Journal of Machine Learning Research 24 (2023) 1-43 Submitted 11/21; Revised 8/22; Published 1/23

Generalization Bounds for Noisy Iterative Algorithms Using Properties of Additive Noise Channels

Hao Wang HAO_WANG@G.HARVARD.EDU
Harvard University

Rui Gao RUI.GAO@MCCOMBS.UTEXAS.EDU
The University of Texas at Austin

Flavio P. Calmon FLAVIO@SEAS.HARVARD.EDU
Harvard University

Editor: Gabor Lugosi

Abstract

Machine learning models trained by different optimization algorithms under different data distributions can exhibit distinct generalization behaviors. In this paper, we analyze the generalization of models trained by noisy iterative algorithms. We derive distribution-dependent generalization bounds by connecting noisy iterative algorithms to additive noise channels found in communication and information theory. Our generalization bounds shed light on several applications, including differentially private stochastic gradient descent (DP-SGD), federated learning, and stochastic gradient Langevin dynamics (SGLD). We demonstrate our bounds through numerical experiments, showing that they can help understand recent empirical observations of the generalization phenomena of neural networks.

Keywords: Information theory, algorithmic generalization bound, differential privacy, stochastic gradient Langevin dynamics, federated learning.

1. Introduction

Many learning algorithms aim to solve the following (possibly non-convex) optimization problem:

$$\min_{\boldsymbol{w}\in\mathcal{W}} L_\mu(\boldsymbol{w}), \quad \text{where } L_\mu(\boldsymbol{w}) \triangleq \mathbb{E}\left[\ell(\boldsymbol{w}, \mathrm{Z})\right] = \int_{\mathcal{Z}} \ell(\boldsymbol{w}, \boldsymbol{z}) \mathrm{d}\mu(\boldsymbol{z}), \tag{1}$$

where $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbb{R}^d$ is the model parameter (e.g., weights of a neural network) to optimize; μ is the underlying data distribution that generates Z; and $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ is the loss function (e.g., 0-1 loss). In the context of supervised learning, Z is often composed by a feature vector X and its corresponding label Y. Since the data distribution μ is unknown, $L_\mu(\boldsymbol{w})$ cannot be computed directly. In practice, a data set $\mathrm{S} \triangleq (\mathrm{Z}_1, \cdots, \mathrm{Z}_n)$ containing n i.i.d. points $\mathrm{Z}_i \sim \mu$ is used to minimize an empirical risk:

$$\min_{\boldsymbol{w}\in\mathcal{W}} L_{\mathrm{S}}(\boldsymbol{w}), \quad \text{where } L_{\mathrm{S}}(\boldsymbol{w}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{w}, \mathrm{Z}_i). \tag{2}$$

We consider the following (projected) noisy iterative algorithm for solving the empirical risk optimization in (2). The parameter $\boldsymbol{w}$ is initialized with a random point $\mathrm{W}_0 \in \mathcal{W}$ and

updated using the following rule:

$$\mathrm{W}_t = \mathsf{Proj}_{\mathcal{W}}\left(\mathrm{W}_{t-1} - \eta_t \cdot g(\mathrm{W}_{t-1}, \{\mathrm{Z}_i\}_{i\in\mathcal{B}_t}) + m_t \cdot \mathrm{N}_t\right), \tag{3}$$

where η_t is the learning rate; N_t is an additive noise drawn independently from a distribution P_{N}; m_t is the magnitude of the noise; $\mathcal{B}_t \subseteq [n]$ contains the indices of the data points used at the current iteration and $b_t \triangleq |\mathcal{B}_t|$; g is the direction for updating the parameter (e.g., gradient of the loss function); and

$$g(\mathrm{W}_{t-1}, \{\mathrm{Z}_i\}_{i\in\mathcal{B}_t}) \triangleq \frac{1}{b_t}\sum_{i\in\mathcal{B}_t} g(\mathrm{W}_{t-1}, \mathrm{Z}_i). \tag{4}$$

At the end of each iteration, the parameter is projected onto the domain $\mathcal{W}$, i.e., $\mathsf{Proj}_{\mathcal{W}}(\boldsymbol{w}) \triangleq \operatorname{argmin}_{\boldsymbol{w}'\in\mathcal{W}} \|\boldsymbol{w}' - \boldsymbol{w}\|$. The recursion in (3) is run T iterations and the final output is a random variable W_T.

The goal of this paper is to provide an upper bound for the *expected generalization gap*:

$$\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_{\mathrm{S}}(\mathrm{W}_T)\right], \tag{5}$$

where the expectation is taken over the randomness of the training data set S and of the noisy iterative algorithm.

Noisy iterative algorithms are used in different practical settings due to their many attractive properties (see e.g., Li et al., 2016; Zhang et al., 2017b; Raginsky et al., 2017; Xu et al., 2018). For example, differentially private SGD (DP-SGD) algorithm (see e.g., Song et al., 2013; Abadi et al., 2016), one kind of noisy iterative algorithm, is often used to train machine learning models while protecting user privacy (Dwork et al., 2006). Recently, it has been implemented in open-source libraries, including Opacus (Facebook AI, 2020) and TensorFlow Privacy (Radebaugh and Erlingsson, 2019). The additive noise in iterative algorithms may also mitigate overfitting for deep neural networks (DNNs) (Neelakantan et al., 2015). From a theoretical perspective, noisy iterative algorithms can escape local minima (Kleinberg et al., 2018) or saddle points (Ge et al., 2015) and generalize well (Pensia et al., 2018).

We derive generalization bounds for noisy iterative algorithms. Although these bounds may be vacuous, when calculated numerically, they maintain a high correlation with the generalization gap. As a result, they shed light on many empirical observations of neural networks that are not explained by uniform notions of hypothesis class complexity (Vapnik and Chervonenkis, 1971; Valiant, 1984). For example, a neural network trained using true labels exhibits better generalization ability than a network trained using corrupted labels even when the network architecture is fixed and perfect training accuracy is achieved (Zhang et al., 2017a). Distribution-independent bounds is unable to capture this phenomenon because they are invariant to both true data and corrupted data.[1] In contrast, our bounds capture this empirical observation, exhibiting a lower value on networks trained on true

1. Note that there are generalization bounds (see e.g., Bartlett, 1998; Bartlett et al., 2017; Koltchinskii and Panchenko, 2002) that implicitly depend on data distribution through, e.g., margin and/or weight matrices' norm. Since different training data result in distinct weight matrices, these bounds may capture some generalization phenomena, such as label corruption.

labels compared to ones trained on corrupted labels (Figure 1). Another example is that a wider network often has a more favourable generalization capability (Neyshabur et al., 2015). This may seem counter-intuitive at first glance since one may expect that wider networks have a higher VC-dimension and, consequently, would have a higher generalize gap. Our bounds capture this behaviour and are decreasing with respect to the neural network width (Figure 2).

We present three generalization bounds for the noisy iterative algorithms (Section 4). These bounds rely on different kinds of f-divergence but are proved in a uniform manner by exploring properties of additive noise channels (Section 3). Among them, the KL-divergence bound can deal with sampling with replacement; the total variation bound is often the tightest one; and the χ^2-divergence bound requires the mildest assumption. We apply our results to applications, including DP-SGD, federated learning, and SGLD (Section 5). Under these applications, our generalization bounds can be significantly simplified and estimated from the training data. Finally, we demonstrate our bounds through numerical experiments (Section 6), showing that they can predict the behavior of the true generalization gap.

Our generalization bounds incorporate a time-decaying factor. This decay factor tightens the bounds by enabling the impact of early iterations to reduce with time. Our analysis is motivated by a line of recent works (Feldman et al., 2018; Balle et al., 2019; Asoodeh et al., 2020) which observed that data points used in the early iterations enjoy stronger differential privacy guarantees than those occurring late. Accordingly, we prove that if a data point is used at an early iteration, its contribution to our generalization bounds is decreasing with time due to the cumulative effect of the additive noise.

The proof techniques of this paper are based on fundamental tools from information theory. We first use an information-theoretic framework, proposed by Russo and Zou (2016) and Xu and Raginsky (2017) and further tightened by Bu et al. (2020), for deriving algorithmic generalization bounds. This framework relates the generalization gap in (5) with the f-information[2] $I_f(\mathrm{W}_T;\mathrm{Z}_i)$ between the algorithmic output W_T and each individual data point Z_i. However, estimating this f-information from data is intractable since the underlying distribution is unknown. Given this major challenge, we connect the noisy iterative algorithms with additive noise channels, a fundamental model used in data transmission. As a result, we further upper bound the f-information by a quantity that can be estimated from data by developing new properties of additive noise channels. Moreover, we incorporate a time-decaying factor into our bounds. This factor is established by strong data processing inequalities (Dobrushin, 1956; Cohen et al., 1998) and has an intuitive interpretation: the dependence between algorithmic output W_T and the data points used in the early iterations is decreasing with time due to external additive noise (i.e., $I_f(\mathrm{W}_T;\mathrm{Z}_i)$ is decreasing with T for a fixed Z_i).

1.1 Related Works

There are significant recent works which adopt the information-theoretic framework (Xu and Raginsky, 2017) for analyzing the generalization capability of noisy iterative algorithms. Among them, Pensia et al. (2018) initially derived a generalization bound in Corollary 1

2. The f-information (see (11) for its definition) includes a family of measures, such as mutual information, which quantify the dependence between two random variables.

and their bound was extended in Proposition 3 of Bu et al. (2020) for the SGLD algorithm. Although the framework in Pensia et al. (2018) can be applied to a broad class of noisy iterative algorithms, their bound in Corollary 1 and Proposition 3 in Bu et al. (2020) rely on the Lipschitz constant of the loss function, which makes them independent of the data distribution. Distribution-independent bounds can be potentially loose since the Lipschitz constant may be large and may not capture some empirical observations (e.g., label corruption (Zhang et al., 2017a)). Specifically, this Lipschitz constant only relies on the architecture of the network instead of the weight matrices or the data distribution so it is the same for a network trained from corrupted data and a network trained from true data.

To obtain a distribution-dependent bound, Negrea et al. (2019) improved the analysis in Pensia et al. (2018) by replacing the Lipschitz constant with a gradient prediction residual when analyzing the SGLD algorithm. Their follow-up work (Haghifam et al., 2020) investigated the Langevin dynamics algorithm (i.e., full batch SGLD), which was later extended by Rodríguez-Gálvez et al. (2021) to SGLD, and observed a time-decaying phenomenon in their experiments. Specifically, (Haghifam et al., 2020) incorporated a quantity, namely the squared error probability of the hypothesis test, into their bound in Theorem 4.2 and this quantity decays with the number of iterations. This seems to suggest that earlier iterations have a larger impact on their generalization bound. In contrast, our decay factor indicates that the impact of earlier iterations is reducing with the total number of iterations. Furthermore, the bound in their Theorem 4.2 requires a bounded loss function while our χ^2-based generalization bound only needs the variance of the loss function to be bounded. More broadly, Neu et al. (2021) investigated the generalization properties of SGD. However, the generalization bound in their Proposition 3 suffers from a weaker order $O(1/\sqrt{n})$ when the analysis is applied to the SGLD algorithm.

In addition to the works discussed above, there is a line of papers on deriving SGLD generalization bounds (Mou et al., 2018; Li et al., 2020) through other proof techniques. Among them, Mou et al. (2018) introduced two generalization bounds. The first one (Theorem 1 of Mou et al., 2018), a stability-based bound, achieves $O(1/n)$ rate in terms of the sample size n but relies on the Lipschitz constant of the loss function which makes it distribution-independent. The second one (Theorem 2 of Mou et al., 2018), a PAC-Bayes bound, replaces the Lipschitz constant by an expected-squared gradient norm but suffers from a slower rate $O(1/\sqrt{n})$. In contrast, our SGLD bound in Proposition 15 has order $O(1/n)$ and tightens the expected-squared gradient norm by the variance of gradients. The PAC-Bayes bound in Mou et al. (2018) also incorporates an explicit time-decaying factor. However, their analysis seems to heavily rely on the Gaussian noise. In contrast, our generalization bounds include a decay factor for a broad class of noisy iterative algorithms. A follow-up work by Li et al. (2020) combined the algorithmic stability approach with PAC-Bayesian theory and presented a bound which achieves order $O(1/n)$. However, their bound requires the scale of the learning rate to be upper bounded by the inverse Lipschitz constant of the loss function, which could result in negligible learning rates in practice. In contrast, we do not need any assumptions on the learning rate.

A standard approach (see e.g., He et al., 2021) of deriving a generalization bound for the DP-SGD algorithm follows two steps: (i) prove that DP-SGD satisfies the (ϵ, δ)-DP guarantees (Song et al., 2013; Wu et al., 2017; Feldman et al., 2018; Balle et al., 2019; Asoodeh et al., 2020); (ii) derive/apply a generalization bound that holds for *any* (ϵ, δ)-DP algorithm

(Dwork et al., 2015; Bassily et al., 2021; Jung et al., 2021). However, generalization bounds obtained from this procedure are distribution-independent since DP is robust with respect to the data distribution. In contrast, our bounds in Section 5.1 are distribution-dependent. We extend our analysis and derive a generalization bound in the setting of federated learning in Section 5.2. A previous work by Yagli et al. (2020) also proved a generalization bound for federated learning in their Theorem 3 but their bound involves a mutual information which could be hard to estimate from data.

The conference version (Wang et al., 2021) of this work investigates the generalization of the SGLD and DP-SGD algorithms. In this paper, we study a broader class of noisy iterative algorithms, including SGLD and DP-SGD as two examples. This extension requires a significant improvement of the proof techniques presented in Wang et al. (2021). Specifically, we introduce a unified framework for deriving generalization bounds through f-divergence while the analysis in the prior work (Wang et al., 2021) is tailored to the KL-divergence. In the context of the DP-SGD algorithm, we prove that the total variation distance leads to a tighter generalization bound and the χ^2-divergence leads to a bound requiring milder assumptions compared with the results in Wang et al. (2021). Finally, we derive a new generalization bound in the context of federated learning as an application of our framework.

2. Preliminaries

2.1 Notations

For a positive integer n, we define the set $[n] \triangleq \{1, \cdots, n\}$. We denote by $\|\cdot\|_1$ and $\|\cdot\|_2$ the 1-norm and 2-norm of a vector, respectively. A random variable X is σ-sub-Gaussian if $\log \mathbb{E}\left[\exp \lambda(\mathrm{X} - \mathbb{E}\left[\mathrm{X}\right])\right] \leq \sigma^2\lambda^2/2$ for any $\lambda \in \mathbb{R}$. For a random vector $\mathrm{X} = (\mathrm{X}_1, \cdots, \mathrm{X}_d)$, we define its variance and minimum mean absolute error (MMAE) as

$$\mathsf{Var}\left(\mathrm{X}\right) \triangleq \inf_{\boldsymbol{a} \in \mathbb{R}^d} \mathbb{E}\left[\|\mathrm{X} - \boldsymbol{a}\|_2^2\right], \tag{6}$$

$$\mathsf{mmae}(\mathrm{X}) \triangleq \inf_{\boldsymbol{a} \in \mathbb{R}^d} \mathbb{E}\left[\|\mathrm{X} - \boldsymbol{a}\|_1\right]. \tag{7}$$

The vector $\boldsymbol{a}$ which minimizes (6) and (7) are

$$\operatorname*{argmin}_{\boldsymbol{a} \in \mathbb{R}^d} \mathbb{E}\left[\|\mathrm{X} - \boldsymbol{a}\|_2^2\right] = (\mathbb{E}\left[\mathrm{X}_1\right], \cdots, \mathbb{E}\left[\mathrm{X}_d\right]), \tag{8}$$

$$\operatorname*{argmin}_{\boldsymbol{a} \in \mathbb{R}^d} \mathbb{E}\left[\|\mathrm{X} - \boldsymbol{a}\|_1\right] = (\mathsf{median}(\mathrm{X}_1), \cdots, \mathsf{median}(\mathrm{X}_d)), \tag{9}$$

where $\mathsf{median}(\mathrm{X}_i)$ is the median of the random variable X_i.

In order to measure the difference between two probability distributions, we recall Csiszár's f-divergence (Csiszár, 1967). Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ and P, Q be two probability distributions over a set $\mathcal{X} \subseteq \mathbb{R}^d$. The f-divergence between P and Q is defined as

$$\mathrm{D}_f(P\|Q) \triangleq \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q. \tag{10}$$

Examples of f-divergence include KL-divergence ($f(t) = t\log t$), total variation distance ($f(t) = |t-1|/2$), and χ^2-divergence ($f(t) = t^2 - 1$). The f-divergence motivates a way of measuring dependence between a pair of random variables (X, Y). Specifically, the f-information between (X, Y) is defined as

$$I_f(\mathrm{X};\mathrm{Y}) \triangleq \mathrm{D}_f(P_{\mathrm{X},\mathrm{Y}} \| P_{\mathrm{X}} \otimes P_{\mathrm{Y}}) = \mathbb{E}\left[\mathrm{D}_f(P_{\mathrm{Y}|\mathrm{X}} \| P_{\mathrm{Y}})\right], \tag{11}$$

where $P_{\mathrm{X},\mathrm{Y}}$ is the joint distribution, P_{X}, P_{Y} are the marginal distributions, $P_{\mathrm{Y}|\mathrm{X}}$ is the conditional distribution, and the expectation is taken over $\mathrm{X} \sim P_{\mathrm{X}}$. In particular, if the KL-divergence is used in (11), the corresponding f-information is the well-known mutual information (Shannon, 1948).

