ON THE GENERALIZATION OF MODELS TRAINED WITH SGD: INFORMATION-THEORETIC BOUNDS AND IMPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper follows up on a recent work of Neu (2021) and presents new and tighter information-theoretic upper bounds for the generalization error of machine learning models, such as neural networks, trained with SGD. We apply these bounds to analyzing the generalization behaviour of linear and two-layer ReLU networks. Experimental study based on these bounds provide some insights on the SGD training of neural networks. They also point to a new and simple regularization scheme which we show performs comparably to the current state of the art.

1 INTRODUCTION

The observation that high capacity deep neural networks trained with mini-batched stochastic gradient descent, referred to SGD in this paper, tend to generalize well (Zhang et al., 2017) contradicts the classical wisdom in statistical learning theory (e.g., Vapnik (1998)) and has stimulated intense research interest in understanding the generalization behaviour of modern neural networks.

In this direction, generalization bounds for over-parameterized neural networks are obtained (Du et al., 2018; Allen-Zhu et al., 2019; Bartlett et al., 2017; Neyshabur et al., 2015; 2018a;b; Arora et al., 2018; 2019) and a curious "double descent" phenomenon is observed and analyzed (Belkin et al., 2019; Nakkiran et al., 2019; Yang et al., 2020). Built on a connection between stability and generalization (Bousquet & Elisseeff, 2002), a stability-based bound is first presented in Hardt et al. (2016), followed by a surge of research effort exploiting similar approaches (London, 2017; Chen et al., 2018; Feldman & Vondrak, 2019; Lei & Ying, 2020; Bassily et al., 2020). Information-theoretic bounding techniques established recently (Russo & Zou, 2016; 2019; Xu & Raginsky, 2017; Asadi et al., 2018; Bu et al., 2020; Steinke & Zakynthinou, 2020; Dwork et al., 2015; Bassily et al., 2018; Asadi & Abbe, 2020; Hafez-Kolahi et al., 2020; Zhou et al., 2020) have also demonstrated great power in analyzing SGD-like algorithms. For example, Pensia et al. (2018) is the first to utilize information-theoretic bound in analyzing the generalization ability of SGLD (Gelfand & Mitter, 1991; Welling & Teh, 2011). The bound was subsquently improved by Negrea et al. (2019); Haghifam et al. (2020); Rodríguez-Gálvez et al. (2020); Wang et al. (2021). Inspired by the work of Pensia et al. (2018), Neu (2021) presents an information- theoretic analysis of the models trained with SGD. The analysis of Neu (2021) constructs an auxiliary weight process parallel to SGD training and upper-bounds the generalization error through this auxiliary process.

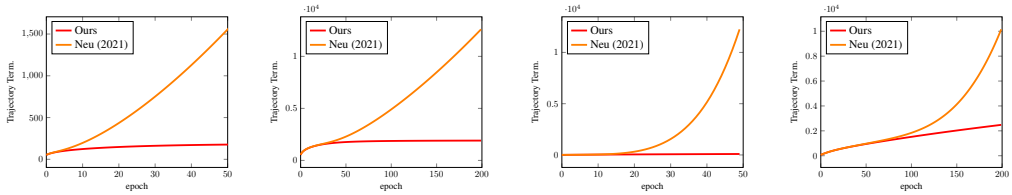
Another line of research connects the generalization of neural networks with the flatness of loss minima (Hochreiter & Schmidhuber, 1997) found by SGD or its variant (Keskar et al., 2017; Dinh et al., 2017; Dziugaite & Roy, 2017; Neyshabur et al., 2017; Chaudhari et al., 2017; Jastrzebski et al., 2017; Jiang et al., 2019; Zheng et al., 2021; Foret et al., 2020). This understanding has led to the discovery of new SGD-based training algorithms for improved generalization. For example, in a concurrent development by Zheng et al. (2021) and Foret et al. (2020), a local "max-pooling" operation is applied to the loss landscape prior to the SGD updates. This approach, referred to as AMP (Zheng et al., 2021) or SAM (Foret et al., 2020), is shown to make SGD favor flatter minima and achieve the state-of-the-art performance among various competitive regularization schemes.

In this paper, we focus on investigating the generalization of machine learning models trained with SGD. Although we are primarily motivated by the curiosity to understanding neural networks, the results of this paper in fact broadly apply broadly to any model trained with SGD.

Our development builds upon that of Neu (2021). Following the same construction of the auxiliary weight process in Neu (2021), we present upper bounds of generalization error that improve upon Neu (2021) in two ways. This improvement is mainly due to removing an unnecessary term in the bound of Neu (2021), which the author refers to as "local gradient sensitivity", by invoking the HWI inequality (Raginsky & Sason, 2018) (i.e., Lemma 4 of this paper). Additionally, the improvement also benefits from replacing a sample-level mutual information term in Neu (2021) with an instance-level mutual information term, exploiting a recent result of Bu et al. (2020). The bounds we obtain decompose into two terms, one measuring the impact of training trajectories ("the trajectory term") and the other measuring the impact of the flatness of the found solution ("the flatness term").

Remarkably, the generalization bound given in Neu (2021) also contains a trajectory term and a flatness term, where the flatness term is identical to ours due to the same construction of the auxillary weight process. However, our trajectory term is much tighter than that in Neu (2021). Figure 1 shows an experimental comparison of the trajectory term in our bound (Theorem 2) with that in the bound in Neu (2021) (re-stated as Lemma 3 in this paper[1]) for two neural network models. The trajectory terms are compared for two different settings of σ, which is the variance parameter of the noise used in the construction of the auxiliary weight process.

After the the initial preprint of Neu (2021), a revised version of the paper Neu et al. (2021) is presented, which the bound is revised by fixing a coefficient error in the bound of Neu (2021) and adopts a slightly different form. It is particularly note-worthy that the development of the bound in Neu et al. (2021) also makes uses of the HWI inequality (or Lemma 4), but the unnecessary local gradient sensitivity term remains in the bound. This is because the development of Neu et al. (2021) does not make full use of the condition in Lemma 4, which requires no independence in the involved random variables. We will elaborate on this in an appropriate context.



(a) $\sigma = 10^{-5}$ (MNIST) (b) $\sigma = 10^{-6}$ (MNIST) (c) $\sigma = 10^{-5}$ (CIFAR10) (d) $\sigma = 10^{-6}$ (CIFAR10)

Figure 1: Comparison of the trajectory term between our bound (Theorem 2) and the bound in Neu (2021). (a)(b) MLP trained on MNIST. (c)(d) CNN trained on CIFAR-10.

We also provide an application of our bounds is in analyzing the generalization behaviour of linear and two-layer ReLU networks, where we show that the activation state in ReLU networks plays an important role in generalization.

It is remarkable that removing the local gradient sensitivity term not only makes the bound much tighter, it also makes it much easier to evaluate. In this paper, we empirically validate the derived bounds, and provide various insights pertaining to the generalization behavior of models trained with SGD. For example, a particular interesting observation from our experiments is that a key quantity arising in the trajectory term of the bounds, which we refer to as *gradient dispersion*[2], reveals a double descent phenomenon with respect to training epochs. Most intriguingly, the valley in the double descent curve appears to mark the great divide between the "generalization regime" and the "memorization regime" of training. Furthering from this observation, we also show that it is possible to reduce the memorization effect by dynamically clipping the gradient and reducing its dispersion.

Our bounds also inspire a natural and simple solution to alleviate generalization error. Specifically, we propose a new training scheme, referred to as *Gaussian model perturbation* (GMP), aiming at reducing the flatness term of the bounds. This scheme effectively applies a local "average pooling" to the empirical risk surface prior to SGD, greatly resembling the "max-pooling" approach adopted in AMP (Zheng et al., 2021). We demonstrate experimentally that GMP achieves a competitive performance with the current art of regularization schemes.

[1]The original bound in Neu (2021) has an error in a coefficient, which we fixed in this paper.

[2]The quantity is often referred to as gradient variance in the literature, but we prefer "dispersion" to "variance" so as to better comply with the mathematical conventions and avoid possible confusion.

Proofs, additional discussions and experimental results are presented in Appendices.

Other Related Literature Gradient dispersion is mostly studied from optimization perspectives (Bottou et al., 2018; Roux et al., 2012; Johnson & Zhang, 2013; Wen et al., 2020; Faghri et al., 2020). Prior to this work, only a few works relate gradient dispersion with the generalization behaviour of the networks. In Neu (2021) and Wang et al. (2021), gradient dispersion also appears in the generalization bounds. In Jiang et al. (2019), gradient dispersion is argued to capture a notion of "flatness" of the local minima of the loss landscape, thereby correlating with generalization. Injecting noise in the training process has been proposed in various regularization schemes (Bishop, 1995; Camuto et al., 2020; 2021; Srivastava et al., 2014; Wei et al., 2020). But unlike GMP derived in this paper, where noise is injected to the model parameters, noise in those schemes is injected either to the training data or to the network activation.

2 PRELIMINARIES

Population Risk, Empirical Risk and Generalization Error Unless otherwise noted, a random variable will be denoted by a capitalized letter (e.g., Z), and its realization denoted by the corresponding lower-case letter (e.g. z). Let $\mathcal{Z}$ be the instance space of interest and μ be an unknown distribution on $\mathcal{Z}$, specifying random variable Z. Let $\mathcal{W} \subseteq \mathbb{R}^d$ be the space of hypotheses. Suppose that a training sample $S = (Z_1, Z_2, \ldots, Z_n)$ is drawn i.i.d. from μ and that a stochastic learning algorithm $\mathcal{A}$ takes S as its input and outputs a hypothesis $W \in \mathcal{W}$ according to some conditional distribution $P_{W|S}$ mapping $\mathcal{Z}^n$ to $\mathcal{W}$. Let $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ be a loss function, where $\ell(w, z)$ measures the "unfitness" or "error" of any $z \in \mathcal{Z}$ with respect to a hypothesis $w \in \mathcal{W}$. The population risk, for any $w \in \mathcal{W}$, is defined as

$$L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)].$$

The goal of learning is to find a hypothesis w that minimizes the population risk. But since μ is only partially accessible via the sample S, in practice, we instead turn to the empirical risk, defined as

$$L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i).$$

The expected generalization error of the learning algorithm $\mathcal{A}$ is then defined as

$$\mathrm{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)],$$

where the expectation is taken over the joint distribution of (S, W) (i.e., $\mu^n \otimes P_{W|S}$).

Throughout this paper, we take ℓ as a continuous function (adopting the usual notion "surrogate loss"). Additionally, we assume that ℓ is differentiable almost everywhere with respect to both w and z. Furthermore we assume that $\ell(w, Z)$ is R-subgaussian[3] for any $w \in \mathcal{W}$. Note that a bounded loss is guaranteed to be subgaussian. Let $I(X; Y)$ denote the mutual information (Cover & Thomas, 2012) between any pair of random variables (X, Y). The following results are known.

Lemma 1 (Xu & Raginsky (2017, Theorem 1.)). *The generalization error of $\mathcal{A}$ is bounded by*

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

Lemma 2 (Bu et al. (2020, Proposition 1.)). *The generalization error of $\mathcal{A}$ is bounded by*

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2R^2 I(W; Z_i)},$$

Stochastic Gradient Descent We now restrict the learning algorithm $\mathcal{A}$ to be the mini-batched SGD algorithm for empirical risk minimization. For each training epoch, the dataset S is randomly split into m disjoint mini-batches, each having size b, namely, $n = mb$. Based on each batch, one parameter update is performed. Specifically, let B_t denote the batch used for the tth update. Define

$$g(w, B_t) \triangleq \frac{1}{b}\sum_{z\in B_t} \nabla_w \ell(w, z),$$

[3] A random variable X is R-subgaussian if for any ρ, $\log \mathbb{E}\exp\left(\rho\left(X - \mathbb{E}X\right)\right) \leq \rho^2 R^2/2$.

namely, $g(w, B_t)$ is the average gradient computed for the batch B_t with respect to parameter w. The rule for the tth parameter update is then

$$W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t),$$

where λ_t is the learning rate at the step t. The initial parameter setting W_0 is assumed to be drawn from the zero-mean spherical Gaussian $\mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$ with variance σ_0^2 in each dimension. We will assume that the SGD algorithm stops after T updates and outputs W_T as the learned model parameter.

Given the training sample S, let ξ govern the randomness in the sequence $(B_1, B_2, \ldots, B_T)$ of batches. For the simplicity of notion, we will fix the configuration of ξ. That is, we will assume a fixed "batching trajectory", or a fixed way to shuffle the example indices $\{1, \ldots, n\}$ and divide them into m batches in each epoch. The presented generalization bounds of this paper can be extended to the case where the batching trajectory is uniformly random (as we set up above). This merely involves averaging over all batching trajectories or taking expectation over ξ.

