Generalization Bounds for Randomized Learning with Application to Stochastic Gradient Descent

Ben London
blondon@amazon.com

1 Introduction

Randomized algorithms are central to modern machine learning. In the presence of massive datasets, researchers often turn to stochastic optimization to solve learning problems. Of particular interest is *stochastic gradient descent* (SGD), a first-order method that approximates the learning objective and gradient by a random point estimate. A classical question in learning theory is, if a randomized learner has access to a finite training sample, will the resulting learned model *generalize* to the data's generating distribution?

This question has been addressed by two notable studies: Elisseeff et al. [5] analyzed bagging methods, and Hardt et al. [6] analyzed SGD. Both analyses crucially rely on the *algorithmic stability* of the learning algorithm, which measures sensitivity to perturbations in the training data. A (randomized) learning algorithm that is stable enjoys good generalization properties. Our work addresses two gaps in the previous studies: Hardt et al.'s results for SGD hold in expectation over draws of the training data and example sequence, whereas one typically prefers results that hold with high probability; Elisseeff et al. derived high-probability generalization bounds, but did not prove the necessary stability conditions for SGD; moreover, their bounds only apply to static distributions on the algorithm, such as a fixed, uniform distribution. We would ideally like high-probability bounds for SGD that support non-uniform, data-dependent sampling.

In this excerpt from a longer, ongoing project, we prove several key results. First, we show that SGD on a smooth, strongly convex objective is *uniformly stable*, in a stronger sense than proven by Hardt et al. We then prove two high-probability generalization bounds for randomized learning algorithms: one that holds for fixed, data-independent product measures (such as a uniform distribution); and another, combining PAC-Bayesian theory with stability analysis, that holds for arbitrary, data-dependent distributions. When the distribution is a product measure, the latter bound holds with high probability over draws of both the data and training randomization. When combined with our stability result, one obtains new, high-probability generalization bounds for SGD. Moreover, the possibilities for data-dependent training randomization are intriguing; our analysis could provide a better theoretical understanding of weighted sampling strategies, such as importance sampling [12, 16, 14, 1] and curriculum learning [2], or lead to new algorithms that balance faster empirical risk minimization with generalization.

2 Preliminaries

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote a closed, compact input space, and let $\mathcal{Y}$ denote a set of labels. For convenience, let $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ denote their Cartesian product. We assume some unknown distribution, $\mathbb{D}$, on $\mathcal{Z}$. Given a sequence of training examples, $S \triangleq (z_1, \ldots, z_n) = ((x_1, y_1), \ldots, (x_n, y_n))$, drawn independently and identically from $\mathbb{D}$, we wish to learn a predictor, $h : \mathcal{X} \to \mathcal{Y}$, from some class of models, $\mathcal{H}$. We assume access to a deterministic learning algorithm, $\mathcal{A} : \mathcal{Z}^n \times \Theta \to \mathcal{H}$, which, given access to S, and an instantiation of some hyperparameters, $\theta \in \Theta$, produces a hypothesis, $h \in \mathcal{H}$.

The work was presented at the NIPS workshop on Optimizing the Optimizers. Barcelona, Spain, 2016.

A randomized learning algorithm can be viewed as a deterministic learning algorithm whose hyperparameters are randomized. For instance, SGD is a learning algorithm that takes as input a training set, S, and a sequence of random indices, sampled with replacement from the set $\{1, \ldots, n\}$. The important thing to note is that, given the sequence of indices, SGD is deterministic. Thus, if $\mathbb{P}$ is a distribution on $\Theta = \{1, \ldots, n\}^T$—i.e., the set of all index sequences of length T—then SGD can be viewed as drawing $\theta \in \Theta$ according to $\mathbb{P}$, then running a deterministic learning algorithm, $\mathcal{A}(S, \theta)$.

For a bounded loss function, $L : \mathcal{H} \times \mathcal{Z} \to [0, M]$, let $L\left(\mathcal{A}(S, \theta), z\right)$ denote the loss of a predictor that was output by $\mathcal{A}(S, \theta)$ when applied to an example z. Ultimately, we are interested in minimizing the expected loss over draws of a random example, $z \sim \mathbb{D}$. For a given instantiation of hyperparameters, $\theta \in \Theta$, we denote this quantity by $R(S, \theta) \triangleq \mathbb{E}_{z \sim \mathbb{D}}\left[L\left(\mathcal{A}(S, \theta), z\right)\right]$. (The learning algorithm should always be clear from context.) The expectation can be approximated by the sample average over the training set; i.e., the *empirical risk*, $\hat{R}(S, \theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} L\left(\mathcal{A}(S, \theta), z_i\right)$. By upper-bounding the difference of the two, $G(S, \theta) \triangleq R(S, \theta) - \hat{R}(S, \theta)$, which we refer to as the *generalization error*, we obtain an upper bound on $R(S, \theta)$.

Since learning is randomized, we also want to bound the expected risk over draws of random hyperparameters. We therefore overload the above notation for a distribution, $\mathbb{P}$, on Θ; let $R(S, \mathbb{P}) \triangleq \mathbb{E}_{\theta \sim \mathbb{P}}\left[R(S, \theta)\right]$, $\hat{R}(S, \mathbb{P}) \triangleq \mathbb{E}_{\theta \sim \mathbb{P}}[\hat{R}(S, \theta)]$, and $G(S, \mathbb{P}) \triangleq R(S, \mathbb{P}) - \hat{R}(S, \mathbb{P})$. Note that, by linearity of expectation, the outer expectation over $\theta \sim \mathbb{P}$ can be pushed inside the inner expectation over $z \sim \mathbb{D}$ in the risk, or the summation over i in the empirical risk.

Relationship to PAC-Bayes Conditioned on the training set, a distribution on the hyperparameter space, Θ, induces a distribution on the hypothesis space, $\mathcal{H}$. One could ignore the learning algorithm altogether and just consider a distribution on $\mathcal{H}$ directly, as is done in PAC-Bayesian analysis. In the PAC-Bayes learning framework, we fix a prior distribution, $\mathbb{P}$, on $\mathcal{H}$, then learn a posterior distribution, $\mathbb{Q}$, conditioned on the training data. The complexity of learning is captured by the KL divergence between $\mathbb{Q}$ and $\mathbb{P}$. PAC-Bayes analyzes the *Gibbs loss*, $\mathbb{E}_{h \sim \mathbb{Q}}\left[L\left(h, z\right)\right]$, which is the expected loss of a random hypothesis, $h \in \mathcal{H}$, drawn from $\mathbb{Q}$. This quantity is similar to the expected loss of a randomized algorithm, $\mathbb{E}_{\theta \sim \mathbb{Q}}\left[L\left(\mathcal{A}(S, \theta), z\right)\right]$. The subtle distinction between the two losses is that the latter is explicitly a function of the training data, S, whereas the former quantity may only implicitly depend on S—for instance, if the distribution on $\mathcal{H}$ is induced by a randomized learning algorithm. The advantage of making the learning algorithm explicit is that it isolates the source of randomness, which may help in analyzing the distribution of learned hypotheses. Indeed, it may be difficult to map the output of a randomized learning algorithm to a distribution on the hypothesis space. The disadvantage of making the learning algorithm explicit is that, due to the loss' dependence on the training data, the generalization error could be sensitive to certain examples. This condition is studied in *stability* analysis, which we discuss in the next section.

