A Review of Information-Theoretic Generalization Bounds for Stochastic Gradient Langevin Dynamics

Justin Kang
Department of Electrical and Computer Engineering
University of Toronto
10 King's College Road, Toronto, Ontario M5S3G4, Canada
js.kang@mail.utoronto.ca

***Abstract*—In this review, we study the progression of information-theoretic generalization bounds for Stochastic Gradient Langevin Dynamics (SGLD). SGLD is an important optimization algorithm with many applications in statistical learning. We discuss the formulation of SGLD and its applications. We review the first information-theoretic generalization bounds by Russo et al. and Xu et al. which apply to a more general class of learning algorithms, as well as newer work by Bu et al. on this subject. After surveying these fundamental works we conduct a review of the more specific works by Pensia et al. which focus on the class of iterative learning algorithms, of which SGLD is a part. We also review the work Haghifam, Negrea et al., which presents a new frontier in information theoretic bounds for SGLD by formulating the data-dependant estimation framework. Finally, we present a simple novel information-theoretic method to bound the generalization error of a particular formulation of SGLD with a square error loss function.**

I. Introduction

In recent years, the development of statistical learning algorithms has seen tremendous progress. The development of these algorithms has coincided with the availability of large amounts of data. This "big-data" contains information about complex and important systems such as the internet, human language and even biological systems. With such large datasets and powerful algorithms, important predictions can be made about the systems from which the data is drawn. Let us consider an algorithm which takes some data and produces a model. A key open problem in statistical learning theory is how to evaluate the efficacy of such a model on data which was not used to generate the model. This is important, because even if model predictions are consistent with the given data, it is still possible that the predictions will be inaccurate when compared against new data. When this occurs, we often say that the algorithm has produced a model which suffers from overfitting.

How a model performs on data outside of the input data is often referred to as the *generalization* of that model. In their book [1], Abu-Mostafa et al. state that the "holy grail" of statistical learning is an in-sample estimate of the out-of-sample error, i.e. the generalization error. This problem is by no means new and many different methods have been proposed for attempting to estimate generalization error over the years. One of the first methods was the study of Vapnik–Chervonenkis dimension of learning algorithms. Though the bounds associated with this analysis are useful, particular with simple learning algorithms, there has recently been a significant movement to find other ways to bound generalization error. In particular, information theory is a new tool being used. In this review paper, we look at the development of information theoretic generalization bounds for Stochastic Gradient Langevin Dymanics (SGLD), which is a type of learning algorithm. SGLD was first proposed in [2], as a way to mix stochastic gradient decent methods with Bayesian techniques. SGLD is described by an iterative process, with an update equation characterized by both additive white Gaussian noise and a decreasing step size η_t, as well as a term which depends on the gradient:

$$w_{t+1} = w_t - \eta_t \nabla_{w_t} R(w_t, \mathbf{Z}_t) + \sqrt{\frac{2\eta_t}{\beta}}\xi_t. \tag{1}$$

R is a risk function, w_t is an element of our hypothesis space, and $\mathbf{Z}_t$ represents data which is considered in the t^{th} time step. This update function is given more context in the following section. The conclusion of [2] is that in the initial phase the stochastic gradient noise will dominate and the algorithm effectively behaves like an efficient stochastic gradient descent algorithm, while in the later phase the injected noise ξ_t will dominate, so the algorithm will imitate a Langevin dynamics algorithm with a smooth transition between the two.

Since it's inception, SGLD has found practical application in an extremely wide variety of optimization and learning problems. For example [3] makes use of SGLD to study antibiotic resistance in bacteria. These kinds of applications in data analysis and predictive modeling makes the study of SGLD an important topic for research.

1) Notation: Throughout this paper, boldface symbols such as $\mathbf{S}$ denote sets. $\mathbb{E}_{X\sim\Lambda}[f(X)]$ denotes the expectation of $f(x)$ with respect to the argument X under the distribution Λ. The absolute value function is denoted by $|\cdot|$. $P(X|Y)$ is the conditional distribution of X given Y. $I(X;Y)$ is the mutual information between random variables X and Y. $h(X)$ represents the entropy of a continuous random variable X. Furthermore, we denote the KL-divergence as $\mathrm{KL}(P/\!/Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$.

