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

Vitaly Feldman[*]
Google Brain

Jan Vondrák
Stanford University

Editors: Alina Beygelzimer and Daniel Hsu

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 for algorithms that are uniformly stable with $\gamma = O(1/\sqrt{n})$, estimation error is essentially the same as the sampling error. 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

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

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

0. Extended abstract. Full version appears as (Feldman and Vondrák, 2019, v2).

© 2019 V. Feldman & J. Vondrák.

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 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 $\ell_Y : Y \times Y \to \mathbb{R}_+$, 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 (1978); Devroye and Wagner (1979a,b). 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. (Bousquet and Elisseeff, 2002; Mukherjee et al., 2006; Shalev-Shwartz et al., 2010)). 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 (1994) for several specific learning algorithms. In a seminal work Bousquet and Elisseeff (2002) 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 (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010). More recently, Hardt et al. (2016) 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 (Shalev-Shwartz et al., 2010; Feldman, 2016). 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 (Bousquet and Elisseeff, 2002) 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 (Shalev-Shwartz et al., 2010). 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 (Hardt et al., 2016). 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 (Shalev-Shwartz et al., 2010; Hardt et al., 2016).

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. (2010) for learning algorithms that minimize the empirical risk. For such algorithms they showed that

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

allowing them to apply Markov's inequality.

Generalization properties of uniform stability were addressed in a recent work by the authors (Feldman and Vondrák, 2018). 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 (Bousquet and Elisseeff, 2002).

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 (Bousquet and Elisseeff, 2002; Feldman and Vondrák, 2018) give similar generalization guarantees only when $\gamma = O(1/n)$.

Proof approach: The high-probability generalization result in (Bousquet and Elisseeff, 2002) (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 (Feldman and Vondrák, 2018) is based on a technique developed in (Bassily et al., 2016) 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 (2018) who used a bound on the second moment from (Feldman and Vondrák, 2018) 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 (Feldman and Vondrák, 2018) (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.

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 $4/(\lambda n)$ (Bousquet and Elisseeff, 2002). We therefore obtain high-probability generalization bounds on ERM in this case that improve on the known results for any $\lambda = o(1)$.

Using stability of ERM for strongly convex functions, Shalev-Shwartz et al. (2010) 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 (Hardt et al., 2016). 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 (Hardt et al., 2016) 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. 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.

For comparison, a recent work of London (2017) considers extension of the generalization guarantees in (Hardt et al., 2016) 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 ((London, 2017) 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 (Dwork and Feldman, 2018). 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 (Dwork and Feldman, 2018) (that relies on models of unbounded complexity).

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 (Devroye and Wagner, 1979a) where high probability bounds on the generalization error of k-NN are proved). See (Devroye et al., 1996) 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 (Blum et al., 1999; Kale et al., 2011; Kumar et al., 2013)).

A long line of work focuses on the relationship between various notions of stability and learnability in supervised setting (see (Kearns and Ron, 1999; Poggio et al., 2004; Shalev-Shwartz et al., 2010) for an overview). This work employs relatively weak notions of average stability and derives a variety of asymptotic equivalence results. The results in (Bousquet and Elisseeff, 2002) on

uniform stability and their applications to generalization properties of strongly convex ERM algorithms have been extended and generalized in several directions (e.g. (Zhang, 2003; Wibisono and Poggio, 2009)). Maurer (2017) 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 (2018) 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 (2018) and Celisse and Guedj (2016) gives high probability generalization bounds similar to those in (Bousquet and Elisseeff, 2002) 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 et al., 2017; Rivasplata et al., 2018; Charles and Papailiopoulos, 2018; Chen et al., 2018). We also remark that all these works are based on techniques different from ours.

Uniform stability has several additional important connections to differential privacy (Dwork et al., 2006). 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 (Dwork et al., 2014; Bassily et al., 2016). 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 (Steinke and Ullman, 2017; Nissim and Stemmer, 2017) and also in (Feldman and Vondrák, 2018) 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. (Chaudhuri et al., 2011; Wu et al., 2017; Dwork and Feldman, 2018)). Hence understanding the generalization properties of uniformly stable algorithms is likely to play an important role in this line of research.

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

Karim T. Abou-Moustafa and Csaba Szepesvári. An exponential tail bound for lq stable learning rules. application to k-folds cross-validation. In *ISAIM*, 2018. URL http://isaim2018.cs.virginia.edu/papers/ISAIM2018_Abou-Moustafa_Szepesvari.pdf.

