Information-Theoretic Generalization Bounds for Stochastic Gradient Descent

Gergely Neu GERGELY.NEU@GMAIL.COM
Universitat Pompeu Fabra, Barcelona, Spain

Gintare Karolina Dziugaite KAROLINA.DZIUGAITE@ELEMENTAI.COM
Element AI / Mila, Montreal, Canada

Mahdi Haghifam MAHDI.HAGHIFAM@MAIL.UTORONTO.CA
University of Toronto / Vector Institute, Toronto, Canada

Daniel M. Roy DANIEL.ROY@UTORONTO.CA
University of Toronto / Vector Institute, Toronto, Canada

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study the generalization properties of the popular stochastic optimization method known as stochastic gradient descent (SGD) for optimizing general non-convex loss functions. Our main contribution is providing upper bounds on the generalization error that depend on local statistics of the stochastic gradients evaluated along the path of iterates calculated by SGD. The key factors our bounds depend on are the variance of the gradients (with respect to the data distribution) and the local smoothness of the objective function along the SGD path, and the sensitivity of the loss function to perturbations to the final output. Our key technical tool is combining the information-theoretic generalization bounds previously used for analyzing randomized variants of SGD with a perturbation analysis of the iterates.

Keywords: stochastic gradient descent, generalization, information-theoretic generalization

1. Introduction

Stochastic gradient descent (SGD) is arguably the single most important algorithmic component of the modern machine-learning toolbox. First proposed by Robbins and Monro (1951) for finding roots of a function using noisy evaluations, stochastic approximation methods like SGD has been broadly adapted for a variety of tasks in signal processing, control and optimization (Kushner and Yin, 1997; Nemirovski et al., 2009). In the context of machine learning, SGD is extremely popular due to its efficient use of computation that naturally enables it to process very large data sets and its inherent ability to handle noisy data (Bottou and Bousquet, 2007). In modern machine learning, SGD is the de facto standard method for training deep neural networks, partly due to the efficient computability of the gradients through the famous backpropagation algorithm (Rumelhart et al., 1986; LeCun et al., 2012). Given the surprising effectiveness of SGD for deep learning, several empirical and theoretical studies have attempted to explain the reasons of its success. In recent years, this research effort has lead to some remarkable results that finally shed some light on the core factors contributing to the effectiveness of SGD. A handful of examples include showing guaranteed convergence to minimizers (Ge et al., 2015; Lee et al., 2016), to minimum-norm solutions under various assumptions (Ma et al., 2018; Jacot et al., 2018; Oymak and Soltanolkotabi, 2019), and even to global optima of overparametrized neural networks (Du et al., 2018; Allen-Zhu et al., 2019).

This paper aims to contribute to a better understanding of the *generalization properties of SGD* for optimizing general non-convex objectives. This is a widely studied question, mostly from the perspective of *stability*, inspired by the understanding that stable learning algorithms are guaranteed to generalize well to test data (Bousquet and Elisseeff, 2002; Feldman and Vondrák, 2019; Bousquet et al., 2020). This line of work was initiated by Hardt, Recht, and Singer (2016), who showed that SGD has strong stability properties when applied to smooth convex loss functions, and is particularly stable when the objective is also strongly convex. They also provided bounds on the generalization error for non-convex losses, with a rate that is polynomial in the number of steps T, with an exponent that depends on the smoothness of the objective. A significant limitation of their results is that they require the loss function to have a bounded Lipschitz constant, which is generally difficult to ensure (especially when training deep neural networks). In recent years, the results of Hardt et al. have been strengthened in a variety of ways, for example by removing the Lipschitz condition in the general non-convex case by Lei and Ying (2020) and by removing the smoothness condition in the convex case by Bassily et al. (2020).

Another line of work studies the generalization properties of a randomized version of SGD called *stochastic gradient Langevin dynamics* (SGLD, cf. Welling and Teh, 2011). As shown by Raginsky et al. (2017), SGLD has strong finite-sample convergence properties for general non-convex learning, which already implies good generalization. Another line of attack aiming to directly understand the generalization of SGLD was initiated by Pensia, Jog, and Loh (2018). Subsequent improvements to this technique were made by Negrea et al. (2019), Haghifam et al. (2020) and Rodríguez-Gálvez et al. (2021), who prove data-dependent generalization bounds that do not depend on the Lipschitz constant of the loss function. The core technical tool of these analyses is bounding the generalization error in terms of the mutual information between inputs and outputs of learning algorithms, previously proposed in a much more general context by Russo and Zou (2016, 2019) and Xu and Raginsky (2017), which are themselves closely connected to classic PAC-Bayesian generalization error bounds (McAllester, 1999, 2013). Indeed, these information-theoretic tools are particularly suitable for analyzing SGD-like algorithms due to a convenient decomposition of the mutual infomation across iterations by an application of the chain rule of the relative entropy. The obvious downside of this technique is that it relies on randomly perturbing the SGD iterates, which empirically hurts the performance of the algorithm and results in underfitting the training loss.

Our main contribution in this paper is demonstrating that it is possible to conduct an information-theoretic analysis for the vanilla version of SGD, without directly perturbing the iterates. Our key observation is that it is sufficient to add carefully constructed random noise to the iterates *only during the analysis*, which then allows using the technique of Pensia et al. (2018) on these "virtual SGLD" iterates. The key challenge in the analysis is designing the perturbations and correctly handling their propagation through the iterations. The bounds we derive depend on three factors: the variance of the gradient updates and the sensitivity of the gradients to perturbations along the path, and the sensitivity of the final output to perturbations. All of these quantities are evaluated locally along the iterates that SGD produces. Another important consequence of our virtual perturbation technique is that our bounds hold simultaneously for all noise distributions, which obviates the need to tune the hyperparameters of the algorithm to optimize the bound.

As is probably obvious from the above discussion, our approach is thoroughly inspired by the insightful work of Pensia, Jog, and Loh (2018), which itself is inspired by the seminal works of Russo and Zou (2016, 2019) and Xu and Raginsky (2017) on information-theoretic generalization bounds. In recent years, their theory has been extended in a variety of directions that resulted in tighter and

tighter bounds. Here, we highlight the works of Asadi et al. (2018); Asadi and Abbe (2020) whose chaining technique can lead to tighter bounds when applied to neural networks, and the work of Steinke and Zakynthinou (2020) whose key idea of appropriately conditioning the mutual information enables proving more refined data-dependent guarantees. In fact, these latter ideas directly inspired the work of Haghifam et al. (2020) and Rodríguez-Gálvez et al. (2021) on tighter generalization bounds for Langevin dynamics and SGLD, respectively. We conjecture that our analysis can be also refined by using such sophisticated information-theoretic techniques. Finally, we also mention that our main idea of using random perturbations to guarantee boundedness of the mutual information has been partially inspired by the work of Zhang et al. (2020).

Besides the works mentioned above, several other theories of generalization have been proposed in the deep learning literature. One particularly widely held belief is that algorithms that find "wide optima" of the loss landscape generalize well to test data. This hypothesis has been first proposed by Hochreiter and Schmidhuber (1997) and later popularized by Keskar et al. (2017), and has attracted some (moderately rigorous) verification and refutation attempts (e.g., Dinh et al., 2017; Izmailov et al., 2018; He et al., 2019; Chaudhari et al., 2019). One common criticism of this theory is that the "width" of a solution is difficult to formally define, and the most common intuitive definitions suffer from being sensitive to reparametrization. While we don't claim to resolve the debate around "wide optima", our results lend some minimal credence to this theory in that our generalization bounds indeed predict an improvement when the loss of the final solution is insensitive to perturbations. As we will show, our bounds allow measuring this sensitivity in terms of arbitrary coordinate systems, which at least addresses the most elementary concerns with parametrization-sensitivity of common definitions. That said, some aspects of our results defy common wisdom: most notably, our bounds show an *improvement* for larger batch sizes, even though this choice empirically leads to worse performance, purportedly due to converging to "sharper minima". Importantly, our analysis does *not* explain why SGD would converge towards solutions that generalize well and how hyperparameters like the batch size could impact the quality of solutions.