II. Problem Setup and Preliminaries

Consider an unknown underlying data distribution Λ. We take n "measurements" $Z_i \sim \Lambda$ from this distribution, which

together make up the training set $\mathbf{S} = \{z_1, \ldots, z_n\}$. Each measurement z_i belongs to an *instance space* denoted $\mathcal{Z}$. The goal of statistical learning, is to estimate this distribution Λ, or some function thereof, using $\mathbf{S}$. We parameterize the potential candidate distributions by an element $w \in \mathcal{W}$. This set is referred to as the *hypothesis space*. Let $\ell(w, z)$ be a loss function. This loss function should in some sense represent the "compatibility" of the input measurement with the chosen distribution in the hypothesis set. The optimal w is chosen such that in expectation, the loss function is minimized:

$$\underset{w \in \mathcal{W}}{\text{minimize}} \quad \mathbb{E}_{Z\sim\Lambda}\left[\ell(w, Z)\right]. \tag{2}$$

The quantity $\mathbb{E}_{Z\sim\Lambda}\left[\ell(w, Z)\right]$ is often referred to as the *out-of-sample* error, or sometimes as the *risk*. However, this quantity is unknown to us as we attempt to select the parameter w. Thus, a common framework for learning is known as *empirical risk minimization*, where we seek to minimize:

$$L_\Lambda(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i), \tag{3}$$

which clearly approximates the out-of-sample error. A *learning algorithm* takes in some training set $\mathbf{S}$ and produces some output hypothesis $w \in \mathcal{W}$. In the empirical risk minimization scheme, this output w is the minimizer of the empirical risk in (3). We can view our learning algorithms as a channel, which is characterized by a conditional distribution $\mathbb{P}_{W|\mathbf{S}}$. In this review, we will look primarily at SGLD, however, some of the earliest results present bounds on a more general class of learning algorithms. We will define the generalization error as:

$$\operatorname{gen}\left(\Lambda, \mathbb{P}_{W|\mathbf{S}}\right) \triangleq \mathbb{E}_{\mathbf{S},W}\left[L_\Lambda(w) - \mathbb{E}_Z\left[\ell(w, Z)\right]\right]. \tag{4}$$

Another useful way to interpret the generalization error equation (4) is to consider it written the following way:

$$E_{out} = \mathbb{E}_Z\left[\ell(w, Z)\right] = \mathbb{E}_{\mathbf{S},W}\left[L_\Lambda(w)\right] + \operatorname{gen}\left(\Lambda, \mathbb{P}_{W|\mathbf{S}}\right). \tag{5}$$

In this form, we have separated out the quantity we are truly interested in, E_{out}. We can see that the out-of-sample error is a function of both the in-sample error and the generalization error term. Since, in general, both cannot be minimized simultaneously, it is important to understand the relationship between these two quantities.

Below we include some preliminary definitions required for the remainder of the paper:

Definition 1. *A random variable X is part of the set of sub-Gaussian random variables with variance proxy σ^2, i.e. $X \in \mathrm{SubG}(\sigma^2)$ if the following inequality holds:*

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \quad \forall \lambda \in \mathbb{R} \tag{6}$$

We will assume that $\ell(w, Z) \in \mathrm{SubG}(\sigma^2)$ with respect to $Z \sim \Lambda$, for every $w \in \mathcal{W}$. In particular, if Λ is Gaussian and $\ell(w, Z)$ is L-Lipschitz, then $\ell(w, Z)$ is known to be sub-Gaussian. Furthermore, if $\ell(w, Z)$ is bounded, it is also sub-Gaussian. Both of these results will be used in the papers we review.