2.2 Information-theoretic Generalization Bounds

A recent work by Xu and Raginsky (2017) provided a framework for analyzing algorithmic generalization. Specifically, they considered a learning algorithm as a channel (i.e., conditional probability distribution) that takes a training set S as input and outputs a parameter W. Furthermore, they derived an upper bound for the expected generalization gap using the mutual information $I(\mathrm{W};\mathrm{S})$. This bound was later tightened by Bu et al. (2020) using an individual sample mutual information. Next, we recall the generalization bound in Bu et al. (2020). By adapting their proof and leveraging variational representations of f-divergence (Nguyen et al., 2010), we present another two generalization bounds based on different kinds of f-information. Note that these two bounds can also be obtained from Corollary 1 in Rodríguez Gálvez et al. (2021) and applying Jensen's inequality to Eq. (15) in the same paper, respectively.

Lemma 1 *Consider a learning algorithm which takes a data set* $\mathrm{S} = (\mathrm{Z}_1, \cdots, \mathrm{Z}_n)$ *as input and outputs* W.

- *(Proposition 1 in Bu et al., 2020) If the loss* $\ell(\boldsymbol{w}, \mathrm{Z})$ *is* σ*-sub-Gaussian under* $\mathrm{Z} \sim \mu$ *for all* $\boldsymbol{w} \in \mathcal{W}$*, then*

$$|\mathbb{E}\left[L_\mu(\mathrm{W}) - L_{\mathrm{S}}(\mathrm{W})\right]| \leq \frac{\sqrt{2}\sigma}{n} \sum_{i=1}^{n} \sqrt{I(\mathrm{W};\mathrm{Z}_i)}, \tag{12}$$

  where $I(\mathrm{W};\mathrm{Z}_i)$ *is the mutual information (i.e.,* f*-information with* $f(t) = t\log t$*).*

- *If the loss* $\ell(\boldsymbol{w}, \mathrm{Z})$ *is upper bounded by a constant* $A > 0$*, then*

$$|\mathbb{E}\left[L_\mu(\mathrm{W}) - L_{\mathrm{S}}(\mathrm{W})\right]| \leq \frac{A}{n} \sum_{i=1}^{n} \mathrm{T}(\mathrm{W};\mathrm{Z}_i), \tag{13}$$

  where $\mathrm{T}(\mathrm{W};\mathrm{Z}_i)$ *is the* T*-information (i.e.,* f*-information with* $f(t) = |t-1|/2$*).*

- *If the variance of the loss function is finite (i.e.,* $\mathsf{Var}\left(\ell(\mathrm{W};\mathrm{Z})\right) < \infty$*), then*

$$|\mathbb{E}\left[L_\mu(\mathrm{W}) - L_{\mathrm{S}}(\mathrm{W})\right]| \leq \frac{\sqrt{\mathsf{Var}\left(\ell(\mathrm{W};\mathrm{Z})\right)}}{n} \sum_{i=1}^{n} \sqrt{\chi^2(\mathrm{W};\mathrm{Z}_i)}, \tag{14}$$

  where $\chi^2(\mathrm{W};\mathrm{Z}_i)$ *is the* χ^2*-information (i.e.,* f*-information with* $f(t) = t^2 - 1$*) and* Z *is a fresh data point which is independent of* W *(i.e.,* $(\mathrm{W}, \mathrm{Z}) \sim P_{\mathrm{W}} \otimes \mu$*).*

Proof See Appendix A.1. ■

We apply Lemma 1 to analyze the generalization capability of noisy iterative algorithms. Estimating the f-information in Lemma 1 from data is intractable. Hence, we further upper bound these f-information by exploring properties of additive noise channels in the next section. Furthermore, we also incorporate a time-decaying factor into our bound, which is established by strong data processing inequalities, recalled in the upcoming subsection.

Although our analysis is applicable for *any* f-information, we focus on the three f-information in Lemma 1 since:

- Mutual information is often easier to work with due to its many useful properties. For example, the chain rule of mutual information plays an important role for handling sampling with replacement (see Section 5.3).

- T-information often yields a tighter bound than (12) and (14). This can be seen by the following f-divergence inequalities (see Eq. 1 and 94 in Sason and Verdú, 2016):

$$\sqrt{2}\mathrm{T}(\mathrm{W};\mathrm{Z}_i) \leq \sqrt{I(\mathrm{W};\mathrm{Z}_i)} \leq \sqrt{\log(1+\chi^2(\mathrm{W};\mathrm{Z}_i))} \leq \sqrt{\chi^2(\mathrm{W};\mathrm{Z}_i)}.$$

  Furthermore, the T-information can be used to analyze a broader class of noisy iterative algorithms. For example, when the additive noise is drawn from a distribution with bounded support, the other two f-information may lead to an infinite generalization bound while the T-information can still give a non-trivial bound (see the last row in Table 1).

- χ^2-information requires the mildest assumptions. Apart from bounded loss functions, it is often hard to verify the sub-Gaussianity of $\ell(\boldsymbol{w},\mathrm{Z})$ for all $\boldsymbol{w}$. The advantage of (14) is that it replaces the sub-Gaussian constant with the variance of the loss function.

Remark 2 *Using f-information for bounding generalization gap has appeared in prior literature (see e.g., Alabdulmohsin, 2015; Jiao et al., 2017; Wang et al., 2019; Esposito et al., 2021; Rodríguez Gálvez et al., 2021; Aminian et al., 2021; Jose and Simeone, 2021a). More broadly, there are significant recent works (see e.g., Raginsky et al., 2016; Asadi et al., 2018; Lopez and Jog, 2018; Steinke and Zakynthinou, 2020; Hellström and Durisi, 2020; Yagli et al., 2020; Hafez-Kolahi et al., 2020; Jose and Simeone, 2021b; Zhou et al., 2021) on deriving new information-theoretic generalization bounds and applying them to different applications. The reason we adopt Lemma 1 for analyzing noisy iterative algorithms is that it enables us to incorporate a time-decaying factor into our bounds.*

2.3 Strong Data Processing Inequalities

In order to characterize the time-decaying phenomenon, we use an information-theoretic tool: strong data processing inequalities (Dobrushin, 1956; Cohen et al., 1998). We start with recalling the data processing inequality.

Lemma 3 *If a Markov chain* $\mathrm{U} \to \mathrm{X} \to \mathrm{Y}$ *holds, then*

$$I_f(\mathrm{U};\mathrm{Y}) \leq I_f(\mathrm{U};\mathrm{X}). \tag{15}$$

The data processing inequality states that no post-processing of X can increase the information about U. Under certain conditions, the data processing inequality can be sharpened, which leads to a strong data processing inequality, often cast in terms of a contraction coefficient. Next, we recall the contraction coefficients of f-divergences and show their connection with strong data processing inequalities.

For a given transition probability kernel $P_{\mathrm{Y|X}} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ where $\mathcal{P}(\mathcal{Y})$ is the set of all distributions on $\mathcal{Y}$, let $P_{\mathrm{Y|X}} \circ P$ be the distribution on $\mathcal{Y}$ induced by the push-forward of the distribution P (i.e., the distribution of Y when the distribution of X is P). The contraction coefficient of $P_{\mathrm{Y|X}}$ for D_f is defined as

$$\eta_f(P_{\mathrm{Y|X}}) \triangleq \sup_{P,Q:P\neq Q} \frac{\mathrm{D}_f(P_{\mathrm{Y|X}} \circ P \| P_{\mathrm{Y|X}} \circ Q)}{\mathrm{D}_f(P\|Q)} \in [0,1].$$

In particular, when the total variation distance is used, the corresponding contraction coefficient $\eta_{\mathrm{TV}}(P_{\mathrm{Y|X}})$ is known as the Dobrushin's coefficient (Dobrushin, 1956), which owns an equivalent expression:

$$\eta_{\mathrm{TV}}(P_{\mathrm{Y|X}}) = \sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \mathrm{D}_{\mathrm{TV}}(P_{\mathrm{Y|X}=\boldsymbol{x}} \| P_{\mathrm{Y|X}=\boldsymbol{x}'}). \tag{16}$$

Note that the Dobrushin's coefficient upper bounds all other contraction coefficients (Cohen et al., 1998):

$$\eta_f(P_{\mathrm{Y|X}}) \leq \eta_{\mathrm{TV}}(P_{\mathrm{Y|X}}).$$

Furthermore, for any Markov chain $\mathrm{U} \to \mathrm{X} \to \mathrm{Y}$, the contraction coefficients satisfy (see Theorem 5.2 in Raginsky, 2016, for a proof)

$$I_f(\mathrm{U};\mathrm{Y}) \leq \eta_f(P_{\mathrm{Y|X}}) \cdot I_f(\mathrm{U};\mathrm{X}). \tag{17}$$

When $\eta_f(P_{\mathrm{Y|X}}) < 1$, the strict inequality $I_f(\mathrm{U};\mathrm{Y}) < I_f(\mathrm{U};\mathrm{X})$ improves the data processing inequality and, hence, is referred to as a strong data processing inequality. We refer the reader to Polyanskiy and Wu (2016) and Raginsky (2016) for a more comprehensive review on strong data processing inequalities and Calmon et al. (2017) for non-linear strong data processing inequalities in Gaussian channels.

3. Properties of Additive Noise Channels

Additive noise channels have a long history in information theory. Here we show two important properties of additive noise channels which will be used for deriving the generalization bounds in the next section. The first property (Lemma 4) leads to a decay factor into our bounds. The second property (Lemma 6) produces computable generalization bounds.

Consider a single use of an additive noise channel. Let (X,Y) be a pair of random variables related by $\mathrm{Y} = \mathrm{X} + m\mathrm{N}$ where $\mathrm{X} \in \mathcal{X}$; $m > 0$ is a constant; and N represents an independent noise. In other words, the conditional distribution of Y given X can be characterized by $P_{\mathrm{Y|X}=\boldsymbol{x}} = P_{\boldsymbol{x}+m\mathrm{N}}$. If $\mathcal{X}$ is a compact set, the contraction coefficients often have a non-trivial upper bound, leading to a strong data processing inequality. This is

formalized in the following lemma whose proof follows directly from the definition of the Dobrushin's coefficient in (16) and the fact that the Dobrushin's coefficient is a universal upper bound of all the contraction coefficients. Note that the function $\delta(A, m)$ in (18) upper bounds the Dobrushin's coefficient of the additive noise channel and has a closed-form expression for many noise distributions (see Table 1 for examples).

Lemma 4 *Let* N *be a random variable which is independent of* (U, X)*. For a given norm* $\|\cdot\|$ *on a compact set* $\mathcal{X} \subseteq \mathbb{R}^d$ *and* $m, A > 0$*, we define*

$$\delta(A, m) \triangleq \sup_{\|\boldsymbol{x}-\boldsymbol{x}'\| \leq A} \mathrm{D}_{\mathrm{TV}}\left(P_{\boldsymbol{x}+m\mathrm{N}} \| P_{\boldsymbol{x}'+m\mathrm{N}}\right). \tag{18}$$

Then the Markov chain $\mathrm{U} \to \mathrm{X} \to \mathrm{X} + m\mathrm{N}$ *holds and*

$$I_f(\mathrm{U}; \mathrm{X} + m\mathrm{N}) \leq \delta(\mathsf{diam}(\mathcal{X}), m) \cdot I_f(\mathrm{U}; \mathrm{X}), \tag{19}$$

where $\mathsf{diam}(\mathcal{X}) \triangleq \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}'\|$ *is the diameter of* $\mathcal{X}$*.*

Remark 5 *In the next section, we use Lemma 4 to incorporate a time-decaying factor into our generalization bounds. The function* $\delta(A, m)$ *in* (18) *yields a closed-form expression of the decay factor. As mentioned earlier, our analysis is inspired by a line of works on privacy amplification by iterations (Feldman et al., 2018; Asoodeh et al., 2020). For example, Feldman et al. (2018) proved that passing two probability distributions through a noisy iterative algorithm would shrink their Rényi divergence, leading to amplification for Rényi differential privacy. Asoodeh et al. (2020) reformulated the definition of differential privacy by using the* E_γ *divergence, which is a certain* f*-divergence. They characterized the contraction coefficient of* E_γ *by generalizing Dobrushin's coefficient (see Section 3 in their paper). In this paper, we introduce a general framework for deriving generalization bounds through* f*-divergences. We leverage Dobrushin's coefficient since it upper bounds all other contraction coefficients for* f*-divergences.*

Computing f-information in general is intractable when the underlying distribution is unknown. Hence, we further upper bound the f-information in Lemma 1 by a quantity which is easier to compute. To achieve this goal, we introduce another property of additive noise channels. Specifically, let $\mathrm{Y} = \mathrm{X} + m\mathrm{N}$ and $\mathrm{Y}' = \mathrm{X}' + m\mathrm{N}$ be the output variables from the same additive noise channel with input variables X and X′, respectively. Then the f-divergence in the output space can be upper bounded by the optimal transport cost in the input space.

Lemma 6 *Let* N *be a random variable which is independent of* $(\mathrm{X}, \mathrm{X}')$*. For* $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ *and* $m > 0$*, we define a cost function*[3]

$$\mathsf{C}_f(\boldsymbol{x}, \boldsymbol{x}'; m) \triangleq \mathrm{D}_f\left(P_{\boldsymbol{x}+m\mathrm{N}} \| P_{\boldsymbol{x}'+m\mathrm{N}}\right). \tag{20}$$

Then for any $m > 0$*, we have*

$$\mathrm{D}_f(P_{\mathrm{X}+m\mathrm{N}} \| P_{\mathrm{X}'+m\mathrm{N}}) \leq \mathbb{W}(P_{\mathrm{X}}, P_{\mathrm{X}'}; m). \tag{21}$$

3. Note that $\mathsf{C}_f(\boldsymbol{x}, \boldsymbol{x}'; m)$ is not necessarily a metric.

Noise Type	$\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x},\boldsymbol{x}';m)$	$\mathsf{C}_{\chi^2}(\boldsymbol{x},\boldsymbol{x}';m)$	$\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x},\boldsymbol{x}';m)$	$\delta(A,m)$
Gaussian	$\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2m^2}$	$\exp\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{m^2}\right)-1$	$\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2}{2m}$	$1-2\bar{\Phi}\left(\frac{A}{2m}\right)$
Laplace	$\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_1}{m}$	$\exp\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_1}{m}\right)-1$	$\sqrt{\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_1}{2m}}$	$1-\exp\left(-\frac{A}{m}\right)$
Uniform on $[-1,1]$	$\infty\mathbb{I}_{[x\neq x']}$	$\infty\mathbb{I}_{[x\neq x']}$	$\min\left\{1,\left\|\frac{x-x'}{2m}\right\|\right\}$	$\min\left\{1,\frac{A}{2m}\right\}$

Table 1: Closed-form expressions (or upper bounds if in blue color) of $\mathsf{C}_f(\boldsymbol{x},\boldsymbol{x}';m)$ (see (20) for its definition) and $\delta(A,m)$ (see (18) for its definition). The function $\delta(A,m)$ is equipped with the 2-norm for Gaussian distribution and 1-norm for Laplace distribution. We denote the Gaussian complementary cumulative distribution function (CCDF) by $\bar{\Phi}(x)\triangleq\int_x^\infty\frac{1}{\sqrt{2\pi}}\exp(-v^2/2)\mathrm{d}v$ and define $\infty\cdot 0=0$ as convention. The proof is deferred to Appendix B.3.

Here $\mathbb{W}(P_{\mathrm{X}},P_{\mathrm{X}'};m)$ is the optimal transport cost:

$$\mathbb{W}(P_{\mathrm{X}},P_{\mathrm{X}'};m)\triangleq\inf\mathbb{E}\left[\mathsf{C}_f(\mathrm{X},\mathrm{X}';m)\right], \tag{22}$$

where the infimum is taken over all couplings (i.e., joint distributions) of the random variables X and X' with marginals P_{X} and $P_{\mathrm{X}'}$, respectively.

Proof See Appendix B.1. ■

Lemmas 4 and 6 show that the functions $\delta(A,m)$ and $\mathsf{C}_f(\boldsymbol{x},\boldsymbol{x}';m)$ can be useful for sharpening the data processing inequality and upper bounding the f-information in Lemma 1. We demonstrate in Table 1 that these functions can be expressed in closed-form for specific additive noise distributions.

