Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates

Jeffrey Negrea[*]
University of Toronto,
Vector Institute

Mahdi Haghifam[*]
University of Toronto,
Element AI

Gintare Karolina Dziugaite
Element AI

Ashish Khisti
University of Toronto

Daniel M. Roy
University of Toronto,
Vector Institute

Abstract

In this work, we improve upon the stepwise analysis of noisy iterative learning algorithms initiated by Pensia, Jog, and Loh (2018) and recently extended by Bu, Zou, and Veeravalli (2019). Our main contributions are significantly improved mutual information bounds for Stochastic Gradient Langevin Dynamics via data-dependent estimates. Our approach is based on the variational characterization of mutual information and the use of data-dependent priors that forecast the mini-batch gradient based on a subset of the training samples. Our approach is broadly applicable within the information-theoretic framework of Russo and Zou (2015) and Xu and Raginsky (2017). Our bound can be tied to a measure of flatness of the empirical risk surface. As compared with other bounds that depend on the squared norms of gradients, empirical investigations show that the terms in our bounds are orders of magnitude smaller.

1 Introduction

Stochastic subgradient methods, especially stochastic gradient descent (SGD), are at the core of recent advances in deep-learning practice. Despite some progress, developing a precise understanding of generalization error for that class of algorithms remains wide open. Concurrently, there has been steady progress for noisy variants of SGD, such as stochastic gradient Langevin dynamics (SGLD) [13, 26, 34] and its full-batch counterpart, the Langevin algorithm [13]. The introduction of Gaussian noise to the iterates of SGD expands the set of theoretical frameworks that can be brought to bear on the study of generalization. In pioneering work, Raginsky, Rakhlin, and Telgarsky [26] exploit the fact that SGLD approximates Langevin diffusion, a continuous time Markov process, in the small step size limit. One drawback of this and related analyses involving Markov processes is the reliance on mixing. We hypothesize that SGLD is not mixing in practice, so results based upon mixing may not be representative of empirical performance.

In recent work, Pensia, Jog, and Loh [24] perform a stepwise analysis of a family of noisy iterative algorithms that includes SGLD and the Langevin algorithm. At the foundation of this work is the framework of Russo and Zou [29] and Xu and Raginsky [35], where mean generalization error is controlled in terms of the mutual information between the dataset and the learned parameters. (See also the study of on-average KL stability by Wang, Lei, and Fienberg [33].) However, because the data distribution is unknown, so is any mutual information involving the data. This presents a significant barrier to understanding generalization in terms of mutual information.

[*]Equal contribution authors, order of names was determined randomly.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

One of the key contributions of Pensia et al. is a bound on the mutual information between the data and the final weights, which they construct from a bound on the mutual information between the data and the entire trajectory of weights. By exploiting properties of mutual information, they express the latter as a sum of conditional mutual informations associated with each gradient step. While these conditional mutual informations are also unknown, Pensia et al. obtain a bound in terms of the Lipschitz constant for the objective function being optimized.

By passing to the full trajectory and exploiting Lipschitz continuity, Pensia et al. circumvent the statistical barrier posed by the unknown mutual information. Their analysis, however, introduces several sources of looseness. In particular, the use of Lipschitz constants, which lead to distribution-*independent* bounds, eradicates any hope that these bounds will be non-vacuous for modern models and datasets. Indeed, for deep neural networks, the Lipschitz constant for the empirical risk would be prohibitively large, or in some cases infinite, and would immediately render any bound that depends on them vacuous in regimes of interest. In order to fully exploit the decomposition proposed by [24], one needs distribution-*dependent* bounds on the incremental mutual information at each step.

In fact, by a small change, the bounds established by Pensia et al. can be made to depend on expected-squared-gradient-norms, rather than Lipschitz constants, producing distribution-dependent bounds. The resulting bound would be similar to a PAC-Bayesian bound due to Mou et al. [22], which we consider to be the SGLD generalization result most similar to the present work. Writing $\sum_{t\le T}\eta_t$ for $\sum_{t=1}^{T}\eta_t$, their bound is $O\big(\sqrt{(\beta/n)\sum_{t\le T}\eta_t}\big)$ and does not place restrictions on the learning rate or Lipschitz continuity of the loss or its gradient. In other related work, Li, Luo, and Qiao [20] derive an $O\big((1/n)\sqrt{\beta\sum_{t\le T}\eta_t}\big)$ generalization bound for SGLD that depends on expected-squared-gradient-norms. However their result requires the learning rate to scale inversely with the inverse temperature and the Lipschitz constant of the loss, severely limiting the applicability of their result to typical learning problems. Empirically, squared gradient norms are very large during training, which suggests that bounds based on these quantities may not explain empirical performance. As we will show, the dependence on the expected-squared-gradient-norm is spurious.

The key contribution of the present work is the observation that variants of the mutual information between the learned parameters and a subset of the data can be estimated using the rest of the data. We refer to such estimates as *data-dependent* due to their intermediate dependence on part of the data. The use of data-dependent estimates leads to distribution-dependent bounds that naturally adapt to the model of interest and the data distribution. In particular, using data-dependent estimates, we arrive at bounds in terms of the *incoherence* of gradients in the dataset. Roughly speaking, the incoherence measures the amount by which batch gradients computed on subsets of the data disagree, as quantified by squared norm. Crucially, the incoherence is never larger than the squared-gradient-norm on average, and the incoherence is 0 for most iterations of SGLD with small batches.

We note that the mutual information between learned parameter and a single data point is used to produce generalization bounds in work by Bu, Zou, and Veeravalli [6], Raginsky et al. [27], and Wang, Lei, and Fienberg [33]. However, in the SGLD analysis of [6], they do not use data-dependent estimates. Instead, they also rely on Lipschitz constants, leading to bounds similar to [24].

In the process of developing tighter distribution-dependent bounds, we also observe that, in some circumstances, one may obtain tighter estimates by working with conditional or disintegrated information-theoretic quantities. In particular, doing so provides more opportunities to exchange expectation and concave functions than are available with previous mutual information bounds. Using their own mutual information bound and the chain rule, [6] improve on the generalization error bound for SGLD from [24] by a factor of $\sqrt{\log n}$ where n is the sample size. The advantage of [6] that enables this improvement is that their bound is only penalized once per epoch at a randomly chosen step. This effectively changes the order of an expectation and square-root, improving the bound. Building upon [6, 29, 35], we develop generalization bounds in terms of disintegrated information-theoretic quantities that extract expectations from concave functions as much as possible.

Finally, much like the stepwise analysis of SGD carried out by Hardt, Recht, and Singer [14], one could consider an analysis in terms of uniform stability, e.g., in terms of average leave-one-out KL stability [12]. Under an assumption of uniform stability, [22] also showed that expected generalization error decays rapidly at a $O(1/n)$ rate. However, uniform stability has poor dependence on the Lipschitz constant, and so, does not even hold in simple settings, like univariate logistic regression. As such, we do not believe this framework is suitable for studying SGLD as applied in modern

machine learning. For other work on information-theoretic analyses generalization error, and on SGLD, see [1, 3, 4, 15, 16, 27, 32].

1.1 Contributions

The present paper makes the following contributions:

- We provide novel information-theoretic generalization bounds that relate a learned parameter to a random subset of the training data. These bounds depend on forms of on-average information stability, but are different from those in existing work due to our use of disintegration.
- We introduce the technique of data-dependent priors for bounding mutual information in data-dependent estimates of expected generalization error. Specifically, we use data-dependent priors to forecast the dynamics of iterative algorithms using a randomly chosen subset of the data. Each possible subset yields a generalization bound for the empirical risk over the complementary subset. Combining this with our information-theoretic generalization bounds, we recover generalization error bounds for the empirical risk on the full dataset.
- We develop bounds for Langevin dynamics and SGLD that depend on a measure of the *incoherence* of empirical gradients. This quantity is typically orders of magnitude smaller than the squared gradient norms or Lipschitz constants that other bounds depend upon. In our experiments, the difference was a multiplicative factor between 10^2 and 10^4.
- Our generalization bound for SGLD is $O\big(\min\big\{\sqrt{(\beta/bn)\sum_{t\le T}\eta_t},\ (1/n)\sum_{t\le T}\sqrt{\beta\eta_t}\big\}\big)$ where η_t is the learning rate at iteration t, T is the number of iterations, β is the inverse temperature, and b is the minibatch size. This bound is currently state of the art for bounds without assumptions on the smoothness of the loss or restrictions on the learning rate.

1.2 Preliminaries

Let $\mathscr{D}$ be an unknown distribution on a space $\mathscr{Z}$ and let $\mathscr{W}$ be a space of parameters. Consider a loss function $\ell : \mathscr{Z}\times\mathscr{W}\to\mathbb{R}$ and the corresponding risk function $R_{\mathscr{D}}(w) = \mathbb{E}\ell(Z,w)$. Given an i.i.d. dataset of size n, $S\sim\mathscr{D}^n$, we may form the empirical risk function $\hat{R}_S(w) = \frac{1}{n}\sum_{i=1}^m \ell(Z_i,w)$, where $S=(Z_1,\dots,Z_n)$. In the setting of classification and continuous parameter spaces, the loss function is discontinuous and the empirical risk function does not convey useful gradient information. For this reason, it is common to work with a *surrogate* loss, such as cross entropy. To that end, let $\tilde{\ell} : \mathscr{Z}\times\mathscr{W}\to\mathbb{R}$ denote a surrogate loss and let $\tilde{R}_{\mathscr{D}}(w) = \mathbb{E}\tilde{\ell}(Z,w)$ and $\tilde{R}_S(w) = \frac{1}{n}\sum_{i=1}^m \tilde{\ell}(Z_i,w)$ be the corresponding surrogate risk and empirical surrogate risk.

Our primary interest is in the generalization performance of learning algorithms. Abstractly, let W be a random element in $\mathscr{W}$ satisfying $W=\mathscr{A}(S,V)$, where V is some auxiliary random element independent from S and $\mathscr{A}$ is a measurable function representing a randomized learning algorithm that maps the data S to a learned parameter W. Our focus will be the **(mean) generalization error** of W, i.e., $\mathbb{E}\left[R_{\mathscr{D}}(W)-\hat{R}_S(W)\right]$. Note that we have averaged over both the choice of dataset and the source of randomness V available to the learning algorithm $\mathscr{A}$.

For random variables X and Y, write $\mathbb{E}^Y X = \mathbb{E}[X|Y]$ and $\mathbb{P}^Y[X]$ for the conditional expectation and (regular) conditional distribution, respectively, of X given Y.[2] Besides the usual notions of KL divergence, mutual information, and conditional mutual information (see Appendix A for formal definitions), we rely on the following less common notion:

Definition 1.1. Let X, Y, and Z be arbitrary random elements. Let $\otimes$ form product measures. The **disintegrated mutual information between X and Y given Z** is

$$I^Z(X;Y) = \mathrm{KL}(\mathbb{P}^Z[(X,Y)]\,\|\,\mathbb{P}^Z[X]\otimes\mathbb{P}^Z[Y]).$$

It follows immediately from definitions that $I(X,Y|Z) = \mathbb{E}I^Z(X,Y)$. Letting ϕ satisfy $\phi(Z) = I^Z(X;Y)$ a.s., define $I(X,Y|Z=z)=\phi(z)$. This notation is necessarily well defined only up to a null set under the marginal distribution of Z.

[2] We fix arbitrary versions and assume regular versions of conditional distributions exist.

2 Methods

In this section, we establish generalization bounds for learning algorithms in terms of information-theoretic quantities (conditional mutual information, disintegrated mutual information, relative entropy) that depend on the unknown data distribution and the probabilistic properties of the learning algorithm. We then describe two complementary strategies that we employ to bound these otherwise intractable quantities. In Section 3, we apply these methods to the study of the Langevin algorithm and SGLD.