In this review we will primarily be interested in $\mathbb{P}_{W|\mathbf{S}}$ given by the SGLD learning algorithm. In this algorithm, the output W_T is generated from the dataset $\mathbf{S}$. by the following series of T update equations:

$$w_{t+1} = w_t - \frac{\eta_t}{n}\sum_{i=1}^{n} \nabla_{w_t} \ell\left(w_t, z_i\right) + \sqrt{\frac{2\eta_t}{\beta}}\xi_t, \tag{7}$$

where $\xi_t \sim \mathcal{N}(0, 1)$.

III. Generalization Bounds for Probabilistic Learning Algorithms

The work on information theoretic generalization bounds began with [4], and was quickly followed by the work of Xu et. al., from which the following bound was derived.

Theorem 1. *[5] Suppose $\ell(w, Z) \in \mathrm{SubG}\left(\sigma^2\right)$, where $Z \sim \Lambda$. Then for any learning algorithm $\mathbb{P}_{W|\mathbf{S}}$, we have:*

$$\left|\operatorname{gen}\left(\Lambda, \mathbb{P}_{W|\mathbf{S}}\right)\right| \leq \sqrt{\frac{2\sigma^2}{n} I(W; \mathbf{S})}. \tag{8}$$

This was the first major result in information-theoretic generalization bounds for learning algorithms. This work was followed by proof of a tighter bound on generalization presented in [6].

Theorem 2. *[6] Suppose $\ell(w, Z) \in \mathrm{SubG}(\sigma^2)$ under $Z \sim \Lambda$ for all $w \in \mathcal{W}$, then*

$$\begin{aligned} |\operatorname{gen}\left(\mu, \mathbb{P}_{W|S}\right)| &\leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2\sigma^2 I\left(W; Z_i\right)} \\ &\leq \sqrt{\frac{2\sigma^2}{n} I(W; \mathbf{S})}. \end{aligned} \tag{9}$$

In Section V, we will use Theorem 2 to bound the generalization error for a particular instance of SGLD.

IV. Generalization Bounds for Iterative Algorithms

In this section, we review the contribution of [7], which specifically focuses on the generalization of stochastic iterative algorithms in the empirical risk minimization framework. We formally define a class of algorithms, *iterative learning algorithms*, for which their theory applies, below.

Definition 2. *An iterative learning algorithm $\mathbb{P}_{W|\mathbf{S}}$ is defined by T update equations which are given by:*

$$W_t = g\left(W_{t-1}\right) - \eta_t F\left(W_{t-1}, \mathbf{Z}_t\right) + \xi_t \quad \forall t \geq 1. \tag{10}$$

Let $W_0 \in \mathcal{W}$ be any starting point and let W_T be the final output of the iterative learning algorithm. At each step, the new W_t is chosen as some function of the previous end point, plus some data-dependant direction vector (typically the risk function gradient), which is scaled by η_t. Note that at each

step $\mathbf{Z}_t \subseteq \mathbf{S}$. *Finally the update is perturbed by noise* $\xi_t \sim N\left(0, \sigma_t^2 I_d\right)$,.

Note that this definition of iterative algorithms implies that

$$P\left(W_{t+1}|W_1, \ldots, W_t, \mathbf{Z}_1, \ldots, \mathbf{Z}_{t+1}, \mathbf{S}\right) = P\left(W_{t+1}|W_t, \mathbf{Z}_{t+1}\right) \quad (11)$$

Our definition of SGLD as given in (1), is consistent with this definition with the update function of $g(W_{t-1}) = W_{t-1}$ and $F(W_{t-1}, \mathbf{Z}_t) = \nabla_{W_{t-1}} R(W_{t-1}, \mathbf{Z})$. For data $\mathbf{Z}_t$, the empirical risk function R is defined to be $R(w, \mathbf{Z}_t) = \frac{1}{|\mathbf{Z}_t|} \sum_{z \in \mathbf{Z}_t} \ell(w, z)$. Since SGLD is of the form of the iterative algorithms defined in this paper, their results can be applied to SGLD.

