \ell_Y(f_{\bar{s}}(x), y)$ we get that the estimation error of M is exactly the difference between the true loss of $f_{\bar{s}}$ and the empirical loss of $f_{\bar{s}}$ on $\bar{s}$ (sometimes referred to as the *generalization gap*).

Stability is a classical approach to proving generalization bounds pioneered by Rogers and Wagner [RW78] and Devroye and Wagner [DW79a; DW79b]. It is based on analysis of the sensitivity of the learning algorithm to changes in the dataset such as leaving one of the data points out or replacing it with a different one. The choice of how to measure the effect of the change and various ways to average over multiple changes give rise to a variety of stability notions that have been examined in the literature (e.g. [BE02; Muk+06; SS+10]). Unfortunately, most stability notions only lead to bounds on the expectation or the second moment of the estimation error over the random choice of the dataset. In contrast, generalization bounds based on uniform convergence show that the estimation error is small with high probability (more formally, the distribution of the error has exponentially decaying tails). Beyond theoretical interest, high-probability generalization bounds are necessary for inferring generalization when the algorithm is used many times (as is common in practice).

High probability generalization bounds based on stability were first obtained by Lugosi and Pawlak [LP94] for several specific learning algorithms. In a seminal work Bousquet and Elisseeff [BE02] developed a general approach based on the notion of *uniform stability* (defined above). While uniform stability is a relatively strong condition, it is satisfied by several well-studied algorithms. For example, for strongly convex Lipschitz losses the ERM is uniformly stable [BE02; SS+10] (we describe the bounds quantitatively in Sec. 4). More recently, Hardt et al. [HRS16] showed that for convex smooth losses the solution obtained via gradient descent is uniformly stable allowing them to give the first generalization guarantees for many variants of (stochastic) gradient descent. Importantly, no other known approaches give comparable generalization bounds for these fundamental algorithms. Moreover, there exist empirical risk minimizing algorithms for convex problems whose generalization error is $\sqrt{d}$ times larger than the generalization bounds obtained via stability, where d is the dimension of the problem [SS+10; Fel16]. This implies that approaches requiring uniform convergence over the set of all models that minimize the empirical loss (such as most model-complexity-based bounds) will not lead to useful generalization guarantees in this case. We remark that continuous optimization methods play a central role in modern machine learning and hence their generalization properties is a topic of intense theoretical and practical interest in recent years.

1.1 Prior work

The main generalization bound for γ-uniformly stable algorithms given in [BE02] states that for some constant c_0,

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}\left[\Delta_{\bar{s}}(M) \geq c_0\left(\gamma\sqrt{n} + \frac{1}{\sqrt{n}}\right)\sqrt{\log(1/\delta)}\right] \leq \delta. \tag{1}$$

This is in contrast to an easy observation that the expectations of $\mathbf{E}_{\mathcal{P}}[M(\bar{s})]$ and $\mathcal{E}_{\bar{s}}[M(\bar{s})]$ are within γ. Namely,

$$\left|\mathop{\mathbf{E}}_{\bar{s}\sim\mathcal{P}^n}\left[\mathop{\mathbf{E}}_{\mathcal{P}}[M(\bar{s})] - \mathcal{E}_{\bar{s}}[M(\bar{s})]\right]\right| \leq \gamma. \tag{2}$$

Thus the bound on estimation error is worse by at least a factor of $\sqrt{n}$ than the expected difference. In terms of lower bounds, note that the term $\frac{\sqrt{\log(1/\delta)}}{\sqrt{n}}$ is necessary since even for an algorithm that outputs a fixed function (or $\gamma = 0$) this is the optimal bound on the sampling error. In addition, estimation error is at least γ since the function can change arbitrarily in this range.

Naturally, for most algorithms the stability parameter needs to be balanced against the guarantees on the empirical loss. For example, ERM solution to convex learning problems can be made uniformly stable by adding a strongly convex term to the objective [SS+10]. This change in the objective introduces an error that may increase the original empirical loss. In the other example, the stability parameter of gradient descent on smooth objectives is determined by the sum of the rates used for all the gradient steps [HRS16]. Limiting the sum limits the empirical loss that can be achieved. In both of those examples the optimal expected loss can is achieved when $\gamma = \Theta(1/\sqrt{n})$. Unfortunately, in this setting, eq. (1) gives a vacuous bound. As a result, in these applications only bounds on the expectation of the true loss are stated. For both of these applications, deriving a high-probability generalization bound is stated as an open problem [SS+10; HRS16].

Note that eq. (2) does not imply that $\mathbf{E}_{\mathcal{P}}[M(\bar{s})] \leq \mathcal{E}_{\bar{s}}[M(\bar{s})] + O(\gamma/\delta)$ with probability at least $1-\delta$ since $\mathbf{E}_{\mathcal{P}}[M(\bar{s})] - \mathcal{E}_{\bar{s}}[M(\bar{s})]$ can be negative and Markov's inequality cannot be used. Such "low-probability" generalization was first derived by Shalev-Shwartz et al. [SS+10] for learning algorithms that minimize the empirical risk. For such algorithms they showed that

$$\mathop{\mathbf{E}}_{\bar{s}\sim\mathcal{P}^n}\left[\Delta_{\bar{s}}(M)\right] \leq O\left(\gamma + \frac{1}{\sqrt{n}}\right), \tag{3}$$

allowing them to apply Markov's inequality.

Generalization properties of uniform stability were addressed in a recent work by the authors [FV18]. There we demonstrated that there exists a constant c_1 such that

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}\left[\Delta_{\bar{s}}(M) \geq c_1\left(\sqrt{\gamma} + \frac{1}{\sqrt{n}}\right)\sqrt{\log(1/\delta)}\right] \leq \delta \tag{4}$$

improving on eq. (1) for $\gamma = \omega(1/n)$. This result reduces the overhead of high-probability generalization from $\sqrt{n}$ to at most $n^{1/4}$ (achieved for $\gamma = 1/\sqrt{n}$). This bound was used to strengthen the generalization guarantees that are known for the convex optimization algorithms described above but only implies that suboptimality of the solution is $O(1/n^{1/3})$ with high-probability (whereas the optimal rate is $O(1/\sqrt{n})$).

Further, we gave an optimal (up to constant factors) bound on the second moment of the estimation error:

$$\mathop{\mathbf{E}}_{\bar{s}\sim\mathcal{P}^n}\left[\Delta_{\bar{s}}(M)^2\right] \leq O\left(\gamma^2 + \frac{1}{n}\right),$$

improving on the $O(\gamma + \frac{1}{n})$ bound in [BE02].

A natural question of whether the high-probability bounds can be strengthened (or a matching lower bound can be proved) still remained open.

1.2 Our contribution

Our main result is a high-probability generalization bound for any γ-uniformly stable algorithm that has only a logarithmic overhead. In particular, it gives an exponential improvement (in terms of the tail bound δ) over prior work.

Theorem 1.1. *Let $M : Z^n \times Z \to [0,1]$ be an algorithm (or a data-dependent function) with uniform stability γ. Then there exists a constant c such that for any probability distribution $\mathcal{P}$ over Z and any $\delta \in (0,1)$:*

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}\left[\Delta_{\bar{s}}(M) \geq c\left(\gamma\log(n)\log(n/\delta) + \frac{\sqrt{\log(1/\delta)}}{\sqrt{n}}\right)\right] \leq \delta.$$

A somewhat surprising implication of this result is that algorithms that are uniformly stable with $\gamma = O(1/\sqrt{n})$ enjoy essentially the same estimation error guarantees as algorithms that do not look at the data and output a fixed function. For $\gamma \leq \sqrt{\log(1/\delta)}/(\sqrt{n}\log(n/\delta)\log(n))$, there is no significant contribution depending on γ and our bound is optimal up to constant factors. In contrast, both previous works [BE02; FV18] give similar generalization guarantees only when $\gamma = O(1/n)$.

Proof approach: The high-probability generalization result in [BE02] (eq. (1)) is based on a simple observation that as a function of $\bar{s}$, the estimation error has sensitivity of at most $2\gamma + 1/n$. Applying McDiarmid's concentration inequality immediately implies concentration with standard deviation of $\sqrt{n}(\gamma + 1/n)$ around the expectation. The expectation, in turn, is at most γ by eq. (2).

The approach in our prior work [FV18] is based on a technique developed in [Bas+16] to prove generalization bounds for differentially private algorithms. It bounds the tail by proving a bound on the expectation of the maximum of many independent copies of the estimation error. The latter is bounded by using a soft-argmax operation. Soft-argmax is itself stable and hence the expectation of the estimation error of the copy it outputs is small. While the bound of $\sqrt{\gamma}$ derived using this approach may appear to be arbitrary, it has been re-derived using other approaches by the authors and also by Weinberger and Rakhlin [WR18] who used a bound on the second moment from [FV18] to bound the moment generating function of the estimation error.

Our approach is based on two new ideas that both rely strongly on the structure of the estimation error. The first idea is to upper bound the estimation error by using the bound on the estimation error over a smaller dataset. This step is very simple technically and can already be used to re-derive the $\sqrt{\gamma}$ bound from our earlier work [FV18] (optimizing the simple bound $\gamma\sqrt{n'} + 1/\sqrt{n'}$ over $n' \leq n$ gives exactly $2\sqrt{\gamma}$).

The second idea is to reduce the range or the output function by subtracting the mean and "clamping" the values outside the range. Uniform stability can be used to ensure that for an appropriately chosen range this procedure will introduce only a small error. The main technical issue is that we need to ensure that the clamping procedure both preserves the stability parameter and does not shift the mean of the estimation error (as the first step requires a zero-mean random variable). Achieving both of these goals requires a more involved "clamping" procedure and delicate analysis.

Combining these procedures decomposes the estimation error into a sum of mixtures of "local" approximations (that is, accurate for specific setting of some of the samples in the dataset). Repeated application of this combination in a recursive way gives the proof of our main result. The $\log n$ levels of recursion are the reason for the $\log n$ overhead of our bound. In Sec. 3.1 we give a more technical overview of the proof.

1.3 Applications

We now apply our bounds on the estimation error to several known uniformly stable algorithms. Our main focus are learning problems that can be formulated as stochastic convex optimization. Specifically, these are problems in which the goal is to minimize the expected loss: $F_{\mathcal{P}}(w) \doteq \mathbf{E}_{z\sim\mathcal{P}}[\ell(w,z)]$ over $w \in \mathcal{K}$ for some convex body $\mathcal{K} \subset \mathbb{R}^d$ and a family of convex losses $\mathcal{F} = \{\ell(\cdot,z)\}_{z\in Z}$. The stochastic convex optimization problem for a family of losses $\mathcal{F}$ over $\mathcal{K}$ is the problem of minimizing $F_{\mathcal{P}}(w)$ for an arbitrary distribution $\mathcal{P}$ over Z. For concreteness, we consider the well-studied setting in which $\mathcal{F}$ contains 1-Lipschitz convex functions with range in $[0,1]$ and $\mathcal{K}$ is included in the unit ball (settings with an arbitrary Lipschitz constant and domain radius can be reduced to this case via scaling).

Strongly convex ERM: In this setting with an additional assumption that loss functions in $\mathcal{F}$ are λ-strongly convex, ERM has uniform stability of $4/(\lambda n)$ [BE02]. We therefore obtain high-probability generalization bounds on ERM in this case that improve on the known results for any $\lambda = o(1)$ (see Corollary 4.2 for details).

Using stability of ERM for strongly convex functions, Shalev-Shwartz et al. [SS+10] showed that even without strong convexity, the stochastic convex optimization problem can be solved by adding a strongly convex regularizer $\frac{\lambda}{2}\|w\|^2$ to the empirical loss with $\lambda = 1/\sqrt{n}$. They demonstrate that the expected loss of this algorithm is optimal and conjecture that high-probability generalization bounds hold as well. Using Thm. 1.1, we show that excess loss (or sub-optimality) of the solution is at most $O(\log(n/\delta)/\sqrt{n})$ with probability at least $1-\delta$, thereby proving the conjecture. (The optimal choice of λ is determined by balancing the estimation error and the error introduced by adding the regularizer and in our result $\lambda = \log(n)/\sqrt{n}$.).