3 Stability

Informally, stability measures the amount of change in the output of a function when the input is perturbed; a function is stable if the change is proportional to the perturbation. A learning algorithm is stable if the loss on its output changes proportionally to perturbations of its inputs. In other words, a learning algorithm should not be overly sensitive to any single input. Stability is crucial for generalization [11], and has also been linked to differentially private learning [13, 15]. In this section, we discuss several notions of stability tailored for randomized learning algorithms. From this point on, let $D_{\text{H}}(\mathbf{v}, \mathbf{v}') \triangleq \sum_{i=1}^{|\mathbf{v}|} \mathbb{1}\{v_i = v_i'\}$ denote the Hamming distance.

The learning-theoretic literature traditionally measures stability with respect to perturbations of the training data. The following definition, attributed to Elisseeff et al. [5], is a modified version of Bousquet and Elisseeff's [2002] *uniform stability*, designed to accommodate randomized algorithms.

Definition 1 (Data Stability)**.** A randomized learning algorithm, $\mathcal{A}$, is $\beta_{\mathcal{Z}}$-uniformly stable with respect to a loss function, L, and a distribution, $\mathbb{P}$, on Θ if, for any two datasets, $S, S' \in \mathcal{Z}^n$: $D_{\text{H}}(S, S') = 1$, which differ at exactly one example,

$$\sup_{z \in \mathcal{Z}} \left| \mathop{\mathbb{E}}_{\theta \sim \mathbb{P}}\left[L\left(\mathcal{A}(S, \theta), z\right) - L\left(\mathcal{A}(S', \theta), z\right)\right] \right| \leq \beta_{\mathcal{Z}}. \tag{1}$$

Note that Equation 1 holds in expectation over draws of $\theta \sim \mathbb{P}$, which is a weaker requirement than uniformity over Θ. For meaningful generalization rates, we will want $\beta_{\mathcal{Z}}$ to be of order $\mathrm{O}(1/n)$.

We also consider stability with respect to changes in the hyperparameters. We will assume that the hyperparameter space, Θ, decomposes into the product of T subspaces, $\prod_{t=1}^{T} \Theta_t$. For example, Θ could be the set of all sequences of example indices, such as one would sample in SGD.

Definition 2 (Hyperparameter Stability)**.** A randomized learning algorithm, $\mathcal{A}$, is β_Θ-uniformly stable with respect to a loss function, L if, for any dataset, $S \in \mathcal{Z}^n$, and any two hyperparameter instantiations, $\theta, \theta' \in \Theta : D_{\mathrm{H}}(\theta, \theta') = 1$, that differ at a single coordinate,

$$\sup_{z \in \mathcal{Z}} \left| L\left(\mathcal{A}(S, \theta), z\right) - L\left(\mathcal{A}(S, \theta'), z\right) \right| \leq \beta_\Theta. \tag{2}$$

When $\mathcal{A}$ is both $\beta_{\mathcal{Z}}$-uniformly and β_Θ-uniformly stable, we say that $\mathcal{A}$ is $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniformly stable. In addition to requiring $\beta_{\mathcal{Z}} = \mathrm{O}(1/n)$, we will also need β_Θ to be of order $\mathrm{O}(1/\sqrt{nT})$. If $T \geq n$ (as it often is), then $\beta_\Theta = \mathrm{O}(1/T)$ suffices. We review one such case in the next section.

3.1 When Stability is Satisfied

Given the above definitions of uniform stability, the natural question is: when are they satisfied? We will focus on weight vectors, $\mathcal{H} \subseteq \{\mathbf{w} \in \mathbb{R}^d : d \geq 1\}$, learned with SGD. SGD involves a sequence of parameter updates of the form $U_t(\mathbf{w}, z) \triangleq \mathbf{w} - \eta_t \nabla L(\mathbf{w}, z)$, where η_t is a step size for iterate t. (When not needed, we drop the iterate subscript.) In the t^{th} round of SGD, the learned weights, $\mathbf{w}_t$, are defined recursively as $\mathbf{w}_t = U_t(\mathbf{w}_{t-1}, z_t)$.

Hardt et al. [6] analyzed the $\beta_{\mathcal{Z}}$-uniform stability of SGD with uniform sampling (with replacement) for various types of loss functions. We defer to their work for $\beta_{\mathcal{Z}}$-uniform stability and instead focus on the stronger condition, $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniform stability. Proving $\beta_\Theta = \mathrm{O}(1/\sqrt{nT})$ (or $\beta_\Theta = \mathrm{O}(1/T)$) is a challenge that Hardt et al. did not address. Elisseeff et al. [5] proved $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniform stability for certain bagging algorithms, but did not consider SGD. As a first step, we consider loss functions that are Lipschitz, smooth and strongly convex (see Appendix A.1 for definitions). Clearly, we are also interested in non-(strongly)-convex losses, but proving $\mathrm{O}(1/T)$-uniform stability for these losses is difficult; we hope to address this in future work. In the following theorem, we prove $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniform stability for SGD with uniform sampling and "staircase" decaying step sizes.

Theorem 1. *Assume the loss function, L, is γ-strongly convex, λ-Lipschitz and σ-smooth. Suppose SGD with uniform sampling is run for T iterations with step sizes $\eta_t \triangleq 1/(\gamma t + \sigma)$. Then, SGD is $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniformly stable with $\beta_{\mathcal{Z}} \leq \frac{2\lambda^2}{\gamma n}$ and $\beta_\Theta \leq \frac{2\lambda^2}{\gamma T}$.*

The proof is given in Appendix A.1. Since $\beta_\Theta = \mathrm{O}(1/T)$, then for $T \geq n$, the stability is dominated by $\mathrm{O}(1/\sqrt{nT})$, which is sufficient for good generalization.

4 Risk Bounds

In this section, we present several new theoretical results concerning the generalization behavior of randomized learning algorithms. While previous work [5] has addressed this topic, to our knowledge, ours is the first analysis of stochastic gradient methods that yields PAC bounds; that is, risk bounds that hold with high probability over draws of a finite training dataset.

Our work is closely related to that of Hardt et al. [6] and Lin et al. [8], who used stability analysis to derive bounds for *generalization in expectation*; i.e., upper bounds on $\mathbb{E}_{S \sim \mathbb{D}^n}\left[G(S, \mathbb{P})\right]$. While such results are useful for gaining insight into generalization behavior, PAC bounds are usually favored. Hardt et al. posited that PAC bounds would require a more sophisticated analysis, and perhaps a stronger condition than $\beta_{\mathcal{Z}}$-uniform stability (Definition 1). In actuality, $\beta_{\mathcal{Z}}$-uniform stability is sufficient for PAC learning in expectation over draws of hyperparameters, $\theta \sim \mathbb{P}$ (as shown in Appendix A.3). However, to prove a risk bound that holds with high probability over draws of both $S \sim \mathbb{D}^n$ and $\theta \sim \mathbb{P}$ indeed requires a stronger condition, $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniform stability (Definition 2).

We begin with such a risk bound for any fixed product measure on Θ. We then present a novel analysis of randomized learning from the PAC-Bayesian perspective. The result is a generalization

bound that holds, with high probability over $S \sim \mathbb{D}^n$, for all posteriors, $\mathbb{Q}$, on Θ, provided the learning algorithm is $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniformly stable with respect to a fixed prior, $\mathbb{P}$. This latter result could have interesting implications for data-dependent sampling strategies for SGD.