Furthermore, throughout [7], the following assumptions are made.

Assumption 1. *The loss function satisfies:*

$$\ell(w, Z) \in \mathrm{SubG}\left(\sigma_z^2\right) \quad (12)$$

for $Z \sim \Lambda$ *for all* $w \in \mathcal{W}$.

Assumption 2. *The directional part of the update function,* F *is bounded:*

$$\sup_{w \in W, z \in Z} \|F(w, z)\|_2 \leq L, \quad L > 0. \quad (13)$$

Assumption 3. *The sampling strategy is agnostic to the previous iterates of the parameter vectors:*

$$P\left(\mathbf{Z}_{t+1}|\mathbf{Z}_1, \ldots, \mathbf{Z}_t, W_1, \ldots, W_t, \mathbf{S}\right) = P\left(\mathbf{Z}_{t+1}|\mathbf{Z}_1, \ldots, \mathbf{Z}_t, \mathbf{S}\right). \quad (14)$$

Combining Assumption 3, with (11), we note that the following condition holds:

$$P\left(W_{t+1}|W_1, \ldots, W_t, \mathbf{Z}_1, \ldots, \mathbf{Z}_T, \mathbf{S}\right) = P\left(W_{t+1}|W_t, \mathbf{Z}_{t+1}\right). \quad (15)$$

Now that we have stated all of the assumptions used in [7], we can state the main result.

Theorem 3. *For an iterative algorithm* $\mathbb{P}_{W|\mathbf{S}}$ *which satisfies Assumptions 1-3, the mutual information between the output* W *and the input data* $\mathbf{S}$ *satisfies:*

$$I(W; \mathbf{S}) \leq \sum_{t=1}^{T} \frac{d}{2} \log\left(1 + \frac{\eta_t^2 L^2}{d\sigma_t^2}\right). \quad (16)$$

In [7] a simple information theoretic proof is given, which is summarized below:

Proof. We begin by using chain rule:

$$\begin{aligned} I(\mathbf{S}; W) \leq & I\left(\mathbf{S}; W_1, \ldots, W_T\right) \\ \leq & I\left(\mathbf{Z}_1, \ldots, \mathbf{Z}_t; W_1, \ldots, W_T\right) \\ = & I\left(\mathbf{Z}_1, \ldots, \mathbf{Z}_T; W_1\right) + \ldots \\ & + I\left(\mathbf{Z}_1, \ldots, \mathbf{Z}_T; W_T|W_1, \ldots, W_{T-1}\right) \end{aligned} \quad (17)$$

Each of the individual terms in the final equation of (17) can be computed based on the previously stated assumptions. This is done by breaking up each term in the sum to be

$$\begin{aligned} I\left(\mathbf{Z}_1, \ldots, \mathbf{Z}_T; W_t|W_1, \ldots, W_{t-1}\right) &= I(W_t; \mathbf{Z_t}|W_{t-1}) \\ &= h\left(W_t|W_{t-1}\right) - h\left(W_t|W_{t-1}, \mathbf{Z}_t\right). \end{aligned} \quad (18)$$

Both of the entropies in the above equation can be bounded from the assumptions. The details of this are left in the appendix of [7]. □

The implication of Theorem 3 in conjunction with Theorem 1 is that for an SGLD algorithm as defined in (1):

$$\left|\mathrm{gen}\left(\Lambda, \mathbb{P}_{W|S}\right)\right| \leq \sqrt{\frac{\sigma^2}{n} \sum_{t=1}^{T} \frac{\eta_t^2 L^2}{\sigma_t^2}}. \quad (19)$$

Furthermore, Theorem 3 from [5] leads to a probabilistic generalization bound, which we state in the following theorem.

