STABILITY BASED GENERALIZATION BOUNDS FOR EXPONENTIAL FAMILY LANGEVIN DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the generalization of noisy stochastic mini-batch based iterative algorithms based on the notion of stability. Recent years have seen key advances in data-dependent generalization bounds for noisy iterative learning algorithms such as stochastic gradient Langevin dynamics (SGLD) based on (Mou et al., 2018; Li et al., 2020) and related approaches (Negrea et al., 2019; Haghifam et al., 2020). In this paper, we unify and substantially generalize stability based generalization bounds and make three technical advances. First, we bound the generalization error of general noisy stochastic iterative algorithms (not necessarily gradient descent) in terms of expected stability, which in turn can be bounded by the expected Le Cam Style Divergence (LSD). Such bounds have a $O(1/n)$ sample dependence unlike many existing bounds with $O(1/\sqrt{n})$ dependence. Second, we introduce Exponential Family Langevin Dynamics (EFLD) which is a substantial generalization of SGLD and which allows exponential family noise to be used with gradient descent. We establish data-dependent expected stability based generalization bounds for general EFLD. Third, we consider an important new special case of EFLD: noisy sign-SGD, which extends sign-SGD by using Bernoulli noise over $\{-1, +1\}$, and we establish optimization guarantees for the algorithm. Further, we present empirical results on benchmark datasets to illustrate the our bounds are non-vacuous and quantitatively much sharper than existing bounds.

1 INTRODUCTION

Recent years have seen renewed interest and advances in characterizing generalization performance of learning algorithms in terms of stability, which considers change in performance of a learning algorithm based on change of a single training point (Hardt et al., 2016; Bousquet & Elisseeff, 2002; Li et al., 2020; Mou et al., 2018). For stochastic gradient descent (SGD), Hardt et al. (2016) established generalization bounds based on uniform stability (Hardt et al., 2016; Bousquet & Elisseeff, 2002), although the analysis needed rather small step sizes $\eta_t = 1/t$ which is not useful in practice. While improving the analysis for SGD has remained a challenge, advances have been made on noisy SGD algorithms, especially stochastic gradient Langevin dynamics (SGLD) (Welling & Teh, 2011; Mou et al., 2018; Li et al., 2020), which adds Gaussian noise to the stochastic gradients of marginal variance σ_t^2. In parallel, there has been key developments on related information-theoretic generalization bounds applicable to SGLD type algorithms (Negrea et al., 2019; Haghifam et al., 2020; Xu & Raginsky, 2017; Russo & Zou, 2016; Pensia et al., 2018).

While these developments have led to major advances in analyzing generalization of noisy SGD algorithms, they each have certain limitations which leave room for further improvements. Using uniform stability, Mou et al. (2018) established a bound for SGLD of the form $\frac{K}{n}\sqrt{\sum_t \eta_t^2/\sigma_t^2}$ which depends on K, the global Lipschitz constant for the loss, and with step size $\eta_t \leq \sigma_t \ln 2/K$. The bound has a desirable dependency of $O(1/n)$ on the samples, but has an undesirable dependence on K, and the step sizes, bounded by σ_t/K, are too small. Mou et al. (2018) also presented another bound which addresses some of these issues, but gets an undesirable $O(1/\sqrt{n})$ sample dependence. By building on the developments of Russo & Zou (2016); Xu & Raginsky (2017); Pensia et al. (2018), Negrea et al. (2019) made advances from the information theoretic perspective and established bounds for SGLD which have the desirable dependence on the norm of gradient incoherence, i.e., difference in gradients over different mini-batches, avoids dependence on Lipschitz constant K, and is applicable to unbounded sub-Gaussian losses, but have an undesirable $O(1/\sqrt{n})$ sample

dependence. Haghifam et al. (2020) made further advances on the problem from an information theoretic perspective based on the conditional mutual information framework of Steinke & Zakynthinou (2020) and obtained generalization bounds based on gradient incoherence with $O(1/n)$ sample dependence, but their analysis holds for full batch Langevin dynamics, not mini-batch SGLD. Li et al. (2020) made advances on such bounds based on the notion of Bayes-stability, by combining ideas from PAC-Bayes bounds into stability, and established a bound of the form $\frac{c}{n}\sqrt{\sum_t \eta_t^2 \mathbf{g}_e(t)/\sigma_t^2}$ for bounded losses, where $\mathbf{g}_e(t)$ is the expected gradient norm square at step t. While the bound avoids dependency on the Lipschitz constant K, the dependence on the gradient norm makes such bounds much weaker than the information theoretic bounds of Negrea et al. (2019); Haghifam et al. (2020) which depend on the norm of gradient incoherence, which are typically orders of magnitude smaller. Further, the analysis of Li et al. (2020) still needs small step sizes, bounded by σ_t/K.

In this paper, we build on the core strengths of such existing approaches, most notably the $O(1/n)$ sample dependence of stability based bounds (Mou et al., 2018; Li et al., 2020) and the dependence on gradient incoherence for information theoretic bounds (Negrea et al., 2019; Haghifam et al., 2020), and develop a framework (Section 2) for developing generalization bounds for noisy stochastic iterative (NSI) algorithms. Our framework considers generalization based on the concept of *expected stability*, rather than uniform stability (Hardt et al., 2016; Bousquet & Elisseeff, 2002; Bousquet et al., 2020; Mou et al., 2018), which yields distribution dependent generalization bounds and avoids the worst-case setting of uniform stability. Building on Li et al. (2020), we show that expected stability of general NSI algorithms can be bounded by the expected Le Cam Style Divergence with dependence on parameter distributions from mini-batches differing by one sample. In Section 3, we introduce Exponential Family Langevin Dynamics (EFLD), a family of noisy gradient descent algorithms based on exponential family noise. Special cases of EFLD include SGLD and noisy versions of Sign-SGD or quantized SGD algorithms widely used in practice (Bernstein et al., 2018a;b; Jin et al., 2020; Alistarh et al., 2017)Our main result provides an expected stability based generalization bound applicable to any EFLD algorithm with a $O(1/n)$ sample dependence and a dependence on gradient incoherence, rather than gradient norms. Existing generalization bounds for SGLD (Li et al., 2020; Negrea et al., 2019) usually use properties of the Gaussian distribution, and our analysis on EFLD illustrates that this was unnecessary. We also consider optimization guarantees for EFLD and establish such results for noisy Sign-SGD and SGLD. Through experiments on benchmark datasets (Section 4), we illustrate that our bounds are non-vacuous and quantitatively much sharper than existing bounds (Li et al., 2020; Negrea et al., 2019).

Related work. Uniform stability has been a classical approach for bounding generalization error (Bousquet & Elisseeff, 2002; Bousquet et al., 2020; Feldman & Vondrak, 2018; 2019), pioneered by Rogers & Wagner (1978); Devroye & Wagner (1979). Beyond the aforementioned work, there has been recent work on differential privacy that analyzes the uniform stability of differentially private SGD (DP-SGD) (Hardt et al., 2016; Bassily et al., 2020). Beyond uniform stability, information-theoretic approaches (Russo & Zou, 2016; Xu & Raginsky, 2017) that bounds the generalization error by the mutual information between the algorithm input S and the algorithm output $\mathbf{w}$, have been used for deriving generalization bounds for noisy iterative algorithms (Pensia et al., 2018; Bu et al., 2019). Along this line of literature, Negrea et al. (2019); Haghifam et al. (2020); Rodríguez-Gálvez et al. (2021) prove data-dependent generalization bounds dropping dependence on the Lipschitz constant. Further, tighter bounds (Haghifam et al., 2020; Zhou et al., 2021; Rodríguez-Gálvez et al., 2021; Neu, 2021; Hellström & Durisi, 2021) are proposed based on conditional mutual information (Steinke & Zakynthinou, 2020; Grünwald et al., 2021; Hellström & Durisi, 2020). Due to space limitations, an extended discussion of the related work is deferred to Appendix A.