4.1 Generalization via Stability

Our first goal is to prove a risk bound for randomized algorithms, such as SGD, that holds with high probability over draws of both $S \sim \mathbb{D}^n$ and $\theta \sim \mathbb{P}$. This is a stronger proposition than generalization in expectation, and as such requires stronger assumptions: namely, that $\mathbb{P}$ is a product measure, and $\mathcal{A}$ has $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniform stability. These conditions let us upper-bound the moment-generating function of $G(S, \theta) - \mathbb{E}[G(S, \theta)]$ (proven in Appendix A.4), which enables the following theorem (which is a refinement of [5, Theorem 15]). Our proof is given in Appendix A.5.

Theorem 2. *Suppose $\mathcal{A}$ is a $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniformly stable learning algorithm with respect to a loss function, L, and a fixed product measure, $\mathbb{P}$, on Θ. Then, for any $n \geq 1$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of a dataset, $S \sim \mathbb{D}^n$, and hyperparameters, $\theta \sim \mathbb{P}$,*

$$R(S, \theta) \leq \hat{R}(S, \theta) + \beta_{\mathcal{Z}} + \sqrt{\frac{((M + 2n\beta_{\mathcal{Z}})^2 + 4nT\beta_\Theta^2) \ln \frac{1}{\delta}}{2n}}. \tag{3}$$

When $T \geq n$, combining Theorem 2 with Theorem 1 yields a generalization bound that is asymptotically dominated by $\mathrm{O}(1/\sqrt{n})$. (We leave this simple variable substitution to the reader.) Further, since Theorem 1 guarantees $\beta_\Theta = \mathrm{O}(1/T)$, increasing T, the number of iterations of SGD, actually *reduces* the bound. This suggests that training for longer periods does not lead to overfitting, provided the loss function, sampling distribution and step sizes satisfy the conditions of Theorem 1.

4.2 A PAC-Bayesian View

In the previous subsection, we assumed that hyperparameters were sampled according to a fixed distribution on Θ; for instance, SGD with uniformly random sampling from S. While this distribution may be sufficient (or even optimal) for certain situations, it may sometimes be advantageous to sample according to a data-dependent distribution. Unfortunately, the previous analysis does not accommodate data-dependent hyperparameter distributions, and would therefore not accommodate, e.g., SGD with data-dependent sampling. This shortcoming motivates the following PAC-Bayesian view of randomized learning. PAC-Bayes risk bounds must hold (with high probability) for all posteriors simultaneously, including those that depend on the data. In our extension for randomized learning algorithms, we fix a prior on Θ—which could be uniform—then determine a posterior on Θ given the training set. Our PAC-Bayes bound holds for all posteriors, including those that depend on S, but penalizes those that diverge significantly the prior.

Theorem 3. *Suppose $\mathcal{A}$ is a $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniformly stable learning algorithm with respect to a loss function, L, and a fixed product measure, $\mathbb{P}$, on Θ. Then, for any $n \geq 1$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$ over draws of a dataset, $S \sim \mathbb{D}^n$, every posterior distribution, $\mathbb{Q}$, on Θ satisfies*

$$R(S, \mathbb{Q}) \leq \hat{R}(S, \mathbb{Q}) + \beta_{\mathcal{Z}} + \sqrt{\frac{2\left((M + 2n\beta_{\mathcal{Z}})^2 + 4nT\beta_\Theta^2\right)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{2}{\delta}\right)}{n}}. \tag{4}$$

The proof is given in Appendix A.6. Note that the stability requirements only need to be satisfied by the fixed prior, such as a uniform distribution. This simple prior can have $(\mathrm{O}(1/n), \mathrm{O}(1/T))$-uniform stability, as indicated by Theorem 1. The penalty for letting the posterior stray from the prior is captured by the KL divergence term. When the KL term is sub-logarithmic in n, we achieve a $\tilde{\mathrm{O}}(1/\sqrt{n})$ generalization rate. When the posterior is a product measure, we can adapt Equation 4 to hold with high probability over both $S \sim \mathbb{D}^n$ and $\theta \sim \mathbb{P}$, as shown in Appendix A.7.

5 Discussion

We have presented several new theoretical results regarding the stability and generalization error of learning with SGD. Most interestingly, our PAC-Bayes bound provides a means of analyzing non-uniform, data-dependent sampling strategies, which could have repercussions for importance sampling or curriculum learning; or, possibly a new SGD variant that explicitly minimizes both the empirical risk and KL divergence terms. We plan to investigate these ideas in forthcoming work.

References

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

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009.

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

[4] M. Donsker and S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

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

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

[7] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

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

[9] B. London, B. Huang, and L. Getoor. Stability and generalization in structured prediction. *Journal of Machine Learning Research*, 17, 2016.

[10] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141:148–188, 1989.

[11] 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. *Advances in Computational Mathematics*, 25:161–193, 2006.

[12] D. Needell, N. Srebro, and R. Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Neural Information Processing Systems*, 2014.

[13] A. Thakurta and A. Smith. Differentially private model selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, 2013.

[14] D. Vainsencher, H. Liu, and T. Zhang. Local smoothness in variance reduced optimization. In *Neural Information Processing Systems*, 2015.

[15] Y. Wang, J. Lei, and S. Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle. *CoRR*, abs/1502.06309, 2015.

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

A Supplemental Material

The following appendices are provided to supplement the paper.

A.1 Proof of Theorem 1

The following definitions, borrowed from Hardt et al., are used to characterize loss functions and update rules.

Definition 3 (Lipschitzness)**.** A loss function, L, is λ-Lipschitz if

$$\sup_{\mathbf{w},\mathbf{w}' \in \mathcal{H}} \sup_{z \in \mathcal{Z}} \frac{|L(\mathbf{w}, z) - L(\mathbf{w}', z)|}{\|\mathbf{w} - \mathbf{w}'\|_2} \leq \lambda. \tag{5}$$

Definition 4 (Smoothness)**.** A loss function, L, is σ-smooth if

$$\sup_{\mathbf{w},\mathbf{w}' \in \mathcal{H}} \sup_{z \in \mathcal{Z}} \frac{\|\nabla L(\mathbf{w}, z) - \nabla L(\mathbf{w}', z)\|_2}{\|\mathbf{w} - \mathbf{w}'\|_2} \leq \sigma. \tag{6}$$

Definition 5 (Expansivity)**.** An update, U, is α-expansive if

$$\sup_{\mathbf{w},\mathbf{w}' \in \mathcal{H}} \sup_{z \in \mathcal{Z}} \frac{\|U(\mathbf{w}, z) - U(\mathbf{w}', z)\|_2}{\|\mathbf{w} - \mathbf{w}'\|_2} \leq \alpha. \tag{7}$$

We say that U is contractive if $\alpha \leq 1$.

Definitions 3 to 5 are essentially variants of Lipschitz continuity: the loss is σ-smooth if its gradient is σ-Lipschitz; likewise, an update is α-expansive if it is α-Lipschitz.

We now prove a fundamental technical lemma that is central to our proof of Theorem 1.

Lemma 1. *Assume that the loss function, L, is λ-Lipschitz. Further, assume that each SGD update, U_t, is α_t-expansive. If SGD is run for T iterations on two sequences of examples that differ at a single step, k, then the resulting learned hypotheses, $\mathbf{w}_T$ and $\mathbf{w}'_T$ satisfy*

$$\|\mathbf{w}_T - \mathbf{w}'_T\|_2 \leq 2\lambda\eta_k \prod_{t=k+1}^{T} \alpha_t. \tag{8}$$

Proof. For the first $k-1$ iterations of SGD, the example sequences are the same; therefore, so are the learned weights. On processing the k^{th} example, the weights may diverge, but we will show that the divergence is bounded, due to the Lipschitz property. For every iteration after k, the weights may continue to follow different trajectories, but the expansivity property lets us bound the difference of the final, learned weights.