Theorem 4. *Let* $I(S; W) \leq \epsilon = \sum_{t=1}^{T} \frac{d}{2} \log\left(1 + \frac{\eta^2 L^2}{d\sigma_t^2}\right)$. *For any* $\alpha > 0$ *and* $0 < \beta \leq 1$, *if we have:*

$$n > \frac{\frac{sR^2}{\alpha^2}}{\left(\xi + \log\left(\frac{2}{\beta}\right)\right)} \quad (20)$$

$$P\left(\left|\mathrm{gen}\left(\Lambda, \mathbb{P}_{W|S}\right)\right| > \alpha\right) \leq \beta \quad (21)$$

As we stated earlier, since these results can be readily applied to any iterative algorithm, they can be applied to SGLD. In [7], this is exactly what is done.

It is established that for a given choice of β, α, taking:

$$n \geq \frac{64R^4}{a^4} \left(\log\left(\frac{2}{\beta}\right)\right)^2, \quad (22)$$

ensures that provided that (21) in satisfied, so long as a total of K epochs are run, where K satisfies:

$$K \leq \frac{1}{ne} \left(\frac{2^{2(\sqrt{n}-1)a}}{b^{-\frac{n}{2}}}\right). \quad (23)$$

This is a valuable result because it takes the abstract information-theoretic bounds, and provides real estimates for the amount of data and computation required to meet a given generalization error with high probability. However, as we will see, the true utility of this is limited by the looseness of these bounds.

V. Generalization Bounds for SGLD with Square Error Loss Functions

We will now use the results of Theorem 2 to prove a generalization bound on a learned distribution using SGLD with the square error loss function. Note that in the formulation which we consider, in each step the entire dataset $\mathbf{S}$ is used. In fact, our formulation hinges on this fact, exploiting the linearity of the update function. In [8], this formulation is referred to as Langevin Dynamics, since the gradient update does not depend on a stochastic choice of batch.

Theorem 5. *Let $\mathbf{S} = \{z_1, \ldots, z_n\} \sim \Lambda^n$ be a set of measurements. Furthermore let $\mathcal{Z} \subset \mathbb{R}$, and let the set $\mathcal{Z}$ be bounded, such that $Z_i \in \mathrm{SubG}(\sigma_z^2)$. We wish to estimate $\mathbb{E}_{Z\sim\Lambda}[Z]$. The loss function for this problem will be $\ell(z, w) = (z - w)^2$. Furthermore, we will assume that $\ell(w, Z) \in \mathrm{SubG}(\sigma^2)$ for all $w \in \mathcal{W}$. To solve this problem, we use SGLD, with a total of T iterations. The update function which we consider is:*

$$w_{t+1} = w_t - \frac{\eta_t}{n}\sum_{i=1}^{n} \nabla_{w_t} \ell\left(w_t, z_i\right) + \sqrt{\frac{2\eta_t}{\beta}}\xi_t, \tag{24}$$

where $\xi_t \sim \mathcal{N}(0, 1)$. Then the generalization error must satisfy:

$$|\operatorname{gen}(\Lambda, \mathbb{P}_{W|S})| < \sigma\sqrt{\log\left(\frac{b^2\sigma_z^2}{\sigma_\xi^2} + 1\right)}. \tag{25}$$

Where

$$\sigma_\xi^2 = \sum_{t=1}^{T}\left[\prod_{t'=t+1}^{T}\left(1 - 2\eta_{t'}\right)\sqrt{\frac{2\eta_t}{\beta}}\right]^2, \tag{26}$$

and,

$$b = \sum_{t=1}^{T}\frac{2\eta_t}{n}\prod_{t'=t+1}^{T}\left(1 - 2\eta_{t'}\right). \tag{27}$$

Proof. From Theorem 2, we have that:

$$|\operatorname{gen}\left(\mu, P_{W|\mathbf{S}}\right)| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I\left(w_T; z_i\right)}. \tag{28}$$

To upper bound this equation, we must upper bound $I(w_T; z_i)$. Since the measurements are independent, we can write:

$$I\left(w_T; z_i\right) \leq I\left(w_T; z_i | z_1, z_2, \ldots, z_{i-1}, z_{i-1}, \ldots, z_n\right). \tag{29}$$