Remark 7 *Let N be drawn from a Gaussian distribution. Substituting the closed-form expression of $\mathsf{C}_{KL}(\boldsymbol{x},\boldsymbol{x}';m)$ from Table 1 into Lemma 6 leads to*

$$\mathrm{D}_{KL}(P_{\mathrm{X}+m\mathrm{N}}\|P_{\mathrm{X}'+m\mathrm{N}})\leq\frac{1}{2m^2}\mathbb{W}_2^2(P_{\mathrm{X}},P_{\mathrm{X}'}) \tag{23}$$

where $\mathbb{W}_2(P_{\mathrm{X}},P_{\mathrm{X}'})$ is the 2-Wasserstein distance equipped with the L_2 cost function:

$$\mathbb{W}_2^2(P_{\mathrm{X}},P_{\mathrm{X}'})\triangleq\inf\mathbb{E}\left[\|\mathrm{X}-\mathrm{X}'\|_2^2\right].$$

This inequality serves as a fundamental building block for proving Otto-Villani's HWI inequality (Otto and Villani, 2000) in the Gaussian case (Raginsky and Sason, 2013; Boucheron et al., 2013).

4. Generalization Bounds for Noisy Iterative Algorithms

In this section, we present our main result—generalization bounds for noisy iterative algorithms. First, by leveraging strong data processing inequalities, we prove that the amount

of information about the data points used in early iterations decays with time. Accordingly, our generalization bounds incorporate a time-decaying factor which enables the impact of early iterations on our bounds to reduce with time. Second, by using properties of additive noise channels developed in the last section, we further upper bound the f-information by a quantity which is often easier to estimate. The above two aspects correspond to Lemma 8 and 9 which are the basis of our main result in Theorem 10.

Before diving into the analysis, we first discuss assumptions made in this paper.

Assumption 1 (Sampling w/o Replacement) *The mini-batch indices* $(\mathcal{B}_1, \cdots, \mathcal{B}_T)$ *in* (3) *are specified before the algorithm is run and data are drawn without replacement.*

If the mini-batches are selected when the algorithm is run, one can analyze the expected generalization gap by first conditioning on $\mathcal{B} \triangleq (\mathcal{B}_1, \cdots, \mathcal{B}_T)$ and then taking an expectation over the randomness of $\mathcal{B}$:

$$\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right] = \mathbb{E}\left[\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T) \mid \mathcal{B}\right]\right].$$

Our analysis can be extended to the case where data are drawn with replacement (see Proposition 15) by using the chain rule for mutual information.

Assumption 2 (Bounded Gradient & Compact Domain) *The parameter domain* $\mathcal{W}$ *is compact and* $\|g(\boldsymbol{w}, \boldsymbol{z})\| \leq K$ *for all* $\boldsymbol{w}, \boldsymbol{z}$. *We denote the diameter of* $\mathcal{W}$ *by* $D \triangleq \sup_{\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}} \|\boldsymbol{w} - \boldsymbol{w}'\|$.

Our generalization bounds rely on the second assumption mildly. In fact, this assumption only affects the time-decaying factor in our bounds which is always upper bounded by 1. If we remove this assumption, our bounds still hold though the decay factor disappears.

Now we are in a position to derive generalization bounds under the above assumptions. As a consequence of strong data processing inequalities, the following lemma indicates that the information of a data point Z_i contained in the algorithmic output W_T will reduce with time T.

Lemma 8 *Under Assumption 1 (Sampling w/o Replacement) and 2 (Bounded Gradient & Compact Domain), if a data point* Z_i *is used in the* t*-th iteration, then*

$$I_f(\mathrm{W}_T; \mathrm{Z}_i) \leq I_f(\mathrm{W}_t; \mathrm{Z}_i) \cdot \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'} K, m_{t'}), \tag{24}$$

where the function $\delta(\cdot, \cdot)$ *is defined in* (18).

Proof For the t-th iteration, we rewrite the recursion in (3) as

$$\mathrm{U}_t = \mathrm{W}_{t-1} - \eta_t \cdot g(\mathrm{W}_{t-1}, \{\mathrm{Z}_i\}_{i \in \mathcal{B}_t}) \tag{25a}$$

$$\mathrm{V}_t = \mathrm{U}_t + m_t \cdot \mathrm{N}_t \tag{25b}$$

$$\mathrm{W}_t = \mathsf{Proj}_{\mathcal{W}}(\mathrm{V}_t). \tag{25c}$$

Let Z_i be a data point used at the t-th iteration. Under Assumption 1 (Sampling w/o Replacement), the following Markov chain holds:

$$\mathrm{Z}_i \to \mathrm{U}_t \to \mathrm{V}_t \to \mathrm{W}_t \to \cdots \to \mathrm{W}_{T-1} \to \mathrm{U}_T \to \mathrm{V}_T \to \mathrm{W}_T. \tag{26}$$

Let $\mathcal{U}_T$ be the range of U_T. By Assumption 2 (Bounded Gradient & Compact Domain) and the triangle inequality,

$$\mathsf{diam}(\mathcal{U}_T) \leq \mathsf{diam}(\mathcal{W}) + 2\eta_T K = D + 2\eta_T K.$$

Now we leverage the strong data processing inequality in Lemma 4 and obtain

$$\begin{aligned} I_f(\mathrm{W}_T; \mathrm{Z}_i) &\leq I_f(\mathrm{V}_T; \mathrm{Z}_i) \\ &\leq \delta(D + 2\eta_T K, m_T) \cdot I_f(\mathrm{U}_T; \mathrm{Z}_i) \\ &\leq \delta(D + 2\eta_T K, m_T) \cdot I_f(\mathrm{W}_{T-1}; \mathrm{Z}_i), \end{aligned}$$

where the first and last steps are due to the data processing inequality (Lemma 3). Applying this procedure recursively leads to the desired conclusion. ■

For many types of noise (e.g., Gaussian or Laplace noise), the function $\delta(\cdot, \cdot)$ is *strictly* smaller than 1 (see Table 1). In this case, the information about the data points used in early iterations is reducing via the multiplicative factor in (24). Furthermore, one can even prove that $I_f(\mathrm{W}_T; \mathrm{Z}_i) \to 0$ as $T \to \infty$ if the magnitude of the additive noise in (3) has a lower bound.

Lemma 8 explains how our generalization bounds in Theorem 10 incorporate a time-decaying factor. However, it still involves an f-information $I_f(\mathrm{W}_t; \mathrm{Z}_i)$, which can be hard to compute from data. Next, we further upper bound this f-information by using properties of additive noise channels developed in the last section (see Lemma 6).

Lemma 9 *Under Assumption 1 (Sampling w/o Replacement), if a data point Z_i is used at the t-th iteration, then*

$$I_f(\mathrm{W}_t; \mathrm{Z}_i) \leq \mathbb{E}\left[\mathsf{C}_f\left(g(\mathrm{W}_{t-1}, \mathrm{Z}), g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}); \frac{m_t b_t}{\eta_t}\right)\right], \tag{27}$$

where the function $\mathsf{C}_f(\cdot, \cdot; \cdot)$ is defined in (20) *and the expectation is taken over $(\mathrm{W}_{t-1}, \mathrm{Z}, \bar{\mathrm{Z}}) \sim P_{\mathrm{W}_{t-1}} \otimes \mu \otimes \mu$.*

Proof Recall the definition of U_t, V_t in (25). The data processing inequality yields

$$I_f(\mathrm{W}_t; \mathrm{Z}_i) \leq I_f(\mathrm{V}_t; \mathrm{Z}_i). \tag{28}$$

By the definition of f-information, we can write

$$I_f(\mathrm{V}_t; \mathrm{Z}_i) = \mathbb{E}\left[\mathrm{D}_f(P_{\mathrm{V}_t|\mathrm{Z}_i} \| P_{\mathrm{V}_t})\right] = \int_{\mathcal{Z}} \mathrm{D}_f(P_{\mathrm{V}_t|\mathrm{Z}_i = \boldsymbol{z}} \| P_{\mathrm{V}_t}) \mathrm{d}\mu(\boldsymbol{z}). \tag{29}$$

Since $\mathrm{V}_t = \mathrm{U}_t + m_t \cdot \mathrm{N}_t$ by its definition, Lemma 6 leads to

$$\mathrm{D}_f\left(P_{\mathrm{V}_t|\mathrm{Z}_i = \boldsymbol{z}} \| P_{\mathrm{V}_t}\right) \leq \mathbb{W}(P_{\mathrm{U}_t|\mathrm{Z}_i = \boldsymbol{z}}, P_{\mathrm{U}_t}; m_t). \tag{30}$$

To further upper bound the above optimal transport cost, we construct a special coupling. Let W_{t-1} be the output of the noisy iterative algorithm at the $(t-1)$-st iteration. Under

Assumption 1 (Sampling w/o Replacement), the data point Z_i is independent of W_{t-1} since it is only used at the t-th iteration. Then we introduce two random variables:

$$\mathrm{U}^*_{\boldsymbol{z}} \triangleq \mathrm{W}_{t-1} - \frac{\eta_t}{b_t}\left(\sum_{j\in\mathcal{B}_t, j\neq i} g(\mathrm{W}_{t-1}, \mathrm{Z}_j) + g(\mathrm{W}_{t-1}, \boldsymbol{z})\right),$$
$$\mathrm{U}^* \triangleq \mathrm{W}_{t-1} - \frac{\eta_t}{b_t}\sum_{j\in\mathcal{B}_t} g(\mathrm{W}_{t-1}, \mathrm{Z}_j).$$

Here $\mathrm{U}^*_{\boldsymbol{z}}$ and U^* have marginals $P_{\mathrm{U}_t|\mathrm{Z}_i=\boldsymbol{z}}$ and P_{U_t}, respectively. By the definition of optimal transport cost in (22), we have

$$\mathbb{W}\left(P_{\mathrm{U}_t|\mathrm{Z}_i=\boldsymbol{z}}, P_{\mathrm{U}_t}; m_t\right) \leq \mathbb{E}\left[\mathsf{C}_f(\mathrm{U}^*_{\boldsymbol{z}}, \mathrm{U}^*; m_t)\right]. \tag{31}$$

The property of $\mathsf{C}_f(\boldsymbol{x}, \boldsymbol{y}; m)$ in Lemma 17 yields

$$\begin{aligned}\mathbb{E}\left[\mathsf{C}_f(\mathrm{U}^*_{\boldsymbol{z}}, \mathrm{U}^*; m_t)\right] &= \mathbb{E}\left[\mathsf{C}_f\left(-\frac{\eta_t}{b_t}g(\mathrm{W}_{t-1}, \boldsymbol{z}), -\frac{\eta_t}{b_t}g(\mathrm{W}_{t-1}, \mathrm{Z}_i); m_t\right)\right] \\ &= \mathbb{E}\left[\mathsf{C}_f\left(\frac{\eta_t}{b_t}g(\mathrm{W}_{t-1}, \mathrm{Z}_i), \frac{\eta_t}{b_t}g(\mathrm{W}_{t-1}, \boldsymbol{z}); m_t\right)\right] \\ &= \mathbb{E}\left[\mathsf{C}_f\left(g(\mathrm{W}_{t-1}, \mathrm{Z}_i), g(\mathrm{W}_{t-1}, \boldsymbol{z}); \frac{m_t b_t}{\eta_t}\right)\right]. \end{aligned} \tag{32}$$

We introduce two independent copies $\mathrm{Z}, \bar{\mathrm{Z}}$ of Z_i such that $(\mathrm{W}_{t-1}, \mathrm{Z}, \bar{\mathrm{Z}}) \sim P_{\mathrm{W}_{t-1}} \otimes \mu \otimes \mu$. Combining (29–32) and using Tonelli's theorem lead to

$$I_f(\mathrm{V}_t; \mathrm{Z}_i) \leq \mathbb{E}\left[\mathsf{C}_f\left(g(\mathrm{W}_{t-1}, \mathrm{Z}), g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}); \frac{m_t b_t}{\eta_t}\right)\right]. \tag{33}$$

Substituting (33) into (28) gives the desired conclusion. ■

With Lemma 1, 8, and 9 in hand, we now present the main result in this section: three generalization bounds for noisy iterative algorithms under different assumptions.

Theorem 10 *Suppose that Assumption 1 (Sampling w/o Replacement) and 2 (Bounded Gradient & Compact Domain) hold.*

- *If the loss $\ell(\boldsymbol{w}, \mathrm{Z})$ is σ-sub-Gaussian under $\mathrm{Z} \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, the expected generalization gap $\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right]$ can be upper bounded by*

$$\frac{\sqrt{2}\sigma}{n}\sum_{t=1}^{T} b_t \sqrt{\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(\mathrm{W}_{t-1}, \mathrm{Z}), g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}); \frac{m_t b_t}{\eta_t}\right)\right] \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'}K, m_{t'})}. \tag{34}$$

- *If the loss function is upper bounded by a constant $A > 0$, the expected generalization gap $\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right]$ can be upper bounded by*

$$\frac{A}{n}\sum_{t=1}^{T} b_t \mathbb{E}\left[\mathsf{C}_{\mathrm{TV}}\left(g(\mathrm{W}_{t-1}, \mathrm{Z}), g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}); \frac{m_t b_t}{\eta_t}\right)\right] \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'}K, m_{t'}). \tag{35}$$

- *If the variance of the loss function is finite (i.e.,* $\mathsf{Var}\left(\ell(\mathrm{W}_T;\mathrm{Z})\right) < \infty$*), the expected generalization gap* $\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right]$ *can be upper bounded by*

$$\frac{\sigma}{n}\sum_{t=1}^{T} b_t \sqrt{\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); \frac{m_t b_t}{\eta_t}\right)\right] \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'}K, m_{t'})}, \quad (36)$$

where $\sigma \triangleq \sqrt{\mathsf{Var}\left(\ell(\mathrm{W}_T;\mathrm{Z})\right)}$ *with* $(\mathrm{W}_T,\mathrm{Z}) \sim P_{\mathrm{W}_T} \otimes \mu$.

Proof See Appendix C.1. ■

Our generalization bounds involve the information (e.g., step size η_t, magnitude of noise m_t, and batch size b_t) at all iterations. Moreover, our bounds depend on the data distribution μ through the expectation terms. Under Assumption 1 (Sampling w/o Replacement), if a data point Z_i is used in the t-th iteration, it is independent of W_{t-1} (i.e., $P_{\mathrm{W}_{t-1},\mathrm{Z}_i} = P_{\mathrm{W}_{t-1}} \otimes \mu$). This is why the expectations are taken over $(\mathrm{W}_{t-1},\mathrm{Z},\bar{\mathrm{Z}}) \sim P_{\mathrm{W}_{t-1}} \otimes \mu \otimes \mu$. Finally, since the function δ is often strictly smaller than 1 (see Table 1 for some examples), the multiplicative factor $\prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'}K, m_{t'})$ enables the impact of early iterations on our bounds to reduce with time T.

The generalization bounds in Theorem 10 may seem contrived at first glance as they rely on the functions δ and C_f defined in (18) and (20). However, in the next section, we will show that these bounds can be significantly simplified when we apply them to real applications. Furthermore, we will also compare the advantage of each bound under these applications.

5. Applications

We demonstrate the generalization bounds in Theorem 10 through several applications in this section.

5.1 Differentially Private Stochastic Gradient Descent (DP-SGD)

Differentially private stochastic gradient descent (DP-SGD) is a variant of SGD where noise is added to a stochastic gradient estimator in order to ensure privacy of each individual record. We recall an implementation of (projected) DP-SGD (see e.g., Algorithm 1 in Feldman et al., 2018). At each iteration, the parameter of the empirical risk is updated using the following rule:

$$\mathrm{W}_t = \mathsf{Proj}_{\mathcal{W}}\left(\mathrm{W}_{t-1} - \eta\left(g(\mathrm{W}_{t-1}, \{\mathrm{Z}_i\}_{i\in\mathcal{B}_t}) + \mathrm{N}_t\right)\right), \quad (37)$$

where N_t is an additive noise drawn independently from a distribution P_N; $\mathcal{B}_t$ contains the indices of the data points used at the current iteration and $b_t \triangleq |\mathcal{B}_t|$; the function g indicates a direction for updating the parameter. The recursion in (37) is run for T iterations and we assume that data are drawn without replacement. At the end of each iteration, the parameter is projected onto a compact domain $\mathcal{W}$. We denote the diameter of $\mathcal{W}$ by D. The output from the DP-SGD algorithm is the last iterate W_T. Finally, we assume that

$$\sup_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{z}\in\mathcal{Z}} \|g(\boldsymbol{w},\boldsymbol{z})\| \le K. \quad (38)$$

This assumption can be satisfied by gradient clipping and is crucial for guaranteeing differential privacy as it controls the sensitivity of each update.

The differential privacy guarantees of the DP-SGD algorithm have been extensively studied in the literature (see e.g., Song et al., 2013; Wu et al., 2017; Feldman et al., 2018; Balle et al., 2019; Asoodeh et al., 2020). Here we consider a different angle: the generalization of DP-SGD. We derive generalization bounds for the DP-SGD algorithm under Laplace and Gaussian mechanisms by using our Theorem 10.

Proposition 11 (Laplace mechanism) *Suppose that the additive noise* N_t *in* (37) *follows a standard multivariate Laplace distribution. Let* $\mathcal{W}$ *be equipped with the 1-norm and* $q \triangleq 1 - \exp\left(-(D + 2\eta K)/\eta\right) \in (0, 1)$.

