Understanding the Generalization Ability of Deep Learning Algorithms: A Kernelized Rényi's Entropy Perspective

Yuxin Dong[1] , **Tieliang Gong**[1] , **Hong Chen**[2] and **Chen Li**[1]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[2]College of Science, Huazhong Agriculture University, Wuhan 430070, China
adidasgtl@gmail.com, dongyuxin@stu.xjtu.edu.cn, chenh@mail.hzau.edu.cn, cli@xjtu.edu.cn

Abstract

Recently, information-theoretic analysis has become a popular framework for understanding the generalization behavior of deep neural networks. It allows a direct analysis for stochastic gradient / Langevin descent (SGD/SGLD) learning algorithms without strong assumptions such as Lipschitz or convexity conditions. However, the current generalization error bounds within this framework are still far from optimal, while substantial improvements on these bounds are quite challenging due to the intractability of high-dimensional information quantities. To address this issue, we first propose a novel information theoretical measure: kernelized Rényi's entropy, by utilizing operator representation in Hilbert space. It inherits the properties of Shannon's entropy and can be effectively calculated via simple random sampling, while remaining independent of the input dimension. We then establish the generalization error bounds for SGD/SGLD under kernelized Rényi's entropy, where the mutual information quantities can be directly calculated, enabling evaluation of the tightness of each intermediate step. We show that our information-theoretical bounds depend on the statistics of the stochastic gradients evaluated along with the iterates, and are rigorously tighter than the current state-of-the-art (SOTA) results. The theoretical findings are also supported by large-scale empirical studies[1].

1 Introduction

Modern deep neural networks (DNNs) achieve astonishing success through their ability to memorize the entire training data while also generalizing well to unseen data. Generalization bounds in conventional statistical learning theory fail to explain this empirical observation since they attribute the generalization to the constrained complexity of hypothesis spaces, which are usually scale-sensitive [Zhang *et al.*, 2021]. Instead, recent studies discovered that the algorithmic choice has a significant influence on the generalization behavior of DNNs [Hardt *et al.*, 2016; Bartlett *et al.*, 2017], raising broad research interests in investigating the theoretical properties of different learning algorithms [Pensia *et al.*, 2018; Neu *et al.*, 2021; Wang *et al.*, 2021; Li and Liu, 2022].

Stochastic gradient descent (SGD) has become the workhorse behind modern DNNs training. Despite its simplicity, SGD also enables high efficiency in complex and non-convex optimization problems [Bottou *et al.*, 2018]. This motivates extensive research into provable generalization bounds for deep learning algorithms. The first line of research employs the concept of uniform stability, beginning with [Hardt *et al.*, 2016] on investigating convergence in expectation and followed by enormous efforts exploiting similar ideas [Bassily *et al.*, 2020; Lei *et al.*, 2021b; Yang *et al.*, 2021b; Yang *et al.*, 2021a]. Another line of research connects the generalization of DNNs with information-theoretic analysis [Xu and Raginsky, 2017], also demonstrating great potential in analyzing noisy and iterative learning algorithms: [Pensia *et al.*, 2018] is the first to investigate the generalization ability of stochastic gradient Langevin dynamics (SGLD, a variant of SGD that injects Gaussian noise at each iteration), whose result is improved by following studies [Negrea *et al.*, 2019; Wang *et al.*, 2021]; [Neu *et al.*, 2021] then establishes information-theoretic bounds for SGD by introducing virtual noises through an auxiliary weight process, whose bounds are subsequently tightened in [Wang and Mao, 2021]. Besides stability and information-theoretic views, researchers also provide PAC-Bayesian [Neyshabur *et al.*, 2018; Yang *et al.*, 2019] and model compression [Arora *et al.*, 2018; Zhou *et al.*, 2018] perspectives for generalization analysis.

Although current efforts on understanding and explaining the generalization of deep learning algorithms have yielded appealing results, these bounds are still restrictive due to their heavy reliance on strong assumptions or dimensionality of hypothesis spaces, making them easily become vacuous when applied to large-scale DNNs. For example, uniform stability-based generalization bounds usually assume Lipschitz continuity and smoothness of the empirical risk function [Hardt *et al.*, 2016; Lei *et al.*, 2021b] or global optimum assumptions such as convexity and the Polyak-Lojasiewicz (PL) condition [Lei *et al.*, 2021a; Li and Liu, 2022] to guarantee convergence, which is hard to meet in practice. On the contrary, information-theoretic generalization results do not

[1]Proofs available at https://github.com/Gamepiaynmo/KRE

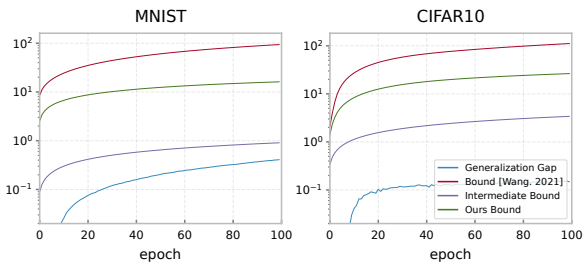

Figure 1: Generalization error of SGLD with MLP and CNN models. We provide more detailed analysis in Section 5.

rely on strong assumptions about the risk function, but the dimensionality of the hypothesis space, which often results in severely over-estimated upper bounds. As shown in Figure 1, there exists a 10^2 to 10^3 gap between the true generalization error and the SOTA information-theoretic generalization bound [Wang *et al.*, 2021]. Furthermore, the intractability of high-dimensional information quantities possesses extra obstacles to further tightening these bounds, since it is impractical to evaluate their tightness compared to the actual value of intermediate information quantities used in the proof, neither numerically nor theoretically.

In this paper, we establish computable information-theoretic bounds for noisy and iterative learning algorithms by adopting an alternative information measure, namely kernelized Rényi's entropy. This new information quantity inherits the elegant properties of the original Shannon's definition, while being directly computable from given samples and independent of the input dimensionality. We then bound the expected generalization error of these learning algorithms with kernelized Rényi's entropy, where each key information quantity in the bound could be directly accessed and visualized by simple random sampling during the training process. Based on our visualization results, we improve previous information-theoretic bounds by strictly tightened ones. As an example, the above-mentioned work [Wang *et al.*, 2021] upper bounds the key mutual information quantity in eq.(1) by the variance of the gradient, which is grossly over-estimated, being 10 to 10^2 times looser than the actual value (Intermediate Bound) as shown in Figure 1. This motivates us to reduce the gap by incorporating covariance between different dimensions of the gradient vector, which also applies to the work of [Pensia *et al.*, 2018; Wang and Mao, 2021], showing significant improvement on multiple deep learning benchmarks. In summary, the key contributions of this work include:

- We propose kernelized Rényi's entropy based on operator representation in Hilbert space. Unlike the classical Shannon's entropy, our information quantity is directly computable regardless of the dimensionality, while still being compatible with existing information-theoretic generalization frameworks.
- We establish mutual information generalization bounds for SGD and SGLD under the notion of kernelized Rényi's entropy and then visualize them on synthetic and real-world learning tasks. Our visualization results indicate multiple potential improvements in previous information-theoretic generalization results.
- We provide improved bounds based on one of our observations by considering correlations between different dimensions of the gradient vector. Empirical studies then demonstrate that our bounds are 5 more times tighter compared to previous SOTA results.

2 Preliminaries

Given random variable X, we denote the corresponding sample space by $\mathcal{X}$, samples by lower-case letter $\mathbf{x}$, and probability distribution function (PDF) by p_X. We write $\|\cdot\|$ to denote the Euclidean norm of a vector or the Frobenius norm of a matrix, and I_d to denote the d-dimensional identity matrix.

2.1 Problem Setting

Let $\mathcal{Z}$ be the instance space of interest and $\mathcal{W}$ be the hypotheses space. Let $S = \{Z_i\}_{i=1}^n$ be a dataset of n i.i.d. samples taking values in $\mathcal{Z}$ and $W \in \mathcal{W} \subset \mathbb{R}^d$ be the output of learning algorithm $\mathcal{A}$ according to some conditional distribution $P_{W|S}$ mapping from $\mathcal{Z}^n$ to $\mathcal{W}$. Let $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ be a loss function. We aim to seek for a parameter $w \in \mathbb{R}^d$ that minimizes the population risk L, defined by

$$L(w) \triangleq \mathbb{E}_Z[\ell(w, Z)].$$

Since the data distribution is usually unknown, we turn to minimize the empirical risk

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

For a learning algorithm $\mathcal{A}$ characterized by $P_{W|S}$, the corresponding generalization error is defined as the expected difference between $L(w)$ and $L_S(w)$, i.e.

$$\mathrm{gen}(W, S) \triangleq \mathbb{E}_{W,S}[L(W) - L_S(W)].$$

We assume throughout that $\ell(w, z)$ is differentiable almost everywhere with respect to w for any Z, and $\ell(w, Z)$ is R-subgaussian for any $w \in \mathcal{W}$. Under these assumptions, [Xu and Raginsky, 2017] shows that the generalization error of any learning algorithm $\mathcal{A}$ is bounded by

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

where $I(S; W)$ is the mutual information between the input dataset S and the output parameter vector W. Due to the high-dimensional nature of modern DNNs, this quantity is generally uncomputable, possessing extra obstacles to derive tightened generalization bounds.

2.2 Rényi's Entropy and Extensions

Recall that the Rényi's α-order entropy $H_\alpha(X)$ is defined on the PDF p_X for a given continuous random variable X in $\mathcal{X}$:

$$H_\alpha(X) \triangleq \frac{1}{1-\alpha} \log \int_{\mathcal{X}} p^\alpha(\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{2}$$

where the limit case $\alpha \to 1$ recovers Shannon's entropy. Exactly calculating this information quantity requires knowledge about the underlying data distribution, which is usually unknown in practice. To alleviate this issue, [Giraldo *et al.*, 2014] proposes a novel measure of entropy by utilizing the Hilbert space representation with finite data points. Specifically, it resembles quantum Rényi's entropy in terms of the eigenspectrum of a normalized Hermitian matrix constructed by projecting data points to a reproducing kernel Hilbert space (RKHS). In this paper, we follow this Hilbert space representation framework, with slight restrictions on the associated reproducing kernel:

Assumption 1. *Let $\kappa(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ be a reproducing kernel, where $\phi\colon \mathcal{X} \mapsto \mathcal{H}$ is the corresponding feature mapping. Assume that κ satisfies*

- *Normalized: $\kappa(\mathbf{x}, \mathbf{x}) = 1$ for any $\mathbf{x} \in \mathcal{X}$;*
- *Shift invariant: $\kappa(\mathbf{x}, \mathbf{x}') = f(\|\mathbf{x} - \mathbf{x}'\|)$ for some function $f\colon \mathbb{R}^+ \mapsto \mathbb{R}^+$;*
- *L_2 integrable: $\forall \mathbf{x} \in \mathcal{X}$, $\int_{\mathcal{X}} \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}' < \infty$.*

Given random variable $X \in \mathcal{X}$, define linear operator $G_X\colon \mathcal{H} \mapsto \mathcal{H}$ as $G_X f \triangleq \mathbb{E}_X[\phi(\mathbf{x})\langle \phi(\mathbf{x}), f \rangle]$. One can verify that $\mathrm{tr}(G_X) = 1$ when the kernel κ is normalized, so that the eigenvalues of G_X constitute a probability distribution which is a natural density estimator for the distribution of X.

3 Kernelized Rényi's Entropy: An Alternative Information Measure

In this section, we introduce kernelized Rényi's Entropy by extending the work of [Giraldo *et al.*, 2014] from finite-sample cases to infinite-sample cases, enabling direct analysis of entropy quantities based on the PDF, uninfluenced by the actual sampling process. Our definition inherits the elegant properties of the original Shannon's entropy by setting $\alpha \to 1$, while still being able to be directly accessed via simple random sampling.