2 Generalization Bounds with Expected Stability

In the setting of statistical learning, there is an instance space $\mathcal{Z}$, a hypothesis space $\mathcal{W}$, and a loss function $\ell : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$. Let D be an unknown distribution of $\mathcal{Z}$ and let $S \sim D^n$ be n i.i.d. draws from D. For any specific hypothesis $\mathbf{w} \in \mathcal{W}$, the population and empirical loss are respectively given by $L_D(\mathbf{w}) \triangleq \mathbb{E}_{z\sim D}[\ell(\mathbf{w}, z)]$, and $L_S(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}, z_i)$. For any distribution P over the hypothesis space, we respectively denote the expected population and empirical loss as

$$L_D(P) \triangleq \mathbb{E}_{z\sim D}\mathbb{E}_{\mathbf{w}\sim P}[\ell(\mathbf{w}, z)] \,, \qquad \text{and} \qquad L_S(P) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{w}\sim P}[\ell(\mathbf{w}, z_i)] \,. \tag{1}$$

Consider a randomized algorithm A which works with $S = \{z_1, \ldots, z_n\} \sim D^n$ and cre-

ates a distribution over the hypothesis space $\mathcal{W}$. For convenience, we will denote the distribution as $A(S)$. The focus of the analysis is to bound the generalization error of A defined as: $\text{gen}(A(S)) \triangleq L_D(A(S)) - L_S(A(S))$. We will assume A is permutation invariant, i.e., the ordering of samples in S do not modify $A(S)$, an assumption satisfied by most learning algorithms. We will focus on developing bounds for the expectation $\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]$, and discuss high-probability bounds in the Appendix B.

2.1 Bounds based on Expected Stability

We start our analysis by noting that the expected generalization error can be upper bounded by *expected stability* based on the Hellinger divergence (Sason & Verdu, 2016; Li et al., 2020): $H^2(P\|P') = \frac{1}{2}\int_{\mathbf{w}}(\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})})^2 d\mathbf{w}$.

Proposition 1. *Let $S_n \sim D^n$ and let S'_n be a dataset obtained by replacing $z_n \in S_n$ with $z'_n \sim D$. Let $A(S_n), A(S'_n)$ respectively denote the distributions over the hypothesis space $\mathcal{W}$ obtained by running randomized algorithm A on S_n, S'_n. Assume that for $S_n \sim D^n, \forall z \in \mathcal{Z}$, $\mathbb{E}_{W\sim A(S_n)}[\ell^2(W,z)] \leq c_0/2, c_0 > 0$. With $H(\cdot,\cdot)$ denoting the Hellinger divergence, we have*

$$|\mathbb{E}_{S_n\sim D^n}[L_D(A(S_n)) - L_S(A(S_n))]| \leq c_0 \mathbb{E}_{S_n\sim D^n}\mathbb{E}_{z'_n\sim D}\sqrt{2H^2\big(A(S_n), A(S'_n)\big)} \,. \tag{2}$$

Remark 2.1. Proposition 1 does not need bounded losses. Just the *second moment* of $\ell(W,z), W \sim A(S_n), S_n \sim D^n, \forall z \in \mathcal{Z}$ need to be bounded. The assumption is satisfied by bounded losses. It is instructive to compare the assumption to that in recent information theoretic bounds (Haghifam et al., 2020; Xu & Raginsky, 2017), where one assumes $\ell(\mathbf{w}, Z), Z \sim D, \forall \mathbf{w} \in \mathcal{W}$ to be *sub-Gaussian*.

Remark 2.2. The bound in Proposition 1 is in terms of *expected stability* where we consider $\mathbb{E}_{S\sim D^n}\mathbb{E}_{z'_n\sim D}[\cdots]$, an important departure from bounds based on *uniform stability* (Elisseeff et al., 2005; Bousquet & Elisseeff, 2002; Mou et al., 2018; Bousquet et al., 2020; Feldman & Vondrak, 2018; 2019) where one considers $\sup_{S,S'\in\mathcal{Z}^n,|S\setminus S'|=1}[\cdots]$. Replacing $\sup$ by $\mathbb{E}$ makes the bounds distribution dependent, and arguably leads to quantitatively tighter bounds.

Note that the Hellinger divergence can be bounded by the KL divergence.

Proposition 2. *For any distributions P and P', $2H^2(P,P') \leq \min\big\{KL(P,P'), \sqrt{\frac{1}{2}KL(P,P')}\big\}$.*

2.2 Expected Stability of Noisy Stochastic Iterative Algorithms

We consider a general family of noisy stochastic iterative (NSI) algorithms. Given $S \sim D^n$, such iterative algorithms have two additional sources of randomness in each iteration t: (a) a stochastic mini-batch of samples S_{B_t}, with $|S_{B_t}| = b$, drawn uniformly at random with replacement from S; and (b) noise $\boldsymbol{\xi}_t$ suitably included in the iterative update. Given a trajectory of past iterates $W_{0:(t-1)} = \mathbf{w}_{0:(t-1)}$, the new iterate W_t is drawn from a distribution $P_{B_t,\boldsymbol{\xi}_t|\mathbf{w}_{0:(t-1)}}$ over $\mathcal{W}$:

$$W_t \sim P_{B_t,\boldsymbol{\xi}_t|\mathbf{w}_{0:(t-1)}}(W) \,. \tag{3}$$

We will drop the conditioning $\mathbf{w}_{0:(t-1)}$ to avoid clutter in the sequel. Let P_T, P'_T denote the marginal distributions over hypotheses $\mathbf{w} \in \mathcal{W}$ after T steps of the algorithm based on S_n, S'_n respectively. Further, let $P_{0:(t-1)}$ denote the joint distribution over $W_{0:(t-1)} = (W_0, \ldots, W_{t-1})$, and let $P_{t|} \equiv P_{B_t,\boldsymbol{\xi}_t|\mathbf{w}_{0:(t-1)}}$ compactly denote the conditional distribution on W_t conditioned on the trajectory $W_{0:(t-1)} = \mathbf{w}_{0:t-1}$. Following (Negrea et al., 2019; Haghifam et al., 2020; Pensia et al., 2018), we use the following chain rule for KL-divergence: $KL(P_T\|P'_T) \leq KL(P_{0:T}\|P'_{0:T}) = \sum_{t=1}^{T}\mathbb{E}_{P_{0:(t-1)}}[KL(P_{t|}\|P'_{t|})]$. Let $\bar{S} \sim D^{n+1}$, and let S_n, S'_n be size n subsets of $\bar{S}$ such that $S_n = \{Z_1, \ldots, Z_{n-1}, Z_n\}$ and $S'_n = \{Z_1, \ldots, Z_{n-1}, Z'_n\}$, where $Z'_n = Z_{n+1}$. Let $S_0 = \{Z_1, \ldots, Z_{n-1}\}$. The algorithms we consider use a mini-batch of size b in each iteration uniformly sampled from n samples. Let the set of all mini-batch index sets be denoted by G. Let the set of all mini-batch index sets A drawn from S_0 be denoted by G_0. Note that $|G_0| = \binom{n-1}{b}$. Let G_1 denote the set of all mini-batch index sets B which includes the last sample, viz. z_n for S with mini-batches and z'_n for S'_n. Note that $|G_1| = \binom{n-1}{b-1}$. Also note that $|G_0| + |G_1| = \binom{n-1}{b} + \binom{n-1}{b-1} = \binom{n}{b} = |G|$.

Following Li et al. (2020), we can bound their conditional KL-divergences $KL(P_{t|}\|P'_{t|})$ in terms of a Le Cam Style Divergence (LSD). While the classical Le Cam divergence (Sason & Verdu,

2016) is $LCD(P\|P') \triangleq \frac{1}{2}\int \frac{(dP-dP')^2}{dP+dP'}$ (where dP denotes the density), our bounds in terms of $LSD(P_{t|}\|\|P'_{t|}) \triangleq \int \frac{(dP_{B_t,\boldsymbol{\xi}_t}-dP'_{B_t,\boldsymbol{\xi}_t})^2}{dP_{A_t,\boldsymbol{\xi}_t}}, B_t \in G_1, A_t \in G_0$. Note that $P_{B_t,\boldsymbol{\xi}_t}$ and $P'_{B_t,\boldsymbol{\xi}_t}$ represent the distribution of W_t for S_n and S'_n respectively since the mini-batch S_{B_t} of S_n and S'_n differs in the n-th sample. Putting everything together, we have the following LSD based bound.

Lemma 1. *Consider a noisy stochastic iterative algorithms of the form* (3) *with mini-batch size* $b \leq n/2$. *Then, with* $c_1 = \sqrt{2}c_0$ *(with* c_0 *as in Proposition 1), we have*

$$|\mathbb{E}_{S_n}[L_D(A(S_n)) - L_{S_n}(A(S_n))]| \leq c_1 \frac{b}{n}\mathbb{E}_{S_n}\mathbb{E}_{z'_n}\sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{W_{0:(t-1)}} \mathop{\mathbb{E}}_{B_t\in G_1} \mathop{\mathbb{E}}_{A_t\in G_0}\left[\int_{\boldsymbol{\xi}_t} \frac{\left(dP_{B_t,\boldsymbol{\xi}_t} - dP'_{B_t,\boldsymbol{\xi}_t}\right)^2}{dP_{A_t,\boldsymbol{\xi}_t}} d\boldsymbol{\xi}_t\right]} . \tag{4}$$

Remark 2.3. Li et al. (2020) essentially has this result for SGLD and inspired our work. Our proofs are significantly simpler and, more importantly, illustrates applicability to general noisy iterative algorithms of the form (3), not just SGLD with Gaussian noise as in Li et al. (2020).