Auxiliary Weight Process We now associate with the SGD algorithm an auxiliary weight process $\{\widetilde{W}_t\}$. Let $\sigma_1, \sigma_2, \ldots, \sigma_T$ be a sequence of positive real numbers. Define

$$\widetilde{W}_0 \triangleq W_0, \quad \text{and} \;\; \widetilde{W}_t \triangleq \widetilde{W}_{t-1} - \lambda_t g(W_{t-1}, B_t) + N_t, \;\text{for}\; t > 0,$$

where $N_t \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I}_d)$ is a Gaussian noise. The relationship between this auxiliary weight process $\{\widetilde{W}_t\}$ and the weight process $\{W_t\}$ in SGD is shown in the Bayesian network below.

$$\begin{array}{ccccccccccc}
 & & N_1 & & N_2 & & \cdots & & N_{T-1} & & N_T \\
 & & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
\widetilde{W}_0 & \rightarrow & \widetilde{W}_1 & \rightarrow & \widetilde{W}_2 & \rightarrow & \cdots & \rightarrow & \widetilde{W}_{T-1} & \rightarrow & \widetilde{W}_T \\
\| & \nearrow & & \nearrow & & \nearrow & & & & \nearrow & \\
W_0 & \rightarrow & W_1 & \rightarrow & W_2 & \rightarrow & \cdots & \rightarrow & W_{T-1} & \rightarrow & W_T
\end{array}$$

Let $\Delta_t = \sum_{\tau=1}^{t} N_\tau$. Noting that the weight updates in $\{\widetilde{W}_t\}$ uses the same gradient signal as that used in $\{W_t\}$ (which depends on W_{t-1} not $\widetilde{W}_{t-1}$), it is immediate that $\widetilde{W}_t = W_t + \Delta_t$. Note that this auxiliary process follows the same construction as Neu (2021), which we will use to study the generalization error of SGD. To that end, let's define *gradient dispersion* at parameter $W = w$ by

$$\mathbb{V}(w) \triangleq \mathbb{E}\left[||\nabla_w \ell(w, Z) - \mathbb{E}\left[\nabla_w \ell(w, Z)\right]||_2^2\right],$$

where the expectation is taken over $Z \sim \mu$. For a given sample $s \in \mathcal{Z}^n$, define

$$\gamma(w, s) \triangleq \mathbb{E}\left[L_s(w + \Delta_T) - L_s(w)\right],$$

where the expectation is taken over Δ_T and $L_s(w)$ is the empirical risk of s at parameter w.

In the remainder of the paper, let S' denote another sample drawn from μ^n, independent of all other random variables. The main generalization bound in Neu (2021) is re-stated below.

Lemma 3 ((Neu, 2021, Theorem 1.))**.** *The generalization error of SGD is upper bounded by*

$$|\mathrm{gen}(\mu, P_{W_T|S})| \leq \sqrt{\frac{4R^2}{n} \sum_{t=1}^{T} \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}\left[\Psi(W_{t-1}) + \frac{\mathbb{V}(W_{t-1})}{b}\right]} + \left|\mathbb{E}\left[\gamma(W_T, S) - \gamma(W_T, S')\right]\right|,$$

where $\Psi(w_{t-1}) \triangleq \mathbb{E}\left[||\mathbb{E}\left[g(w_{t-1}, Z)\right] - \mathbb{E}\left[g(w_{t-1} + \zeta, Z)\right]||_2^2\right]$ *and* $\zeta \sim \mathcal{N}(0, 2\sum_{i=1}^{t-1} \sigma_i^2 \mathrm{I}_d)$.

The term $\Psi(w_{t-1})$ in the bound is referred to as "local gradient sensitivity" in Neu (2021).

3 New Generalization Bounds for SGD

The first improvement offered by our new bound upon the bound in Neu (2021) or Lemma 3 is the removal of the local sensitivity term $\Psi(w_{t-1})$. This requires invoking a vector version of the HWI inequality (Raginsky & Sason, 2018, Lemma 3.4.2), which we first state.

Lemma 4. *Let X and Y be two random vectors in $\mathbb{R}^d$, and let $N \sim \mathcal{N}(0, \mathbf{I}_d)$ be independent of (X, Y). Then, for every $t > 0$, $\mathrm{D}_{\mathrm{KL}}(P_{X+\sqrt{t}N}||P_{Y+\sqrt{t}N}) \leq \frac{1}{2t}\mathbb{W}_2^2(P_X, P_Y)$ where $\mathbb{W}_2$ is the L2-Wasserstein distance and D_{KL} is the KL divergence.*

Note that in this lemma, it is not required that X and Y are independent and the lemma holds for any joint distribution of (X, Y). Exploiting this fact allows us to develop the following generalization bound, in which, compared with Lemma 3 the local gradient sensitivity term $\Phi(W_{t-1})$ is removed. We elaborate on this in Appendix B.1.

Theorem 1. *The generalization error of SGD is upper bounded by*

$$|\mathrm{gen}(\mu, P_{W_T|S})| \leq \sqrt{\frac{4R^2}{nb}\sum_{t=1}^{T}\frac{\lambda_t^2}{\sigma_t^2}\mathbb{E}\left[\mathbb{V}(W_{t-1})\right]} + \left|\mathbb{E}\left[\gamma(W_T, S) - \gamma(W_T, S')\right]\right|.$$

Removing $\Phi(W_{t-1})$ will make the bound much tighter as shown in Figure 1. This remarkable improvement should come at no surprise, since $\Phi(W_{t-1})$ monotonically increases with training epochs (noting that $\Phi(W_{t-1})$ has the cumulative variance $2\sum_{i=1}^{t-1}\sigma_i^2\mathrm{I}_d$) while $\mathbb{V}(W_{t-1})$ appears decreasing (see Figure 4a). Figure 1 also indicates that if the noise variance σ_t is made small, one expects $\mathbb{V}(W_{t-1})$ to dominate the trajectory term in Lemma 3. However, the factor $1/\sigma_t^2$ will become too large, make the bound loose.

The proof of Theorem 1, as Lemma 3, relies on the sample-level mutual information bound in Lemma 1. This theorem can be further tightened by exploiting the fact that the instance-level mutual information bound in Lemma 2 is in fact tighter than the sample-level mutual information bound in Lemma 1, as shown in Bu et al. (2020). The main ingredient to proceed is the following lemma.

Lemma 5. *Let $G_t = -\lambda_t g(W_{t-1}, B_t)$. If $Z_i \in B_t$, then $I(G_t + N_t; Z_i|\widetilde{W}_{t-1}) \leq \frac{2\lambda_t^2}{\sigma_t^2 b^2}\mathbb{E}\left[\mathbb{V}(W_{t-1})\right]$.*

In this lemma, the mutual information $I(G_t + N_t; Z_i|\widetilde{W}_{t-1})$ roughly indicates the degree by which the SGD's updating signal G_t (smoothed with noise) depends on an individual training instance Z_i, when Z_i is used for computing the gradient. When this dependency is strong (giving rise to a high value of the mutual information), the model conceivably tends to overfit the individual training instances. This lemma suggests that the strength of this dependency can be upper-bounded by the expected gradient dispersion at the current weight configuration. In our experiments, we will estimate the expected gradient dispersion and validate this intuition.

It is remarkable that the noise $\{N_t\}$ plays an important role for the bound to hold. To see this, consider $b = 1$ and $\mathcal{Z}$ is countable and large. Then $I(G_t; Z_t|W_{t-1})$ is merely the conditional entropy $H(Z_i|W_{t-1})$, which would grow with sample size n at least as $\log n$. Upper-bounding it with a quantity independent of n would be impossible – This justifies the construction of the auxiliary weight process.

We now state our main theorem. Unlike Lemma 3, which considers a random batching trajectory, this theorem considers a fixed batching trajectory to keep the expression less cluttered. For that batching trajectory, we will use $\mathcal{T}_i$ to denote the set of indices of batches B_t containing instance Z_i.

Theorem 2. *The expected generalization error of SGD is bounded by*

$$|\mathrm{gen}(\mu, P_{W_T|S})| \leq \frac{2R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{\lambda_t^2}{\sigma_t^2}\mathbb{E}\left[\mathbb{V}(W_{t-1})\right]} + \left|\mathbb{E}\left[\gamma(W_T, S) - \gamma(W_T, S')\right]\right|. \tag{1}$$

Further, assume $L_\mu(w_T) \leq \mathbb{E}_\Delta\left[L_\mu(w_T + \Delta_T)\right]$, ℓ is twice differentiable, and σ_t^2 is independent of t. Denote by H_{W_T} the Hessian of the loss with respect to W_T and let $\mathrm{Tr}(\cdot)$ denote trace. Then

$$\mathrm{gen}(\mu, P_{W_T|S}) \leq 3\left(\frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\lambda_t^2\mathbb{E}\left[\mathbb{V}(W_{t-1})\right]}\right)^{\frac{2}{3}}\left(\frac{T}{2}\mathbb{E}\left[\mathrm{Tr}\left(\mathrm{H}_{W_T}(Z)\right)\right]\right)^{\frac{1}{3}} \tag{2}$$

The condition $L_\mu(w_T) \leq \mathbb{E}_\Delta\left[L_\mu(w_T + \Delta_T)\right]$, which is also assumed to derive a PAC-Bayesian bound in Foret et al. (2020), indicates that the perturbation does not decrease the population risk.

In the bound of Eq.1, the first term captures the impact of the training trajectory ("trajectory term"), and the second term captures the impact of the final solution, which in fact measures the flatness for the loss landscape at the found solution ("flatness term"). It is worth noting that using both the gradient dispersion (or the trace of the gradient's covariance matrix), and the trace of Hessian

to characterize generalization is not new. For example, the Takeuchi information criterion used by Thomas et al. (2020) consists of Hessian and gradient dispersion (referred to as "curvature" and "noise" by the authors), which is shown relevant to estimating the generalization gap.

The learning rate and batch size in SGD have explicitly appeared in the trajectory term of Eq.1 in Theorem 2. From the way they appear in the bound, one may be tempted to assert that a small learning rate or large batch size will improve generalization. This would then contradict some previous observations (Jastrzebski et al., 2017; Wu et al., 2018; He et al., 2019), in which increasing the ratio of learning rate to batch size will benefit generalization. We investigate this by performing experiments with varying learning rates and batch sizes (see Figure 7 in Appendix E.2). As seen in the experiments, increasing the learning rate impacts the trajectory term and the flatness term in opposite ways, i.e. making one increase and the other decrease. A similar (but reverted) behaviour is also observed with batch sizes. This makes the generalization bound in Eq.1, have a rather complex relationship with the settings of learning rate and batch size.

Eq.2 follows from Eq.1 by minimizing the bound over σ. The bound in Eq.2 is thus independent of a choice of σ_t's and can be computed easily and efficiently.

Lemma 3 and its tightened version in Theorem 1 also similarly contain a trajectory term and a flatness term. Despite that the flatness term there are identical to that in Theorem 2, we now show the trajectory term in Theorem 2 does improve on its counter-part in Theorem 1.

Lemma 6. *Assume that the training instances are sampled without replacement in every epoch. Then the trajectory term in Eq.1 of Theorem 2 is upper-bounded by* $\sqrt{\frac{4R^2}{nb}\sum_{t=1}^{T}\frac{\lambda_t^2}{\sigma_t^2}\mathbb{E}\left[\mathbb{V}(W_{t-1})\right]}$.

The condition in Lemma 6 is usually satisfied in practice. This lemma then immediately implies that the trajectory term in Theorem 2 is no worse than that in Theorem 1.

To summarize, we remark that these bounds suggest that in order for the model to generalize well, both the trajectory term and the flatness term need to be small — the former involves the interaction of the learning rate and batch size with the gradient dispersion along the training trajectory, whereas the latter depends on the flatness of the empirical risk surface at the found solution.

We now apply Theorem 2 to two neural network models in a regression setting. Let $Z = [X, Y] \in \mathbb{R}^{d_0+1}$. Assume $||X|| = 1$ and $|Y| \leq 1$. We use SGD to train a model $f(W, \cdot) : \mathbb{R}^{d_0} \to \mathbb{R}$ and define the loss as $\ell(W, Z) = 1/2(Y - f(W, X))^2$.

Theorem 3 (Linear Networks)**.** *Let $f(W, X) = W^T X$. Then,*

$$\mathrm{gen}(\mu, P_{W_T|S}) \leq 3\left(\frac{RT^2}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\lambda_t^2\mathbb{E}\left[(4d_0^2\ell(W_{t-1}, Z_i) + 2||W_{t-1}||^2 + 1)\right]}\right)^{\frac{2}{3}}.$$

Theorem 4 (Two-Layer ReLU Networks)**.** *Following Arora et al. (2019), consider $f(W, X) = \frac{1}{\sqrt{m}}\sum_{r=1}^{m} A_r \mathrm{ReLU}(W_r^T X)$ where m is the width of the neural network, $A_r \sim \mathrm{unif}(\{+1, -1\})$ and $\mathrm{ReLU}(\cdot)$ is the ReLU activation. We only train the first layer parameters $W = [W_1, W_2, \ldots, W_m] \in \mathbb{R}^{d_0\times m}$ and fix the second layer parameters $A = [A_1, A_2, \ldots, A_m] \in \mathbb{R}^m$ during training. Then,*

$$\mathrm{gen}(\mu, P_{W_T|S}) \leq 3\left(\frac{T}{m}\sum_{r=1}^{m}\mathbb{E}\left[\mathbb{I}_{r,i}^T\right]\right)^{\frac{1}{3}}\left(\frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\lambda_t^2\mathbb{E}\left[\sum_{r=1}^{m}\frac{\mathbb{I}_{r,i}^{t-1}}{m}\left(4d_0^2\ell(W_{t-1}, Z_i) + \frac{||2W_{t-1,r}||^2}{m} + 1\right)\right]}\right)^{\frac{2}{3}},$$

where $\mathbb{I}_{r,i}^{t-1} = \mathbb{I}\{W_{t-1,r}^T X_i \geq 0\}$, $\mathbb{I}$ is the indicator function and $W_{t-1,r}$ is the rth column of W_{t-1}.

Figure 2 shows that this bound can characterize the dynamics of generalization gap of the ReLU network. We defer the implementation details including the linear network's plot to Appendix.

Compared with Theorem 3, we notice that in the two-layer ReLU network, the ReLU activation state along the training trajectory plays a key role in the bound of Theorem 4. Specifically, the weights of the deactivated neurons do not contribute to the bound of Theorem 4, making the bound not explicitly depend on the model dimension d. This result also suggests that sparsely activated ReLU networks are expected to generalize better. Despite various empirical evidence pointing to this behaviour (see, e.g., Glorot et al. (2011)), to the best of our knowledge, this theorem provides the first theoretical justification in this regard.