Starting at T and recursing backward, we have that

$$\|\mathbf{w}_T - \mathbf{w}'_T\|_2 \leq \left\|\mathbf{w}_{T-1} - \mathbf{w}'_{T-1}\right\|_2 \alpha_T \leq \ldots \leq \|\mathbf{w}_k - \mathbf{w}'_k\|_2 \prod_{t=k+1}^{T} \alpha_t. \tag{9}$$

Then, expanding the k^{th} update,

$$\begin{aligned} \|\mathbf{w}_k - \mathbf{w}'_k\|_2 &= \left\|\mathbf{w}_{k-1} - \eta_k \nabla L(\mathbf{w}_{k-1}, z_k) - \mathbf{w}'_{k-1} + \eta_k \nabla L(\mathbf{w}_{k-1}, z'_k)\right\|_2 \\ &\leq \|\eta_k \nabla L(\mathbf{w}_{k-1}, z_k)\|_2 + \|\eta_k \nabla L(\mathbf{w}_{k-1}, z'_k)\|_2 \\ &\leq 2\eta_k \lambda. \end{aligned} \tag{10}$$

Combining Equations 9 and 10 completes the proof. □

We can now prove Theorem 1. First, note that $\eta_t \leq 1/\sigma$ for all $t = 1, \ldots, T$. As noted by Hardt et al. [6, proof of Theorem 3.9], due to the strong convexity of the loss function, this step size guarantees

that each update is contractive with coefficient $1 - \eta_t\gamma = 1 - \frac{1}{t+\sigma/\gamma}$. Moreover,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[\|\mathbf{w}_T - \mathbf{w}'_T\|_2\right] &\leq \sum_{k=1}^{T}\left(\prod_{t=k+1}^{T}(1-\eta_t\gamma)\right)\eta_k \cdot \frac{2\lambda}{n} \\
&= \sum_{k=1}^{T}\left(\prod_{t=k+1}^{T}\left(1-\frac{1}{t+\sigma/\gamma}\right)\right)\frac{1}{k+\sigma/\gamma}\cdot\frac{2\lambda}{\gamma n} \\
&= \sum_{k=1}^{T}\frac{k+\sigma/\gamma}{T}\cdot\frac{1}{k+\sigma/\gamma}\cdot\frac{2\lambda}{\gamma n} = \frac{2\lambda}{\gamma n}. \qquad (11)
\end{aligned}$$

Combining Equation 11 with the Lipschitz property (Equation 5),

$$\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[L(\mathbf{w}_T, z) - L(\mathbf{w}'_T, z)\right] \leq \lambda \mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[\|\mathbf{w}_T - \mathbf{w}'_T\|_2\right], \qquad (12)$$

we obtain an upper bound on the data stability coefficient, $\beta_{\mathcal{Z}} \leq \frac{2\lambda^2}{\gamma n}$.

Now, suppose the example sequence is perturbed at any index k. Via Lemma 1, we have that

$$\begin{aligned}
\|\mathbf{w}_T - \mathbf{w}'_T\|_2 &\leq 2\lambda\eta_k \prod_{t=k+1}^{T}(1-\eta_t\gamma) \\
&= \frac{2\lambda}{\gamma}\cdot\frac{1}{k+\sigma/\gamma}\prod_{t=k+1}^{T}\left(1-\frac{1}{t+\sigma/\gamma}\right) \\
&= \frac{2\lambda}{\gamma}\cdot\frac{1}{k+\sigma/\gamma}\cdot\frac{k+\sigma/\gamma}{T} = \frac{2\lambda}{\gamma T}, \qquad (13)
\end{aligned}$$

which we combine with Equation 5 to obtain $\beta_\Theta \leq \frac{2\lambda^2}{\gamma T}$.

A.2 Technical Lemmas: Stability of the Generalization Error

To prove our generalization bounds, we will require the following technical lemmas, which connect stability with respect to the loss function to stability with respect to the generalization error.

Lemma 2. *If $\mathcal{A}$ is $\beta_{\mathcal{Z}}$-uniformly stable with respect to L and $\mathbb{P}$, then, for any $S, S' \in \mathcal{Z}^n$: $D_{\text{H}}(S, S') = 1$,*

$$|G(S,\mathbb{P}) - G(S',\mathbb{P})| \leq 2\beta_{\mathcal{Z}} + \frac{M}{n}. \qquad (14)$$

Proof. Observe that the difference of generalization errors decomposes as

$$\begin{aligned}
|G(S,\mathbb{P}) - G(S',\mathbb{P})| &= \left|R(S,\mathbb{P}) - \hat{R}(S,\mathbb{P}) - R(S',\mathbb{P}) + \hat{R}(S',\mathbb{P})\right| \\
&\leq |R(S,\mathbb{P}) - R(S',\mathbb{P})| + \left|\hat{R}(S',\mathbb{P}) - \hat{R}(S,\mathbb{P})\right|, \qquad (15)
\end{aligned}$$

following from the triangle inequality. We will upper-bound the righthand terms separately. First, using linearity of expectation and the $\beta_{\mathcal{Z}}$-uniform stability of $\mathcal{A}$, we have that

$$\begin{aligned}
|R(S,\mathbb{P}) - R(S',\mathbb{P})| &= \left|\mathop{\mathbb{E}}_{z\sim\mathbb{D}}\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[L\left(\mathcal{A}(S,\theta),z\right) - L\left(\mathcal{A}(S',\theta),z\right)\right]\right| \\
&\leq \mathop{\mathbb{E}}_{z\sim\mathbb{D}}\left|\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[L\left(\mathcal{A}(S,\theta),z\right) - L\left(\mathcal{A}(S',\theta),z\right)\right]\right| \\
&\leq \mathop{\mathbb{E}}_{z\sim\mathbb{D}}\beta_{\mathcal{Z}} = \beta_{\mathcal{Z}}. \qquad (16)
\end{aligned}$$

Then, without loss of generality, assume that S differs from S' at the i^{th} example, denoted z_i'. Using the triangle inequality and $\beta_{\mathcal{Z}}$-uniform stability,

$$\begin{aligned}\left|\hat{R}(S,\mathbb{P}) - \hat{R}(S',\mathbb{P})\right| &\leq \frac{1}{n}\sum_{j\neq i}\left|\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}[L(\mathcal{A}(S,\theta),z_j) - L(\mathcal{A}(S',\theta),z_j)]\right| \\ &+ \left|\frac{1}{n}\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}[L(\mathcal{A}(S,\theta),z_i) - L(\mathcal{A}(S',\theta),z_i')]\right| \\ &\leq \frac{1}{n}\sum_{j\neq i}\beta_{\mathcal{Z}} + \frac{M}{n} = \beta_{\mathcal{Z}} + \frac{M}{n}. \end{aligned} \tag{17}$$