We can write this in terms of entropy to be:

$$I\left(w_T; z_i\right) \leq h\left(w_T | z_1, z_0, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\right) - h\left(w_T|\mathbf{S}\right). \tag{30}$$

The second term is going to be easier for us to evaluate. Let us begin by re-writing the update function and grouping terms:

$$\begin{aligned} w_{t+1} &= w_t + \frac{2\eta_t}{n}\sum_{i=1}^{n}\left(z_i - w_t\right) + \sqrt{\frac{2\eta_t}{\beta}}\xi_t \\ &= \left(1 - 2\eta_t\right)w_t - 2\eta_t\left(\frac{1}{n}\sum_{i=1}^{n} z_i\right) + \sqrt{\frac{2\eta_t}{\beta}}\xi_t. \end{aligned} \tag{31}$$

If $\mathbf{S}$ is given, then the middle term in (31) is not random. The key observation here is that because the update function is linear in both w_t and ξ_t, in fact, w_T is a linear combination of sub-Gaussian random variables, and thus is itself sub-Gaussian. From this update function, we see that:

$$h(w_T|\mathbf{S}) = h\left(\sum_{t=1}^{T}\underbrace{\left(\prod_{t'=t+1}^{T}\left(1 - 2\eta_{t'}\right)\right)\sqrt{\frac{2\eta_t}{\beta}}}_{a_t}\xi_t\right). \tag{32}$$

This is the entropy of a random variable which is distributed like $\mathcal{N}\left(0, \sigma_\xi^2 = \sum_{t=1}^{T} a_t^2\right)$. Thus, we find that:

$$h\left(w_T|S\right) = \frac{1}{2}\log\left(2\pi e\sigma_\xi^2\right). \tag{33}$$

The first term in (30) is more complicated, because z_i is not known. Despite this, we can still bound this quantity by making the same key observation as we just did before. w_T is a linear function of z_i and ξ_t. From the update function we can show that:

$$w_T = \sum_{t=1}^{T}\frac{2\eta_t}{n}\prod_{t'=t+1}^{T}\left(1 - 2\eta_{t'}\right)z_i + \sum_{t=1}^{T} a_t^2\xi_t \tag{34}$$

Thus, we can bound the entropy in the first term:

$$h\left(w_T|z_1, z_2, \ldots, z_{i+1}, z_{i+1}, \ldots, z_n\right) = h\left(\underbrace{\sum_{t=1}^{T}\frac{2\eta_t}{n}\prod_{t'=t+1}^{T}\left(1 - 2\eta_{t'}\right)z_i}_{b} + \sum_{t=1}^{T} a_t\varepsilon_t\right). \tag{35}$$

Since $bz_i \in \operatorname{subG}\left(b^2\sigma_z^2\right)$ we have the following bound:

$$h\left(bz_i + \sum_{t=1}^{T} a_t\xi_t\right) \leq \frac{1}{2}\log\left(2\pi e\left(b^2\sigma_z^2 + \sum_{t=1}^{T} a_t^2\right)\right). \tag{36}$$

Putting these results together, we complete the proof:

$$I\left(w_T; z_i\right) \leq \frac{1}{2}\log\left(\frac{b^2\sigma_z^2}{\sum a_t^2} + 1\right). \tag{37}$$

□

For a learning rate of $\eta_t = \frac{1}{t}$, we have:

$$b = \frac{2}{n}\sum_{t=1}^{T}\frac{1}{t}\prod_{t'=t+1}^{T}\left(1 - \frac{2}{t'}\right). \tag{38}$$

The term in the product can be simplified (for $T > 1$):

$$\prod_{t'=t+1}^{T}\left(1 - \frac{2}{t'}\right) = \frac{t(t-1)}{T(T-1)}. \tag{39}$$

Thus,

$$b = \frac{2}{n}\sum_{t=1}^{T}\frac{1}{t}\frac{t(t-1)}{T(T-1)} = \frac{1}{n}. \tag{40}$$