We make repeated use of generalized notions of *priors* and *posteriors*, which arise in the PAC-Bayes literature ([7, 21, 31], etc.) and relate to variational bounds on mutual information, which we will now describe: Consider learned parameters W, data S, and auxiliary variables V, viewed as random elements in $\mathscr{W}$, Z^n, etc., respectively. In PAC-Bayes, a generalized posterior is an arbitrary random measure on $\mathscr{W}$. In our setting, the **posterior, Q, (of W given S and V)** is the conditional distribution of W given S and V. (Formally, Q is a probability kernel, but one can think informally that $Q = f(S,V)$ for some measurable function taking values in the space of Borel probability measures, and so we will simply say that Q is $\sigma(S,V)$-measurable.)

Definition 2.1 (Data-dependent prior)**.** Let Q be a $\sigma(S,V)$-measurable posterior. A **(generalized) prior** P is a random measure on $\mathscr{W}$, measurable with respect to some sub-σ-algebra of $\sigma(S,V)$. A prior P is said to be **data-dependent** if it is not independent of S.

Let P be a $\mathscr{F}$-measurable data-dependent prior, where $\sigma(V) \subset \mathscr{F}$. Using a variational characterization of mutual information (see Appendix B.1), we have

$$\mathbb{E}^{\mathscr{F}}[\mathrm{KL}(Q \,\|\, P)] \geq I^{\mathscr{F}}(W;S) \text{ a.s.}, \tag{1}$$

with equality for $P = \mathbb{P}^{\mathscr{F}}[W]$. Therefore, if the expected KL divergence is small, W contains little information about S beyond what is already captured by $\mathscr{F}$. If the special case where the disintegrated mutual information is zero, then W is independent of S given $\mathscr{F}$. In the context of generalization, this implies that the data S not contained in $\mathscr{F}$ can be used to form an unbiased estimate of the risk of W. The bounds we present below extend this logic to nonzero mutual information.

The utility of using data-dependent priors to control disintegrated mutual information depends on the balance of two effects: On the one hand, $I(W;S) \leq I(W;S|\mathscr{F})$, and so conditioning never improves a theoretical bound and may make it looser. On the other hand, $I(W;S)$ depends on the *unknown* data distribution and so distribution-independent bounds will often be very loose. In contrast, the KL divergence based on P can exploit the information in $\mathscr{F} \subset \sigma(S,V)$ to obtain tighter data-dependent bounds on $I^{\mathscr{F}}(W;S)$.

In order to construct data-dependent priors, we partition the dataset S in two halves, based on a random subset $J \subset \{1,\ldots,n\}$ with $\#J = m$ nonrandom. Let $J = \{j_1,\ldots,j_m\}$, The first half, $S_J = (Z_{j_1},\ldots,Z_{j_m})$, contains m points, which we will use to construct a data-dependent prior P. The second half, S_J^c, containing the remaining $n-m$ points, is independent of P. (Note that S_J and S_J^c are independent of J, since m is nonrandom.)

This particular construction of data-dependent priors allow us to leverage a type of *non-uniform KL-stability*: the prior P may exploit S_J to make a data-dependent forecast of Q, yielding a bound, B, on the conditional expected generalization error (with respect to the remaining $n-m$ data points in S_J^c). Averaging over S_J, we obtain a bound on the (unconditional) expected generalization error.

Definition 2.2. Let S_J, S_J^c be defined as above. Suppose that $\mathscr{F}$ is a σ-field with $\sigma(S_J) \subset \mathscr{F} \perp\!\!\!\perp \sigma(S_J^c)$. An expected generalization error bound based on a **data-dependent estimate** is one of the form

$$\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right] \leq \mathbb{E}[B], \tag{2}$$

where B is $\mathscr{F}$ measurable, and satisfies $\mathbb{E}^{\mathscr{F}}\left[R_{\mathscr{D}}(W) - \hat{R}_{S_J^c}(W)\right] \leq B$.

The idea of using data-dependent priors to obtain tighter bounds is standard in the PAC-Bayes literature [2, 10, 23, 28], but its utility in the present work is brought through by our introduction of data-dependent estimates. In the following section, we derive information-theoretic bounds on expected generalization error that can exploit data-dependent priors to form data-dependent estimates. We will then use these tools to study SGLD, without mixing assumptions.

2.1 Information-Theoretic Generalization Bounds based on Random Subsets of Data

Existing work by Xu and Raginsky [35] bounds the expected generalization error of a learning algorithm in terms of the mutual information between the random parameters and the data. The following result is a simple extension of [35, Thm. 1] that bounds the expected generalization error in terms of the mutual information between the parameters and a random subset of the data.

Theorem 2.3 (Data-Dependent Mutual Information Bound)**.** *Let W be a random element in $\mathscr{W}$, let $S \sim \mathscr{D}^n$, and let $J \subseteq [n]$, $|J| = m$, be uniformly distributed and independent from S and W. Suppose that $\ell(Z,w)$ is σ-subgaussian when $Z \sim \mathscr{D}$, for each $w \in \mathscr{W}$. Let $Q = \mathbb{P}^S[W]$, and let P be a $\sigma(S_J)$-measurable data-dependent prior on $\mathscr{W}$. Then*

$$\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right] \leq \sqrt{2\frac{\sigma^2}{n-m} I(W;S_J^c)} \leq \sqrt{2\frac{\sigma^2}{n-m}\mathbb{E}[\mathrm{KL}(Q\,\|\,P)]}.$$

The proof of this result can be found in Appendix B. When $m = 0$, this recovers [35, Thm. 1].

When the size of the subset is $m = n-1$, this bound is weaker than [6, Prop. 1], due to the order of the concave square-root function and the expectation over the choice datapoint to be left out. This difference is addressed by our next result.

Randomization is one way that learning algorithms can control the mutual information between (a random subsets of) the data and the learned parameter. Let U be a random element independent from S and J, representing some aspect of the source of randomness used by the learning algorithm. Because $S \perp\!\!\!\perp \{J,U\}$ and $S \sim \mathscr{D}^n$, we have $(S_J,U) \perp\!\!\!\perp S_J^c$ and thus

$$I(W;S_J^c) \leq I(W;S_J^c|S_J,U) = \mathbb{E}I^{S_J,U}(W;S_J^c),$$

where the last equality follows from the definition of conditional mutual information. The next result shows that we can pull the expectation over both S_J and U outside the concave square-root function. In the case of SGLD, U will be the sequence of minibatch index sets.

Theorem 2.4 (Data-Dependent Disintegrated Mutual Information Bound)**.** *Let W, S, and J be as in Theorem 2.3, and let U be independent from S and J. Suppose that $\ell(Z,w)$ is σ-subgaussian when $Z \sim \mathscr{D}$, for each $w \in \mathscr{W}$. Let $Q = \mathbb{P}^{S,U}[W]$ and let P be a $\sigma(S_J,U)$-measurable data-dependent prior on $\mathscr{W}$. Then*

$$\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right] \leq \mathbb{E}\sqrt{2\frac{\sigma^2}{n-m} I^{S_J,U}(W;S_J^c)} \leq \mathbb{E}\sqrt{2\frac{\sigma^2}{n-m}\mathbb{E}^{S_J,U}\mathrm{KL}(Q\,\|\,P)}$$

The proof of this result can be found in Appendix B. Since $I^{S_J,U}(W;S_J^c)$ is (S_J,U)-measurable, we may use S_J and U to obtain a data-dependent bound. In the case that $m = n-1$, our bound is similar to, but not strictly comparable to, [6, Prop. 1]. Our bound is incomparable due to our use of disintegrated mutual information, $I^{S_J}(W;S_J^c)$ and the fact that we take the expectations over the dataset outside of the convex square-root function. The disintegrated mutual information cannot be upper bounded by the full mutual information, $I(W,S_J^c)$, which appears in [6] (even by taking expectations under the square root using Jensen's inequality). However, Theorem 2.4 is essentially a disintegrated version of [6, Prop. 1]. In their actual SGLD expected generalization error bound, [6] controls the unconditional mutual information using the Lipschitz constant of the surrogate loss. Hence, one could easily recover the same bound using our result. The conditioning we have done, however, allows us to control the mutual information more carefully in order to achieve a tighter bound for SGLD than is provided by [6].

These bounds allow for a tradeoff: for large m, the mutual information is measured between the parameter and a small random subset of the data, and so we expect the mutual information to be small. (Indeed, this term will decrease monotonically in m.) At the same time, the $\frac{1}{n-m}$ term is larger, reflecting the reduced effect of averaging over only $n-m$ data to form our estimate of the empirical risk. It is unclear without further context whether this bound is tighter in the regime of small, intermediate, and large m. In fact, we find that, for the bounds we derive in our applications, $m = n-1$ is optimal. This difference materially affects the quality and tightness of the bounds, as is discussed in Remark 3.4. However, for $m = n-1$ and bounded loss, the following bound is tighter, while it is incomparable for other values of m.

Theorem 2.5 (Data-Dependent KL Bound)**.** *Let W, S, J, and U be as in Theorem 2.4. Let $Q = \mathbb{P}^{S,U}[W]$ and let P be a $\sigma(S_J, U)$-measurable data-dependent prior on $\mathscr{W}$. Suppose that $\ell(Z,w)$ is $[a_1,a_2]$-bounded a.s. when $Z \sim \mathscr{D}$, for each $w \in \mathscr{W}$.*

$$\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right] \leq \mathbb{E}\sqrt{\frac{(a_2-a_1)^2}{2}\,\mathrm{KL}(Q\,\|\,P)}.$$

The proof of this result can be found in Appendix B. For an analytic comparison of the three bounds in the case that $m = n-1$, see Appendix F. Remark B.2 explains why this result is only stated for bounded loss functions.

2.2 Decomposing KL Divergences and Mutual Information for Sequential Algorithms

Consider an iterative learning algorithm, and let $W_0, W_1, W_2, \ldots W_T \in \mathscr{W}$ be the parameters during the course of T iterations. In light of the variational bound for mutual information, we can obtain a generalization bound for W_T by bounding the expected KL divergences between the conditional distribution $\mathbb{P}^{S_J}[W_T]$ and some S_J-measurable "prior" distribution $P(Z)$. Unfortunately, the first distribution has no known tractable representation. Pensia, Jog, and Loh [24] use monotonicity to bound a mutual information involving the terminal parameter with one involving the full trajectory, then use the chain rule to decompose this into a sum of conditional mutual informations. The same principles allow us to first bound the terminal KL divergence by the KL for the full trajectory, and then decompose the KL divergence for the full trajectory over each individual step.

Setting some notation, let T be a nonnegative integer, let $[T]_0 = \{0,1,2,\ldots,T\}$, let μ be a distribution on $\mathscr{W}^{[T]_0}$, and let X be a random variable with distribution μ. We are interested in naming certain marginal and conditional distributions (disintegrations) related to μ. In particular, for $t \in [T]_0$, let

i) $\mu_t = \mathbb{P}[X_t]$, the marginal law of X_t;
ii) $\mu_{t|} = \mathbb{P}^{X_{0:(t-1)}}[X_t]$, the conditional law of X_t given $X_{0:(t-1)}$; and
iii) $\mu_{0:t} = \mathbb{P}[X_{0:t}]$, the marginal law of $X_{0:t}$.

Proposition 2.6 (Decomposition of KL Divergences)**.** *Let Q, P be probability measures on $\mathscr{W}^{[T]_0}$. Suppose that $Q_0 = P_0$. Then*

$$\mathrm{KL}(Q_T\,\|\,P_T) \leq \mathrm{KL}(Q\,\|\,P) = \textstyle\sum_{t=1}^{T} \mathbb{E}_{Q_{0:(t-1)}}[\mathrm{KL}(Q_{t|}\,\|\,P_{t|})].$$

where, as per Section 1.2, $Q_{t|}$ is the conditional law of t-th iterate given the previous iterates, and so $\mathrm{KL}(Q_{t|}\,\|\,P_{t|})$ is a random variable which depends the $(W_0,\ldots W_{t-1}) \sim Q_{0:t-1}$.

The proof of this result may be found in Appendix B.