Proposition 1. *Given linear operator G_X defined as above on random variable $X \in \mathcal{X}$ with PDF p_X, we have*

$$\lim_{\alpha \to 1} \frac{1}{1-\alpha} \log \mathrm{tr}(G_X^\alpha) = -\mathrm{tr}(G_X \log G_X) = -\iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}'.$$

Proposition 1 directly implies the following definition of kernelized Rényi's entropy of order $\alpha \to 1$:

Definition 1. *Given continuous random variable X and its PDF p_X, the kernelized Rényi's entropy for X of order $\alpha \to 1$ is defined as*

$$S_1(X) \triangleq -C_\kappa \iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}'.$$

where $C_\kappa = 1/\int_{\mathcal{X}} \kappa^2(0, \mathbf{x}) \, \mathrm{d}\mathbf{x} > 0$ is the normalizing factor that let the squared kernel function integrate to 1.

Compared with the classical Shannon's definition which is intractable for high-dimensional distributions, Definition 1 could be directly accessed regardless of the dimensionality. To this end, one can randomly sample m data points $\{\mathbf{x}_i\}_{i=1}^m$ from p_X, and denote $\hat{G}_X\colon \mathcal{H} \mapsto \mathcal{H}$ as an empirical version of G_X by $\hat{G}_X f \triangleq \frac{1}{m} \sum_{i=1}^m \phi(\mathbf{x}_i) \langle \phi(\mathbf{x}_i), f \rangle$. It can be verified that $\hat{G}_X$ is an unbiased estimate of G_X, which further implies the following finite-sample approximation to kernelized Rényi's entropy:

Proposition 2. *Let $\{\mathbf{x}_i\}_{i=1}^m$ be i.i.d. data points sampled from X, and let $K \in \mathbb{R}^{m \times m}$ be the kernel matrix constructed by $K_{ij} = \frac{1}{m} \kappa(\mathbf{x}_i, \mathbf{x}_j)$. Then with confidence $1 - \delta$,*

$$|S_1(X) - \hat{S}_1(X)| \leq \frac{9 C_\kappa \sqrt{2 \log \frac{2}{\delta}}}{\sqrt[3]{m}}, \tag{3}$$

where $\hat{S}_1(X) = -C_\kappa \mathrm{tr}(K \log K)$.

Note that the above concentration result only involves the number of samples while remaining independent of the dimension, which allows our kernelized entropy to be directly accessed in high-dimensional cases. This property is a significant benefit in analyzing the behavior of modern DNNs, which usually involve thousands or even millions of parameters. One can also notice that Definition 1 can be easily extended to multivariate joint entropy by taking $\kappa = \kappa_X \otimes \kappa_Y$ as the kernel function for the joint distribution $P_{X,Y}$. With these settings, kernelized Rényi's divergence and mutual information can be derived accordingly:

Definition 2. *Given probability measures P, Q on $\mathcal{X}$ and their PDF p, q, the kernelized Rényi's divergence between P and Q is defined as:*

$$D_1(P \parallel Q) \triangleq C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x}) \log \frac{p(\mathbf{x}')}{q(\mathbf{x}')} \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}'.$$

Definition 3. *Given normalized kernels κ_X, κ_Y, continuous random variables X, Y and their PDF p_X, p_Y, the kernelized Rényi's mutual information between X and Y is defined as:*

$$I_1(X;Y) \triangleq C_{\kappa_X} C_{\kappa_Y} \iiiint_{\mathcal{Y}^2 \times \mathcal{X}^2} p_{X,Y}(\mathbf{x}, \mathbf{y}) \cdot \log \frac{p_{X,Y}(\mathbf{x}', \mathbf{y}')}{p_X(\mathbf{x}') p_Y(\mathbf{y}')} \kappa_X^2(\mathbf{x}, \mathbf{x}') \kappa_Y^2(\mathbf{y}, \mathbf{y}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{y}'.$$

The main difference between Shannon's entropy and kernelized Rényi's entropy lies in the kernel function κ. Specifically, our definition recovers the original Shannon's entropy when κ is the Dirac-Delta function. To characterize the difference between them, we introduce the discrepancy function

$$u_X^\kappa(\mathbf{x}) \triangleq C_\kappa \int_{\mathcal{X}} [\log p_X(\mathbf{x}) - \log p_X(\mathbf{x}')] \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}',$$

and its expected version

$$E_X^\kappa(p) \triangleq \left| \int_{\mathcal{X}} p(\mathbf{x}) u_X(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right|.$$

We simply denote the **expected discrepancy** $E_X^\kappa(p_X)$ by $E_X^{\kappa\prime}$ and $E_X^\kappa(\hat{p}_X)$ by E_X^κ for convenience, where $\hat{p}_X(\mathbf{x}) \triangleq C_\kappa \int_{\mathcal{X}} p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}'$.

Proposition 3. *Let $\kappa(\mathbf{x}, \mathbf{x}') = \mathbb{1}_{\|\mathbf{x}-\mathbf{x}'\|<c}$. Assume that the PDF $p_X(\cdot)$ satisfies:*

- *Continuous: $\forall \mathbf{x} \in \mathcal{X}$, $\lim_{\mathbf{x}' \to \mathbf{x}} p_X(\mathbf{x}') = p_X(\mathbf{x})$;*
- *Positive: $\forall \mathbf{x} \in \mathcal{X}$, $\lim_{\mathbf{x}' \to \mathbf{x}} p_X(\mathbf{x}') > 0$;*

then we have $\lim_{c\to 0} E_X^\kappa \to 0$ and $\lim_{c\to 0} {E_X^\kappa}' \to 0$.

Proposition 3 indicates that when the normalized kernel function $C_\kappa \cdot \kappa^2$ has a very peaked bump, i.e. c is small, the expected discrepancy terms E_X^κ and ${E_X^\kappa}'$ both tend to 0. As we will show in Proposition 4, setting $c \to 0$ corresponds to the case where kernelized Rényi's entropy recovers the original Shannon's definition. The continuity assumption above is easily satisfied when X is a continuous random variable. The positiveness assumption is also naturally satisfied when X is truncated between some interval $[a, b]$ so that $p_X(x) > 0$ for any $x \in [a, b]$ (e.g. randomly sampled image data are always truncated by $[0, 255]$), or the distribution of X is tailed so that $p_X(\mathbf{x}) > 0$ for any finite $\mathbf{x} \in \mathcal{X}$ (e.g. the Gaussian distribution is widely used for model parameter initialization), which are common cases in modern DNNs.

Proposition 4. *Let $X, X' \in \mathcal{X}$, $Y \in \mathcal{Y}$, $Z \in \mathcal{Z}$ be continuous random variables with probability measures P_X, $P_{X'}$, P_Y and P_Z respectively. Then*

1. $H(X) \le S_1(X) \le H(X) + {E_X^\kappa}'$.
2. $D_1(P_X \parallel P_{X'}) \ge -E_X^\kappa$.
3. $I_1(X;Y) = D_1(P_{X,Y} \parallel P_X \otimes P_Y) \ge 0$.
4. $I_1(X;Y) = S_1(X) + S_1(Y) - S_1(X,Y)$.
5. $I_1(X;Y|Z) = I_1(X;Y,Z) - I_1(X;Z)$.
6. *Let X, Y, Z form Markov chain $X \to Y \to Z$, then $I_1(X;Y) \ge I_1(X;Z)$ and $I_1(Y;Z) \ge I_1(X;Z)$.*

Proposition 4 shows that kernelized Rényi's entropy inherits the essential properties of the original Shannon's entropy, thus guaranteeing compatibility with existing information theoretical analysis frameworks. Property 1 verifies that when $c \to 0$ in Proposition 3, kernelized Rényi's entropy recovers the original Shannon's entropy. Combining with the following properties, this conclusion also applies to divergence and mutual information quantities. Property 2 indicates that although kernelized Rényi's divergence is not guaranteed to be positive, it cannot be significantly less than 0 when the kernel κ is chosen properly. Property 4 and 5 imply that an estimate of the mutual information quantity could be acquired by estimating the value of multiple individual entropy quantities. Property 6 is the kernelized Rényi's entropy version of the data processing inequality.

In the sequel, we will use Gaussian kernel in kernelized Rényi's information quantities to bound the expected generalization error, i.e.

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / 2\sigma_\kappa^2),$$

where σ_κ is the kernel width. Note that there is a trade-off for the choice of σ_κ: A small σ_κ implies $E_X^\kappa \approx 0$ and reduces to Shannon's entropy. However, this will cause a large normalization factor C_κ and result in a large estimation error as shown in Proposition 2. In practice, we usually select σ_κ according to the top 10% to 20% Euclidean distances between all pairwise data points as suggested by [Yu *et al.*, 2019].

4 Generalization Bounds with Kernelized Rényi's Entropy

This section presents information-theoretic generalization bounds for iterative and noisy learning algorithms under kernelized Rényi's entropy. Firstly, we show that the mutual information bound for expected generalization error in eq.(1) also holds for our kernelized one:

Theorem 1. *Suppose that $\ell(w, Z)$ is R-subgaussian with respect to Z for every $w \in \mathcal{W}$, then*

$$|\mathbb{E}_{S,W}[L(W) - L_S(W)]| \le \sqrt{\frac{2R^2 \hat{I}_1(S;W)}{n}}, \quad \textit{and}$$

$$\mathbb{E}_{S,W}[L(W) - L_S(W)]^2 \le \frac{4R^2(\hat{I}_1(S;W) + \log 3)}{n},$$

where $\hat{I}_1(S;W) = I_1(S;W) + E_{S,W}^\kappa$.

Remark 1. *Theorem 1 provides a kernelized Rényi's entropy perspective for information-theoretic generalization [Xu and Raginsky, 2017]. As indicated by Proposition 3, $E_{S,W}^\kappa$ is the expected discrepancy of the joint distribution between S and W associated with κ, which vanishes when $\sigma_\kappa \to 0$. It is worth noting that Theorem 1 upper bounds both the expectation and the variance of $\mathrm{gen}(W, S)$ by the same quantity $I_1(S;W)$, thus also yields high-probability bounds for the generalization error through concentration inequalities e.g. Markov's and Chebyshev's inequalities.*

Next, we apply our generalization result on mini-batched iterative and noisy learning algorithms for empirical risk minimization. Suppose algorithm $\mathcal{A}$ finishes in T steps, and let $W_0 \in \mathcal{W}$ be the initial parameter vector. At the t-th step, a batch of data points $B_t \subset S$ independent from the current parameter vector is randomly selected and used to compute a direction for gradient descent:

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

Then the updating rule can be formalized by

$$W_t = W_{t-1} - \eta_t g(W_{t-1}, B_t) + \xi_t, \tag{4}$$

where W_t denotes the parameter vector at t-th step, η_t is the learning rate and $\xi_t \in \mathcal{W}$ is a random vector independent from W_{t-1} and B_t. Obviously, $W_0 \to W_1 \to \cdots \to W_T$ forms a Markov chain.

4.1 Stochastic Gradient Langevin Dynamics

The SGLD algorithm is a variant of the classical SGD algorithm by injecting random noises in each gradient update as shown in eq.(4). A common choice is the isotropic Gaussian noise, i.e. $\xi_t \sim N(0, \sigma_t^2 I_d)$, since it has the maximum entropy for a fixed variance σ_t^2 which leads to the tightest upper bound. [Pensia *et al.*, 2018] derived the following information-theoretic generalization bound for SGLD:

Lemma 1. *Let W_T be the parameter vector acquired by the SGLD algorithm after T updates, then*

$$I(W_T; S) \le \sum_{t=1}^{T} \frac{d}{2} \log\left(\frac{\eta_t^2 L}{d\sigma_t^2} + 1\right), \tag{5}$$

where $L = \max_{w \in \mathcal{W}, z \in \mathcal{Z}} \|g(w, z)\|_2^2$.