In the remainder of this paper, we will study more complex network architectures beyond the linear and two-layer ReLU networks and implications of our bounds in those settings.

4 EXPERIMENTAL STUDY

Bound Verification We first verify our bound of Eq.2 in Theorem 2 by training an MLP (with one hidden layer) and an AlexNet (Krizhevsky et al., 2012) on MNIST and CIFAR10 (Krizhevsky, 2009), respectively. To simplify estimation, we fix the weight initialization and set σ_t and λ_t to be constants σ and λ, respectively. To compute $\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\mathbb{E}\left[\mathbb{V}(W_{t-1})\right]}$, we compute the gradient dispersion as its empirical estimate from a batch, utilizing a PyTorch (Paszke et al., 2019) library BackPack (Dangel et al., 2020). To compute $\mathrm{Tr}\left(\mathbb{E}\left[\mathrm{H}_{W_T}(Z)\right]\right)$, we use the PyHessian library (Yao et al., 2020) to compute the Hessian.

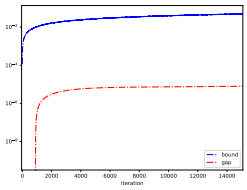
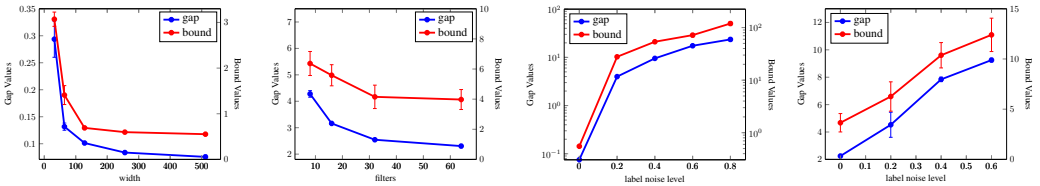


Figure 2: Generalization gap of a two-layer ReLU network v.s. Theorem 4.

We perform experiments with varying network width and varying levels of label noise. Specifically, noise level ϵ refers to the setting where we replace the labels of ϵ fraction of the training and testing instances with random labels. The estimated bound is compared against the true generalization gap, namely, the difference between the training loss and testing loss, and is shown in Figure 3.



(a) MLP on MNIST (b) AlexNet on CIFAR10 (c) MLP on MNIST (d) AlexNet on CIFAR10

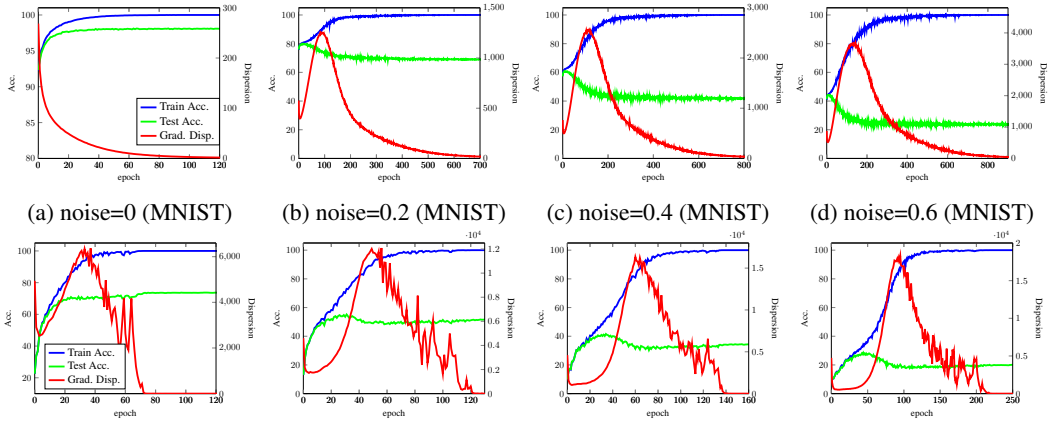
Figure 3: Estimated bound and empirical generalization gap ("gap") as functions of network width ((a) and (b)) and label noise level ((c) and (d)). Left Y-axis: gap value; right Y-axis: bound value.

In Figure 3, we see that in all cases the estimated bound follows closely the trend of the true generalization gap. The fact that the bound curve consistently tracks the gap curve under various label noise levels indicates that our bound very well captures the changes of the data distribution. Note that in Figure 3 (a) and (b), our bound decays with the increase of the model size, showing a trend as opposite to the bounds obtained in classical learning theory. But such a trend clearly better explains the generalization behaviour of modern neural networks.

Epoch-wise Double Descent of Gradient Dispersion We experimentally investigate the impact of gradient dispersion on the training of the neural networks by fixing the learning rate, batch size and weight initialization for the each model (MLP for MNIST, AlexNet for CIFAR10). For each model and various label noise levels, we plot in Figure 4 the evolution of the (empirical) gradient dispersion $\widehat{\mathbb{V}}(w_t)$, training accuracy and testing accuracy across training epochs.

An intriguing epoch-wise "double descent" phenomenon is observed, particularly when the labels are noisy. According to the double descent curve, the training may be split into three phases (e.g., Figure 4 (h)). In the first phase, the gradient dispersion rapidly descends and maintains a very low level. In this phase, both training and test accuracies increase while maintaining a very small generalization gap. This suggests that the network in this phase is extracting useful patterns and generalizes well. In the second phase, the gradient dispersion starts increasing until it reaches a peak value. In this phase, the training and testing accuracies gradually diverge, marking the model entering an overfitting or "memorization" regime – when the data contains the noisy labels, the network mostly tries to memorize the labels in the training set. In the third phase, the gradient dispersion descends again, reaching a low value. In this phase, the model continuously overfits the training data, until the training and testing curves reach their respective maximum and minimum. It appears that the timing of the three phases depends on the dataset and the label noise level. For simpler data (e.g. MNIST) and cleaner datasets , the first phase may be shorter. This is arguably

because in these datasets, extracting useful patterns is relatively easier. Nonetheless, the valley in the double-descent curve appears to mark a "great divide" between generalization and memorization.



(a) noise=0 (MNIST) (b) noise=0.2 (MNIST) (c) noise=0.4 (MNIST) (d) noise=0.6 (MNIST)

(e) noise=0 (CIFAR10) (f) noise=0.2 (CIFAR10) (g) noise=0.4 (CIFAR10) (h) noise=0.6 (CIFAR10)

Figure 4: Dynamics of gradient dispersion, in relation to training/testing accuracies.

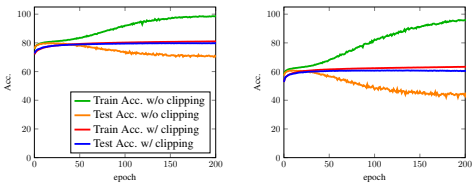
Dynamic Gradient Clipping Inspired by our generalization bounds and above observations, one way to reduce the generalization error is to control the trajectory term of the bounds by reducing the gradient dispersion in each training step. Here we investigate a simple scheme that dynamically clips the gradient norm so as to reduce the gradient dispersion. Specifically, whenever the current gradient norm is larger than the gradient norm K steps earlier, or $||g(W_t, B_t)||_2 > ||g(W_{t-K}, B_{t-K})||_2$ (i.e., the model is expected to have entered the "memorization" regime), we reduce the norm of the current gradient $g(W_t, B_t)$ to α fraction of $||g(W_{t-K}, B_{t-K})||_2$, for some prescribed value $\alpha < 1$. The effectiveness of this scheme is best demonstrated when the labels contain noise. As shown in Figure 5 (and Figure 9 in Appendix E.4), dynamic gradient clipping significantly closes the gap between the training accuracy and the testing accuracy. The models trained with this scheme maintain a near-optimal testing accuracy (e.g., about 80% when the label noise level of MNIST is 0.2), without suffering from the severe memorization effect as seen in models trained without this scheme. Further understanding of the double-descent phenomenon of the gradient dispersion may enable more delicate design of such a dynamic clipping scheme and potentially lead to novel and powerful regularization techniques.

5 A Practical Implication: Gaussian Model Perturbation

The appearance of the flatness term in our generalization bounds suggests that for an empirical risk minimizer w^* to generalize well, it is necessary that the empirical risk surface at w^* is flat, or insensitive to a small perturbation of w^*. This naturally motivates a training scheme using the following regularized loss:

$$\min_w L_s(w) + \rho \mathop{\mathbb{E}}_{\Delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)} [L_s(w + \Delta) - L_s(w)],$$

where ρ is a hyper-parameter. Replacing the expectation above with its stochastic approximation using k realizations of Δ gives rise to the following optimization problem.



(a) noise=0.2 (MNIST) (b) noise=0.4 (MNIST)

Figure 5: Dynamic Gradient Clipping.

$$\min_w \frac{1}{b} \sum_{z \in B} \left((1-\rho)\ell(w, z) + \rho \frac{1}{k} \sum_{i=1}^{k} (\ell(w + \delta_i, z)) \right).$$

We refer to the SGD training scheme using this loss as *Gaussian model perturbation* or GMP. Notably, GMP requires $k + 1$ forward passes for every parameter update. Empirical evidence shows that a small k, for example, $k = 3$, already gives competitive performance. Implementing the $k + 1$ forward passes on parallel processors further reduces the computation load.

Method	SVHN	CIFAR-10	CIFAR-100
ERM	96.86±0.060	93.68±0.193	72.16±0.297
Dropout	97.04±0.049	93.78±0.147	72.28±0.337
L. S.	96.93±0.070	93.71±0.158	72.51±0.179
Flooding	96.85±0.085	93.74±0.145	72.07±0.271
MixUp	96.91±0.057	**94.52±0.112**	73.19±0.254
Adv. Tr.	97.06±0.091	93.51±0.130	70.88±0.145
AMP	**97.27±0.015**	94.35±0.147	74.40±0.168
GMP^3	97.18±0.057	94.33±0.094	74.45±0.256
GMP^{10}	97.09±0.068	94.45±0.158	**75.09±0.285**

Table 1: Classification acc.(%) of VGG16.

We compare GMP with several major regularization schemes in the current art, including Dropout (Srivastava et al., 2014), label smoothing (L.S.) (Szegedy et al., 2016), Flooding (Ishida et al., 2020), MixUp (Zhang et al., 2018), adversarial training (Adv. Tr.) (Goodfellow et al., 2015), and AMP (Zheng et al., 2021). The compared schemes are evaluated on three popular benchmark image classification datasets SVHN (Netzer et al., 2011), CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). Two representative deep architectures VGG16 (Simonyan & Zisserman, 2015) and PreActResNet18 (He et al., 2016) are taken as the underlying model. We train the models for 200 epochs by SGD. The learning rate is initialized as 0.1 and divided by 10 after 100 and 150 epochs. For all compared models, the batch size is set to 50 and weight decay is set to 10^{-4}. For GMP, we choose $\rho = 0.5$ and set the standard deviation of the Gaussian noise Δ to 0.03. The value of k is chosen as 3 and 10 respectively (referred to as GMP^3 and GMP^{10}).

The performances of all compared schemes are given in Table 1 (and the results of PreActResNet18 are shown in Appendix E.6). For the compared regularization schemes except GMP, we directly report their performances as given in Zheng et al. (2021). Table 1 demonstrates the effectiveness of GMP. Overall GMP performs comparably to the current art of regularization schemes, although appearing slighly inferior to AMP on SVHN and MixUp on CIFAR-10, respectively.

6 Conclusion and Outlook

This paper presents new generalization bounds for models, such as neural networks, trained with SGD, improving upon the results of Neu (2021). Our bounds naturally point to new and effective regularization schemes. At the same time, they reveal interesting phenomena in the SGD training of neural networks. While these phenomena deserve further investigation in their own right, we here suggest another direction for improving the bounds, namely, via the use of strong data-processing inequalities (DPI) (Polyanskiy & Wu, 2015; 2017; Raginsky, 2016).

For any Markov chain $U \to X \to Y$, we will denote by $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{Y}$ the spaces in which U, X, Y take values, respectively. For any distribution P on $\mathcal{X}$, we will use $P_{Y|X} \circ P$ to denote the distribution on $\mathcal{Y}$ induced by the push-forward of the distribution P by $P_{Y|X}$, namely, for any $y \in \mathcal{Y}$, $\left(P_{Y|X} \circ P\right)(y) \triangleq \int P_{Y|X}(y|x)P(x)dx$. Let $\mathcal{S}(U)$ be the support of P_U and $\mathcal{H}(U, P_{X|U})$ be the convex hull of $\{P_{X|U=u} : u \in \mathcal{S}(U)\}$. Define

$$\eta(U \to X \to Y) \triangleq \sup_{P,Q \in \mathcal{H}(U,P_{X|U})} \frac{\mathrm{D_{KL}}(P_{Y|X} \circ P || P_{Y|X} \circ Q)}{\mathrm{D_{KL}}(P||Q)}.$$

Lemma 7. *For any Markov chain $U \to X \to Y$, $I(U;Y) \leq \eta(U \to X \to Y) I(U;X)$.*

Here $\eta(U \to X \to Y)$ serves as the "contraction coefficient" for the stochastic kernel $P_{Y|X}$, characterizing the greatest extent by which the kernel may bring closer any two distributions on $\mathcal{S}(U)$ in its output space. It is easy to see that $\eta(U \to X \to Y) \leq 1$, giving rise to a stronger DPI.