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059, 2016.

Avrim Blum, Adam Kalai, and John Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *COLT*, pages 203–208, 1999.

Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

Alain Celisse and Benjamin Guedj. Stability revisited: new generalisation bounds for the leave-one-out. *arXiv preprint arXiv:1608.06412*, 2016.

Zachary B. Charles and Dimitris S. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 744–753, 2018. URL http://proceedings.mlr.press/v80/charles18a.html.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

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

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

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

Luc Devroye and Terry J. Wagner. Distribution-free performance bounds with the resubstitution error estimate (corresp.). *IEEE Trans. Information Theory*, 25(2):208–210, 1979b.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. *CoRR*, abs/1803.10266, 2018. URL http://arxiv.org/abs/1803.10266. Extended abstract in COLT 2018.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.

Vitaly Feldman. Generalization of ERM in stochastic convex optimization: The dimension strikes back. *CoRR*, abs/1608.04414, 2016. URL http://arxiv.org/abs/1608.04414. Extended abstract in NIPS 2016.

Vitaly Feldman and Jan Vondrák. Generalization bounds for uniformly stable algorithms. In *Proceedings of NeurIPS*, pages 9770–9780, 2018. URL http://papers.nips.cc/paper/8182-generalization-bounds-for-uniformly-stable-algorithms.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *CoRR*, abs/1902.10710, 2019. URL http://arxiv.org/abs/1902.10710.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, pages 1225–1234, 2016. URL http://jmlr.org/proceedings/papers/v48/hardt16.html.

Satyen Kale, Ravi Kumar, and Sergei Vassilvitskii. Cross-validation and mean-square stability. In *Innovations in Computer Science - ICS*, pages 487–495, 2011. URL http://conference.itcs.tsinghua.edu.cn/ICS2011/content/papers/31.html.

Michael J. Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

Ravi Kumar, Daniel Lokshtanov, Sergei Vassilvitskii, and Andrea Vattani. Near-optimal bounds for cross-validation via loss stability. In *ICML*, pages 27–35, 2013. URL http://jmlr.org/proceedings/papers/v28/kumar13a.html.

Ilja Kuzborskij and Christoph H. Lampert. Data-dependent stability of stochastic gradient descent. In *ICML*, pages 2820–2829, 2018. URL http://proceedings.mlr.press/v80/kuzborskij18a.html.

Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *ICML*, pages 2159–2167, 2017. URL http://proceedings.mlr.press/v70/liu17c.html.

Ben London. A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. In *NIPS*, pages 2935–2944, 2017. URL http://papers.nips.cc/paper/6886-a-pac-bayesian-analysis-of-randomized-learning-with-application-to-stocha

Gábor Lugosi and Miroslaw Pawlak. On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Trans. Information Theory*, 40(2):475–481, 1994.

Andreas Maurer. A second-order look at stability and generalization. In *COLT*, pages 1461–1475, 2017. URL http://proceedings.mlr.press/v65/maurer17a.html.

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

Kobbi Nissim and Uri Stemmer. Concentration bounds for high sensitivity functions through differential privacy. *CoRR*, abs/1703.01970, 2017. URL http://arxiv.org/abs/1703.01970.

Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

Omar Rivasplata, Csaba Szepesvari, John S Shawe-Taylor, Emilio Parrado-Hernandez, and Shiliang Sun. Pac-bayes bounds for stable algorithms with instance-dependent priors. In *Advances in Neural Information Processing Systems*, pages 9234–9244, 2018.

W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3):506–514, 1978. URL http://www.jstor.org/stable/2958555.

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

Thomas Steinke and Jonathan Ullman. Subgaussian tail bounds via stability arguments. *arXiv preprint arXiv:1701.03493*, 2017. URL https://arxiv.org/abs/1701.03493.

Nir Weinberger and Alexander Rakhlin. On high probability bounds for uniformly stable learning algorithms, 2018. Unpublished manuscript.

Rosasco Lorenzo Wibisono, Andre and Tomaso Poggio. Sufficient conditions for uniform stability of regularization algorithms. Technical Report MIT-CSAIL-TR-2009-060, MIT, 2009.

Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *(SIGMOD)*, pages 1307–1322, 2017.

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