Note that the constant L in Lemma 1 is upper bounded by the square of the Lipschitz constant if $\ell(w, z)$ is Lipschitz continuous with regard to w, and the bound is dimension-dependent. This result is then improved in [Wang *et al.*, 2021] by removing the Lipschitz assumption:

Lemma 2. *Under the same conditions of Lemma 1:*

$$I(W_T; S) \leq \sum_{t=1}^{T} \frac{\eta_t^2 V_t}{2\sigma_t^2}, \tag{6}$$

*where V_t is the **gradient variance** at step t, defined by*

$$V_t \triangleq \mathbb{E}_{W_{t-1}, B_t}\left[\|g(W_{t-1}, B_t) - \mathbb{E}_{B_t}[g(W_{t-1}, B_t)]\|_2^2\right].$$

At first glance, the above bound does not depend on the dimensionality d. The gradient variance V_t, however, actually relies on d since it is the summation of the variance raised by each dimension of the stochastic gradient vector. The main reason is that they use isotropic Gaussian distributions to upper bound the entropy of stochastic gradient, being severely over-estimated according to our empirical results (see Figure 2 and 3). To address this issue, we consider using the correlation between different dimensions of the gradient, which yields strictly tighter bounds for SGLD:

Theorem 2. *Under the same conditions of Lemma 1:*

$$I_1(W_T; S) \leq \sum_{t=1}^{T} I_1(W_t; B_t | W_{t-1}) \tag{7}$$

$$\leq \sum_{t=1}^{T} \left(\frac{1}{2} \log \left| \frac{\eta_t^2}{\sigma_t^2} \mathbb{V}_t + I \right| + E^{\kappa}_{W_t | W_{t-1}} \right), \tag{8}$$

$$I(W_T; S) \leq \sum_{t=1}^{T} \frac{1}{2} \log \left| \frac{\eta_t^2}{\sigma_t^2} \mathbb{V}_t + I \right|, \tag{9}$$

*where $\mathbb{V}_t = \mathrm{Cov}[g(W_{t-1}, B_t)]$ is the **gradient covariance** matrix and $|\cdot|$ denotes the matrix determinant.*

Remark 2. *Theorem 2 asserts that the kernelized Rényi's mutual information $I_1(W_T; S)$ is upper bounded by the determinant of the gradient covariance matrix, which involves the full correlation between different dimensions of the gradient vector. The limit case $\sigma_\kappa \to 0$ implies an upper bound for Shannon's mutual information $I(W_T; S)$ in eq.(9). Note that the kernelized Rényi's information quantities in eq.(7) can be directly calculated from B_t, enabling us to validate the tightness of these intermediate bounds. Combining with Theorem 1, one can obtain upper bounds for the expected generalization error of SGLD.*

The following proposition shows that our bound is strictly tighter than that of Lemma 1 and 2.

Proposition 5. *Given $\mathbb{V}_t$, V_t and L defined as above, let $\{c_i\}_{i=1}^r$ be a disjoint partition of $\{n\}$, i.e. $c_1 \cup \cdots \cup c_r = \{n\}$ and $c_i \cap c_j = \Phi$ for any $i \neq j$. Let $\mathbb{V}_t^i$ be the sub-matrix of $\mathbb{V}_t$ with columns and rows indexed by c_i, and define*

$$\theta_c(\mathbb{V}) = \frac{1}{2} \log \left| \frac{\eta_t^2}{\sigma_t^2} \mathbb{V} + I \right|, \quad \theta_v(V) = \frac{d}{2} \log \left(\frac{\eta_t^2 V}{d\sigma_t^2} + 1 \right),$$

$$\textit{then} \qquad \theta_c(\mathbb{V}_t) \leq \sum_{i=1}^{r} \theta_c(\mathbb{V}_t^i) \leq \theta_v(V_t) \leq \theta_v(L).$$

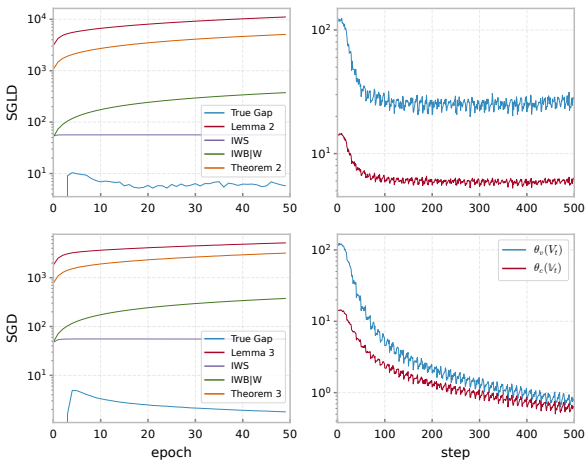

Figure 2: Comparison of generalization bounds on synthetic data.

Remark 3. *The quantities θ_c and θ_v correspond to the upper bounds in Theorem 2 and Lemma 2 respectively. When the model size d is large, it is infeasible to calculate the entire covariance matrix $\mathbb{V}_t$ due to limited memory. Proposition 5 suggests an alternative upper bound for $\theta_c(\mathbb{V}_t)$, which could be calculated using much lower memory by dividing the parameter vector into different groups according to their correlation (e.g. let each layer of the model be a group), calculating $\theta_c(\mathbb{V}_t^i)$ for each group and then summing them up. In the limit case where every single parameter of W represents a group, calculating θ_c requires no more memory than θ_v, while still being strictly tighter than the latter one.*

4.2 Stochastic Gradient Descent

Unlike SGLD, the SGD algorithm does not involve random noises in each update, i.e. $\xi_t = 0$. This actually causes extra difficulty for information-theoretic generalization analysis as the strategy used to derive the SGLD bound is no longer available. To circumvent the issue, [Neu *et al.*, 2021] proposes to introduce an auxiliary weight process $\tilde{W}_t$ that manually includes virtual noises, and then bridges the differences between these two processes (W_t and $\tilde{W}_t$). Let

$$\tilde{W}_0 = W_0, \quad \text{and}$$
$$\tilde{W}_t = \tilde{W}_{t-1} - \eta_t g(W_{t-1}, B_t) + \tilde{\xi}_t, \quad \text{for } t > 0,$$

where $\tilde{\xi}_t \sim N(0, \sigma_t^2 I)$ are random Gaussian vectors. Obviously, we have $\tilde{W}_t = W_t + \Delta_t$, where $\Delta_t = \sum_{i=1}^{t} \tilde{\xi}_i$. The recent work [Wang and Mao, 2021] establishes the following information-theoretic generalization bound for SGD:

Lemma 3. *Assume that $L(W_T) \leq \mathbb{E}_{\Delta_t}[L(W_T + \Delta_t)]$ and ℓ is twice differentiable. Then for any $\sigma_1, \cdots, \sigma_T > 0$, we have*

$$\mathrm{gen}(W_T; S) \leq \frac{1}{2} \sum_{t=1}^{T} \sigma_t^2 \mathbb{E}_{W_T}[\mathbb{H}(W_T)] + |\mathrm{gen}(\tilde{W}_T; S)|,$$

$$\textit{and} \qquad I\left(\tilde{W}_T; S\right) \leq \sum_{t=1}^{T} \frac{d}{2} \log \left(\frac{\eta_t^2 V_t}{d\sigma_t^2} + 1 \right), \tag{10}$$

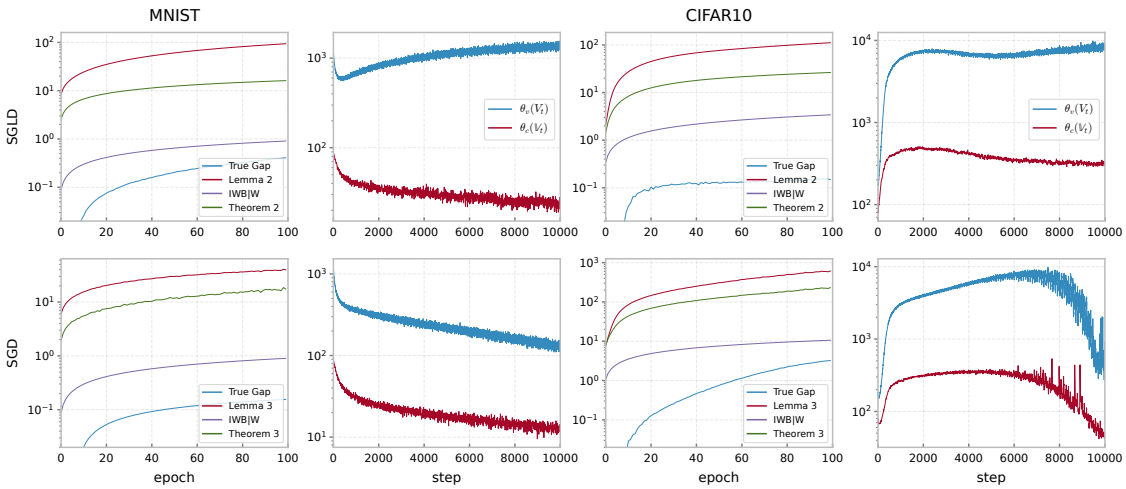

Figure 3: Visualization and comparison of information-theoretic generalization bounds for SGLD and SGD on MNIST and CIFAR10.

where $\mathbb{H}(W_T) = \mathbb{E}_Z[\mathrm{tr}(H_{W_T}(Z))]$ *and* $H_{W_T}(Z)$ *is the Hessian matrix of the loss* $\ell(W_T, Z)$ *with respect to* W_T.

Again, the above mutual information bound relies on d. We adopt the same strategy that was explored in Theorem 2 to alleviate this issue:

Theorem 3. *Assume* κ *satisfies Assumption 1 and under the same conditions of Lemma 3:*

$$I_1\Big(\tilde{W}_T; S\Big) \leq \sum_{t=1}^{T}\left(\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right| + E^{\kappa}_{\tilde{W}_t|\tilde{W}_{t-1}}\right),$$

$$I\Big(\tilde{W}_T; S\Big) \leq \sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right|. \tag{11}$$

Remark 4. *Theorem 3 establishes information-theoretic generalization bound for the SGD algorithm within the framework of kernelized Rényi's entropy, where eq.(11) corresponds to the limit case* $\sigma_\kappa \to 0$. *We highlight that our result upper bounds the generalization error by gradient covariance, which is strictly tighter than that of Lemma 3 since* $\theta_c(\mathbb{V}_t) \leq \theta_v(V_t)$ *as shown in Proposition 5. Note that the bounds in Theorem 2 and 3 could be further improved by taking into account higher order moments of the stochastic gradient, but this would impose a significant computational burden because the s-th moment tensor is of size* d^s, *making the potential improvement less meaningful in practice.*

5 Empirical Studies

In this section, we visualize the computable generalization bounds for SGLD/SGD derived in the previous sections, and verify the tightness of previous results as well as our improved ones in Theorem 2 and 3. For simplicity, we use constant values for learning rates $\eta_t = \eta$ and Gaussian noises $\sigma_t = \sigma$. We ignore the expected discrepancy E^{κ}_X terms in computation since they tend to 0 by taking appropriate σ_κ values, and are in fact not computable. Advanced tuning techniques such as momentum, weight decay, and batch normalization are not adopted. To compute R in Theorem 1, we collect the loss values of each batch in each epoch and let $R = \frac{1}{2}[\max_t \ell(W_{t-1}, B_t) - \min_t \ell(W_{t-1}, B_t)]$. To compute $\mathbb{V}_t$ and V_t in the mutual information upper bounds above, we use the BackPack Pytorch library [Dangel *et al.*, 2020] to acquire an empirical estimate of $\mathbb{V}_t$ and V_t from each batch input of data. To compute $\mathbb{H}(W_T)$ in Theorem 3, we use the PyHessian library to acquire the Hessian matrix. Each experiment is repeated 100 times to acquire i.i.d. samples of W_t and B_t, which are then used to construct the kernel matrix K in eq.(3) to compute the kernelized Rényi's mutual information in our information-theoretic upper bounds.