Furthermore,

$$\sigma_\xi^2 = \frac{2}{\beta}\sum_{t=1}^{T}\left(\frac{t(t-1)}{T(T-1)}\right)^2\frac{1}{t} = \frac{2}{\beta T^2(T-1)^2}\sum_{t=1}^{T}t(t-1)^2 = \frac{(T+1)(3T-2)}{\beta 6T(T-1)}. \quad (41)$$

Thus, we can bound the generalization by:

$$|\operatorname{gen}(\Lambda, \mathbb{P}_{W|S})| < \sigma\sqrt{\log\left(\frac{b^2\sigma_z^2}{\sigma_\xi^2}+1\right)}. \quad (42)$$

Since $\log(x+1) < x \quad \forall x > 0$,

$$|\operatorname{gen}(\Lambda, \mathbb{P}_{W|S})| < \frac{b\sigma\sigma_z}{\sigma_\xi} = \frac{\sigma\sigma_z\sqrt{6\beta T(T-1)}}{n\sqrt{(T+1)(3T-2)}}. \quad (43)$$

VI. Generalization Bounds for SGLD using Data Dependent Measurements

In this section, we review the contributions of [8]. This work is motivated by the looseness of the bounds from [7]. The authors note that the bounds of Theorem 3 have several shortcomings. In particular, it is established that the use of Assumption 2 makes the bound of Theorem 3 ineffective in practice, as most interesting modern problems would require a prohibitively large Lipshitz constant L. The key contribution of [8] is the understanding that many of the shortcoming of Theorem 3 come from the fact that the bound is distribution independent.

In [8], it is proposed that this shortcoming can be overcome by the use of data-dependent estimates. The key idea is to split the initial data $\mathbf{S}$ into two parts. These two parts are denoted as $\mathbf{S}_J$ and $\mathbf{S}_J^c$ which satisfy $\mathbf{S}_J \cup \mathbf{S}_J^c = \mathbf{S}$. Let $|\mathbf{S}| = n$ and $|\mathbf{S}_J| = m$. One of these subsets can be used to make a data-dependant prior, which is independent of the other subset. Using this fact, the authors establish the following theorem.

Theorem 6. *Let $W \in \mathcal{W}$ be a random element, let $\mathbf{S} \sim \Lambda^n$, and let $\mathbf{J} \subseteq [n], |\mathbf{J}| = m$, be uniformly distributed and independent from $\mathbf{S}$ and W. If $\ell(w, Z) \in \operatorname{SubG}(\sigma^2)$ with $Z \sim \Lambda \ \ \forall w \in \mathcal{W}$ Let $Q = \mathbb{P}_{W|S}$, and let P be a $\boldsymbol{\sigma}(\mathbf{S}_J)$ measurable data-dependent prior on $\mathcal{W}$. Then:*

$$|\operatorname{gen}(Q, \Lambda)| \leq \sqrt{2\frac{\sigma^2}{n-m}I\left(W; \mathbf{S}_j^c\right)} \leq \sqrt{2\frac{\sigma^2}{n-m}\mathbb{E}[\operatorname{KL}(Q /\!\!/ P)]}. \quad (44)$$

The complete proof is available in Appendix B of [8]. The primary tools used in the proof is the Donsker-Varadhan lemma and a clever re-writing in terms of the cumulant generating function.

This key insight and initial result is further expanded upon with application to SGLD. Due to space limitations, in this review we are unable to study the theory of [8] in full form. Instead, we will focus on a simple example of (24). The following analysis can be found in the appendix of [8].