Corollary 1.2. *Let $\mathcal{K}$ be a convex body of radius 1, $\mathcal{F} = \{\ell(\cdot, z) \mid z \in Z\}$ be a family of convex 1-Lipschitz loss functions over $\mathcal{K}$ with range in $[0,1]$. For a dataset $\bar{s} \in Z^n$ let $w_{\bar{s}}$ denote the empirical minimizer of regularized loss on $\bar{s}$: $w_{\bar{s},\lambda} = \operatorname{argmin}_{w\in\mathcal{K}} \sum_{i\in[n]} \ell(w, s_i) + \frac{\lambda n}{2}\|w\|_2^2$. There exist a constant c such that for every distribution $\mathcal{P}$ over Z, $\delta > 0$ and $\lambda = \log(n)/\sqrt{n}$:*

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}\left[F_{\mathcal{P}}(w_{\bar{s},\lambda}) \geq \min_{w\in\mathcal{K}} F_{\mathcal{P}}(w) + \frac{c\log(n/\delta)}{\sqrt{n}}\right] \leq \delta.$$

(Stochastic) gradient descent: Another fundamental application of uniform stability is proving generalization bounds for (stochastic) gradient descent on sufficiently smooth convex loss functions [HRS16]. Importantly, in this case the estimation error can be bounded without any assumptions on how close the output of the algorithm is to the empirical minimum. Therefore this approach can be used to give generalization bounds for variants of SGD used in practice (as opposed to those prescribed by theoretical analysis). For most versions of SGD no alternative analyses of the estimation error are known. The analysis in [HRS16] focuses on the stochastic gradient descent and derives uniform stability for the expectation of the loss (over the randomness of the algorithm). From this result they obtained generalization in expectation over both randomness of the algorithm and the choice of the dataset. Obtaining bounds that hold with high-probability was left as an open problem.

Theorem 1.1 ensures that the bounds on estimation error hold with high probability over the choice of the dataset. This suffices to get generalization with high probability for deterministic variants of gradient descent. As an example application we derive nearly optimal generalization bounds for full gradient descent (see Corollary 4.4). To obtain generalization bounds for SGD we additionally observe that for most standard choices of picking batches randomly, the uniform stability of the gradient descent as a function of the randomness of SGD is highly concentrated around its mean. As a result we can obtain a bound on the estimation error that holds with high probability over the randomness of SGD and is worse than the bound that holds in expectation by at most a logarithmic factor. As an example application of this technique we derive nearly optimal generalization bounds for stochastic gradient descent that uses sampling with replacement for each gradient and batch size of 1 (see Corollary 4.7).

For comparison, a recent work of London [Lon17] considers extension of the generalization guarantees in [HRS16] to high-probability over the randomness in the choice of samples. The approach there relies on sensitivity of the estimation error to the choices of random samples. It requires independent sampling at each step and the resulting bound on the estimation error has an overhead of $\sqrt{T}$, where T is the number of iterations. As a result it gives much weaker bounds in the setting we consider ([Lon17] focuses on the smooth and strongly convex case).

Prediction privacy: Finally, we show that our results can be used to improve the recent bounds on estimation error of learning algorithms with differentially private prediction. These are algorithms introduced to model privacy-preserving learning in the settings where users only have black-box access to the learned model via a prediction interface [DF18] (see Def. 4.9). The properties of differential privacy imply that the expectation over the randomness of a predictor $K\colon (X \times Y)^n \times X$ of the loss of K at any point $x \in X$ is uniformly stable. Specifically, for an ϵ-differentially private prediction algorithm, every loss function $\ell_Y\colon Y \times Y \to [0,1]$, two datasets $\bar{s}, \bar{s}' \in (X \times Y)^n$ that differ in a single element and $(x,y) \in X \times Y$:

$$\left| \mathop{\mathbf{E}}_{K}[\ell_Y(K(\bar{s},x),y)] - \mathop{\mathbf{E}}_{M}[\ell_Y(K(\bar{s}',x),y)] \right| \leq e^{\epsilon} - 1.$$

Therefore, our generalization bounds can be directly applied to the data-dependent function $M(\bar{s},(x,y)) \doteq \mathbf{E}_K[\ell_Y(K(\bar{s},x),y)]$. These bounds can, in turn, be used to get nearly optimal generalization bounds for an algorithm for learning linear thresholds given in [DF18] (that relies on models of unbounded complexity). The details of these applications appear in Section 4.

1.4 Other related work

Early work on stability focused on obtaining generalization guarantees for "local" algorithms such as k-nearest neighbor. The bounds were also primarily on variance of the estimation error (a notable exception is [DW79a] where high probability bounds on the generalization error of k-NN are proved). See [DGL96] for an overview. Stability is also used in a similar spirit for bounding the estimation error of other estimators of true loss such as leave-one-out and k-fold cross-validation estimators (for example [BKL99; KKV11; Kum+13]).

A long line of work focuses on the relationship between various notions of stability and learnability in supervised setting (see [KR99; Pog+04; SS+10] for an overview). This work employs relatively weak notions of average stability and derives a variety of asymptotic equivalence results. The results in [BE02] on uniform stability and their applications to generalization properties of strongly convex ERM algorithms have been extended and generalized in several directions (e.g. [Zha03; WP09]). Maurer [Mau17] considers generalization bounds for a special case of linear regression with a strongly convex regularizer and a sufficiently smooth loss function. Their bounds are data-dependent and are potentially stronger for large values of the regularization parameter (and hence stability). However the bound is vacuous when the stability parameter is larger than $n^{-1/4}$ and hence is not directly comparable to ours. Kuzborskij and Lampert [KL18] give data-dependent generalization bounds for SGD on smooth convex and non-convex losses based on stability. They use on-average stability that does not imply generalization bounds with high probability. Recent work of Abou-Moustafa and Szepesvári [AS18] and Celisse and Guedj [CG16] gives high probability generalization bounds similar to those in [BE02] but using a bound on a high-order moment of stability instead of the uniform stability. Recent applications of stability to generalization can be found for example in [Liu+17; Riv+18; CP18; CJY18]. We also remark that all these works are based on techniques different from ours.

Uniform stability has several additional important connections to differential privacy [Dwo+06]. First, differential privacy is itself a type of worst-case stability guarantee that bounds the effect of every data point on the output distribution of the algorithm. Our work is in part inspired by the recent progress showing that differential privacy implies generalization with high probability [Dwo+14; Bas+16]. Both the assumptions and guarantees given in this line of work are different from ours and we do not know a way to relate between those. For example, the generalization guarantees obtained in work on differential privacy hold with high probability over the randomness of the algorithm, whereas our results when applied to a differentially private algorithm would only give generalization of the expectation over the algorithm's randomness. We remark that the techniques developed in this line of work were used to re-derive and extend several standard concentration inequalities [SU17; NS17] and also in [FV18] to give an improved generalization bound for uniform stability.

Uniformly stable algorithms also play an important role in privacy-preserving learning since a differentially private learning algorithm can usually be obtained by adding noise to the output of a uniformly stable one (e.g. [CMS11; Wu+17; DF18]). Hence understanding the generalization properties of uniformly stable algorithms is likely to play an important role in this line of research.

2 Preliminaries

For a domain Z, a dataset $\bar{s} \in Z^n$ in an n-tuple of elements in Z. We refer to element with index i by s_i and by $\bar{s}^{i\leftarrow z}$ to the dataset obtained from $\bar{s}$ by setting the element with index i to z. We refer to a function that takes as an input a dataset $\bar{s} \in Z^n$ and a point $z \in Z$ as a *data-dependent function* over Z.

We think of data-dependent functions as outputs of an algorithm that takes $\bar{s}$ as an input. For example in supervised learning Z is the set of all possible labeled examples $Z = X \times Y$ and the algorithm M is defined as estimating some loss function $\ell_Y : Y \times Y \to \mathbb{R}_+$ of the model $f_{\bar{s}}$ output by a learning algorithm $A(\bar{s})$ on example $z = (x, y)$. That is $M(\bar{s}, z) = \ell_Y(f_{\bar{s}}(x), y)$. Note that in this setting $\mathcal{E}_{\mathcal{P}}[M(\bar{s})]$ is exactly the true loss of $f_{\bar{s}}$ on data distribution $\mathcal{P}$, whereas $\mathcal{E}_{\bar{s}}[M(\bar{s})]$ is the empirical loss of $f_{\bar{s}}$.

Definition 2.1. *A data-dependent function $M\colon Z^n \times Z \to \mathbb{R}$ has uniform stability γ if for all $\bar{s} \in Z^n$, $i \in [n]$, $s, z \in Z$, $|M(\bar{s}, z) - M(\bar{s}^{i\leftarrow s}, z)| \leq \gamma$.*

Two natural ways to use this definition for randomized algorithms are (1) consider stability of the expectation over the algorithm's randomness $\mathbf{E}_M[M(S, z)]$; (2) consider stability of $M(S, z)$ for a fixed setting of M's random bits. The first approach is simpler and usually results in a better stability parameter but only leads to generalization guarantees for $\mathbf{E}_M[M(S)]$ (for examples see [EEP05; HRS16] and Sec. 4.2). This approach is necessary for obtaining a non-trivial bound on uniform stability in classification problems [LP94; BE02]. In contrast, the second approach may lead to generalization with high-probability over M's randomness if the stability parameter can be upper-bounded with high probability (over M's randomness).

Uniform stability is equivalent to $M(\bar{s}, z)$ as a function of $\bar{s}$ having *sensitivity* γ or γ-bounded differences for all $z \in Z$.

Definition 2.2. *A real-valued function $f : Z^n \to \mathbb{R}$ has sensitivity at most γ if for all $\bar{s} \in Z^n$, $i \in [n]$, $s \in Z$, $|f(\bar{s}) - f(\bar{s}^{i\leftarrow s})| \leq \gamma$.*

We will use McDiarmid's inequality for functions of bounded sensitivity.

Lemma 2.3. *Let $f : Z^n \to \mathbb{R}$ be a function with sensitivity of at most γ. Then for any distribution $\mathcal{P}$ over Z, $\mu \doteq \mathbf{E}_{\bar{s}\sim\mathcal{P}^n}[f(\bar{s})]$ and any $t > 0$,*

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}[f(\bar{s}) \geq \mu + t] \leq e^{-2t^2/(n\gamma^2)}.$$

Estimation error: For convenience, as in [FV18], we reduce bounds on $\Delta_{\bar{s}}(M)$ to bounds on the leave-one-out estimation error for the unbiased version of M (we include the details here for completeness).

Specifically, we define $L(\bar{s}, z) \doteq M(\bar{s}, z) - \mathbf{E}_{z\sim\mathcal{P}}[M(\bar{s}, z)]$. Clearly, L is *unbiased* with respect to $\mathcal{P}$ in the sense that for every $\bar{s} \in Z^n$, $\mathcal{E}_{\mathcal{P}}[L(\bar{s})] = \mathbf{E}_{z\sim P}[L(\bar{s}, z)] = 0$. Note that if the range of M is $[0, 1]$ then the range of L is $[-1, 1]$. Further, L has uniform stability of at most 2γ since for two datasets $\bar{s}$ and $\bar{s}'$ that differ in a single element,

$$|L(\bar{s}, z) - L(\bar{s}', z)| \leq |M(\bar{s}, z) - M(\bar{s}', z)| + \mathop{\mathbf{E}}_{z\sim\mathcal{P}}[|M(\bar{s}, z) - M(\bar{s}', z)|] \leq 2\gamma.$$

Observe that

$$\Delta_{\bar{s}}(M) = \left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{E}_{\mathcal{P}}[M(\bar{s})] - M(\bar{s}, s_i)\right)\right| = \left|-\frac{1}{n}\sum_{i=1}^{n} L(\bar{s}, s_i)\right| = \left|\mathcal{E}_{\bar{s}}[L(\bar{s})]\right|.$$

The leave-one-out version of the estimation error is defined as follows. For any $z \in Z$ let

$$\mathcal{E}_{\bar{s}}^{\leftarrow z}[L] := \frac{1}{n}\sum_{i\in[n]} L(\bar{s}^{i\leftarrow z}, s_i).$$

Observe that the uniform stability of L implies that for every $\bar{s}$ and every z,

$$\begin{aligned}\left|\mathcal{E}_{\bar{s}}[L(\bar{s})] - \mathcal{E}_{\bar{s}}^{\leftarrow z}[L]\right| &= \left|\frac{1}{n}\sum_{i\in[n]} L(\bar{s}, s_i) - \frac{1}{n}\sum_{i\in[n]} L(\bar{s}^{i\leftarrow z}, s_i)\right| \\ &\leq \frac{1}{n}\sum_{i\in[n]} \left|L(\bar{s}, s_i) - L(\bar{s}^{i\leftarrow z}, s_i)\right| \leq \gamma. \end{aligned} \tag{5}$$

Tail bounding function: The goal of our analysis is to bound the following function.