5.1 Synthetic Data

Our first experiment incorporates a simple linear regression problem

$$y = \mathbf{w}^\top \mathbf{x} + \varepsilon,$$

where $\mathbf{x}$ is the 10 dimensional input vector, y is the regression target, $\mathbf{w}$ is the linear coefficient and ε is some zero-mean random noise. We train an MLP with one hidden layer of width 10. For each of the 100 independent training processes, we generate an independent training dataset of size $n = 100$ using the same strategy. The comparison of existing theoretical generalization bounds against the true generalization gap is shown in the left side of Figure 2, in which we denote the information-theoretic bounds by the corresponding key information quantities: $I_1(W_T; S)$ by IWS, and the summation of $I_1(W_t; B_t|W_{t-1})$ by IWB|W. Note that although eq.(7) is derived under the context of the SGLD algorithm, it still holds for the SGD algorithm since the only prerequisite of this inequality is the Markov chain relationship $S \to \{B_t\}_{t=1}^T \to \{W_t\}_{t=1}^T$.

As can be seen, IWS is always smaller than IWB|W, and both of them consistently fall in the interval between the curve of the True Gap and the bound of eq.(9) for SGLD (or eq.(11) for SGD), indicating that our approximations successfully reflect the actual behavior of these information-theoretic quantities $I(W_t; S)$ and $I(W_t; B_t|W_{t-1})$. We can gain several important insights from the visualization results:

1) The gap between IWS and the true generalization gap indicates that the sub-gaussian constant R of the loss function $\ell(w, Z)$ with respect to Z is over-estimated. It is natural to assume that well-trained DNNs yield lower loss than a random initialization, and thus the constant R is expected to decrease along with the training process. This observation could be adopted to further tighten the bound in eq.(1) and Theorem 1.

2) IWS quickly reaches the peak and then turns to decrease along with the training process (one can refer to the Appendix for more details), whose behavior matches that of the True Gap curve, while IWB|W keeps increasing since each $I(W_t; B_t|W_{t-1})$ is always strictly positive. This observation indicates that although the model always learns some knowledge from a new batch (i.e. $I(W_t; B_t|W_{t-1}) \geq 0$), the total information that W contains about the dataset S (i.e. $I(W_t; S)$) quickly reaches the upper limit: the network is actually forgetting information that learned previously. This "forget" behavior is not captured by the current work of information-theoretic generalization bounds, resulting in a gradually increasing gap between IWS and IWB|W when the number of training epochs grows large.

3) The remaining gap between IWB|W and our improved bounds indicates that current bounds are still far from optimal even if the full correlation of the noisy gradient is considered. This observation is supported by the recent works [Gurbuzbalaban *et al.*, 2021; Camuto *et al.*, 2021], who claim that the stochastic gradient vector generated by SGD is heavy-tailed and their entropy is significantly over-estimated by assuming Gaussian distributions. Another conjecture is that some implicit self-regularization mechanisms exist in DNNs [Mahoney and Martin, 2019; Martin and Mahoney, 2021], resulting in the information captured by the weights being much lower than their theoretical capacity.

The right side of Figure 2 provides an intuitive comparison between the upper bounds θ_c (gradient covariance matrix) in Theorem 2 and θ_v (gradient variance) in Lemma 2. It can be seen that $\theta_v(V_t)$ is always larger than $\theta_c(\mathbb{V}_t)$, especially at the beginning of the training process. This observation verifies our claim in Proposition 5.

5.2 Real-world Data

We then visualize our computable generalization bounds on real-world datasets to demonstrate the scalability of kernelized Rényi's entropy. Following the experiment settings in [Wang *et al.*, 2021], we train an MLP with a wider hidden layer on MNIST and a 4-layer CNN on CIFAR10. In each of the 100 individual training processes, a portion of data pairs is uniformly sampled from the entire dataset as the training dataset to simulate the randomness of S. Detailed experiment settings can be found in Appendix.

Similarly, the comparisons between different generalization bounds are reported in Figure 3. As can be seen, the curve of IWB|W still consistently falls in the correct interval between adjacent bounds, and perfectly reflect the increasing trend of the gradient covariance upper bound. The gradient variance bound in eq.(6) is still grossly over-estimated compared to IWB|W. For comparison, our tightened bound of eq.(9) (or eq.(11)) covers a large portion of this gap between the curves of IWB|W and Lemma 2 for SGLD (or Lemma 3 for SGD). Moreover, it can be seen that during the training process of the SGLD algorithm, the curve of θ_c quickly stops increasing and starts to decrease in the latter epochs, while the curve of θ_v is consistently increasing along with the whole training process. This observation further verifies the tightness of our improved generalization bounds.

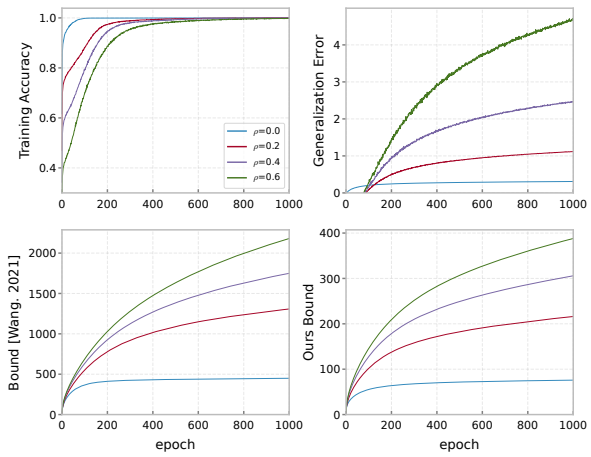

Figure 4: Random label experiment on MNIST.

Next, we conduct random label experiments to demonstrate the tightness of our improved generalization bounds under different levels of label noises. We keep the same experiment settings as above, while randomly replacing the training labels with noisy labels with a certain probability specified by the hyper-parameter ρ. As shown in Figure 4, higher levels of label noise lead to higher generalization errors. While both the bound of [Wang and Mao, 2021] and ours successfully reflect the trend of the true generalization gap alongside the training process, our bound is 5 more times tighter than theirs in Lemma 3. We refer the readers to the Appendix for extra experimental results on CIFAR10 and varying model sizes.

6 Conclusion

In this work, we address the common issue that Shannon's information quantities are intractable for estimation in practice. This possesses extra obstacles for information-theoretic generalization analysis, since it is impossible to evaluate the tightness of any intermediate information quantities ($I(W; S)$, $I(W_t; B_t|W_{t-1})$) used by previous generalization bounds [Wang *et al.*, 2021; Wang and Mao, 2021], resulting in the corresponding upper bounds (Lemma 2 and 3) being severely over-estimated. To address this issue, we propose an alternative measure of entropy named kernelized Rényi's entropy, which could be directly estimated regardless of the dimensionality, and still be compatible with existing generalization analysis frameworks. We successfully apply it to derive and visualize information-theoretic generalization bounds for noisy and iterative learning algorithms, indicating multiple potential directions for further improvement. We then prove tightened bounds for SGLD and SGD based on one of these findings, demonstrating significant improvement over existing works on multiple deep learning benchmarks.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2021ZD0110700), National Natural Science Foundation of China (62106191, 12071166, 62192781, 61721002), the Research Council of Norway (RCN) under grant 309439, Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Centre for Engineering Science and Technology and Project of Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for The Belt and Road Training in MOOC China).

References

[Arora *et al.*, 2018] 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*, pages 254–263. PMLR, 2018.

[Bartlett *et al.*, 2017] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

[Bassily *et al.*, 2020] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

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

[Camuto *et al.*, 2021] Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gurbuzbalaban, and Umut Simsekli. Asymmetric heavy tails and implicit bias in gaussian noise injections. In *International Conference on Machine Learning*, pages 1249–1260. PMLR, 2021.

[Dangel *et al.*, 2020] Felix Dangel, Frederik Kunstner, and Philipp Hennig. Backpack: Packing more into backprop. In *International Conference on Learning Representations*, 2020.

[Giraldo *et al.*, 2014] Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.

[Gurbuzbalaban *et al.*, 2021] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pages 3964–3975. PMLR, 2021.

[Hardt *et al.*, 2016] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[Harutyunyan *et al.*, 2021] Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems*, 34:24670–24682, 2021.

[Lei *et al.*, 2021a] Yunwen Lei, Ting Hu, and Ke Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *Journal of Machine Learning Research*, 22:1–41, 2021.

[Lei *et al.*, 2021b] Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186. PMLR, 2021.

[Li and Liu, 2022] Shaojie Li and Yong Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning*, pages 12931–12963. PMLR, 2022.

[Mahoney and Martin, 2019] Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.

[Martin and Mahoney, 2021] Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *J. Mach. Learn. Res.*, 22(165):1–73, 2021.

[Negrea *et al.*, 2019] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

[Neu *et al.*, 2021] Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

[Neyshabur *et al.*, 2018] 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*, 2018.

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

[Wang and Mao, 2021] Ziqiao Wang and Yongyi Mao. On the generalization of models trained with sgd: Information-theoretic bounds and implications. In *International Conference on Learning Representations*, 2021.

[Wang *et al.*, 2021] Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of sgld using properties of gaussian channels. *Advances in Neural Information Processing Systems*, 34:24222–24234, 2021.

[Xu and Raginsky, 2017] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

[Yang *et al.*, 2019] Jun Yang, Shengyang Sun, and Daniel M Roy. Fast-rate pac-bayes generalization bounds via shifted rademacher processes. *Advances in Neural Information Processing Systems*, 32, 2019.

[Yang *et al.*, 2021a] Zhenhuan Yang, Yunwen Lei, Siwei Lyu, and Yiming Ying. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss. In *International Conference on Artificial Intelligence and Statistics*, pages 2026–2034. PMLR, 2021.

[Yang *et al.*, 2021b] Zhenhuan Yang, Yunwen Lei, Puyu Wang, Tianbao Yang, and Yiming Ying. Simple stochastic and online gradient descent algorithms for pairwise learning. *Advances in Neural Information Processing Systems*, 34:20160–20171, 2021.

[Yu *et al.*, 2019] Shujian Yu, Luis Gonzalo Sanchez Giraldo, Robert Jenssen, and Jose C Principe. Multivariate extension of matrix-based rényi's α-order entropy functional. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2960–2966, 2019.

[Zhang *et al.*, 2021] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[Zhou *et al.*, 2018] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. In *Advances in Neural Information Processing Systems*, 2018.

Appendix of "Understanding the Generalization Ability of Deep Learning Algorithms: A Kernelized Rényi's Entropy Perspective"

A Proof of Section 3

A.1 Proof of Proposition 1

Proposition 1 (Restate)**.** *Given linear operator G_X defined as above on random variable $X \in \mathcal{X}$ with PDF p_X, we have*

$$\lim_{\alpha\to 1}\frac{1}{1-\alpha}\log \mathrm{tr}(G_X^\alpha) = -\mathrm{tr}(G_X \log G_X) = -\iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}')\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'.$$

Proof. Let $\{\psi_i\}_{i=1}^{N_\mathcal{H}}$ be a complete orthogonal basis for $\mathcal{H}$. Then by L'Hopital's rule:

$$\begin{aligned}
\lim_{\alpha\to 1}\frac{1}{1-\alpha}\log \mathrm{tr}(G_X^\alpha) &= \lim_{\alpha\to 1} -\frac{\frac{\partial}{\partial\alpha}\mathrm{tr}(G_X^\alpha)}{\mathrm{tr}(G_X^\alpha)} = -\mathrm{tr}(G_X \log G_X)\\
&= -\sum_{i=1}^{N_\mathcal{H}} \langle G_X\psi_i, \log G_X\psi_i\rangle\\
&= -\sum_{i=1}^{N_\mathcal{H}} \int_\mathcal{X} p_X(\mathbf{x})\langle\psi_i,\phi(\mathbf{x})\rangle\langle\phi(\mathbf{x}),\log G_X\psi_i\rangle\,\mathrm{d}\mathbf{x}\\
&= -\int_\mathcal{X} p_X(\mathbf{x})\langle\log G_X\phi(\mathbf{x}), \sum_{i=1}^{N_\mathcal{H}}\langle\psi_i,\phi(\mathbf{x})\rangle\psi_i\rangle\,\mathrm{d}\mathbf{x}\\
&= -\int_\mathcal{X} p_X(\mathbf{x})\langle\log G_X\phi(\mathbf{x}),\phi(\mathbf{x})\rangle\,\mathrm{d}\mathbf{x}\\
&= -\int_\mathcal{X} p_X(\mathbf{x})\langle\int_\mathcal{X}\log p_X(\mathbf{x}')\phi(\mathbf{x}')\langle\phi(\mathbf{x}'),\phi(\mathbf{x})\rangle\,\mathrm{d}\mathbf{x}',\phi(\mathbf{x})\rangle\,\mathrm{d}\mathbf{x}\\
&= -\iint_{\mathcal{X}^2} p_X(\mathbf{x})\log p_X(\mathbf{x}')\langle\phi(\mathbf{x}'),\phi(\mathbf{x})\rangle\langle\phi(\mathbf{x}'),\phi(\mathbf{x})\rangle\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'\\
&= -\iint_{\mathcal{X}^2} p_X(\mathbf{x})\log p_X(\mathbf{x}')\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'.
\end{aligned}$$

□

A.2 Proof of Proposition 2

Proposition 2 (Restate)**.** *Let $\{\mathbf{x}_i\}_{i=1}^m$ be i.i.d. data points sampled from X, and let $K \in \mathbb{R}^{m\times m}$ be the kernel matrix constructed by $K_{ij} = \frac{1}{m}\kappa(\mathbf{x}_i,\mathbf{x}_j)$. Then with confidence $1-\delta$,*

$$|S_1(X) - \hat{S}_1(X)| \le \frac{9C_\kappa\sqrt{2\log\frac{2}{\delta}}}{\sqrt[3]{m}},$$

where $\hat{S}_1(X) = -C_\kappa \mathrm{tr}(K\log K)$.