Notation. For two distributions P and Q satisfying $P \ll Q$, we denote their relative entropy by $\mathcal{D}(P\|Q) = \int_x \mathrm{d}P(x) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right)$. The distribution of a random variable X will be denoted by P_X and the product distribution between P_X and P_Y will be denoted by $P_X \otimes P_Y$. With this notation, the mutual information between two random variables is defined as $I(X;Y) = \mathcal{D}(P_{X,Y}\|P_X \otimes P_Y)$. Whenever possible, we use capital letters to denote random variables and use lowercase letters to denote their realizations. We use $\|\cdot\|$ to denote the Euclidean norm on $\mathbb{R}^d$ and $\mathcal{S}_+$ to denote the set of symmetric positive definite matrices in $\mathbb{R}^{d\times d}$. For any $u \in \mathbb{R}^d$ and $\Sigma \in \mathcal{S}_+$, we will use $\mathcal{N}(u,\Sigma)$ to denote the multivariate normal distribution with mean u and covariance Σ.

2. Background

We let $S = \{Z_1, Z_2, \ldots, Z_n\}$ denote a data set of n i.i.d. samples taking value in the set $\mathcal{Z}$ and $W = \mathcal{A}(S)$ be the output of a (potentially randomized) learning algorithm run on data set S. We assume that this output is a d-dimensional real-valued vector, representing parameters of a potentially nonlinear model such as a neural network. The performance of a learning algorithm is evaluated in terms of a loss function mapping data points and parameter vectors to positive real numbers, with $\ell(w,z)$ giving the loss of the model with parameters $w \in \mathbb{R}^d$ evaluated on data point $z \in \mathcal{Z}$. We will assume throughout that $\ell(w,z)$ is differentiable with respect to w for

all z, and let $g(w,z)$ denote its gradient evaluated at w. Furthermore, we will assume that the distribution of $Z \sim Z_1$ is such that $\ell(w,Z)$ is R-subgaussian for any $w \in \mathbb{R}^d$ in the sense that the inequality $\mathbb{E}\left[\exp(y\left(\ell(w,Z) - \mathbb{E}\left[\ell(w,Z)\right]\right))\right] \leq \exp(R^2y^2/2)$ is satisfied for all $y \in \mathbb{R}$. This condition clearly holds if the loss function is bounded on the support of the data distribution. Letting $S' = \{Z'_1, Z'_2, \ldots, Z'_n\}$ be an independent data set of the same distribution as S, and denoting the average loss of w on data set $s = \{z_1, z_2, \ldots, z_n\}$ by $L(w,s) = \frac{1}{n}\sum_{i=1}^n \ell(w,z_i)$, we define our main object of interest, the *expected generalization error* of algorithm $\mathcal{A}$ on the data set S as

$$\mathrm{gen}(W,S) = \mathbb{E}\left[L(W,S') - L(W,S)\right].$$

The expected generalization error (that we will often simply call *generalization error*) measures the difference between the training loss $L(W,S)$ and the test loss $\mathbb{E}\left[L(W,S')|\,W\right]$ on expectation with respect to the randomness of the data set S and the output W. Our techniques will be based on the now-classic results of Russo and Zou (2016); Xu and Raginsky (2017) that show that, under the conditions stated above, the generalization error of any learning algorithm can be bounded as

$$|\mathrm{gen}(W,S)| \leq \sqrt{\frac{2R^2 I(W;S)}{n}}, \tag{1}$$

where $I(W;S)$ denotes the mutual information between the random variables W and S.

We will consider the classic stochastic gradient descent (SGD) algorithm originally proposed by Robbins and Monro (1951) and applied here to approximately minimize the empirical loss $L(w,S)$ in terms of w on the data set S. This iterative algorithm operates by making sequential updates to a parameter vector in the direction opposite to the gradient of the loss function evaluated at a minibatch of data points. Such a minibatch estimator will be denoted with a slight abuse of notation as $g(w,Z_J) = \frac{1}{|J|}\sum_{i\in J} g(w,Z_i)$, where $J \subseteq [n]$ is a set of indices. Then, the iterates of SGD are given by drawing W_1 from an arbitrary fixed distribution independent of S, and then updating the parameters recursively as

$$W_{t+1} = W_t - \eta_t G_t = W_t - \eta_t g(W_t, Z_{J_t}) \tag{2}$$

for all $t = 1, \ldots, T$, where η_t is a positive learning-rate parameter and $J_t \subseteq [n]$ is the set of b_t indices of the minibatch selected for the t-th update, and $G_t = g(W_t, Z_{J_t})$. We will assume that η_t and J_t are chosen independently of the history or the data set, but otherwise make no restrictions about them (e.g., we allow increasing and cyclic stepsizes, and randomized minibatch schedules such as random shuffling). We will often denote the minibatch Z_{J_t} by B_t for brevity, and we will sometimes refer to the quantity $\sum_{t=1}^T b_t/n$ as the number of passes over the data set.

3. Generalization-error bounds for SGD

Our main result is a bound on the generalization error of the final iterate W_T produced by SGD as defined in the previous section. The key quantities appearing in the bound are the following:

- The *local value sensitivity* of the loss function at $w \in \mathbb{R}^d$, defined on data set $s \in \mathcal{Z}^n$ at level σ as
$$\Delta_\sigma(w,s) = \mathbb{E}\left[L(w,s) - L(w+\xi,s)\right]$$
where $\xi \sim \mathcal{N}(0,\sigma^2 I)$. When the loss is μ-smooth, this quantity is bounded by $\mu\sigma^2 d/2$.

- The *local gradient sensitivity* of the loss function, defined at $w \in \mathbb{R}^d$ and at level σ as

$$\Gamma_\sigma(w) = \mathbb{E}\left[\|\overline{g}(w) - \overline{g}(w+\xi)\|^2\right],$$

  where $\xi \sim \mathcal{N}(0, \sigma^2 I)$, and $\overline{g}(w) = \mathbb{E}[g(w, Z)]$ is the population gradient evaluated at w. This quantity characterizes the sensitivity of the gradients of the expected loss function to perturbations around w, and is upper-bounded by $\mu^2\sigma^2 d$ when the loss function is globally μ-smooth.

- The *local gradient variance* of the loss function defined at $w \in \mathbb{R}^d$ as

$$V_t(w) = \mathbb{E}\left[\|g(w, B_t) - \overline{g}(w)\|^2 \,\middle|\, W_t = w\right].$$

  It is important to note that this is a non-standard notion of variance in that it measures the expected squared Euclidean distance from the *population gradient* $\mathbb{E}[g(w, Z)]$ instead of the conditional expectation of the gradient $\mathbb{E}[g(w, Z)|\, W_t = w]$.

Our main result is stated as follows:

Theorem 1 *Fix any sequence $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_T)$ of positive real numbers and define $\sigma_{1:t} = \sqrt{\sum_{k=1}^{t-1} \sigma_k^2}$ for all t. Then, the generalization error of the final iterate of SGD satisfies*

$$|\mathrm{gen}(W_T, S)| \leq \sqrt{\frac{4R^2}{n}\sum_{t=1}^{T} \frac{\eta_t^2}{\sigma_t^2}\mathbb{E}\left[\Gamma_{\sigma_{1:t}}(W_t) + V_t(W_t)\right]} + \left|\mathbb{E}\left[\Delta_{\sigma_{1:T}}(W_T, S') - \Delta_{\sigma_{1:T}}(W_T, S)\right]\right|.$$

The proof is provided in Section 4. There are several interesting properties of this result. One of its most important merits is that it depends on *pathwise* statistics of the SGD iterates, as the local gradient sensitivity and variance parameters are evaluated along the path taken by SGD. Consequently, the bound becomes small whenever the gradients demonstrate little variability along the path, both in terms of varying w and Z. Finally, the bound also depends on the sensitivity of the loss function to perturbations around the final output W_T, measured both on the training set S and the test set S'. Intuitively, this term becomes small if SGD outputs a parameter vector in a flat area of the loss surface.