Considering the KL between full trajectories may yield a loose upper bound on the KL between terminal parameters (in particular, when the trajectory cannot be inferred from the terminus). We gain, however, analytical tractability, as we will see in the next section when we analyze particular algorithms stepwise. In fact, many bounds that appear in the literature implicitly require this form of incrementation. Our approach based on the KL divergence and data-dependent priors gives us much tighter control of the KL divergence contribution of each step.

3 Generalization Bounds for Specific Algorithms

Now that we have all of the theoretical tools required, we may establish bounds on the generalization error of specific noisy iterative learning algorithms by inventing sensible data-dependent priors. The use of a data-dependent prior which closely forecasts the true algorithm in each step is key in establishing tighter generalization bounds. We first consider the stochastic gradient Langevin dynamics (SGLD) algorithm [34], then handle its full batch counterpart the (unadjusted) Langevin algorithm [9, 11], which we will refer to informally as Langevin dynamics (LD). Note that the loss and risk functions used for training, $(\tilde{\ell}, \tilde{R}_{\mathscr{D}}, \tilde{R}_S)$, need not be the same loss functions used for assessing performance and generalization error, $(\ell, R_{\mathscr{D}}, \hat{R}_S)$, as explained in Section 1.2.

3.1 Stochastic Gradient Langevin Dynamics

Let η_t to be the learning rate at time t; β_t be the inverse temperature at time t; and ε_t, i.i.d. $\mathscr{N}(0, \mathbb{I}_d)$. Let b_t be the minibatch size at time t. We are interested in stochastic gradient Langevin dynamics,

whose iterates are given by

$$W_{t+1} = W_t - \eta_t \nabla \tilde{R}_{S_t}(W_t) + \sqrt{2\eta_t/\beta_t}\,\varepsilon_t. \tag{3}$$

where $\tilde{R}_{S_t}(w) = \frac{1}{b_t}\sum_{z\in S_t}\tilde{\ell}(w,z)$, and S_t is a subset of S of size b_t sampled uniformly at random with a sampling procedure which is independent of S, and independent of $\{\varepsilon_t\}_{t\geq 0}$. The b_t data points in S_t are chosen *without replacement*.

3.1.1 A data-dependent prior for SGLD

Let S_J be a random subset of S, of size m, chosen independently from $W_0, W_1, \ldots$, and independently of the sequence of minibatches, $\{S_t\}_{t\geq 0}$. Let the set of indices appearing in the t-th minibatch be denoted by K_t, so that $S_t = S_{K_t}$ for each t. By assumption, each K_t is a uniformly random subset of $\{1,\ldots,n\}$ of size b_t. We set $U = (K_1, \ldots K_T)$, as to match the notation in the theorems of Section 2.1. Let $S_{J_t} = S_J \cap S_t = S_{J\cap K_t}$ and let $b'_t = \#S_{J_t}$. Let $S^c_t = S_t \setminus S_J = S_{K_t\setminus J}$ and $b^c_t = b_t - b'_t$. Define

$$\xi_t = \frac{b^c_t}{b_t}\left(\nabla \tilde{R}_{S^c_t}(W_t) - \nabla \tilde{R}_{S_J}(W_t)\right). \tag{4}$$

Let $Q(S,U)$ be the joint law of $(W_0,\ldots,W_T)$ given a dataset S and minibatch sequence U. Then $Q(S,U)$ is a random measure as it depends on the random dataset S and the sequence of indices U. It follows from Eq. (3) that $Q(S,U)_{t|}$ is multivariate normal with mean $\mu_{Q,t}(S,U) = W_t - \eta_t \nabla \tilde{R}_S(W_t)$ and covariance $2\frac{\eta_t}{\beta_t}\mathbb{I}_d$. Consider the data-dependent prior defined so that its conditional $P_{t|}(S_J,U)$ is a multivariate normal with covariance $2\frac{\eta_t}{\beta}\mathbb{I}_d$, and with mean

$$\mu_{P,t}(S_J,U) = W_t - \eta_t\left(\frac{b'_t}{b_t}\nabla \tilde{R}_{S_{J_t}}(W_t) + \frac{b_t - b'_t}{b_t}\nabla \tilde{R}_{S_J}(W_t)\right).$$

Note that $\mu_{Q,t}(S,U) - \mu_{P,t}(S_J,U) = \eta_t \xi_t(S,\mathrm{idx})$. Thus the one-step KL divergence satisfies

$$2\mathrm{KL}(Q_{t+1|}(S,\mathrm{idx})\,\|\,P_{t+1|}(S_J,U)) = \frac{\beta_t\eta_t}{4}\|\xi_t\|_2^2$$

Applying Proposition 2.6, we have (almost surely over the choice of (S,J,U))

$$2\mathrm{KL}(Q_T(S,U)\,\|\,P_T(S_J,U)) \leq \sum_{t=1}^T \mathbb{E}^{S,J,U}\mathrm{KL}(Q_{t|}(S,U)\,\|\,P_{t|}(S_J,U)) = \sum_{t=1}^T \mathbb{E}^{S,J,U}\frac{\beta_t\eta_t}{4}\|\xi_t\|_2^2.$$

Note that ξ_t depends on the exact weight sequence, and hence is $\sigma(S,J,U,W_{t-1})$-measurable, but not $\sigma(S,J,U)$-measurable. Hence, $\mathbb{E}^{S,J,U}\frac{\beta_t\eta_t}{8}\|\xi_t\|_2^2$ is a $\sigma(S,J,U)$-measurable for each t.

3.1.2 Expected Generalization Error Bounds for SGLD

Theorem 3.1 (Expected Generalization Error Bounds for SGLD)**.** *Let $\{W_t\}_{t\in[T]}$ denote the iterates of SGLD. Let the batch size be constant, $b_t = b$. If $\ell(Z,w)$ is σ-subgaussian for each $w\in\mathscr{W}$, then*

$$\mathbb{E}(R_{\mathscr{D}}(W_T) - R_S(W_T)) \leq \mathbb{E}\sqrt{\frac{\sigma^2}{n-m}\sum_{t=1}^T \frac{\beta_t\eta_t}{4}\mathbb{E}^{S_J,J,U}\|\xi_t\|_2^2} \leq \frac{\sigma}{2}\sqrt{\frac{n}{(n-1)^2}\sum_{t=1}^T\left(\tfrac{1}{b} + \tfrac{1}{n}\tfrac{n-m-1}{m}\right)\beta_t\eta_t\mathrm{tr}(\mathbb{E}[\hat{\Sigma}_t(S)])} \tag{5}$$

and if $\ell(Z,w)$ is $[a_1,a_2]$-bounded, and if $m = n-1$, then

$$\mathbb{E}(R_{\mathscr{D}}(W_T) - R_S(W_T)) \leq \mathbb{E}\sqrt{\frac{(a_2-a_1)^2}{4}\sum_{t=1}^T\frac{\beta_t\eta_t}{4}\mathbb{E}^{S,J,U}\|\xi_t\|_2^2} \leq \left[\frac{(a_2-a_1)^2 n}{4(n-1)^2 b}\right]^{1/2}\mathbb{E}\sqrt{\sum_{t=1}^T\frac{\beta_t\eta_t}{4}\mathrm{tr}(\mathbb{E}^S[\hat{\Sigma}_t(S)])} \tag{6}$$

where $\hat{\Sigma}_t(S) = \underset{Z\sim Unif(S)}{Var^{W_t,S}}(\nabla\tilde{R}_Z(W_t))$ is the finite population variance matrix of surrogate gradients.

Proof. The results are the direct combinations of Theorem 2.4 and Propositions 2.6 and B.1; and Theorem 2.5 and Proposition 2.6, respectively, with our data-dependent prior. Jensen's inequality is used to move expectations under $\sqrt{\cdot}$. Lemma D.2 expresses the results in terms of $\hat{\Sigma}$. □

Remark 3.2. Suppose that $\beta_t = \beta$, $b_t = b$, and $m = n-1$. Under uniform moment conditions on $\mathbb{E}^{S_J,J,U}\|\xi_t\|_2^2$, our generalization error bounds in Eq. (5) is clearly $O\big(\sqrt{(\beta/bn)\sum_{t\le T}\eta_t}\big)$. Since $\xi_t = 0$ whenever $K_t \subset J$, we find that our first bound in Eq. (5) is also $O\big((1/n)\sum_{t\le T}\sqrt{\beta\eta_t}\big)$. To see this, notice that for non-negative random variables C_t and $B_t \sim \mathrm{Ber}(p)$,

$$\mathbb{E}\sqrt{\textstyle\sum_{t=1}^T B_t C_t} \le \mathbb{E}[\textstyle\sum_{t=1}^T B_t\sqrt{C_t}] = p\textstyle\sum_{t=1}^T \mathbb{E}[\sqrt{C_t}|B_t = 1].$$

When $m = n-1$, taking $B_t = I_{\xi_t \neq 0}$, $p = b/n$, $C_t = \frac{\beta_t\eta_t}{8}\mathbb{E}^{S_J,J,U}\|\xi_t\|_2^2$ yields the stated rate. ◁

3.2 Langevin Dynamics

Under the same notation as above, the iterates of the Langevin dynamics algorithm are given by

$$W_{t+1} = W_t - \eta_t\nabla\tilde{R}_S(W_t) + \sqrt{2\eta_t/\beta_t}\,\varepsilon_t. \tag{7}$$

3.2.1 Expected Generalization Error Bounds for LD

We can recover bounds generalization error bounds for LD as a special case of SGLD when the batch size is the dataset size, $b_t = n$ for all t. The data-dependent prior is the same as for SGLD.

Theorem 3.3 (Expected Generalization Error Bounds for Langevin Dynamics)**.** *Let $\{W_t\}_{t\in[T]}$ denote the iterates of the Langevin dynamics algorithm. If $\ell(Z,w)$ is σ-subgaussian for each $w \in \mathscr{W}$, then*

$$\mathbb{E}(R_{\mathscr{D}}(W_T) - R_S(W_T))4 \le \sqrt{\frac{\sigma^2}{(n-1)m}\sum_{t=1}^T \frac{\beta_t\eta_t}{4}\mathbb{E}\mathrm{tr}(\hat{\Sigma}_t(S))}, \tag{8}$$

and if $\ell(Z,w)$ is $[a_1,a_2]$-bounded and $m = n-1$, then

$$\mathbb{E}(R_{\mathscr{D}}(W_T) - R_S(W_T)) \le \mathbb{E}\sqrt{\frac{(a_2-a_1)^2}{4}\sum_{t=1}^T \frac{\beta_t\eta_t}{4}\mathbb{E}^{S_J}\|\xi_t\|_2^2} \le \frac{a_2-a_1}{2(n-1)}\mathbb{E}\sqrt{\sum_{t=1}^T \frac{\beta_t\eta_t}{4}\mathbb{E}^S\mathrm{tr}(\hat{\Sigma}_t(S))},$$

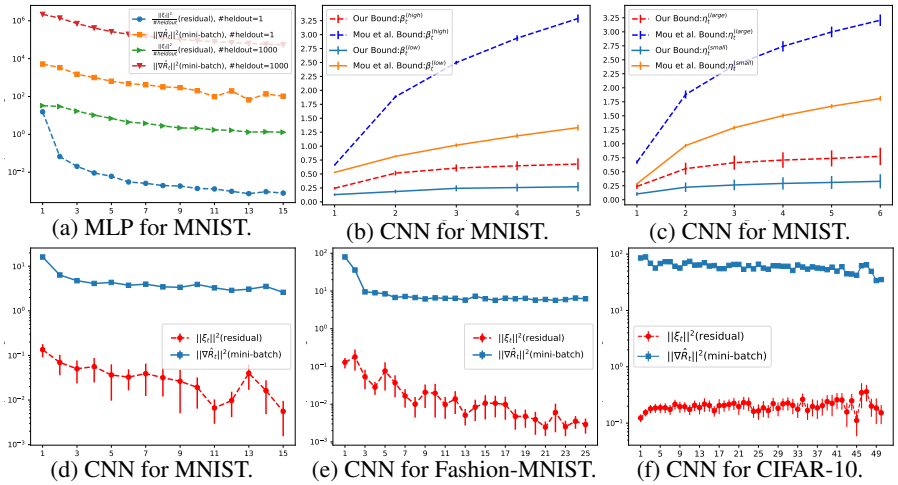
where $\hat{\Sigma}_t(S) = \underset{Z\sim Unif(S)}{Var^{W_t,S}}(\nabla\tilde{R}_Z(W_t))$ is the finite population variance matrix of surrogate gradients.