Substituting Equations 16 and 17 into Equation 15 completes the proof. □

Lemma 3. *If $\mathcal{A}$ is β_Θ-uniformly stable with respect to L, then, for any $S \in \mathcal{Z}^n$ and $\theta, \theta' \in \Theta$: $D_{\text{H}}(\theta,\theta') = 1$,*

$$|G(S,\theta) - G(S,\theta')| \leq 2\beta_\Theta. \tag{18}$$

Proof. The proof is almost identical to that of Lemma 2. First, we decompose the generalization error:

$$|G(S,\theta) - G(S,\theta')| \leq |R(S,\theta) - R(S,\theta')| + \left|\hat{R}(S,\theta) - \hat{R}(S,\theta')\right|. \tag{19}$$

Then, we upper-bound the difference of risk terms:

$$|R(S,\theta) - R(S,\theta')| \leq \mathop{\mathbb{E}}_{z\sim\mathbb{D}}[|L(\mathcal{A}(S,\theta),z) - L(\mathcal{A}(S,\theta'),z)|] \leq \beta_\Theta. \tag{20}$$

Then, we upper-bound the difference of empirical risk terms:

$$\left|\hat{R}(S,\theta) - \hat{R}(S,\theta')\right| \leq \frac{1}{n}\sum_{i=1}^{n}|L(\mathcal{A}(S,\theta),z_i) - L(\mathcal{A}(S,\theta'),z_i)| \leq \beta_\Theta. \tag{21}$$

Combining Equations 19 to 21 completes the proof. □

A.3 A Risk Bound for $\beta_{\mathcal{Z}}$-Uniform Stability

As a warm-up to Theorem 2, we will show that $\beta_{\mathcal{Z}}$-uniform stability (Definition 1) is indeed a sufficient condition for PAC-learning with randomized algorithms, such as SGD. The following theorem holds with high probability over draws of a training dataset, $S \sim \mathbb{D}^n$, and in expectation over draws of hyperparameters, $\theta \sim \mathbb{P}$, according to a fixed distribution, $\mathbb{P}$.

Theorem 4. *Suppose $\mathcal{A}$ is a $\beta_{\mathcal{Z}}$-uniformly stable learning algorithm with respect to a loss function, L, and a fixed distribution, $\mathbb{P}$, on Θ. Then, for any $n \geq 1$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of a dataset, $S \sim \mathbb{D}^n$,*

$$R(S,\mathbb{P}) \leq \hat{R}(S,\mathbb{P}) + \beta_{\mathcal{Z}} + (M + 2n\beta_{\mathcal{Z}})\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}. \tag{22}$$

By combining Theorem 4 with a $\mathrm{O}(1/n)$-uniform stability bound (such as those by Hardt et al. [6], or Theorem 1), we obtain an upper bound on the generalization error that decays at a rate of $\mathrm{O}(1/\sqrt{n})$.

Proof. Our proof uses a simple adaptation of the canonical technique pioneered by Bousquet and Elisseeff [3]. First, we review a cornerstone of stability-based generalization analysis, commonly known as *McDiarmid's inequality* [10]. The following is a specialized version of the general theorem. Suppose $\varphi : \mathcal{Z}^n \to \mathbb{R}$ is a function for which there exists a constant, $\beta > 0$, such that

$$\forall S, S' \in \mathcal{Z}^n : D_{\text{H}}(S,S') = 1,\ \ |\varphi(S) - \varphi(S')| \leq \beta. \tag{23}$$

Then, for any $\epsilon > 0$,

$$\Pr_{S\sim\mathcal{Z}^n}\{\varphi(S) - \mathbb{E}\,\varphi(S) \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{n\beta^2}\right). \tag{24}$$

An important special case is when $\beta = \mathrm{O}(1/n)$, in which case the righthand side of Equation 24 becomes $\mathrm{O}\left(\exp(-2n\epsilon^2)\right)$, which decays rapidly.

By Lemma 2, $G(\,\cdot\,, \mathbb{P})$ satisfies Equation 23 with $\beta = 2\beta_{\mathcal{Z}} + M/n$. We therefore have that

$$\Pr_{S\sim\mathbb{D}^n}\left\{G(S,\mathbb{P}) - \mathbb{E}\left[G(S,\mathbb{P})\right] \geq (M + 2n\beta_{\mathcal{Z}})\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right\}$$
$$\leq \exp\left(\frac{-2\left((M + 2n\beta_{\mathcal{Z}})\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right)^2}{n\left(2\beta_{\mathcal{Z}} + M/n\right)^2}\right) = \delta.$$

Thus, with probability at least $1-\delta$ over draws of $S \sim \mathbb{D}^n$,

$$G(S,\mathbb{P}) \leq \mathop{\mathbb{E}}_{S\sim\mathbb{D}^n}\left[G(S,\mathbb{P})\right] + (M + 2n\beta_{\mathcal{Z}})\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}. \tag{25}$$

Let us pause at this point to recognize that the generalization error, $G(S,\mathbb{P})$, is not a zero-mean random variable. This is because the learning algorithm—hence, the loss composed with the training algorithm—is a function of the entire training set and does not decompose over individual examples. (In contrast, the generalization error of a given hypothesis has mean zero.) Therefore, to finish the proof, we must upper-bound the expected generalization error, $\mathbb{E}_{S\sim\mathbb{D}^n}\left[G(S,\mathbb{P})\right]$. To do so, we use linearity of expectation, and the fact that each example is i.i.d.:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathbb{D}^n}\left[G(S,\mathbb{P})\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{S\sim\mathbb{D}^n}\mathop{\mathbb{E}}_{z_i'\sim\mathbb{D}}\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[L\left(\mathcal{A}(S,\theta), z_i'\right) - L\left(\mathcal{A}(S,\theta), z_i\right)\right] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{S\sim\mathbb{D}^n}\mathop{\mathbb{E}}_{z_i'\sim\mathbb{D}}\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[L\left(\mathcal{A}(S',\theta), z_i'\right) - L\left(\mathcal{A}(S,\theta), z_i\right)\right] + \beta_{\mathcal{Z}} \\
&= \beta_{\mathcal{Z}}.
\end{aligned} \tag{26}$$

In the first inequality, we formed a new dataset, S', by replacing z_i with z_i'; the difference of losses, $\mathbb{E}_{\theta\sim\mathbb{P}}\left[L\left(\mathcal{A}(S,\theta), z_i'\right) - L\left(\mathcal{A}(S',\theta), z_i'\right)\right]$, is upper-bounded by $\beta_{\mathcal{Z}}$, via Definition 1. The last line follows from symmetry; since S and S' are both distributed according to $\mathbb{D}^n$, and θ is independent of S and S', the expected losses cancel out. Combining Equations 25 and 26 completes the proof. □

A.4 An Upper-bound on the Moment-generating Function

The proofs of Theorems 2 and 3 require an upper-bound on the moment-generating function of the random variable $G(S,\theta) - \mathbb{E}\left[G(S,\theta)\right]$.