Denote $V_t \triangleq \widetilde{W}_{t-1} + G_t$. It can be verified that $Z_i \to V_t \to \widetilde{W}_t$ form a Markov chain. Denote $\eta_{i,t} \triangleq \eta(Z_i \to V_t \to \widetilde{W}_t)$, and $\Gamma_i^t = \{t+1, t+2, \ldots, T\} \setminus \mathcal{T}_i$. Theorem 2 can be improved to:

Theorem 5. *The expected generalization error of SGD is bounded by*

$$|\mathrm{gen}(\mu, P_{W_T|S})| \leq \frac{2R}{nb} \sum_{i=1}^{n} \sqrt{\sum_{t \in \mathcal{T}_i} \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}\left[\mathbb{V}(W_{t-1})\right] \cdot \prod_{\tau \in \Gamma_i^t} \eta_{i,\tau}} + \left|\mathbb{E}\left[\gamma(W_T, S) - \gamma(W_T, S')\right]\right|.$$

It remains to characterize the contraction coefficient $\eta_{i,\tau}$ in a computable form. Simply bounding it via the Dobrushin's coefficient (Dobrushin, 1956), as suggested in Wang et al. (2021) for analyzing SGLD, is unlikely to make the bound in this theorem significantly tighter than that in Theorem 2.

REFERENCES

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

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Amir R 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 R Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7245–7254, 2018.

Peter L Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6241–6250, 2017.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *Algorithmic Learning Theory*, pp. 25–55. PMLR, 2018.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

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

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130, 2020.

Alexander Camuto, Matthew Willetts, Umut Simsekli, Stephen J. Roberts, and Chris C. Holmes. Explicit regularisation in gaussian noise injections. In *Advances in Neural Information Processing Systems*, 2020.

Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gürbüzbalaban, and Umut Şimşekli. Asymmetric heavy tails and implicit bias in gaussian noise injections. *arXiv preprint arXiv:2102.07006*, 2021.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *5th International Conference on Learning Representations*. OpenReview.net, 2017.

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

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

Felix Dangel, Frederik Kunstner, and Philipp Hennig. Backpack: Packing more into backprop. In *8th International Conference on Learning Representations*. OpenReview.net, 2020.

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

Roland L Dobrushin. Central limit theorem for nonstationary markov chains. i. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

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

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2350–2358, 2015.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Fartash Faghri, David Duvenaud, David J Fleet, and Jimmy Ba. A study of gradient variance in deep learning. *arXiv preprint arXiv:2007.04532*, 2020.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pp. 1270–1279. PMLR, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

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.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *14th International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 315–323, 2011.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33, 2020.

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. *arXiv preprint arXiv:2004.12983*, 2020.

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

Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems 32*, pp. 1141–1150, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645, 2016.

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

Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Do we need zero training loss after achieving zero training error? In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations*. OpenReview.net, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

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

Ben London. A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2935–2944, 2017.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. In *International Conference on Learning Representations*, 2018.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 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*, pp. 11013–11023, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Gergely Neu. Information-theoretic generalization bounds for stochastic gradient descent. *arXiv preprint arXiv:2102.00931*, 2021.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *COLT*, 2021.

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

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

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*, 2018a.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2018b.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32: 8026–8037, 2019.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, 2018.

Yury Polyanskiy and Yihong Wu. Dissipation of information in channels with input constraints. *IEEE Transactions on Information Theory*, 62(1):35–55, 2015.

Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, pp. 211–249. Springer, 2017.

Maxim Raginsky. Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications and coding. *Foundations and Trends in Communications and Information Theory; NOW Publishers: Boston, MA, USA*, 2018.

Borja Rodríguez-Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm. *arXiv preprint arXiv:2010.10994*, 2020.

Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2663–2671, 2012.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pp. 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.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3437–3452. PMLR, 2020.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 2818–2826, 2016.

Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3503–3513. PMLR, 2020.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.

Hao Wang, Yizhe Huang, Rui Gao, and Flavio P Calmon. Learning while dissipating information: Understanding the generalization capability of sgld. *arXiv preprint arXiv:2102.02976*, 2021.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International Conference on Machine Learning*, pp. 10181–10192. PMLR, 2020.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Yeming Wen, Kevin Luk, Maxime Gazeau, Guodong Zhang, Harris Chan, and Jimmy Ba. An empirical study of stochastic gradient descent with structured covariance noise. In *International Conference on Artificial Intelligence and Statistics*, pp. 3621–3631. PMLR, 2020.

Lei Wu, Chao Ma, and E Weinan. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 2018: 8279–8288, 2018.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 2017:2525–2534, 2017.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777. PMLR, 2020.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 581–590. IEEE, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations*. OpenReview.net, 2017.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR*, 2018.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2019.

Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *CVPR*, 2021.

Ruida Zhou, Chao Tian, and Tie Liu. Individually conditional individual mutual information bound on generalization error. *arXiv preprint arXiv:2012.09922*, 2020.

Appendix

Table of Contents

A HWI Inequality: Proof of Lemma 4 **15**

B Improved Sample-Based Mutual Information Bounds: Proof of Theorem 1 **16**
B.1 Lemma 8 16
B.2 Proof of Theorem 1 17

C Instance-Based Mutual Information Bounds: Proof of Theorem 2 and an Additional Corollary **18**
C.1 Proof of Lemma 5 18
C.2 Proof of Theorem 2 20
C.3 Proof of Lemma 6 21
C.4 Corollary 1 21

D Application in Neural Networks: Proof of Theorems 3 and 4 **22**
D.1 Proof of Theorem 3 22
D.2 Proof of Theorem 4 23

E Experiment Details **23**
E.1 Architectures and Hyperparameters 23
E.2 Learning Rate and Batch Size. 24
E.3 Implementation Details of Figure 2 24
E.4 Algorithm of Dynamic Gradient Clipping and Additional Results 24
E.5 Discussion on Gradient Dispersion of Models Trained on Clean Datasets 25
E.6 Algorithm of Gaussian Model Perturbation and Additional Results 26

F Background on Strong Data Processing Inequality **26**

G Omitted Proofs in Section 6 **28**
G.1 Proof of Theorem 5 29

A HWI Inequality: Proof of Lemma 4

Proof. The proof given here is a simple extension of the proof of Raginsky & Sason (2018, Lemma 3.4.2), which is a special instance of the (weak) HWI inequality.

$$
\begin{aligned}
\mathrm{D_{KL}}(P_{X+\sqrt{t}N}||P_{Y+\sqrt{t}N}) \leq & \mathrm{D_{KL}}(P_{X,Y,X+\sqrt{t}N}||P_{X,Y,Y+\sqrt{t}N}) && (1)\\
= & \mathop{\mathbb{E}}_{X,Y}\left[\mathrm{D_{KL}}(P_{X+\sqrt{t}N|X,Y}||P_{Y+\sqrt{t}N|X,Y})\right] \\
= & \mathop{\mathbb{E}}_{X,Y}\left[\mathrm{D_{KL}}(\mathcal{N}(X,t)||\mathcal{N}(Y,t))|X,Y\right] && (2)\\
= & \frac{1}{2t}\mathop{\mathbb{E}}_{X,Y}\left[||X-Y||^2\right], && (3)
\end{aligned}
$$

where Eq.1 is by the chain rule of the KL divergence, Eq.2 holds since N is independent of (X, Y) and Eq.3 is a special case of the equality

$$
\mathcal{N}(\mu_1,\Sigma_1)||\mathcal{N}(\mu_2,\Sigma_2)) = \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{Tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right].
$$

Eq.3 holds for any joint distribution of (X, Y), so by the definition of Wasserstein distance, we have

$$\mathrm{D_{KL}}(P_{X+\sqrt{t}N}||P_{Y+\sqrt{t}N}) \leq \frac{1}{2t}\mathbb{W}_2^2(P_X, P_Y). \tag{4}$$

This completes the proof. □

B IMPROVED SAMPLE-BASED MUTUAL INFORMATION BOUNDS: PROOF OF THEOREM 1

In this section, we provide a proof for Theorem 1, while elaborating on why the local gradient sensitivity term in Lemma 3 can be removed. The key ingredient that makes this possible is a construction given in the proof of Lemma 8.

B.1 LEMMA 8

Lemma 8. *Let $G_t = -\lambda_t g(W_{t-1}, B_t)$. Then, $I(G_t + N_t; S|\widetilde{W}_{t-1}) \leq \frac{2\lambda_t^2}{\sigma_t^2 b}\mathbb{E}\left[\mathbb{V}(W_{t-1})\right]$.*

Proof. The LHS of the inequality can be rewritten as

$$\begin{aligned}
&I\left(-g(W_{t-1}, B_t) + \frac{1}{\lambda_t}N_t; S|\widetilde{W}_{t-1}\right) \\
&= \int_{\tilde{w}_{t-1}} \left(\int_s \mathrm{D_{KL}}\left(P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|S=s,\widetilde{W}_{t-1}=\tilde{w}_{t-1}}||P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}\right) dP_{S|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}(s)\right) dP_{\widetilde{W}_{t-1}}(\tilde{w}_{t-1}) \\
&= \mathop{\mathbb{E}}_{\widetilde{W}_{t-1},S}\left[\mathrm{D_{KL}}\left(P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|S=s,\widetilde{W}_{t-1}=\tilde{w}_{t-1}}||P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}\right)\right].
\end{aligned} \tag{5}$$

Since N_t is independent of S and $\widetilde{W}_{t-1}$, by Lemma 4, we have

$$\begin{aligned}
&\mathrm{D_{KL}}\left(P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|S=s,\widetilde{W}_{t-1}=\tilde{w}_{t-1}}||P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}\right) \\
\leq &\frac{\lambda_t^2}{2\sigma_t^2}\mathbb{W}_2^2\left(P_{-g(W_{t-1},B_t)|S=s,\widetilde{W}_{t-1}=\tilde{w}_{t-1}}, P_{-g(W_{t-1},B_t)|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}\right).
\end{aligned} \tag{6}$$

Let S' be an independent copy of S so B_t' is an independent copy of B_t. We now consider a special coupling: define two random vectors $\widetilde{G}_t^s$ and $\widetilde{G}_t$ as follows,

$$\widetilde{G}_t^s \triangleq -\frac{1}{b}\left(\sum_{Z_j \in B_t} \nabla_w \ell(\widetilde{w}_{t-1} - \Delta_{t-1}, Z_j)\right), \quad \widetilde{G}_t' \triangleq -\frac{1}{b}\left(\sum_{Z_j' \in B_t'} \nabla_w \ell(\widetilde{w}_{t-1} - \Delta_{t-1}, Z_j')\right).$$

Thus, $\widetilde{G}_t^s$ and $\widetilde{G}_t'$ have marginals $P_{-g(W_{t-1},B_t)|S=s,\widetilde{W}_{t-1}=\tilde{w}_{t-1}}$ and $P_{-g(W_{t-1},B_t)|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}$, respectively. This will subsequently allow us to invoke the HWI inequality, namely, plugging in the distributions of $\widetilde{G}_t^s$ and $\widetilde{G}_t'$ into Eq.6.

Notice that in this construction, we do not need $\widetilde{G}_t'$ to be independent of $\widetilde{G}_t^s$. In fact, they depend on the same random variable Δ_{t-1}. This is a major departure from the proof in Neu (2021). This construction, although different from that in Neu (2021), is sufficient for invoking the HWI inequality or Eq.6, since Eq.6 only makes reference to two pairs of marginal distributions without referring to the joint distribution of each pair. To see better the difference between this construction and that in Neu (2021), in the latter, $\widetilde{G}_t^s$ is defined in the same way as ours, but $\widetilde{G}_t'$ is defined as $\widetilde{G}_t' \triangleq -\frac{1}{b}\left(\sum_{Z_j' \in B_t'} \nabla_w \ell(\widetilde{w}_{t-1} - \Delta_{t-1}', Z_j')\right)$, namely that Δ_{t-1} is replaced by an independent copy Δ_{t-1}'. Although this construction in Neu (2021) gives rise to same pair of marginals as ours, it is precisely due to the use of Δ_{t-1}' that causes the local gradient sensitivity term to appear in their bounds.

Returning to continue with the proof, we now take the expectation over $S \sim P_{S|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}$ for the Wasserstein distance in Eq.6,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{S}\left[\frac{\lambda_t^2}{2\sigma_t^2}\mathbb{W}_2^2\left(P_{-g(W_{t-1},B_t)|S=s,\widetilde{W}_{t-1}=\tilde{w}_{t-1}}, P_{-g(W_{t-1},B_t)|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}\right)\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right] \\
\leq&\frac{\lambda_t^2}{2\sigma_t^2}\mathop{\mathbb{E}}_{S,\Delta_{t-1},S'}\left[||\widetilde{G}_t^s-\widetilde{G}_t'||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right] \qquad (7)\\
=&\frac{\lambda_t^2}{2\sigma_t^2}\mathop{\mathbb{E}}_{S,\Delta_{t-1},S'}\left[||g(\widetilde{w}_{t-1}-\Delta_{t-1},B_t')-g(\widetilde{w}_{t-1}-\Delta_{t-1},B_t)||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right] \qquad (8)\\
=&\frac{\lambda_t^2}{2\sigma_t^2}\mathop{\mathbb{E}}_{\Delta_{t-1},S,S'}\left[||g(W_{t-1},B_t')-\mathop{\mathbb{E}}_{Z}[g(W_{t-1},Z)]+\mathop{\mathbb{E}}_{Z}[g(W_{t-1},Z)]-g(W_{t-1},B_t)||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right]\\
\leq&\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{\Delta_{t-1},S}\left[||g(\widetilde{w}_{t-1}-\Delta_{t-1},B_t)-\mathop{\mathbb{E}}_{Z}[g(\widetilde{w}_{t-1}-\Delta_{t-1},Z)]||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right], \qquad (9)
\end{aligned}
$$

where Eq.7 is by the definition of Wasserstein distance and Eq.9 is by the triangle inequality (i.e. $||a-b||_2^2 \leq 2||a||_2^2+2||b||_2^2$). **We stress that if we had followed the construction of Neu (2021), Eq.8 would appear as $\frac{\lambda_t^2}{2\sigma_t^2}\mathop{\mathbb{E}}_{S,\Delta_{t-1},S'}\left[||g(\widetilde{w}_{t-1}-\Delta_{t-1}',B_t')-g(\widetilde{w}_{t-1}-\Delta_{t-1},B_t)||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right]$, and the appearance of Δ_{t-1}' would lead to the additional local gradient sensitivity term**.