For asymptotic properties of this bound when $\tilde{\ell}$ is L-Lipschitz, as in [24], see Appendix E. For a simple analytic worked example of mean estimation using Langevin dynamics, refer to Appendix G.

Remark 3.4 (Dependence of our bounds on the subset size, m). The choice of $m \in \{1,\dots,n\}$ can make a material difference in the quality of the bound and whether it is vacuous or not. As seen in Eq. (8), if m is $\Omega(n)$ then the upper bound on expected generalization error is $O(\beta/n)$. If β is $\Omega(\sqrt{n})$, as is typical in practice, then overall, the bound is $O(n^{-1/2})$. If, on the other hand, m is $o(n)$ then the order of the bound with respect to n would be lower—in particular if m is $O(\sqrt{n})$ then our bound would not be decreasing in n for β of order $\Omega(\sqrt{n})$. ◁

4 Empirical Results

We have developed bounds that depend on the gradient prediction residual of our data dependent priors (which we call the *incoherence* of the gradients), rather than on the gradient norms (as in [22]) or Lipschitz constants (as in [6, 24]). The extent to which this represents an advance is, however, an empirical question. The functional form of our bounds and those in the cited work are nearly identical. The first key differences between our work and others is the replacement of gradient norms ($\|\nabla\tilde{R}_t\|^2$) and Lipschitz constants in other work with gradient prediction residual, ($\|\xi_t\|$) in our work. The second key difference is the order of expectations and square-roots, which favor our bounds due to Jensen's inequality. In this section, we perform an empirical comparison of the gradient prediction residual of our data dependent priors and the gradient norm across various architectures and datasets. This illustrates the first of the differences, the quantities appearing in the bound. Our results indicate that that our data-dependent priors yield significantly tighter results, as the sum of square gradient incoherences of our data dependent priors are between 10^2 and 10^4 times smaller than the sum of square gradient norms in the experiments we ran.



(a) MLP for MNIST. (b) CNN for MNIST. (c) CNN for MNIST.

(d) CNN for MNIST. (e) CNN for Fashion-MNIST. (f) CNN for CIFAR-10.

Figure 1: Numerical results for various datasets and architectures. All x-axes show the number of Epochs of training. Fig. 1a shows the effect of different amounts of heldout data on the summands appearing in our bound, and what those would be if we upper bounded the *incoherence* $\|\xi\|$ by $\|\nabla\hat{R}\|$ when it is not 0. Fig. 1b compares a Monte Carlo estimate of our bound with that of [22] and shows the effect of inverse temperature on each. Fig. 1c compares a Monte Carlo estimate of our bound with that of [22] and shows the effect of learning rate on each. Figs. 1d to 1f compare the summands appearing in our bound and those of [22] across datasets.

In Fig. 1, we compare $\|\xi_t\|^2$ and $\|\nabla\tilde{R}_t\|^2$ in order to assess the improvement our methods bring over existing results for SGLD. Specifically, the values of each plot are the averages of $\sqrt{\eta\beta}\|\xi_t\|/b$ and $\sqrt{\eta\beta}\|\nabla\tilde{R}_{S_t}\|/b$ over an epoch. These serve as estimates of the per-epoch contributions to the respective summations in our Theorem 3.1 and the bound of Mou et al. (Thm. 2 therein, when there is no L_2-regularization). The average and standard error of both expressions taken over multiple runs are displayed. Bounds from related work that depend on Lipschitz constants would further upper bound what we show for [22], by replacing $\|\nabla\tilde{R}_t\|$ with a Lipschitz constant. The Lipschitz constant could be lower bounded by the largest observed gradient norm, and would be off the chart.

From Fig. 1a, we see that the empirical performance reflects our analytical results that the bound is tighter for large m. As can be inferred from Eq. (4), the difference between $\|\xi_t\|^2$ and $\|\nabla\tilde{R}_t\|^2$ increases with m. From Figs. 1d to 1f we see that the squared gradient incoherence, $\|\xi_t\|^2$, are between 100 and 10,000 times smaller than the squared gradient norms, $\|\nabla\tilde{R}\|^2$ in all of these examples.

Using Monte Carlo simulation, we compared estimates of our expected generalization error bounds with (coupled) estimates of the bound from [22]. The results, in Figs. 1b and 1c, show that our bounds are materially tighter, and remain non-vacuous after many more epochs. Fig. 1b also compares the two generalization error bounds for different inverse temperature schedules. Fig. 1c compares the two generalization error bounds based for different learning rate schedules. It can inferred from Figs. 1b and 1c that our proposed bound yields to tighter values when the learning rate and the inverse temperature are small. However, it should be noted that with small learning rate and the inverse temperature, it would be difficult to have a very low training error when the empirical risk minimization is performed using SGLD.

The details of our model architectures, temperature, learning rate schedules and hyperparameter selections may be found in Appendix H. We did not aim to achieve the state-of-the art predictive performance. With further tuning, the prediction results could be improved.

Acknowledgments

JN is supported by an NSERC Vanier Canada Graduate Scholarship, and by the Vector Institute. MH was supported by a MITACS Accelerate Fellowship with Element AI. DMR is supported by an NSERC Discovery Grant and an Ontario Early Researcher Award. This research was carried out in part while GKD and DMR were visiting the Simons Institute for the Theory of Computing.

References

[1] A. Lopez and V. Jog. "Generalization error bounds using Wasserstein distances". In: *IEEE Information Theory Workshop*. 2018.

[2] A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. "Tighter PAC-Bayes bounds". In: *Advances in Neural Information Processing Systems*. 2007, pp. 9–16.

[3] A. Asadi, E. Abbe, and S. Verdú. "Chaining mutual information and tightening generalization bounds". In: *Advances in Neural Information Processing Systems*. 2018, pp. 7234–7243.

[4] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff. "Learners that Use Little Information". In: *Algorithmic Learning Theory*. 2018, pp. 25–55.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[6] Y. Bu, S. Zou, and V. V. Veeravalli. "Tightening Mutual Information Based Bounds on Generalization Error". In: *IEEE International Symposium on Information Theory (ISIT)*. To appear. 2019. arXiv: 1901.04609.

[7] O. Catoni. "PAC-Bayesian supervised classification: the thermodynamics of statistical learning". In: *Institute of Mathematical Statistics Lecture Notes-Monograph Series*. Vol. 56. 2007. arXiv: 1901.04609.

[8] M. D. Donsker and S. S. Varadhan. "Asymptotic evaluation of certain Markov process expectations for large time, I". *Communications on Pure and Applied Mathematics* 28.1 (1975), pp. 1–47.

[9] A. Durmus and E. Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". *The Annals of Applied Probability* 27.3 (2017), pp. 1551–1587.

[10] G. K. Dziugaite and D. M. Roy. "Data-dependent PAC-Bayes priors via differential privacy". In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 29. Cambridge, MA: MIT Press, 2018. arXiv: 1802.09583.

[11] D. L. Ermak. "A computer simulation of charged particles in solution. I. Technique and equilibrium properties". *The Journal of Chemical Physics* 62.10 (1975), pp. 4189–4196.

[12] V. Feldman and T. Steinke. "Calibrating Noise to Variance in Adaptive Data Analysis". In: *Conference On Learning Theory*. 2018, pp. 535–544.

[13] S. B. Gelfand and S. K. Mitter. "Recursive stochastic algorithms for global optimization in R^d". *SIAM Journal on Control and Optimization* 29.5 (1991), pp. 999–1018.

[14] M. Hardt, B. Recht, and Y. Singer. "Train faster, generalize better: Stability of stochastic gradient descent". In: *International Conference on Machine Learning*. 2016. arXiv: 1509.01240.

[15] A. E. I. Ibrahim and M. Gastpar. *Strengthened Information-theoretic Bounds on the Generalization Error*. 2019. arXiv: 1903.03787.

[16] J. Jiao, Y. Han, and T. Weissman. "Dependence measures bounding the exploration bias for general measurements". In: *IEEE International Symposium on Information Theory*. 2017.

[17] O. Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

[18] J. Kemperman. "On the Shannon capacity of an arbitrary channel". In: *Indagationes Mathematicae (Proceedings)*. Vol. 77. 2. North-Holland. 1974, pp. 101–115.

[19] Y. LeCun, C. Cortes, and C. J. C. Burges. *MNIST handwritten digit database*. http://yann.lecun.com/exdb/mnist/. 2010.

[20] J. Li, X. Luo, and M. Qiao. *On generalization error bounds of noisy gradient methods for non-convex learning*. 2019. arXiv: 1902.00621.

[21] D. A. McAllester. "Some PAC-Bayesian Theorems". *Machine Learning* 37.3 (Dec. 1999), pp. 355–363. ISSN: 1573-0565.

[22] W. Mou, L. Wang, X. Zhai, and K. Zheng. "Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 605–638.

[23] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. "PAC-Bayes bounds with data dependent priors". *Journal of Machine Learning Research* 13.Dec (2012), pp. 3507–3531.

[24] A. Pensia, V. Jog, and P.-L. Loh. "Generalization error bounds for noisy, iterative algorithms". In: *2018 IEEE International Symposium on Information Theory (ISIT)*. 2018, pp. 546–550.

[25] B. Poole, S. Ozair, A. v. d. Oord, A. A. Alemi, and G. Tucker. "On variational bounds of mutual information" (2019). arXiv: 1905.06922.

[26] M. Raginsky, A. Rakhlin, and M. Telgarsky. "Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis". In: *Proc. Conference on Learning Theory (COLT)*. 2017. arXiv: 1702.03849.

[27] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu. "Information-theoretic analysis of stability and bias of learning algorithms". In: *2016 IEEE Information Theory Workshop (ITW)*. IEEE. 2016, pp. 26–30.

[28] 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. 9214–9224.

[29] D. Russo and J. Zou. *How much does your data exploration overfit? Controlling bias via information usage*. 2015. arXiv: 1511.05219.

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

[31] J. Shawe-Taylor and R. C. Williamson. "A PAC analysis of a Bayesian estimator". In: *Proceedings of the tenth annual conference on Computational learning theory*. ACM. 1997, pp. 2–9.

[32] V. Thomas, F. Pedregosa, B. van Merriënboer, P.-A. Mangazol, Y. Bengio, and N. L. Roux. "Information matrices and generalization". *arXiv preprint arXiv:1906.07774* (2019).

[33] Y.-X. Wang, J. Lei, and S. E. Fienberg. "On-Average KL-Privacy and Its Equivalence to Generalization for Max-Entropy Mechanisms". In: *Privacy in Statistical Databases*. Ed. by J. Domingo-Ferrer and M. Pejić-Bach. Cham: Springer International Publishing, 2016, pp. 121–134. ISBN: 978-3-319-45381-1.

[34] M. Welling and Y. W. Teh. "Bayesian learning via stochastic gradient Langevin dynamics". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 681–688.

[35] A. Xu and M. Raginsky. "Information-theoretic analysis of generalization capability of learning algorithms". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2524–2533.

A Common Definitions

In this appendix, we collect together a few standard definitions from information theory. Let P, Q be probability measures on a common measurable space. Write $Q \ll P$ when Q is absolutely continuous with respect to P, i.e., for all measurable subsets A, $Q(A) = 0$ if $P(A) = 0$. By the Radon–Nikodym theorem, when $Q \ll P$, there exists a measurable function $\frac{\mathrm{d}Q}{\mathrm{d}P}$, called a Radon–Nikodym derivative or density, such that $Q(A) = \int_A \frac{\mathrm{d}Q}{\mathrm{d}P} \mathrm{d}P$ for all measurable subsets A. The **KL divergence** (or **relative entropy**) **of Q with respect to P**, written $\mathrm{KL}(Q \,\|\, P)$, is defined to be $\int \log \frac{\mathrm{d}Q}{\mathrm{d}P} \mathrm{d}Q$ when $Q \ll P$ and is defined to be infinity otherwise.