Remark 2.4. Note that the bound does not assume the loss to be bounded, depends on expectations over samples S_n, z'_n, trajectories $\mathbf{w}_{0:(t-1)}$, and mini-batches B_t, A_t. Further, the bound depends on the distribution discrepancy as captured by the expected LSD.

Remark 2.5. The bound seems to worsen with b, the size of the mini-batch. As we shown in Section 3, the expected LSD term has a $\frac{1}{b^2}$ dependence for the Exponential Family Langevin dynamics (EFLD) models we introduce, so the leading b is neutralized.

3 Exponential Family Langevin Dynamics

In recent years, considerable advances have been made in establishing generalization bounds for stochastic gradient Langevin dynamics (SGLD) (Li et al., 2020; Pensia et al., 2018; Negrea et al., 2019; Haghifam et al., 2020). As an example of NSI algorithms of the form (3), SGLD adds an isotropic Gaussian noise at every step of SGD: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, S_{B_t}) + \mathcal{N}\left(0, \sigma_t^2 \mathbb{I}_d\right)$, where $\nabla \ell(\mathbf{w}_t, S_{B_t})$ is the stochastic gradient on mini-batch B_t, η_t is the step size, and σ_t^2 is noise vairance.

In this paper, we introduce a substantial generalization of SGLD called Exponential Family Langevin Dynamics (EFLD) which uses general exponential family noise in noisy iterative updates of the form (3). In addition to being a mathematical generalization of the popular SGLD, the proposed EFLD provides flexibility to use noise gradient algorithms with different representation of the gradient, e.g., Bernoulli noise for Sign-SGD, discrete distribution for quantized or finite precision SGD, etc. (Canonne et al., 2020; Alistarh et al., 2017; Jiang & Agrawal, 2018; Yang et al., 2019).

3.1 Exponential Family Langevin Dynamics (EFLD)

Exponential families (Barndorff-Nielsen, 2014; Brown, 1986; Wainwright & Jordan, 2008) constitute a large family of parametric distributions which include Gaussian, Bernoulli, gamma, Poisson, Dirichlet, etc., as special cases. Exponential families are typically represented in terms of natural parameters $\boldsymbol{\theta}$, and we consider component-wise independent distributions with scaled natural parameter $\boldsymbol{\theta}_\alpha = \boldsymbol{\theta}/\alpha$ with scaling $\alpha > 0$, i.e.,

$$p_\psi(\boldsymbol{\xi}, \boldsymbol{\theta}_\alpha) = \exp(\langle \boldsymbol{\xi}, \boldsymbol{\theta}_\alpha\rangle - \psi(\boldsymbol{\theta}_\alpha))\boldsymbol{\pi}_0(\boldsymbol{\xi}) = \prod_{j=1}^{p} \exp(\xi_j \theta_{j\alpha} - \psi_j(\theta_{j\alpha}))\pi_0(\xi_j) ,$$

where $\boldsymbol{\xi}$ is the sufficient statistic, $\psi(\boldsymbol{\theta}_\alpha) = \sum_{j=1}^{p} \psi_j(\theta_{j\alpha})$ is the log-partition function, and $\boldsymbol{\pi}_0(\xi) = \prod_{j=1}^{p} \pi_0(\xi_j)$ is the base measure. Note that $\alpha = 1$ gives the canonical form of the exponential family distributions. For general scaling $\alpha > 0$, for some cases the base measure π_0 may depend on the scaling, i.e., $\pi_{0,\alpha}$. A scaling $\alpha > 0$ is valid as long as $\exp(\langle \boldsymbol{\xi}, \boldsymbol{\theta}_\alpha\rangle$ is integrable, i.e., $\int_{\boldsymbol{\xi}} \exp(\langle \boldsymbol{\xi}, \boldsymbol{\theta}_\alpha\rangle \pi_0(\boldsymbol{\xi}) d\boldsymbol{\xi} < \infty$. Further, ψ is a smooth function by construction (Barndorff-Nielsen, 2014; Banerjee et al., 2005; Wainwright & Jordan, 2008) and the smoothness of ψ implies $\nabla^2 \psi(\boldsymbol{\theta}_\alpha) \leq c_2 \mathbb{I}$.

Exponential family Langevin dynamics (EFLD) uses noisy stochastic gradient updates similar to SGLD, but using exponential family noise rather than Gaussian noise as in SGLD. In particular, for mini-batch S_{B_t}, EFLD updates are as follows: with step size $\rho_t > 0$

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \rho_t \boldsymbol{\xi}_t , \qquad \boldsymbol{\xi}_t \sim p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B_t,\alpha_t}) , \tag{5}$$

where

$$p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B_t,\alpha_t}) = \exp(\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B_t,\alpha_t}\rangle - \psi(\boldsymbol{\theta}_{B_t,\alpha_t}))\boldsymbol{\pi}_0(\boldsymbol{\xi}) \,, \qquad \boldsymbol{\theta}_{B_t,\alpha_t} \triangleq \frac{\boldsymbol{\theta}_{B_t}}{\alpha_t} = \frac{\nabla \ell(\mathbf{w}_{t-1}, S_{B_t})}{\alpha_t} \,. \quad (6)$$

For EFLD, the natural parameter $\boldsymbol{\theta}_{B_t,\alpha_t}$ at step t is simply a scaled version of the mini-batch gradient $\nabla \ell(\mathbf{w}_{t-1}, S_{B_t})$. We first show that EFLD becomes SGLD when the exponential family is Gaussian, and becomes a noisy version of sign-SGD (Bernstein et al., 2018a;b) when the exponential family is Bernoulli over $\{-1, +1\}$. More details and examples are in Appendix C.1.

Example 3.1 (SGLD)**.** SGLD uses scaled Gaussian noise with $\psi(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2/2, \alpha_t = \sqrt{\sigma_t/\eta_t}$, so that $p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B_t,\alpha_t}) = \mathcal{N}(\boldsymbol{\theta}_{B_t}, \alpha_t^2 \mathbb{I}_d)$. By taking $\rho_t = \sqrt{\eta_t \sigma_t}$, the update (5) based on $\rho_t \boldsymbol{\xi}_t$ is distributed as $\mathcal{N}(\rho_t \boldsymbol{\theta}_{B_t}, \rho_t^2 \alpha_t^2 \mathbb{I}_d) = \mathcal{N}(\eta_t \nabla \ell(\mathbf{w}_{t-1}, S_{B_t}), \sigma_t^2 \mathbb{I}_d)$. Thus the EFLD update reduces to the SGLD update: $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \nabla \ell(\mathbf{w}_{t-1}, S_{B_t}) + \mathcal{N}\left(0, \sigma_t^2 \mathbb{I}_d\right)$. □

Example 3.2 (Noisy Sign-SGD)**.** By taking $\rho_t = \eta_t$ and component-wise $\xi_j \in \{-1, 1\}$, $\pi_0(\xi_j) = 1$, $\psi(\theta) = \log(\exp(-\theta) + \exp(\theta))$ in exponential family update equation (5), the j-th component of exponential family distribution $p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B_t,\alpha_t})$ becomes $p_{\boldsymbol{\theta}_{B_t,\alpha_t,j}}(\xi_j) = \frac{\exp(\xi_j \boldsymbol{\theta}_{B_t,\alpha_t,j})}{\exp(-\boldsymbol{\theta}_{B_t,\alpha_t,j}) + \exp(\boldsymbol{\theta}_{B_t,\alpha_t,j})}$. Thus, the EFLD update reduces to a noisy version of Sign-SGD: $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \boldsymbol{\xi}_t$, $\boldsymbol{\xi}_{t,j} \sim p_{\boldsymbol{\theta}_{B_t,\alpha_t,j}}(\xi_j)$, $j \in [d]$, where $\boldsymbol{\theta}_{B_t,\alpha_t} = \nabla \ell(\mathbf{w}_{t-1}, S_{B_t})/\alpha_t$ is the scaled mini-batch gradient. □

3.2 Expected Stability of Exponential Family Langevin Dynamics

From Lemma 1, conditioned on a trajectory $\mathbf{w}_{0:(t-1)}$, mini-batches S_{B_t}, S_{A_t}, we can get generalization bound by suitably bounding the Le Cam Style Divergence (LSD) given by: $I_{A_t,B_t} = \int_{\boldsymbol{\xi}_t} \frac{\left(dP_{B_t,\boldsymbol{\xi}_t} - dP'_{B_t,\boldsymbol{\xi}_t}\right)^2}{dP_{A_t,\boldsymbol{\xi}_t}} d\boldsymbol{\xi}_t$. For EFLD, the density functions $dP_{B_t,\boldsymbol{\xi}_t}$ are exponential family densities $p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B_t,\alpha_t})$ as in (5)-(6), and we have the following bound on the per step LSD:

Theorem 1. *For a given set $\bar{S} \sim D^{n+1}$ and $\mathbf{w}_{t-1}$ at iteration $(t-1)$, let $\Delta_{t|\mathbf{w}_{t-1}}(\bar{S}) = \max_{z,z' \in \bar{S}} \|\nabla \ell(\mathbf{w}_{t-1}, z) - \nabla \ell(\mathbf{w}_{t-1}, z')\|_2$. Further, for a c_2-smooth log-partition function ψ, let the scaling $\alpha_{t|\mathbf{w}_{t-1}}$ be data-dependent such that $\alpha_{t|\mathbf{w}_{t-1}}^2 \geq 8c_2 \Delta_{t|\mathbf{w}_{t-1}}^2(S_{n+1})$. Then, we have*

$$I_{A_t,B_t} \leq 5c_2 \|\boldsymbol{\theta}_{B_t,\alpha_t} - \boldsymbol{\theta}_{B'_t,\alpha_t}\|_2^2 = \frac{5c_2}{2\alpha_{t|\mathbf{w}_{t-1}}^2} \left[\left\|\nabla \ell(\mathbf{w}_{t-1}, S_{B_t}) - \nabla \ell\left(\mathbf{w}_{t-1}, S'_{B_t}\right)\right\|_2^2\right] \,, \quad (7)$$

Note that S_{B_t} and S'_{B_t} only differ at samples z_n and z'_n. The above bound can now be directly applied to Lemma 1 to get expected stability based generalization bounds for any EFLD algorithm.

Theorem 2. *Consider an exponential family Langevin dynamics (EFLD) algorithm of the form* (5)-(6) *with a c_2-smooth log-partition function ψ. Then, for mini-batch size $b \leq n/2$, with $c = c_0\sqrt{5c_2}$ (with c_0 as in Lemma 1) and $\alpha_{t|}^2 \geq 8c_2 \Delta_{t|}^2(S_{n+1})$ (as in Theorem 1, with the conditioning on $\mathbf{w}_{t-1}$ hidden to avoid clutter), we have*

$$|\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]| \leq c\frac{1}{n} \mathop{\mathbb{E}}_{S_{n+1}} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{W_{0:(t-1)}} \frac{1}{\alpha_{t|}^2} \left[\|\nabla \ell(\mathbf{w}_{t-1}, z_n) - \nabla \ell(\mathbf{w}_{t-1}, z'_n)\|_2^2\right]} \,. \quad (8)$$

Remark 3.1. Theorem 2 captures the generalization error of SGLD, which is a special case of EFLD. Our bound has the same dependence on n, T, step size η_t as the bound in Li et al. (2020). However, our bound is numerically sharper because we replace the *gradient norms*, i.e., $\frac{1}{n}\sum_{z \in S} \|\ell(\mathbf{w}_t, z)\|$ in Li et al. (2020) and with gradient discrepancy $\|\nabla \ell(\mathbf{w}_t, z) - \nabla \ell(\mathbf{w}_t, z')\|$, which is quantitatively smaller than gradient norms as we show in the experiment section. The bound in Negrea et al. (2019) depends on *gradient incoherence* which is empirically smaller than gradient discrepancy as observed in the experiment section, their bound depends on $1/\sqrt{n}$, which is worse than the $1/n$ dependence in our bound.

Remark 3.2. EFLD can be extended to work with anisotropic noise by using $\boldsymbol{\theta}_{B_t,\boldsymbol{\alpha}_t} = \nabla \ell(\mathbf{w}_{t-1}, S_{B_t}) \oslash \boldsymbol{\alpha}_t$ in (6) where $\boldsymbol{\alpha}_t \in \mathbb{R}^p$ determines different scaling for each dimension and $\oslash$ denotes Hadamard division. Theorems 1 and 2 can be extended to such anisotropic noise by using $\boldsymbol{\alpha}$-scaled norms for the gradient discrepancy, i.e., $\|\mathbf{g} - \mathbf{g}'\|_{2,\boldsymbol{\alpha}}^2 = \sum_j (g_j - g'_j)^2/\alpha_j^2$. □

3.3 Proof Sketches of Main Results: Theorems 1 and 2

We focus on Theorem 1. To avoid clutter, we drop the subscript t for the analysis and note that the analysis holds for any step t. When the density $dP_{B,\boldsymbol{\xi}} = p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B,\alpha})$, by mean-value theorem,

for each $\boldsymbol{\xi}$, we have $p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B,\alpha}) - p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B',\alpha}) = \langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}, \nabla_{\tilde{\boldsymbol{\theta}}_{B,\alpha}} p_\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle$, for some $\tilde{\boldsymbol{\theta}}_{B,\alpha} = \gamma_{\boldsymbol{\xi}}\boldsymbol{\theta}_{B,\alpha} + (1-\gamma_{\boldsymbol{\xi}})\boldsymbol{\theta}'_{B,\alpha}$ where $\gamma_{\boldsymbol{\xi}} \in [0,1]$. Then,

$$I_{A,B} = \int_{\boldsymbol{\xi}} \frac{\left(p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B,\alpha}) - p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B',\alpha})\right)^2}{p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi} = \int_{\boldsymbol{\xi}} \frac{\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}, \boldsymbol{\xi} - \nabla_{\tilde{\boldsymbol{\theta}}_{B,\alpha}} \psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle^2 \, p_\psi^2(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})}{p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi} \,,$$

since $p_\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha}) = \exp(\langle \boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}_{B,\alpha}\rangle - \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}))\pi_0(\boldsymbol{\xi})$.

Handling Distributional Dependence of $\tilde{\boldsymbol{\theta}}_B$. Note that we cannot proceed with the analysis with the density term depending on $\tilde{\boldsymbol{\theta}}_{B,\alpha}$ since $\tilde{\boldsymbol{\theta}}_{B,\alpha}$ depends on $\boldsymbol{\xi}$. So, we first bound the density term depending on $\tilde{\boldsymbol{\theta}}_{B,\alpha}$ in terms of exponential family densities with parameters $\boldsymbol{\theta}_{B,\alpha}$ and $\boldsymbol{\theta}_{B,\alpha}$ using c_2-smoothness of ψ.

Lemma 2. *For some $\gamma_{\boldsymbol{\xi}} \in [0,1]$, $\tilde{\boldsymbol{\theta}}_{B,\alpha} = \gamma_{\boldsymbol{\xi}}\boldsymbol{\theta}_{B,\alpha} + (1-\gamma_{\boldsymbol{\xi}})\boldsymbol{\theta}'_{B,\alpha}$, we have*

$$\frac{\exp\left[\langle \boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}_{B,\alpha}\rangle - \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha})\right]}{\max\left(\exp\left[\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}_{B,\alpha})\right], \exp\left[\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B',\alpha}\rangle - \psi(\boldsymbol{\theta}_{B',\alpha})\right]\right)} \leq \exp\left[c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2\right] \,.$$

Bounding the Density Ratio. Next we focus on the density ratio $p_\psi^2(\boldsymbol{\xi},\tilde{\boldsymbol{\theta}}_{B,\alpha})/p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})$. By Lemma 2, it suffices to focus on $p_\psi^2(\boldsymbol{\xi},\boldsymbol{\theta}_{B,\alpha})/p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})$ or the equivalent term for $\boldsymbol{\theta}_{B',\alpha}$. We show that the density ratio can be bounded by another exponential family with parameters $(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})$.

Lemma 3. *For any $\boldsymbol{\xi}$, we have*

$$\frac{\exp\left[\langle \boldsymbol{\xi}, 2\boldsymbol{\theta}_{B,\alpha}\rangle - 2\psi(\boldsymbol{\theta}_{B,\alpha})\right]}{\exp\left[\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{A,\alpha}\rangle - \psi(\boldsymbol{\theta}_{A,\alpha})\right]} \leq \exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right] \exp\left[\langle \boldsymbol{\xi}, (2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\rangle - \psi(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})\right].$$

The analysis for the term $p_\psi^2(\boldsymbol{\xi},\boldsymbol{\theta}_{B',\alpha})/p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})$ is exactly the same.