While it may feel unsatisfying that the tradeoffs involving $\boldsymbol{\sigma}$ are not characterized explicitly, one can find consolation in the remarkable fact that the bound simultaneously holds for all possible choices of $\boldsymbol{\sigma}$. Indeed, this is a very useful property given that the bound presents some delicate tradeoffs involving the parameters $\boldsymbol{\sigma}$: large values of σ_t^2 decrease the first term in the bound related to the variability of the gradients, at the expense of increasing the second term related to the flatness of the loss around the final iterate. More generally, the optimal choice of these parameters can depend on the local properties of the loss functions along the expected SGD path. One important limitation is that the bound only holds for *fixed* sequences, and in particular that σ_t is not allowed to depend on the past iterates $W_{1:t}$. Similarly, the theorem crucially requires the sequence of learning rate schedule to be fixed independently of the data. We provide a more detailed discussion of this issue at the end of Section 5.3. On the other hand, it is possible to further extend the flexibility of the bound by allowing more general perturbation covariances; details are provided in Section 5.1.

3.1. Generalization-error guarantees for smooth loss functions

To provide some intuition about the magnitude of the terms in the bound, we state the following corollary that provides a simpler bound under some concrete assumptions on the loss function and the parameters:

Corollary 2 *Suppose that $\eta_t = \eta$ and $b_t = b$ for all t and the minibatches are chosen so that for each $i \in [n]$, there is exactly one index t such that $i \in J_t$. Furthermore, suppose that $\mathbb{E}\left[\|g(w,Z) - \overline{g}(w)\|^2\right] \leq v$ for all w and that ℓ is globally μ-smooth in the sense that the inequality $\|g(w,z) - g(w+u,z)\| \leq \mu \|u\|$ holds for all $w, u \in \mathbb{R}^d$ and all $z \in \mathcal{Z}$. Then, the generalization error of the final iterate of SGD satisfies the following bound for any σ:*

$$|\mathrm{gen}(W_T, S)| = O\left(\sqrt{\frac{R^2\eta^2 T}{n}\left(\mu^2 dT + \frac{v}{b\sigma^2}\right)} + \mu\sigma^2 dT\right).$$

The proof follows from noticing that

$$\Gamma_{\sigma_{1:t}}(w) = \mathbb{E}\left[\|g(w+\xi_t, z) - g(w,z)\|^2\right] \leq \mu^2 \mathbb{E}\left[\|\xi_t\|^2\right] = \mu^2 d\sigma_{1:t}^2 = \mu^2 d\sigma^2 t$$

and that $\Delta_{\sigma_{1:T}}(W_T, S)$ and $\Delta_{\sigma_{1:T}}(W_T, S')$ can be bounded using

$$|\mathbb{E}\left[\ell(W_T, z) - \ell(W_T + \xi_T, z)\right]| \leq |\mathbb{E}\left[\langle \nabla \ell(w,z), \xi_T\rangle\right]| + \frac{\mu}{2}\mathbb{E}\left[\|\xi_T\|^2\right] = \frac{\mu\sigma_{1:T}^2 d}{2} = \frac{\mu\sigma^2 dT}{2},$$

where the equality follows from using $\mathbb{E}\left[\xi_T\right] = 0$ and the independence of W_T and ξ_T. Furthermore, the independence of Z_{J_t} and W_t implies that $V_t(W_t) = \frac{1}{b}\mathbb{E}\left[\|g(W_t, Z) - \mathbb{E}\left[g(W_t, Z')\right]\|^2\right] \leq \frac{v}{b}$.

The rates become better as T and η are decreased, but doing so can hurt the fit on the training data. Furthermore, the rates also improve as b is increased, although only until a critical value where the term $\mu^2 dT$ becomes dominant. Such tradeoffs involving the batch size are not uncommon in the related literature (see, e.g., Lin et al., 2020). The noise variance σ^2 still influences the bound in a relatively complex way, but should be tuned as a function of the minibatch gradient variance v/b: as this quantity approaches zero, one can afford to set smaller perturbations and improve the last sensitivity term in the bound. Once again, we highlight that the optimal tuning of σ does not require prior knowledge of problem parameters like v and μ.

We discuss the rates that can be derived from the bound in some important settings:

Small-batch SGD: Setting $T = O(n)$ and $b = O(1)$, it is not possible derive a bound that vanishes with large n under the classic stepsize choice $\eta = O(1/\sqrt{n})$. That said, using $\eta = O(1/n)$ and $\sigma = \Theta(n^{-4/3})$, it is possible to guarantee the vanishing rate $|\mathrm{gen}(W;S)| = O(n^{-1/3})$. of

Large-batch SGD: Setting $T = O(\sqrt{n})$ and $b = \Omega(\sqrt{n})$, the stepsize as $\eta = O(1/T) = O(1/\sqrt{n})$, and $\sigma = \Theta(1/\sqrt{n})$ we obtain a rate of $|\mathrm{gen}(W;S)| = O(1/\sqrt{n})$.

Interestingly, the rates for the latter case saturate when setting $b = \Theta(\sqrt{n})$ and do not improve any further even when setting $b = \omega(\sqrt{n})$. Thus, even if the bounds of Corollary 2 remain qualitatively true for larger batch sizes, no further improvement is to be expected. In both cases discussed above, the rates are obviously far from being tight in general, especially in the small-batch case where using stepsizes of order $1/\sqrt{n}$ are known to lead to bounds of order $1/\sqrt{n}$ for general convex functions,

and even better rates are known when a smoothness assumption is also in place. That said, the above examples show that it is indeed possible to achieve generalization-error bounds that are vanishing in n, for parameter settings that are not entirely unrealistic. Notably, unlike the rates proved by Hardt et al. (2016), the exponent of the rates we can guarantee is independent of the smoothness parameter and the rates themselves are independent of the Lipschitz constant of the loss function. One downside of our guarantees is their direct dependence on the dimension d which can be attenuated by using more general perturbation distributions that are better adapted to the geometry of the loss function (cf. Section 5.1).

3.2. Comparison with SGLD

To put our result into perspective, we now also describe the Stochastic Gradient Langevin Dynamics (SGLD) algorithm that is essentially a variant of SGD that adds isotropic Gaussian noise to its iterates (Gelfand and Mitter, 1991; Welling and Teh, 2011). Specifically, the iterates of SGLD are given by the recursion

$$W_{t+1} = W_t - \eta_t g(W_t, Z_{J_t}) + \varepsilon_t, \tag{3}$$

for all t, where $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2 I)$, and the hyperparameters η_t, σ_t and J_t are chosen according to an arbitrary rule oblivious to the data set S. As first shown by Pensia et al. (2018), the generalization error of this algorithm can be directly bounded in terms of the mutual information between W_T and S, which itself can be shown to be of order $C^2 \sum_{t=1}^{T} \eta_t^2/\sigma_t^2$, under the assumption that the loss function is C-Lipschitz. This bound has been improved by Negrea et al. (2019) and Haghifam et al. (2020), who replace the Lipschitz constant by a data-dependent quantity that is often orders of magnitude smaller. Instead of reproducing their rather involved definitions, we state the following simple guarantee that can be obtained via a straightforward modification of the proof of our Theorem 1:

Proposition 3 *The generalization error of the final iterate of SGLD satisfies*

$$\operatorname{gen}(W_T, S) \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^{T} \frac{\eta_t^2}{\sigma_t^2} \mathbb{E}\left[V_t(W_t)\right]}.$$

While this bound can be weaker than the data-dependent ones mentioned above, they are quite directly comparable to our results concerning SGD. One key difference is that the sensitivity terms disappear from the generalization bound, which can be attributed to SGLD being inherently more stable than SGD. This improved generalization-error guarantee however comes at the price of a worse training error, owing to the presence of the perturbations in the updates. While this degradation is difficult to quantify in general, it is qualitatively related to the effect captured by the sensitivity terms in our bound for SGD: intuitively, large sensitivity of the gradients to perturbations along the path can negatively impact the convergence speed in convex settings, and can result in large variations of the final iterate in nonconvex settings. Thus, in a certain sense, the sensitivity terms in our bound for SGD are pushed into the training error of SGLD. That said, there are known cases (e.g., strongly convex loss functions) where adding perturbations to the iterates incrementally is known to result in significantly better excess-risk guarantees than perturbing the final iterate (Feldman et al., 2018). The extent to which this holds for general non-convex functions is unclear.