Given random elements X and Y, the **mutual information between X and Y**, written $I(X;Y)$ is

$$I(X;Y) = \mathrm{KL}(\mathbb{P}[(X,Y)] \,\|\, \mathbb{P}[X] \otimes \mathbb{P}[Y]),$$

where $\otimes$ forms the product measure. Given another random element Z, the **conditional mutual information between X and Y given Z** is defined to be $I(X;Y|Z) = I(X;(Y,Z)) - I(X;Z) = I((X,Z);Y) - I(Z;Y)$.

Relative entropy and mutual information satisfy many well-known properties: For example, relative entropy and mutual information are nonnegative; $X \perp\!\!\!\perp Y \iff I(X;Y) = 0$; and $I(X;Y) \leq I(X;(Y,Z))$. From this last inequality, one may deduce that $I(X;Y) \leq I(X;Y|Z)$ when $X \perp\!\!\!\perp Z$.

B Proofs of Results

B.1 Bounding Mutual Information by KL Divergence

The following is a well-known result that allows one to bound mutual information by the expectation of the KL divergence of a "posterior" with respect to a"prior" (where these terms are taken to have their more general interpretation from PAC-Bayesian theory, as opposed to the classical Bayesian theory).

Proposition B.1 (Variational Representation of Mutual Information)**.** *Let X and Y be random elements. Then, for all probability measures P on the same space as Y,*

$$I(X;Y) \leq \mathbb{E}[\mathrm{KL}(\mathbb{P}^X[Y] \,\|\, P)],$$

with equality for $P = \mathbb{E}[\mathbb{P}^X[Y]] = \mathbb{P}[Y]$.

The result is implicit in [18] and is considered folklore in the literature (e.g., it is referenced without proof in [7]). For a simple derivation, see [25, Eq. (1)]. Given another random element Z, it follows immediately by the disintegration theorem [17, Thm. 6.4] that, for all Z-measurable random probability measures P on the same space as Y,

$$I^Z(X;Y) \leq \mathbb{E}^Z[\mathrm{KL}(\mathbb{P}^{X,Z}[Y] \,\|\, P)] \text{ a.s.},$$

with a.s. equality for $P = \mathbb{E}^Z[\mathbb{P}^{X,Z}[Y]] = \mathbb{P}^Z[Y]$.

B.2 Proofs of Main Results

Proof of Theorem 2.3. Let $\tilde{W}$ be a random element in $\mathscr{W}$ such that $W \stackrel{\mathrm{d}}{=} \tilde{W}$ and $\tilde{W} \perp\!\!\!\perp S_J^c$. Let $\mathscr{G}$ denote the class of all functions g such that $\mathbb{E}\exp(g(\tilde{W}, S_J^c)) < \infty$. Then

$$I(W;S_J^c) = \mathrm{KL}(\mathbb{P}(W,S_J^c) \,\|\, \mathbb{P}(\tilde{W},S_J^c)) \tag{9}$$

$$= \sup_{g \in \mathscr{G}} \mathbb{E}g(W,S_J^c) - \log \mathbb{E}e^{g(\tilde{W},S_J^c)} \tag{10}$$

where the second equality follows from the Donsker–Varadhan variational formula [5, Prop. 4.15] (see also [8]). Let $f(w,s) = R_{\mathscr{D}}(w) - \hat{R}_s(w)$ so that $\mathbb{E}f(W,S_J^c) = \mathbb{E}R_{\mathscr{D}}(W) - \mathbb{E}\hat{R}_{S_J^c}(W)$ and $\mathbb{E}f(\tilde{W},S_J^c) = 0$. Let ψ be the cumulant generating function of $f(\tilde{W},S_J^c)$ and let D be the domain on which this cumulant generating function is defined. Then $\lambda f \in \mathscr{G}$ exactly when $\lambda \in D$. Then, for

every $\lambda \in D$,

$$\sup_{g \in \mathscr{G}} \mathbb{E} g(W, S_J^c) - \log \mathbb{E} e^{g(\tilde{W}, S_J^c)} \geq \lambda \mathbb{E} f(W, S_J^c) - \log \mathbb{E} e^{\lambda f(\tilde{W}, S_J^c)} \tag{11}$$

$$= \lambda \mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_{S_J^c}(W)\right] - \psi(\lambda). \tag{12}$$

By rearranging and optimizing over λ, we find that

$$\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_{S_J^c}(W)\right] \leq \inf_{\lambda \in D} \frac{\psi(\lambda) + I(W; S_J^c)}{\lambda}.$$

Because the subset J is random and independent of (S, W), we have $\mathbb{E}\hat{R}_{S_J^c}(W) = \mathbb{E}\hat{R}_S(W)$. Hence,

$$\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right] = \mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_{S_J^c}(W)\right] \leq \inf_{\lambda \in D}\left[\frac{\psi(\lambda) + I(W; S_J^c)}{\lambda}\right].$$

At this point we have established a slightly more abstract result that permits applications beyond the subgaussian case. By the subgaussian hypothesis, $f(w, S_J^c)$ is itself σ_{n-m}-subgaussian for each $w \in \mathscr{W}$, and so the bound above reduces to

$$\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right] \leq \sqrt{2\sigma_{n-m}^2 I(W; S_J^c)}$$

using the same optimization argument as in [6], [35], etc. From the proof of Theorem C.1, $\sigma_{n-m} \leq \frac{\sigma}{\sqrt{n-m}}$, completing the proof. □

Proof of Theorem 2.4. Let $\tilde{W}$ be a random element in $\mathscr{W}$ such that $(W, S_J, U) \overset{\mathrm{d}}{=} (\tilde{W}, S_J, U)$ and $\tilde{W} \perp\!\!\!\perp S_J^c \mid \{S_J, U\}$. Let Q and P satisfy $Q(S_J, U) = \mathbb{P}^{S_J, U}[W, S_J^c]$ and $P(S_J, U) = \mathbb{P}^{S_J, U}[\tilde{W}, S_J^c]$ a.s. By the Donsker–Varadhan variational formula [5, Prop. 4.15] and the disintegration theorem [17, Thm. 6.4], with probability one, for all measurable functions g such that $P(S_J, U)(\exp g) < \infty$,

$$\begin{aligned} I^{S_J, U}(W; S_J^c) &= \mathrm{KL}(Q(S_J, U) \,\|\, P(S_J, U)) \\ &\leq Q(S_J, U)(g) - \log P(S_J, U)(\exp g). \end{aligned}$$

Let $f(w, s) = R_{\mathscr{D}}(w) - \hat{R}_s(w)$. Note that, a.s., $P(S_J, U)(f) = \mathbb{E}^{S_J, U}[f(\tilde{W}, S_J^c)] = 0$ and

$$Q(S_J, U)(f) = \mathbb{E}^{S_J, U}[f(W, S_J^c)] = \mathbb{E}^{S_J, U}[R_{\mathscr{D}}(W) - \hat{R}_{S_J^c}(W)].$$

Let ψ be the cumulant generating function of $P(S_J, U)$, i.e., $\psi(\lambda; S_J, U) = \log P(S_J, U)(\exp\{\lambda f\})$. Let $D(S_J, U) = \{\lambda \in \mathbb{R} : \psi(\lambda; S_J, U) < \infty\}$. Then, with probability one, for all $\lambda \in D(S_J, U)$,

$$I^{S_J, U}(W; S_J^c) \geq \lambda \mathbb{E}^{S_J, U}\left[R_{\mathscr{D}}(W) - \hat{R}_{S_J^c}(W)\right] - \psi(\lambda; S_J, U).$$

Rearranging, with probability one,

$$\mathbb{E}^{S_J, U}\left[R_{\mathscr{D}}(W) - \hat{R}_{S_J^c}(W)\right] \leq \inf_{\lambda \in D(S_J, U)} \frac{I^{S_J, U}(W; S_J^c) + \psi(\lambda; S_J, U)}{\lambda}.$$

Because $W \perp\!\!\!\perp J$ and the subset J is random and uniformly distributed,

$$\begin{aligned} \mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right] &= \mathbb{E}\, \mathbb{E}^{S_J, U}\left[R_{\mathscr{D}}(W) - \hat{R}_{S_J^c}(W)\right] \\ &\leq \mathbb{E}\left[\inf_{\lambda \in D(S_J, U)} \frac{I^{S_J, U}(W; S_J^c) + \psi(\lambda; S_J, U)}{\lambda}\right]. \end{aligned}$$

At this point we have established a slightly more abstract result that permits applications beyond the subgaussian case. By the subgaussian hypothesis, $f(w, S_J^c)$ is itself σ_{n-m}-subgaussian for each $w \in \mathscr{W}$, and so the bound above reduces to

$$\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right] \leq \mathbb{E}\sqrt{2\sigma_{n-m}^2 I^{S_J, U}(W; S_J^c)}$$

using the same optimization argument as in [6], [35], etc. From the proof of Theorem C.1, $\sigma_{n-m} \leq \frac{\sigma}{\sqrt{n-m}}$, completing the proof. □

Proof of Theorem 2.5. For any two random measures $P(S_J,U), Q(S,U)$, the Donsker–Varadhan variational formula [5, Prop. 4.15] and the disintegration theorem [17, Thm. 6.4], give that with probability one

$$\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U)) \geq \sup_{g\in\mathscr{G}} \left(Q(S,U)(g) - P(S_J,U)(g) - \log\left[P(S_J,U)\left(\exp(g - P(S_J,U)(g))\right)\right]\right),$$

where $\mathscr{G}(S_J,U) = \{g : P(S_J,U)(\exp g) < \infty\}$.

Taking $g(w) = \lambda\left(R_{\mathscr{D}}(w) - \hat{R}_{S_J^c}(w)\right)$, and letting

$$\begin{aligned} R_{\mathscr{D}}(Q) &= Q(S,U)(R_{\mathscr{D}}) & R_{\mathscr{D}}(P) &= P(S_J,U)(R_{\mathscr{D}}) \\ \hat{R}_{S_J^c}(Q) &= Q(S,U)(\hat{R}_{S_J^c}) & \hat{R}_{S_J^c}(P) &= P(S_J,U)(\hat{R}_{S_J^c}) \end{aligned}$$

where, for brevity, we have used the short hand $Q = Q(S,U)$ and $P = P(S_J,U)$. Then, with probability one

$$\begin{aligned} &\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U)) \\ &\quad\geq \lambda\left(R_{\mathscr{D}}(Q) - \hat{R}_{S_J^c}(Q) - \left(R_{\mathscr{D}}(P) - \hat{R}_{S_J^c}(P)\right)\right) \\ &\qquad - \log\left[P(S_J,U)\left(\exp\left(\lambda\left(R_{\mathscr{D}} - \hat{R}_{S_J^c} - \left(R_{\mathscr{D}}(P) - \hat{R}_{S_J^c}(P)\right)\right)\right)\right)\right] \end{aligned}$$

Let

$$\psi(\lambda; S,J,U) = \log\left[P(S_J,U)\left(\exp\left(\lambda\left(R_{\mathscr{D}} - \hat{R}_{S_J^c} - \left(R_{\mathscr{D}}(P) - \hat{R}_{S_J^c}(P)\right)\right)\right)\right)\right],$$

and $D(S,J,U) = \{\lambda \in \mathbb{R} : \psi(\lambda; S,J,U) < \infty\}$. With probability one

$$\left(R_{\mathscr{D}}(Q) - \hat{R}_{S_J^c}(Q) - \left(R_{\mathscr{D}}(P) - \hat{R}_{S_J^c}(P)\right)\right) \leq \inf_{\lambda\in D(S,J,U)} \frac{\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U)) + \psi(\lambda; S,J,U)}{\lambda}$$

Since $P(S_J,U)$ is independent of S_J^c then we have $\mathbb{E}^{S_J,J,U}\left[R_{\mathscr{D}}(P) - \hat{R}_{S_J^c}(P)\right] = 0$. Hence, by averaging over S_J^c (equivalently, taking the conditional expectation conditional on (S_J,J,U)) we have, with probability one

$$\begin{aligned} \mathbb{E}^{S_J,J,U}\left[R_{\mathscr{D}}(Q) - \hat{R}_{S_J^c}(Q)\right] &= \mathbb{E}^{S_J,J,U}\left[R_{\mathscr{D}}(Q) - \hat{R}_{S_J^c}(Q) - \left(R_{\mathscr{D}}(P) - \hat{R}_{S_J^c}(P)\right)\right] \\ &\leq \mathbb{E}^{S_J,J,U}\left[\inf_{\lambda\in D(S,J,U)} \frac{\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U)) + \psi(\lambda; S,J,U)}{\lambda}\right] \end{aligned}$$

Finally, by taking the full expectation, since $J \perp\!\!\!\perp Q(S,U)$ we get:

$$\mathbb{E}\left[R_{\mathscr{D}}(Q(S,U)) - \hat{R}_S(Q(S,U))\right] \leq \mathbb{E}\left[\inf_{\lambda>0} \frac{\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U)) + \psi_{S,J,U}(\lambda)}{\lambda}\right]$$

where the final $\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U))$ on the right hand side is between two random measures, and hence is a random variable depending on (S,J,U); and the expectation on the right hand side integrates over (S,J,U).