Proof. Let λ_i and μ_i, $i\in[1,m]$ be the eigenvalues of G_X and $\hat{G}_X$ respectively. Following the proof of Theorem 6.2 in [Giraldo *et al.*, 2014] while taking $\varphi(x) = |x|$, we have that with probability $1-\delta$,

$$\sum_{i=1}^m |\lambda_i - \mu_i| \le C\sqrt{\frac{2\log\frac{2}{\delta}}{m}},$$

where $C = \max_{\mathbf{x}\in\mathcal{X}}\kappa(\mathbf{x},\mathbf{x}) = 1$. Then for any $s>0$, we have

$$\left|\mathrm{tr}(G_X\log G_X) - \mathrm{tr}(\hat{G}_X\log\hat{G}_X)\right| = \left|\sum_{i=1}^m \lambda_i\log\lambda_i - \sum_{i=1}^m \mu_i\log\mu_i\right| \le \sum_{i=1}^m |\lambda_i\log\lambda_i - \mu_i\log\mu_i|$$

$$\leq \sum_{i=1}^{m} \max(-|\lambda_i - \mu_i| \log |\lambda_i - \mu_i|, -(1 - |\lambda_i - \mu_i|) \log(1 - |\lambda_i - \mu_i|)) \tag{12}$$

$$\leq \sum_{i=1}^{m} -\sqrt{\frac{2 \log \frac{2}{\delta}}{m^3}} \log \sqrt{\frac{2 \log \frac{2}{\delta}}{m^3}} = \sqrt{\frac{2 \log \frac{2}{\delta}}{m}} \log \sqrt{\frac{m^3}{2 \log \frac{2}{\delta}}}$$

$$\leq s \sqrt{\frac{2 \log \frac{2}{\delta}}{m}} \left(\frac{m^3}{2 \log \frac{2}{\delta}} \right)^{\frac{1}{2s}} \leq s m^{\frac{3}{2s} - \frac{1}{2}} \sqrt{2 \log \frac{2}{\delta}}, \tag{13}$$

where (12) is maximized when $|\lambda_1 - \mu_1| = \cdots = |\lambda_n - \mu_n| = \frac{1}{m} \sqrt{\frac{2 \log \frac{2}{\delta}}{m}}$, and (13) follows by the fact that for any $t > 0$, $\log x \leq x^t / t$. By taking $s = 9$, we have

$$\begin{aligned} |S_1(X) - \hat{S}_1(X)| &= C_\kappa |\mathrm{tr}(G_X \log G_X) - \mathrm{tr}(K \log K)| \\ &= C_\kappa \left| \mathrm{tr}(G_X \log G_X) - \mathrm{tr}(\hat{G}_X \log \hat{G}_X) \right| \leq \frac{9 C_\kappa \sqrt{2 \log \frac{2}{\delta}}}{\sqrt[3]{m}}. \end{aligned}$$

□

A.3 Proof of Proposition 3

Proposition 3 (Restate). *Let $\kappa(\mathbf{x}, \mathbf{x}') = \mathbb{1}_{\|\mathbf{x} - \mathbf{x}'\| < c}$. Assume that the PDF $p_X(\cdot)$ satisfies:*

- *Continuous: $\forall \mathbf{x} \in \mathcal{X}$, $\lim_{\mathbf{x}' \to \mathbf{x}} p_X(\mathbf{x}') = p_X(\mathbf{x})$;*
- *Positive: $\forall \mathbf{x} \in \mathcal{X}$, $\lim_{\mathbf{x}' \to \mathbf{x}} p_X(\mathbf{x}') > 0$;*

then we have $\lim_{c \to 0} E_X^\kappa \to 0$ and $\lim_{c \to 0} {E_X^\kappa}' \to 0$.

Proof. For any PDF $p(\cdot)$ defined on $\mathcal{X}$:

$$\begin{aligned} \lim_{c \to 0} E_X^\kappa(p) &= \lim_{c \to 0} C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x}) [\log p_X(\mathbf{x}) - \log p_X(\mathbf{x}')] \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\ &= \lim_{c \to 0} C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x}) [\log p_X(\mathbf{x}) - \log p_X(\mathbf{x} + \mathbf{x}')] \kappa^2(0, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\ &= C_\kappa \int_{\mathcal{X}} \left\{ \lim_{c \to 0} \int_{\|\mathbf{x} - \mathbf{x}'\| < c} p(\mathbf{x}) [\log p_X(\mathbf{x}) - \log p_X(\mathbf{x} + \mathbf{x}')] \, \mathrm{d}\mathbf{x} \right\} \mathrm{d}\mathbf{x}' \\ &= C_\kappa \int_{\mathcal{X}} 0 \, \mathrm{d}\mathbf{x}' = 0. \end{aligned}$$

□

A.4 Proof of Proposition 4

Proposition 4 (Restate). *Let $X, X' \in \mathcal{X}$, $Y \in \mathcal{Y}$, $Z \in \mathcal{Z}$ be continuous random variables with probability measures P_X, $P_{X'}$, P_Y and P_Z respectively. Then*

1. *$H(X) \leq S_1(X) \leq H(X) + {E_X^\kappa}'$.*
2. *$D_1(P_X \parallel P_{X'}) \geq -E_X^\kappa$.*
3. *$I_1(X; Y) = D_1(P_{X,Y} \parallel P_X \otimes P_Y) \geq 0$.*
4. *$I_1(X; Y) = S_1(X) + S_1(Y) - S_1(X, Y)$.*
5. *$I_1(X; Y | Z) = I_1(X; Y, Z) - I_1(X; Z)$.*
6. *Let X, Y, Z form Markov chain $X \to Y \to Z$, then $I_1(X; Y) \geq I_1(X; Z)$ and $I_1(Y; Z) \geq I_1(X; Z)$.*

Proof of Property 1.

$$\begin{aligned} S_1(X) - H(X) &= \int_{\mathcal{X}} p_X(\mathbf{x}) \log p_X(\mathbf{x}) \, \mathrm{d}\mathbf{x} - C_\kappa \iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\ &= \int_{\mathcal{X}} p_X(\mathbf{x}) \log p_X(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \int_{\mathcal{X}} p_X(\mathbf{x}) \left(C_\kappa \int_{\mathcal{X}} \log p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}' \right) \mathrm{d}\mathbf{x} \end{aligned}$$

$$
\begin{aligned}
&\geq \int_{\mathcal{X}} p_X(\mathbf{x}) \log p_X(\mathbf{x}) \,\mathrm{d}\mathbf{x} - \int_{\mathcal{X}} p_X(\mathbf{x}) \left(\log C_\kappa \int_{\mathcal{X}} p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x}' \right) \mathrm{d}\mathbf{x} \qquad (14) \\
&= \int_{\mathcal{X}} p_X(\mathbf{x}) \log \frac{p_X(\mathbf{x})}{C_\kappa \int_{\mathcal{X}} p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x}'} \,\mathrm{d}\mathbf{x} \\
&= \mathrm{KL}(P_X \parallel Q_X) \geq 0,
\end{aligned}
$$

where (14) follows by Jensen's inequality, and Q_X is the probability measure with PDF $q_X(\mathbf{x}) = C_\kappa \int_{\mathcal{X}} p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x}'$. Meanwhile,

$$
\begin{aligned}
S_1(X) &= -C_\kappa \iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x} \,\mathrm{d}\mathbf{x}' \\
&= -\int_{\mathcal{X}} p_X(\mathbf{x}) \left(C_\kappa \int_{\mathcal{X}} \log p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x}' \right) \mathrm{d}\mathbf{x} \\
&= -\int_{\mathcal{X}} p_X(\mathbf{x}) \log p_X(\mathbf{x}) \,\mathrm{d}\mathbf{x} - \int_{\mathcal{X}} p_X(\mathbf{x}) \left[C_\kappa \int_{\mathcal{X}} (\log p_X(\mathbf{x}') - \log p_X(\mathbf{x})) \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x}' \right] \mathrm{d}\mathbf{x} \\
&\leq H(X) + {E_X^\kappa}'.
\end{aligned}
$$

□

Proof of Property 2. Let p and q be the PDF of X and X' respectively. Then consider the following functional on $q(\mathbf{x})$:

$$
J(q) = C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x}) \log \frac{p(\mathbf{x}')}{q(\mathbf{x}')} \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x} \,\mathrm{d}\mathbf{x}' - \eta_0 \left(\int_{\mathcal{X}} q(\mathbf{x}) \,\mathrm{d}\mathbf{x} - 1 \right),
$$

where η_0 is the Lagrange multiplier that ensure $q(\mathbf{x})$ is a probability distribution. The divergence $D_1(P \parallel Q)$ attains an extremum when the functional derivative is equal to zero:

$$
\frac{\partial J}{\partial q} = -C_\kappa \int_{\mathcal{X}} \frac{p(\mathbf{x})}{q(\mathbf{x}')} \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x} - \eta_0 = -\frac{C_\kappa \int_{\mathcal{X}} p(\mathbf{x}) \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x}}{q(\mathbf{x}')} - \eta_0 = 0,
$$

which indicates that the minimizer $\hat{q}(\mathbf{x}')$ satisfies $\hat{q}(\mathbf{x}') \propto C_\kappa \int_{\mathcal{X}} p(\mathbf{x}) \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x}$. Combing with $\int_{\mathcal{X}} \hat{q}(\mathbf{x}') \,\mathrm{d}\mathbf{x}' = 1$, we have

$$
\hat{q}(\mathbf{x}') = C_\kappa \int_{\mathcal{X}} p(\mathbf{x}) \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x}.
$$

Therefore,

$$
\begin{aligned}
D_1(P \parallel Q) &\geq C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x}) \log \frac{p(\mathbf{x}')}{\hat{q}(\mathbf{x}')} \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x} \,\mathrm{d}\mathbf{x}' \\
&= \int_{\mathcal{X}} \left(C_\kappa \int_{\mathcal{X}} p(\mathbf{x}) \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x} \right) \log \frac{p(\mathbf{x}')}{\hat{q}(\mathbf{x}')} \,\mathrm{d}\mathbf{x}' \\
&\geq -\int_{\mathcal{X}} \hat{q}(\mathbf{x}') \left(C_\kappa \int_{\mathcal{X}} (\log p(\mathbf{x}') - \log p(\mathbf{x})) \kappa^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x} \right) \mathrm{d}\mathbf{x}' \qquad (15) \\
&\geq -E_X^\kappa.
\end{aligned}
$$

where (15) follows by Jensen's inequality. □

Proof of Property 3. The first equality directly follows from the definition of kernelized divergence (Definition 2) and mutual information (Definition 3). The positiveness of $I_1(X;Y)$ follows by setting $n \to \infty$ in Proposition 4.1 of [Giraldo *et al.*, 2014]. □