A major downside of SGLD is that it requires prior commitment to the sequence of perturbation parameters $\boldsymbol{\sigma}$. This is made complicated by the poorly-understood tradeoffs between the generalization error and the training error: while the former is improved by setting large perturbation variances,

the latter is clearly hurt by it. Since this tradeoff is characterized even less explicitly than in our main result regarding SGD, it is virtually impossible to set $\boldsymbol{\sigma}$ in a way that optimizes both terms of the excess risk. Thus, even if the sensitivity terms in our bound are worse than the excess training loss of SGLD for a fixed $\boldsymbol{\sigma}$, our bounds still have the major advantage of simultaneously holding for all noise distributions, thus obviating the need to tune hyperparameters of the algorithm itself. Based on the above discussion, we find it plausible that SGLD may generally be able to achieve better excess risk than SGD, although with the major caveat that tuning its hyperparameters is prohibitively complex as compared to SGD.

4. Analysis

The core idea of our analysis is applying the generalization bound (1) to a perturbed version of the output W_T, making sure that the mutual information between the input and the perturbed output is bounded. To be precise, we define a perturbed version of the output as $\widetilde{W}_T = W_T + \xi_T$, where ξ_T is a random perturbation independent from the data and W_T. For the proof of our main theorem, we will use perturbations from a zero-mean isotropic Gaussian distribution with variance $\sigma_{1:T}^2$, that is, $\xi_T \sim \mathcal{N}(0, \sigma_{1:T}^2 I)$. It is then straightforward to apply the generalization-error guarantee (1) to the pair $(\widetilde{W}_T, S)$ and bound the generalization error as

$$\begin{aligned} \operatorname{gen}(W_T, S) &= \operatorname{gen}(\widetilde{W}_T, S) + \mathbb{E}\left[L(W_T, S') - L(\widetilde{W}_T, S')\right] + \mathbb{E}\left[L(\widetilde{W}_T, S) - L(W_T, S)\right]. \\ &\leq \sqrt{\frac{2R^2 I(\widetilde{W}_T; S)}{n}} + \mathbb{E}\left[\Delta_{\sigma_{1:T}}(W_T, S') - \Delta_{\sigma_{1:T}}(W_T, S)\right]. \end{aligned} \tag{4}$$

The key challenge in the proof is controlling the mutual information $I(\widetilde{W}_T; S)$. In order to bound this term, our analysis makes direct use of techniques first introduced by Pensia et al. (2018), with subtle modifications made to account for the fact that our perturbations do not appear as part of the algorithm, but are only defined to aid the analysis.

Our main technical idea is constructing the perturbation ξ_T in an incremental manner, and using these incremental perturbations to define a *perturbed SGD path*. In particular, we define the perturbations $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2 I)$ and the perturbed SGD iterates through the recursion

$$\widetilde{W}_{t+1} = \widetilde{W}_t - \eta_t G_t + \varepsilon_t = \widetilde{W}_t - \eta_t g(W_t, Z_{J_t}) + \varepsilon_t. \tag{5}$$

This procedure can be seen to produce $\widetilde{W}_t = W_t + \xi_t$ with $\xi_t = \sum_{k=1}^{t-1} \varepsilon_k \sim \mathcal{N}(0, \sigma_{1:t}^2)$, eventually yielding $\widetilde{W}_T = W_T + \xi_T$ as output. Intuitively, these iterates can be thought of as being between the SGD iterates (2) and the SGLD iterates (3) in that they add random perturbations to each update, but the gradients themselves are evaluated at the unperturbed SGD iterates. This results in a solution path that is strongly coupled with the SGD iterates, yet is still amenable to analysis techniques introduced by Pensia et al. (2018) due to the presence of the perturbations.

For the analysis, it will be also useful to define the "ghost SGD" iterates and their perturbed counterpart as

$$W'_{t+1} = W'_t - \eta_t G'_t \qquad \text{and} \qquad \widetilde{W}'_{t+1} = \widetilde{W}'_t - \eta_t G'_t + \varepsilon'_t$$

that use the independently drawn data set S' and minibatch B'_t indexed by J'_t to construct the gradient estimates $G'_t = g(W'_t, B'_t)$, and the independent perturbation $\varepsilon'_t \sim \mathcal{N}(0, \sigma_t^2 I)$. As we will see,

bounding the mutual information between $\widetilde{W}_T$ and S can be reduced to bounding the relative entropy between the conditional distribution of $\widetilde{W}_T$ given S and the marginal distribution of $\widetilde{W}'_T$ (which matches the marginal distribution of W_T by definition).

The analysis crucially relies on the following general lemma that quantifies the effect of random perturbations on the relative entropy of random variables:

Lemma 4 *Let X and Y be random variables taking values in $\mathbb{R}^d$ with bounded second moments and let $\sigma > 0$. Letting $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ be independent of X and Y, the relative entropy between the distributions of $X + \varepsilon$ and $Y + \varepsilon$ is bounded as*

$$\mathcal{D}\left(P_{X+\varepsilon} \| P_{Y+\varepsilon}\right) \leq \frac{1}{2\sigma^2} \mathbb{E}\left[\|X - Y\|^2\right].$$

The bound is tight in the sense that it holds with equality if X and Y are constants, although we also note that it can be arbitrarily loose in some cases (e.g., the left-hand side is obviously zero when X and Y are identically distributed, whereas the right-hand side can be positive if they are independent). This looseness can be addressed by observing that the bound allows for arbitrary dependence between X and Y, so one can pick the coupling between these random variables that minimizes the bound. In other words, the term on the right-hand side can be replaced by the squared 2-Wasserstein distance. Essentially the same result has been previously shown as Lemma 3.4.2 by Raginsky and Sason (2013) and a similar (although much less general) connection between the relative entropy and the Wasserstein distance has been made by Zhang et al. (2020). We provide the straightforward proof of Lemma 4 in Appendix A.

We are now in position to state and prove our most important technical result that, together with the inequality (4), will immediately imply the statement of Theorem 1:

Theorem 5 *Fix any sequence $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_T)$ of positive real numbers and let $\sigma_{1:t} = \sqrt{\sum_{k=1}^{t-1} \sigma_k^2}$ for all t. Then, the mutual information between $\widetilde{W}_T$ and S satisfies*

$$I(\widetilde{W}_T; S) \leq \sum_{t=1}^{T} \frac{2\eta_t^2}{\sigma_t^2} \mathbb{E}\left[\Gamma_{\sigma_{1:t}}(W_t) + V_t(W_t)\right].$$