If, for $(V \mid S_J,U) \sim P(S_J,U)$ it is the case that $\left(R_{\mathscr{D}}(V) - \hat{R}_{S_J^c}(V)\right)$ is σ-subgaussian for any (S,J,U), then this can be optimized to get

$$\mathbb{E}\left[R_{\mathscr{D}}(Q(S,U)) - \hat{R}_S(Q(S,U))\right] \leq \mathbb{E}\sqrt{2\sigma^2\,\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U))}$$

When the loss is $[a_1,a_2]$-bounded then $R_{\mathscr{D}}(V) - \hat{R}_{S_J^c}(V)$ is $\frac{a_2-a_1}{2}$ subgaussian which completes the proof.

□

Remark B.2 (Why does Theorem 2.5 use a boundedness assumption instead of a subgaussian assumption?). Note that we needed the boundedness assumption because even if, for $Z \sim \mathscr{D}$, $\ell(Z,w)$ was subgaussian (uniformly in $w \in \mathscr{W}$) it may not be the case that for $(V \mid S_J, U) \sim P(S_J, U)$, $\left(R_{\mathscr{D}}(V) - \hat{R}_{S_J^c}(V)\right)$ is subgaussian. In contrast, in the proofs of Theorem 2.3, Theorem 2.4, and Theorem C.1 the expectations over S_J^c included in the definition of the required cumulant generating functions let us take advantage of the subgaussian property of $\ell(Z,w)$. ◁

Proof of Proposition 2.6.

$$\mathrm{KL}(Q_T \,\|\, P_T) \le \mathrm{KL}(Q_T \,\|\, P_T) + \mathbb{E}\mathrm{KL}(Q_{|T} \,\|\, P_{|T}) = \mathrm{KL}(Q \,\|\, P).$$

This tells us that the KL divergence between marginal distributions of the terminal parameter is upper bounded by the KL between the distributions of the full trajectories.

Assuming $Q_0 = P_0$, we may decompose $\mathrm{KL}(Q \,\|\, P)$ across iterations, obtaining

$$\mathrm{KL}(Q \,\|\, P) = \mathop{\mathbb{E}}_{W\sim Q}\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P}(W)\right] = \mathop{\mathbb{E}}_{W\sim Q}\left[\sum_{t=1}^{T} \log \frac{\mathrm{d}Q_{t|}}{\mathrm{d}P_{t|}}(W)\right] = \sum_{t=1}^{T} \mathbb{E}_{Q_{0:(t-1)}}[\mathrm{KL}(Q_{t|} \,\|\, P_{t|})]. \tag{13}$$

□

C Mutual Information Bound for Subgaussian Losses

Theorem C.1 (Xu and Raginsky's Theorem 1)**.** *Suppose that $\ell(w,Z)$ is σ-subgaussian when $Z \sim \mu$, for all $w \in \mathscr{W}$, Then*

$$\left|\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right]\right| \le \sqrt{\frac{2\sigma^2}{n} I(S;W)}$$

A proof of this result is found in [35]. However, one may use the arguments therein to establish the further conclusion that $\ell(W,Z)$ or $R_S(W)$ is also subgaussian, which is not generally true. In this section we briefly describe the flaw in that logic and provide a clarification of their proof under the same assumptions. [29] give a proof for discrete parameter spaces, which does not contain this flaw. While it is straightforward to cast their proof into measure-theoretic language, we give the details for completeness.

The discussion in [35] preceding the theorem asserts that if $f : \mathscr{W} \times \mathscr{S}$ is such that $f(w,S)$ is σ subgaussian for all $w \in \mathscr{W}$ and if $W \perp\!\!\!\perp S$ then $f(W,S)$ is σ-subgaussian. A simple counter example is given by $\mathscr{W} = \mathscr{S} = \mathbb{R}$, with $f(w,s) = w + s$, and $(W,S) \sim \text{Cauchy} \times N(0,1)$. In this case $f(w,S)$ is clearly 1-subgaussian for each $w \in \mathscr{W}$, while $f(W,S)$ does not even have bounded absolute first moment, let alone a moment generating function defined in any open ball about 0.

The main issue in the argument establishing subgaussianity of $f(W,S)$ is failing to properly use a version of the conditional variance formula (modified to apply for moment generating functions as opposed to variances). The intuition of the conditional variance formula is useful in reconciling the final result with our counterexample, but is not sufficient for a general proof as the subgaussian parameter is not generally a standard deviation. The conditional variance formula asserts that

$$\mathrm{Var}(f(W,S)) = \mathbb{E}\left[\mathrm{Var}^W f(W,S)\right] + \mathrm{Var}\left(\mathbb{E}^W f(W,S)\right).$$

The argument by which one would conclude that $f(W,S)$ is subgaussian only acknowledges the first term, thus assuming that the second term is 0 (which would only hold when $\mathbb{E}^W f(W,S)$ is a.s. constant in W).

More precisely, since we are working with subgaussian parameters instead of true standard deviations:

$$\begin{aligned}
&\log \mathbb{E}\exp(t(f(W,S) - \mathbb{E}f(W,S))) \\
&\quad = \log \mathbb{E}\left[\exp(t(\mathbb{E}^W f(W,S) - \mathbb{E}f(W,S)))\mathbb{E}^W \exp(t(f(W,S) - \mathbb{E}^W f(W,S)))\right] \\
&\quad \le \log \exp(t^2\sigma^2/2)\mathbb{E}[\exp(t(\mathbb{E}^W f(W,S) - \mathbb{E}f(W,S)))] \\
&\quad = t^2\sigma^2/2 + \log \mathbb{E}[\exp(t(\mathbb{E}^W f(W,S) - \mathbb{E}f(W,S)))]
\end{aligned}$$

The RHS is $\geq t^2\sigma^2/2$ with equality if and only if $(\mathbb{E}^W f(W,S) - \mathbb{E}f(W,S))$ is constant (by Jensen' inequality). The first inequality is an equality when $f(w,S)$ is normal with variance σ^2 for all $w \in \mathscr{W}$.

Ergo, the assertion that $f(W,S)$ is σ-subgaussian holds exactly when $(\mathbb{E}_S f(W,S) - \mathbb{E}f(W,S))$ is constant. This situation is not generally of interest in learning theory; this amounts to saying that all parameter vectors lead to the same expected generalization error, and hence there is no purpose to learning from the data!

The final result is, of course, still valid and may be proven directly via the Donsker–Varadhan variational formula.

Proof. As in [35] we will leverage the fact that for each $w \in \mathscr{W}$, $f(w,S) = \frac{1}{n}\sum_{i=1}^n \ell(w,Z_i)$ is $\tau = \sigma/\sqrt{n}$ subgaussian, *however these variable may have different means for each value of* w. Let $\check{f}(w,s) = f(w,s) - \mathbb{E}f(w,S)$.

By Donsker–Varadhan and the fact that $\mathbb{E}^W \check{f}(\bar{W},\bar{S}) = 0$ a.s.,

$$\begin{aligned} I(W;S) &\geq \mathbb{E}\lambda \check{f}(W,S) - \log \mathbb{E}\exp(\lambda \check{f}(\bar{W},\bar{S})) \\ &\geq \mathbb{E}\lambda \check{f}(W,S) - \log \mathbb{E}\mathbb{E}^W \exp(\lambda \check{f}(\bar{W},\bar{S})) \\ &\geq \lambda \mathbb{E}\check{f}(W,S) - \log \mathbb{E}\exp(\lambda^2\tau^2/2) \\ &\geq \lambda \mathbb{E}\check{f}(W,S) - \lambda^2\tau^2/2. \end{aligned}$$

Optimizing over λ now yields the desired result, because

$$|\mathbb{E}\check{f}(W,S)| = |\mathbb{E}\left[f(W,S) - \mathbb{E}^W f(W,\bar{S})\right]| = |\mathbb{E}\left[R_{\mathscr{D}}(W) - \hat{R}_S(W)\right]|\,.$$

□

D Properties of the Hypergeometric Distribution and of Finite Population Variances

In this section, we enumerate a number of well-known results, and also derive some particular ones for our application.

D.1 Properties of the Hypergeometric Distribution

Let $n,m,b \in \mathbb{N}$, $m,b \leq n$. Write $B \sim \mathrm{HG}(n,m,b)$ when

$$\mathbb{P}(B=j) = \frac{\binom{m}{j}\binom{n-m}{b-j}}{\binom{n}{b}}, \quad j \in \{0 \vee b+m-n, \ldots, n \wedge m\}.$$

It follows that

$$\mathbb{E}(B) = b\frac{m}{n} \qquad \mathrm{Var}(B) \qquad = b\frac{m}{n}\frac{n-m}{n}\frac{n-b}{n-1} \leq b\frac{m(n-m)}{n^2}$$

D.2 Finite Population Statistics with Disjoint Samples

In this section we compute the covariance of the sample means for each population, and provide a formula for the variance of a linear combination of the two estimators.

Lemma D.1 (Variance for disjoint finite population statistics)**.** *Suppose that there is a finite population of size, N, $S = (y_1, \ldots, y_N)$. Consider two disjoint subsets of sizes n_1 and n_2 are chosen uniformly at random from S. Let $\bar{Y}_i$ be the sample mean on the ith sample. Let Σ be the population variance matrix. Then*

$$Var\begin{pmatrix}\bar{Y}_1 \\ \bar{Y}_2\end{pmatrix} = \frac{1}{N-1}\begin{bmatrix}(N-n_1)/n_1 & -1 \\ -1 & (N-n_2)/n_2\end{bmatrix} \otimes \Sigma$$

$$Var(a\bar{Y}_1 - b\bar{Y}_2) = \frac{1}{(N-1)}\left(-(a-b)^2 + N(a^2/n_1 + b^2/n_2)\right)\Sigma$$

Proof. Let ζ_i be an indicator for whether y_i appears in the first sample, and let W_i be an indicator for whether y_i appears in the second sample.