- *If the loss* $\ell(\boldsymbol{w}, \mathrm{Z})$ *is* σ*-sub-Gaussian under* $\mathrm{Z} \sim \mu$ *for all* $\boldsymbol{w} \in \mathcal{W}$*, then*

$$\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right] \leq \frac{2\sigma}{n} \sum_{t=1}^{T} \sqrt{b_t \cdot \mathsf{mmae}\left(g(\mathrm{W}_{t-1}, \mathrm{Z})\right) \cdot q^{T-t}}. \tag{39}$$

- *If the loss function is upper bounded by* $A > 0$*, then*

$$\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right] \leq \frac{\sqrt{2}A}{n} \sum_{t=1}^{T} \sqrt{b_t} \cdot \mathbb{E}\left[\sqrt{\left\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - \boldsymbol{e}\right\|_1}\right] \cdot q^{T-t}, \tag{40}$$

  where $\boldsymbol{e} \triangleq \mathsf{median}\left(g(\mathrm{W}_{t-1}, \mathrm{Z})\right)$.

- *If the variance of the loss function is bounded (i.e.,* $\mathsf{Var}\left(\ell(\mathrm{W}_T; \mathrm{Z})\right) < \infty$*), then*

$$\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right] \leq \frac{\sigma}{n} \sum_{t=1}^{T} \sqrt{b_t \cdot \mathbb{E}\left[\exp\left(2\left\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - \boldsymbol{e}\right\|_1\right) - 1\right] \cdot q^{T-t}}, \tag{41}$$

  where $\sigma = \sqrt{\mathsf{Var}\left(\ell(\mathrm{W}_T; \mathrm{Z})\right)}$ *and* $\boldsymbol{e} \triangleq \mathsf{median}\left(g(\mathrm{W}_{t-1}, \mathrm{Z})\right)$.

Proof See Appendix D.1. ■

Proposition 12 (Gaussian mechanism) *Suppose that the additive noise* N_t *in* (37) *follows a standard multivariate Gaussian distribution. Let* $\mathcal{W}$ *be equipped with the 2-norm and* $q \triangleq 1 - 2\bar{\Phi}\left((D + 2\eta K)/2\eta\right) \in (0, 1)$ *with* $\bar{\Phi}(\cdot)$ *being the Gaussian CCDF.*

- *If the loss* $\ell(\boldsymbol{w}, \mathrm{Z})$ *is* σ*-sub-Gaussian under* $\mathrm{Z} \sim \mu$ *for all* $\boldsymbol{w} \in \mathcal{W}$*, then*

$$\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right] \leq \frac{2\sigma}{n} \sum_{t=1}^{T} \sqrt{\mathsf{Var}\left(g(\mathrm{W}_{t-1}, \mathrm{Z})\right) \cdot q^{T-t}}. \tag{42}$$

- *If the loss function is upper bounded by* $A > 0$*, then*

$$\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right] \leq \frac{A}{n} \sum_{t=1}^{T} \mathbb{E}\left[\left\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - \boldsymbol{e}\right\|_2\right] \cdot q^{T-t}, \tag{43}$$

  where $\boldsymbol{e} \triangleq \mathbb{E}\left[g(\mathrm{W}_{t-1}, \mathrm{Z})\right]$.

- *If the variance of the loss function is bounded (i.e.,* $\mathsf{Var}\left(\ell(\mathrm{W}_T;\mathrm{Z})\right) < \infty$*), then*

$$\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right] \leq \frac{\sigma}{n}\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[\exp\left(4\left\|g(\mathrm{W}_{t-1},\mathrm{Z}) - \boldsymbol{e}\right\|_2^2\right) - 1\right]\cdot q^{T-t}}, \quad (44)$$

where $\sigma = \sqrt{\mathsf{Var}\left(\ell(\mathrm{W}_T;\mathrm{Z})\right)}$ *and* $\boldsymbol{e} \triangleq \mathbb{E}\left[g(\mathrm{W}_{t-1},\mathrm{Z})\right]$.

Proof See Appendix D.1. ■

Our Theorem 10 leads to three generalization bounds for each DP-SGD mechanism. We discuss the advantage of each bound in the following remark by focusing on the Gaussian mechanism.

Remark 13 *We first assume that the loss function is upper bounded by* A*, leading to an* $A/2$*-sub-Gaussian loss* $\ell(\boldsymbol{w},\mathrm{Z})$ *and* $\sqrt{\mathsf{Var}\left(\ell(\mathrm{W}_T;\mathrm{Z})\right)} \leq A/2$*. Since*

$$\mathbb{E}\left[\mathrm{X}\right] \leq \sqrt{\mathbb{E}\left[\mathrm{X}^2\right]} \leq \frac{1}{2}\sqrt{\mathbb{E}\left[\exp(4\mathrm{X}^2) - 1\right]},$$

then for $\boldsymbol{e} \triangleq \mathbb{E}\left[g(\mathrm{W}_{t-1},\mathrm{Z})\right]$

$$\mathbb{E}\left[\left\|g(\mathrm{W}_{t-1},\mathrm{Z}) - \boldsymbol{e}\right\|_2\right] \leq \sqrt{\mathsf{Var}\left(g(\mathrm{W}_{t-1},\mathrm{Z})\right)} \leq \frac{1}{2}\sqrt{\mathbb{E}\left[\exp\left(4\left\|g(\mathrm{W}_{t-1},\mathrm{Z}) - \boldsymbol{e}\right\|_2^2\right) - 1\right]}.$$

Therefore, we have

$$(43) \leq (42) \leq (44).$$

In other words, the total-variation bound in (35) *yields the tightest generalization bound* (43) *for the DP-SGD algorithm. On the other hand, the* χ^2*-divergence bound in* (36) *leads to a bound* (44) *that requires the mildest assumption. At this moment, it seems unclear what the advantage of the KL-divergence bound is. Nonetheless, we will show in Section 5.3 that the nice properties of mutual information (e.g., chain rule) help extend our analysis to the general setting where data are drawn with replacement.*

A standard approach (see e.g., He et al., 2021) for analyzing the generalization of the DP-SGD algorithm often follows two steps: establish (ϵ,δ)-differential privacy guarantees for the DP-SGD algorithm and prove/apply a generalization bound that holds for *any* (ϵ,δ)-differentially private algorithms. However, generalization bounds obtained in this manner are distribution-independent since differential privacy is robust with respect to the data distribution. As observed in existing literature (see e.g., Zhang et al., 2017a) and our Figure 1, machine learning models trained under different data distributions can exhibit completely different generalization behaviors. Our bounds take into account the data distribution through the expectation terms (or mmae, variance).

Our generalization bounds can be estimated from data. Take the bound in (42) as an example. If sufficient data are available at each iteration, we can estimate the variance term by the population variance of $\{g(\mathrm{W}_{t-1},\mathrm{Z}_i) \mid i \in \mathcal{B}_t\}$ since W_{t-1} is independent of Z_i for $i \in \mathcal{B}_t$. Alternatively, we can draw a hold-out set for estimating the variance term at each iteration.

5.2 Federated Learning (FL)

Federated learning (FL) (McMahan et al., 2017) is a setting where a model is trained across multiple clients (e.g., mobile devices) under the management of a central server while the training data are kept decentralized. We recall the federated averaging algorithm with local-update DP-SGD in Algorithm 1 and refer the readers to Kairouz et al. (2021) for a more comprehensive review.

Algorithm 1 Federated averaging (local DP-SGD).

Input:
Total number of clients N and clients per round C
Total global updates T and local updates M
DP-SGD learning rate η
Initialize: W_0 randomly selected from $\mathcal{W}$
for $t = 1, \cdots, T$ global steps **do**
  Server chooses a subset $\mathcal{S}_t$ of C clients
  Server sends W_{t-1} to all selected clients
  for each client $k \in \mathcal{S}_t$ in parallel **do**
    Initialize $\mathrm{W}_{t,0}^k \leftarrow \mathrm{W}_{t-1}$
    for $j = 1, \cdots, M$ local steps **do**
      Draw b fresh data points $\{\mathrm{Z}_i^k\}_{i\in[b]}$ and noise $\mathrm{N}_{t,j}^k \sim N(0, \mathbf{I}_d)$
      Update the parameter $\mathrm{W}_{t,j}^k \leftarrow \mathsf{Proj}_{\mathcal{W}}\left(\mathrm{W}_{t,j-1}^k - \eta\left(g\left(\mathrm{W}_{t,j-1}^k, \{\mathrm{Z}_i^k\}_{i\in[b]}\right) + \mathrm{N}_{t,j}^k\right)\right)$
    end for
    Send $\mathrm{W}_{t,M}^k$ back to the server
  end for
  Server aggregates the parameter $\mathrm{W}_t = \frac{1}{C}\sum_{k\in\mathcal{S}_t} \mathrm{W}_{t,M}^k$
end for
Output: W_T

It is crucial to be able to *monitor* the performance of the global model on each client. Although the global model could achieve a desirable performance on average, it may fail to achieve high accuracy for each local client. This is because in the federated learning setting, data are typically unbalanced (different clients own different number of samples) and not identically distributed (data distribution varies across different clients). Since in practice clients may not have an extra hold-out data set to evaluate the performance of the global model, they can instead compute the loss of the model on their training set and compensate the mismatch by the generalization gap (or its upper bound). It is worth noting that this approach of monitoring model performance is completely decentralized as the clients do not need to share their data with the server and all the computation can be done locally. As discussed in Remark 13, the total variation bound in (35) often leads to the tightest generalization bound so we recast it under the setting of FL.

Proposition 14 *Let $\mathcal{T}_k \subset [T]$ contain the indices of global iterations in which the k-th client interacts with the server. If the loss function is upper bounded by $A > 0$, the expected generalization gap of the k-th client has an upper bound:*

$$\mathbb{E}\left[L_{\mu_k}(\mathrm{W}_T) - L_{\mathrm{S}_k}(\mathrm{W}_T)\right] \leq \frac{A}{n_k}\sum_{t\in\mathcal{T}_k}\sum_{j=1}^{M}\mathbb{E}\left[\|g(\mathrm{W}_{t,j-1}^k, \mathrm{Z}^k) - \boldsymbol{e}\|_2\right]\cdot q^{M(T+1-t)-j},$$

where n_k is the number of training data from the k-th client, $\boldsymbol{e} \triangleq \mathbb{E}\left[g(\mathrm{W}_{t,j-1}^k, \mathrm{Z}^k)\right]$, and

$$q \triangleq 1 - 2\bar{\Phi}\left(\frac{\sqrt{C}(D + 2\eta K)}{2\eta}\right) \in (0,1)$$

with D being the diameter of $\mathcal{W}$, $K \triangleq \sup_{\boldsymbol{w},\boldsymbol{z}} \|g(\boldsymbol{w},\boldsymbol{z})\|_2$, and $\bar{\Phi}(\cdot)$ being the Gaussian CCDF.

Proof See Appendix D.2. ■

Yagli et al. (2020) introduced a generalization bound in the context of federated learning. However, their bound in Theorem 3 involves a mutual information. Here we replace the mutual information with an expectation term. This improvement allows local clients to compute our bound from their training data reliably.

5.3 Stochastic Gradient Langevin Dynamics (SGLD)

We analyze the generalization gap of the stochastic gradient Langevin dynamics (SGLD) algorithm (Gelfand and Mitter, 1991; Welling and Teh, 2011). We start by recalling a standard framework of SGLD. The data set S is first divided into m disjoint mini-batches:

$$\mathrm{S} = \bigcup_{j=1}^{m} \mathrm{S}_j, \quad \text{where } |\mathrm{S}_j| = b \text{ and } \mathrm{S}_j \cap \mathrm{S}_k = \emptyset \text{ for } j \neq k.$$

We initialize the parameter of the empirical risk with a random point $\mathrm{W}_0 \in \mathcal{W}$ and update using the following rule:

$$\mathrm{W}_t = \mathrm{W}_{t-1} - \eta_t \nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \mathrm{S}_{B_t}) + \sqrt{\frac{2\eta_t}{\beta_t}} \mathrm{N}_t, \tag{45}$$

where η_t is the learning rate; β_t is the inverse temperature; N_t is drawn independently from a standard Gaussian distribution; $B_t \in [m]$ is the mini-batch index; $\hat{\ell}$ is a surrogate loss (e.g., hinge loss); and

$$\nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \mathrm{S}_{B_t}) \triangleq \frac{1}{b} \sum_{\mathrm{Z} \in \mathrm{S}_{B_t}} \nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \mathrm{Z}). \tag{46}$$

We study a general setting where the output from SGLD can be any function of the parameters across all iterations (i.e., $\mathrm{W} = f(\mathrm{W}_1, \cdots, \mathrm{W}_T)$), including the setting considered before where $\mathrm{W} = \mathrm{W}_T$. For example, the output can be an average of all iterates (i.e., Polyak averaging) $\mathrm{W} = \frac{1}{T}\sum_t \mathrm{W}_t$ or the parameter which achieves the smallest value of the loss function $\mathrm{W} = \operatorname{argmin}_{\mathrm{W}_t} L_\mu(\mathrm{W}_t)$.

Alas, Theorem 10 cannot be applied directly to the SGLD algorithm because the Markov chain in (26) does not hold any more when data are drawn with replacement. In order to circumvent this issue, we develop a different proof technique by using the chain rule for mutual information.

Proposition 15 *If the loss function $\ell(\boldsymbol{w}, \mathrm{Z})$ is σ-sub-Gaussian under $\mathrm{Z} \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, then*

$$\mathbb{E}\left[L_\mu(\mathrm{W}) - L_\mathrm{S}(\mathrm{W})\right] \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \mathrm{S}_j)\right)},$$

where the set $\mathcal{T}_j$ contains the indices of iterations in which the mini-batch S_j is used.

Proof See Appendix D.3. ■

Our bound incorporates the gradient variance which measures a particular kind of "flatness" of the loss landscape. We note that a recent work (Jiang et al., 2020) has observed empirically that the variance of gradients is predictive of and highly correlated with the generalization gap of neural networks. Here we evidence this connection from a theoretical viewpoint by incorporating the gradient variance into the generalization bound.

Unfortunately, our generalization bound does not incorporate a decay factor anymore.[4] To understand why it happens, let us imagine an extreme scenario in which the SGLD algorithm outputs all the iterates (i.e., $\mathrm{W} = (\mathrm{W}_1, \cdots, \mathrm{W}_T)$). For a data point Z_i used at the t-th iteration, the data processing inequality implies that

$$I(\mathrm{W}_1, \cdots, \mathrm{W}_T; \mathrm{Z}_i) \geq I(\mathrm{W}_t; \mathrm{Z}_i).$$

Hence, it is impossible to have $I(\mathrm{W}_1, \cdots, \mathrm{W}_T; \mathrm{Z}_i) \to 0$ as $T \to \infty$ unless $I(\mathrm{W}_t; \mathrm{Z}_i) = 0$.

Many existing SGLD generalization bounds (e.g., Mou et al., 2018; Li et al., 2020; Pensia et al., 2018; Negrea et al., 2019) are expressed as a sum of errors associated with each training iteration. In order to compare with these results, we present an analogous bound in the following corollary. This bound is obtained by combining a key lemma for proving Proposition 15 with Minkowski inequality and Jensen's inequality so it is often much weaker than Proposition 15.

Corollary 16 *If the loss function $\ell(\boldsymbol{w}, \mathrm{Z})$ is σ-sub-Gaussian under $\mathrm{Z} \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, the expected generalization gap of the SGLD algorithm can be upper bounded by*

$$\frac{\sqrt{2}\sigma}{2} \min\left\{ \frac{1}{n} \sum_{t=1}^{T} \sqrt{\beta_t \eta_t \cdot \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \mathrm{Z}_t^\dagger)\right)}, \sqrt{\frac{1}{bn} \sum_{t=1}^{T} \beta_t \eta_t \cdot \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \mathrm{Z}_t^\dagger)\right)} \right\},$$

where $\mathrm{Z}_t^\dagger$ is any data point used in the t-th iteration.

Proof See Appendix D.4. ■

Our bound is distribution-dependent through the variance of gradients in contrast with Corollary 1 of Pensia et al. (2018), Proposition 3 of Bu et al. (2020), and Theorem 1 of Mou et al. (2018), which rely on the Lipschitz constant: $\sup_{\boldsymbol{w}, \boldsymbol{z}} \|\nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}, \boldsymbol{z})\|_2$. These bounds fail

4. We note that the analysis in Mou et al. (2018) requires $\mathrm{W} = \mathrm{W}_T$. Hence, in the setting we consider (i.e., W is a function of $\mathrm{W}_1, \cdots, \mathrm{W}_T$), it is unclear if it is possible to include a decay factor in the bound.

to explain some generalization phenomena of DNNs, such as label corruption (Zhang et al., 2017a), because the Lipschitz constant takes a supremum over all possible weight matrices $\boldsymbol{w}$ and data points $\boldsymbol{z}$. In other words, this Lipschitz constant only relies on the architecture of the network instead of the weight matrices or data distribution. Hence, it is the same for a network trained from corrupted data and a network trained from true data. We remark that the Lipschitz constant used by Pensia et al. (2018); Bu et al. (2020); Mou et al. (2018) is different from the Lipschitz constant of the function corresponding to a network w.r.t. the input variable. The latter one has been used in the literature (see e.g., Bartlett et al., 2017) for deriving generalization bounds and, to some degree, can capture generalization phenomena, such as label corruption.