Proof of Property 4. Notice that

$$
\begin{aligned}
S_1(X) &= -C_{\kappa_X} \iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}') \kappa_X^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x} \,\mathrm{d}\mathbf{x}' \\
&= -C_{\kappa_X} \iiint_{\mathcal{Y} \times \mathcal{X}^2} p_{X,Y}(\mathbf{x}, \mathbf{y}) \log p_X(\mathbf{x}') \kappa_X^2(\mathbf{x}, \mathbf{x}') \,\mathrm{d}\mathbf{x} \,\mathrm{d}\mathbf{x}' \,\mathrm{d}\mathbf{y} \\
&= -C_{\kappa_X} C_{\kappa_Y} \iiiint_{\mathcal{Y}^2 \times \mathcal{X}^2} p_{X,Y}(\mathbf{x}, \mathbf{y}) \log p_X(\mathbf{x}') \kappa_X^2(\mathbf{x}, \mathbf{x}') \kappa_Y^2(\mathbf{y}, \mathbf{y}') \,\mathrm{d}\mathbf{x} \,\mathrm{d}\mathbf{x}' \,\mathrm{d}\mathbf{y} \,\mathrm{d}\mathbf{y}'.
\end{aligned}
$$

Similarly, we have $S_1(Y) = -C_{\kappa_X} C_{\kappa_Y} \iiiint_{\mathcal{Y}^2 \times \mathcal{X}^2} p_{X,Y}(\mathbf{x}, \mathbf{y}) \log p_Y(\mathbf{y}') \kappa_X^2(\mathbf{x}, \mathbf{x}') \kappa_Y^2(\mathbf{y}, \mathbf{y}') \,\mathrm{d}\mathbf{x} \,\mathrm{d}\mathbf{x}' \,\mathrm{d}\mathbf{y} \,\mathrm{d}\mathbf{y}'$. Combining the expressions above finishes the proof. □

Proof of Property 5. Following the definition of kernelized mutual information, we can derive the expression of kernelized conditional mutual information as follows:

$$I_1(X;Y|Z) = C_{\kappa_X} C_{\kappa_Y} C_{\kappa_Z} \iint_{\mathcal{Z}^2} \iint_{\mathcal{Y}^2} \iint_{\mathcal{X}^2} p_{X,Y,Z}(x,y,z) \log \frac{p_{X,Y|Z}(x',y'|z')}{p_{X|Z}(x'|z') p_{Y|Z}(y'|z')} \tag{16}$$

$$\cdot \kappa_X^2(x,x') \kappa_Y^2(y,y') \kappa_Z^2(z,z') \,\mathrm{d}x \,\mathrm{d}x' \,\mathrm{d}y \,\mathrm{d}y' \,\mathrm{d}z \,\mathrm{d}z'. \tag{17}$$

Then the proof of property 5 directly follows by the definition of conditional and unconditional kernelized mutual information. □

Proof of Property 6. Following property 5, we have that

$$I_1(X;Y,Z) = I_1(X;Y|Z) + I_1(X;Z) = I_1(X;Z|Y) + I_1(X;Y).$$

From the Markov condition, X and Z are conditionally independent given Y, i.e. $p_{X,Z|Y}(x,z|y) = p_{X|Y}(x|y) p_{Z|Y}(z|y)$. By the definition of conditional mutual information, we have $I_1(X;Z|Y) = 0$. Then the first inequality of property 6 follows by the positiveness of $I_1(X;Y|Z)$. Similarly, one can prove the second part of property 6. □

B Proof of Section 4

B.1 Proof of Theorem 1

Lemma 4. *Let P, Q be probability measures defined on the same measurable space, where P is absolutely continuous with respect to Q. Then*

$$D_1(P \parallel Q) + E_P^\kappa \geq \mathbb{E}_P[X] - \log \mathbb{E}_Q[e^X].$$

where X is any random variable such that e^X is Q-integrable and $\mathbb{E}_P[X]$ exists.

Proof. Define Q^X be a probability measure such that

$$Q^X(\Omega) = \int_\Omega \frac{e^X}{\mathbb{E}_Q[e^X]} \,\mathrm{d}Q,$$

then Q is absolutely continuous with respect to Q^X. Observe that

$$\begin{aligned}
D_1(P \parallel Q) + E_P^\kappa &= D_1(P \parallel Q^X) + E_P^\kappa + C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log \frac{e^{x'}}{\mathbb{E}_Q[e^X]} \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' \\
&\geq C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log e^{x'} \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' - C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log \mathbb{E}_Q[e^X] \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' \\
&= \int_{\mathcal{X}} p_X(x) \left(C_\kappa \int_{\mathcal{X}} x' \kappa^2(x,x') \,\mathrm{d}x' \right) \mathrm{d}x - \log \mathbb{E}_Q[e^X] \int_{\mathcal{X}} p_X(x) \left(C_\kappa \int_{\mathcal{X}} \kappa^2(x,x') \,\mathrm{d}x' \right) \mathrm{d}x \\
&= \int_{\mathcal{X}} p_X(x) x \,\mathrm{d}x - \log \mathbb{E}_Q[e^X] \int_{\mathcal{X}} p_X(x) \,\mathrm{d}x = \mathbb{E}_P[X] - \log \mathbb{E}_Q[e^X].
\end{aligned}$$

□

Lemma 5. *(Lemma 2 in [Harutyunyan* et al.*, 2021]) Let X be a zero-mean random variable that is R-subgaussian, then $\forall \lambda \in \left[0, \frac{1}{4R^2}\right)$:*

$$\mathbb{E}\left[e^{\lambda X^2}\right] \leq 1 + 8\lambda R^2.$$

Lemma 6. *(Lemma 3 in [Harutyunyan* et al.*, 2021]) Let X and Y be independent random variables. Let f be a measurable function such that $f(x,Y)$ is R-subgaussian and $\mathbb{E}_Y[f(x,Y)] = 0$ for all $x \in \mathcal{X}$, then $f(X,Y)$ is also R-subgaussian.*

Theorem 1 (Restate)**.** *Suppose that $\ell(w,Z)$ is R-subgaussian with respect to Z for every $w \in \mathcal{W}$, then*

$$|\mathbb{E}_{S,W}[L(W) - L_S(W)]| \leq \sqrt{\frac{2R^2 \hat{I}_1(S;W)}{n}}, \quad \textit{and}$$

$$\mathbb{E}_{S,W}[L(W) - L_S(W)]^2 \leq \frac{4R^2(\hat{I}_1(S;W) + \log 3)}{n},$$

where $\hat{I}_1(S;W) = I_1(S;W) + E_{S,W}^\kappa$.

Proof. Let $f(w,s) = L(w) - L_s(w)$ and let W' and S' be independent copies of W and S and $\lambda \in [0,\infty)$, then

$$\begin{aligned} I_1(W;S) + E^{\kappa}_{W,S} &= D_1(P_{W,S} \parallel P_W \otimes P_S) + E^{\kappa}_{W,S} \\ &\geq \mathbb{E}_{W,S}[\lambda f(W,S)] - \log \mathbb{E}_{W',S'}\left[e^{\lambda f(W',S')}\right] \\ &= \mathbb{E}_{W,S}[\lambda f(W,S)] - \log \mathbb{E}_{W,S'}\left[e^{\lambda f(W,S')}\right] \end{aligned} \tag{18}$$

by Lemma 4 and the fact that W, W', S' are independent. Notice that $f(w,S)$ is $R/\sqrt{n}$-subgaussian for each $w \in \mathcal{W}$, since $L_S(w)$ is the average of n i.i.d. R-subgaussian random variables. Moreover, $\mathbb{E}_S[f(w,S)] = 0$ for any fixed w. Then by Lemma 6, $f(W,S)$ is $R/\sqrt{n}$-subgaussian. Therefore,

$$\log \mathbb{E}_{W,S'}\left[e^{\lambda f(W,S') - \lambda \mathbb{E}_{W,S'}[f(W,S')]}\right] \leq \frac{\lambda^2 R^2}{2n},$$
$$\log \mathbb{E}_{W,S'}\left[e^{\lambda f(W,S')}\right] \leq \frac{\lambda^2 R^2}{2n}.$$

Plugging into (18), we have

$$I_1(W;S) + E^{\kappa}_{W,S} \geq \lambda \mathbb{E}_{W,S}[f(W,S)] - \frac{\lambda^2 R^2}{2} \geq \frac{n}{2R^2}\mathbb{E}^2_{W,S}[f(W,S)].$$

This finishes the proof of the first part. For the second part, let $\tilde{f}(w,s) = (L(w) - L_s(w))^2$ and $\lambda \in \left[0, \frac{1}{4R^2}\right)$, then

$$\begin{aligned} I_1(W;S) + E^{\kappa}_{W,S} &\geq \mathbb{E}_{W,S}\left[\lambda \tilde{f}(W,S)\right] - \log \mathbb{E}_{W',S'}\left[e^{\lambda \tilde{f}(W',S')}\right] \\ &\geq \mathbb{E}_{W,S}\left[\lambda (L(W) - L_S(W))^2\right] - \log(1 + 8\lambda R^2) \\ &\geq \frac{1}{4R^2}\mathbb{E}_{W,S}\left[(L(W) - L_S(W))^2\right] - \log 3, \end{aligned}$$

by Lemma 4, 5 and taking $\lambda = \frac{1}{4R^2}$. This finishes the proof of the second part. □

B.2 Proof of Theorem 2

Lemma 7. *Let random variable $X \sim N(0,\Sigma)$ and X' be any continuous random variable that satisfies $\mathbb{E}[X'] = 0$ and $\mathrm{Cov}[X'] = \Sigma$, then $S_1(X') \leq S_1(X) + E^{\kappa}_{X'}$. Furthermore if κ is Gaussian with kernel width σ_κ, then $S_1(X) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log|\Sigma| + \frac{\sigma_\kappa^2}{4}\mathrm{tr}[\Sigma^{-1}]$.*

Proof. Let $p(\cdot)$ and $q(\cdot)$ be the PDF of X and X' respectively. Notice that $p_\kappa = C_\kappa \kappa^2(0,\cdot)$ integrates to 1, thus could be treated as a probability distribution whose covariance matrix is $\frac{1}{2}\sigma_\kappa^2 I$. Then we have

$$\begin{aligned} S_1(X) - S_1(X') &= C_\kappa \iint_{\mathcal{X}^2} \left[q(\mathbf{x})\left(\log p(\mathbf{x}') + \log \frac{q(\mathbf{x}')}{p(\mathbf{x}')}\right) - p(\mathbf{x}) \log p(\mathbf{x}')\right] \kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}' \\ &= C_\kappa \iint_{\mathcal{X}^2} [q(\mathbf{x}) - p(\mathbf{x})] \log p(\mathbf{x}') \kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}' + C_\kappa \iint_{\mathcal{X}^2} q(\mathbf{x}) \log \frac{q(\mathbf{x}')}{p(\mathbf{x}')} \kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}' \\ &= C_\kappa \iint_{\mathcal{X}^2} [q(\mathbf{x}) - p(\mathbf{x})]\left(-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathbf{x}'^{\top}\Sigma^{-1}\mathbf{x}'\right)\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}' + D_1(Q \parallel P) \\ &\geq -\left(\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma|\right)\int_{\mathcal{X}} [q(\mathbf{x}) - p(\mathbf{x})]\left(C_\kappa \int_{\mathcal{X}} \kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x} \\ &\quad - \frac{1}{2}\mathrm{tr}\left\{\left[\int_{\mathcal{X}} [q(\mathbf{x}) - p(\mathbf{x})]\left(C_\kappa \int_{\mathcal{X}} \mathbf{x}'\mathbf{x}'^{\top}\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x}\right]\Sigma^{-1}\right\} - E^{\kappa}_{X'} \\ &= -\left(\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma|\right)\int_{\mathcal{X}} [q(\mathbf{x}) - p(\mathbf{x})]\,\mathrm{d}\mathbf{x} \\ &\quad - \frac{1}{2}\mathrm{tr}\left\{\left[\int_{\mathcal{X}} [q(\mathbf{x}) - p(\mathbf{x})]\left(\mathbf{x}\mathbf{x}^{\top} + \frac{1}{2}\sigma_\kappa^2 I\right)\mathrm{d}\mathbf{x}\right]\Sigma^{-1}\right\} - E^{\kappa}_{X'} \\ &= -\frac{1}{2}\mathrm{tr}\left[\left(\int_{\mathcal{X}} [q(\mathbf{x}) - p(\mathbf{x})]\mathbf{x}\mathbf{x}^{\top}\,\mathrm{d}\mathbf{x}\right)\Sigma^{-1}\right] - E^{\kappa}_{X'} = -\frac{1}{2}\mathrm{tr}\left[(\Sigma - \Sigma)\Sigma^{-1}\right] - E^{\kappa}_{X'} = -E^{\kappa}_{X'}. \end{aligned}$$