Combining Eq.5, Eq.6 and Eq.9, we have

$$
\begin{aligned}
I(G_t+N_t;S|\widetilde{W}_{t-1}) \leq&\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{\widetilde{W}_{t-1},\Delta_{t-1},S}\left[||g(\widetilde{W}_{t-1}-\Delta_{t-1},B_t)-\mathop{\mathbb{E}}_{Z}\left[g(\widetilde{W}_{t-1}-\Delta_{t-1},Z)\right]||_2^2\right]\\
=&\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathop{\mathbb{E}}_{S}\left[||g(w_{t-1},B_t)-\mathop{\mathbb{E}}_{Z}[g(w_{t-1},Z)]||_2^2\Big|W_{t-1}=w_{t-1}\right]\right]\\
=&\frac{2\lambda_t^2}{b\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathop{\mathbb{E}}_{Z}\left[||g(w_{t-1},Z)-\mathop{\mathbb{E}}_{Z}[g(w_{t-1},Z)]||_2^2\Big|W_{t-1}=w_{t-1}\right]\right] \qquad (10)\\
=&\frac{2\lambda_t^2}{b\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})],
\end{aligned}
$$

where Eq.10 is due to the fact that the gradients $g(w_{t-1},Z_j)$ are independent of each other in a batch given $W_{t-1}=w_{t-1}$.

This completes the proof. □

B.2 Proof of Theorem 1

Proof. We first follow the similar decomposition of the expected generalization error and apply Lemma 1 as in Neu (2021),

$$
\begin{aligned}
|\mathrm{gen}(\mu,P_{W_T|S})| =& \left|\mathrm{gen}(\mu,P_{\widetilde{W}_T|S})+\mathop{\mathbb{E}}_{W_T,\Delta_T}\left[L_\mu(W_T)-L_\mu(\widetilde{W}_T)\right]+\mathop{\mathbb{E}}_{W_T,\Delta_T,S}\left[L_S(\widetilde{W}_T)-L_S(W_T)\right]\right|\\
\leq& \sqrt{\frac{2R^2}{n}I(\widetilde{W}_T;S)}+\left|\mathop{\mathbb{E}}_{W_T,S,S'}[\gamma(W_T,S)-\gamma(W_T,S')]\right|. \qquad (11)
\end{aligned}
$$

The remaining task is to bound the mutual information $I(\widetilde{W}_T;S)$.

Notice that

$$
\begin{aligned}
I(\widetilde{W}_T; S) =& I(\widetilde{W}_{T-1} - \lambda_T g(W_{T-1}, B_T) + N_t; S) \\
\leq& I(\widetilde{W}_{T-1}, -\lambda_T g(W_{T-1}, B_T) + N_T; S) && (12)\\
=& I(\widetilde{W}_{T-1}; S) + I(-\lambda_T g(W_{T-1}, B_T) + N_T; S|\widetilde{W}_{T-1}) && (13)\\
\leq& I(\widetilde{W}_{T-2}; S) + I(-\lambda_T g(W_{T-2}, B_{T-1}) + N_{T-1}; S|\widetilde{W}_{T-2}) + I(-\lambda_T g(W_{T-1}, B_T) + N_T; S|\widetilde{W}_{T-1}) && (14)\\
&\vdots \\
\leq& \sum_{t=1}^{T} I(-\lambda_t g(W_{t-1}, B_t) + N_t; S|\widetilde{W}_{t-1}), && (15)
\end{aligned}
$$

where Eq.12 is by $I(f(X); Y) \leq I(X; Y)$, Eq.13 is by the chain rule of mutual information, Eq.14 is by applying the similar procedure (namely, Eq.12-Eq.13) to $I(\widetilde{W}_{T-1}; S)$, and Eq.15 is due to the fact that $\widetilde{W}_0$ is independent of S (i.e. $I(\widetilde{W}_0; S) = 0$). Applying Lemma 8 and putting everything together, we have

$$
|\mathrm{gen}(\mu, P_{W_T|S})| \leq \sqrt{\frac{4R^2}{nb} \sum_{t=1}^{T} \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}\left[\mathbb{V}(W_{t-1})\right]} + \left|\mathbb{E}\left[\gamma(W_T, S) - \gamma(W_T, S')\right]\right|.
$$

This completes the proof. □

C Instance-Based Mutual Information Bounds: Proof of Theorem 2 and an Additional Corollary

We note that similar to Theorem 1, our instance-based mutual information bounds does not include the local gradient sensitivity term. This is because of a similar construction (to that in Lemma 8) is used establishing Lemma 5. Indeed, Lemma 5 plays a similar role in proving Theorem 2 as Lemma 8 in proving Theorem 1.

C.1 Proof of Lemma 5

Proof. Let Z_i be used in t^{th} step, the LHS of the inequality can be rewritten as

$$
\begin{aligned}
& I\left(-g(W_{t-1}, B_t) + \frac{1}{\lambda_t} N_t; Z_i|\widetilde{W}_{t-1}\right) \\
=& \int_{\tilde{w}_{t-1}} \left(\int_z \mathrm{D_{KL}}\left(P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|Z_i=z,\widetilde{W}_{t-1}=\tilde{w}_{t-1}} || P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}\right) dP_{Z_i|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}(z) \right) dP_{\widetilde{W}_{t-1}}(\tilde{w}_{t-1}) \\
=& \mathop{\mathbb{E}}_{\widetilde{W}_{t-1}, Z_i} \left[\mathrm{D_{KL}}\left(P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|Z_i=z,\widetilde{W}_{t-1}=\tilde{w}_{t-1}} || P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}\right) \right]. && (16)
\end{aligned}
$$

Since N_t is independent of Z_i and $\widetilde{W}_{t-1}$, by Lemma 4, we have

$$
\begin{aligned}
& \mathrm{D_{KL}}\left(P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|Z_i=z,\widetilde{W}_{t-1}=\tilde{w}_{t-1}} || P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}\right) \\
\leq& \frac{\lambda_t^2}{2\sigma_t^2} \mathbb{W}_2^2\left(P_{-g(W_{t-1},B_t)|Z_i=z,\widetilde{W}_{t-1}=\tilde{w}_{t-1}}, P_{-g(W_{t-1},B_t)|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}\right). && (17)
\end{aligned}
$$

Let Z_i' be an independent copy of Z_i such that $(\widetilde{W}_{t-1}, Z_i', Z_i) \sim P_{\widetilde{W}_{t-1}} \otimes P_{Z_i|W_{t-1}} \otimes P_{Z_i'|W_{t-1}}$. Let $B_t^i = B_t \setminus \{Z_i\}$. We now consider a special coupling: define two random vectors $\widetilde{G}_t^z$ and $\widetilde{G}_t$ as

follows,

$$\widetilde{G}_t^z \triangleq -\frac{1}{b}\left(\sum_{Z_j\in B_t^i}\nabla_w\ell(\widetilde{w}_{t-1}-\Delta_{t-1},Z_j)+\nabla_w\ell(\widetilde{w}_{t-1}-\Delta_{t-1},z)\right),$$

$$\widetilde{G}_t \triangleq -\frac{1}{b}\left(\sum_{Z_j\in B_t^i}\nabla_w\ell(\widetilde{w}_{t-1}-\Delta_{t-1},Z_j))+\nabla_w\ell(\widetilde{w}_{t-1}-\Delta_{t-1},Z_i')\right).$$

Thus, $\widetilde{G}_t^z$ and $\widetilde{G}_t$ have marginals $P_{-g(W_{t-1},B_t)|Z_i=z,\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}$ and $P_{-g(W_{t-1},B_t)|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}$, respectively.

We now take the expectation over $Z_i\sim P_{Z_i|\widetilde{W}_{t-1}=\tilde{w}_{t-1}}$ for the Wasserstein distance in Eq. 17,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Z_i}\left[\mathbb{W}_2^2\left(P_{-g(W_{t-1},B_t)|Z_i=z,\widetilde{W}_{t-1}=\widetilde{w}_{t-1}},P_{-g(W_{t-1},B_t)|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}\right)\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right]\\
\leq&\mathop{\mathbb{E}}_{Z_i,\Delta_{t-1},Z_i',B_t^i}\left[||\widetilde{G}_t^z-\widetilde{G}_t||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right] &(18)\\
=&\mathop{\mathbb{E}}_{\Delta_{t-1},Z_i,Z_i'}\left[\frac{1}{b^2}||\nabla_w\ell(\tilde{w}_{t-1}-\Delta_{t-1},Z_i)-\nabla_w\ell(\tilde{w}_{t-1}-\Delta_{t-1},Z_i')||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right] &(19)\\
=&\frac{1}{b^2}\mathop{\mathbb{E}}_{\Delta_{t-1},Z_i,Z_i'}\left[||\nabla_w\ell(W_{t-1},Z_i)-\mathop{\mathbb{E}}_{Z}\left[\nabla_w\ell(W_{t-1},Z)\right]+\mathop{\mathbb{E}}_{Z}\left[\nabla_w\ell(W_{t-1},Z)\right]-\nabla_w\ell(W_{t-1},Z_i')||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right]\\
\leq&\frac{2}{b^2}\mathop{\mathbb{E}}_{\Delta_{t-1},Z_i}\left[||\nabla_w\ell(W_{t-1},Z_i)-\mathop{\mathbb{E}}_{Z}\left[\nabla_w\ell(W_{t-1},Z)\right]||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right]\\
&+\frac{2}{b^2}\mathop{\mathbb{E}}_{\Delta_{t-1},Z_i'}\left[||\nabla_w\ell(W_{t-1},Z_i')-\mathop{\mathbb{E}}_{Z}\left[\nabla_w\ell(W_{t-1},Z)\right]||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right] &(20)\\
=&\frac{4}{b^2}\mathop{\mathbb{E}}_{\Delta_{t-1},Z}\left[||\nabla_w\ell(\tilde{w}_{t-1}-\Delta_{t-1},Z)-\mathop{\mathbb{E}}_{Z}\left[\nabla_w\ell(\tilde{w}_{t-1}-\Delta_{t-1},Z)\right]||_2^2\Big|\widetilde{W}_{t-1}=\tilde{w}_{t-1}\right], &(21)
\end{aligned}$$

where Eq. 18 is by the definition of Wasserstein distance and Eq. 20 is by the triangle inequality (i.e. $||a-b||_2^2\leq 2||a||_2^2+2||b||_2^2$).

Combining Eq.16, Eq.17 and Eq.21, we have

$$\begin{aligned}
I\left(-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t};Z_i|\widetilde{W}_{t-1}\right)\leq&\frac{2\lambda_t^2}{\sigma_t^2b^2}\mathop{\mathbb{E}}_{\widetilde{W}_{t-1},\Delta_{t-1}}\left[\mathop{\mathbb{E}}_{Z}\left[||\nabla_w\ell(w_{t-1},Z_i)-\mathop{\mathbb{E}}_{Z}\left[[\nabla_w\ell(w_{t-1},Z)]\right]||_2^2\right]\Big|W_{t-1}=w_{t-1}\right]\right]\\
=&\frac{2\lambda_t^2}{\sigma_t^2b^2}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathbb{V}(W_{t-1})\right].
\end{aligned}$$

This completes the proof. □

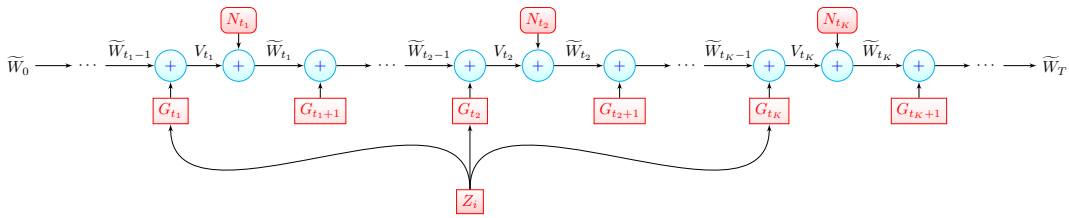


Figure 6: The relationship between Z_i, $\widetilde{W}_t$'s V_t's, and G_t's where $G_t \triangleq -\lambda_t g(W_{t-1},B_t)$, and $V_t\triangleq\widetilde{W}_{t-1}+G_t$.

C.2 PROOF OF THEOREM 2

Proof. Given a specific sample path, let the step t_K be the last iteration that Z_i is used (e.g., see Figure 6). By using the data processing inequality, we can easily have

$$I(\widetilde{W}_T; Z_i) \leq I(\widetilde{W}_{t_K}; Z_i). \tag{22}$$

Then notice that

$$\begin{aligned} I(\widetilde{W}_{t_K}; Z_i) &= I\left(\widetilde{W}_{t_K-1} + G_{t_K} + N_{t_K}; Z_i\right) \\ &\leq I\left(\widetilde{W}_{t_K-1}, G_{t_K} + N_{t_K}; Z_i\right) \qquad (23) \\ &= I(\widetilde{W}_{t_K-1}; Z_i) + I\left(-g(W_{t_K-1}, B_{t_K}) + \frac{1}{\lambda_{t_K}} N_{t_K}; Z_i | \widetilde{W}_{t_K-1}\right), \qquad (24) \end{aligned}$$

where $G_{t_K} = -\lambda_{t_K} g(W_{t_K-1}, B_{t_K})$, Eq.23 is by $I(f(X); Y) \leq I(X; Y)$ and Eq.24 is by the chain rule of mutual information. The second term in Eq.24 can be upper bounded by using Lemma 5. The first term $I(\widetilde{W}_{t_K-1}; Z_i)$ can be upper bounded by following the similar procedure recursively, namely, Eq.22-24. Let t_1 be the first time that Z_i is used in training. Given the fact that Z_i is independent of $\widetilde{W}_t$ when $t < t_1$, it's easy to see

$$I(\widetilde{W}_T; Z_i) \leq \sum_{t \in \mathcal{T}_i} \frac{2\lambda_t^2}{\sigma_t^2 b^2} \mathop{\mathbb{E}}_{W_{t-1}} [\mathbb{V}(W_{t-1})],$$

where $\mathcal{T}_i$ is the set of all indices of iterations that contain Z_i.