Proof We start by defining some useful notation. For all t, we let $Q_{t|X}$ and $\widetilde{Q}_{t|X}$ respectively denote the distributions of W_t and $\widetilde{W}_t$ conditioned on the random variable X, and $\widetilde{Q}_{1:T|X}$ denote the joint distribution of $\widetilde{W}_{1:T} = (\widetilde{W}_1, \ldots, \widetilde{W}_T)$ given X. Further, we will denote the distribution of the data set as ν. Then, the mutual information between $\widetilde{W}_T$ and S can be bounded as

$$I(\widetilde{W}_T; S) = \mathbb{E}\left[\mathcal{D}\left(\widetilde{Q}_{T|S} \middle\| \widetilde{Q}_T\right)\right] \leq \mathbb{E}\left[\mathcal{D}\left(\widetilde{Q}_{1:T|S} \middle\| \widetilde{Q}_{1:T}\right)\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathcal{D}\left(\widetilde{Q}_{t|\widetilde{W}_{1:t-1},S} \middle\| \widetilde{Q}_{t|\widetilde{W}_{1:t-1}}\right)\right],$$

where the inequality follows from the data-processing inequality, and the last step uses the chain rule of the relative entropy. Using the notation introduced above, we rewrite each term as

$$\begin{aligned}
\mathbb{E}\left[\mathcal{D}\left(\widetilde{Q}_{t+1|\widetilde{W}_{1:t},S} \middle\| \widetilde{Q}_{t+1|\widetilde{W}_{1:t}}\right)\right] &= \int_s \int_{\widetilde{w}_{1:t}} \mathcal{D}\left(\widetilde{Q}_{t+1|\widetilde{w}_{1:t},s} \middle\| \widetilde{Q}_{t+1|\widetilde{w}_{1:t}}\right) \mathrm{d}\widetilde{Q}_{1:t|s}(\widetilde{w}_{1:t}) \mathrm{d}\nu(s) \\
&= \int_s \int_{\widetilde{w}_{1:t}} \Psi_t(\widetilde{w}_{1:t}, s) \mathrm{d}\widetilde{Q}_{1:t|s}(\widetilde{w}_{1:t}) \mathrm{d}\nu(s), \qquad (6)
\end{aligned}$$

where we defined $\Psi_t(\widetilde{w}_{1:t}, s)$ as the relative entropy between $\widetilde{W}_{t+1}$ and $\widetilde{W}'_{t+1}$ conditioned on the previous perturbed iterates $\widetilde{W}_{1:t} = \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}$ and the data set $S = s$. It remains to bound these terms for all t.

To proceed, notice that under these conditions, the updates can be written as

$$\widetilde{W}_{t+1} = \widetilde{w}_t - \eta_t G_t + \varepsilon_t \qquad \text{and} \qquad \widetilde{W}'_{t+1} = \widetilde{w}_t - \eta_t G'_t + \varepsilon'_t.$$

Thus, given the condition $\widetilde{W}_{1:t} = \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}$, the relative entropy between $\widetilde{W}_{t+1}|S$ and $\widetilde{W}'_{t+1}$ equals the relative entropy between $\eta_t G_t - \varepsilon_t|S, \widetilde{w}_{1:t}$ and $\eta_t G'_t - \varepsilon'_t|\widetilde{w}_{1:t}$. Furthermore, under the same condition, we have $G_t = g(W_t, B_t) = g(\widetilde{w}_t - \xi_t, B_t)$ and $G'_t = g(W'_t, B'_t) = g(\widetilde{w}_t - \xi'_t, B'_t)$, so we can appeal to Lemma 4 to obtain the bound

$$\Psi_t(\widetilde{w}_{1:t}, s) \leq \frac{\eta_t^2}{2\sigma_t^2} \mathbb{E}\left[\left\| g(\widetilde{W}_t - \xi_t, B_t) - g(\widetilde{W}_t - \xi'_t, B'_t) \right\|^2 \middle| \widetilde{W}_{1:t} = \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}, S = s \right].$$

Introducing the population gradient $\overline{g}(w) = \mathbb{E}\left[g(w, Z)\right]$, and still using the condition that $\widetilde{W}_t = \widetilde{W}'_t$, we upper-bound the term in the above expectation as follows:

$$\begin{aligned}
\left\| g(\widetilde{W}_t - \xi_t, B_t) - g(\widetilde{W}_t - \xi'_t, B'_t) \right\|^2 &= \left\| g(\widetilde{W}_t - \xi_t, B_t) - \overline{g}(\widetilde{W}_t) + \overline{g}(\widetilde{W}_t) - g(\widetilde{W}_t - \xi'_t, B'_t) \right\|^2 \\
&\leq 2 \left\| g(\widetilde{W}_t - \xi_t, B_t) - \overline{g}(\widetilde{W}_t) \right\|^2 + 2 \left\| \overline{g}(\widetilde{W}_t) - g(\widetilde{W}_t - \xi'_t, B'_t) \right\|^2 \\
&\leq 4 \left\| g(\widetilde{W}_t - \xi_t, B_t) - \overline{g}(\widetilde{W}_t - \xi_t) \right\|^2 + 4 \left\| \overline{g}(\widetilde{W}_t - \xi_t) - \overline{g}(\widetilde{W}_t) \right\|^2 \\
&\quad + 4 \left\| \overline{g}(\widetilde{W}_t - \xi'_t) - g(\widetilde{W}_t - \xi'_t, B'_t) \right\|^2 + \left\| \overline{g}(\widetilde{W}_t - \xi'_t) - \overline{g}(\widetilde{W}_t) \right\|^2 \\
&= 4 \left\| g(W_t, B_t) - \overline{g}(W_t) \right\|^2 + 4 \left\| \overline{g}(W_t) - \overline{g}(W_t + \xi_t) \right\|^2 \\
&\quad + 4 \left\| \overline{g}(W'_t) - g(W'_t, B'_t) \right\|^2 + 4 \left\| \overline{g}(W'_t) - \overline{g}(W'_t + \xi'_t) \right\|^2,
\end{aligned}$$

where each inequality follows from an application of Cauchy–Schwartz. Plugging the result into Equation (6), we are left with integrating all terms with respect to the joint distribution of $(\widetilde{W}_{1:t}, S)$.

To proceed, note that

$$\begin{aligned}
&\mathbb{E}\left[\left\| g(W_t, B_t) - \overline{g}(W_t) \right\|^2 \middle| \widetilde{W}_{1:t} = \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}, S = s \right] \\
&\qquad = \mathbb{E}\left[\left\| g(W_t, B_t) - \overline{g}(W_t) \right\|^2 \middle| \widetilde{W}_{1:t} = \widetilde{w}_{1:t}, S = s \right]
\end{aligned}$$

due to the independence of W_t, B_t from $\widetilde{W}'_t$. Thus, we have

$$\begin{aligned}
&\int_s \int_{\widetilde{w}_{1:t}} \mathbb{E}\left[\left\| g(W_t, B_t) - \overline{g}(W_t) \right\|^2 \middle| \widetilde{W}_{1:t} = \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}, S = s \right] \mathrm{d}\widetilde{Q}_{1:t|s}(\widetilde{w}_{1:t}) \mathrm{d}\nu(s) \\
&\int_s \int_{\widetilde{w}_{1:t}} \mathbb{E}\left[\left\| g(W_t, B_t) - \overline{g}(W_t) \right\|^2 \middle| \widetilde{W}_{1:t} = \widetilde{w}_{1:t}, S = s \right] \mathrm{d}\widetilde{Q}_{1:t|s}(\widetilde{w}_{1:t}) \mathrm{d}\nu(s) \\
&\quad = \mathbb{E}\left[\left\| g(W_t, B_t) - \overline{g}(W_t) \right\|^2 \right].
\end{aligned}$$