Definition 2.4. *For an integer n, real $R, \gamma > 0$ and $\delta > 0$, let $D_\delta(n, R, \gamma)$ be the maximum value D such that for every domain Z, probability distribution $\mathcal{P}$ over Z, and a data-dependent γ-uniformly stable function $L : Z^n \times Z \to [-R, R]$ such that $\mathbf{E}_{z\sim\mathcal{P}}[L(\bar{s}, z)] = 0$,*

$$\Pr[|\mathcal{E}_{\bar{s}}^{\leftarrow z}[L]| \geq D] \leq \delta.$$

Observe that by simple scaling, the range and stability in the definition of D_δ can be adjusted by an arbitrary factor. That is, for an arbitrary factor $\alpha > 0$,

$$D_\delta(n, R, \gamma) = \frac{1}{\alpha} D_\delta(n, \alpha R, \alpha\gamma). \tag{6}$$

3 Proof of the Main Result

In this section, we prove our main concentration bound with exponential tails.

3.1 Overview of our approach

Let us recall the main parameters of our problem: the dataset size n, the range $[-R, R]$ of the function L, and the uniform stability parameter γ. Our approach is based on two operations, which reduce the bound on $D(n, R, \gamma)$ (ignoring for the moment the dependence on the tail probability δ as defined at the end of Section 2) to a bound on $D(n', R', \gamma)$ for some n' and R'. For simplicity, we ignore some details and logarithmic factors in the following.

Range reduction. If $\gamma < R/\sqrt{n}$, then by McDiarmid's inequality, $L(\bar{s}, z)$ for each z is concentrated in a window of size $R' = \gamma\sqrt{n} < R$. So we can "center" the function by subtracting the fixed function $\phi(z) = \mathbf{E}_{\bar{s}}[L(\bar{s}, z)]$ and then "clamp" the function $L(\bar{s}, z) - \phi(z)$ to range $R' = \gamma\sqrt{n}$. We will need to deal with two additional errors: the sampling error for ϕ which is on the order of $R/\sqrt{n}$, and the contribution of the values that we have "clamped off".

Dataset size reduction. For any setting of parameters, we can consider the dataset $[n]$ as partitioned into b blocks of size $n' = n/b$. (For the moment ignoring divisibility issues.) Since the estimation error, $\frac{1}{n}\sum_{i=1}^{n} L(\bar{s}, s_i)$, is an average over all coordinates, we can view it as an average of averages over each block. The expression for each block, conditioned on 'the variables outside the block, is just like the estimation error for a smaller problem with dataset size n'. Again we obtain some additional error terms, but roughly speaking we can reduce the dataset size from n to n' without significant change in the estimation error.

Let us assume for now we can do both of these operations in a way that preserves the stability parameter γ and keeps the function unbiased (by which we mean the condition $\mathbf{E}_z[L(\bar{s}, z)] = 0$). Interestingly, applying these two operations repeatedly essentially proves the concentration inequality that we want. We sketch the argument below focusing on $\gamma = R/\sqrt{n}$. This is effectively the hardest regime (the result for other values of γ is implied by applying one of the operations above once).

- Suppose that the estimation error as a function of n, R, γ is $D(n, R, \gamma)$. We want to prove that for $\gamma = R/\sqrt{n}$, $D(n, R, \gamma) = \tilde{O}(\gamma)$.
- Starting with parameters n, R, γ such that $\gamma = R/\sqrt{n}$, we can use the block partitioning argument to decrease the dataset size from n to $n' = n/b$ for some parameter $b > 1$. The stability parameter is unchanged, and equal to $\gamma = R/\sqrt{bn'}$.
- Since $\gamma = R/\sqrt{bn'} < R/\sqrt{n'}$, we can use the range reduction argument to clamp the function $L(\bar{s}, z)$ to a range of $R' = \gamma\sqrt{n'} = R/\sqrt{b}$. The stability parameter remains (hopefully) unchanged, $\gamma = R'/\sqrt{n'}$. Hence we are back in the situation similar to the one we started from, with the stability parameter being equal to the range divided by square root of the dataset size.
- Let's assume by induction that the estimation error for the function we obtained is $D(n', R', \gamma) = \tilde{O}(\gamma)$. By reversing the two operations that we performed, we obtain that, up to the additional error terms, the estimation error $D(n, R, \gamma)$ remains roughly the same, $D(n, R, \gamma) = \tilde{O}(\gamma))$. This leads to an inductive argument proving that the estimation error is indeed $\tilde{O}(\gamma)$.

Remaining issues. The main issue that need to be resolved in order to carry out this strategy is that simple "clamping" to a fixed range $[-R', R']$ might cause the function to be no longer unbiased. The bias can be eliminated by subtracting the mean $\mathbf{E}_z[L(\bar{s}, z)]$ again. Unfortunately, this operation may double the stability parameter. As a result, using this simple fix leads to a worse bound on the estimation error: $\gamma 2^{O(\sqrt{\log n})}$. To avoid this large overhead one needs to design a "clamping" operation that preserves both $\mathbf{E}_z[L(\bar{s}, z)] = 0$ and the uniform stability parameter. It turns out that both requirements can be satisfied simultaneously by a $\bar{s}$-dependent shift of the range, $[-R' + b_{\bar{s}}, R' + b_{\bar{s}}]$ that however requires a rather delicate analysis. We present this construction in Section 3.2.

The unbiased property needs to be also preserved in the block partitioning operation. Here, we obtain it essentially "for free", since the property $\forall \bar{s}; \mathbf{E}_z[L(\bar{s}, z)] = 0$ is preserved under conditioning on a subset of the coordinates in $\bar{s}$. Thus the block partitioning operation is technically rather simple.

3.2 Range reduction

We start by designing a procedure which reduces the range of a function to a desired width while preserving the expectation and stability at the same time.

Lemma 3.1. *Let $K : Z^n \times Z \to [-r, r]$ be a function, $\mathcal{P}$ a distribution on Z, and $w, \beta > 0$ such that*

- *for every $\bar{s} \in Z^n$, $\mathbf{E}_{z\sim\mathcal{P}}[K(\bar{s}, z)] = 0$,*
- *$\mathbf{E}_{(\bar{s},z)\sim\mathcal{P}^{n+1}}[(K(\bar{s}, z) - w)_+] \leq \beta$,*
- *$\mathbf{E}_{(\bar{s},z)\sim\mathcal{P}^{n+1}}[((-w - K(\bar{s}, z))_+] \leq \beta$,*
- *$K(\bar{s}, z)$ has uniform stability γ.*

Then for every $\bar{s} \in Z^n$, there exists $b_{\bar{s}} \in [-w, w]$ such that $\tilde{K}(\bar{s}, z) \doteq \mathtt{clamp}_{[b_{\bar{s}}-w, b_{\bar{s}}+w]}(K(\bar{s}, z))$ has the following properties:

- *for every $\bar{s} \in Z^n$, $\mathbf{E}_{z\sim\mathcal{P}}[\tilde{K}(\bar{s}, z)] = 0$,*
- *$\mathbf{E}_{(\bar{s},z)\sim\mathcal{P}^{n+1}}[|\tilde{K}(\bar{s}, z) - K(\bar{s}, z)|] \leq 4\beta$,*
- *$\tilde{K}(\bar{s}, z)$ has uniform stability γ.*

Note that in this lemma, the magnitude of the range r does not play any role; we just need the fact that $K(\bar{s}, z)$ is bounded.

Proof. Fix $\bar{s} \in Z^n$ and consider the function $\tilde{K}_x(\bar{s}, z) \doteq \mathtt{clamp}_{[x-w,x+w]}(K(\bar{s}, z))$. In other words,

$$\tilde{K}_x(\bar{s}, z) = K(\bar{s}, z) - (K(\bar{s}, z) - (x + w))_+ + ((x - w) - K(\bar{s}, z))_+.$$

Therefore,

$$\mathop{\mathbf{E}}_{z\sim\mathcal{P}}[\tilde{K}_x(\bar{s}, z)] = \mathop{\mathbf{E}}_{z\sim\mathcal{P}}[K(\bar{s}, z)] + \psi_{\bar{s}}(x),$$

where

$$\psi_{\bar{s}}(x) \doteq - \mathop{\mathbf{E}}_{z\sim\mathcal{P}}[(K(\bar{s}, z) - (x + w))_+] + \mathop{\mathbf{E}}_{z\sim\mathcal{P}}[((x - w) - K(\bar{s}, z))_+].$$

$\psi_{\bar{s}}(x)$ is continuous, since the expressions inside the expectations are uniformly continuous (in fact 1-Lipschitz) in x. Also, since $K(\bar{s}, z) \in [-r, r]$ and $w > 0$, we have $\psi_{\bar{s}}(-r) \leq 0$, $\psi_{\bar{s}}(r) \geq 0$, and $\psi_{\bar{s}}(x)$ is obviously non-decreasing. By the intermediate value theorem, there exists $b_{\bar{s}} \in [-r, r]$ such that $\psi_{\bar{s}}(b_{\bar{s}}) = 0$. This means that $\mathbf{E}_{z\sim\mathcal{P}}[(K(\bar{s}, z) - (b_{\bar{s}} + w))_+] = \mathbf{E}_{z\sim\mathcal{P}}[((b_{\bar{s}} - w) - K(\bar{s}, z))_+]$. We can define $\tilde{K}(\bar{s}, z) \doteq \mathtt{clamp}_{[b_{\bar{s}}-w, b_{\bar{s}}+w]}(K(\bar{s}, z))$ and we get $\mathbf{E}_{z\sim\mathcal{P}}[\tilde{K}(\bar{s}, z)] = \mathbf{E}_{z\sim\mathcal{P}}[K(\bar{s}, z)] = 0$. In addition, we observe that we must actually have $b_{\bar{s}} \in [-w, w]$, because otherwise it would not be possible that a function bounded by $[b_{\bar{s}} - w, b_{\bar{s}} + w]$ has expectation 0.

Let $\beta^+_{\bar{s}} \doteq \mathbf{E}_{z\sim\mathcal{P}}[(K(\bar{s}, z) - w)_+]$ and $\beta^-_{\bar{s}} \doteq \mathbf{E}_{z\sim\mathcal{P}}[(-w - K(\bar{s}, z))_+]$. If $b_{\bar{s}} \geq 0$, then

$$\mathop{\mathbf{E}}_{z\sim\mathcal{P}}[((b_{\bar{s}} - w) - K(\bar{s}, z))_+] = \mathop{\mathbf{E}}_{z\sim\mathcal{P}}[(K(\bar{s}, z) - (b_{\bar{s}} + w))_+] \leq \beta^+_{\bar{s}}.$$

Conversely, if $b_{\bar{s}} \leq 0$, then

$$\mathop{\mathbf{E}}_{z \sim \mathcal{P}}[(K(\bar{s}, z) - (b_{\bar{s}} + w))_+] = \mathop{\mathbf{E}}_{z \sim \mathcal{P}}[((b_{\bar{s}} - w) - K(\bar{s}, z))_+] \leq \beta_{\bar{s}}^- .$$

Either way,

$$\mathop{\mathbf{E}}_{z \sim \mathcal{P}}[(K(\bar{s}, z) - (b_{\bar{s}} + w))_+] + \mathop{\mathbf{E}}_{z \sim \mathcal{P}}[((b_{\bar{s}} - w) - K(\bar{s}, z))_+] \leq \max\{2\beta_{\bar{s}}^+, 2\beta_{\bar{s}}^-\}.$$

Note that

$$|\tilde{K}(\bar{s}, z) - K(\bar{s}, z)| = (K(\bar{s}, z) - (b_{\bar{s}} + w))_+ + ((b_s - w) - K(\bar{s}, z))_+.$$

Therefore, taking the expectation over both $\bar{s}$ and z, we obtain

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\bar{s}, z) \sim \mathcal{P}^{n+1}}[|\tilde{K}(\bar{s}, z) - K(\bar{s}, z)|] &= \mathop{\mathbf{E}}_{(\bar{s}, z) \sim \mathcal{P}^{n+1}}[(K(\bar{s}, z) - (b_{\bar{s}} + w))_+] + \mathop{\mathbf{E}}_{(\bar{s}, z) \sim \mathcal{P}^{n+1}}[((b_s - w) - K(\bar{s}, z))_+] \\
&= \mathop{\mathbf{E}}_{\bar{s} \sim \mathcal{P}^n}[\mathop{\mathbf{E}}_{z \sim \mathcal{P}}[(K(\bar{s}, z) - (b_{\bar{s}} + w))_+] + \mathop{\mathbf{E}}_{z \sim \mathcal{P}}[((b_{\bar{s}} - w) - K(\bar{s}, z))_+]] \\
&\leq \mathop{\mathbf{E}}_{\bar{s} \sim \mathcal{P}^n}[\max\{2\beta_{\bar{s}}^-, 2\beta_{\bar{s}}^+\}] \\
&\leq \mathop{\mathbf{E}}_{\bar{s} \sim \mathcal{P}^n}[2\beta_{\bar{s}}^- + 2\beta_{\bar{s}}^+].
\end{aligned}$$