We use $m = n-1$ and $\{i^\star\} = \mathbf{J}$. It follows from the results that:

$$\operatorname{KL}\left(Q_{t+1}(\mathbf{S}) /\!\!/ P_{t+1}\left(\mathbf{S}_J\right)\right) = \frac{\left(\mu_{t+1} - \mu'_{t+1}\right)^2}{4\eta_t/\beta} = \frac{\beta}{n^2}z_i^2\eta_t. \quad (45)$$

Thus, we can apply the following generalization bound:

$$\left|\operatorname{gen}\left(\Lambda, \mathbb{P}_{W|S}\right)\right| \leq \mathbb{E}\sqrt{2\sigma^2 KL\left(Q_T(S) \| P_T\left(S_J\right)\right)} \leq \mathbb{E}\sqrt{2\sigma^2\frac{\beta}{n^2}z_{i^*}^2\sum_{t=0}^{T-1}\eta_t} = \mathbb{E}\left[|z_i|\right]\left(\sqrt{2\sigma^2\frac{\beta}{n^2}\sum_{t=0}^{T-1}\eta_t}\right) \quad (46)$$

However, when one applies the methods of [7] we find a bound of:

$$\left|\operatorname{gen}\left(\Lambda, \mathbb{P}_{W|\mathbf{S}}\right)\right| \leq \sqrt{\frac{2\sigma^2}{n}\sum_{t=0}^{T-1}I\left(\bar{W}_{t+1}; \mathbf{S}|W_1^t\right)} \leq \sqrt{2\sigma^2\frac{\beta}{n^2}E\left[z_i^2\right]\sum_{t=0}^{T-1}\eta_t}. \quad (47)$$

Comparing these two, we can see we see that (47) is larger since $E\left[|z_i|\right] \leq \sqrt{E\left[z_i^2\right]}$ by Jensen's inequality. It is not immediately clear how these bounds compare to Theorem 5, and this is left for future study.

VII. Conclusion

In this review we considered some recent developments in information-theoretic generalization bounds for SGLD. We motivated this study by establishing why such bounds are considered the "holy-grail" of statistical learning [1]. After this we establish the initial information-theoretic frameworks proposed in [4], [5], as well as the improvements considered by [6]. These initial works provided a powerful new tool for analyzing the generalization of stochastic statistical learning algorithms. They are based on the intuitive notion that the more an algorithm's output depends on the input, the less likely the output is to generalize. This is formalized by considering the information-theoretic notion of mutual information between the output and input. Next we reviewed [7], which focuses on the generalization of iterative algorithms. SGLD falls into their framework, and their analysis allows us to make guarantees about the performance of SGLD with certain choices of parameters. Their work primarily lies on upper bounding the mutual information of the final output of an iterative algorithm and the input data by considering the entire trajectory of the of the output at each iteration. Furthermore, they make a Lipshitz assumption in order to eliminate data distribution dependence. Though this technique presents some interesting ideas, the shortcomings are also apparent.

In [8], the authors note the impracticality of these bounds for modern problems. In particular, they espouse the idea

that data-dependant bounds are needed to develop ways of bounding generalization error in modern problems. To do this, they consider a framework where the initial data is split into two parts. Part of the data is used for generating a prior, while the pother is not, thus enabling the development of generalization bounds which depend on the underlying data distribution. Finally, we develop a bound for a particular case of SGLD with a square-error loss function. This bound does not follow the procedure of [7], but instead uses the linearity of the update function. Future work remains to validate the new proposed bounds and form a proper comparison.

REFERENCES

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012.

[2] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, Madison, WI, USA, 2011, p. 681–688.

[3] M.-N. Hamid and I. Friedberg, "Reliable uncertainty estimate for antibiotic resistance classification with stochastic gradient langevin dynamics," 2018.

[4] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, A. Gretton and C. C. Robert, Eds., 2016.

[5] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 2524–2533. [Online]. Available: http://papers.nips.cc/paper/6846-information-theoretic-analysis-of-generalization-capability-of-learning-algorithms.pdf

[6] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 587–591.

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

[8] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-theoretic generalization bounds for sgld via data-dependent estimates," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 015–11 025. [Online]. Available: http://papers.nips.cc/paper/9282-information-theoretic-generalization-bounds-for-sgld-via-data-dependent-estimates.pdf