Bounding the Integral. Ignoring multiplicative terms which do not depend on $\boldsymbol{\xi}$ for the moment, the analysis needs to bound an integral term of the form $\int_{\boldsymbol{\xi}} \langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}, \boldsymbol{\xi} - \nabla\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle^2 \; p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) d\boldsymbol{\xi}$, and a similar term with $p_\psi^2(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha})$. First, note that $\nabla\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha}) = \tilde{\boldsymbol{\mu}}_{B,\alpha}$, the expectation parameter for $p_\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})$ Wainwright & Jordan (2008); Banerjee et al. (2005). The integral, however, is with respect to $p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})$. We handle this discrepancy by using $\boldsymbol{\xi} - \nabla\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha}) = (\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]) + (\mathbb{E}[\boldsymbol{\xi}] - \nabla\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha}))$, and decomposing as sum-of-squares. Quadratic form for the first term yields the covariance $\mathbb{E}[(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^T] = \nabla^2\psi(\boldsymbol{\theta}_{2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}}) \leq c_2\mathbb{I}$, by smoothness. The second term depends on the difference of gradients $\nabla\psi(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) - \nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha})$ which, using smoothness and additional analysis, can be bounded by the norm of $(\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})$. All the pieces can be put together to get the bound in Theorem 1, which when used in Lemma 1 yields Theorem 2.

3.4 Optimization Guarantees for EFLD

We now establish optimization guarantees for two examples of EFLD, i.e., Noisy Sign-SGD with Bernoulli noise over $\{-1,+1\}$ and SGLD with Gaussian noise.

Noisy Sign-SGD. For mini-batch B_t and scaling α_t, mini-batch Noisy Sign-SGD updates the parameters as $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t\boldsymbol{\xi}_t$, where each component $j \in [d]$

$$\boldsymbol{\xi}_{t,j} \sim p_{\boldsymbol{\theta}_{B_t,\alpha_t,j}}(x) = \frac{\exp(x\boldsymbol{\theta}_{B_t,\alpha_t,j})}{\exp(-\boldsymbol{\theta}_{B_t,\alpha_t,j}) + \exp(\boldsymbol{\theta}_{B_t,\alpha_t,j})}, \; x \in \{-1,+1\} \tag{9}$$

where $\boldsymbol{\theta}_{B_t,\alpha_t} = \nabla\ell(\mathbf{w}_{t-1}, S_{B_t})/\alpha_t$ is the scaled mini-batch gradient. The full-batch version uses parameters $\mathbb{E}_{B_t}[\boldsymbol{\theta}_{B_t,\alpha_t}] = \nabla L_S(\mathbf{w}_{t-1})$ For the optimization analysis, we assume that the loss is smooth and mini-batch gradients are unbiased, symmetric, and sub-Gaussian.

Assumption 1. *The loss function L_S satisfies: for all $\mathbf{w}$ and $\mathbf{w}'$, for some non-negative constant $\vec{K} := [K_1, \ldots, K_d]$, we have $L_S(\mathbf{w}) \leq L_S(\mathbf{w}') + \nabla L_S(\mathbf{w}')^T(\mathbf{w} - \mathbf{w}') + \frac{1}{2}\sum_i K_i(\mathbf{w}_i - \mathbf{w}'_i)^2$.*

Assumption 2. *Given $\mathbf{w}_{t-1}$, the mini-batch gradient $\nabla\ell(\mathbf{w}_{t-1}, S_{B_t})$ is (a) unbiased, i.e., $\mathbb{E}_{B_t|\mathbf{w}_{t-1}}\nabla\ell(\mathbf{w}_{t-1}, S_{B_t}) = \nabla L_S(\mathbf{w}_{t-1})$; (b) symmetric, i.e., the density $p(x)$ of $x \equiv \nabla\ell(\mathbf{w}_{t-1}, S_{B_t})$ is symmetric around its expectation $L_S(\mathbf{w}_{t-1})$: $p(x) = p(2\nabla L_S(\mathbf{w}_{t-1}) - x)$ and (c) sub-Gaussian, i.e., for any $\lambda > 0$, any $\mathbf{v}$ s.t. $\|\mathbf{v}\|_2 = 1$, $\mathbb{E}_{B_t|\mathbf{w}_{t-1}} \exp \lambda\langle \mathbf{v}, \nabla\ell(\mathbf{w}_{t-1}, S_{B_t}) - \nabla L_S(\mathbf{w}_{t-1})\rangle \leq \exp(\lambda^2\kappa_t^2/2)$ for some constant $\kappa_t > 0$.*

Based on the assumptions, we have the following optimization guarantee for mini-batch noisy Sign-SGD. We defer the optimization guarantee for full-batch noisy Sign-SGD to Appendix D.

Theorem 3. *Under Assumption 1 and 2, for mini-batch noisy Sign-SGD with step size $\eta_t = 1/\sqrt{T}$, α_t satisfying $c \geq \alpha_t \geq \max[\sqrt{2}\kappa_t, 4\|\nabla L_S(\mathbf{w}_{t-1})\|_\infty]$, we have for any S and any initialization $\mathbf{w}_0$*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2^2\right] \leq \frac{4c}{\sqrt{T}}\left(L_S(\mathbf{w}_0) - L_S(\mathbf{w}^*) + \frac{1}{2}\|\vec{K}\|_1\right), \tag{10}$$

where the expectation is taken over the randomness of algorithm.

SGLD. We acknowledge that the following optimization result of SGLD exists in various forms, as noisy gradient descent algorithms have been studied in literature such as differential privacy, where SGLD can be viewed as DP-SGD (Bassily et al., 2014; Wang & Xu, 2019) and the proof technique boils down to bounding the stochastic variance of the noisy gradient (Shamir & Zhang, 2013).

Theorem 4. *Under Assumption 1 and 2, with $K_i = K, \forall i \in [d]$, for any S, SGLD (EFLD with step size $\rho_t = \sqrt{\eta_t\sigma_t}$, $\alpha_t = \sqrt{\sigma_t/\eta_t}$), $|B_t| = b$, and $\eta_t = \frac{1}{\sqrt{T}}$, can achieve*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L_S(\mathbf{w}_t)\|^2 \leq O\left(\frac{1}{\sqrt{T}}\right) + O\left(K\frac{p\sum_{t=1}^{T}\alpha_t^4 + \log T}{\sqrt{T}}\right), \tag{11}$$

where the expectation is over the randomness of the algorithm.

The error rate of SGLD depends on the noise variance α_t. One can choose a decaying noise variance such as $\alpha_t = 1/\sqrt[4]{t}$ to guarantee the convergence. Then the rate will become $O(\log T/\sqrt{T})$. We note that similar to the optimization guarantees of DP-SGD, the convergence rate depends on the dimension of the gradient p due to the isotropic Gaussian noise. Special noise structures such as anisotropic noise that aligned with the gradient structure can reduce the dependence on dimension (Kairouz et al., 2020; Zhang et al., 2021; Asi et al., 2021; Zhou et al., 2020).