The order of our generalization bound in Corollary 16 is $\min\left(\frac{1}{n}\sum_{t=1}^{T}\sqrt{\beta\eta_t}, \sqrt{\frac{\beta}{bn}\sum_{t=1}^{T}\eta_t}\right)$. It is tighter than Theorem 2 of Mou et al. (2018) whose order is $\sqrt{\frac{\beta}{n}\sum_{t=1}^{T}\eta_t}$. Our bound is applicable regardless of the choice of learning rate while the bound in Li et al. (2020) requires the scale of the learning rate to be upper bounded by the reciprocal of the Lipschitz constant. Our Corollary 16 has the same order with Negrea et al. (2019) but we incorporate an additional decay factor when applying our bounds to the DP-SGD algorithm (see Proposition 12) and numerical experiments suggest that our bound is more favourably correlated with the true generalization gap (see Table 2).

6. Numerical Experiments

In this section, we demonstrate our generalization bound (Proposition 15) through numerical experiments on the MNIST data set (LeCun et al., 1998), CIFAR-10 data set (Krizhevsky et al., 2009), and SVHN data set (Netzer et al., 2011), showing that it can predict the behavior of the true generalization gap.

6.1 Corrupted Labels

As observed in Zhang et al. (2017a), DNNs have the potential to memorize the entire training data set even when a large portion of the labels are corrupted. For networks with identical architecture, those trained using true labels have better generalization capability than those ones trained using corrupted labels, although both of them achieve perfect training accuracy. Unfortunately, distribution-independent bounds, such as the ones using VC-dimension, may not be able to capture this phenomenon because they are invariant for both true data and corrupted data. In contrast, our bound quantifies this empirical observation, exhibiting a lower value on networks trained on true labels compared to ones trained on corrupted labels (Figure 1).

In our experiment, we randomly select 5000 samples as our training data set and change the label of $\alpha \in \{0\%, 25\%, 50\%, 75\%\}$ of the training samples. Then we use the SGLD algorithm to train a neural network under different corruption level. The training process continues until the training accuracy is 1.0 (see Figure 1 Left). We compare our generalization bound with the generalization gap in Figure 1 Middle and Right. As shown, both our bound and the generalization gap are increasing w.r.t. the corruption level in the last epoch. Furthermore, the curve of our bound has very similar shape with the generalization gap. Finally, we observe that the generalization gap tends to be stable since the algorithm

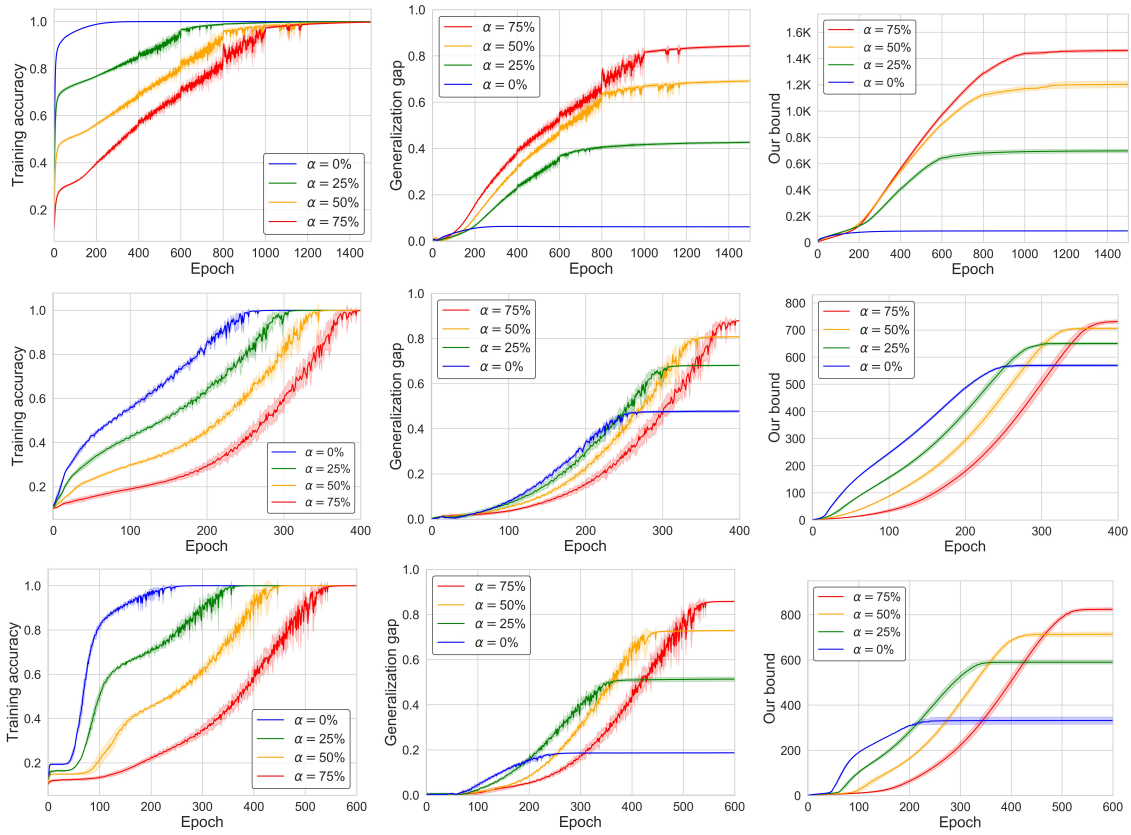

Figure 1: Illustration of our generalization bound in Proposition 15. We use the SGLD algorithm to train 3-layer neural networks on MNIST (top row) and convolutional neural networks on CIFAR-10 (middle row) and SVHN (bottom row) when the training data have different label corruption level $\alpha \in \{0\%, 25\%, 50\%, 75\%\}$. Left column: training accuracy. Middle column: empirical generalization gap. Right column: empirical generalization bound.

converges (Figure 1 Middle). Our generalization bound captures this phenomenon (Figure 1 Right) as the variance of gradients becomes negligible when the algorithm starts converging. The intuition is that the variance of gradients reflects the flatness of the loss landscape and as the algorithm converges, the loss landscape becomes flatter.

6.2 Network Width

As observed by several recent studies (see e.g., Neyshabur et al., 2015; Jiang et al., 2020), wider networks can lead to a smaller generalization gap. This may seem contradictory to the traditional wisdom as one may expect that a class of wider networks has a higher VC-dimension and, hence, would have a higher generalize gap. In our experiment, we use the SGLD algorithm to train neural networks with different widths. The training process runs for 400 epochs until the training accuracy is 1.0. We compare our generalization bound

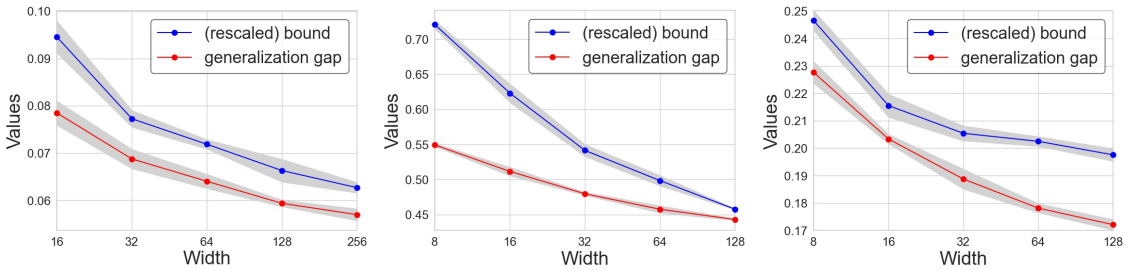

Figure 2: The generalization gap of neural networks with varying widths trained on MNIST (left), CIFAR-10 (middle), and SVHN (right) using the SGLD algorithm. To compare with, we plot the generalization bound (Proposition 15) rescaled by multiplying 7e-4 for MNIST, 9e-4 for CIFAR-10, and 7e-4 for SVHN.

data set	method	lr	width	depth	τ	Ψ	MI
MNIST	Ours (Proposition 15)	**0.70**	**1.00**	**0.56**	**0.50**	**0.75**	**0.34**
	Negrea et al. (2019)	0.26	0.26	0.48	0.25	0.33	0.12
CIFAR-10	Ours (Proposition 15)	**0.41**	**0.93**	**1.00**	**0.45**	**0.78**	**0.25**
	Negrea et al. (2019)	0.33	0.41	0.85	0.38	0.53	0.16

Table 2: We adopt the three evaluation criteria proposed in Jiang et al. (2020) for comparing our generalization bound with the benchmark method (Theorem 3.1 of Negrea et al., 2019): (i) Kendall's rank-correlation coefficient (τ), (ii) Granulated Kendall's coefficient (Ψ), and (iii) conditional independent test (MI). All scores, except MI, are within $[-1, 1]$ and the score of MI is normalized to $[0, 1]$. We also report the correlations when a single hyper-parameter (e.g., learning rate (lr)) is varying.

with the generalization gap in Figure 2. As shown, both the generalization gap and our bound are decreasing with respect to the network width.

6.3 Comparison with Benchmarks

To evaluate our bound, we adopt the three criteria proposed in Jiang et al. (2020): (i) Kendall's rank-correlation coefficient (τ) (Kendall, 1938), (ii) Granulated Kendall's coefficient (Ψ), and (iii) conditional independent test via mutual information (MI) (Verma and Pearl, 1991). In our experiments, we select 3 commonly used hyper-parameters (i.e., learning rate (lr), width, depth), which are believed to influence the generalization gap, and let each hyper-parameter choose three different values. We train 27 neural networks under all combinations of hyper-parameters and assess the correlations between the generalization bound and the generalization gap.

We compare our generalization bound with the gradient-prediction-residual bound in Theorem 3.1 of Negrea et al. (2019) under the above three evaluation criteria. As shown in

Table 2, our generalization bound is highly correlated with the true generalization gap and outperforms the benchmark under all the criteria suggested in Jiang et al. (2020).

7. Conclusion

In this paper, we investigate the generalization of noisy iterative algorithms and derive three generalization bounds based on different f-divergence. We establish a unified framework and leverage fundamental tools from information theory (e.g., strong data processing inequalities and properties of additive noise channels) for proving these bounds. We demonstrate our generalization bounds through applications, including DP-SGD, FL, and SGLD, in which our bounds own a simple form and can be estimated from data. Numerical experiments suggest that our bounds can predict the behavior of the true generalization gap.

Acknowledgments

The authors would like to thank the anonymous reviewers and the action editor for their careful reading of our paper and their valuable suggestions. The work of H. Wang and F. P. Calmon is supported in part by the National Science Foundation under grants CAREER 1845852, IIS 1926925, and FAI 2040880 and F. P. Calmon also acknowledges a Google Faculty Research Award and an Amazon Research Award.

Appendix A. Proofs for Section 2

A.1 Proof of Lemma 1

Inequalities (13) and (14) can be proved by combining the proof technique introduced in Bu et al. (2020) and variational representations of f-divergence (Nguyen et al., 2010). Note that these two bounds can also be obtained from Corollary 1 in Rodríguez Gálvez et al. (2021) and applying Jensen's inequality to Eq. (15) in the same paper, respectively. We provide the proof here for the sake of completeness.

Proof Recall that (see Example 6.1 and 6.4 in Wu, 2020) for any two probability distributions P and Q over a set $\mathcal{X} \subseteq \mathbb{R}^d$ and a constant $A > 0$,

$$\mathrm{D}_{\mathrm{TV}}(P\|Q) = \sup_{\substack{h:\mathcal{X}\to\mathbb{R} \\ 0\le\|h\|_\infty\le A}} \left|\frac{\mathbb{E}_P\left[h(\mathrm{X})\right] - \mathbb{E}_Q\left[h(\mathrm{X})\right]}{A}\right|, \tag{47}$$

$$\mathrm{D}_{\chi^2}(P\|Q) = \sup_{h:\mathcal{X}\to\mathbb{R}} \frac{\left(\mathbb{E}_P\left[h(\mathrm{X})\right] - \mathbb{E}_Q\left[h(\mathrm{X})\right]\right)^2}{\mathsf{Var}_Q\left(h(\mathrm{X})\right)}. \tag{48}$$

On the other hand, the expected generalization gap can be written as

$$\mathbb{E}\left[L_\mu(\mathrm{W}) - L_{\mathrm{S}}(\mathrm{W})\right] = \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}\left[\ell(\mathrm{W},\bar{\mathrm{Z}}_i)\right] - \mathbb{E}\left[\ell(\mathrm{W},\mathrm{Z}_i)\right]\right),$$

where $(\mathrm{W},\bar{\mathrm{Z}}_i) \sim P_{\mathrm{W}} \otimes \mu$. Consequently,

$$\left|\mathbb{E}\left[L_\mu(\mathrm{W}) - L_{\mathrm{S}}(\mathrm{W})\right]\right| \le \frac{1}{n}\sum_{i=1}^n \left|\mathbb{E}\left[\ell(\mathrm{W},\bar{\mathrm{Z}}_i)\right] - \mathbb{E}\left[\ell(\mathrm{W},\mathrm{Z}_i)\right]\right|.$$

If the loss function is upper bounded by $A > 0$, taking $P = P_{\mathrm{W},\mathrm{Z}_i}$ and $Q = P_{\mathrm{W}} \otimes \mu$ in (47) yields

$$\left|\mathbb{E}\left[L_\mu(\mathrm{W}) - L_{\mathrm{S}}(\mathrm{W})\right]\right| \le \frac{A}{n}\sum_{i=1}^n \mathrm{D}_{\mathrm{TV}}(P_{\mathrm{W},\mathrm{Z}_i}\|P_{\mathrm{W}} \otimes \mu) = \frac{A}{n}\sum_{i=1}^n \mathrm{T}(\mathrm{W};\mathrm{Z}_i).$$

Similarly, taking $P = P_{\mathrm{W},\mathrm{Z}_i}$ and $Q = P_{\mathrm{W}} \otimes \mu$ in (48) leads to

$$\left|\mathbb{E}\left[L_\mu(\mathrm{W}) - L_{\mathrm{S}}(\mathrm{W})\right]\right| \le \frac{1}{n}\sum_{i=1}^n \sqrt{\mathsf{Var}\left(\ell(\mathrm{W};\mathrm{Z})\right) \cdot \chi^2(\mathrm{W};\mathrm{Z}_i)}$$

where $(\mathrm{W},\mathrm{Z}_i) \sim P_{\mathrm{W},\mathrm{Z}_i}$ and $(\mathrm{W},\mathrm{Z}) \sim P_{\mathrm{W}} \otimes \mu$. ■

Appendix B. Proofs for Section 3

B.1 Proof of Lemma 6

Lemma 6 follows from a slight tweak of the proof of Theorem 4 in Polyanskiy and Wu (2016).

Proof First, we choose a coupling $P_{\mathrm{X},\mathrm{X}'}$, which has marginals P_{X} and $P_{\mathrm{X}'}$. The probability distribution of $\mathrm{X}+m\mathrm{N}$ can be written as the convolution of P_{X} and $P_{m\mathrm{N}}$. Specifically, for any measurable set $\mathcal{A} \subseteq \mathcal{X}$,

$$P_{\mathrm{X}+m\mathrm{N}}(\mathcal{A}) = \int_{\boldsymbol{y}\in\mathcal{A}} \int_{\boldsymbol{x}\in\mathcal{X}} \mathrm{d}P_{m\mathrm{N}}(\boldsymbol{y}-\boldsymbol{x})\mathrm{d}P_{\mathrm{X}}(\boldsymbol{x}) = \int_{\boldsymbol{y}\in\mathcal{A}} \int_{\boldsymbol{x}\in\mathcal{X}} \int_{\boldsymbol{x}'\in\mathcal{X}} \mathrm{d}P_{\boldsymbol{x}+m\mathrm{N}}(\boldsymbol{y})\mathrm{d}P_{\mathrm{X},\mathrm{X}'}(\boldsymbol{x},\boldsymbol{x}').$$

Similarly, we have

$$P_{\mathrm{X}'+m\mathrm{N}}(\mathcal{A}) = \int_{\boldsymbol{y}\in\mathcal{A}} \int_{\boldsymbol{x}\in\mathcal{X}} \int_{\boldsymbol{x}'\in\mathcal{X}} \mathrm{d}P_{\boldsymbol{x}'+m\mathrm{N}}(\boldsymbol{y})\mathrm{d}P_{\mathrm{X},\mathrm{X}'}(\boldsymbol{x},\boldsymbol{x}').$$

Since the mapping $(P,Q) \to \mathrm{D}_f(P\|Q)$ is convex (see Theorem 6.1 in Polyanskiy and Wu, 2019, for a proof), Jensen's inequality yields

$$\begin{aligned}\mathrm{D}_f\left(P_{\mathrm{X}+m\mathrm{N}}\|P_{\mathrm{X}'+m\mathrm{N}}\right) &\le \int_{\boldsymbol{x}\in\mathcal{X}} \int_{\boldsymbol{x}'\in\mathcal{X}} \mathrm{D}_f(P_{\boldsymbol{x}+m\mathrm{N}}\|P_{\boldsymbol{x}'+m\mathrm{N}})\mathrm{d}P_{\mathrm{X},\mathrm{X}'}(\boldsymbol{x},\boldsymbol{x}') \\ &= \mathbb{E}\left[\mathsf{C}_f(\mathrm{X},\mathrm{X}';m)\right], \end{aligned} \tag{49}$$

where the last step follows from the definition. The left-hand side of (49) only relies on the marginal distributions of X and X′, so taking the infimum on both sides of (49) over all couplings $P_{\mathrm{X},\mathrm{X}'}$ leads to the desired conclusion. ■

B.2 A Useful Property

We derive a useful property of $\mathsf{C}_f(\boldsymbol{x},\boldsymbol{y};m)$ which will be used in our proofs.