Let $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ and let $\Sigma = \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)(y_i-\mu)'$

Then for any $i \neq j$:

$$\zeta_i \sim \text{Ber}(n_1/N) \qquad W_i \sim \text{Ber}(n_2/N)$$
$$\text{Var}(\zeta_i) = \frac{n_1(N-n_1)}{N^2} \qquad \text{Var}(W_i) = \frac{n_2(N-n_2)}{N^2}$$

$$\begin{aligned}\text{Cov}(\zeta_i,\zeta_j) &= \mathbb{E}[\zeta_i\zeta_j] - \frac{n_1^2}{N^2} \\ &= \mathbb{P}[\zeta_i=\zeta_j=1] - \frac{n_1^2}{N^2} \\ &= \frac{n_1(n_1-1)}{N(N-1)} - \frac{n_1^2}{N^2} \\ &= -\frac{n_1}{N}\left(1-\frac{n_1}{N}\right)\frac{1}{N-1}\end{aligned}$$

$$\begin{aligned}\text{Cov}(W_i,W_j) &= \mathbb{E}[W_iW_j] - \frac{n_2^2}{N^2} \\ &= \text{PP}[W_i=W_j=1] - \frac{n_2^2}{N^2} \\ &= \frac{n_2(n_2-1)}{N(N-1)} - \frac{n_2^2}{N^2} \\ &= -\frac{n_2}{N}\left(1-\frac{n_2}{N}\right)\frac{1}{N-1}\end{aligned}$$

$$\begin{aligned}\text{Cov}(\zeta_i,W_i) &= \mathbb{E}[\zeta_iW_i] - \frac{n_1n_2}{N^2} \\ &= \mathbb{P}[\zeta_i=W_j=1] - \frac{n_1n_2}{N^2} \\ &= 0 - \frac{n_1n_2}{N^2} \\ &= -\frac{n_1n_2}{N^2}\end{aligned}$$

$$\begin{aligned}\text{Cov}(\zeta_i,W_j) &= \mathbb{E}[\zeta_iW_j] - \frac{n_1n_2}{N^2} \\ &= \text{PP}[\zeta_i=W_j=1] - \frac{n_1n_2}{N^2} \\ &= \frac{n_1n_2}{N(N-1)} - \frac{n_1n_2}{N^2} \\ &= \frac{n_1n_2}{N^2(N-1)}.\end{aligned}$$

$$(\bar{Y}_1,\bar{Y}_2) = \sum_{i=1}^{N} y_i(\zeta_i/n_1, W_i/n_2)$$

$$\begin{aligned}\text{Var}(\bar{Y}_1) &= \text{Var}\left(\sum_{i=1}^{N}\frac{y_i}{n_1}\zeta_i\right) \\ &= \frac{1}{n_1^2}\left(\sum_{i=1}^{N} y_iy_i'\frac{n_1(N-n_1)}{N^2} - \sum_{i\neq j} y_iy_j'\frac{n_1(N-n_1)}{N^2(N-1)}\right) \\ &= \frac{(N-n_1)}{n_1N^2}\left(\sum_{i=1}^{N} y_iy_i' - \sum_{i\neq j} y_iy_j'\frac{1}{N-1}\right) \\ &= \frac{(N-n_1)}{n_1(N-1)N}\sum_{i=1}^{N}(y_i-\mu)(y_i-\mu)' \\ &= \frac{(N-n_1)}{n_1(N-1)}\Sigma\end{aligned}$$

Similarly

$$\text{Var}(\bar{Y}_2) = \frac{(N-n_2)}{n_2(N-1)}\Sigma$$

Now, for the less well known part:

$$\begin{aligned}
\mathrm{Cov}(\bar{Y}_1, \bar{Y}_2) &= \mathrm{Cov}\left(\sum_{i=1}^{N} \frac{y_i}{n_1}\zeta_i, \sum_{i=1}^{N} \frac{y_i}{n_2} W_i\right) \\
&= \sum_{i=1}^{N} \frac{y_i y_i'}{n_1 n_2}\mathrm{Cov}(\zeta_i, W_i) + \sum_{i\neq j} \frac{y_i y_j'}{n_1 n_2}\mathrm{Cov}(\zeta_i, W_j) \\
&= -\sum_{i=1}^{N} \frac{y_i y_i'}{n_1 n_2}\frac{n_1 n_2}{N^2} + \sum_{i\neq j} \frac{y_i y_j'}{n_1 n_2}\frac{n_1 n_2}{N^2(N-1)} \\
&= -\frac{1}{N^2}\left(\sum_{i=1}^{N} y_i y_i' - \sum_{i\neq j} y_i y_j' \frac{1}{N-1}\right) \\
&= -\frac{1}{N-1}\Sigma
\end{aligned}$$

Hence

$$\mathrm{Var}\begin{pmatrix}\bar{Y}_1 \\ \bar{Y}_2\end{pmatrix} = \frac{1}{N-1}\begin{bmatrix}(N-n_1)/n_1 & -1 \\ -1 & (N-n_2)/n_2\end{bmatrix} \otimes \Sigma$$

For our application we need $\mathrm{Var}(a\bar{Y}_1 - b\bar{Y}_2)$:

$$\begin{aligned}
\mathrm{Var}(a\bar{Y}_1 - b\bar{Y}_2) &= a^2\frac{(N-n_1)}{n_1(N-1)}\Sigma + b^2\frac{(N-n_2)}{n_2(N-1)}\Sigma + 2ab\frac{1}{N-1}\Sigma \\
&= \frac{1}{(N-1)}\left(a^2\frac{N-n_1}{n_1} + 2ab + b^2\frac{N-n_2}{n_2}\right)\Sigma \\
&= \frac{1}{(N-1)}\left(-(a-b)^2 + N(a^2/n_1 + b^2/n_2)\right)\Sigma
\end{aligned}$$

□

Lemma D.2 (Bounding $\mathbb{E}\mathbb{E}^{S_J,J,U}\|\xi_t\|_2^2$ for SGLD)**.** *In the setting of Section 3.1*

$$\mathbb{E}\mathbb{E}^{S_J,J,U}\|\xi_t\|_2^2 = \frac{n(n-m)}{(n-1)^2 b_t}\left(1 + \frac{b_t}{n}\frac{n-m-1}{m}\right)\mathbb{E}[\hat{\Sigma}_t(S)]$$

Proof. Applying the conditional variance formula gives:

$$\begin{aligned}
\mathbb{E}\mathbb{E}^{S_J,J,U}\|\xi_t\|_2^2 &= \mathbb{E}\mathrm{Var}^{S,W_t}(\mathbb{E}^{b_t,W_t,S}[\xi_t]) + \mathbb{E}\mathbb{E}^{S,W_t}[\mathrm{Var}^{b_t,W_t,S}(\xi_t)] \\
&= 0 + \mathbb{E}\mathbb{E}^{S,W_t}\left[\mathrm{Var}^{b_t,W_t,S}\left(\frac{b_t^c}{b_t}\nabla\tilde{R}_{S_t^c}(W_t) - \frac{b_t^c}{b_t}\nabla\tilde{R}_{S_J}(W_t)\right)\right] \\
&= \mathbb{E}\mathbb{E}^{S,W_t}\left[\frac{(b_t^c)^2}{b_t^2}\mathrm{Var}^{b_t,W_t,S}\left(\nabla\tilde{R}_{S_t^c}(W_t) - \nabla\tilde{R}_{S_J}(W_t)\right)\right]
\end{aligned}$$

Applying Lemma D.1 further yields

$$
\begin{aligned}
&\mathbb{E}^{S,W_t}\left[\frac{(b_t^c)^2}{b_t^2}\mathrm{Var}^{b_t,W_t}\left(\nabla\tilde{R}_{S_t^c}(W_t)-\nabla\tilde{R}_{S_J}(W_t)\right)\right] \\
&\quad=\mathbb{E}^{S,W_t}\left[\frac{(b_t^c)^2}{b_t^2}\frac{1}{(n-1)}\left(\frac{n}{b_t^c}+\frac{n}{m}\right)\hat{\Sigma}_t(S)\right] \\
&\quad=\frac{n}{(n-1)b_t^2}\mathbb{E}^{S,W_t}\left[b_t^c+(b_t^c)^2\frac{1}{m}\right]\mathbb{E}^{S,W_t}[\hat{\Sigma}_t(S)] \\
&\quad=\frac{n}{(n-1)b_t^2}\left(b_t\frac{n-m}{n}+\left(\frac{(n-m)^2}{n^2}b_t^2+b_t\frac{m}{n}\frac{n-m}{n}\frac{n-b_t}{n-1}\right)\frac{1}{m}\right)\mathbb{E}^{S,W_t}[\hat{\Sigma}_t(S)] \\
&\quad=\frac{n}{(n-1)b_t^2}\left(b_t\frac{n-m}{n-1}+b_t^2\frac{(n-m)(n-m-1)}{n(n-1)m}\right)\mathbb{E}^{S,W_t}[\hat{\Sigma}_t(S)] \\
&\quad=\frac{n}{(n-1)b_t^2}\left(b_t\frac{n-m}{n-1}+b_t^2\frac{(n-m)(n-m-1)}{n(n-1)m}\right)\mathbb{E}^{S,W_t}[\hat{\Sigma}_t(S)] \\
&\quad=\frac{n(n-m)}{(n-1)^2b_t}\left(1+\frac{b_t}{n}\frac{n-m-1}{m}\right)\mathbb{E}^{S,W_t}[\hat{\Sigma}_t(S)]
\end{aligned}
$$

□

E Asymptotic Results

E.1 Langevin Dynamics

In this section we continue from the end of Section 3.2.1. Under the assumption that $\tilde{\ell}$ is L-Lispchitz (the same assumption as in [6]) we have the following results which portray the asymptotic behavior of the expected generalization error of the Langevin diffusion algorithm for ℓ being the 0-1 loss (which is $1/2$-subgaussian):

$$
\mathop{\mathbb{E}}_{W_T\sim Q_T}\left(R_{\mathscr{D}}(W_T)-R_S(W_T)\right)\leq\frac{L}{2(n-1)}\sqrt{\sum_{t=1}^{T}\beta_t\eta_t}
$$

E.1.1 Geometrically Decaying Learning Rate

Under an assumption of L-Lipschitz loss and geometrically decaying learning rate and a temperature that ramps up to a polynomial in n ($\eta_t=\eta_0\rho^t$ for $0<\rho<1$ and that $\beta_t=\beta_0(n-1)^\theta(1-\nu^t)$ for some $0<\theta<1$) then we have the following bound:

$$
\sup_{T\geq 0}\left[\mathop{\mathbb{E}}_{W_T\sim Q_T}\left(R_{\mathscr{D}}(W_T)-R_S(W_T)\right)\right]\leq\frac{L}{2(n-1)^{1-\theta}}\sqrt{\beta_0\eta_0\frac{\rho(1-\nu)}{(1-\rho)(1-\rho\nu)}}
$$

E.1.2 Polynomial Decaying Learning Rate

Under an assumption of L-Lipschitz loss and polynomial decaying learning rate and temperature that is polynomial in n ($\eta_t=\eta_0 t^{-\alpha}$ for $\alpha>0$ and that $\beta_t=\beta_0(n-1)^p$ for some $0<p<1$) then we have the following bound:

$$
\left[\mathop{\mathbb{E}}_{W_T\sim Q_T}\left(R_{\mathscr{D}}(W_T)-R_S(W_T)\right)\right]\leq\begin{cases}\frac{L}{2(n-1)^{1-p}}\sqrt{1+\frac{1}{\alpha-1}T^{1-\alpha}} & \alpha<1\\ \frac{L}{2(n-1)^{1-p}}\sqrt{1+\log(T)} & \alpha=1\\ \frac{L\alpha}{2(n-1)^{1-p}(\alpha-1)} & \alpha>1\end{cases}
$$

F Comparing Theorems 2.3 to 2.5 when $m=n-1$

Let $V\sim P(S_J,U)$, $W\sim Q(S,U)$, and $\tilde{W}\sim Q(S,U)$ independently of W. In the case of $[a_1,a_2]$-bounded loss, $\left(R_{\mathscr{D}}(V)-\hat{R}_{S\setminus S}(V)\right)$ is $(a_2-a_1)/2$-subgaussian, so that Theorem 2.5 yields:

$$
\mathbb{E}\left[R_{\mathscr{D}}(W)-\hat{R}_S(\tilde{W})\right]\leq\mathbb{E}\sqrt{(a_2-a_1)^2\,\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U))\,/2}.
$$

Using KL divergence based upper bounds for mutual information (Proposition B.1), Theorem 2.4 gives us

$$\mathbb{E}\left[R_{\mathscr{D}}(W)-\hat{R}_S(W)\right] \leq \mathbb{E}\sqrt{(a_2-a_1)^2\mathbb{E}^{S_J,U}[\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U))]\,/2},$$

while Theorem 2.3 yields:

$$\mathbb{E}\left[R_{\mathscr{D}}(W)-\hat{R}_S(W)\right] \leq \sqrt{(a_2-a_1)^2\mathbb{E}\left[\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U))\right]/2}$$

for $m=n-1$, the bounds are ranked as 2.3 $\geq$ 2.4 $\geq$ 2.5 (by Jensen's inequality for each conditional expectation being passed into $\sqrt{\cdot}$). When $\mathrm{KL}(Q(S)\,\|\,P(S_J))$ has a large variance then the difference can be quite material.