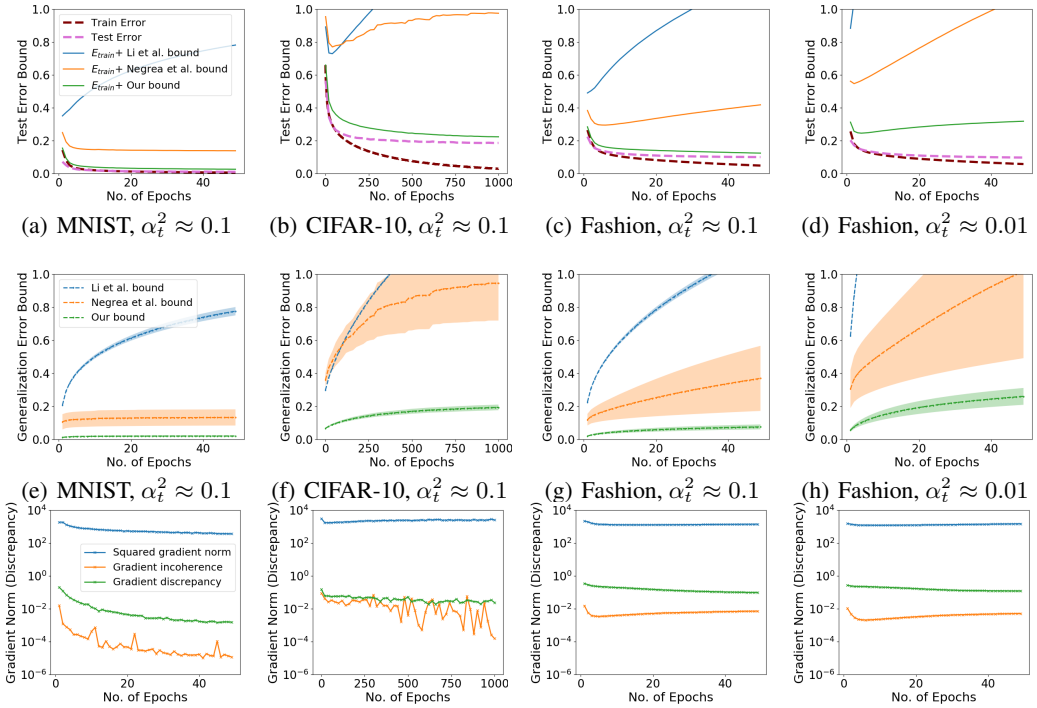
4 Experiments

In this section, we conduct a series of experiments to evaluate our generalization error bounds. For SGLD, we aim to compare the proposed bound in Theorem 2 with existing bounds in Li et al. (2020) and Negrea et al. (2019) for various datasets. Note that we do not consider the bound in Haghifam et al. (2020) as a baseline because the bound in Haghifam et al. (2020) only applies to full-batch setting, but we focus on noisy mini-batch gradient descent. We note that the squared norm of "two-sample incoherence" appeared in Haghifam et al. (2020) (Theorem 4.2) roughly has the same order of magnitude as the gradient incoherence appeared in Negrea et al. (2019). We also evaluate the optimization performance of proposed Noisy Sign-SGD by comparing it with the original sign-SGD (Bernstein et al., 2018a) and present the corresponding generalization bound in Theorem 2.

The details of our model architectures, learning rate scheduling, hyper-parameter selections and additional experimental results can be found in Appendix E. We acknowledge that we did not achieve the state-of-the art predictive performance, mainly due to the simplicity of our model architectures. With more complex model and further tuning, the prediction results could be improved.

4.1 Stochastic Gradient Langevin Dynamics

Comparison with existing work. We have derived theoretical generalization error bounds that depend on the data-dependent quantity *gradient discrepancy*, i.e., $\|\nabla\ell(\mathbf{w}_t, z_n) - \nabla\ell(\mathbf{w}_t, z_n')\|_2^2$. Existing bounds in Li et al. (2020) and Negrea et al. (2019) have also improved the Lipschitz constant in Mou et al. (2018) to a data-dependent quantity. As shown in Figure 1 (a)-(d), by combining with the empirical training error, all three generalization error bounds can be used to bound the empirical test error, but our bound is able to generate a much tighter upper bound. Such difference is mainly due to the fact that we replace the squared *gradient norm* in Li et al. (2020) and the squared

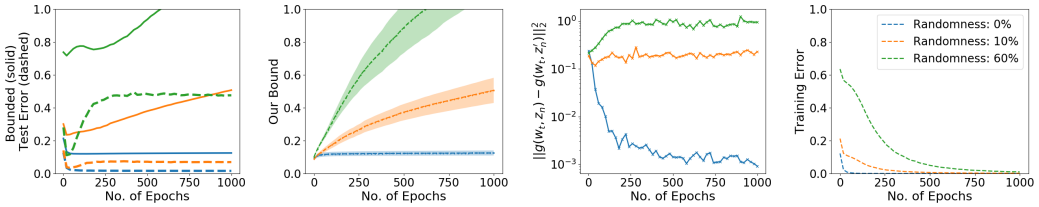


(a) MNIST, $\alpha_t^2 \approx 0.1$ (b) CIFAR-10, $\alpha_t^2 \approx 0.1$ (c) Fashion, $\alpha_t^2 \approx 0.1$ (d) Fashion, $\alpha_t^2 \approx 0.01$

(e) MNIST, $\alpha_t^2 \approx 0.1$ (f) CIFAR-10, $\alpha_t^2 \approx 0.1$ (g) Fashion, $\alpha_t^2 \approx 0.1$ (h) Fashion, $\alpha_t^2 \approx 0.01$

(i) MNIST, $\alpha_t^2 \approx 0.1$ (j) CIFAR-10, $\alpha_t^2 \approx 0.1$ (k) Fashion, $\alpha_t^2 \approx 0.1$ (l) Fashion, $\alpha_t^2 \approx 0.01$

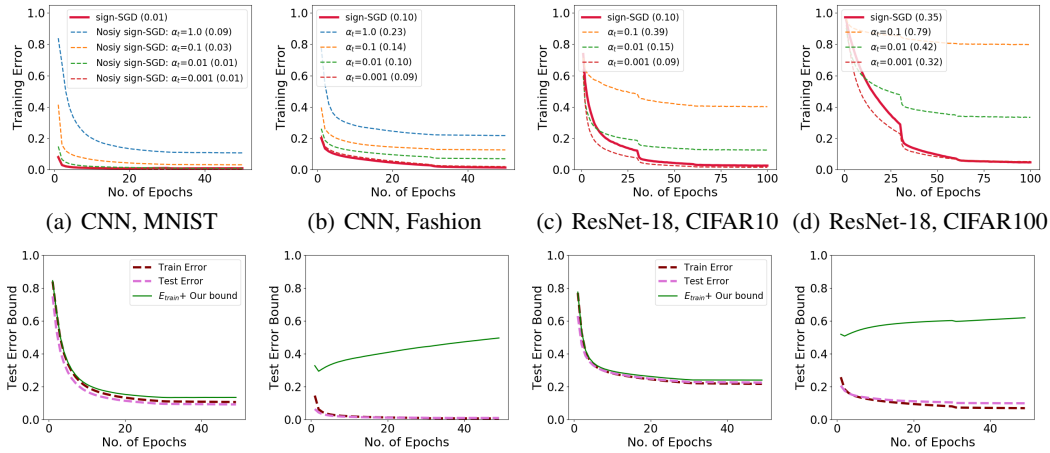
Figure 1: Numerical results for training CNN using SGLD ($\sigma_t = \sqrt{2\eta_t/\beta_t}$) on MNIST, Fashion-MNIST and CIFAR-10. X-axis shows the number of training epochs. (a)-(d) shows our bound is non-vacuous and can be used to bound the empirical test error. (e)-(h) compare our bound with the existing bounds and show the effect on α_t^2. (i)-(l) show the key factors in each bound, i.e., the squared gradient norm in Li et al. (2020), the gradient incoherence in Negrea et al. (2019), and the gradient discrepancy in our bound. Our bounds are numerically sharper than existing bounds, and larger α_t^2 leads to tighter generalization bounds which is consistent with the theoretical analysis.



(a) Bounded Test Error (b) Our Bound (c) Gradient Discrepancy (d) Training Error

Figure 2: Numerical results for training CNN using SGLD ($\sigma_t = 0.2\eta_t$) on a subset of MNIST ($n = 10000$) with different randomness on labels. (a) demonstrates that, as the randomness increases, the empirical test error (dashed lines) increases but still can be bounded by our generalization bound by combining the empirical training error (solid lines). (b) presents our bound in Theorem 2. (c) shows the gradient discrepancy $\|\nabla \ell(\mathbf{w}_t, z_n) - \nabla \ell(\mathbf{w}_t, z_n')\|_2^2$. (d) plots the training error. The gradient discrepancy increases as randomness increases, so does our generalization bound.

norm of *gradient incoherence* in Negrea et al. (2019) with the gradient discrepancy. Results in Figure 1 (e)-(h) show that our bounds are much sharper than those of Li et al. (2020) because our gradient discrepancy (Figure 1 (i)-(l)) is usually 2-4 order of magnitude smaller than the squared gradient norms appeared in Li et al. (2020). Our bounds are also sharper than those of Negrea et al. (2019) due to an improved dependence on n from an order of $1/\sqrt{n}$ to $1/n$. Note that, even though the gradient incoherence in Negrea et al. (2019) is about 1 to 2 order of magnitude smaller than the gradient discrepancy for simple problems such as MNIST and Fashion-MNIST, the difference between the gradient incoherence and our gradient discrepancy reduces as the problem becomes harder (see results for CIFAR-10 in Figure 1(j)).



(a) CNN, MNIST (b) CNN, Fashion (c) ResNet-18, CIFAR10 (d) ResNet-18, CIFAR100

(e) MNIST, $\alpha_t = 1$ (f) MNIST, $\alpha_t = 0.01$ (g) Fashion, $\alpha_t = 1$ (h) Fashion, $\alpha_t = 0.01$

Figure 3: (a)-(d) show the training dynamics of CNN on MNIST and Fashion-MNIST, and ResNet-18 on CIFAR-10 and CIFAR-100 using noisy sign-SGD with different scaling α_t. Legends indicate the choice of α_t and the numbers in brackets are test errors at convergence. As $\alpha_t \to 0$, Nosiy sign-SGD matches both the optimization trajectory as well as the final test accuracy of the original sign-SGD (Bernstein et al., 2018a). (e)-(f) show that empirical test error can be bounded by our bound and the corresponding training error. The larger α_t is the sharper our bound is.