Lemma 17 *For any $\boldsymbol{z} \in \mathbb{R}^d$ and $a > 0$, the function $\mathsf{C}_f(\boldsymbol{x},\boldsymbol{x}';m)$ in (20) satisfies*

$$\mathsf{C}_f(a\boldsymbol{x}+\boldsymbol{z}, a\boldsymbol{x}'+\boldsymbol{z}; m) = \mathsf{C}_f\left(\boldsymbol{x},\boldsymbol{x}';\frac{m}{a}\right), \quad \mathsf{C}_f(\boldsymbol{x},\boldsymbol{x}';m) = \mathsf{C}_f(-\boldsymbol{x}',-\boldsymbol{x};m).$$

Proof For simplicity, we assume that N is a continuous random variable in $\mathbb{R}^d$ with probability density function (PDF) $p(\boldsymbol{w})$. Then the PDFs of $a\boldsymbol{x}+\boldsymbol{z}+m\mathrm{N}$ and $a\boldsymbol{x}'+\boldsymbol{z}+m\mathrm{N}$ are

$$\frac{1}{m^d}\cdot p\left(\frac{\boldsymbol{w}-a\boldsymbol{x}-\boldsymbol{z}}{m}\right) \quad \text{and} \quad \frac{1}{m^d}\cdot p\left(\frac{\boldsymbol{w}-a\boldsymbol{x}'-\boldsymbol{z}}{m}\right).$$

By definition,

$$\begin{aligned}\mathsf{C}_f(a\boldsymbol{x}+\boldsymbol{z}, a\boldsymbol{x}'+\boldsymbol{z}; m) &= \mathrm{D}_f(P_{a\boldsymbol{x}+\boldsymbol{z}+m\mathrm{N}}\|P_{a\boldsymbol{x}'+\boldsymbol{z}+m\mathrm{N}}) \\ &= \frac{1}{m^d}\int_{\mathbb{R}^d} p\left(\frac{\boldsymbol{w}-a\boldsymbol{x}'-\boldsymbol{z}}{m}\right) f\left(\frac{p\left(\frac{\boldsymbol{w}-a\boldsymbol{x}-\boldsymbol{z}}{m}\right)}{p\left(\frac{\boldsymbol{w}-a\boldsymbol{x}'-\boldsymbol{z}}{m}\right)}\right)\mathrm{d}\boldsymbol{w}. \end{aligned} \tag{50}$$

Let $\boldsymbol{v} = (\boldsymbol{w}-\boldsymbol{z})/a$. Then (50) is equal to

$$\frac{a^d}{m^d}\int_{\mathbb{R}^d} p\left(\frac{\boldsymbol{v}-\boldsymbol{x}'}{m/a}\right) f\left(\frac{p\left(\frac{\boldsymbol{v}-\boldsymbol{x}}{m/a}\right)}{p\left(\frac{\boldsymbol{v}-\boldsymbol{x}'}{m/a}\right)}\right)\mathrm{d}\boldsymbol{v} = \mathsf{C}_f\left(\boldsymbol{x},\boldsymbol{x}';\frac{m}{a}\right).$$

Therefore, $\mathsf{C}_f(a\boldsymbol{x}+\boldsymbol{z}, a\boldsymbol{x}'+\boldsymbol{z}; m) = \mathsf{C}_f\left(\boldsymbol{x}, \boldsymbol{x}'; \frac{m}{a}\right)$. By choosing $a=1$ and $\boldsymbol{z} = -\boldsymbol{x}-\boldsymbol{x}'$, we have $\mathsf{C}_f(-\boldsymbol{x}', -\boldsymbol{x}; m) = \mathsf{C}(\boldsymbol{x}, \boldsymbol{x}'; m)$. ■

B.3 Proof of Table 1

We first derive closed-form expressions (or upper bounds) for the function $\delta(A, m)$. The closed-form expression of $\delta(A, m)$ for uniform distribution can be naturally obtained from its definition so we skip the proof. The closed-form expressions for standard multivariate Gaussian distribution and standard univariate Laplace distribution can be found at Polyanskiy and Wu (2016) and Asoodeh et al. (2020). In what follows, we provide an upper bound for $\delta(A, m)$ when N follows a standard multivariate Laplace distribution.

Proof For a given positive number A and a random variable N which follows a standard multivariate Laplace distribution, consider the following optimization problem:

$$\sup_{\|\boldsymbol{v}\|_1 \le A} \mathrm{D}_{\mathrm{TV}}\left(P_{\mathrm{N}} \| P_{\boldsymbol{v}+\mathrm{N}}\right) = \sup_{\|\boldsymbol{v}\|_1 \le A} \mathbb{E}\left[\left(1 - \frac{\exp(-\|\mathrm{N}-\boldsymbol{v}\|_1)}{\exp(-\|\mathrm{N}\|_1)}\right) \mathbb{I}_{\|\mathrm{N}-\boldsymbol{v}\|_1 \ge \|\mathrm{N}\|_1}\right].$$

By exchanging the supremum and the expectation, we have

$$\sup_{\|\boldsymbol{v}\|_1 \le A} \mathrm{D}_{\mathrm{TV}}\left(P_{\mathrm{N}} \| P_{\boldsymbol{v}+\mathrm{N}}\right) \le \mathbb{E}\left[\sup_{\|\boldsymbol{v}\|_1 \le A}\left\{\left(1 - \frac{\exp(-\|\mathrm{N}-\boldsymbol{v}\|_1)}{\exp(-\|\mathrm{N}\|_1)}\right) \mathbb{I}_{\|\mathrm{N}-\boldsymbol{v}\|_1 \ge \|\mathrm{N}\|_1}\right\}\right]. \tag{51}$$

Note that

$$\begin{aligned}
&\sup_{\|\boldsymbol{v}\|_1 \le A}\left\{\left(1 - \frac{\exp(-\|\mathrm{N}-\boldsymbol{v}\|_1)}{\exp(-\|\mathrm{N}\|_1)}\right) \mathbb{I}_{\|\mathrm{N}-\boldsymbol{v}\|_1 \ge \|\mathrm{N}\|_1}\right\} \\
&= 1 - \exp\left(-\sup_{\|\boldsymbol{v}\|_1 \le A}\left\{\|\mathrm{N}-\boldsymbol{v}\|_1 - \|\mathrm{N}\|_1\right\}\right) = 1 - \exp(-A).
\end{aligned}$$

Substituting this equality into (51) gives

$$\sup_{\|\boldsymbol{x}-\boldsymbol{x}'\|_1 \le A} \mathrm{D}_{\mathrm{TV}}\left(P_{\boldsymbol{x}+\mathrm{N}} \| P_{\boldsymbol{x}'+\mathrm{N}}\right) = \sup_{\|\boldsymbol{v}\|_1 \le A} \mathrm{D}_{\mathrm{TV}}\left(P_{\mathrm{N}} \| P_{\boldsymbol{v}+\mathrm{N}}\right) \le 1 - \exp(-A),$$

which leads to $\delta(A, 1) \le 1 - \exp(-A)$. Finally, we have

$$\delta(A, m) = \delta\left(\frac{A}{m}, 1\right) \le 1 - \exp\left(-\frac{A}{m}\right).$$

■

Now we consider the function $\mathsf{C}_f(\boldsymbol{x}, \boldsymbol{x}'; m)$.

Proof By Lemma 17, we have

$$\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}'; m) = \mathsf{C}_{\mathrm{KL}}\left(0, \frac{\boldsymbol{x}'-\boldsymbol{x}}{m}; 1\right).$$

We denote $(\boldsymbol{x}' - \boldsymbol{x})/m$ by $\boldsymbol{v}$. Since all the coordinates of $\mathrm{N} = (\mathrm{N}_1 \cdots, \mathrm{N}_d)$ are mutually independent, $P_{\mathrm{N}} = P_{\mathrm{N}_1} \cdots P_{\mathrm{N}_d}$ and $P_{\boldsymbol{v}+\mathrm{N}} = P_{v_1+\mathrm{N}_1} \cdots P_{v_d+\mathrm{N}_d}$. By the chain rule of KL-divergence, we have

$$\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}'; m) = \mathrm{D}_{\mathrm{KL}}(P_{\mathrm{N}} \| P_{\boldsymbol{v}+\mathrm{N}}) = \sum_{i=1}^{d} \mathrm{D}_{\mathrm{KL}}(P_{\mathrm{N}_i} \| P_{v_i+\mathrm{N}_i}). \tag{52}$$

Hence, we only need to calculate $\mathrm{D}_{\mathrm{KL}}(P_{\mathrm{N}} \| P_{v+\mathrm{N}})$ for a constant $v \in \mathbb{R}$ and a random variable $\mathrm{N} \in \mathbb{R}$.
(1) If N follows a standard Gaussian distribution, then

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(P_{\mathrm{N}} \| P_{v+\mathrm{N}}) &= \mathbb{E}\left[\log \frac{\exp(-\mathrm{N}^2/2)}{\exp(-(\mathrm{N}-v)^2/2)}\right] \\ &= \frac{1}{2}\mathbb{E}\left[(\mathrm{N}-v)^2 - \mathrm{N}^2\right] = \frac{v^2}{2}.\end{aligned}$$

Substituting this equality into (52) gives

$$\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}'; m) = \frac{\|\boldsymbol{v}\|_2^2}{2} = \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2m^2},$$

where the last step is due to the definition of $\boldsymbol{v}$.
(2) If N follows a standard Laplace distribution, then

$$\mathrm{D}_{\mathrm{KL}}(P_{\mathrm{N}} \| P_{v+\mathrm{N}}) = \mathbb{E}\left[\log \frac{\exp(-|\mathrm{N}|)}{\exp(-|\mathrm{N}-v|)}\right] = |v| + \exp(-|v|) - 1.$$

Substituting this equality into (52) gives

$$\begin{aligned}\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}'; m) &= \sum_{i=1}^{d} |v_i| + \exp(-|v_i|) - 1 = \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_1}{m} + \sum_{i=1}^{d}\left(\exp\left(-\frac{|x_i - x_i'|}{m}\right) - 1\right) \\ &\leq \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_1}{m}.\end{aligned}$$

Similarly, by Lemma 17, we have

$$\mathsf{C}_{\chi^2}(\boldsymbol{x}, \boldsymbol{x}'; m) = \mathsf{C}_{\chi^2}\left(0, \frac{\boldsymbol{x}' - \boldsymbol{x}}{m}; 1\right).$$

We denote $(\boldsymbol{x}' - \boldsymbol{x})/m$ by $\boldsymbol{v}$. By the property of χ^2-divergence (see Section 2.4 in Tsybakov, 2009), we have

$$\mathsf{C}_{\chi^2}(\boldsymbol{x}, \boldsymbol{x}'; m) = \mathrm{D}_{\chi^2}(P_{\mathrm{N}} \| P_{\boldsymbol{v}+\mathrm{N}}) = \prod_{i=1}^{d}\left(1 + \mathrm{D}_{\chi^2}(P_{\mathrm{N}_i} \| P_{v_i+\mathrm{N}_i})\right) - 1. \tag{53}$$

Hence, we only need to calculate $\mathrm{D}_{\chi^2}(P_{\mathrm{N}} \| P_{v+\mathrm{N}})$ for $v \in \mathbb{R}$ and $\mathrm{N} \in \mathbb{R}$.
(1) If N follows a standard Gaussian distribution, then

$$\begin{aligned}\mathrm{D}_{\chi^2}(P_{\mathrm{N}} \| P_{v+\mathrm{N}}) &= \mathbb{E}\left[\frac{\exp(-\mathrm{N}^2/2)}{\exp(-(\mathrm{N}-v)^2/2)}\right] - 1 \\ &= \exp(v^2/2)\mathbb{E}\left[\exp(-vN)\right] - 1 = \exp(v^2) - 1.\end{aligned}$$

Substituting this equality into (53) gives

$$\mathsf{C}_{\chi^2}(\boldsymbol{x},\boldsymbol{x}';m)=\exp(\|\boldsymbol{v}\|_2^2)-1=\exp\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{m^2}\right)-1.$$

(2) If N follows a standard Laplace distribution, then

$$\begin{aligned}\mathrm{D}_{\chi^2}(P_{\mathrm{N}}\|P_{v+\mathrm{N}})&=\mathbb{E}\left[\frac{\exp(-|\mathrm{N}|)}{\exp(-|\mathrm{N}-v|)}\right]-1\\&=\frac{2}{3}\exp(|v|)+\frac{1}{3}\exp(-2|v|)-1.\end{aligned}$$

Substituting this equality into (53) gives

$$\begin{aligned}\mathsf{C}_{\chi^2}(\boldsymbol{x},\boldsymbol{x}';m)&=\prod_{i=1}^{d}\left(\frac{2}{3}\exp\left(\frac{|x_i-x_i'|}{m}\right)+\frac{1}{3}\exp\left(\frac{-2|x_i-x_i'|}{m}\right)\right)-1\\&\le\exp\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_1}{m}\right)-1.\end{aligned}$$

Finally, we use Pinsker's inequality (see Theorem 4.5 in Wu, 2020, for a proof) for proving an upper bound of $\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x},\boldsymbol{x}';m)$:

$$\begin{aligned}\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x},\boldsymbol{x}';m)&=\mathrm{D}_{\mathrm{TV}}\left(P_{\boldsymbol{x}+m\mathrm{N}}\|P_{\boldsymbol{x}'+m\mathrm{N}}\right)\\&\le\sqrt{\frac{\mathrm{D}_{\mathrm{KL}}\left(P_{\boldsymbol{x}+m\mathrm{N}}\|P_{\boldsymbol{x}'+m\mathrm{N}}\right)}{2}}\\&=\sqrt{\frac{\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x},\boldsymbol{x}';m)}{2}}.\end{aligned}$$

Hence, any upper bound of $\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x},\boldsymbol{x}';m)$ can be naturally translated into an upper bound for $\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x},\boldsymbol{x}';m)$. This is how we obtain the upper bounds of $\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x},\boldsymbol{x}';m)$ under Gaussian or Laplace distribution in Table 1. On the other hand, if N follows a uniform distribution on $[-1,1]\subseteq\mathbb{R}$, by Lemma 17 we have

$$\mathsf{C}_{\mathrm{TV}}(x,x';m)=\mathsf{C}_{\mathrm{TV}}\left(0,\frac{x'-x}{m};1\right)=\min\left\{1,\left|\frac{x-x'}{2m}\right|\right\}.$$

Note that in this case $\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}$ so we write them as x,x'. ■

Remark 18 *We used Pinsker's inequality for deriving an upper bound of $\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x},\boldsymbol{x}';m)$ in the above proof. One can potentially tighten this bound by exploring other f-divergence inequalities (see e.g., Eq. 4 in Sason and Verdú, 2016).*

Appendix C. Proofs for Section 4

C.1 Proof of Theorem 10

Proof Combining Lemma 8 and 9 together leads to an upper bound of $I_f(\mathrm{W}_T;\mathrm{Z}_i)$ for any data point Z_i used at the t-th iteration:

$$I_f(\mathrm{W}_T;\mathrm{Z}_i) \leq \mathbb{E}\left[\mathsf{C}_f\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); \frac{m_t b_t}{\eta_t}\right)\right] \cdot \prod_{t'=t+1}^{T} \delta(D+2\eta_{t'}K, m_{t'}). \tag{54}$$

Additionally, if the loss $\ell(\boldsymbol{w},\mathrm{Z})$ is σ-sub-Gaussian for all $\boldsymbol{w}\in\mathcal{W}$, Lemma 1 and Assumption 1 (Sampling w/o Replacement) altogether yield

$$\begin{aligned}\left|\mathbb{E}\left[L_\mu(\mathrm{W}_T) - L_\mathrm{S}(\mathrm{W}_T)\right]\right| &\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(\mathrm{W}_T;\mathrm{Z}_i)} \\ &= \frac{1}{n}\sum_{t=1}^{T}\sum_{i\in\mathcal{B}_t}\sqrt{2\sigma^2 I(\mathrm{W}_T;\mathrm{Z}_i)}. \end{aligned}\tag{55}$$

Substituting (54) into (55) yields the following upper bound of the expected generalization gap:

$$\begin{aligned}&\frac{\sqrt{2}\sigma}{n}\sum_{t=1}^{T}\sum_{i\in\mathcal{B}_t}\sqrt{\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); \frac{m_t b_t}{\eta_t}\right)\right] \cdot \prod_{t'=t+1}^{T} \delta(D+2\eta_{t'}K, m_{t'})} \\ &= \frac{\sqrt{2}\sigma}{n}\sum_{t=1}^{T} b_t \sqrt{\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); \frac{m_t b_t}{\eta_t}\right)\right] \cdot \prod_{t'=t+1}^{T} \delta(D+2\eta_{t'}K, m_{t'})}.\end{aligned}$$

Similarly, we can obtain another two generalization bounds using Lemma 1 and the upper bound in (54). ■

Appendix D. Proofs for Section 5

D.1 Proof of Proposition 11 and 12

In the setting of DP-SGD, the three generalization bounds in Theorem 10 become

$$\frac{\sqrt{2}\sigma}{n}\sum_{t=1}^{T} b_t \sqrt{\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); b_t\right)\right] \cdot \left(\delta(D+2\eta K,\eta)\right)^{T-t}} \tag{56}$$

where σ is the sub-Gaussian constant;

$$\frac{A}{n}\sum_{t=1}^{T} b_t \mathbb{E}\left[\mathsf{C}_{\mathrm{TV}}\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); b_t\right)\right] \cdot \left(\delta(D+2\eta K,\eta)\right)^{T-t} \tag{57}$$

where A is an upper bound of the loss function; and

$$\frac{\sigma}{n}\sum_{t=1}^{T} b_t \sqrt{\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); b_t\right)\right] \cdot \left(\delta(D+2\eta K,\eta)\right)^{T-t}} \tag{58}$$

where $\sigma \triangleq \sqrt{\mathsf{Var}\left(\ell(\mathrm{W}_T;\mathrm{Z})\right)}$.