We recall that by assumption,

$$\mathop{\mathbf{E}}_{\bar{s} \sim \mathcal{P}^n}[\beta_{\bar{s}}^+] = \mathop{\mathbf{E}}_{(\bar{s}, z) \sim \mathcal{P}^{n+1}}[(K(\bar{s}, z) - w)_+] \leq \beta$$

and

$$\mathop{\mathbf{E}}_{\bar{s} \sim \mathcal{P}^n}[\beta_{\bar{s}}^-] = \mathop{\mathbf{E}}_{(\bar{s}, z) \sim \mathcal{P}^{n+1}}[(-w - K(\bar{s}, z))_+] \leq \beta.$$

Hence we conclude that

$$\mathop{\mathbf{E}}_{(\bar{s}, z) \sim \mathcal{P}^{n+1}}[|\tilde{K}(\bar{s}, z) - K(\bar{s}, z)|] \leq 4\beta.$$

Finally, we need to argue about the stability of $\tilde{K}(\bar{s}, z)$. We assume that $K(\bar{s}, z)$ has stability γ. Suppose that $\bar{s}'$ is obtained from $\bar{s}$ by changing one coordinate. We claim that $|b_{\bar{s}'} - b_{\bar{s}}| \leq \gamma$. If not, suppose w.l.o.g. that $b_{\bar{s}'} > b_{\bar{s}} + \gamma$: Then we would have

$$\mathop{\mathbf{E}}_{z \sim \mathcal{P}}[(K(\bar{s}', z) - (b_{\bar{s}'} + w))_+] < \mathop{\mathbf{E}}_{z \sim \mathcal{P}}[((K(\bar{s}, z) + \gamma) - (b_{\bar{s}} + \gamma + w))_+] = \mathop{\mathbf{E}}_{z \sim \mathcal{P}}[(K(\bar{s}, z) - (b_{\bar{s}} + w))_+].$$

By a similar argument,

$$\mathop{\mathbf{E}}_{z \sim \mathcal{P}}[((b_{\bar{s}'} - w) - K(\bar{s}', z))_+] > \mathop{\mathbf{E}}_{z \sim \mathcal{P}}[((b_{\bar{s}} + \gamma + w) - (K(\bar{s}, z) + \gamma))_+] = \mathop{\mathbf{E}}_{z \sim \mathcal{P}}[((b_{\bar{s}} + w) - K(\bar{s}, z))_+].$$

However, by construction the left-hand sides should be equal, and also the right-hand sides should be equal, which is a contradiction.

Now we can prove that by changing one coordinate of $\bar{s}$, the value of $\tilde{K}(\bar{s}, z)$ cannot change by more than γ. Since both $K(\bar{s}, z)$ and $b_{\bar{s}}$ can change by at most γ when switching from $\bar{s}$ to $\bar{s}'$, we have

$$\begin{aligned}
\tilde{K}(\bar{s}', z) &= \mathtt{clamp}_{[b_{\bar{s}'} - w, b_{\bar{s}'} + w]}(K(\bar{s}', z)) \\
&\geq \mathtt{clamp}_{[b_{\bar{s}} - \gamma - w, b_{\bar{s}} - \gamma + w]}(K(\bar{s}, z) - \gamma) \\
&= \mathtt{clamp}_{[b_{\bar{s}} - w, b_{\bar{s}} + w]}(K(\bar{s}, z)) - \gamma \\
&= \tilde{K}(\bar{s}, z) - \gamma.
\end{aligned}$$

Similarly, we prove that $\tilde{K}(\bar{s}', z) \leq \tilde{K}(\bar{s}, z) + \gamma$. Therefore, $\tilde{K}(\bar{s}, z)$ has stability γ. □

Next, we use adaptive clamping to argue that, roughly speaking, when γ is small enough, we can reduce the range of $L(\bar{s}, z)$ without changing γ and affecting the estimation error significantly. An additional error term will appear due to the need to center the function $L(\bar{s}, z)$ for each fixed z before this operation; this additional error cannot be avoided.

Lemma 3.2. *Let $n \geq 4$, $\delta \leq \frac{1}{e}$, and $\gamma, R > 0$ such that $\gamma < \frac{R}{2\sqrt{n \ln(n/\delta)}}$. Then*

$$D_{4\delta}(n, R, \gamma) \leq D_\delta(n, R', \gamma) + \frac{2R}{\sqrt{n}} \sqrt{\ln(1/\delta)}$$

where $R' = 2\gamma\sqrt{n \ln(n/\delta)}$.

Proof. Let $L : Z^n \times Z \to [-R, R]$ be a function of stability γ, $\mathcal{P}$ a probability distribution over Z, and $\mathbf{E}_{z\sim\mathcal{P}}[L(\bar{s}, z)] = 0$. First, let us shift the function for each fixed z to make the expectation over $\bar{s}$ equal to 0. We define $\phi(z) \doteq \mathbf{E}_{\bar{s}\sim\mathcal{P}^n}[L(\bar{s}, z)]$ and $K(\bar{s}, z) = L(\bar{s}, z) - \phi(z)$. Since for every $z \in Z$, $K(\bar{s}, z)$ has sensitivity γ in $\bar{s}$ and $\mathbf{E}_{\bar{s}\sim\mathcal{P}^n}[K(\bar{s}, z)] = 0$, McDiarmid's inequality (Lemma 2.3) implies that

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}[K(\bar{s}, z) \geq \gamma\sqrt{n \ln(n/\delta)}] \leq \frac{\delta^2}{n^2},$$

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}[K(\bar{s}, z) \leq -\gamma\sqrt{n \ln(n/\delta)}] \leq \frac{\delta^2}{n^2}.$$

Since the range of K is bounded by $[-2R, 2R]$, this implies that

$$\mathop{\mathbf{E}}_{\bar{s}\sim\mathcal{P}^n}[(K(\bar{s}, z) - \gamma\sqrt{n \ln(n/\delta)})_+] \leq \frac{2R\delta^2}{n^2},$$

$$\mathop{\mathbf{E}}_{\bar{s}\sim\mathcal{P}^n}[(-\gamma\sqrt{n \ln(n/\delta)} - K(\bar{s}, z))_+] \leq \frac{2R\delta^2}{n^2}.$$

Obviously the same bounds remain valid when we also take the expectation over $z \sim \mathcal{P}$.

Next, we apply Lemma 3.1, with $w = \gamma\sqrt{n \ln(n/\delta)}$, $r = 2R$ and $\beta = \frac{2R\delta^2}{n^2}$. Hence there exists $b_{\bar{s}} \in [-w, w]$ for each $\bar{s} \in Z^n$ such that $\tilde{K}(\bar{s}, z) = \texttt{clamp}_{[b_{\bar{s}}-w, b_{\bar{s}}+w]}(K(\bar{s}, z))$ satisfies

- for every $\bar{s} \in Z^n$, $\mathbf{E}_{z\sim\mathcal{P}}[\tilde{K}(\bar{s}, z)] = 0$,
- $\mathbf{E}_{(\bar{s},z)\sim\mathcal{P}^{n+1}}[|\tilde{K}(\bar{s}, z) - K(\bar{s}, z)|] \leq 4\beta = \frac{8R\delta^2}{n^2}$,
- $\tilde{K}(\bar{s}, z)$ has uniform stability γ.

Also, since $b_{\bar{s}} \in [-w, w]$, the function $\tilde{K}(\bar{s}, z)$ is bounded by $[-2w, 2w] = [-R', R']$, where $R' = 2w = 2\gamma\sqrt{n \ln(n/\delta)}$ as in the statement of the lemma.

To summarize the relationship between $\tilde{K}(\bar{s}, z)$ and $L(\bar{s}, z)$, we have

$$L(\bar{s}, z) = K(\bar{s}, z) + \phi(z) = \tilde{K}(\bar{s}, z) + \phi(z) + (K(\bar{s}, z) - \tilde{K}(\bar{s}, z)).$$

We want to bound the leave-one-out estimation error of $L(\bar{s}, z)$,

$$\mathcal{E}_{\bar{s}}^{\leftarrow z}[L] = \frac{1}{n}\sum_{i=1}^{n} L(\bar{s}^{i\leftarrow z}, s_i).$$

From here, we can write

$$\mathcal{E}_{\bar{s}}^{\leftarrow z}[L] = \mathcal{E}_{\bar{s}}^{\leftarrow z}[\tilde{K}] + \frac{1}{n}\sum_{i=1}^{n}\phi(s_i) + \frac{1}{n}\sum_{i=1}^{n}(K(\bar{s}^{i\leftarrow z}, s_i) - \tilde{K}(\bar{s}^{i\leftarrow z}, s_i))$$

and

$$|\mathcal{E}_{\bar{s}}^{\leftarrow z}[L]| \leq |\mathcal{E}_{\bar{s}}^{\leftarrow z}[\tilde{K}]| + \left|\frac{1}{n}\sum_{i=1}^{n}\phi(s_i)\right| + \frac{1}{n}\sum_{i=1}^{n}\left|K(\bar{s}^{i\leftarrow z}, s_i) - \tilde{K}(\bar{s}^{i\leftarrow z}, s_i)\right|.$$

By the definition of D_δ, since $\tilde{K}(\bar{s}, z)$ has range $[-R', R']$ and uniform stability γ, we get

$$\mathbf{Pr}[|\mathcal{E}_{\bar{s}}^{\leftarrow z}[\tilde{K}]| \geq D_\delta(n, R', \gamma)] \leq \delta.$$

Let us bound the remaining two terms. The expression $\frac{1}{n}\sum_{i=1}^{n}\phi(s_i)$ is the average of n independent samples in the range $[-R, R]$, which can be viewed as a function of n independent random variables of sensitivity $\frac{2R}{n}$. Also, we know that $\mathbf{E}_{z\sim\mathcal{P}}[\phi(z)] = \mathbf{E}_{\bar{s}\sim\mathcal{P}^n}\mathbf{E}_{z\sim\mathcal{P}}[L(\bar{s}, z)] = 0$. Therefore by Lemma 2.3 (applied to $\frac{1}{n}\sum_i \phi(s_i)$ and $-\frac{1}{n}\sum_i \phi(s_i)$),

$$\mathbf{Pr}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\phi(s_i)\right| \geq \frac{R}{\sqrt{n}}\sqrt{2\ln\frac{1}{\delta}}\right] \leq 2\delta.$$

Finally, the expression $\mathbf{E}_{(\bar{s},z)\sim\mathcal{P}^{n+1}}[|K(\bar{s}^{i\leftarrow z}, s_i) - \tilde{K}(\bar{s}^{i\leftarrow z}, s_i)|]$ for each fixed i is bounded by $4\beta = \frac{8R\delta^2}{n^2}$ as we argued above. Hence

$$\mathop{\mathbf{E}}_{(\bar{s},z)\sim\mathcal{P}^{n+1}}\left[\frac{1}{n}\sum_{i=1}^{n}|K(\bar{s}^{i\leftarrow z}, s_i) - \tilde{K}(\bar{s}^{i\leftarrow z}, s_i)|\right] \leq \frac{8R\delta^2}{n^2}.$$

In this case, we simply use Markov's inequality, saying that

$$\mathop{\mathbf{Pr}}_{(\bar{s},z)\sim\mathcal{P}^{n+1}}\left[\frac{1}{n}\sum_{i=1}^{n}|K(\bar{s}^{i\leftarrow z}, s_i) - \tilde{K}(\bar{s}^{i\leftarrow z}, s_i)| \geq \frac{8R\delta}{n^2}\right] \leq \delta.$$

Therefore, by (3.2) and the union bound we obtain

$$\mathop{\mathbf{Pr}}_{(\bar{s},z)\sim\mathcal{P}^{n+1}}\left[|\mathcal{E}_{\bar{s}}^{\leftarrow z}[L]| \geq D_\delta(n, R', \gamma) + \frac{R}{\sqrt{n}}\sqrt{2\ln\frac{1}{\delta}} + \frac{8R\delta}{n^2}\right] \leq 4\delta.$$

For $\delta \leq 1/e$ and $n \geq 4$, we have $\frac{8R\delta}{n^2} \leq \frac{R}{e\sqrt{n}}\sqrt{\ln(1/\delta)}$, and $(\sqrt{2} + \frac{1}{e})\frac{R}{\sqrt{n}}\sqrt{\ln(1/\delta)} < \frac{2R}{\sqrt{n}}\sqrt{\ln(1/\delta)}$. Since this holds for every unbiased function $L(\bar{s}, z)$ of range $[-R, R]$ and stability γ, we obtain the statement of the lemma. □

3.3 Dataset size reduction

Here we show by a block partitioning argument that increasing the dataset size n cannot increase the estimation error significantly.