Effect of Randomness. Motivated by Zhang et al. (2017), we train CNN with SGLD on a smaller subset of MNIST dataset ($n = 10000$) with randomly corrupted labels. The corruption fraction varies from 0% (without label corruption) to 60%. As shown in Figure 2 (d), for long enough training time, all experiments with different level of label randomness can achieve almost zero training error. However, the one with higher level of randomness has higher generalization/test error (Figure 2 (a) dashed lines). Our generalization bound also becomes larger as the randomness increases since the corresponding gradient discrepancy increases.

4.2 Noisy Sign-SGD

Optimization. Figure 3 (a)-(d) show the training dynamics of Noisy Sign-SGD under various selections of α_t. As $\alpha_t \to 0$, Noisy Sign-SGD matches both the optimization trajectory as well as the final test accuracy of the original Sign-SGD (Bernstein et al., 2018a). However, as α_t increases, the probability of getting 1 approaches 0.5, and ξ_t approximates a uniform distribution. As a result, the corresponding Noisy Sign-SGD still converges, but the generalization performance is much worse.

Generalization Bound. Figure 3(e)-(f) show that our bound successfully bounds the empirical test error. The larger α_t is the sharper the upper bound is. However, larger α_t would slow down and adversely affect the optimization, e.g., Figure 3 (a)-(d) blue and orange lines. In practice, one needs to balance the optimization error and generalization by choosing a suitable scaling α_t.

5 Conclusions

Inspired by recent advances in stability based and information theoretic approaches to generalization bounds (Mou et al., 2018; Pensia et al., 2018; Negrea et al., 2019; Li et al., 2020; Haghifam et al., 2020), we have presented a framework for developing such bounds based on expected stability for noisy stochastic iterative (NSI) learning algorithms. We have also introduced Exponential Family Langevin Dynamics (EFLD), a family of noisy gradient descent algorithms based on exponential family noise, including SGLD and Noisy Sign-SGD as two special cases. We have developed an expected stability based generalization bound applicable to any EFLD algorithm with a $O(1/n)$ sample dependence and a dependence on gradient incoherence, rather than gradient norms. Further, we have provided optimization guarantees for EFLD and establish such results for Noisy Sign-SGD and SGLD. Our experiments on various benchmarks illustrate that our bounds are non-vacuous and quantitatively much sharper than existing bounds (Li et al., 2020; Negrea et al., 2019).

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 1709–1720. Curran Associates, Inc., 2017.

Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pp. 383–392. PMLR, 2021.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

Ole Barndorff-Nielsen. *Information and exponential families: in statistical theory*. John Wiley & Sons, 2014.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 2019.

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

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018a.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2018b.

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

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

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

Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 587–591. IEEE, 2019.

Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 74–86, 2018.

Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. In *NeurIPS*, 2020.

Xiangyi Chen, Tiancong Chen, Haoran Sun, Zhiwei Steven Wu, and Mingyi Hong. Distributed training with heterogeneous data: Bridging median-and mean-based algorithms. *arXiv preprint arXiv:1906.01736*, 2019.

Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9770–9780, 2018.

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

Peter Grünwald, Thomas Steinke, and Lydia Zakynthinou. Pac-bayes, mac-bayes and conditional mutual information: Fast rate bounds that handle general vc classes. *arXiv preprint arXiv:2106.09683*, 2021.

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. *Advances in Neural Information Processing Systems*, 2020.

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

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.

Fredrik Hellström and Giuseppe Durisi. Fast-rate loss bounds via conditional information measures with applications to neural networks. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 952–957. IEEE, 2021.

Peng Jiang and Gagan Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2525–2536. Curran Associates, Inc., 2018.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pp. 1724–1732, 2017.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *arXiv preprint arXiv:1902.04811*, 2019.

Richeng Jin, Yufan Huang, Xiaofan He, Tianfu Wu, and Huaiyu Dai. Stochastic-sign sgd for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

Peter Kairouz, Mónica Ribero, Keith Rush, and Abhradeep Thakurta. Dimension independence in unconstrained private erm via adaptive preconditioning. *arXiv preprint arXiv:2008.06570*, 2020.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical Report Vol. 1. No. 4., University of Toronto, 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.

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. URL `https://openreview.net/forum?id=SkxxtgHKPS`.

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*, pp. 605–638. PMLR, 2018.

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*, 2019.

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

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

David Pollard. *A user's guide to measure theoretic probability*. Number 8. Cambridge University Press, 2002.

Borja Rodríguez-Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm. In *2020 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2021.

William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pp. 506–514, 1978.

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

Igal Sason and Sergio Verdu. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62, 2016.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pp. 71–79. PMLR, 2013.

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

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Di Wang and Jinhui Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1182–1189, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, ICML '11, pp. 681–688, 2011.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

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

Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, and Christopher De Sa. Swalp: Stochastic weight averaging in low-precision training. *36th International Conference on Machine Learning (ICML)*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Sy8gdB9xx`.

Huanyu Zhang, Ilya Mironov, and Meisam Hejazinia. Wide network learning with differential privacy. *arXiv preprint arXiv:2103.01294*, 2021.

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

Yingxue Zhou, Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private sgd with gradient subspace identification. In *International Conference on Learning Representations*, 2020.

A Related Work

Uniform stability. Uniform stability has been a classical approach for bounding generalization error (Bousquet & Elisseeff, 2002; Bousquet et al., 2020; Feldman & Vondrak, 2018; 2019), pioneered by Rogers & Wagner (1978); Devroye & Wagner (1979). Recently, uniform stability has been shown powerful in analyzing the stability of SGLD (Welling & Teh, 2011). Mou et al. (2018) prove the uniform stability of SGLD by showing that the uniform stability can be bounded by the squared Hellinger distance, and further they establish discretized Fokker-Planck equations for analyzing the squared Hellinger distance. Then they provide the uniform stability bound of SGLD as $\frac{L}{n}\sqrt{\sum_t \eta_t^2/\sigma_t^2}$ which depends on L, the Lipschitz constant for the gradient, and the step size $\eta_t \leq \frac{\sigma_t \ln 2}{L}$ (Mou et al., 2018; Li et al., 2020). Recently, Li et al. (2020) follow up on Mou et al. (2018) and derive a data-dependent bound using based on Bayes-stability, and got a bound of the form $\frac{c}{n}\sqrt{\sum_t \eta_t^2 \mathbf{g}_e(t)/\sigma_t^2}$, where $\mathbf{g}_e(t)$ is the expected gradient norm square at step t. Their bound improves the Lipschitz constant to the expected gradient norm square. Recently, Bassily et al. (2019) analyze the uniform stability of differentially private SGD (DP-SGD) for convex optimization by showing the gradient update is a non-expansive operation, which is the key fact in proving the stability of SGD (Hardt et al., 2016). The approach in Hardt et al. (2016) can extend to non-convex setting as well, however it requires fast decaying in step size as $\eta_t = O(1/t)$. Bassily et al. (2020) provide stability analysis of SGD for convex and non-smooth functions.

Information-theoretic bound. Besides the works mentioned above, other theories of deriving generalization bounds for noisy iterative algorithms have been proposed via information-theoretic approaches (Russo & Zou, 2016; Xu & Raginsky, 2017). Such results show that the generalization error of any learning algorithm can be bounded as $O(\sqrt{I(S;\mathbf{w})/n})$, where $I(S;\mathbf{w})$ is the mutual information between the algorithm input S and the algorithm output $\mathbf{w}$. Thus, recent works following this approach focus on bounding the mutual information for a broad class of iterative algorithms, including SGLD which obtains a $O(\sqrt{\log T/n})$ generalization bound by choosing $\eta_t = O(1/t)$, where T is the total number of iterations (Pensia et al., 2018; Bu et al., 2019). Subsequent improvements to this technique were made by Negrea et al. (2019); Haghifam et al. (2020); Rodríguez-Gálvez et al. (2021) to prove data-dependent generalization bounds that do not depend on the Lipschitz constant of the loss function and obtain $\sqrt{\sum_t \eta_t/n}$ bounds. Especially, Haghifam et al. (2020); Zhou et al. (2021) introduce generalization bounds based on conditional mutual information inspired by Steinke & Zakynthinou (2020), leading to tighter bounds than those based on mutual information. Recently, Rodríguez-Gálvez et al. (2021) extend the result of Haghifam et al. (2020) from full-batch gradient to stochastic setting. Neu (2021) extend this information-theoretic approach to derive generalization bound for vanilla SGD. Hellström & Durisi (2021) provide a fast-rate bound for bounded loss functions based via Conditional Information Measures (Grünwald et al., 2021; Hellström & Durisi, 2020).