Consider the case that kernel κ is Gaussian, i.e. $\kappa^2(\mathbf{x},\mathbf{x}') = \exp\big(-\|\mathbf{x}-\mathbf{x}'\|_2^2/\sigma_\kappa^2\big)$ and $C_\kappa = (\pi\sigma_\kappa^2)^{-d/2}$, we have

$$\begin{aligned}
S_1(X) &= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\frac{1}{\sqrt{(\pi\sigma_\kappa^2)^d}}\iint_{\mathcal{X}^2}\exp\left(-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right)\left[\frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|+\frac{1}{2}\mathbf{x}'^\top\Sigma^{-1}\mathbf{x}'\right]\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma_\kappa^2}\right)\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}' \\
&= \frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|+\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{\mathcal{X}}\frac{1}{2}\exp\left(-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right)\left(\frac{1}{\sqrt{(\pi\sigma_\kappa^2)^d}}\int_{\mathcal{X}}\mathbf{x}'^\top\Sigma^{-1}\mathbf{x}'\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma_\kappa^2}\right)\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x} \\
&= \frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|+\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{\mathcal{X}}\frac{1}{2}\exp\left(-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right)\mathrm{tr}\left[\left(\mathbf{x}\mathbf{x}^\top+\frac{1}{2}\sigma_\kappa^2 I\right)\Sigma^{-1}\right]\mathrm{d}\mathbf{x} \\
&= \frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|+\frac{1}{2}\mathrm{tr}\left[\left(\Sigma+\frac{1}{2}\sigma_\kappa^2 I\right)\Sigma^{-1}\right] = \frac{d}{2}\log(2\pi e)+\frac{1}{2}\log|\Sigma|+\frac{\sigma_\kappa^2}{4}\mathrm{tr}[\Sigma^{-1}].
\end{aligned}$$

□

Lemma 8. *Let X, Y, Δ and $\xi \sim N(0,\sigma^2 I)$ be independent random variables. Let $f\colon \mathcal{W}\times\mathcal{Z}^b \to \mathcal{W}$ be a determinant function and let $\Omega(\cdot) = \mathbb{E}_X[f(\cdot,X)]$. Then*

$$I_1(f(Y+\Delta,X)+\xi;X|Y) \le \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)]+I\right| + E^\kappa_{f(Y+\Delta,X)-\Omega(Y+\Delta)+\xi|Y,\Delta}.$$

Proof. Let $g(Y,\Delta,X) = f(Y+\Delta,X)-\Omega(Y+\Delta)$, then

$$\begin{aligned}
&I_1(f(Y+\Delta,X)+\xi;X|Y=\mathbf{y},\Delta=\delta) && (19)\\
&= I_1(g(Y,\Delta,X)+\xi;X|Y=\mathbf{y},\Delta=\delta) \\
&= S_1(g(Y,\Delta,X)+\xi|Y=\mathbf{y},\Delta=\delta) - S_1(g(Y,\Delta,X)+\xi|X,Y=\mathbf{y},\Delta=\delta) \\
&= S_1(g(Y,\Delta,X)+\xi|Y=\mathbf{y},\Delta=\delta) - S_1(\xi) \\
&= S_1(g(Y,\Delta,X)+\xi|Y=\mathbf{y},\Delta=\delta) - \frac{d}{2}\log(2\pi e\sigma^2) - \frac{d\sigma_\kappa^2}{4\sigma^2} \\
&\le \frac{d}{2}\log(2\pi e)+\frac{1}{2}\log\big|\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\sigma^2 I\big| - \frac{d}{2}\log(2\pi e\sigma^2) + E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta} \\
&\quad + \frac{\sigma_\kappa^2}{4}\mathrm{tr}\Big[\big(\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\sigma^2 I\big)^{-1}\Big] - \frac{d\sigma_\kappa^2}{4\sigma^2} && (20)\\
&\le \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+I\right| + E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta} && (21)
\end{aligned}$$

where (20) follows by Lemma 7 and noticing that

$$\begin{aligned}
\mathrm{Cov}[g(Y,\Delta,X)+\xi|Y=\mathbf{y},\Delta=\delta] &= \mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\mathrm{Cov}[\xi] \\
&= \mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\sigma^2 I,
\end{aligned}$$

and (21) follows by the fact that

$$\mathrm{tr}\Big[\big(\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\sigma^2 I\big)^{-1}\Big] \le \mathrm{tr}\Big[\big(\sigma^2 I\big)^{-1}\Big].$$

This leads to the following bound:

$$\begin{aligned}
I_1(f(Y+\Delta,X)+\xi;X|Y) &\le I_1(f(Y+\Delta,X)+\xi,\Delta;X|Y) && (22)\\
&= I_1(f(Y+\Delta,X)+\xi,\Delta;X|Y) - I_1(\Delta;X|Y) && (23)\\
&= I_1(f(Y+\Delta,X)+\xi;X|Y,\Delta) && (24)\\
&= \mathbb{E}_{Y,\Delta}[I_1(f(Y+\Delta,X)+\xi;X|Y=\mathbf{y},\Delta=\delta)] \\
&\le \mathbb{E}_{Y,\Delta}\left[\frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+I\right|\right] + E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta} && (25)\\
&\le \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathbb{E}_{Y,\Delta}[\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]]+I\right| + E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta} && (26)\\
&= \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)]+I\right| + E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta}, && (27)
\end{aligned}$$

where (23) follows by the fact that Δ and X are independent, (25) follows by applying (21), (26) follows by Jensen's inequality and the concavity of the log-determinant function, and (27) follows by the law of total variance and noticing that

$$\mathrm{Cov}[\mathbb{E}_X[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]] = \mathrm{Cov}[\mathbb{E}_X[f(Y,\Delta,X)-\mathbb{E}_X[f(Y,\Delta,X)]|Y=\mathbf{y},\Delta=\delta]] = 0.$$

□

Theorem 2 (Restate)**.** *Under the same conditions of Lemma 1:*

$$I_1(W_T;S) \le \sum_{t=1}^{T} I_1(W_t;B_t|W_{t-1}) \le \sum_{t=1}^{T}\left(\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right| + E^{\kappa}_{W_t|W_{t-1}}\right),$$

$$I(W_T;S) \le \sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right|,$$

where $\mathbb{V}_t = \mathrm{Cov}[g(W_{t-1},B_t)]$ is the ***gradient covariance*** *matrix and $|\cdot|$ denotes the matrix determinant.*

Proof. Notice that $S \to (B_1,\cdots,B_T) \to (W_1,\cdots,W_T)$ form a Markov chain, then

$$\begin{aligned} I_1(W_T;S) &\le I_1(W_T;B_1,\cdots,B_T) \le I_1(W_0,W_1,\cdots,W_T;B_1,\cdots,B_T) \\ &= I_1(W_0;B_1,\cdots,B_T) + I_1(W_1;B_1,\cdots,B_T|W_0) + I_1(W_2;B_1,\cdots,B_T|W_0,W_1) \\ &\quad + \cdots + I_1(W_T;B_1,\cdots,B_T|W_0,\cdots,W_T). \end{aligned}$$

For each $t\in[1,T]$, we have

$$\begin{aligned} I_1(W_t;B_1,\cdots,B_t|W_0,\cdots,W_{t-1}) &= S_1(W_t|W_0,\cdots,W_{t-1}) - S_1(W_t|B_1,\cdots,B_t,W_0,\cdots,W_{t-1}) \\ &= S_1(W_t|W_{t-1}) - S_1(W_t|B_t,W_{t-1}) \\ &= I_1(W_t;B_t|W_{t-1}). \end{aligned}$$

Let $X=B_t$, $Y=W_{t-1}$, $\Delta=0$, $\xi=\xi_t$ and $f(W_{t-1},B_t) = -\eta_t g(W_{t-1},B_t)$ in Lemma 8, we then have

$$\begin{aligned} I_1(W_t;B_t|W_{t-1}) &= I_1(W_t-W_{t-1};B_t|W_{t-1}) = I_1(-\eta_t g(W_{t-1},B_t)+\xi_t;B_t|W_{t-1}) \\ &\le \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathrm{Cov}[g(W_{t-1},B_t)] + I\right| + E^{\kappa}_{W_t|W_{t-1}}, \end{aligned}$$

which finishes the proof. □

B.3 Proof of Proposition 5

Lemma 9. *Let V be an $n\times n$ symmetric positive definite matrix partitioned by*

$$V = \begin{bmatrix} A & C^\top \\ C & B \end{bmatrix},$$

where A, B are symmetric matrices of size $n_1\times n_1$ and $n_2\times n_2$ respectively. Then $|V|\le|A||B|$.

Proof. Notice that

$$V = D\begin{bmatrix} A & 0 \\ 0 & B-CA^{-1}C^\top \end{bmatrix}D^\top, \quad \text{where } D = \begin{bmatrix} I_{n_1} & 0 \\ CA^{-1} & I_{n_2} \end{bmatrix},$$

then we have $|V| = |D||A||B-CA^{-1}C^\top||D^\top| = |A||B-CA^{-1}C^\top|$. Let $D^\dagger$ be the pseudo-inverse of D, then for any column vector $\mathbf{x}$ of length n:

$$\mathbf{x}^\top\begin{bmatrix} A & 0 \\ 0 & B-CA^{-1}C^\top \end{bmatrix}\mathbf{x} = (\mathbf{x}^\top D^\dagger)V(\mathbf{x}^\top D^\dagger)^\top \ge 0,$$

therefore $B-CA^{-1}C^\top$ is positive semi-definite. Similarly, we can prove that $CA^{-1}C^\top$ is positive semi-definite. Let λ_i, μ_i, ν_i, $i\in\{1,\cdots,n_2\}$ be the eigenvalues of $B-CA^{-1}C^\top$, B and $CA^{-1}C^\top$ respectively in descending order, then by Weyl's inequality, we have $\lambda_i \le \mu_i - \nu_{n_2} \le \mu_i$ for all $i\in\{1,\cdots,n_2\}$, which implies that $|B-CA^{-1}C^\top| \le |B|$. The proof is complete by combining the result above: $|V| = |A||B-CA^{-1}C^\top| \le |A||B|$. □

Proposition 5 (Restate)**.** *Given $\mathbb{V}_t$, V_t and L defined as above, let $\{c_i\}_{i=1}^r$ be a partition of $\{n\}$, i.e. $c_1\cup\cdots\cup c_r = \{n\}$ and $c_i\cap c_j = \Phi$ for any $1\le i<j\le r$. Let $\mathbb{V}_t^i$ be the sub-matrix of $\mathbb{V}_t$ with columns and rows indexed by c_i, and define*

$$\theta_c(\mathbb{V}) = \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}+I\right|, \quad \theta_v(V) = \frac{d}{2}\log\left(\frac{\eta_t^2 V}{d\sigma_t^2}+1\right),$$

$$\textit{then} \qquad \theta_c(\mathbb{V}_t) \le \sum_{i=1}^{r}\theta_c(\mathbb{V}_t^i) \le \theta_v(V_t) \le \theta_v(L).$$

Table 1: Hyper-parameters used to train deep learning models.