Lemma 3.3. *For positive integers $k, n' \geq 1$, $n = kn'$, and real $R, \gamma > 0$, $\delta > 0$, let $L \colon Z^n \times Z \to [-R, R]$ be a data-dependent, γ-uniformly stable function unbiased relative a distribution $\mathcal{P}$. Then*

$$D_\delta(n, R, \gamma) \leq D_{\delta/k}(n/k, R, \gamma).$$

Proof. Assume first $n = kn'$. Let $L(\bar{s}, z)$ be any function as described in the lemma. For a set of indices $I \subseteq [n]$ and $\bar{s} \in Z^n$ we denote $\bar{s}_I = (s_i)_{i \in I}$. Similarly, we denote

$$\mathcal{E}_{\bar{s}_I}^{\leftarrow z}[L] \doteq \frac{1}{|I|} \sum_{i \in I} L(\bar{s}^{i \leftarrow z}, s_i).$$

We partition the set $[n]$ into k blocks of size n': $B_1 = \{1, \dots, n'\}, B_2 = \{n' + 1, \dots, 2n'\}$, etc. Observe that

$$\mathcal{E}_{\bar{s}}^{\leftarrow z}[L] = \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in B_j} L(\bar{s}^{i \leftarrow z}, s_i) = \frac{1}{k} \sum_{j=1}^{k} \mathcal{E}_{\bar{s}_{B_j}}^{\leftarrow z}[L]. \tag{7}$$

If we condition on the values of s_i for $i \notin B_j$, we can view the quantity $\mathcal{E}_{\bar{s}_{B_j}}^{\leftarrow z}[L]$ as the estimation error for the function L restricted to the n' variables $s_i, i \in B_j$. Hence, for any fixed choice of the values $s_i, i \notin B_j$, we have by definition

$$\mathop{\mathbf{Pr}}_{\bar{s}_{B_j}, z \sim \mathcal{P}^{n'+1}} \left[\left| \mathcal{E}_{\bar{s}_{B_j}}^{\leftarrow z}[L] \right| \geq D_{\delta/k}(n', R, \gamma) \mid s_i : i \notin B_j \right] \leq \frac{\delta}{k}.$$

Since the bound is independent of the values of $s_i, i \notin B_j$, it remains valid if we remove the conditioning:

$$\mathop{\mathbf{Pr}}_{\bar{s}_{B_j}, z \sim \mathcal{P}^{n'+1}} \left[\left| \mathcal{E}_{\bar{s}_{B_j}}^{\leftarrow z}[L] \right| \geq D_{\delta/k}(n', R, \gamma) \right] \leq \frac{\delta}{k}.$$

By (7) and the union bound,

$$\mathop{\mathbf{Pr}}_{\bar{s}, z \sim \mathcal{P}^{n+1}} \left[|\mathcal{E}_{\bar{s}}^{\leftarrow z}[L]| \geq D_{\delta/k}(n', R, \gamma) \right] \leq \mathbf{Pr} \left[\exists j \in [k], \left| \mathcal{E}_{\bar{s}_{B_j}}^{\leftarrow z}[L] \right| \geq D_{\delta/k}(n', R, \gamma) \right] \leq \delta.$$

This means that $D_\delta(n, R, \gamma) \leq D_{\delta/k}(n', R, \gamma)$. □

We will need another version of this inequality, for the case where n is not divisible by n'.

Lemma 3.4. *For positive integers $n \geq n'$, and real $R, \gamma > 0$, $\delta > 0$, let $L \colon Z^n \times Z \to [-R, R]$ be a data-dependent, γ-uniformly stable function unbiased relative a distribution $\mathcal{P}$. Then*

$$D_\delta(n, R, \gamma) \leq D_{\delta/n}(n', R, \gamma).$$

Proof. We use the same argument as above, except that we use n overlapping blocks of size n': $B_1 = \{1, \dots, n'\}, B_2 = \{2, \dots, n' + 1\}$, etc. (using indices modulo n). Since each element appears in exactly n' blocks, we obtain

$$\mathcal{E}_{\bar{s}}^{\leftarrow z}[L] = \frac{1}{n'n} \sum_{j=1}^{n} \sum_{i \in B_j} L(\bar{s}^{i \leftarrow z}, s_i) = \frac{1}{n} \sum_{j=1}^{k} \mathcal{E}_{\bar{s}_{B_j}}^{\leftarrow z}[L]$$

instead of (7). The rest of the proof is exactly the same. We lose a factor of n in the δ parameter because of a union bound over n blocks. □

3.4 The inductive bound on estimation error

Here we combine the two reduction steps to prove our main bound on estimation error. It is convenient to state the inductive statement as follows.

Lemma 3.5. *For any $\delta \le \frac{1}{e}$, $a \in \mathbb{N}$, $n = 4^a$, $\gamma = \frac{1}{\sqrt{n}}$ and $R = 8\sqrt{\ln(n/\delta)}$,*

$$D_\delta(n, R, \gamma) \le \frac{8}{\sqrt{n}} \ln\left(\frac{n^2}{\delta}\right) \log_2 n.$$

Proof. We proceed by induction on a. The base case, $a = 1$, holds trivially, because here $n = 4, \gamma = \frac{1}{2}$, and the desired bound is $D_\delta(n, R, \gamma) \le 8 \ln \frac{16}{\delta}$ which holds because $D_\delta(n, R, \gamma) \le R = 8\sqrt{\ln \frac{4}{\delta}}$.

For the inductive step, consider $n = 4^{a+1} \ge 16$ and $\gamma = \frac{1}{\sqrt{n}} = \frac{1}{2^{a+1}}$. We use the two main ingredients that we proved above.

From the range reduction step (Lemma 3.2), since $\gamma = \frac{1}{\sqrt{n}} = \frac{R}{8\sqrt{n \ln(n/\delta)}}$, we obtain

$$D_\delta(n, R, \gamma) \le D_{\delta/4}(n, R/4, \gamma) + \frac{2R}{\sqrt{n}}\sqrt{\ln \frac{4}{\delta}}.$$

Next, we consider $n' = n/4 = 4^a$, $\gamma' = 2\gamma = \frac{1}{\sqrt{n'}}$ and $R' = 8\sqrt{\ln(n'/\delta)}$. Using the basic scaling identity (6) and $R' = 8\sqrt{\ln(n/(2\delta))} \ge 4\sqrt{\ln(n/\delta)} = R/2$, we obtain

$$D_\delta(n, R, \gamma) \le \frac{1}{2} D_{\delta/4}(n, R/2, 2\gamma) + \frac{2R}{\sqrt{n}}\sqrt{\ln \frac{4}{\delta}} \le \frac{1}{2} D_{\delta/4}(n, R', \gamma') + \frac{16}{\sqrt{n}} \ln \frac{n}{\delta}.$$

From the dataset size reduction step (Lemma 3.3), since $n = 4n'$, we obtain that

$$D_{\delta/4}(n, R', \gamma') \le D_{\delta/16}(n', R', \gamma').$$

Now we have an expression which is bounded by the inductive hypothesis:

$$D_{\delta/16}(n', R', \gamma') \le \frac{8}{\sqrt{n'}} \ln\left(\frac{n'^2}{\delta/16}\right) \log_2 n' = \frac{16}{\sqrt{n}} \ln\left(\frac{n^2}{\delta}\right)(-2 + \log_2 n).$$

Therefore we conclude that

$$D_\delta(n, R, \gamma) \le \frac{1}{2} D_{\delta/16}(n', R', \gamma') + \frac{16}{\sqrt{n}} \ln\left(\frac{n}{\delta}\right) \le \frac{8}{\sqrt{n}} \ln\left(\frac{n^2}{\delta}\right) \log_2 n.$$

□

From here, we can derive our main result by reducing it to Lemma 3.5. We first deal with the case when $\gamma \ge \frac{1}{4\sqrt{n \log(n/\delta)}}$.

Theorem 3.6. *For any $\delta \le \frac{1}{e}$, $n \ge 4$ and $\gamma \ge \frac{1}{4\sqrt{n \log(n/\delta)}}$,*

$$D_\delta(n, 1, \gamma) \le 16\gamma \ln\left(\frac{n^3}{\delta}\right) \log_2 n.$$

Proof. Let us scale the function by a factor of $R = 4\sqrt{\ln(n/\delta)}$, so we obtain a function with range R and uniform stability $\gamma' = R\gamma \geq \frac{1}{\sqrt{n}}$. Let $a' = \lfloor \log_4(1/\gamma'^2) \rfloor$. I.e, $n' = 4^{a'}$ is the largest power of 4 below $1/\gamma'^2 \leq n$. Let also $\gamma'' = \frac{1}{\sqrt{n'}} \geq \gamma'$. Since the range is $R = 4\sqrt{\ln(n/\delta)} \leq 8\sqrt{\ln(n'/\delta)}$, by Lemma 3.5,

$$D_\delta(n', R, \gamma'') \leq \frac{8}{\sqrt{n'}} \ln\left(\frac{n'^2}{\delta}\right) \log_2 n' = 8\gamma'' \ln\left(\frac{n'^2}{\delta}\right) \log_2 n'.$$

Since $\gamma'' \geq \gamma'$ and $n \geq 1/\gamma'^2 \geq n'$, we get by monotonicity in n (Lemma 3.4) and monotonicity in γ (obvious),

$$D_\delta(n, R, \gamma') \leq D_{\delta/n}(n', R, \gamma'') \leq 8\gamma'' \ln\left(\frac{n^3}{\delta}\right) \log_2 n.$$

Since n' is within a factor of 4 from $1/\gamma'^2$, we have $\gamma'' = \frac{1}{\sqrt{n'}} \leq 2\gamma'$. So,

$$D_\delta(n, R, \gamma') \leq 16\gamma' \ln\left(\frac{n^3}{\delta}\right) \log_2 n.$$

Finally, scaling back by a factor of $1/R$ (see (6)), we conclude that

$$D_\delta(n, 1, \gamma) \leq 16\gamma \ln\left(\frac{n^3}{\delta}\right) \log_2 n.$$

□

We also obtain a bound for smaller values of γ.

Theorem 3.7. *For any $\delta \leq \frac{1}{e}$, $n \geq 4$ and $\gamma < \frac{1}{4\sqrt{n \ln(n/\delta)}}$,*

$$D_\delta(n, 1, \gamma) \leq 16\gamma \ln\left(\frac{4n^3}{\delta}\right) \log_2 n + \frac{2}{\sqrt{n}}\sqrt{\ln(4/\delta)}.$$

Proof. Since $R = 1$ and $\gamma < \frac{1}{4\sqrt{n \ln(n/\delta)}}$, for $R' = 2\gamma\sqrt{n \ln(n/\delta)}$ we obtain from Lemma 3.2 that

$$D_\delta(n, R, \gamma) \leq D_{\delta/4}(n, R', \gamma) + \frac{2}{\sqrt{n}}\sqrt{\ln(4/\delta)}.$$

Now $\gamma = \frac{R'}{2\sqrt{n \ln(n/\delta)}}$, so we get from Theorem 3.6 and the scaling identity (6),

$$D_{\delta/4}(n, R', \gamma) = R' \, D_{\delta/4}(n, 1, \gamma/R') \leq 16\gamma \ln\left(\frac{4n^3}{\delta}\right) \log_2 n.$$

□

Finally we show how this implies Theorem 1.1. Let $M : Z^n \times Z \to [0, 1]$ be a data-dependent function of uniform stability γ. In Theorem 1.1, we have the quantity

$$\Delta_{\bar{s}}(M) = \left| \mathop{\mathbf{E}}_{\mathcal{P}}[M(\bar{s})] - \mathcal{E}_{\bar{s}}[M(\bar{s})] \right| = |\mathcal{E}_{\bar{s}}[L(\bar{s})]|$$

which differs by at most 2γ from the quantity $|\mathcal{E}_{\bar{s}}^{\leftarrow z}[L]|$, where $L(\bar{s}, z)$ has uniform stability 2γ (see Section 2). By definition, we have

$$\mathbf{Pr}\left[|\mathcal{E}_{\bar{s}}^{\leftarrow z}[L]| \geq D_\delta(n, 1, 2\gamma)\right] \leq \delta$$

and hence

$$\mathbf{Pr}\left[\Delta_{\bar{s}}(M) \geq D_\delta(n, 1, 2\gamma) + 2\gamma\right] \leq \delta.$$

By Theorems 3.6 and 3.7, we have

$$\begin{aligned} D_\delta(n, 1, 2\gamma) + 2\gamma &\leq 32\gamma \ln\left(\frac{4n^3}{\delta}\right) \log_2 n + \frac{2}{\sqrt{n}}\sqrt{\ln(4/\delta)} + 2\gamma \\ &\leq 32\gamma \ln\left(\frac{5n^3}{\delta}\right) \log_2 n + \frac{2}{\sqrt{n}}\sqrt{\ln(4/\delta)}. \end{aligned}$$

This proves Theorem 1.1.