Lemma 4. *Fix a product measure, $\mathbb{P}(\theta) = \prod_{t=1}^{T}\mathbb{P}_t(\theta_t)$, and suppose $\mathcal{A}$ is a $(\beta_{\mathcal{Z}}, \beta_{\Theta})$-uniformly stable with respect to L and $\mathbb{P}$. Let $\bar{\beta}_{\mathcal{Z}} \triangleq 2\beta_{\mathcal{Z}} + M/n$, and recall, from Lemmas 2 and 3, that $\mathcal{A}$ is therefore $(\bar{\beta}_{\mathcal{Z}}, 2\beta_{\Theta})$-uniformly stable with respect to G and $\mathbb{P}$. Then, with*

$$\Phi(S,\theta) \triangleq G(S,\theta) - \mathbb{E}\left[G(S,\theta)\right], \tag{27}$$

for any $\epsilon > 0$,

$$\mathop{\mathbb{E}}_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{P}}}\left[\exp\left(\epsilon\,\Phi(S,\theta)\right)\right] \leq \exp\left(\frac{\epsilon^2}{8}\left(n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_{\Theta}^2\right)\right). \tag{28}$$

Proof. To reduce notation, we omit the subscript notation from expectations. Further, let $z_{i:j} \triangleq z_i, \ldots, z_j$ and $\theta_{i:j} \triangleq \theta_i, \ldots, \theta_j$. (Interpret $z_{1:0}$ and $\theta_{1:0}$ as the empty set.) We start by constructing a Doob martingale as follows:

$$V_i \triangleq \begin{cases}
\mathbb{E}[G(S,\theta) \mid z_1] - \mathbb{E}[G(S,\theta)] & \text{for } i = 1 \\
\mathbb{E}[G(S,\theta) \mid z_{1:i}] - \mathbb{E}[G(S,\theta) \mid z_{1:i-1}] & \text{for } i \in \{2,\ldots,n\} \\
\mathbb{E}[G(S,\theta) \mid S, \theta_{1:i}] - \mathbb{E}[G(S,\theta) \mid S, \theta_{1:i-1}] & \text{for } i \in \{n+1,\ldots,n+T\}
\end{cases}$$

Observe that $\mathbb{E}[V_i] = 0$ and $\sum_{i=1}^{n+T} V_i = \Phi(S, \theta)$. Thus, using the *law of total expectation* (alternatively, *iterated expectations*),

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\epsilon\,\Phi(S,\theta)\right)\right] &= \mathbb{E}\left[\exp\left(\epsilon \sum_{i=1}^{n+T} V_i\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n+T} e^{\epsilon V_i}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{n+T-1} e^{\epsilon V_i}\ \mathbb{E}\left[e^{\epsilon V_{n+T}} \mid S, \theta_{1:T-1}\right]\right] \\
&\leq \mathbb{E}\left[\prod_{i=1}^{n+T-1} e^{\epsilon V_i} \sup_{S,\theta_{1:T-1}} \mathbb{E}\left[e^{\epsilon V_{n+T}} \mid S, \theta_{1:T-1}\right]\right].
\end{aligned}$$

By iteratively applying this upper bound, we obtain

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\epsilon\,\Phi(S,\theta)\right)\right] \leq\ & \left(\prod_{i=1}^{n} \sup_{z_{1:i-1}} \mathbb{E}\left[e^{\epsilon V_i} \mid z_{1:i-1}\right]\right) \\
& \times \left(\prod_{j=1}^{T} \sup_{S,\theta_{1:j-1}} \mathbb{E}\left[e^{\epsilon V_{n+j}} \mid S, \theta_{1:j-1}\right]\right).
\end{aligned}$$

If each V_i is bounded over all $z_{1:i-1}$, and each V_j is bounded over all S and $\theta_{1:j-1}$, then *Hoeffding's lemma* [7] can be used to upper-bound their respective moment-generating functions. Hoeffding's lemma states that, if X is a zero-mean random variable, such that $a \leq X \leq b$ almost surely, then, for all $\epsilon \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\epsilon X}\right] \leq \exp\left(\frac{\epsilon^2 (b-a)^2}{8}\right).$$

We therefore need to show that:

$$\begin{aligned}
&\forall i \in 1, \ldots, n,\ \exists c_i : \sup V_i - \inf V_i \\
&\qquad = \sup_{z_{1:i}, z_i'} \mathbb{E}[G(S,\theta) \mid z_{1:i}] - \mathbb{E}[G(S',\theta) \mid z_{1:i-1}, z_i'] \leq c_i; \qquad (29) \\
&\forall j \in 1, \ldots, T,\ \exists c_j : \sup V_{n+j} - \inf V_{n+j} \\
&\qquad = \sup_{S, \theta_{1:j}, \theta_j'} \mathbb{E}[G(S,\theta) \mid S, \theta_{1:j}] - \mathbb{E}[G(S,\theta') \mid S, \theta_{1:j-1}, \theta_j'] \leq c_j, \qquad (30)
\end{aligned}$$

To prove Equation 29, we use the fact that $\mathcal{A}$ is $\bar{\beta}_{\mathcal{Z}}$-uniformly stable with respect to G, as well as the mutual independence between examples and hyperparameters:

$$\begin{aligned}
&\sup_{z_{1:i}, z_i'} \mathbb{E}[G(S,\theta) \mid z_{1:i}] - \mathbb{E}[G(S',\theta) \mid z_{1:i-1}, z_i'] \\
&\qquad = \sup_{z_{1:i}, z_i'} \sum_{z_{i+1:n}} \mathop{\mathbb{E}}_{\theta \sim \mathbb{P}} \left[G(S,\theta) - G(S',\theta)\right] \mathbb{D}(z_{i+1:n}) \\
&\qquad \leq \sum_{z_{i+1:n}} \bar{\beta}_{\mathcal{Z}}\, \mathbb{D}(z_{i+1:n}) = \bar{\beta}_{\mathcal{Z}}.
\end{aligned}$$

(For notational simplicity, the expectation over $z_{i+1:n}$ is denoted by a summation.) Similarly, to prove Equation 30, we use β_Θ-uniform stability and independence between hyperparameters:

$$\begin{aligned}
&\sup_{S, \theta_{1:j}, \theta_j'} \mathbb{E}[G(S,\theta) \mid S, \theta_{1:j}] - \mathbb{E}[G(S,\theta') \mid S, \theta_{1:j-1}, \theta_j'] \\
&\qquad = \sup_{S, \theta_{1:j}, \theta_j'} \sum_{\theta_{j+1:T}} \left(G(S,\theta) - G(S,\theta')\right) \mathbb{P}(\theta_{j+1:T}) \\
&\qquad \leq \sum_{\theta_{j+1:T}} 2\beta_\Theta\, \mathbb{P}(\theta_{j+1:T}) = 2\beta_\Theta.
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\exp\left(\epsilon\,\Phi(S,\theta)\right)\right] \leq \left(\prod_{i=1}^{n}\exp\left(\frac{\epsilon^2\bar{\beta}_{\mathcal{Z}}^2}{8}\right)\right)\left(\prod_{j=1}^{T}\exp\left(\frac{\epsilon^2(2\beta_\Theta)^2}{8}\right)\right)$$
$$= \exp\left(\frac{\epsilon^2}{8}\left(n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_\Theta^2\right)\right),$$

which establishes Equation 28. □

A.5 Proof of Theorem 2

The proof proceeds similarly to that of Theorem 4, but we first need to show that $G(S,\theta)$ concentrates around its mean. Define $\Phi(S,\theta)$ as in Equation 27 and observe that, for any $t > 0$ and $\epsilon > 0$,

$$\Pr_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{P}}}\left\{\Phi(S,\theta)\geq t\right\} = \Pr_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{P}}}\left\{e^{\epsilon\Phi(S,\theta)}\geq e^{\epsilon t}\right\} \leq e^{-\epsilon t}\mathop{\mathbb{E}}_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{P}}}\left[e^{\epsilon\Phi(S,\theta)}\right].$$