Hyper-parameter	Synthetic	MNIST	CIFAR10
learning rate (η)	0.001	0.01	0.01
size of training dataset (n)	100	5000	5000
epochs	50	100	100
batch size	10	50	50
steps (T)	500	10000	10000
variance of the noise (σ_t^2)	10^{-3}	10^{-5}	10^{-5}

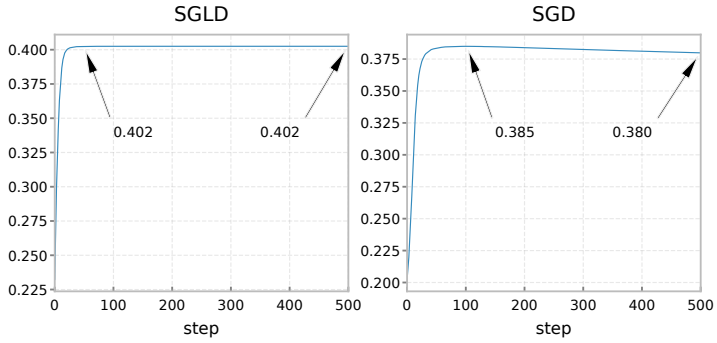

Figure 5: Behavior of $I_1(W_T; S)$ on synthetic data.

Proof. Notice that $V_t = \mathrm{tr}[\mathbb{V}_t]$. Since the covariance matrix is always symmetric positive semi-definite, we can denote the eigenvalues of $\mathbb{V}_t$ by $\lambda_1, \cdots, \lambda_d \geq 0$, then

$$\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right| = \log\left[\prod_{i=1}^{d}\left(\frac{\eta_t^2\lambda_i}{\sigma_t^2}+1\right)\right] \leq \log\left[\frac{1}{d}\sum_{i=1}^{d}\left(\frac{\eta_t^2\lambda_i}{\sigma_t^2}+1\right)\right]^d = d\log\left[\frac{\eta_t^2}{d\sigma_t^2}\sum_{i=1}^{d}\lambda_i + 1\right] = d\log\left[\frac{\eta_t^2 V_t}{d\sigma_t^2}+1\right],$$

where the only inequality follows by the fact that the geometric mean is always less than the arithmetic mean. Let $V_t^i = \mathrm{tr}[\mathbb{V}_t^i]$, then through the same strategy, one can prove that for all $i \in \{1, \cdots, r\}$:

$$\theta_c(\mathbb{V}_t^i) \leq \theta_v(V_t^i),$$

then by Jensen's inequality, we have:

$$\sum_{i=1}^{r}\theta_c(\mathbb{V}_t^i) \leq \sum_{i=1}^{r}\theta_v(V_t^i) = \sum_{i=1}^{r} d\log\left[\frac{\eta_t^2 V_t^i}{d\sigma_t^2}+1\right] \leq d\log\left[\frac{\eta_t^2}{d\sigma_t^2}\sum_{i=1}^{r}V_t^i + 1\right] = d\log\left[\frac{\eta_t^2}{d\sigma_t^2}V_t + 1\right] = \theta_v(V_t).$$

Next, by applying Lemma 9 recursively, we can prove that

$$\theta_c(\mathbb{V}_t) = \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right| \leq \frac{1}{2}\log\prod_{i=1}^{r}\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t^i + I\right| = \frac{1}{2}\sum_{i=1}^{r}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t^i + I\right| = \sum_{i=1}^{r}\theta_c(\mathbb{V}_t^i).$$

To prove the last inequality of Proposition 5, notice that

$$\begin{aligned} V_t &= \mathbb{E}_{B_t}[\|g(W_{t-1}, B_t) - \mathbb{E}_{B_t}[g(W_{t-1}, B_t)]\|_2^2] = \mathbb{E}_{B_t}[\|g(W_{t-1}, B_t)\|_2^2] - \|\mathbb{E}_{B_t}[g(W_{t-1}, B_t)]\|_2^2 \\ &\leq \mathbb{E}_{B_t}[\|g(W_{t-1}, B_t)\|_2^2] \leq \max_{w\in\mathcal{W}, z\in\mathcal{Z}}\|g(w,z)\|_2^2 = L, \end{aligned}$$

which finishes the proof by the monotonicity of the log function. □

B.4 Proof of Theorem 3

Lemma 10. *The mutual information* $I_1(\tilde{W}_T; S) \leq \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1}, B_t) + \tilde{\xi}_t; B_t|\tilde{W}_{t-1})$.

Proof.

$$I_1(\tilde{W}_T; S) = I_1(\tilde{W}_{T-1} - \eta_T g(W_{T-1}, B_T) + \tilde{\xi}_T; S)$$

$$
\begin{aligned}
&\leq I_1(\tilde{W}_{T-1}, -\eta_T g(W_{T-1}, B_T) + \tilde{\xi}_T; S) && (28)\\
&= I_1(\tilde{W}_{T-1}; S) + I_1(-\eta_T g(W_{T-1}, B_T) + \tilde{\xi}_T; S|\tilde{W}_{T-1}) && (29)\\
&\leq I_1(\tilde{W}_{T-2}; S) + I_1(-\eta_{T-1} g(W_{T-2}, B_{T-1}) + \tilde{\xi}_{T-1}; S|\tilde{W}_{T-2})\\
&\quad + I_1(-\eta_T g(W_{T-1}, B_T) + \tilde{\xi}_T; S|\tilde{W}_{T-1})\\
&\leq \cdots\\
&\leq I_1(\tilde{W}_0; S) + \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1}, B_t) + \tilde{\xi}_t; S|\tilde{W}_{t-1}) && (30)\\
&= \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1}, B_t) + \tilde{\xi}_t; S|\tilde{W}_{t-1}) && (31)\\
&\leq \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1}, B_t) + \tilde{\xi}_t; B_t|\tilde{W}_{t-1}) && (32)\\
&= \sum_{t=1}^{T} I_1(\tilde{W}_t - \tilde{W}_{t-1}; B_t|\tilde{W}_{t-1}) = \sum_{t=1}^{T} I_1(\tilde{W}_t; B_t|\tilde{W}_{t-1}),
\end{aligned}
$$

where (28) follows by noticing that $Z \to (X, Y) \to f(X, Y)$ forms a Markov chain and then apply property 6 in Proposition 4, (29) follows by property 5 in Proposition 4, (30) follows by repeating the steps above (29), (31) follows by noticing that $\tilde{W}_0$ and S are independent, and (32) follows by the fact that $S \to B_t \to -\eta_t g(w, B_t) + \tilde{\xi}_t | w = \tilde{W}_{t-1}$ form a Markov chain. □

Theorem 3 (Restate)**.** *Under the same conditions of Lemma 3:*

$$
I_1\Big(\tilde{W}_T; S\Big) \leq \sum_{t=1}^{T}\bigg(\frac{1}{2}\log\bigg|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\bigg| + E^{\kappa}_{\tilde{W}_t|\tilde{W}_{t-1}}\bigg),
$$

$$
I\Big(\tilde{W}_T; S\Big) \leq \sum_{t=1}^{T}\frac{1}{2}\log\bigg|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\bigg|.
$$

Proof. Applying Lemma 10, we have

$$
I_1(\tilde{W}_T; S) \leq \sum_{t=1}^{T}\bigg(\frac{1}{2}\log\bigg|\frac{\eta_t^2}{\sigma_t^2}\mathrm{Cov}[g(W_{t-1}, B_t)] + I\bigg| + E^{\kappa}_{\tilde{W}_t|\tilde{W}_{t-1}}\bigg),
$$

by applying Lemma 8 with $X = B_t$, $Y = \tilde{W}_{t-1}$, $\Delta = -\Delta_{t-1}$, $\xi = \tilde{\xi}_t$ and $f(W_{t-1}, B_t) = -\eta_t g(W_{t-1}, B_t)$. □

C Experiment Details

Deep learning models are trained with an Intel Xeon CPU (2.10GHz, 48 cores), 256GB memory, and 4 Nvidia Tesla V100 GPU (32GB). For the MNIST data set, we train an MLP with one hidden layer of size 128. For the CIFAR10 dataset, we train a CNN with 4 convolution layers (32, 32, 48, 48) of size 3×3 followed by two fully connected layers of size 48. All of the layers above use ReLU as the activation function. The hyper-parameters used for training are listed in Table 1.

As can be seen in Figure 5, the curves of IWS soon stop increasing after several epochs and start to decrease. This is consistent with the behavior of the true generalization error, supporting our claim that the estimate of $I_1(W_T; S)$ successfully reflects the behavior of $I(W_T; S)$.

In figure 6, we conduct the random label experiment on CIFAR10, and in Figure 7, we test the generalization behavior of MLP on the MNIST dataset with different dimensionality of the hidden layer. It can be seen that our bound is still 5 more times tighter than the bound of Lemma 3.

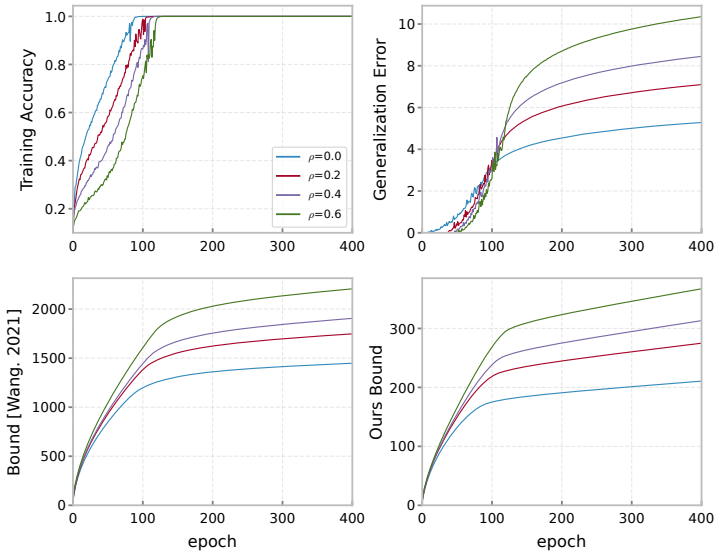

Figure 6: Random label experiment on CIFAR10.

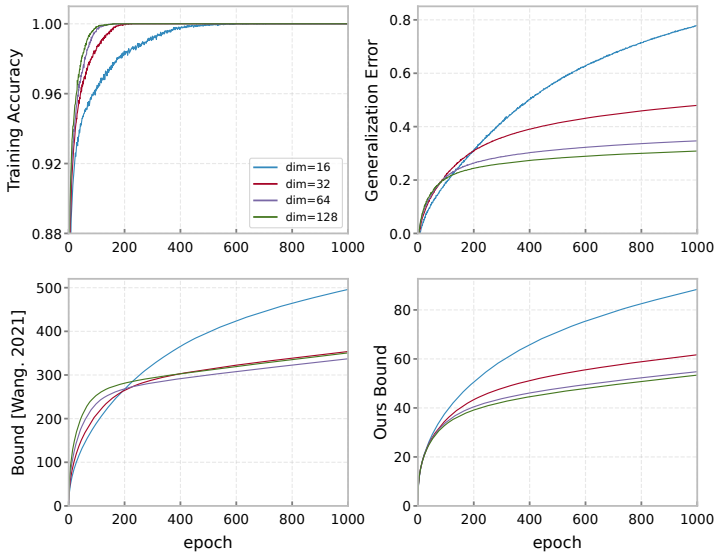

Figure 7: Generalization of MLP on MNIST with different dimensionality of the hidden layer.