Thus, the following bound holds,

$$\begin{aligned} |\mathrm{gen}(\mu, P_{W_T|S})| &= \left| \mathrm{gen}(\mu, P_{\widetilde{W}_T|S}) + \mathop{\mathbb{E}}_{W_T, \Delta_T} \left[L_\mu(W_T) - L_\mu(\widetilde{W}_T) \right] + \mathop{\mathbb{E}}_{W_T, \Delta_T, S} \left[L_S(\widetilde{W}_T) - L_S(W_T) \right] \right| \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2R^2 I(\widetilde{W}_T; Z_i)} + \left| \mathop{\mathbb{E}}_{W_T, S, S'} [\gamma(W_T, S) - \gamma(W_T, S')] \right| \qquad (25) \\ &\leq \frac{2R}{nb} \sum_{i=1}^{n} \sqrt{\sum_{t \in \mathcal{T}_i} \frac{\lambda_t^2}{\sigma_t^2} \mathop{\mathbb{E}}_{W_{t-1}} [\mathbb{V}(W_{t-1})]} + \left| \mathop{\mathbb{E}}_{W_T, S, S'} [\gamma(W_T, S) - \gamma(W_T, S')] \right|, \end{aligned}$$

where Eq. 25 is by Lemma 2.

Next, to handle the mismatch between the outputs of perturbed SGD and SGD, we apply Taylor expansion around $\Delta_T = \vec{0}$,

$$\begin{aligned} \mathop{\mathbb{E}}_{W_T, S, \Delta_T} [L_S(W_T + \Delta_T) - L_S(W_T)] &= \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{W_T, Z_i, \Delta_T} [\ell(W_T + \Delta_T, Z_i) - \ell(W_T, Z_i)] \\ &\approx \mathop{\mathbb{E}}_{W_T, Z, \Delta_T} \left[\langle \nabla_w \ell(W_T, Z), \Delta_T \rangle + \frac{1}{2} \Delta_T^T \mathrm{H}_{W_T}(Z) \Delta_T \right] \\ &= \mathop{\mathbb{E}}_{W_T, Z, \Delta_T} \left[\frac{1}{2} \Delta_T^T \mathrm{H}_{W_T}(Z) \Delta_T \right] \qquad (26) \\ &= \frac{1}{2} \langle \mathop{\mathbb{E}}_{W_T, Z} [\mathrm{H}_{W_T}(Z)], \mathop{\mathbb{E}}_{\Delta_T} \left[\Delta_T \Delta_T^T \right] \rangle \\ &= \frac{1}{2} \langle \mathop{\mathbb{E}}_{W_T, Z} [\mathrm{H}_{W_T}(Z)], \mathrm{diag}(\sum_{t=1}^{T} \sigma_t^2) \rangle \qquad (27) \\ &= \frac{\sum_{t=1}^{T} \sigma_t^2}{2} \mathrm{Tr}(\mathop{\mathbb{E}}_{W_T, Z} [\mathrm{H}_{W_T}(Z)]), \end{aligned}$$

where Eq.26 is by the zero mean of the perturbation, Eq.27 is by the independence of the coordinates of Δ_T, $\langle \cdot, \cdot \rangle$ denotes the inner product of two matrices, $\mathrm{diag}(A)$ is the diagonal matrix with element A and $\mathrm{Tr}(\cdot)$ is the trace of a matrix.

Under the condition $\mathop{\mathbb{E}}_{W_T,S'}[\gamma(W_T,S')] \geq 0$, we now bound $\mathrm{gen}(\mu, \mathrm{P}_{\widetilde{W}_T|S})$ instead of its absolute value, $|\mathrm{gen}(\mu, \mathrm{P}_{\widetilde{W}_T|S})|$. The following is straightforward,

$$\begin{aligned}
\mathrm{gen}(\mu, \mathrm{P}_{\widetilde{W}_T|S}) \leq & \frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]} + \mathop{\mathbb{E}}_{W_T,S,S'}[\gamma(W_T,S)-\gamma(W_T,S')] \\
\leq & \frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]} + \mathop{\mathbb{E}}_{W_T,S}[\gamma(W_T,S)] \\
= & \frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]} + \frac{\sum_{t=1}^{T}\sigma_t^2}{2}\mathrm{Tr}(\mathop{\mathbb{E}}_{W_T,Z}[\mathrm{H}_{W_T}(Z)]).
\end{aligned}$$

Since every choice of σ gives a valid generalization bound. The optimal bound can be found by simply utilizing the fact $A/\sigma+\sigma^2 B \geq 3(A/2)^{2/3}B^{1/3}$ for any positive A and B, where the equality is achieved by the optimal σ. This completes the proof. □

C.3 Proof of Lemma 6

Proof. Notice that if each instance is trained only once for every epoch, it's easy to see that

$$\sum_{i=1}^{n}\sum_{t\in\mathcal{T}_i}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathbb{V}_{W_{t-1}}\right] = b\sum_{t=1}^{T}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathbb{V}_{W_{t-1}}\right].$$

By $\sqrt{\sum_i x_i} \leq \sum_i \sqrt{x_i}$, we have

$$\frac{2R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]} \leq \frac{2R}{n}\sum_{t=1}^{T}\sqrt{\frac{\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]}. \tag{28}$$

Or alternatively, square root is a concave function. By Jensen's inequality, we have

$$\frac{2R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]} \leq \sqrt{\frac{4R^2}{nb}\sum_{t=1}^{T}\frac{\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]}. \tag{29}$$

This completes the proof. □

C.4 Corollary 1

Corollary 1. *If the loss function is differentiable and β-smooth with respect to w, then under the condition of Lemma 6,*

$$|\mathrm{gen}(\mu, P_{W_T|S})| \leq \min\left\{\frac{2R}{n}\sum_{t=1}^{T}\sqrt{\frac{\lambda_t^2}{\sigma_t^2}\mathbb{E}\left[\mathbb{V}(W_{t-1})\right]}, \sqrt{\frac{4R^2}{nb}\sum_{t=1}^{T}\frac{\lambda_t^2}{\sigma_t^2}\mathbb{E}\left[\mathbb{V}(W_{t-1})\right]}\right\} + \beta d\sum_{t=1}^{T}\sigma_t^2.$$

Proof. Recall the smoothness implies $f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v}-\mathbf{w}\rangle + \frac{\beta}{2}||\mathbf{v}-\mathbf{w}||^2$ for all $\mathbf{v}$ and $\mathbf{w}$. By the triangle inequality, we have

$$|\mathbb{E}\left[L_\mu(W_T) - L_\mu(W_T+\Delta_T)\right]| \leq |\mathbb{E}\left[\langle \nabla_w \ell(W_T, Z), \Delta_T\rangle\right]| + \frac{\beta}{2}\mathbb{E}\left[||\Delta_T||^2\right] = \frac{\beta d\sum_{t=1}^{T}\sigma_t^2}{2}$$

Thus, we can see that $|\mathbb{E}\left[L_\mu(W_T) - L_\mu(W_T+\Delta_T)\right]| + |\mathbb{E}\left[L_S(W_T+\Delta_T) - L_S(W_T)\right]| \leq \beta d\sum_{t=1}^{T}\sigma_t^2$.

The first term in the bound comes directly from Eq.28 and Eq.29. □

In Corollary 1, we note that the dependence of d in the bound results from the spherical Gaussian noise used in the construction of the weight process $\widetilde{W}_T$. It is possible to replace the spherical Gaussian with a Gaussian noise having a non-diagonal covariance that reflects the geometry of the loss landscape. With this replacement, the dimension d in the flatness term will be replaced by the trace of $\sum_{t=1}^{T} \kappa_t$, where κ_t is the covariance matrix of the noise added at step t. Please refer to Neu (2021) for a similar development.

D Application in Neural Networks: Proof of Theorems 3 and 4

D.1 Proof of Theorem 3

Proof. From Eq.19 in the proof of Lemma 5, we find that the square norm of gradient incoherence $||\nabla_w \ell(W, Z_i) - \nabla_w \ell(W, Z_j)||^2$ plays a key role in the derivation. Indeed, this quantity is upper bounded by using the gradient dispersion in the proof. With a slight modification in the proof of Lemma 5 and Theorem 2, we can have the following bound,

$$\mathrm{gen}(\mu, P_{W_T|S}) \leq 3 \left(\frac{R}{2nb} \sum_{i=1}^{n} \sqrt{\sum_{t \in \mathcal{T}_i} \lambda_t^2 \mathop{\mathbb{E}}_{W_{t-1}, Z_i, Z_j} [||\nabla_w \ell(W_{t-1}, Z_i) - \nabla_w \ell(W_{t-1}, Z_j)||^2]} \right)^{\frac{2}{3}} \left(\frac{T}{2} \mathop{\mathbb{E}}_{W_T, Z} [\mathrm{Tr}\left(\mathrm{H}_{W_T}(Z)\right)] \right)^{\frac{1}{3}}.$$

By using the first order Taylor expansion, the square norm of gradient incoherence is

$$\begin{aligned} ||\nabla_w \ell(W, Z_i) - \nabla_w \ell(W, Z_j)||^2 \approx & ||\nabla_w \ell(W, Z_i) + \nabla_z \nabla_w \ell(W, Z_i)(Z_j - Z_i) - \nabla_w \ell(W, Z_i)||^2 \\ = & ||\nabla_z \nabla_w \ell(W, Z_i)(Z_j - Z_i)||^2. \end{aligned} \tag{30}$$

Notice that $\nabla_w \ell(W, Z) = (W^T X - Y)X$. Let $\hat{Y} = f(W, X)$. Then,

$$\nabla_z \nabla_w \ell(W, Z) = \begin{pmatrix} \hat{Y} - Y + W_1 X_1 & \cdots & \hat{Y} - Y + W_1 X_{d_0} & -X_1 \\ \hat{Y} - Y + W_2 X_1 & \cdots & \hat{Y} - Y + W_2 X_{d_0} & -X_2 \\ \vdots & \vdots & \vdots & \vdots \\ \hat{Y} - Y + W_{d_0} X_1 & \cdots & \hat{Y} - Y + W_{d_0} X_{d_0} & -X_{d_0} \end{pmatrix} \in \mathbb{R}^{d_0 \times (d_0+1)}.$$

Thus,

$$\begin{aligned} ||\nabla_w \ell(W, Z_i) - \nabla_w \ell(W, Z_j)||^2 \leq & ||\nabla_z \nabla_w \ell(W, Z_i)||_F^2 ||Z_j - Z_i||^2 \\ = & \left(\sum_{j,k=1}^{d_0} \left(\hat{Y}_i - Y_i + W_j X_{i,k} \right)^2 + \sum_{j=1}^{d_0} X_{i,j}^2 \right) ||Z_j - Z_i||^2 \\ \leq & \left(4 \sum_{j,k=1}^{d_0} \ell(W, Z_i) + 2 \sum_{j,k=1}^{d_0} (W_j X_{i,k})^2 + 1 \right) ||Z_j - Z_i||^2 \\ = & \left(4 d_0^2 \ell(W, Z_i) + 2||W||^2 + 1 \right) ||Z_j - Z_i||^2, \end{aligned} \tag{31}$$

where $|| \cdot ||_F$ is the Frobenius norm.

For the Hessian matrix, it's easy to see that $\mathrm{Tr}(\mathrm{H}_{W_T}(Z)) = 1$.

And also notice that $||Z_j - Z_i||^2 \leq 2||Z_j||^2 + 2||Z_i||^2 = 2(1 + Y_j^2) + 2(1 + Y_i^2) \leq 8$.