The last inequality uses Markov's inequality. Using Lemma 4 to upper-bound the moment-generating function, and taking

$$\beta \triangleq n\left(2\beta_{\mathcal{Z}} + M/n\right)^2 + 4T\beta_\Theta^2, \quad \text{and} \quad \epsilon \triangleq \frac{4t}{\beta},$$

we then have that

$$\Pr_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{P}}}\left\{\Phi(S,\theta)\geq t\right\} \leq \exp\left(-\frac{4t^2}{\beta}\right)\exp\left(\left(\frac{4t}{\beta}\right)^2\frac{\beta}{8}\right) = \exp\left(-\frac{2t^2}{\beta}\right)$$
$$= \exp\left(-\frac{2t^2}{n\left(2\beta_{\mathcal{Z}} + M/n\right)^2 + 4T\beta_\Theta^2}\right).$$

Therefore, recalling the definition of $\Phi(S,\theta)$, and solving for t, we have that

$$\Pr_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{P}}}\left\{G(S,\theta) - \mathbb{E}\left[G(S,\theta)\right] \geq \sqrt{\frac{(M+2n\beta_{\mathcal{Z}})^2 + 4nT\beta_\Theta^2}{2n}\ln\frac{1}{\delta}}\right\}$$
$$\leq \exp\left(\frac{-2\left(\sqrt{\frac{(M+2n\beta_{\mathcal{Z}})^2+4nT\beta_\Theta^2}{2n}\ln\frac{1}{\delta}}\right)^2}{n\left(2\beta_{\mathcal{Z}} + M/n\right)^2 + 4T\beta_\Theta^2}\right) = \delta.$$

Thus, with probability at least $1-\delta$ over draws of both $S\sim\mathbb{D}^n$ and $\theta\sim\mathbb{P}$,

$$G(S,\theta) \leq \mathbb{E}\left[G(S,\theta)\right] + \sqrt{\frac{((M+2n\beta_{\mathcal{Z}})^2 + 4nT\beta_\Theta^2)\ln\frac{1}{\delta}}{2n}}.$$

Since $\mathbb{E}\left[G(S,\theta)\right] = \mathbb{E}\left[G(S,\mathbb{P})\right]$, we apply Equation 26 to finish the proof.

A.6 Proof of Theorem 3

PAC-Bayesian analysis typically requires a key step known as *change of measure*, attributed to Donsker and Varadhan [4]. If X is a random variable taking values in Ω, then for any two distributions, $\mathbb{P}$ and $\mathbb{Q}$, on Ω,

$$\mathop{\mathbb{E}}_{X\sim\mathbb{Q}}\left[X\right] \leq D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\mathop{\mathbb{E}}_{X\sim\mathbb{P}}\left[e^X\right]. \tag{31}$$

Let $\epsilon > 0$ denote a free parameter, which we will define later. Via Equation 31, we have that

$$G(S,\mathbb{Q}) = \frac{1}{\epsilon}\mathop{\mathbb{E}}_{\theta\sim\mathbb{Q}}\left[\epsilon\,G(S,\theta)\right]$$
$$\leq \frac{1}{\epsilon}\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[\exp\left(\epsilon\,G(S,\theta)\right)\right]\right). \tag{32}$$

By Markov's inequality, with probability at least $1-\delta$ over draws of $S \sim \mathbb{D}^n$,

$$\begin{aligned}\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[\exp\left(\epsilon\, G(S,\theta)\right)\right] &\leq \frac{1}{\delta}\mathop{\mathbb{E}}_{S\sim\mathbb{D}^n}\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[\exp\left(\epsilon\, G(S,\theta)\right)\right] \\ &= \frac{1}{\delta}\mathop{\mathbb{E}}_{S\sim\mathbb{D}^n}\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[\exp\left(\epsilon\,\mathbb{E}\left[G(S,\theta)\right] + \epsilon\,\Phi(S,\theta)\right)\right] \\ &= \frac{1}{\delta}\exp\left(\epsilon \mathop{\mathbb{E}}_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{P}}}\left[G(S,\theta)\right]\right)\mathop{\mathbb{E}}_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{P}}}\left[\exp\left(\epsilon\,\Phi(S,\theta)\right)\right].\end{aligned}$$

The second line uses the definition of $\Phi(S,\theta)$ from Equation 27, where the inner expectation is over $S \sim \mathbb{D}^n$ and $\theta \sim \mathbb{P}$. The last line uses the fact that the inner expectation is constant with respect to the outer expectation. We can now bound the righthand terms separately. Using Equation 26, we have that

$$\exp\left(\epsilon \mathop{\mathbb{E}}_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{P}}}\left[G(S,\theta)\right]\right) = \exp\left(\epsilon \mathop{\mathbb{E}}_{S\sim\mathbb{D}^n}\left[G(S,\mathbb{P})\right]\right) \leq \exp\left(\epsilon\,\beta_{\mathcal{Z}}\right).$$

We use Lemma 4 to upper-bound $\mathbb{E}\left[\exp\left(\epsilon\,\Phi(S,\theta)\right)\right]$. Putting the pieces together, and letting $\bar{\beta}_{\mathcal{Z}} = 2\beta_{\mathcal{Z}} + M/n$, we thus have that with probability at least $1-\delta$,

$$\mathop{\mathbb{E}}_{\theta\sim\mathbb{P}}\left[\exp\left(\epsilon\, G(S,\theta)\right)\right] \leq \frac{1}{\delta}\exp\left(\epsilon\,\beta_{\mathcal{Z}} + \frac{\epsilon^2}{8}\left(n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_{\Theta}^2\right)\right),$$

which implies (via Equation 32)

$$G(S,\mathbb{Q}) \leq \beta_{\mathcal{Z}} + \frac{1}{\epsilon}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta}\right) + \frac{\epsilon}{8}\left(n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_{\Theta}^2\right). \tag{33}$$

What remains is to optimize ϵ to minimize the bound. Minimizing an expression of the form $a/\epsilon + b\epsilon$ is fairly straightforward; the optimal value is $\epsilon^\star = \sqrt{a/b}$. However, if we were to apply this formula to Equation 33, the optimal ϵ would depend on $\mathbb{Q}$ via the KL divergence term. Since we want the bound to hold simultaneously for all $\mathbb{Q}$, we need to define ϵ such that it does not depend on $\mathbb{Q}$. To do so, we construct a sequence of discrete values:

$$\forall i = 0,1,2,\ldots,\ \epsilon_i \triangleq 2^i\sqrt{\frac{8\ln\frac{2}{\delta}}{n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_{\Theta}^2}}. \tag{34}$$

For each ϵ_i, we assign $\delta_i \triangleq \delta 2^{-(i+1)}$ probability to the probability that Equation 33 does not hold, substituting (ϵ_i, δ_i) for (ϵ, δ). Thus, with probability at least $1 - \sum_{i=0}^{\infty}\delta_i = 1 - \delta\sum_{i=0}^{\infty}2^{-(i+1)} = 1-\delta$, all $i = 0,1,2,\ldots$ satisfy

$$G(S,\mathbb{Q}) \leq \beta_{\mathcal{Z}} + \frac{1}{\epsilon_i}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_i}\right) + \frac{\epsilon_i}{8}\left(n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_{\Theta}^2\right). \tag{35}$$

For any $\mathbb{Q}$, we select the optimal index, $i^\star$, as

$$i^\star = \left\lfloor \frac{1}{2\ln 2}\ln\left(\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)} + 1\right)\right\rfloor. \tag{36}$$