Similarly, we observe that

$$\mathbb{E}\left[\left\|g(W'_t, B'_t) - \overline{g}(W'_t)\right\|^2 \middle| \widetilde{W}_{1:t} = \widetilde{w}_{1:t}, \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}, S = s\right] \\ = \mathbb{E}\left[\left\|g(W'_t, B'_t) - \overline{g}(W'_t)\right\|^2 \middle| \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}\right]$$

due to the independence of $\widetilde{W}'_t, B'_t$ from $\widetilde{W}_t$ and S, so we can write

$$\begin{aligned}
&\int_s \int_{\widetilde{w}_{1:t}} \mathbb{E}\left[\left\|g(W'_t, B'_t) - \overline{g}(W'_t)\right\|^2 \middle| \widetilde{W}_{1:t} = \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}, S = s\right] \mathrm{d}\widetilde{Q}_{1:t|s}(\widetilde{w}_{1:t})\mathrm{d}\nu(s) \\
&= \int_s \int_{\widetilde{w}_{1:t}} \mathbb{E}\left[\left\|g(W'_t, B'_t) - \overline{g}(W'_t)\right\|^2 \middle| \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}\right] \mathrm{d}\widetilde{Q}_{1:t|s}(\widetilde{w}_{1:t})\mathrm{d}\nu(s) \\
&= \int_{\widetilde{w}_{1:t}} \mathbb{E}\left[\left\|g(W'_t, B'_t) - \overline{g}(W'_t)\right\|^2 \middle| \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}\right] \mathrm{d}\widetilde{Q}_{1:t}(\widetilde{w}_{1:t}) \\
&= \mathbb{E}\left[\left\|g(W'_t, B'_t) - \overline{g}(W'_t)\right\|^2\right] = \mathbb{E}\left[\left\|g(W_t, B_t) - \overline{g}(W_t)\right\|^2\right],
\end{aligned}$$

where the last step follows from noticing that the marginal distributions of W_t and W'_t are the same.

As for the remaining terms, we first have the following:

$$\begin{aligned}
&\int_s \int_{\widetilde{w}_{1:t}} \mathbb{E}\left[\left\|\overline{g}(W_t + \xi_t) - \overline{g}(W_t)\right\|^2 \middle| \widetilde{W}_{1:t} = \widetilde{w}_{1:t}, S = s\right] \mathrm{d}\widetilde{Q}_{1:t|s}(\widetilde{w}_{1:t})\mathrm{d}\nu(s) \\
&= \int_{\widetilde{w}_t} \mathbb{E}\left[\left\|\overline{g}(W_t + \xi_t) - \overline{g}(W_t)\right\|^2 \middle| \widetilde{W}_t = \widetilde{w}_t\right] \mathrm{d}\widetilde{Q}_t(\widetilde{w}_t) \\
&= \int_{w_t} \int_{\widetilde{w}_t} \mathbb{E}\left[\left\|\overline{g}(w_t + \xi_t) - \overline{g}(w_t)\right\|^2 \middle| \widetilde{W}_t = \widetilde{w}_t, W_t = w_t\right] \mathrm{d}\widetilde{Q}_{t|w_t}(\widetilde{w}_t)\mathrm{d}Q_t(w_t) \\
&= \int_{w_t} \mathbb{E}\left[\left\|\overline{g}(w_t + \xi_t) - \overline{g}(w_t)\right\|^2 \middle| W_t = w_t\right] \mathrm{d}Q_t(w_t) = \mathbb{E}\left[\Gamma_{\sigma_{1:t}}(W_t)\right],
\end{aligned}$$

where the final step follows from noticing that the conditional distribution of ξ_t given W_t is a zero-mean Gaussian with covariance $\sigma^2_{1:t} I$, and recalling the definition of Γ. A similar derivation gives

$$\begin{aligned}
&\int_s \int_{\widetilde{w}_{1:t}} \mathbb{E}\left[\left\|\overline{g}(W'_t + \xi'_t) - \overline{g}(W'_t)\right\|^2 \middle| \widetilde{W}'_{1:t} = \widetilde{w}_{1:t}, S = s\right] \mathrm{d}\widetilde{Q}_{1:t|s}(\widetilde{w}_{1:t})\mathrm{d}\nu(s) \\
&= \int_{\widetilde{w}_t} \mathbb{E}\left[\left\|\overline{g}(W'_t + \xi'_t) - \overline{g}(W'_t)\right\|^2 \middle| \widetilde{W}'_t = \widetilde{w}_t\right] \mathrm{d}\widetilde{Q}_t(\widetilde{w}_t) \\
&= \int_{w'_t} \int_{\widetilde{w}_t} \mathbb{E}\left[\left\|\overline{g}(w'_t + \xi'_t) - \overline{g}(w'_t)\right\|^2 \middle| \widetilde{W}'_t = \widetilde{w}_t, W'_t = w'_t\right] \mathrm{d}\widetilde{Q}_{t|w'_t}(\widetilde{w}_t)\mathrm{d}Q_t(w'_t) \\
&= \int_{w'_t} \int_{\widetilde{w}_t} \mathbb{E}\left[\left\|\overline{g}(w'_t + \xi'_t) - \overline{g}(w'_t)\right\|^2 \middle| W'_t = w'_t\right] \mathrm{d}Q_t(w'_t) = \mathbb{E}\left[\Gamma_{\sigma_{1:t}}(W'_t)\right] = \mathbb{E}\left[\Gamma_{\sigma_{1:t}}(W_t)\right],
\end{aligned}$$

where we again used that the marginal distribution of W'_t matches that of W_t. The proof is concluded by putting everything together. ■

5. Extensions

In this section, we discuss some additional results that can be derived using the techniques developed in previous parts of the paper, as well as propose some open problems for future research.

5.1. Geometry-aware guarantees

One potential criticism regarding the bound of Theorem 1 is that it heavily depends on the parametrization of the loss function. Indeed, measuring the sensitivity of the values and the gradients of the loss function in terms of isotropic Gaussian perturbations is somewhat arbitrary, and can result in very conservative bounds. In particular, the loss function may have better smoothness properties and the gradients could have lower variance when measured in terms of different norms, and the final optimum may be more sensitive to perturbations in certain directions than in other ones. Luckily, our framework allows addressing these issues by using perturbations of a more general form, specifically Gaussian perturbations with general covariance matrices. The key technical result that allows us to take advantage of this generalization is the following simple variant of Lemma 4:

Lemma 6 *Let X and Y be random variables taking values in $\mathbb{R}^d$ with bounded second moments and let Σ be an arbitrary symmetric positive definite matrix. Letting $\varepsilon \sim \mathcal{N}(0, \Sigma)$ be independent of X and Y, the relative entropy between the distributions of $X + \varepsilon$ and $Y + \varepsilon$ is bounded as*

$$\mathcal{D}\left(P_{X+\varepsilon} \| P_{Y+\varepsilon}\right) \leq \frac{1}{2}\mathbb{E}\left[\|X - Y\|^2_{\Sigma^{-1}}\right].$$

Accordingly, we can define the generalized local gradient sensitivity and variance functions

$$\Gamma_{\Sigma,\Sigma'}(w) = \mathbb{E}\left[\|\overline{g}(w) - \overline{g}(w+\xi)\|^2_{\Sigma^{-1}}\right] \quad \text{and} \quad V_{t,\Sigma}(w) = \mathbb{E}\left[\|g(w, B_t) - \overline{g}(w)\|^2_{\Sigma^{-1}} \,\middle|\, W_t = w\right],$$

where $\xi \sim \mathcal{N}(0, \Sigma')$, and adapt the definition of Δ as $\Delta_\Sigma(w, s) = \mathbb{E}\left[L(w, s) - L(w + \xi, s)\right]$ with $\xi \sim \mathcal{N}(0, \Sigma)$. Then, we can prove the following refined version of Theorem 1:

Theorem 7 *Fix any sequence $\boldsymbol{\Sigma} = (\Sigma_1, \Sigma_2, \ldots, \Sigma_T)$ of symmetric positive definite matrices and let $\Sigma_{1:t} = \sum_{k=1}^{t-1} \Sigma_k$ for all t. Then, the generalization error of the final iterate of SGD satisfies*

$$|\mathrm{gen}(W_T, S)| \leq \sqrt{\frac{4R^2}{n}\sum_{t=1}^{T}\eta_t^2 \mathbb{E}\left[\Gamma_{\Sigma_t,\Sigma_{1:t}}(W_t) + V_{t,\Sigma_t}(W_t)\right]} + \left|\mathbb{E}\left[\Delta_{\Sigma_{1:T}}(W_T, S') - \Delta_{\Sigma_{1:T}}(W_T, S)\right]\right|.$$

The proof is left as a straightforward exercise for the reader: one only needs to use the perturbations $\varepsilon_t \sim \mathcal{N}(0, \Sigma_t)$ and apply Lemma 6 instead of Lemma 4 in the proof of Theorem 1. As before, this guarantee comes with the attractive property of simultaneously holding for all possible choices of $\boldsymbol{\Sigma}$, and is thus able to take advantage of potentially hidden geometric properties of the loss landscape.