Putting everything together, we have

$$\mathrm{gen}(\mu, P_{W_T|S}) \leq 3 \left(\frac{RT^2}{nb} \sum_{i=1}^{n} \sqrt{\sum_{t \in \mathcal{T}_i} \lambda_t^2 \mathop{\mathbb{E}}_{W_{t-1}, Z_i, Z_j} [(4 d_0^2 \ell(W_{t-1}, Z_i) + 2||W_{t-1}||^2 + 1)]} \right)^{\frac{2}{3}}.$$

This completes the proof. □

D.2 Proof of Theorem 4

Proof. Consider the square norm of gradient incoherence,

$$||\nabla_w \ell(W, Z_i) - \nabla_w \ell(W, Z_j)||^2 \approx ||\nabla_z \nabla_w \ell(W, Z_i)(Z_j - Z_i)||^2.$$

Since $\nabla_{w_r} \ell(W, Z_i) = \frac{1}{\sqrt{m}} A_r (f(W, X_i) - Y_i) X_i \mathbb{I}_{r,i}$, where $\mathbb{I}_{r,i} = \mathbb{I}\{W_r^T X_i \geq 0\}$, we have

$$\begin{aligned}
&||\nabla_w \ell(W, Z_i) - \nabla_w \ell(W, Z_j)||^2 \\
\leq &||\nabla_z \nabla_w \ell(W, Z_i)||_F^2 ||Z_j - Z_i||^2 \\
= &\sum_{r=1}^{m} \left(\sum_{j,k=1}^{d_0} (\frac{A_r \mathbb{I}_{r,i}}{\sqrt{m}})^2 \left(\hat{Y}_i - Y_i + \frac{A_r \mathbb{I}_{r,i}}{\sqrt{m}} W_{r,j} X_{i,k} \right)^2 + \sum_{j=1}^{d_0} (\frac{A_r \mathbb{I}_{r,i}}{\sqrt{m}} X_{i,j})^2 \right) ||Z_j - Z_i||^2 \\
\leq &\sum_{r=1}^{m} \left(4 d_0^2 \frac{\mathbb{I}_{r,i}}{m} \ell(W, Z_i) + \frac{2\mathbb{I}_{r,i}}{m^2} ||W_r||^2 + \frac{\mathbb{I}_{r,i}}{m} \right) ||Z_j - Z_i||^2 \\
= &\frac{1}{m} \sum_{r=1}^{m} \mathbb{I}_{r,i} \left(4 d_0^2 \ell(W, Z_i) + \frac{2||W_r||^2}{m} + 1 \right).
\end{aligned} \tag{32}$$

In addition, we notice that $\mathrm{Tr}(\mathrm{H}_{W_T}(Z)) = \frac{1}{m} \sum_{r=1}^{m} \mathbb{I}_{r,i}$ and $||Z_j - Z_i||^2 \leq 8$. Putting everything together, we have

$$\mathrm{gen}(\mu, P_{W_T|S}) \leq 3 \left(\frac{T}{m} \sum_{r=1}^{m} \mathbb{E}\left[\mathbb{I}_{r,i}\right] \right)^{\frac{1}{3}} \left(\frac{R}{nb} \sum_{i=1}^{n} \sqrt{\sum_{t \in \mathcal{T}_i} \lambda_t^2 \mathbb{E}\left[\sum_{r=1}^{m} \frac{\mathbb{I}_{r,i}}{m} \left(4 d_0^2 \ell(W, Z_i) + \frac{2||W_r||^2}{m} + 1 \right) \right]} \right)^{\frac{2}{3}}$$

This completes the proof. □

E Experiment Details

E.1 Architectures and Hyperparameters

In Section 4, MLP has one hidden layer with 512 hidden units, and AlexNet has five convolution layers (conv. 3×3 (64 filters) $\rightarrow$ max-pool 3×3 $\rightarrow$ conv. 5×5 (192 filters) $\rightarrow$ max-pool 3×3 $\rightarrow$ conv. 3×3 (384 filters) $\rightarrow$ conv. 3×3 (256 filters) $\rightarrow$ conv. 3×3 (256 filters) $\rightarrow$ max-pool 3×3) followed by two fully connected layers both with 4096 units and a 10-way linear layer as the output layer. All of the convolution layers and the fully connected layers use standard rectified linear activation functions (ReLU).

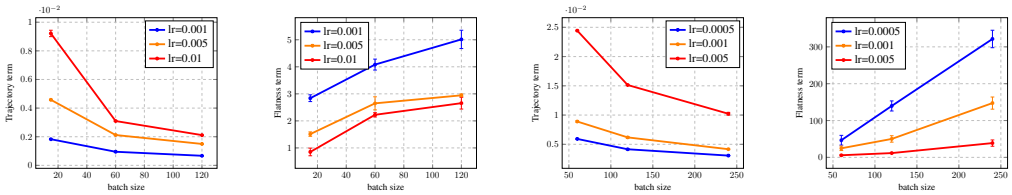
The fixed learning rates used for MLP and AlexNet are 0.01 and 0.001, respectively. The batch size is set to 60. For the corrupted label experiment, we train the models until the models achieve 100% training accuracy. For other cases, we train the neural networks until the training loss converges (e.g., < 0.0001). Other settings are either described in Section 4 or apparent in the figures. Standard techniques such as weight decay and batch normalization are not used.

In Section 5, we compare GMP with other advanced regularization methods. The results of other methods are reported directly from Zheng et al. (2021), and we now give their hyperparameter settings here for completeness. For Dropout, 10% of neurons are randomly selected to be deactivated in each layer. For label smoothing, the coefficient is 0.2. For flooding, the level is set to 0.02. For MixUp, we lineally combine random pairs of training data where the coefficient is drawn from $\mathrm{Beta}(1, 1)$. For adversarial training, the perturbation size is 1 for each pixel and we take one step to generate adversarial examples. For AMP, the number of inner iteration is 1, and the L_2 norm ball radius values are 0.5 for PreActResNet18 and 0.1 for VGG16, respectively.

The implementation in this paper is on PyTorch, and all the experiments are carried out on NVIDIA Tesla V100 GPUs (32 GB).

E.2 Learning Rate and Batch Size.

The learning rate and batch size in SGD have explicitly appeared in the trajectory term of the bound in Theorem 2. We now investigate this by performing experiments with varying learning rates and batch sizes. In our experiments, the model is continuously updated until the average training loss drops below 0.0001. We separate trajectory and flatness terms of the bound and plot them in Figure 7.



(a) MLP on MNIST (b) MLP on MNIST (c) AlexNet on CIFAR10 (d) AlexNet on CIFAR10

Figure 7: The impact of learning rate and batch size on the trajectory term and the flatness term.

A key observation in Figure 7 is that the learning rate impacts the trajectory term and the flatness term in opposite ways, as seen, for example, in (a) and (b), where the two set of curves swap their orders in the two figures. On the other hand, the batch size also impacts the two terms in opposite ways, as seen in (a) and (b) where curves decrease in (a) but increase in (b). This makes the generalization bound, i.e., the sum of the two terms, have a rather complex relationship with the settings of learning rate and batch size. This relationship is further complicated by the fact that a small learning rate requires a longer training time, or a larger number T of training iterations, which increases the number that are summed over in the trajectory term. Nonetheless, we do observe that a smaller batch size gives a lower value of the flatness term ((b) and (d)), confirming the previous wisdom that small batch sizes enable the neural network to find a flat minima (Keskar et al., 2017).

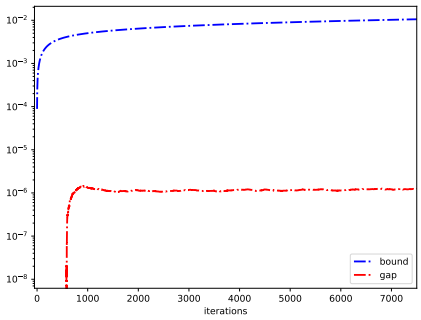


Figure 8: Generalization gap of a linear network v.s. Theorem 3. Note y-axe is log-scale.

E.3 Implementation Details of Figure 2

We let $d_0 = 200$ and use a two-layer ReLU network with hidden units 10000 to generate Y. Moreover, we apply a $\tanh$ function to the output of this network so that $|Y| \leq 1$. The input data $X \sim \mathcal{N}(0, \mathrm{I})$ and we normalize X before training so that $||X|| = 1$. For the training phase, we choose $m = 100$, let data size be $20,000$, batch size be 100 and learning rate be 0.5. Figure 8 shows the similar result of Theorem 3.

E.4 Algorithm of Dynamic Gradient Clipping and Additional Results

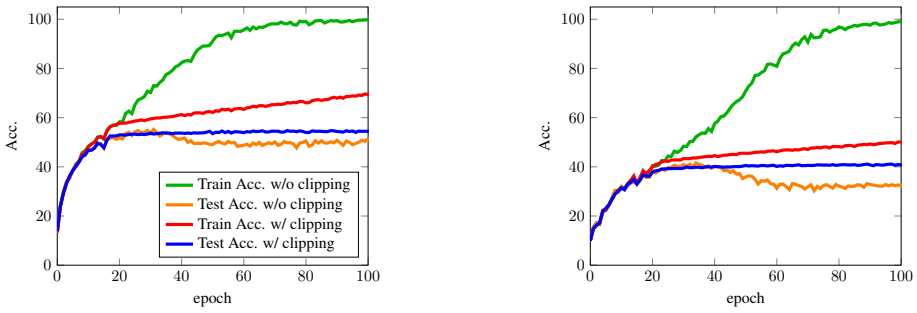
Gradient clipping is a common technique for preventing gradient exploding (see, e.g., Merity et al. (2018); Peters et al. (2018)). This technique is also used in Zhang et al. (2019) to accelerate training. In this paper, gradient clipping is used to investigate and control the impact of gradient dispersion on generalization error.

The dynamic gradient clipping algorithm is described in Algorithm 1. For both MLP and AlexNet, we let $\alpha = 0.1$. The start step for clipping, T_c, is also an important hyperparameter. However, it can be removed by detecting the evolution of the average gradient norm for each epoch. Specifically, whenever the average gradient norm of epoch j is larger than the average gradient norm of epoch $j-1$, the clipping operation begins.

Algorithm 1 Dynamic Gradient Clipping

Require: Training set S, Batch size b, Loss function ℓ, Initial model parameter $\boldsymbol{w}_0$, Learning rate λ, Initial minimum gradient norm $\mathcal{G}$, Number of iterations T, Clipping parameter α, Clipping step T_c

for $t \leftarrow 1$ to T **do**
  Sample $\mathcal{B} = \{\boldsymbol{z}_i\}_{i=1}^{b}$ from training set S
  Compute gradient:
  $g_{\mathcal{B}} \leftarrow \sum_{i=1}^{b} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_{t-1}, \boldsymbol{z}_i)/b$
  if $t > T_c$ **then**
    if $||g_{\mathcal{B}}||_2 > \mathcal{G}$ **then**
      $g_{\mathcal{B}} \leftarrow \alpha \cdot \mathcal{G} \cdot g_{\mathcal{B}}/||g_{\mathcal{B}}||_2$
    else
      $\mathcal{G} \leftarrow ||g_{\mathcal{B}}||_2$
    end if
  end if
  Update parameter: $\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} - \lambda \cdot g_{\mathcal{B}}$
end for



(a) noise=0.2 (CIFAR10)

(b) noise=0.4 (CIFAR10)

Figure 9: Dynamic Gradient Clipping (AlexNet).

From Figure 5 and Figure 9 we can see that dynamic gradient clipping effectively alleviates overfitting by conspicuously slowing down the transition of training to the memorization regime, without changing the convergence speed of testing accuracy. Unfortunately, the current design of the dynamic gradient clipping algorithm does not provide a significant improvement for models trained on a clean dataset (without label noise). Designing better regularization algorithms may require understanding the epoch-wise double descent curve of gradient dispersion where the model is trained on a clean dataset.

E.5 Discussion on Gradient Dispersion of Models Trained on Clean Datasets

In the case of no noise injected, Figure 4a shows that the model with good generalization property has a exponentially-decaying gradient dispersion. This is consistent with our discussion of Lemma 5 in Section 3, that is, small $I(G_t + N_t; Z_i|\widetilde{W}_{t-1})$ indicates good generalization. Notably, gradient dispersion of AlexNet trained on the ture CIFAR10 data still has a epoch-wise double descent curve. The difference between Figure 4e with Figure 4f-4h is that the testing accuracy does not decrease in the second phase/memorization regime for AlexNet trained on the true CIFAR10 data. Loosely speaking, we conjugate that memorizing random labels will hurt the performance on unseen clean data but memorizing clean (or true) labels will not. This may explain why dynamic gradient clipping

Method	SVHN	CIFAR-10	CIFAR-100
ERM	97.05±0.063	94.98±0.212	75.69±0.303
Dropout	97.20±0.065	95.14±0.148	75.52±0.351
L.S.	97.22±0.087	95.15±0.115	77.93±0.256
Flooding	97.16±0.047	95.03±0.082	75.50±0.234
MixUp	97.26±0.044	95.91±0.117	78.22±0.210
Adv. Tr.	97.23±0.080	95.01±0.085	74.77±0.229
AMP	**97.70±0.025**	**96.03±0.091**	**78.49±0.308**
GMP^3	97.43±0.037	95.64±0.053	78.05±0.208
GMP^{10}	97.34±0.058	95.71±0.073	78.07±0.170

Table 2: Top-1 classification accuracy acc.(%) of PreActResNet18. We run experiments 10 times and report the mean and the standard deviation of the testing accuracy. Superscript denotes the number of sampled Gaussian noises during training.

or preventing the training entering the memorization regime cannot improve the performance on a clean dataset.

E.6 Algorithm of Gaussian Model Perturbation and Additional Results

Algorithm 2 Gaussian Model Perturbation Training

Require: Training set S, Batch size b, Loss function ℓ, Initial model parameter $\boldsymbol{w}_0$, Learning rate λ, Number of noise k, Standard deviation of Gaussian distribution σ, Lagrange multiplier ρ
while $\boldsymbol{w}_t$ not converged **do**
  Update iteration: $t \leftarrow t+1$
  Sample $\mathcal{B} = \{\boldsymbol{z}_i\}_{i=1}^b$ from training set S
  Sample $\Delta_j \sim \mathcal{N}(0, \sigma^2)$ for $j \in [k]$
  Compute gradient:
  $g_{\mathcal{B}} \leftarrow \sum_{i=1}^b \left(\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_t, z_i) + \rho \sum_{j=1}^k \left(\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_t + \Delta_j, z_i) - \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_t, z_i) \right) / k \right) / b$
  Update parameter: $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \lambda \cdot g_{\mathcal{B}}$
end while

The GMP algorithm is given in Algorithm 2. Table 1 and Table 2 show that our method is competitive to the state-of-the-art regularization techniques. Specifically, our method achieves the best performance on SVHN for both models and on CIFAR-100 where VGG16 is employed. Particularly, testing accuracy is improved by nealy 2% on the CIFAR-100 dataset with VGG16. For other tasks, GMP is always able to achieve the top-3 performance. In addition, we find that increasing the number of sampled noises does not guarantee the improvement of testing accuracy and may even degrade the performance on some datasets (e.g., SVHN). This hints that we can use small number of noises to reduce the running time without losing performance. Moreover, we observe that **GMP with $k = 3$ usually takes around $1.76\times$ that of ERM training time**, which is affordable.