4 Applications

We now apply our bounds on the estimation error to several known uniformly stable algorithms. Additional applications can be derived in a similar manner.

4.1 Learning via Stochastic Convex Optimization

We consider learning problems that can be formulated as stochastic convex optimization. Specifically, these are problems in which the goal is to minimize the expected loss:

$$F_{\mathcal{P}}(w) \doteq \mathop{\mathbf{E}}_{z \sim \mathcal{P}}[\ell(w, z)],$$

over $w \in \mathcal{K} \subset \mathbb{R}^d$ for some convex body $\mathcal{K}$ and a family of convex losses $\mathcal{F} = \{\ell(\cdot, z)\}_{z \in Z}$. The stochastic convex optimization problem for $\mathcal{F}$ is the problem of minimizing $F_{\mathcal{P}}(w)$ over $\mathcal{K}$ for an arbitrary distributions $\mathcal{P}$ over Z. The *excess loss* of a vector $\tilde{w}$ is $F_{\mathcal{P}}(\tilde{w}) - \min_{w \in \mathcal{K}} F_{\mathcal{P}}(w)$. We also denote the empirical loss by $F_{\bar{s}}(w) \doteq \frac{1}{n}\sum_{i \in [n]} \ell(w, s_i)$.

Many learning problems can be expressed in or relaxed to this general form. As a result many optimization algorithms are known and the optimal (excess) error rates are understood for a variety of families of convex functions. However most of these results are obtained via algorithm-specific techniques such as online-to-batch conversion [CCG04] and stability-based arguments rather than uniform convergence. As it turns out, this is unavoidable. This was first pointed out in the seminal work of Shalev-Shwartz et al. [SS+10] who showed that there is exists a gap between the bounds that can be obtained via uniform convergence (or ERM algorithms) and bounds achievable via alternative approaches.

For concreteness, let $\mathcal{F}$ be the family of all convex 1-Lipschitz losses over the unit Euclidean ball in d dimension (denoted by $\mathcal{B}_2^d(1)$). It is well-known that in this case the stochastic convex optimization problem can be solved with expected excess error $1/\sqrt{n}$ via projected SGD. At the same time it was shown in [SS+10] that there exists an algorithm that minimizes the empirical loss while having the worst case excess loss of $\Omega\left(\frac{\log d}{n}\right)$. This has been subsequently strengthened to $\Omega\left(\frac{d}{n}\right)$ by Feldman [Fel16] who also showed a lower bound of $\Omega\left(\sqrt{\frac{d}{n}}\right)$ for obtaining uniform convergence in this setting. Further, with Lipschitzness assumption replaced by the assumption that functions have range in $[0, 1]$ the gap becomes infinite even for $d = 2$ [Fel16].

4.1.1 Strongly convex ERM

We now revisit the stability results known for this basic setting [BE02; SS+10] (for simplicity and without loss of generality we will scale the domain and functions to 1).

Theorem 4.1 ([SS+10]). *Let $\mathcal{K} \subseteq \mathcal{B}_2^d(1)$ be a convex body, $\mathcal{F} = \{\ell(\cdot, z) \mid z \in Z\}$ be a family of* 1-*Lipschitz, λ-strongly convex loss functions over $\mathcal{K}$ with range in* $[0,1]$. *For a dataset $\bar{s} \in Z^n$ let $w_{\bar{s}}$ denote the empirical minimizer of loss on $\bar{s}$: $w_{\bar{s}} = \mathrm{argmin}_{w \in \mathcal{K}} F_{\bar{s}}(w)$. Then the algorithm $M(\bar{s}, z)$ that evaluates $\ell(w_{\bar{s}}, z)$ has uniform stability $\frac{4}{\lambda n}$.*

As an immediate corollary of this result and Theorem 1.1 we obtain:

Corollary 4.2. *In the setting of Thm. 4.1, there exists a constant c such that for every $\delta > 0$:*

$$\mathop{\mathbf{Pr}}_{\bar{s} \sim \mathcal{P}^n}\left[F_{\mathcal{P}}(w_{\bar{s}}) \geq \min_{w \in \mathcal{K}} F_{\mathcal{P}}(w) + \frac{c \log(n)\log(n/\delta)}{\lambda n} + \frac{c\sqrt{\log(1/\delta)}}{\sqrt{n}}\right] \leq \delta.$$

Theorem 4.1 requires strong convexity. As pointed out in [SS+10], it is possible to add a strongly convex regularizing term $\frac{\lambda}{2}\|w\|^2$ to the objective function that has sufficiently small effect on the loss function while ensuring stability (and generalization). Specifically, the objective function will change by at most λ since w is assumed to be in a ball of radius 1. By choosing $\lambda = \frac{\log n}{\sqrt{n}}$, we obtain a excess loss of $O\left(\frac{\log(n/\delta)}{\sqrt{n}}\right)$ (Corollary 1.2). This improves on the $O\left(\frac{\sqrt{\log(1/\delta)}}{n^{1/3}}\right)$ bound on excess loss obtained from the results in [FV18] by choosing $\lambda = 1/n^{2/3}$. We also remark that it is well-known that the same (up to a constant factor) stability bounds — and hence generalization bounds – apply to algorithms that minimize the (regularized) empirical loss within $1/n$. Therefore Corollary 4.2 leads to an efficient algorithm for solving the problem.

4.1.2 (Deterministic) gradient descent

We now recall the results of Hardt et al. [HRS16] for convex and smooth functions. These results derive their guarantees from the fact that a gradient step on a sufficiently smooth loss function is non-expansive. That is, for any pair of points w and w', any σ-smooth (that is, having a σ-Lipschitz gradient) convex function f, and $0 \leq \eta \leq 2/\sigma$,

$$\|(w - \eta\nabla f(w)) - (w' - \eta\nabla f(w'))\| \leq \|w - w'\|. \tag{8}$$

Projection to a convex body is also non-expansive. This implies that uniform stability can be proved for projected gradient descent of the following general form. For a vector $\bar{\eta}_t = (\eta_{t,1}, \ldots, \eta_{t,n})$ a gradient step with rate vector $\bar{\eta}_t$ is the update

$$w_{t+1} \leftarrow \mathtt{proj}_{\mathcal{K}}\left(w_t - \sum_{i \in [n]} \eta_{t+1,i}\nabla\ell(w_t, s_i)\right), \tag{9}$$

where $\mathtt{proj}_{\mathcal{K}}$ denotes projection to $\mathcal{K}$. For example, if a batch of size k is used in a gradient step with rate η_t then for each point s_i in the batch $\eta_{t,i} = \eta_t/k$ and for each point not in the batch $\eta_{t,i} = 0$. The non-expansiveness of the gradient steps and projections implies that the effect of each datapoint s_i on the loss of the solution can be bounded by $\sum_t \eta_{t,i}\|\nabla\ell(w_t, s_i)\|$. More formally, the following lemma follows directly from eq. (8) (and is a simple generalization of analysis in [HRS16] that only considers updates on a single data sample). We include the proof for completeness.

Lemma 4.3. *Let $\mathcal{K} \subseteq \mathcal{B}_2^d(1)$ be a convex body, $\mathcal{F} = \{\ell(\cdot, z) \mid z \in Z\}$ be a family of convex 1-Lipschitz and σ-smooth loss functions over $\mathcal{K}$ with range in $[0,1]$. For a dataset $\bar{s}$, number of iterations T and a sequence of rate vectors $\bar{\eta}_1, \ldots, \bar{\eta}_T$ let PGD$(w_0, (\bar{\eta}_t)_{t\in[T]}, \bar{s})$ denote the output of the algorithm that starting from $w_0 \in \mathcal{K}$, performs T updates according to eq. (9) and returns w_T. If for every $t \in [T]$, $\eta_t \doteq \|\bar{\eta}_t\|_1 \le 2/\sigma$ then, for every w_0, the algorithm $M(\bar{s}, z)$ that evaluates $\ell(w_T, z)$ on the output w_T of PGD$(w_0, (\bar{\eta}_t)_{t\in[T]}, \bar{s})$ has uniform stability $2 \cdot \left\|(\bar{\eta})_{t\in[T]}\right\|_{1,\infty}$, where*

$$\left\|(\bar{\eta})_{t\in[T]}\right\|_{1,\infty} \doteq \max_{i\in[n]} \sum_{t\in[T]} \eta_{t,i}.$$

Proof. Let $\bar{s}$ and $\bar{s}'$ be two datasets that differ in a single element at index i^*. Let $w_T = \mathrm{PGD}(w_0, (\bar{\eta}_t)_{t\in[T]}, \bar{s})$ and $w'_T = \mathrm{PGD}(w_0, (\bar{\eta}_t)_{t\in[T]}, \bar{s}')$. We prove the following claim by induction on T.

$$\|w_T - w'_T\|_2 \le 2 \sum_{t\in[T]} \eta_{t,i^*}.$$

The lemma will then follow from the definition of stability and 1-Lipschitness of $\ell(w_T, z)$. Clearly, the claim holds for $T = 0$. Now, assume that the claim holds for all $T \le \tau$.

$$\begin{aligned}
\|w_{\tau+1} - w'_{\tau+1}\|_2 &= \left\| \mathtt{proj}_{\mathcal{K}}\left(w_\tau - \sum_{i\in[n]} \eta_{\tau+1,i} \nabla \ell(w_\tau, s_i)\right) - \mathtt{proj}_{\mathcal{K}}\left(w'_\tau - \sum_{i\in[n]} \eta_{\tau+1,i} \nabla \ell(w'_\tau, s'_i)\right) \right\|_2 \\
&\le \left\| \left(w_\tau - \sum_{i\in[n]} \eta_{\tau+1,i} \nabla \ell(w_\tau, s_i)\right) - \left(w'_\tau - \sum_{i\in[n]} \eta_{\tau+1,i} \nabla \ell(w'_\tau, s'_i)\right) \right\|_2 \\
&\le \left\| \left(w_\tau - \sum_{i\in[n]\setminus\{i^*\}} \eta_{\tau+1,i} \nabla \ell(w_\tau, s_i)\right) - \left(w'_\tau - \sum_{i\in[n]\setminus\{i^*\}} \eta_{\tau+1,i} \nabla \ell(w'_\tau, s_i)\right) \right\|_2 \\
&\quad + \left\| \eta_{\tau+1,i^*} \nabla \ell(w_\tau, s_{i^*}) - \eta_{\tau+1,i^*} \nabla \ell(w_\tau, s'_{i^*}) \right\|_2 \\
&\le \|w_\tau - w'_\tau\|_2 + \left\| \eta_{\tau+1,i^*} \nabla \ell(w_\tau, s_{i^*}) - \eta_{\tau+1,i^*} \nabla \ell(w_\tau, s'_{i^*}) \right\|_2 \\
&\le 2 \sum_{t\in[\tau]} \eta_{t,i^*} + 2\eta_{\tau+1,i^*} = 2 \sum_{t\in[\tau+1]} \eta_{t,i^*},
\end{aligned}$$

where we used eq. (8) for gradient step at rate $\eta_{\tau+1} = \|\bar{\eta}_{\tau+1}\|_1$ on the function

$$f(w) = \sum_{i\in[n]\setminus\{i^*\}} \frac{\eta_{\tau+1,i}}{\eta_{\tau+1}} \ell(w, s_i)$$

to obtain the fifth line. Note that f is a convex combination of functions from $\mathcal{F}$ and therefore is σ-smooth and by our assumption $\eta_{\tau+1} \le 2/\sigma$. □

Lemma 4.3 together with Theorem 1.1 immediately implies generalization bounds for a variety of versions of gradient descent with different rates, arbitrary batch sizes and multiple passes over the data. For most such algorithms no alternative analyses of estimation error are known. Importantly, the estimation error can be bounded without any assumptions on how close the output of the algorithm is to the empirical minimum. Therefore this approach gives generalization bounds for algorithms used in practice as opposed to rates and number of iterations that are necessary for a theoretical proof of convergence (but are rarely used in practice).

As a concrete example we give a corollary for running full gradient descent with standard rates that guarantee convergence to within $1/\sqrt{n}$ of the empirical minimum. We are not aware of any other approaches to proving generalization guarantees for this algorithm in this general setting. Let PGD$(T, \eta, \bar{s})$ denote PGD$(w_0, (\bar{\eta}_t)_{t\in[T]}, \bar{s})$ for w_0 being the origin and $\eta_{t,i} = \eta/n$ for all $i \in [n]$ and $t \in [T]$. Standard analysis of gradient descent on σ-smooth functions (*e.g.* [Bub15]) implies that in the setting of Lemma 4.3, PGD$(T, 1/\sigma, \bar{s})$ outputs $w_{\bar{s}}$ such that

$$F_{\bar{s}}(w_{\bar{s}}) \leq \min_{w\in\mathcal{K}} F_{\bar{s}}(w) + \frac{2}{\eta T}. \tag{10}$$

By optimizing the choice of T, the best previous bound in [FV18] gives an upper bound of $O\left(\frac{\sqrt{\log(1/\delta)}}{n^{1/3}}\right)$ on excess loss. Similarly, applying our improved bounds gives the following statement.