5.2. Bounds for general learning algorithms

Our core idea of conducting a perturbation analysis of the output of SGD can be easily generalized to prove generalization guarantees for a much broader family of algorithms. In fact, the following bound can be proved without making any assumptions about how the output W is constructed:

Proposition 8 *Let W and W' be the $\mathbb{R}^d$-dimensional outputs of a learning algorithm $\mathcal{A}$ run on the independent and i.i.d. data sets S and S'. Then, the generalization error of $\mathcal{A}$ is bounded as*

$$|\mathrm{gen}(W,S)| \leq \inf_{\Sigma \in \mathcal{S}_+} \left\{ \sqrt{\frac{R^2}{n} \mathbb{E}\left[\|W - W'\|^2_{\Sigma^{-1}} \right]} + \left| \mathbb{E}\big[\Delta_\Sigma(W,S') - \Delta_\Sigma(W,S) \big] \right| \right\}.$$

Notably, this bound only depends on the norms of the parameter vectors and does not involve undesirable quantities like Lipschitz constants or the total number of parameters. While we are aware of generalization bounds with similar qualities for specific families of deep neural networks (e.g., Neyshabur et al., 2015; Bartlett et al., 2017; Neyshabur et al., 2018; Long and Sedghi, 2019), we are not aware of any guarantee of comparable generality in the literature, and we believe that it might be of independent interest. We also believe that this bound can be made tighter in special cases by using the chaining techniques of Asadi et al. (2018) and Asadi and Abbe (2020).

It is tempting to use the above guarantee to study the generalization properties of SGD. As we show below, this gives significantly weaker bounds than our main result. To see this, let us consider the special case $\Sigma = \sigma^2_{1:T} I = \sigma^2 T \cdot I$ and write

$$\frac{1}{2\sigma^2 T} \mathbb{E}\left[\left\| W_T - W'_T \right\|^2 \right] = \frac{1}{2\sigma^2 T} \mathbb{E}\left[\left\| \sum_{t=1}^T \eta_t (G_t - G'_t) \right\|^2 \right] \leq \frac{1}{2\sigma^2} \sum_{t=1}^T \eta_t^2 \mathbb{E}\left[\left\| G_t - G'_t \right\|^2 \right],$$

where the last step follows from Jensen's inequality. While superficially similar to the bound of Theorem 1, this bound is in fact much weaker since it depends on the *marginal* variance of the gradients, which can be very large in general. Nevertheless, assuming that the loss function is C-Lipschitz, we can simply bound $\|G_t - G'_t\| \leq C$ and obtain a mutual-information bound that is comparable to the one proved for SGLD by Pensia et al. (2018): for constant $\eta_t = \eta$ and $\sigma_t = \sigma$, both bounds are of order $\eta^2 C^2 T / \sigma^2$. Still, as previously discussed, the sensitivity term in our bound can be much larger than the excess empirical risk of SGLD, making the guarantee derived using this generic technique weaker overall.

5.3. Variations on SGD

It is natural to ask if our techniques are applicable for analyzing other iterative algorithms besides the vanilla SGD variant considered in the previous sections. We discuss a few possible extensions below.

Momentum, extrapolation, and iterate averaging. Following Pensia et al. (2018), it is easy to incorporate momentum into our updates by simply redefining w as the concatenation of the parameter vector and the rolling geometric average of the gradients, and redefining the function g appropriately. Gradient extrapolation can be also handled similarly—we refer the interested reader to Sections 4.3–4.4 of Pensia et al. (2018) for details of these extensions. An extension not covered in their work is iterate averaging, which can be written as the recursion

$$\begin{bmatrix} W_{t+1} \\ U_{t+1} \end{bmatrix} = \begin{bmatrix} W_t - \eta_t g(W_t, Z_{J_t}) \\ \gamma_t U_t + (1-\gamma_t) W_t \end{bmatrix} = \begin{bmatrix} I & 0 \\ (1-\gamma_t) I & \gamma_t I \end{bmatrix} \begin{bmatrix} W_{t+1} \\ U_{t+1} \end{bmatrix} - \begin{bmatrix} \eta_t g(W_t, Z_{J_t}) \\ 0 \end{bmatrix},$$

where γ_t is a sequence of weights in $[0,1]$ (some common choices being $\frac{t-1}{t}$ leading to uniform averaging or a constant γ that leads to geometric tail-averaging). This SGD variant outputs $U_T =$

$\sum_{t=1}^{T}(1-\gamma_t)\prod_{k=t}^{T}\gamma_k W_t$. It is easy to deduce that our guarantees continue to hold for all these extensions, with the pathwise gradient statistics defined in terms of the appropriately redefined update function g. Thus, our bounds do not show a qualitative improvement in terms of generalization error for such methods, which may seem counterintuitive since interpolation and iterate averaging are known to offer stabilization properties in a variety of settings (Neu and Rosasco, 2018; Mücke et al., 2019; Lin et al., 2020). Not being able to account for this effect seems to be an inherent limitation of our technique, most likely caused by upper bounding the mutual information between $\widetilde{W}_T$ and S by the mutual information between the entire SGD path and S in the very first step of the proof. We do not believe that a simple fix is possible.

Adaptive learning rates and perturbations. While our analysis uses a fixed sequence of learning rates η_t, it may be possible to replace these with adaptive stepsizes and even preconditioners such as the ones used in AdaGrad or ADAM (Duchi et al., 2011; Kingma and Ba, 2015). A straightforward way to do this is including these as part of the function g, but extra care needs to be taken due to the long-term dependence of these stepsizes of the past gradients. We believe that the most common incrementally defined stepsize rules can be incorporated in our framework by appropriately conditioning the distribution of W_t on them, but we leave working out the details of this extension as future work. Similarly, we believe that using adaptive perturbation distributions (i.e., choosing each σ_t as a function of the history) is possible, but is subject to the same challenges. One caveat is that it may be difficult to argue that $\Delta_{\sigma_{1:T}}(W_T, S)$ would be small in this case, due to the complicated dependence between the perturbations, W_T, and S.

6. Discussion

While our work has arguably shed some new light on previously unknown aspects of SGD generalization, our results are definitely far from being truly satisfactory. Indeed, while the key terms concerning the perturbation-sensitivity of the loss function and the variance of the gradients have clear intuitive meaning, it is entirely unclear if they are actually the right quantities for characterizing the generalization properties of SGD. In fact, we believe that it may be extremely challenging to verify our findings empirically and we find it unlikely that matching lower bounds could be shown. Even if the quantities we identify turn out to correctly characterize generalization, our results fail to explain *why* running SGD would ever result in trajectories with these quantities being small and thus good generalization. For these reasons, we prefer to think of our work as a mere first step of a potentially interesting line of research, rather than truly a mature contribution.

One important limitation of our bounds is that they feature a non-standard definition of gradient variance, which can make interpretation of the results somewhat difficult. Indeed, while $V_t(w)$ truly corresponds to the variance of the stochastic gradient evaluated at w in the single-pass case, its meaning is much less clear when performing multiple passes over the data set due to the complicated dependence between the current iterate W_t and the previously sampled data points. Thus, in this case, the effect of the choice of hyperparameters like the minibatch size and the number of passes is difficult to quantify and requires further investigation. On a similar note, we remark that the particular choice of $\overline{g}(W_t)$ as the population gradient in the definitions of Γ and V is not the only possible one, and in fact it can be replaced by any $\sigma(W_{1:t})$-measurable function. A natural choice would be the time-dependent $\overline{g}_t = \mathbb{E}\left[g(w,Z)\middle|\, W_t = w\right]$ which would result in a more natural definition of gradient variance, but a much less interpretable notion of gradient sensitivity. We leave the exploration of other alternatives for future work.