Since

$$\frac{1}{2}\sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)} + 1} \leq 2^{i^\star} \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)} + 1},$$

with a bit arithmetic, we have that

$$\sqrt{\frac{2(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta})}{n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_{\Theta}^2}} \leq \epsilon_{i^\star} \leq \sqrt{\frac{8(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta})}{n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_{\Theta}^2}}. \tag{37}$$

It can also be shown [9] that

$$D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{i^\star}} \leq \frac{3}{2}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right). \tag{38}$$

Thus, with probability at least $1-\delta$, every posterior, $\mathbb{Q}$, satisfies

$$\begin{aligned} G(S,\mathbb{Q}) &\leq \beta_{\mathcal{Z}} + \frac{1}{\epsilon_{i^\star}}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{i^\star}}\right) + \frac{\epsilon_{i^\star}}{8}\left(n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_\Theta^2\right) \\ &\leq \beta_{\mathcal{Z}} + \sqrt{\frac{n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_\Theta^2}{2(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta})}} \cdot \frac{3}{2}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right) \\ &\quad + \sqrt{\frac{8(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta})}{n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_\Theta^2}} \cdot \frac{n\bar{\beta}_{\mathcal{Z}}^2 + 4T\beta_\Theta^2}{8} \\ &= \beta_{\mathcal{Z}} + \sqrt{\frac{2\left(n^2\bar{\beta}_{\mathcal{Z}}^2 + 4nT\beta_\Theta^2\right)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right)}{n}}. \end{aligned} \tag{39}$$

Substituting the definition of $G(S,\mathbb{Q})$ and $\bar{\beta}_{\mathcal{Z}}$, we obtain Equation 4.

A.7 A Derandomized PAC-Bayes Bound for Product Posteriors

Equation 4 holds with high probability over draws of the training data, but the risk is in expectation over draws of hyperparameters. To obtain a bound similar to Theorem 3, that holds with high probability over draws of both the training data, $S \sim \mathbb{D}^n$, and the hyperparameters, $\theta \sim \mathbb{Q}$, we consider posteriors that are product measures.

Theorem 5. *Suppose $\mathcal{A}$ is a $(\beta_{\mathcal{Z}}, \beta_\Theta)$-uniformly stable learning algorithm with respect to a loss function, L, and a fixed product measure, $\mathbb{P}$, on Θ. Then, for any $n \geq 1$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of both a dataset, $S \sim \mathbb{D}^n$, and hyperparameters, $\theta \sim \mathbb{Q}$, from any posterior product measure, $\mathbb{Q}$, on Θ,*

$$R(S,\theta) \leq \hat{R}(S,\theta) + \beta_{\mathcal{Z}} + \beta_\Theta\sqrt{2\,T\ln\frac{2}{\delta}} + \sqrt{\frac{2\left((M+2n\beta_{\mathcal{Z}})^2 + 4nT\beta_\Theta^2\right)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{4}{\delta}\right)}{n}}. \tag{40}$$

Note that, if $\beta_\Theta = \mathrm{O}(1/T)$, the term $\beta_\Theta\sqrt{2\,T\ln\frac{1}{\delta_2}}$ vanishes at a rate of $\mathrm{O}(1/\sqrt{T})$. And if $T \geq n$, this term is dominated by $\mathrm{O}(1/\sqrt{n})$.

Proof. To accommodate all posteriors that might arise from a draw of $S \sim \mathbb{D}^n$, it helps to consider $\mathbb{Q}$ as a function of S. Accordingly, we let $\mathbb{Q}(S)$ denote the distribution induced by S. With $\delta_1 \triangleq \delta/2$, let

$$E_1(S) \triangleq \left\{\exists\mathbb{Q} : G(S,\mathbb{Q}) \geq +\beta_{\mathcal{Z}} + \sqrt{\frac{2\left((M+2n\beta_{\mathcal{Z}})^2 + 4nT\beta_\Theta^2\right)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta_1}\right)}{n}}\right\}$$

denote the event that there exists a posterior for which Equation 4 does not hold. With $\delta_2 \triangleq \delta/2$, let

$$E_2(S,\theta) \triangleq \left\{G(S,\theta) \geq G(S,\mathbb{Q}(S)) + \beta_\Theta\sqrt{2\,T\ln\frac{1}{\delta_2}}\right\}$$

denote the event that the generalization error of a given θ exceeds the expected generalization error under the posterior $\mathbb{Q}(S)$ by more than $\beta_\Theta\sqrt{2\,T\ln\frac{1}{\delta_2}}$.

The probability that we wish to upper-bound is

$$\begin{aligned} \Pr_{\substack{S\sim\mathbb{D}^n \\ \theta\sim\mathbb{Q}(S)}}\left\{E_1(S) \vee E_2(S,\theta)\right\} &\leq \Pr_{S\sim\mathbb{D}^n}\left\{E_1(S)\right\} \;+\; \Pr_{\substack{S\sim\mathbb{D}^n \\ \theta\sim\mathbb{Q}(S)}}\left\{E_2(S,\theta)\right\} \\ &\leq \Pr_{S\sim\mathbb{D}^n}\left\{E_1(S)\right\} \;+\; \sup_{S\in\mathcal{Z}^n}\Pr_{\theta\sim\mathbb{Q}(S)}\left\{E_2(S,\theta)\,|\,S\right\}. \end{aligned}$$

By Theorem 3, $\Pr_{S\sim\mathbb{D}^n}\{E_1(S)\} \leq \delta_1$. To upper-bound $\Pr_{\theta\sim\mathbb{Q}(S)}\{E_2(S,\theta)\,|\,S\}$, it suffices to show that $G(S,\theta)$ concentrates tightly around $G(S,\mathbb{Q}(S))$.

Recall that $\mathcal{A}$ is $2\beta_\Theta$-uniformly stable with respect to L, independent of the posterior. Remember also that Lemma 2 implies G satisfies McDiarmid's stability condition (Equation 23) with $\beta \triangleq 2\beta_\Theta$. Since $\mathbb{Q}(S)$ is a product measure, we can therefore apply McDiarmid's inequality (Equation 24) with $\epsilon \triangleq \beta_\Theta\sqrt{2\,T\ln\frac{1}{\delta_2}}$:

$$\Pr_{\theta\sim\mathbb{Q}(S)}\{E_2(S,\theta)\,|\,S\} \leq \exp\left(\frac{-2\left(\beta_\Theta\sqrt{2\,T\ln\frac{1}{\delta_2}}\right)^2}{T\left(2\beta_\Theta\right)^2}\right) = \delta_2.$$

Thus,

$$\Pr_{\substack{S\sim\mathbb{D}^n\\ \theta\sim\mathbb{Q}(S)}}\{E_1(S)\vee E_2(S,\theta)\} \leq \delta_1+\delta_2 = \delta;$$

so, with probability at least $1-\delta$,

$$\begin{aligned} G(S,\theta) &\leq \beta_\Theta\sqrt{2\,T\ln\frac{1}{\delta_2}} + G(S,\mathbb{Q}(S)) \\ &\leq \beta_\Theta\sqrt{2\,T\ln\frac{1}{\delta_2}} + \beta_{\mathcal{Z}} + \sqrt{\frac{2\left((M+2n\beta_{\mathcal{Z}})^2+4nT\beta_\Theta^2\right)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\ln\frac{2}{\delta_1}\right)}{n}}. \end{aligned}$$

When we add $\hat{R}(S,\theta)$ to both sides of the inequality, and replace δ_1 and δ_2 with $\delta/2$, we obtain Equation 40. □