For PreActResNet18, the performance of GMP appears slighly inferior to the most recent record given by AMP (Zheng et al., 2021). Noting that the key ingredient of AMP, "max-pooling" in the parameter space, greatly resembles regularization term in GMP, which may be seen as "average-pooling" in the same space.

One potential extension of GMP is letting the variance of the noise distribution be a function of the iteration step t. In other words, using the time-dependent σ_t instead of a constant σ.

F Background on Strong Data Processing Inequality

For a Markov chain $U \to X \to Y$, the well-known data-processing inequality (Cover & Thomas, 2012) states that $I(U;Y) \leq I(U;X)$. It turns out that this inequality can often be tightened in the

form of

$$I(U;Y) \leq \eta I(U;X) \tag{33}$$

for some $\eta < 1$. Such an inequality is referred to as a strong data-processing inequality (SDPI). Note the coefficient η fundamentally depends on the contraction property of the stochastic kernel $P_{Y|X}$, characterizing the extent by which the kernel brings two different distributions P_X and $P_{X'}$ closer after the stochastic mapping. As the reader may refer to Polyanskiy & Wu (2015; 2017); Raginsky (2016) for a full exposition of this subject, here for self-containedness, we here make an modest revision of the settings in Polyanskiy & Wu (2015; 2017); Raginsky (2016) and develop slightly finer results concerning the contraction properties of $P_{Y|X}$. Specifically, the contraction coefficient η in our setting will not only depend on the kernel $P_{Y|X}$, it will also depend on the "effective input space" which the kernel acts upon.

To that end, we will denote by $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{Y}$ the spaces in which U, X, Y take values, respectively. For any distribution P on $\mathcal{X}$, we will use $P_{Y|X} \circ P$ to denote the distribution on $\mathcal{Y}$ induced by the push-forward of the distribution P by $P_{Y|X}$, namely, for any $y \in \mathcal{Y}$,

$$\left(P_{Y|X} \circ P\right)(y) \triangleq \int P_{Y|X}(y|x)P(x)dx.$$

For the Markov chain $U \to X \to Y$, we will denote by $\Omega(U)$ the support of the distribution P_U. That is, $\mathcal{S}(U)$ is the subset of $\mathcal{U}$ on which P_U is strictly positive. Let $\mathcal{H}(U, P_{X|U})$ be the convex hull of $\{P_{X|U=u} : u \in \mathcal{S}(U)\}$, namely, $\mathcal{H}(U, P_{X|U})$ contains all distributions on $\mathcal{X}$ which can be expressed as $P_{X|U} \circ P$ for some distribution P on $\mathcal{U}$ whose support is a subset (not necessarily proper) of $\mathcal{S}(U)$. It is apparent that $P_X \in \mathcal{H}(U, P_{X|U})$.

Given the Markov chain, $U \to X \to Y$, we now define the contraction coefficient η as

$$\eta(U \to X \to Y) \triangleq \sup_{P,Q \in \mathcal{H}(U,P_{X|U})} \frac{\mathrm{D_{KL}}(P_{Y|X} \circ P || P_{Y|X} \circ Q)}{\mathrm{D_{KL}}(P||Q)}, \tag{34}$$

where $\mathrm{D_{KL}}$ denotes the KL divergence. We note that this definition of the contraction coefficient differs from the standard definition (Polyanskiy & Wu, 2015; 2017; Raginsky, 2016) in that the supremization in the latter is over all P, Q which are distributions on $\mathcal{X}$, making the contraction coefficient only depends on $P_{Y|X}$. In our definition (34), the coefficient η also depends on the "effective input space" of the kernel, namely, $\mathcal{H}(U, P_{X|U})$.

Some standard results concerning the contraction coefficient can be easily extended to this revised definition of η in Eq.34, which we state below.

Lemma 9. *For any Markov chain $U \to X \to Y$,*

$$I(U;Y) \leq \eta(U \to X \to Y) I(U;X). \tag{35}$$

Additionally, if $Y = X + N$ for a Gaussian noise $N \sim \mathcal{N}(0, \delta^2 \mathbf{I})$ independent of (U, X) then

$$\eta(U \to X \to Y) \leq 1 - 2\mathsf{Q}\left(\frac{\mathrm{Diam}(\mathcal{H}(U, P_{X|U}))}{2\delta}\right), \tag{36}$$

where $\mathrm{Diam}(\mathcal{H}(U, P_{X|U}))$ refers to the diameter of $\mathcal{H}(U, P_{X|U})$, measured under L2-distance, and Q is standard Q-function, or the complementary of Gaussian CDF.

We note that the SDPI (35) follows easily from the definition of mutual information and that of η. The Eq.36 is an adaptation, to this context, of the result that the contraction coefficient is upper bounded by the *Dobrushin's coefficient* of the kernel $P_{Y|X}$ (Dobrushin, 1956). More details about the proof of Lemma 9 can be found in Polyanskiy & Wu (2015).

Concerning the contraction coefficients $\eta_{i,t}$, we note that from a purely theoretical perspective, it is rather difficult to estimate them or upper-bound them by a quantity smaller than 1. This is because of the Gaussian noise N_t injected at each step t having unbounded support. As a consequence, $\mathcal{H}(Z_i, P_{\widehat{W}_t|V_t})$ is unbounded, making the restriction of the range of supremization to $\mathcal{H}(Z_i, P_{\widehat{W}_t|V_t})$ useless.

G OMITTED PROOFS IN SECTION 6

In this section, we will prove Theorem 5 that is a stronger version of Theorem 2. To prove Theorem 2, we only need to let the contraction coefficient η be 1, which means using a weak version of the data-processing inequality.

We first apply the SDPI in Lemma 9 to auxiliary weight process and the SGD weight process.

For notational convenience, we denote

$$G_t \triangleq -\lambda_t g(W_{t-1}, B_t), \text{ and } V_t \triangleq \widetilde{W}_{t-1} + G_t. \tag{37}$$

Thus V_t and $\widetilde{W}_t$ differ only by the noise N_t, or $\widetilde{W}_t = V_t + N_t$.

Consider a fixed example Z_i in the training set S. Let $\mathcal{T}_i$ denote the of all batch indices t for which B_t contains Z_i. We will assume that $\mathcal{T}_i$ contains K indices and consider $\mathcal{T}_i$ as $\{t_1, t_2, \ldots, t_K\}$. The relationship between Z_i, $\widetilde{W}_t$'s, V_t's, and G_t's are shown in Figure 6. It is clear that $Z_i \to V_t \to \widetilde{W}_t$ form a Markov chain[4]. Denote

$$\eta_{i,t} \triangleq \eta(Z_i \to V_t \to \widetilde{W}_t).$$

We have the following result.

Lemma 10. *Suppose that $I(G_t + N_t; Z_i|\widetilde{W}_{t-1}) \leq C_t$ for some positive value C_t. Then*

$$I(\widetilde{W}_T; Z_i) \leq \sum_{k=1}^{K} C_{t_k} \prod_{\tau \in \Gamma_i^k} \eta_{i,t},$$

where $\Gamma_i^k = \{t_k + 1, t_k + 2, \ldots, T\} \setminus \mathcal{T}_i$.

Proof. For the step t_K, which is the last time Z_i appeared in the training process, we have

$$\begin{aligned} I(\widetilde{W}_T; Z_i) &\leq \eta_{i,T} I(V_T; Z_i) && (38)\\ &\leq \eta_{i,T} I(\widetilde{W}_{T-1}; Z_i) && (39)\\ &\leq \left(\prod_{\tau=t_K+1}^{T} \eta_{i,\tau} \right) I(\widetilde{W}_{t_K}; Z_i), && (40) \end{aligned}$$

where Eq.38 and Eq.39 is due to Eq.35 and the data processing inequality, respectively. Eq.40 is by applying these two steps recursively.

Notice that

$$\begin{aligned} I(\widetilde{W}_{t_K}; Z_i) = I(\widetilde{W}_{t_K-1} + G_{t_K} + N_{t_K}; Z_i) &\leq I(\widetilde{W}_{t_K-1}, G_{t_K} + N_{t_K}; Z_i) && (41)\\ &= I(\widetilde{W}_{t_K-1}; Z_i) + I(G_{t_K} + N_{t_K}; Z_i|\widetilde{W}_{t_K-1}) && (42)\\ &\leq I(\widetilde{W}_{t_K-1}; Z_i) + C_{t_K}, && (43) \end{aligned}$$

where Eq.41 is by $I(f(X); Y) \leq I(X; Y)$ and Eq.42 is by the chain rule of mutual information. Combine Eq.40 and Eq.43, we have

$$I(\widetilde{W}_T; Z_i) \leq \left(\prod_{\tau=t_K+1}^{T} \eta_{i,\tau} \right) \left(I(\widetilde{W}_{t_K-1}; Z_i) + C_{t_K} \right). \tag{44}$$

Then we can apply the similar procedure, namely Eq.38-43, to $I(\widetilde{W}_{t_K-1}; Z_i)$ and get

$$I(\widetilde{W}_{t_K-1}; Z_i) \leq \left(\prod_{\tau=t_{K-1}+1}^{t_K-1} \eta_{i,\tau} \right) \left(I(\widetilde{W}_{t_{K-1}-1}; Z_i) + C_{t_{K-1}} \right), \tag{45}$$

[4]Note that even for $t < t_1$, in which Z_i and V_t are independent, the Markov chain still holds, although it degenerates.

where t_{K-1} is the second-to-last time that Z_i is used in the training process. Plugging Eq.45 into Eq.44,

$$I(\widetilde{W}_T; Z_i) \leq C_{t_K} \cdot \prod_{\tau=t_K+1}^{T} \eta_{i,\tau} + \left(C_{t_{K-1}} + I(\widetilde{W}_{t_{K-1}-1}; Z_i)\right) \cdot \prod_{\substack{\tau=t_{K-1}+1 \\ \tau \neq t_K}}^{T} \eta_{i,\tau}. \tag{46}$$

Finally, we apply this procedure recursively and given the fact that $I(\widetilde{W}_t; Z_i) = 0$ for $t < t_1$, we have

$$I(\widetilde{W}_T; Z_i) \leq \sum_{k=1}^{K} C_{t_k} \prod_{\tau \in \Gamma_i^k} \eta_{i,\tau},$$

where $\Gamma_i^k = \{t_k + 1, t_k + 2, \ldots, T\} \setminus \mathcal{T}_i$. □

G.1 Proof of Theorem 5

Proof. Let the generalization error of SGD be decomposed by

$$\mathrm{gen}(\mu, P_{W_T|S}) = \mathrm{gen}(\mu, P_{\widetilde{W}_T|S}) + \mathbb{E}\left[L_\mu(W_T) - L_\mu(\widetilde{W}_T)\right] + \mathbb{E}\left[L_S(\widetilde{W}_T) - L_S(W_T)\right]. \tag{47}$$

Then we use Lemma 2 to bound the first term,

$$\mathrm{gen}(\mu, P_{\widetilde{W}_T|S}) \leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2R^2 I(\widetilde{W}_T; Z_i)} + \mathbb{E}\left[L_\mu(W_T) - L_\mu(\widetilde{W}_T)\right] + \mathbb{E}\left[L_S(\widetilde{W}_T) - L_S(W_T)\right]. \tag{48}$$

Given a specific sample path, let the step t be the last iteration that Z_i is used. Then the following Markov chain holds,

$$Z_i \to V_t \to \widetilde{W}_t \to V_{t+1} \to \widetilde{W}_{t+1} \to \cdots \to V_T \to \widetilde{W}_T.$$

For mutual information $I(\widetilde{W}_T; Z_i)$ in Eq.48, by using the strong data processing inequality and the data processing inequality, we have

$$I(\widetilde{W}_T; Z_i) \leq I(V_{t+1}; Z_i) \cdot \prod_{\tau=t+1}^{T} \eta_{i,\tau} \leq I(\widetilde{W}_t; Z_i) \cdot \prod_{\tau=t+1}^{T} \eta_{i,\tau}. \tag{49}$$

Further, we would like to bound the term $I(\widetilde{W}_t; Z_i)$. Notice that

$$\begin{aligned} I(\widetilde{W}_t; Z_i) &= I\left(\widetilde{W}_{t-1} + G_t + N_t; Z_i\right) \\ &\leq I\left(\widetilde{W}_{t-1}, G_t + N_t; Z_i\right) \\ &= I(\widetilde{W}_{t-1}; Z_i) + I\left(-g(W_{t-1}, B_t) + \frac{1}{\lambda_t} N_t; Z_i | \widetilde{W}_{t-1}\right). \end{aligned} \tag{50}$$

Thus, applying Lemma 10 (i.e., by letting $C_t = \frac{\lambda_t^2}{\sigma_t^2 b^2} \mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]$), we have

$$I(\widetilde{W}_T; Z_i) \leq \sum_{t \in \mathcal{T}_i} \frac{\lambda_t^2}{\sigma_t^2 b^2} \mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})] \cdot \prod_{\tau \in \Gamma_i^t} \eta_{i,\tau},$$

where $\mathcal{T}_i$ is the set of all indices of iterations that contains Z_i. Thus, the following bound holds,

$$|\mathrm{gen}(\mu, \mathrm{P}_{\widetilde{W}_T|S})| \leq \frac{R}{nb} \sum_{i=1}^{n} \sqrt{\sum_{t \in \mathcal{T}_i} \frac{2\lambda_t^2}{\sigma_t^2} \mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})] \cdot \prod_{\tau \in \Gamma_i^t} \eta_{i,\tau}} + \left| \mathop{\mathbb{E}}_{W_T, S, S'}[\gamma(W_T, S) - \gamma(W_T, S')] \right|,$$

which concludes the proof. □