Corollary 4.4. *Let $\mathcal{K} \subseteq \mathcal{B}_2^d(1)$ be a convex body, $\mathcal{F} = \{\ell(\cdot, z) \mid z \in Z\}$ be a family of convex 1-Lipschitz and σ-smooth loss functions over $\mathcal{K}$ with range in $[0, 1]$. For every distribution $\mathcal{P}$ over Z, $\delta > 0$, $w_{\bar{s}} \doteq$ PGD$(T, \eta, \bar{s})$ for $\eta = 1/\sigma$ and $T = \lfloor \sigma\sqrt{n}/\log n \rfloor$, and some fixed constant c*

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}\left[F_{\mathcal{P}}(w_{\bar{s}}) \geq \min_{w\in\mathcal{K}} F_{\mathcal{P}}(w) + \frac{c\log(n/\delta)}{\sqrt{n}}\right] \leq \delta.$$

Proof. We first note that, by Lemma 4.3, PGD$(T, \eta, \bar{s})$ with $\eta = 1/\sigma$ and $T = \lfloor \sigma\sqrt{n}/\log n \rfloor$ is $\frac{2}{\sqrt{n}\log n}$ uniformly stable (here we can assume that $\sigma \geq \log n/\sqrt{n}$ since otherwise $T = 0$ and w_0 has the desired property since the range of every loss function is within $\log n/\sqrt{n}$ of some constant).

We next denote $w^* \doteq \text{argmin}_{w\in\mathcal{K}} F_{\mathcal{P}}(w)$ and use the following standard decomposition of excess loss:

$$F_{\mathcal{P}}(w_{\bar{s}}) - F_{\mathcal{P}}(w^*) \leq |F_{\mathcal{P}}(w_{\bar{s}}) - F_{\bar{s}}(w_{\bar{s}})| + \left|F_{\bar{s}}(w_{\bar{s}}) - \min_{w\in\mathcal{K}} F_{\bar{s}}(w)\right| + \min_{w\in\mathcal{K}} F_{\bar{s}}(w) - F_{\mathcal{P}}(w^*).$$

Theorem 1.1 gives an upper-bound of $\frac{c_0\log(n/\delta)}{\sqrt{n}}$ (for some constant c_0) on the first term that holds with probability $1 - \delta/2$. Equation (10) upper bounds the second term by $\frac{4\log n}{\sqrt{n}}$ (we use an additional factor of 2 to account for the $\lfloor\cdot\rfloor$ operation). Finally, $\min_{w\in\mathcal{K}} F_{\bar{s}}(w)$ has sensitivity of $1/n$ and

$$\mathop{\mathbf{E}}_{\bar{s}\sim\mathcal{P}^n}\left[\min_{w\in\mathcal{K}} F_{\bar{s}}(w)\right] \leq \mathop{\mathbf{E}}_{\bar{s}\sim\mathcal{P}^n}[F_{\bar{s}}(w^*)] = F_{\mathcal{P}}(w^*).$$

Therefore, by McDiarmid's inequality (Lemma 2.3),

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}\left[\min_{w\in\mathcal{K}} F_{\bar{s}}(w) - F_{\mathcal{P}}(w^*) \geq \frac{\sqrt{2\ln(2/\delta)}}{\sqrt{n}}\right] \leq \delta/2.$$

Combining the upper bounds on the three terms and using the union bound we obtain the claim. □

4.1.3 Stochastic gradient descent

The analysis above applies to gradient descent with the rates chosen deterministically. In practice, a variety of randomized strategies for picking the batches are use with the most common ones being random shuffling and random sampling with replacement. We describe a strategy for picking which samples to used by a distribution $\mathcal{U}$ over sequences of rate vectors that result from this strategy. We also denote by $\mathrm{PSGD}(w_0, \mathcal{U}, \bar{s})$ the corresponding stochastic gradient descent algorithm: sample $(\bar{\eta}_t)_{t\in[T]}$ from $\mathcal{U}$ and run $\mathrm{PGD}(w_0, (\bar{\eta}_t)_{t\in[T]}, \bar{s})$.

A simple way to obtain generalization bounds for stochastic gradient descent is to consider the expectation $\mathbf{E}_{\mathcal{U}}[(\bar{\eta}_t)_{t\in[T]}]$. By convexity of the absolute value function, the uniform stability parameter of the expectation of the loss over the randomness of PSGD is upper-bounded by $2 \cdot \left\|\mathbf{E}_{\mathcal{U}}[(\bar{\eta}_t)_{t\in[T]}]\right\|_{1,\infty}$. Most standard sampling schemes used by SGD are symmetric with respect to samples and therefore the expected rate vector $\mathbf{E}_{\mathcal{U}}[\bar{\eta}_t] = \frac{\eta_t}{n}(1, \ldots, 1)$, where η_t is the rate of the batch used at step t [HRS16]. Combining this simple analysis with Theorem 1.1 gives generalization with high probability over the dataset but in expectation over the sampling of points.

To obtain high-probability bounds we simply observe that for most common sampling schemes, $\left\|(\bar{\eta}_t)_{t\in[T]}\right\|_{1,\infty}$ is highly concentrated around its expectation. In particular, the norm can be upper-bounded with high probability with relatively low overhead. More formally, we state the following general form of bounds on the estimation error of PSGD.

Theorem 4.5. *Let $\mathcal{K} \subseteq \mathcal{B}_2^d(1)$ be a convex body, $\mathcal{F} = \{\ell(\cdot, z) \mid z \in Z\}$ be a family of convex 1-Lipschitz and σ-smooth loss functions over $\mathcal{K}$ with range in $[0, 1]$. For a number of iterations T let $\mathcal{U}$ be a distribution over sequences of rate vectors of length T and assume that for every $(\bar{\eta}_t)_{t\in[T]}$ in the support $\mathcal{U}$, $\|\bar{\eta}_t\|_1 \leq 2/\sigma$ for all $t \in [T]$.*

Assume that for some $\beta \geq 0$,

$$\Pr_{(\bar{\eta}_t)_{t\in[T]}\sim\mathcal{U}}\left[\left\|(\bar{\eta}_t)_{t\in[T]}\right\|_{1,\infty} \geq \zeta\right] \leq \beta.$$

Then there exist a constant c such that for every distribution $\mathcal{P}$ over Z and $w_0 \in \mathcal{K}$,

$$\Pr_{\bar{s}\sim\mathcal{P}^n,\ w_{\bar{s},T}=\mathit{PSGD}(w_0,\mathcal{U},\bar{s})}\left[|F_{\mathcal{P}}(w_{\bar{s},T}) - F_{\bar{s}}(w_{\bar{s},T})| \geq c\left(\zeta\log(n)\log(n/\delta) + \frac{\sqrt{\log(1/\delta)}}{\sqrt{n}}\right)\right] \leq \beta + \delta.$$

Random shuffling: We first consider random shuffling based schemes (also referred to as sampling without replacement). In such schemes the dataset is split into batches randomly and uniformly. All the batches are used to update the gradient in every pass over the dataset. For every pass over the data each of the samples is used exactly once. Hence the contribution of each pass over the data to the stability parameter is upper bounded by the largest rate used in that pass. Specifically, if the batch size is k, r passed are performed, and in each pass i the largest rate used for a batch is η_i, then for $\zeta = \frac{1}{k}\sum_{i\in[r]}\eta_i$, $\left\|(\bar{\eta}_t)_{t\in[T]}\right\|_{1,\infty}$ is upper bounded by ζ with probability 1. In particular, Theorem 4.5 can be applied with ζ defined as above and $\beta = 0$. While our results give a bound on the estimation error, unfortunately very little is known about the empirical error of gradient descent with random shuffling. In particular, known results for random shuffling are in more restrictive settings [RR12; GOP15; Sha16; LR16; PVRB18] and in most cases only apply to function classes simple enough that one can appeal to complexity-based generalization bounds instead of stability.

Sampling with replacement: Another common sampling scheme uses random and independent sampling with replacement: that is for every iteration a batch of k samples is chosen randomly, uniformly and independently of previous batches. For each of the T iterations, sample s_i is included with probability k/n and therefore the sum of rates for sample i is distributed as $\frac{1}{k}\sum_{t\in[T]}\eta_t B(k/n)$, where $B(k/n)$ is the Bernoulli random variable with bias k/n. We can now use standard concentration inequalities and the union bound to upper bound the largest sum of rates. For example, if $T = O(n/k)$ (which corresponds to a constant number of passes) and the rate is fixed to η then, by (the multiplicative) Chernoff bound, for some constant c_0,

$$\mathop{\mathbf{Pr}}_{(\bar{\eta}_t)_{t\in[T]}\sim\mathcal{U}}\left[\left\|(\bar{\eta}_t)_{t\in[T]}\right\|_{1,\infty} \geq \frac{c_0\eta\log(n/\beta)}{k}\right] \leq \beta. \tag{11}$$

Hence even with a constant number of passes and batches of size 1 the overhead of getting generalization with high probability over the randomness of the algorithm is at most logarithmic. The overhead becomes (relatively) smaller as the number of passes grows.

As a concrete corollary we give high-probability generalization bounds for PSGD with $k = 1$ and fixed rate $\eta = 1/\sqrt{T}$. We denote this rate distribution by $\mathcal{U}_{1,T}$ and denote the origin in $\mathbb{R}^d$ by $\bar{0}$ (and thus the algorithm can is exactly PSGD($\bar{0}, \mathcal{U}_{1,T}, \bar{s}$)). To get a high-probability bound on the empirical loss of this algorithm we note that sampling with replacement corresponds to drawing i.i.d. samples from the uniform distribution over the samples in $\bar{s}$. In particular, the expected loss function in this case is exactly $F_{\bar{s}}$. Standard high-probability generalization bounds for PSGD imply that it minimizes the empirical loss with high-probability but require outputting the average of the iterates (these results are obtained via online-to-batch conversion [CCG04]). Theorem 4.5 applies to the average of the iterates since Lemma 4.3 applies to it as well (or any other convex combination of the iterates). To get an upper-bound for PSGD($\bar{0}, \mathcal{U}_{1,T}, \bar{s}$) (which outputs the last iterate) we use a recent work of Harvey et al. [Har+18] that shows high-probability bound on suboptimality of the last iterate of PSGD[1] with a slightly worse rate.

Lemma 4.6. *[Har+18; Har18] Let $\mathcal{K} \subseteq \mathcal{B}_2^d(1)$ be a convex body, $\mathcal{F} = \{\ell(\cdot, z) \mid z \in Z\}$ be a family of convex 1-Lipschitz loss functions over $\mathcal{K}$ with range in $[0,1]$. There exists a constant c such that for every $\bar{s} \in Z^n$ and $\delta > 0$,*

$$\mathop{\mathbf{Pr}}_{w_T=\textit{PSGD}(\bar{0},\mathcal{U}_{1,T},\bar{s})}\left[F_{\bar{s}}(w_T) \geq \min_{w\in\mathcal{K}} F_{\bar{s}}(w) + \frac{c\log(T)\log(1/\delta)}{\sqrt{T}}\right] \leq \delta.$$

Combining Lemma 4.6 with Theorem (4.5) (used with eq. 11), we obtain the following bound on the generalization error of PSGD($\bar{0}, \mathcal{U}_{1,T}, \bar{s}$) for $T = n$.

Corollary 4.7. *Let $\mathcal{K} \subseteq \mathcal{B}_2^d(1)$ be a convex body, $\mathcal{F} = \{\ell(\cdot, z) \mid z \in Z\}$ be a family of convex 1-Lipschitz $2\sqrt{n}$-smooth loss functions over $\mathcal{K}$ with range in $[0,1]$. There exists a constant c such that for every distribution $\mathcal{P}$ over Z and $\delta > 0$,*

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n,\ w_n=\textit{PSGD}(\bar{0},\mathcal{U}_{1,n},\bar{s})}\left[F_{\mathcal{P}}(w_T) \geq \min_{w\in\mathcal{K}} F_{\mathcal{P}}(w) + \frac{c\log(n)\log^2(n/\delta)}{\sqrt{n}}\right] \leq \delta.$$

Remark 4.8. *Finally, we note that the results in this section can be extended to non-smooth functions by applying a smoothing operation to each convex loss function before optimization. A variety of approaches*

[1] The results stated there are for the decaying rate $\eta_t = 1/\sqrt{t}$, but the same result applies to the fixed rate we use here [Har18].