G An analytically tractable example

We present a simple analytic example, where our upper bound is a clear improvement over existing work when similar simplifications are performed. Let $S=\{z_1,\ldots,z_n\}\sim\mathscr{D}^n$ be a sample from the distribution $\mathscr{D}$ on $\mathbb{R}$. We wish to estimate the mean of $\mathscr{D}$, μ. We will use the loss function $\ell(z,w)=\tilde{\ell}(z,w)=(z-w)^2$ where $w\in\mathscr{W}=\mathbb{R}$. The distribution $\mathscr{D}$, is assumed to satisfy the sub-Gaussianity assumption in Theorems 2.3 and 2.5 for this loss. Upon specializing the SGLD update rule (7) to this setting:

$$W_{t+1}=W_t-\eta_t\frac{d}{dW_t}\tilde{R}_S(W_t)+\sqrt{\frac{2\eta_t}{\beta}}\varepsilon_t=\left(1-\frac{2\eta_t}{n}\right)W_t+\frac{2\eta_t}{n}\sum_{i=1}^{n}z_i+\sqrt{\frac{2\eta_t}{\beta}}\varepsilon_t. \tag{14}$$

We will apply the data-dependent generalization bound in Theorem 2.5 with $m=\#S_J=n-1$ and set $\{i^\star\}=J$. Since we are working with LD, we set the random variable U to a constant (trivial random variable). It follows that:

$$\mathrm{KL}(Q_{t+1|}(S)\,\|\,P_{t+1|}(S_J))=\frac{(\mu_{t+1}-\mu'_{t+1})^2}{4\eta_t/\beta}=\frac{\beta}{n^2}z_{i^\star}^2\eta_t. \tag{15}$$

Thus the expected generalization error is upper bounded by:

$$\mathbb{E}\sqrt{2\sigma^2 KL(Q_T(S)||P_T(S_J))}\leq\mathbb{E}\sqrt{2\sigma^2\frac{\beta}{n^2}z_{i^\star}^2\sum_{t=0}^{T-1}\eta_t}=\mathbb{E}[|z_i|]\left(\sqrt{2\sigma^2\frac{\beta}{n^2}\sum_{t=0}^{T-1}\eta_t}\right). \tag{16}$$

When one applies the results in [24, 35], the upper bounded on the generalization error can be shown to be:

$$\sqrt{\frac{2\sigma^2}{n}I(W_T;S)}\leq\sqrt{\frac{2\sigma^2}{n}\sum_{t=0}^{T-1}I(\bar{W}_{t+1};S|W_1^t)}\leq\sqrt{2\sigma^2\frac{\beta}{n^2}E[z_i^2]\sum_{t=0}^{T-1}\eta_t} \tag{17}$$

Comparing with (16) we see that this bound can be is larger since $E[|z_i|]\leq\sqrt{E[z_i^2]}$ from Jensen's inequality. The discrepancy can be made arbitrarily large based on the choice of $\mathscr{D}$.

H Experiment Details

The first architecture and dataset we consider is a three-layer multilayer perceptron (MLP), with 600 hidden units per hidden layer and rectified linear unit (ReLU) activation functions, trained on MNIST [19]. In Fig. 1a, we compare the bound for two amounts of held out data, $n-m=\#S_J^c$. We see that the empirical performance reflects our analytical results that the bound is tighter for large m. As can be inferred from Eq. (4), the difference between $\|\xi_t\|^2$ and $\|\nabla\tilde{R}_t\|^2$ increases with m.

The remainder of our experiments consider convolutional neural networks (CNNs). For MNIST and Fashion-MNIST, we use a standard network configuration with two convolutional layers (with 32 and 64 filters of size 5×5, respectively, followed by 2×2 max pooling after each convolutional layer), followed by two fully connected layers (1024 nodes each) with ReLU activations.

Our final experiment uses the CIFAR-10 dataset. The CNN architecture has two convolutional layers and three fully connected layers. Both convolutional layers use 64 filters of size 5×5. After each convolutional layer there is a 2×2 max pooling layer. Then, we have three fully connected layers with the number of neurons 384, 192, and 10 respectively.

H.1 Evaluation of the generalization bound

We estimate our generalization error bound and that of [22] using nested Monte Carlo simulations. We use the results of Theorem 3.1, specifically Eq. (6). In order to evaluate this bound we perform two Monte Carlo estimate: one for $\mathbb{E}^{S,J,U}\|\xi_t\|_2^2$, and then for the full expectation (outside of the $\sqrt{\cdot}$). For our bound, for each hyperparameter combination, we have used 10 simulations for the outer expectation, each using 10 simulations to estimate the inner expectation. For the generalization bound in Mou et al. [22] we have used their 100 simulations to evaluate the bound given by their Theorem 10.

H.2 Learning Rate and Inverse Temperatures for Figs. 1b and 1c

In Fig. 1b, we use

$$\beta_t^{(\text{high})} = 100 \times \max\{\exp\left(\frac{\mathrm{t}}{100}\right), 55000\} \tag{18}$$

$$\beta_t^{(\text{low})} = 100 \times \max\{\exp\left(\frac{\mathrm{t}}{100}\right), 5000\} \tag{19}$$

where t denotes the iteration number.

All other parameters are the same and are outlined in Table 2.

For the "high' inverse temperature schedule, at Iteration 5 the training error is 3.21% and the generalization error is 0.9%, while for the "low' inverse temperature schedule, at epoch 5 the training error is 5.18% and the generalization error is 0.16%.

In Fig. 1c we consider $\eta_t^{(\text{small})} = 8 \times 10^{-4} \times 0.96^{\left(\frac{\mathrm{t}}{2000}\right)}$ and $\eta_t^{(\text{large})} = 2 \times 10^{-3} \times 0.96^{\left(\frac{\mathrm{t}}{2000}\right)}$, and the rest of the parameters are the same and are outlined in Table 2. For the "small" learning rate, the training error and the test-set generalization error at Epoch 6 for the small learning rate scenario are 7.62% and 1.1% , respectively,; while for the "large" learning rate the training error and the test-set generalization error at Epoch 6 are 6.3% and 1.0%, respectively.

H.3 Hyperparameters of our experiments

In Tables 1 to 4, we provide the hyperparameter and training details of the experiments that were presented in Section 4.

Parameter	Values
Dataset	MNIST
Architecture	MLP with 3 hidden layers
Batch size	100
Learning rate	`learning rate=`8×10^{-3}`,decay steps=600, decay rate=0.95`
Beta schedule	$\min\{10 \times \exp(\text{iter}/400), 2000\}$
Number of epochs	15
Average Final training error	1.40%
Average Final test error	4.12%
# training examples	55000
Number of runs	50

Table 1: Details of Experiments reported in Fig. 1a for MNIST with MLP

I High Probability PAC-Bayes Bounds

We can leverage the methods used to provide bounds on the expected generalization error above to also derive high probability bounds for the generalization error. We will give an example of this here for completeness, though more work can be done to select a tighter bound from more recent literature and to tune the parameters available to optimize the bound further. For example, in our setting we could optimally tune the level of data dependence for the bound to be tightened. We will

Parameter	Values
Dataset	MNIST
Architecture	CNN with 2 conv. layers
Batch size	100
Learning rate	`learning rate=`4×10^{-3}`,decay steps=2000, decay rate=0.96`
Beta schedule	$\min\{10\times\exp(\text{iter}/100), 55000\}$
Number of epochs	15
Average Final training error	1.81%
Average Final test error	2.03%
# training examples	55000
Number of runs	50

Table 2: Details of Experiments reported in Figs. 1b to 1d for MNIST with CNN

Parameter	Values
Dataset	Fashion-MNIST
Architecture	CNN with 2 conv. layers
Batch size	100
Learning rate	`learning rate=`4×10^{-3}`,decay steps=3500, decay rate=0.93`
Beta schedule	$\min\{10\times\exp(\text{iter}/100), 55000\}$
Number of epochs	25
Average Final training error	8.3%
Average Final test error	10.83%
# training examples	60000
Number of runs	20

Table 3: Details of Experiments reported in Fig. 1e for Fashion-MNIST

make use of Shalev-Shwartz and Ben-David [30] (theorem 31.1 therein), which we state here under the notation and definitions of our work, and in the context of Section 3.1.

Proposition I.1 ([30] Theorem 31.1)**.** *Suppose that the loss function is* $[0,1]$*-bounded. Let* P *be any prior distribution. With probability at least* $(1-\delta)$ *(over the choice of* $S\sim\mathscr{D}^n$*) for any posterior distribution* Q *(even those depending on* S*) with* $W\sim Q$*,*

$$\mathbb{E}^S\left[R_{\mathscr{D}}(W_T)-\hat{R}_S(W_T)\right]\leq\sqrt{\frac{\mathrm{KL}(Q\,\|\,P)+\log(n/\delta)}{2(n-1)}}$$

In our setting P will be allowed to depend on m data points chosen uniformly at random, while Q will depend on the full dataset, so we can apply this result conditional on the subset upon which P depends. Therefore, for any $S_J\in\mathscr{Z}^m$ and any $U\in\mathscr{U}$ and for any kernel $P:\mathscr{Z}^n\times\mathscr{U}\to\mathscr{M}_1(\mathscr{W}^T)$ be any prior distribution which depends on S_J, with probability at least $(1-\delta)$ (over the choice of $S_J^c\sim\mathscr{D}^{n-m}$) for any posterior distribution Q (even those depending on S_J^c) with $W\sim Q$,

$$\begin{aligned}\mathbb{E}^S\left[R_{\mathscr{D}}(W_T)-\hat{R}_{S_J^c}(W_T)\right]&\leq\sqrt{\frac{\mathrm{KL}(Q(S,U)\,\|\,P(S_J,U))+\log((n-m)/\delta)}{2(n-m-1)}}\\&\leq\sqrt{\frac{\sum_{t=1}^T\frac{\beta_t\eta_t}{8}\mathbb{E}^{S,J,U}\|\xi_t\|_2^2+\log((n-m)/\delta)}{2(n-m-1)}}\end{aligned}$$

In the case of Langevin dynamics when using worst case, Lipschitz constant base upper bounds, this gives

$$\mathbb{E}^S\left[R_{\mathscr{D}}(W_T)-\hat{R}_{S_J^c}(W_T)\right]\leq\sqrt{\frac{\frac{L^2}{(n-1)(m-1)}\sum_{t=1}^T\frac{\beta_t\eta_t}{8}+\log((n-m)/\delta)}{2(n-m-1)}}$$

which provides a less trivial tradeoff between m and $n-m$ compared to the expected generalization error bound. One could further take expectations over U and/or J to get high probability bounds for the generalization error based on the full empirical loss.

Parameter	Values
Dataset	CIFAR-10
Architecture	CNN with 2 conv. layers
Batch size	200
Learning rate	`learning rate=`5×10^{-3}`,decay steps=2000, decay rate=0.95`
Beta schedule	$\min\{10 \times \exp(\text{iter}/100), 55000\}$
Number of epochs	50
Average Final training error	6.9%
Average Final test error	29.9%
$\lvert S_J \rvert$	`len(training_set)-1`
# training examples	50000
Number of runs	30

Table 4: Details of Experiments reported in Fig. 1f for CIFAR-10

We intend to investigate such bounds further in future work, and this section serves merely to illustrate the possibility and nature of such high-probability bounds based on data-dependent estimates of mutual information and data-dependent PAC-Bayes priors. We acknowledge that these are not the tightest such bounds possible.