Proof We prove Proposition 12 first.

- If the additive noise follows a standard multivariate Gaussian distribution, Table 1 shows that

$$\delta(D+2\eta K,\eta) = 1 - 2\bar{\Phi}\left(\frac{D+2\eta K}{2\eta}\right), \tag{59}$$

$$\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); b_t\right)\right] = \frac{1}{2b_t^2}\mathbb{E}\left[\|g(\mathrm{W}_{t-1},\mathrm{Z}) - g(\mathrm{W}_{t-1},\bar{\mathrm{Z}})\|_2^2\right]. \tag{60}$$

We introduce a constant vector $\boldsymbol{e}$ whose value will be specified later. Since $\|\boldsymbol{a}-\boldsymbol{b}\|_2^2 \le 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$, we have

$$\begin{aligned}
&\frac{1}{2b_t^2}\mathbb{E}\left[\|g(\mathrm{W}_{t-1},\mathrm{Z}) - g(\mathrm{W}_{t-1},\bar{\mathrm{Z}})\|_2^2\right] \\
&\le \frac{1}{b_t^2}\left(\mathbb{E}\left[\|g(\mathrm{W}_{t-1},\mathrm{Z}) - \boldsymbol{e}\|_2^2\right] + \mathbb{E}\left[\|g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}) - \boldsymbol{e}\|_2^2\right]\right) \\
&= \frac{2}{b_t^2}\mathbb{E}\left[\|g(\mathrm{W}_{t-1},\mathrm{Z}) - \boldsymbol{e}\|_2^2\right],
\end{aligned} \tag{61}$$

where the last step is because W_{t-1} is independent of $(\mathrm{Z},\bar{\mathrm{Z}})$ and $\mathrm{Z},\bar{\mathrm{Z}}$ follow the same distribution. By choosing the constant vector $\boldsymbol{e} = \mathbb{E}\left[g(\mathrm{W}_{t-1},\mathrm{Z})\right]$, we have

$$\mathbb{E}\left[\|g(\mathrm{W}_{t-1},\mathrm{Z}) - \boldsymbol{e}\|_2^2\right] = \mathsf{Var}\left(g(\mathrm{W}_{t-1},\mathrm{Z})\right). \tag{62}$$

Combining (60–62) gives

$$\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); b_t\right)\right] \le \frac{2}{b_t^2}\mathsf{Var}\left(g(\mathrm{W}_{t-1},\mathrm{Z})\right). \tag{63}$$

Substituting (59), (63) into (56) leads to the generalization bound in (42).

- Similarly, Table 1 shows for Gaussian noise

$$\mathbb{E}\left[\mathsf{C}_{\mathrm{TV}}\left(g(\mathrm{W}_{t-1},\mathrm{Z}), g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}); b_t\right)\right] \le \frac{1}{2b_t}\mathbb{E}\left[\|g(\mathrm{W}_{t-1},\mathrm{Z}) - g(\mathrm{W}_{t-1},\bar{\mathrm{Z}})\|_2\right]. \tag{64}$$

Furthermore, by the triangle inequality,

$$\begin{aligned}
&\frac{1}{2b_t}\mathbb{E}\left[\|g(\mathrm{W}_{t-1},\mathrm{Z}) - g(\mathrm{W}_{t-1},\bar{\mathrm{Z}})\|_2\right] \\
&\le \frac{1}{2b_t}\left(\mathbb{E}\left[\|g(\mathrm{W}_{t-1},\mathrm{Z}) - \boldsymbol{e}\|_2\right] + \mathbb{E}\left[\|g(\mathrm{W}_{t-1},\bar{\mathrm{Z}}) - \boldsymbol{e}\|_2\right]\right) \\
&= \frac{1}{b_t}\mathbb{E}\left[\|g(\mathrm{W}_{t-1},\mathrm{Z}) - \boldsymbol{e}\|_2\right].
\end{aligned} \tag{65}$$

By choosing the constant vector $\boldsymbol{e} = \mathbb{E}\left[g(\mathrm{W}_{t-1}, \mathrm{Z})\right]$ and combining (64) with (65), we have

$$\mathbb{E}\left[\mathsf{C}_{\mathrm{TV}}\left(g(\mathrm{W}_{t-1}, \mathrm{Z}), g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}); b_t\right)\right] \leq \frac{1}{b_t}\mathbb{E}\left[\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - \boldsymbol{e}\|_2\right]. \tag{66}$$

Substituting (59), (66) into (57) leads to the generalization bound in (43).

- Finally, Table 1 shows for Gaussian noise

$$\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(\mathrm{W}_{t-1}, \mathrm{Z}), g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}); b_t\right)\right] = \mathbb{E}\left[\exp\left(\frac{\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}})\|_2^2}{b_t^2}\right)\right] - 1. \tag{67}$$

The Cauchy-Schwarz inequality implies that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}})\|_2^2}{b_t^2}\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\frac{2\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - \boldsymbol{e}\|_2^2}{b_t^2}\right)\exp\left(\frac{2\|g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}) - \boldsymbol{e}\|_2^2}{b_t^2}\right)\right] \\
&\leq \sqrt{\mathbb{E}\left[\exp\left(\frac{4\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - \boldsymbol{e}\|_2^2}{b_t^2}\right)\right]\mathbb{E}\left[\exp\left(\frac{4\|g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}) - \boldsymbol{e}\|_2^2}{b_t^2}\right)\right]} \\
&= \mathbb{E}\left[\exp\left(\frac{4\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - \boldsymbol{e}\|_2^2}{b_t^2}\right)\right].
\end{aligned} \tag{68}$$

By choosing the constant vector $\boldsymbol{e} = \mathbb{E}\left[g(\mathrm{W}_{t-1}, \mathrm{Z})\right]$ and combining (67) with (68), we have

$$\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(\mathrm{W}_{t-1}, \mathrm{Z}), g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}); b_t\right)\right] \leq \mathbb{E}\left[\exp\left(\frac{4\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - \boldsymbol{e}\|_2^2}{b_t^2}\right)\right] - 1. \tag{69}$$

Since for any $x \geq 0$ and $b \geq 1$,

$$\exp\left(\frac{x}{b}\right) - 1 \leq \frac{\exp(x) - 1}{b},$$

the inequality in (69) can be further upper bounded as

$$\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(\mathrm{W}_{t-1}, \mathrm{Z}), g(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}); b_t\right)\right] \leq \frac{1}{b_t^2}\left(\mathbb{E}\left[\exp\left(4\|g(\mathrm{W}_{t-1}, \mathrm{Z}) - \boldsymbol{e}\|_2^2\right)\right] - 1\right). \tag{70}$$

Substituting (59), (70) into (58) leads to the generalization bound in (44).

■

By a similar analysis, we can prove the generalization bounds in Proposition 11 for the Laplace mechanism.

D.2 Proof of Proposition 14

Proof Within the t-th global update, we can rewrite the local updates conducted by the client $l \in \mathcal{S}_t$ as follows. The parameter is initialized by $\mathrm{W}_{t,0}^l = \mathrm{W}_{t-1}$ and for $j \in [M]$,

$$\mathrm{U}_{t,j}^l = \mathrm{W}_{t,j-1}^l - \eta \cdot g\left(\mathrm{W}_{t,j-1}^l, \{\mathrm{Z}_i^l\}_{i\in[b]}\right) \tag{71a}$$

$$\mathrm{V}_{t,j}^l = \mathrm{U}_{t,j}^l + \eta \cdot \mathrm{N}_{t,j}^l \tag{71b}$$

$$\mathrm{W}_{t,j}^l = \mathsf{Proj}_{\mathcal{W}}\left(\mathrm{V}_{t,j}^l\right) \tag{71c}$$

where $\{\mathrm{Z}_i^l\}_{i\in[b]}$ are drawn independently from the data distribution μ_l and $\mathrm{N}_{t,j}^l \sim N(0, \mathbf{I}_d)$. If a data point Z_i^k is used at the t-th global update, j-th local update, by the client $k \in \mathcal{S}$, then the following Markov chain holds:

$$\underbrace{\mathrm{Z}_i^k \to \{\mathrm{U}_{t,j}^l\}_{l\in\mathcal{S}_t} \to \{\mathrm{V}_{t,j}^l\}_{l\in\mathcal{S}_t} \to \{\mathrm{W}_{t,j}^l\}_{l\in\mathcal{S}_t} \to \cdots \to \{\mathrm{W}_{t,M}^l\}_{l\in\mathcal{S}_t}}_{\text{local}} \underbrace{\to \mathrm{W}_t}_{\text{global}} \to \cdots \to \mathrm{W}_T$$

Let $\mathcal{U}_{t,j}^l$ be the range of $\mathrm{U}_{t,j}^l$. Note that $\mathsf{diam}\left(\mathcal{U}_{t,j}^l\right) \le \mathsf{diam}(\mathcal{W}) + 2\eta K = D + 2\eta K$. Since $|\mathcal{S}_t| = C$, then

$$\mathsf{diam}\left(\prod_{l\in\mathcal{S}_t} \mathcal{U}_{t,j}^l\right) \le \sqrt{\sum_{l\in\mathcal{S}_t} \mathsf{diam}\left(\mathcal{U}_{t,j}^l\right)^2} \le \sqrt{C}(D + 2\eta K).$$

Following a similar analysis in the proof of Lemma 8, we have

$$\begin{aligned} \mathrm{T}(\mathrm{W}_T; \mathrm{Z}_i^k) &\le q^{M(T-t)} \cdot \mathrm{T}(\mathrm{W}_t; \mathrm{Z}_i^k) \\ &\le q^{(M-j)+M(T-t)} \cdot \mathrm{T}(\{\mathrm{W}_{t,j}^l\}_{l\in\mathcal{S}_t}; \mathrm{Z}_i^k), \end{aligned} \tag{72}$$

where the constant q is defined as

$$q \triangleq 1 - 2\bar{\Phi}\left(\frac{\sqrt{C}(D + 2\eta K)}{2\eta}\right).$$

Analogous to the proof of Lemma 9, we have

$$\mathrm{T}(\{\mathrm{W}_{t,j}^l\}_{l\in\mathcal{S}_t}; \mathrm{Z}_i^k) \le \frac{1}{b}\mathbb{E}\left[\|g(\mathrm{W}_{t,j-1}^k, \mathrm{Z}^k) - \boldsymbol{e}\|_2\right] \tag{73}$$

where $\boldsymbol{e} \triangleq \mathbb{E}\left[g(\mathrm{W}_{t,j-1}^k, \mathrm{Z}^k)\right]$. Since the data point Z_i^k is only used by the client $k \in \mathcal{S}_t$, the right-hand side of (73) does not involve $\{\mathrm{W}_{t,j-1}^l\}_{l\in\mathcal{S}_t\setminus\{k\}}$. Combining (72), (73) with the T-information bound in Lemma 1 yields the desired generalization bound for the k-th client. ■

D.3 Proof of Proposition 15

We first present the following lemma whose proof follows by using the technique in Section II. E of Guo et al. (2005).

Lemma 19 *Let* X *be a random variable which is independent of* $\mathrm{N} \sim N(0, \mathbf{I}_d)$. *Then for any* $m > 0$ *and deterministic function* f

$$I(f(\mathrm{X}) + m\mathrm{N}; \mathrm{X}) \leq \frac{1}{2m^2}\mathsf{Var}\left(f(\mathrm{X})\right). \tag{74}$$

More generally, if Z *is another random variable which is independent of* N, *then for any fixed* $\boldsymbol{z}$

$$I(f(\mathrm{X}) + m\mathrm{N}; \mathrm{X} \mid \mathrm{Z} = \boldsymbol{z}) \leq \frac{1}{2m^2}\mathsf{Var}\left(f(\mathrm{X}) \mid \mathrm{Z} = \boldsymbol{z}\right). \tag{75}$$

Proof By the property of mutual information (see Theorem 2.3 in Polyanskiy and Wu, 2019),

$$I(f(\mathrm{X}) + m\mathrm{N}; \mathrm{X}) = I\left(\frac{f(\mathrm{X}) - \boldsymbol{e}}{m} + \mathrm{N}; \mathrm{X}\right) \tag{76}$$

where $\boldsymbol{e} \triangleq \mathbb{E}\left[f(\mathrm{X})\right]$. We denote

$$g(\boldsymbol{x}) \triangleq \frac{f(\boldsymbol{x}) - \boldsymbol{e}}{m}. \tag{77}$$

The golden formula (see Theorem 3.3 in Polyanskiy and Wu, 2019, for a proof) yields

$$\begin{aligned} I\left(g(\mathrm{X}) + \mathrm{N}; \mathrm{X}\right) &= \mathrm{D}_{\mathrm{KL}}\left(P_{g(\mathrm{X})+\mathrm{N}|\mathrm{X}} \| P_{\mathrm{N}} | P_{\mathrm{X}}\right) - \mathrm{D}_{\mathrm{KL}}\left(P_{g(\mathrm{X})+\mathrm{N}} \| P_{\mathrm{N}}\right) \\ &\leq \mathrm{D}_{\mathrm{KL}}\left(P_{g(\mathrm{X})+\mathrm{N}|\mathrm{X}} \| P_{\mathrm{N}} | P_{\mathrm{X}}\right). \end{aligned} \tag{78}$$

Furthermore, since X and N are independent, we have

$$\mathrm{D}_{\mathrm{KL}}\left(P_{g(\mathrm{X})+\mathrm{N}|\mathrm{X}=\boldsymbol{x}} \| P_{\mathrm{N}}\right) = \mathrm{D}_{\mathrm{KL}}\left(P_{g(\boldsymbol{x})+\mathrm{N}} \| P_{\mathrm{N}}\right) = \frac{\|g(\boldsymbol{x})\|_2^2}{2},$$

where the last step is due to the closed-form expression of the KL-divergence between two Gaussian distributions. Finally, by the definition of conditional divergence, we have

$$\mathrm{D}_{\mathrm{KL}}\left(P_{g(\mathrm{X})+\mathrm{N}|\mathrm{X}} \| P_{\mathrm{N}} | P_{\mathrm{X}}\right) = \frac{1}{2}\mathbb{E}\left[\|g(\mathrm{X})\|_2^2\right] = \frac{1}{2m^2}\mathsf{Var}\left(f(\mathrm{X})\right), \tag{79}$$

where the last step is due to the definition of g in (77). Combining (76–79) leads to the desired conclusion. Finally, it is straightforward to obtain (75) by conditioning on $\mathrm{Z} = \boldsymbol{z}$ and repeating our above derivations. ■

Next, we present the second lemma which will be used for proving Proposition 15.

Lemma 20 *If the loss function $\ell(\boldsymbol{w}, \mathrm{Z})$ is σ-sub-Gaussian under $\mathrm{Z} \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, the expected generalization gap of the SGLD algorithm can be upper bounded by*

$$\frac{\sqrt{2}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}_j)\right)},$$

where the set $\mathcal{T}_j$ contains the indices of iterations in which the mini-batch S_j is used and the variance is over the randomness of $(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}_j) \sim P_{\mathrm{W}_{t-1}, \bar{\mathrm{Z}}_j}$ with $\bar{\mathrm{Z}}_j$ being any data point in the mini-batch S_j.