*to smoothing are known (*e.g. *[BT12]). For our purposes it suffices to observe that for every convex* 1*-Lipschitz function f over $\mathcal{K}$ of radius 1, the standard smoothing via Moreau envelope can be used to obtain a σ-smooth 1-Lipschitz function $\tilde{f}$ such that $|\tilde{f}(w) - f(w)| \leq 1/(2\sigma)$ for all $w \in \mathcal{K}$. Thus we can apply the optimization to $\sqrt{n}$-smooth loss functions that are within $1/\sqrt{n}$ (in infinity norm) of the corresponding functions in $\mathcal{F}$. Note that this level of smoothness and additional error suffice to extend Corollary 4.4 (with $T = n/\log n$) and Corollary 4.7 to non-smooth functions with essentially the same bound on the excess loss.*

4.2 Privacy-Preserving Prediction

Our results can also be used to improve the bounds on generalization error of learning algorithms with differentially private prediction. These are algorithms introduced to model privacy-preserving learning in the settings where users only have black-box access to the model via a prediction interface [DF18]. Formally,

Definition 4.9 ([DF18])**.** *Let K be an algorithm that given a dataset $\bar{s} \in (X \times Y)^n$ and a point $x \in X$ produces a value in Y. Then K is ϵ-*differentially private prediction *algorithm if for every $x \in X$, the output $K(\bar{s}, x)$ is ϵ-differentially private with respect to $\bar{s}$.*

The properties of differential privacy imply that the expectation over the randomness of K of the loss of K at any point is uniformly stable. Specifically, for every ϵ-differentially private prediction algorithm, every loss function $\ell_Y \colon Y \times Y \to [0,1]$, two datasets $\bar{s}$ and $\bar{s}'$ that differ in a single element and $(x,y) \in X \times Y$ we have that

$$\mathop{\mathbf{E}}_{K}[\ell_Y(K(\bar{s},x),y)] \leq e^{\epsilon} \cdot \mathop{\mathbf{E}}_{K}[\ell_Y(K(\bar{s}',x),y)].$$

In particular, this implies that

$$\left|\mathop{\mathbf{E}}_{K}[\ell_Y(K(\bar{s},x),y)] - \mathop{\mathbf{E}}_{K}[\ell_Y(K(\bar{s}',x),y)]\right| \leq e^{\epsilon} - 1.$$

Therefore our generalization bounds can be applied to the data-dependent function $M(\bar{s},(x,y)) \doteq \mathbf{E}_K[\ell_Y(K(\bar{s},x),y)]$. This gives the following corollary of Theorem 1.1:

Theorem 4.10. *For $\epsilon \in (0,1)$, let $K : (X \times Y)^n \times X \to Y$ be an ϵ-differentially private prediction and $\ell_Y \colon Y \times Y \to [0,1]$ be an arbitrary loss function. Let $M(\bar{s},(x,y)) \doteq \mathbf{E}_K[\ell_Y(K(\bar{s},x),y)]$. Then there exists a constant c such for any probability distribution $\mathcal{P}$ over Z and any $\delta \in (0,1)$:*

$$\mathop{\mathbf{Pr}}_{\bar{s}\sim\mathcal{P}^n}\left[\left|\mathop{\mathbf{E}}_{\mathcal{P}}[M(\bar{s})] - \mathcal{E}_{\bar{s}}[M(\bar{s})]\right| \geq c\epsilon\log(n)\log(n/\delta) + \frac{\sqrt{2\ln(4/\delta)}}{\sqrt{n}}\right] \leq \delta.$$

These bounds are stronger than those obtained in [DF18] in several parameter regimes (but are more generally incomparable since bounds in [DF18] are multiplicative).

Dwork and Feldman [DF18] describe an algorithm for agnostically learning threshold functions on a line with differentially private prediction. Their analysis of the generalization error of this algorithm relies crucially on the generalization properties of differentially private prediction. Their weaker generalization bound does not give the high-probability bound on the generalization error that is necessary for satisfying the standard definition of agnostic learning. By plugging Thm. 4.10 we obtain a bound on generalization error that holds with high probability and achieves the optimal rate (up to logarithmic factors). We omit more formal details since they require several additional definitions and the application itself is straightforward.

Acknowledgments

We thank Nick Harvey, Tomer Koren, Mehryar Mohri, Sasha Rakhlin, Yoram Singer, Karthik Sridharan, Csaba Szepesvari and Kunal Talwar for thoughtful discussions and insightful comments about this work.

References

[AS18] K. T. Abou-Moustafa and C. Szepesvári. "An Exponential Tail Bound for Lq Stable Learning Rules. Application to k-Folds Cross-Validation". In: *ISAIM*. 2018.

[Bas+16] R. Bassily, K. Nissim, A. D. Smith, T. Steinke, U. Stemmer, and J. Ullman. "Algorithmic stability for adaptive data analysis". In: *STOC*. 2016, pp. 1046–1059.

[BE02] O. Bousquet and A. Elisseeff. "Stability and generalization". In: *JMLR* 2 (2002), pp. 499–526.

[BKL99] A. Blum, A. Kalai, and J. Langford. "Beating the Hold-Out: Bounds for K-fold and Progressive Cross-Validation". In: *COLT*. 1999, pp. 203–208.

[BT12] A. Beck and M. Teboulle. "Smoothing and first order methods: A unified framework". In: *SIAM Journal on Optimization* 22.2 (2012), pp. 557–580.

[Bub15] S. Bubeck. "Convex Optimization: Algorithms and Complexity". In: *Foundations and Trends in Machine Learning* 8.3-4 (2015), pp. 231–357.

[CCG04] N. Cesa-Bianchi, A. Conconi, and C. Gentile. "On the Generalization Ability of On-Line Learning Algorithms". In: *IEEE Transactions on Information Theory* 50.9 (2004), pp. 2050–2057.

[CG16] A. Celisse and B. Guedj. "Stability revisited: new generalisation bounds for the Leave-one-Out". In: *arXiv preprint arXiv:1608.06412* (2016).

[CJY18] Y. Chen, C. Jin, and B. Yu. "Stability and convergence trade-off of iterative optimization algorithms". In: *arXiv preprint arXiv:1804.01619* (2018).

[CMS11] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. "Differentially Private Empirical Risk Minimization". In: *Journal of Machine Learning Research* 12 (2011), pp. 1069–1109.

[CP18] Z. B. Charles and D. S. Papailiopoulos. "Stability and Generalization of Learning Algorithms that Converge to Global Optima". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. 2018, pp. 744–753.

[DF18] C. Dwork and V. Feldman. "Privacy-preserving Prediction". In: *CoRR* abs/1803.10266 (2018). Extended abstract in COLT 2018. arXiv: `1803.10266`.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[DW79a] L. Devroye and T. J. Wagner. "Distribution-free inequalities for the deleted and holdout error estimates". In: *IEEE Trans. Information Theory* 25.2 (1979), pp. 202–207.

[DW79b] L. Devroye and T. J. Wagner. "Distribution-free performance bounds with the resubstitution error estimate (Corresp.)" In: *IEEE Trans. Information Theory* 25.2 (1979), pp. 208–210.

[Dwo+06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. "Calibrating noise to sensitivity in private data analysis". In: *TCC*. 2006, pp. 265–284.

[Dwo+14] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. "Preserving Statistical Validity in Adaptive Data Analysis". In: *CoRR* abs/1411.2664 (2014). Extended abstract in STOC 2015.

[EEP05] A. Elisseeff, T. Evgeniou, and M. Pontil. "Stability of Randomized Learning Algorithms". In: *Journal of Machine Learning Research* 6 (2005), pp. 55–79.

[Fel16] V. Feldman. "Generalization of ERM in Stochastic Convex Optimization: The Dimension Strikes Back". In: *CoRR* abs/1608.04414 (2016). Extended abstract in NIPS 2016.

[FV18] V. Feldman and J. Vondrák. "Generalization Bounds for Uniformly Stable Algorithms". In: *Proceedings of NeurIPS*. 2018, pp. 9770–9780.

[GOP15] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. "Why random reshuffling beats stochastic gradient descent". In: *arXiv preprint arXiv:1510.08560* (2015).

[Har18] N. Harvey. Personal communication. 2018.

[Har+18] N. J. A. Harvey, C. Liaw, Y. Plan, and S. Randhawa. "Tight Analyses for Non-Smooth Stochastic Gradient Descent". In: *CoRR* abs/1812.05217 (2018). arXiv: `1812.05217`.

[HRS16] M. Hardt, B. Recht, and Y. Singer. "Train faster, generalize better: Stability of stochastic gradient descent". In: *ICML*. 2016, pp. 1225–1234.

[KKV11] S. Kale, R. Kumar, and S. Vassilvitskii. "Cross-Validation and Mean-Square Stability". In: *Innovations in Computer Science - ICS*. 2011, pp. 487–495.

[KL18] I. Kuzborskij and C. H. Lampert. "Data-Dependent Stability of Stochastic Gradient Descent". In: *ICML*. 2018, pp. 2820–2829.

[KR99] M. J. Kearns and D. Ron. "Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation". In: *Neural Computation* 11.6 (1999), pp. 1427–1453.

[Kum+13] R. Kumar, D. Lokshtanov, S. Vassilvitskii, and A. Vattani. "Near-Optimal Bounds for Cross-Validation via Loss Stability". In: *ICML*. 2013, pp. 27–35.

[Liu+17] T. Liu, G. Lugosi, G. Neu, and D. Tao. "Algorithmic Stability and Hypothesis Complexity". In: *ICML*. 2017, pp. 2159–2167.

[Lon17] B. London. "A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent". In: *NIPS*. 2017, pp. 2935–2944.

[LP94] G. Lugosi and M. Pawlak. "On the posterior-probability estimate of the error rate of nonparametric classification rules". In: *IEEE Trans. Information Theory* 40.2 (1994), pp. 475–481.

[LR16] J. Lin and L. Rosasco. "Optimal Learning for Multi-pass Stochastic Gradient Methods". In: *NIPS*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. 2016, pp. 4556–4564.

[Mau17] A. Maurer. "A Second-order Look at Stability and Generalization". In: *COLT*. 2017, pp. 1461–1475.

[Muk+06] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. "Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization". In: *Advances in Computational Mathematics* 25.1-3 (2006), pp. 161–193.

[NS17] K. Nissim and U. Stemmer. "Concentration Bounds for High Sensitivity Functions Through Differential Privacy". In: *CoRR* abs/1703.01970 (2017). arXiv: `1703.01970`.

[Pog+04] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. “General conditions for predictivity in learning theory”. In: *Nature* 428.6981 (2004), pp. 419–422.

[PVRB18] L. Pillaud-Vivien, A. Rudi, and F. Bach. “Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes”. In: *NeurIPS*. 2018, pp. 8125–8135.

[Riv+18] O. Rivasplata, C. Szepesvari, J. S. Shawe-Taylor, E. Parrado-Hernandez, and S. Sun. “PAC-Bayes bounds for stable algorithms with instance-dependent priors”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 9234–9244.

[RR12] B. Recht and C. Ré. “Toward a Noncommutative Arithmetic-geometric Mean Inequality: Conjectures, Case-studies, and Consequences”. In: *COLT*. 2012, pp. 11.1–11.24.

[RW78] W. H. Rogers and T. J. Wagner. “A Finite Sample Distribution-Free Performance Bound for Local Discrimination Rules”. In: *The Annals of Statistics* 6.3 (1978), pp. 506–514.

[Sha16] O. Shamir. “Without-Replacement Sampling for Stochastic Gradient Methods: Convergence Results and Application to Distributed Optimization”. In: *CoRR* abs/1603.00570 (2016).

[SS+10] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. “Learnability, stability and uniform convergence”. In: *The Journal of Machine Learning Research* 11 (2010), pp. 2635–2670.

[SU17] T. Steinke and J. Ullman. “Subgaussian Tail Bounds via Stability Arguments”. In: *arXiv preprint arXiv:1701.03493* (2017).

[WP09] R. L. Wibisono Andre and T. Poggio. *Sufficient conditions for uniform stability of regularization algorithms*. Tech. rep. MIT-CSAIL-TR-2009-060. MIT, 2009.

[WR18] N. Weinberger and A. Rakhlin. *On High Probability Bounds for Uniformly Stable Learning Algorithms*. Unpublished manuscript. 2018.

[Wu+17] X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton. “Bolt-on differential privacy for scalable stochastic gradient descent-based analytics”. In: *(SIGMOD)*. 2017, pp. 1307–1322.

[Zha03] T. Zhang. “Leave-One-Out Bounds for Kernel Methods”. In: *Neural Computation* 15.6 (2003), pp. 1397–1437.