Arguably, the most interesting aspect of our work is the analysis technique we introduced for proving Theorem 5. Of course, our proof technique builds on several elements that are familiar from previous work. In particular, the core idea of adding noise to the output of learning algorithms to ease analysis has been used in numerous works in the past decades. Indeed, early versions of this idea have been proposed by Hinton and van Camp (1993) and Langford and Caruana (2002) from the perspective of PAC-Bayesian generalization bounds (McAllester, 1999, 2013), which approach has been adapted to modern deep neural networks by Dziugaite and Roy (2017). Since then, PAC-Bayesian bounds have been successfully applied to prove a range of generalization bounds for this important setting (see, e.g., Neyshabur et al., 2018; Dziugaite et al., 2021). More broadly, the idea of adding noise to induce stability (and thus better generalization) has also been studied in the literature on differential privacy (Dwork et al., 2006a,b; Chaudhuri et al., 2011; Bassily et al., 2014) and adaptive data analysis (Dwork et al., 2015b,a; Feldman and Steinke, 2017, 2018). That said, we believe that the particular idea of analyzing SGD through the noise decomposition in Equation (5) is indeed novel, and we expect that this idea may have a chance to inspire future analyses of iterative algorithms.

Acknowledgments

G. Neu was supported by "la Caixa" Banking Foundation through the Junior Leader Postdoctoral Fellowship Programme, a Google Faculty Research Award, and the Bosch AI Young Researcher Award. M. Haghifam is supported by the Vector Institute, Mitacs Accelerate Fellowship, and Ewing Rae Scholarship. D. M. Roy was supported, in part, by an NSERC Discovery Grant, Ontario Early Researcher Award, and a stipend provided by the Charles Simonyi Endowment. The first author thanks Gábor Lugosi for illuminating discussions during the preparation of this work, and an anonymous referee who invested serious effort into reviewing the paper and caught a mistake in a previous version of the proof of the main theorem. This bug was fixed by using an improved version of Lemma 4 suggested independently by Tor Lattimore, who the first author also wishes to thank. Finally, the authors are grateful for the suggestions of Borja Rodríguez Gálvez that helped improve the presentation of the results and the rigor of several technical details.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, pages 242–252. PMLR, 2019.

Amir Asadi and Emmanuel Abbe. Chaining meets chain rule: Multilevel entropic regularization and training of neural networks. *Journal of Machine Learning Research*, 21(139):1–32, 2020.

Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7234–7243, 2018.

Peter L Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6241–6250, 2017.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS)*, pages 464–473, 2014.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4381–4391, 2020.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 161–168, 2007.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory (COLT)*, pages 610–626, 2020.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

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

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, pages 1019–1028, 2017.

Simon S Du, Xiyu Zhai, Barnabas Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, pages 486–503, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pages 265–284, 2006b.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2350–2358, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Symposium on Theory of Computing (STOC)*, pages 117–126, 2015b.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes. In *Artificial Intelligence and Statistics (AISTATS)*, pages 604–612, 2021.

Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In *Conference on Learning Theory (COLT)*, pages 728–757, 2017.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory (COLT)*, pages 535–544, 2018.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory (COLT)*, pages 1270–1279, 2019.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory (COLT)*, pages 797–842, 2015.

Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, pages 1225–1234, 2016.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2553–2564, 2019.

Geoffrey E Hinton and Drew van Camp. Keeping neural networks simple by minimising the description length of weights. In *Computational Learning Theory (COLT)*, pages 5–13, 1993.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Uncertainty in Artificial Intelligence (UAI)*, pages 876–885, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8580–8589, 2018.

J. H. B. Kemperman. On the Shannon capacity of an arbitrary channel. In *Indagationes Mathematicae (Proceedings)*, volume 77, pages 101–115, 1974.

Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Harold Kushner and George Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York, 1997.

John Langford and Rich Caruana. (not) bounding the true error. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 809–816, 2002.

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory (COLT)*, pages 1246–1257, 2016.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, pages 5809–5819, 2020.

Tao Lin, Lingjing Kong, Sebastian Stich, and Martin Jaggi. Extrapolation for large-batch training in deep learning. In *International Conference on Machine Learning (ICML)*, pages 6094–6104, 2020.

Philip M Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning (ICML)*, pages 3325–3334, 2018.

David McAllester. A PAC-Bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013.

David A McAllester. PAC-Bayesian model averaging. In *Computational Learning Theory (COLT)*, pages 164–170, 1999.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating SGD saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12568–12577, 2019.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11013–11023, 2019.

Arkadii Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Conference on Learning Theory (COLT)*, pages 3222–3242, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory (COLT)*, pages 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning (ICML)*, pages 4951–4960, 2019.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *International Symposium on Information Theory (ISIT)*, pages 546–550. IEEE, 2018.

Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1-2):1–246, 2013. ISSN 1567-2190.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory (COLT)*, pages 1674–1703, 2017.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

Borja Rodríguez-Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm. In *Information Theory Workshop (ITW)*, pages 1–5, 2021.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error backpropagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Volume 1: Foundations*, chapter 8. The MIT Press, Cambridge, MA, 1986.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1232–1240, 2016.

Daniel Russo and James Zou. How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory (COLT)*, pages 3437–3452, 2020.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2524–2533, 2017.

Mingtian Zhang, Peter Hayes, Thomas Bird, Raza Habib, and David Barber. Spread divergence. In *International Conference on Machine Learning (ICML)*, pages 11106–11116, 2020.

Appendix A. The proof Lemma 4

Let us denote the joint distribution of X, Y by $P_{X,Y}$ and observe that the respective distributions of $X+\varepsilon$ and $Y+\varepsilon$ can be written as

$$P_{X+\varepsilon} = \int_{x,y} \mathcal{N}(x, \sigma^2 I) \mathrm{d}P_{X,Y}(x,y) \quad \text{and} \quad P_{Y+\varepsilon} = \int_{x,y} \mathcal{N}(y, \sigma^2 I) \mathrm{d}P_{X,Y}(x,y),$$

where $\mathcal{N}(x, \sigma^2 I)$ is the Gaussian distribution with mean x and covariance $\sigma^2 I$. Using this observation, we can write

$$\begin{aligned}
\mathcal{D}\left(P_{X+\varepsilon} \| P_{Y+\varepsilon}\right) &= \mathcal{D}\left(\int_{x,y} \mathcal{N}(x, \sigma^2 I) \mathrm{d}P_{X,Y}(x,y) \middle\| \int_{x,y} \mathcal{N}(y, \sigma^2 I) \mathrm{d}P_{X,Y}(x,y)\right) \\
&\leq \int_{x,y} \mathcal{D}\left(\mathcal{N}(x, \sigma^2 I) \middle\| \mathcal{N}(y, \sigma^2 I)\right) \mathrm{d}P_{X,Y}(x,y) \\
&= \mathbb{E}_{X,Y}\left[\mathcal{D}\left(\mathcal{N}(X, \sigma^2 I) \middle\| \mathcal{N}(Y, \sigma^2 I)\right)\right] = \frac{1}{2\sigma^2} \mathbb{E}_{X,Y}\left[\|X-Y\|^2\right],
\end{aligned}$$

where the second line uses Jensen's inequality and the joint convexity of $\mathcal{D}\left(\cdot \| \cdot\right)$ in its arguments, and the last line follows from noticing that $\mathcal{D}\left(\mathcal{N}(x, \Sigma) \| \mathcal{N}(y, \Sigma)\right) = \frac{1}{2}\|x-y\|^2_{\Sigma^{-1}}$ for any x, y and any symmetric positive definite covariance matrix Σ. ■