Proof We denote $\mathrm{Z}^{(k)} \triangleq (\mathrm{Z}_1, \cdots, \mathrm{Z}_k)$ for $k \in [n]$ and $\mathrm{W}^{(t)} \triangleq (\mathrm{W}_1, \cdots, \mathrm{W}_t)$ for $t \in [T]$. For simplicity, in what follows we only provide an upper bound for $I(\mathrm{W}; \mathrm{Z}_n)$. Since W is a function of $\mathrm{W}^{(T)} = (\mathrm{W}_1, \cdots, \mathrm{W}_T)$, the data processing inequality yields

$$I(\mathrm{W}; \mathrm{Z}_n) \leq I(\mathrm{W}^{(T)}; \mathrm{Z}_n) \leq I(\mathrm{W}^{(T)}, \mathrm{Z}^{(n-1)}; \mathrm{Z}_n). \tag{80}$$

By the chain rule,

$$I(\mathrm{W}^{(T)}, \mathrm{Z}^{(n-1)}; \mathrm{Z}_n) = I(\mathrm{W}_T; \mathrm{Z}_n \mid \mathrm{W}^{(T-1)}, \mathrm{Z}^{(n-1)}) + I(\mathrm{W}^{(T-1)}, \mathrm{Z}^{(n-1)}; \mathrm{Z}_n). \tag{81}$$

Let $\boldsymbol{w} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{T-1})$ and $\boldsymbol{z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{n-1})$ be any two vectors. If Z_n is not used at the T-th iteration, without loss of generality we assume that the data points $\mathrm{Z}_1, \cdots, \mathrm{Z}_b$ are used in this iteration. Then

$$\begin{aligned}
&I(\mathrm{W}_T; \mathrm{Z}_n \mid \mathrm{W}^{(T-1)} = \boldsymbol{w}, \mathrm{Z}^{(n-1)} = \boldsymbol{z}) \\
&= I\left(\boldsymbol{w}_{T-1} - \frac{\eta_T}{b} \sum_{i=1}^{b} \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, \boldsymbol{z}_i) + \sqrt{\frac{2\eta_T}{\beta_T}} \mathrm{N}_T; \mathrm{Z}_n \mid \mathrm{W}^{(T-1)} = \boldsymbol{w}, \mathrm{Z}^{(n-1)} = \boldsymbol{z}\right) \\
&= I\left(\mathrm{N}_T; \mathrm{Z}_n \mid \mathrm{W}^{(T-1)} = \boldsymbol{w}, \mathrm{Z}^{(n-1)} = \boldsymbol{z}\right) \\
&= 0.
\end{aligned} \tag{82}$$

On the other hand, if Z_n is used at the T-th iteration, without loss of generality we assume that the other $b-1$ data points which are also used in this iteration are $\mathrm{Z}_1, \cdots, \mathrm{Z}_{b-1}$. Then

$$\begin{aligned}
&I(\mathrm{W}_T; \mathrm{Z}_n \mid \mathrm{W}^{(T-1)} = \boldsymbol{w}, \mathrm{Z}^{(n-1)} = \boldsymbol{z}) \\
&= I\left(\boldsymbol{w}_{T-1} - \frac{\eta_T}{b}\left(\sum_{i=1}^{b-1} \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, \boldsymbol{z}_i) + \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, \mathrm{Z}_n)\right) + \sqrt{\frac{2\eta_T}{\beta_T}} \mathrm{N}_T; \mathrm{Z}_n \mid \mathrm{W}^{(T-1)} = \boldsymbol{w}, \mathrm{Z}^{(n-1)} = \boldsymbol{z}\right) \\
&= I\left(-\frac{\eta_T}{b} \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, \mathrm{Z}_n) + \sqrt{\frac{2\eta_T}{\beta_T}} \mathrm{N}_T; \mathrm{Z}_n \mid \mathrm{W}^{(T-1)} = \boldsymbol{w}, \mathrm{Z}^{(n-1)} = \boldsymbol{z}\right).
\end{aligned} \tag{83}$$

By Lemma 19, we have

$$\begin{aligned}
&I\left(-\frac{\eta_T}{b} \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, \mathrm{Z}_n) + \sqrt{\frac{2\eta_T}{\beta_T}} \mathrm{N}_T; \mathrm{Z}_n \mid \mathrm{W}^{(T-1)} = \boldsymbol{w}, \mathrm{Z}^{(n-1)} = \boldsymbol{z}\right) \\
&\leq \frac{\beta_T \eta_T}{4b^2} \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, \mathrm{Z}_n) \mid \mathrm{W}^{(T-1)} = \boldsymbol{w}, \mathrm{Z}^{(n-1)} = \boldsymbol{z}\right).
\end{aligned} \tag{84}$$

Substituting (84) into (83) gives

$$I(\mathrm{W}_T;\mathrm{Z}_n \mid \mathrm{W}^{(T-1)}=\boldsymbol{w},\mathrm{Z}^{(n-1)}=\boldsymbol{z}) \leq \frac{\beta_T\eta_T}{4b^2}\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(\boldsymbol{w}_{T-1},\mathrm{Z}_n) \mid \mathrm{W}^{(T-1)}=\boldsymbol{w},\mathrm{Z}^{(n-1)}=\boldsymbol{z}\right).$$

Taking expectation w.r.t. $(\mathrm{W}^{(T-1)},\mathrm{Z}^{(n-1)})$ on both sides of the above inequality and using the law of total variance lead to

$$I(\mathrm{W}_T;\mathrm{Z}_n \mid \mathrm{W}^{(T-1)},\mathrm{Z}^{(n-1)}) \leq \frac{\beta_T\eta_T}{4b^2}\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(\mathrm{W}_{T-1},\mathrm{Z}_n)\right). \tag{85}$$

To summarize, (82) and (85) can be rewritten as

$$\begin{aligned} &I(\mathrm{W}_T;\mathrm{Z}_n \mid \mathrm{W}^{(T-1)},\mathrm{Z}^{(n-1)}) \\ &\leq \begin{cases} \frac{\beta_T\eta_T}{4b^2}\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(\mathrm{W}_{T-1},\mathrm{Z}_n)\right) & \text{if } \mathrm{Z}_n \text{ is used at the } T\text{-th iteration,} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{86}$$

Assume that the data point Z_n belongs to the j-th mini-batch S_j. Now substituting (86) into (81) and doing this procedure recursively lead to

$$I(\mathrm{W}^{(T)},\mathrm{Z}^{(n-1)};\mathrm{Z}_n) \leq \sum_{t\in\mathcal{T}_j}\frac{\beta_t\eta_t}{4b^2}\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(\mathrm{W}_{t-1},\mathrm{Z}_n)\right),$$

where the set $\mathcal{T}_j$ contains the indices of iterations in which the mini-batch S_j is used. Hence, this upper bound along with (80) naturally gives

$$I(\mathrm{W};\mathrm{Z}_n) \leq \sum_{t\in\mathcal{T}_j}\frac{\beta_t\eta_t}{4b^2}\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(\mathrm{W}_{t-1},\mathrm{Z}_n)\right). \tag{87}$$

By symmetry, for any data point in S_j besides Z_n, the mutual information between W and this data point can be upper bound by the right-hand side of (87) as well. Finally, recall that Lemma 1 provides an upper bound for the expected generalization gap:

$$\frac{\sqrt{2}\sigma}{n}\sum_{i=1}^{n}\sqrt{I(\mathrm{W}_T;\mathrm{Z}_i)} = \frac{\sqrt{2}\sigma}{n}\sum_{j=1}^{m}\sum_{\mathrm{Z}\in\mathrm{S}_j}\sqrt{I(\mathrm{W}_T;\mathrm{Z})}. \tag{88}$$

By substituting (87) into the above expression, we know the expected generalization gap can be further upper bounded by

$$\frac{\sqrt{2}\sigma}{2n}\sum_{j=1}^{m}\sqrt{\sum_{t\in\mathcal{T}_j}\beta_t\eta_t\cdot\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(\mathrm{W}_{t-1},\bar{\mathrm{Z}}_j)\right)},$$

where $\bar{\mathrm{Z}}_j$ is any data point in the mini-batch S_j. ■

Finally, we are in a position to prove Proposition 15.

Proof Consider a new loss function and the gradient of a new surrogate loss:

$$\ell(\boldsymbol{w}, \mathrm{S}_j) \triangleq \frac{1}{b} \sum_{\mathrm{Z} \in \mathrm{S}_j} \ell(\boldsymbol{w}, \mathrm{Z}), \quad \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}, \mathrm{S}_j) \triangleq \frac{1}{b} \sum_{\mathrm{Z} \in \mathrm{S}_j} \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}, \mathrm{Z}).$$

Then $\ell(\boldsymbol{w}, \mathrm{S}_j)$ is $\sigma/\sqrt{b}$-sub-Gaussian under $\mathrm{S}_j \sim \mu^{\otimes b}$ for all $\boldsymbol{w} \in \mathcal{W}$. We view each mini-batch S_j as a data point and view $\ell(\boldsymbol{w}, \mathrm{S}_j)$ as a new loss function. By using Lemma 20, we obtain:

$$\left|\mathbb{E}\left[L_\mu(\mathrm{W}) - L_{\mathrm{S}}(\mathrm{W})\right]\right| \leq \frac{\sqrt{2}\sigma}{2m\sqrt{b}} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathsf{Var}\left[\nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \mathrm{S}_j)\right]}. \tag{89}$$

Since the data set contains n data points and is divided into m disjoint mini-batches with size b, we have $n = mb$. Substituting this into (89) leads to the desired conclusion. ■

D.4 Proof of Corollary 16

Proof The Minkowski inequality implies that for any non-negative x_i, the inequality $\sqrt{\sum_i x_i} \leq \sum_i \sqrt{x_i}$ holds. Therefore, we can further upper bound the generalization bound in Lemma 20 by

$$\frac{\sqrt{2}\sigma}{2n} \sum_{j=1}^{m} \sum_{t \in \mathcal{T}_j} \sqrt{\beta_t \eta_t \cdot \mathsf{Var}\left[\nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \bar{\mathrm{Z}}_j)\right]} = \frac{\sqrt{2}\sigma}{2n} \sum_{t=1}^{T} \sqrt{\beta_t \eta_t \cdot \mathsf{Var}\left[\nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \mathrm{Z}_t^\dagger)\right]}.$$

Alternatively, by Jensen's inequality and $n = mb$, we can further upper bound the generalization bound in Lemma 20 by

$$\frac{\sqrt{2}\sigma}{2} \sqrt{\frac{1}{bn} \sum_{t=1}^{T} \beta_t \eta_t \cdot \mathsf{Var}\left[\nabla_{\boldsymbol{w}} \hat{\ell}(\mathrm{W}_{t-1}, \mathrm{Z}_t^\dagger)\right]}.$$

■

Appendix E. Supporting Experimental Results

Recall that our generalization bound in Proposition 15 involves the variance of gradients. To estimate this quantity from data, we repeat our experiments 4 times and record the batch gradient at each iteration. This batch gradient is the one used for updating the parameters in the SGLD algorithm so it does not require any additional computations. Then we estimate the variance of gradients by using the population variance of the recorded batch gradients. Finally, we repeat the above procedure 4 times for computing the standard deviation, leading to e.g., the shaded areas in Figure 1. We provide experimental details in Table 3 and 4 and code in https://github.com/yih117/Analyzing-the-Generalization-Capability-of-SGLD-Using-Properties-of-Gaussian-Channels for reproducing our experiments.

Parameter	Details
Data set	MNIST
Number of training data	5000
Batch size	500
Learning rate	Initialization = 0.03, decay rate = 0.96, decay steps=2000
Inverse temperature	$\beta_t = 10^6/(2\eta_t)$
Architecture	MLP with ReLU activation
Depth	3 layers
Width	64 hidden units
Objective function	Cross-entropy loss
Loss function	0-1 loss

Table 3: Experiment details of Figure 1 and 2 on the MNIST data set. The network width is varying among $\{16, 32, 64, 128, 256\}$ hidden units for Figure 2.

Parameter	Details
Data set	CIFAR-10 and SVHN
Number of training data	5000
Batch size	500
Learning rate	Initialization = 0.03, decay rate = 0.96, decay steps = 2000
Inverse temperature	$\beta_t = 10^6/(2\eta_t)$
Architecture	$\mathsf{conv}(5, 32)$ $\mathsf{pool}(2)$ $\mathsf{conv}(5, 32)$ $\mathsf{pool}(2)$ $\mathsf{fc}(120)$ $\mathsf{fc}(84)$ $\mathsf{fc}(10)$
Objective function	Cross-entropy loss
Loss function	0-1 loss

Table 4: Experiment details of Figure 1 and 2 on the CIFAR-10 and SVHN data sets. Here $\mathsf{conv}(k, w)$ is a $k \times k$ convolutional layer with w filters; $\mathsf{pool}(k)$ is a $k \times k$ max pooling layer; and $\mathsf{fc}(k)$ is a fully connected layer with k units. The convolutional layers and the fully connected layers all use ReLU activation function. The network width (i.e., number of filters in CNN) is varying among $\{8, 16, 32, 64, 128\}$ for Figure 2.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

Ibrahim M Alabdulmohsin. Algorithmic stability and uniform generalization. In *Advances in Neural Information Processing Systems*, volume 28, pages 19–27, 2015.

Gholamali Aminian, Laura Toni, and Miguel RD Rodrigues. Jensen-Shannon information based characterization of the generalization error of learning algorithms. In *IEEE Information Theory Workshop*, pages 1–5, 2021.

Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shahab Asoodeh, Mario Diaz, and Flavio P Calmon. Privacy amplification of iterative algorithms via contraction coefficients. In *IEEE International Symposium on Information Theory*, pages 896–901, 2020.

Borja Balle, Gilles Barthe, Marco Gaboardi, and Joseph Geumlek. Privacy amplification by mixing and diffusion mechanisms. In *Advances in Neural Information Processing Systems*, pages 13298–13308, 2019.

Peter L Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2), 1998.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. *SIAM Journal on Computing*, pages 377–405, 2021.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Flavio P Calmon, Yury Polyanskiy, and Yihong Wu. Strong data processing inequalities for input constrained additive noise channels. *IEEE Transactions on Information Theory*, 64(3):1879–1892, 2017.

Joel Cohen, Johannes HB Kempermann, and Gheorghe Zbaganu. *Comparisons of stochastic matrices with applications in information theory, statistics, economics and population*. Springer Science & Business Media, 1998.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

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

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

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

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 2021.

Facebook AI. Introducing Opacus: A high-speed library for training pytorch models with differential privacy, 2020.

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

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

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

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51 (4):1261–1282, 2005.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 9925–9935, 2020.

Fengxiang He, Bohan Wang, and Dacheng Tao. Tighter generalization bounds for iterative differentially private learning algorithms. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, 2020.

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*, 2020.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *IEEE International Symposium on Information Theory*, pages 1475–1479, 2017.

Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic bounds on transfer generalization gap based on Jensen-Shannon divergence. In *European Signal Processing Conference*, pages 1461–1465, 2021a.

Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1):126, 2021b.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *ACM Symposium on Theory of Computing*, pages 9–9, 2021.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *International Conference on Machine Learning*, pages 2698–2707. PMLR, 2018.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *AAAI Conference on Artificial Intelligence*, pages 1788–1794, 2016.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.

Adrian Tovar Lopez and Varun Jog. Generalization error bounds using Wasserstein distances. In *IEEE Information Theory Workshop*, pages 1–5. IEEE, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638, 2018.

Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

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*, pages 11015–11025, 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 *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

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.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations (Workshop)*, 2015.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

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

Yury Polyanskiy and Yihong Wu. Dissipation of information in channels with input constraints. *IEEE Transactions on Information Theory*, 62(1):35–55, 2016.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for 6.441 (MIT), ECE 563 (UIUC), STAT 364 (Yale)*, 2019. URL `http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf`.

Carey Radebaugh and Ulfar Erlingsson. Introducing TensorFlow privacy: Learning with differential privacy for training data, 2019.

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*, 10(1-2):1–246, 2013.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *IEEE Information Theory Workshop*, pages 26–30, 2016.

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

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

Borja Rodríguez Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. Tighter expected generalization error bounds via Wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 34, pages 19109–19121, 2021.

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

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, pages 245–248, 2013.

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

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer-Verlag New York, 2009.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

Thomas Verma and Judea Pearl. *Equivalence and synthesis of causal models.* UCLA, Computer Science Department, 1991.

Hao Wang, Mario Diaz, José Cândido S Santos Filho, and Flavio P Calmon. An information-theoretic view of generalization via Wasserstein distance. In *IEEE International Symposium on Information Theory*, pages 577–581, 2019.

Hao Wang, Yizhe Huang, Rui Gao, and Flavio P Calmon. Analyzing the generalization capability of SGLD using properties of Gaussian channels. In *Advances in Neural Information Processing Systems*, volume 34, pages 24222–24234, 2021.

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

Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *ACM International Conference on Management of Data*, pages 1307–1322, 2017.

Yihong Wu. Lecture notes on: Information-theoretic methods for high-dimensional statistics. 2020. URL `http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf`.

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

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.

Semih Yagli, Alex Dytso, and H Vincent Poor. Information-theoretic bounds on the generalization error and privacy leakage in federated learning. In *IEEE International Workshop on Signal Processing Advances in Wireless Communications*, pages 1–5, 2020.

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

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022, 2017b.

Ruida Zhou, Chao Tian, and Tie Liu. Individually conditional individual mutual information bound on generalization error. In *IEEE International Symposium on Information Theory*, pages 670–675, 2021.