Noisy iterative algorithms. Introducing additional noise in the stochastic gradient has been popular in training deep nets. Noisy iterative methods have proven to be useful for machine learning applications, especially for deep neural networks in terms of escaping from saddle points (Jin et al., 2017; 2019), preserving privacy (Bassily et al., 2020; Wang & Xu, 2019), boosting generalization and stability (Mou et al., 2018; Li et al., 2020). SGLD (Welling & Teh, 2011) has been one of the most popular noisy iterative algorithms for non-convex learning problems, where an isotropic Gaussian noise is added to the stochastic gradient. As a special case, SGLD can be viewed as a differentially private SGD algorithm (DP-SGD) (Bassily et al., 2020; Wang & Xu, 2019) with constant noise variance. Uniform stability has also been popular in the differential privacy literature for analyzing the generalization error bound of DP-SGD algorithms. Recently, noise has been proven to be useful in Sign-SGD (Bernstein et al., 2018a;b; Chen et al., 2019) which is proposed to reduce communication

cost in distributed learning. Algorithm such as noisy sign-SGD will first add symmetric noises to the stochastic gradient, then take the sign of the noisy stochastic gradient to update the parameters (Chen et al., 2019; Jin et al., 2020). Bernstein et al. (2018b); Chen et al. (2019); Jin et al. (2020) have shown that when noise is unimodal and symmetric, sign-SGD can guarantee convergence to stationary point.

B Analysis and Proofs for Expected Stability (Section 2)

Proposition 1. *Let $S_n \sim D^n$ and let S_n' be a dataset obtained by replacing $z_n \in S_n$ with $z_n' \sim D$. Let $A(S_n), A(S_n')$ respectively denote the distributions over the hypothesis space $\mathcal{W}$ obtained by running randomized algorithm A on S_n, S_n'. Assume that for $S_n \sim D^n, \forall z \in \mathcal{Z}$, $\mathbb{E}_{W\sim A(S_n)}[\ell^2(W,z)] \leq c_0/2, c_0 > 0$. With $H(\cdot,\cdot)$ denoting the Hellinger divergence, we have*

$$|\mathbb{E}_{S_n\sim D^n}[L_D(A(S_n)) - L_S(A(S_n))]| \leq c_0 \mathbb{E}_{S_n\sim D^n}\mathbb{E}_{z_n'\sim D}\sqrt{2H^2\big(A(S_n), A(S_n')\big)}\,. \tag{2}$$

Proof. Let $S = (z_1, \ldots, z_n)$ and $\tilde{S}' = (z_1', \ldots, z_n')$ two independent random samples and let $S^{(i)} = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)$ be the sample that is identical to S except in the i-th example where we replace z_i with z_i'.

$$\begin{aligned}
&\mathbb{E}_{S\sim D^n}[L_D(A(S)) - L_S(A(S))] \\
&= \mathbb{E}_S\mathbb{E}_{\tilde{S}'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^n\Big(\ell(A(S^{(i)}), z_i') - \ell(A(S), z_i')\Big)\right] \\
&= \mathbb{E}_S\mathbb{E}_{\tilde{S}'}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_A\Big(\ell(A(S^{(i)}), z_i') - \ell(A(S), z_i')\Big) \\
&\overset{(a)}{=} \mathbb{E}_S\mathbb{E}_{\tilde{S}'}\frac{1}{n}\sum_{i=1}^n\left(\int_{\mathbb{R}^d}\ell(\mathbf{w}; z_i')p^{(i)}(\mathbf{w})d\mathbf{w} - \int_{\mathbb{R}^d}\ell(\mathbf{w}; z_i')p(\mathbf{w})d\mathbf{w}\right) \\
&= \mathbb{E}_S\mathbb{E}_{\tilde{S}'}\frac{1}{n}\sum_{i=1}^n\int_{\mathbb{R}^d}\ell(\mathbf{w}; z_i')\Big(p^{(i)}(\mathbf{w})d\mathbf{w} - p(\mathbf{w})\Big)d\mathbf{w} \\
&= \mathbb{E}_S\mathbb{E}_{\tilde{S}'}\frac{1}{n}\sum_{i=1}^n\int_{\mathbb{R}^d}\ell(\mathbf{w}; z_i')\left(\sqrt{p^{(i)}(\mathbf{w})} + \sqrt{p(\mathbf{w})}\right)\left(\sqrt{p^{(i)}(\mathbf{w})} - \sqrt{p(\mathbf{w})}\right)d\mathbf{w} \\
&\overset{(b)}{\leq} \mathbb{E}_S\mathbb{E}_{\tilde{S}'}\frac{1}{n}\sum_{i=1}^n\left\{\left(\int_{\mathbb{R}^d}\ell(\mathbf{w}; z_i')^2\left(\sqrt{p^{(i)}(\mathbf{w})} + \sqrt{p(\mathbf{w})}\right)^2 d\mathbf{w}\right)^{\frac{1}{2}}\left(\int_{\mathbb{R}^d}\left(\sqrt{p^{(i)}(\mathbf{w})} - \sqrt{p(\mathbf{w})}\right)^2 d\mathbf{w}\right)^{\frac{1}{2}}\right\}
\end{aligned}$$

where (a): $p^{(i)}(\mathbf{w})$ is the distribution obtained by $A(S^{(i)})$, and (b) follows by Cauchy-Schwartz inequality. Focusing on the first integral, we have:

$$\begin{aligned}
\left(\int_{\mathbb{R}^d}\ell(\mathbf{w}; z_i')^2\left(\sqrt{p^{(i)}(\mathbf{w})} + \sqrt{p(\mathbf{w})}\right)^2 d\mathbf{w}\right)^{1/2} &\leq \left(2\int_{\mathbb{R}^d}\ell(\mathbf{w}; z)^2 p^{(i)}(\mathbf{w})d\mathbf{w} + 2\int_{\mathbb{R}^d}\ell(\mathbf{w}; z)^2 p(\mathbf{w})d\mathbf{w}\right)^{1/2} \\
&\leq \left(c_0^2 + c_0^2\right)^{1/2} = \sqrt{2}c_0\,.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}_{S\sim D^n}[L_D(A(S)) - L_S(A(S))] &\leq \mathbb{E}_S\mathbb{E}_{\tilde{S}'}\frac{1}{n}\sum_{i=1}^n c_0\sqrt{2H^2\left(A(S), A(S_n')\right)} \\
&= c_0\mathbb{E}_{S\sim D^n}\mathbb{E}_{z_n'\sim D}\sqrt{2H^2\big(A(S), A(S_n')\big)}\,.
\end{aligned}$$

That completes the proof. □

Proposition 2. *For any distributions P and P', $2H^2(P,P') \leq \min\left\{KL(P,P'), \sqrt{\frac{1}{2}KL(P,P')}\right\}$.*

Proof. For the first part, note that:

$$\begin{aligned}
KL(P, P') &= \int \log \frac{p(\mathbf{w})}{p'(\mathbf{w})} p(\mathbf{w})\, d\mathbf{w} \\
&= 2 \int \left\{ -\log \sqrt{\frac{p'(\mathbf{w})}{p(\mathbf{w})}} \right\} p(\mathbf{w})\, d\mathbf{w} \\
&\overset{(a)}{\geq} 2 \int \left\{ 1 - \sqrt{\frac{p'(\mathbf{w})}{p(\mathbf{w})}} \right\} p(\mathbf{w})\, d\mathbf{w} \\
&= 2 \int \left(p(\mathbf{w}) - \sqrt{p(\mathbf{w}) p'(\mathbf{w})} \right))\, d\mathbf{w} \\
&= \int \left(p(\mathbf{w}) + p'(\mathbf{w}) - 2\sqrt{p(\mathbf{w}) p'(\mathbf{w})} \right)\, d\mathbf{w} \\
&= \int \left(\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})} \right)^2\, d\mathbf{w} \\
&= 2H^2(P, P')\ ,
\end{aligned}$$

where (a) follows since for $x > 0$, $-\log x = -\log(1 - (1 - x)) \geq 1 - x$.

For the second part, note that:

$$\begin{aligned}
2H^2(P, P') &= \int_{\mathbf{w}} (\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})})^2 d\mathbf{w} \\
&= \int_{\mathbf{w}} |\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})}||\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})}|\, d\mathbf{w} \\
&\leq \int_{\mathbf{w}} |\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})}||\sqrt{p(\mathbf{w})} + \sqrt{p'(\mathbf{w})}|\, d\mathbf{w} \\
&= \int_{\mathbf{w}} |p(\mathbf{w}) - p'(\mathbf{w})|\, d\mathbf{w} \\
&= TV(P, P')\ ,
\end{aligned}$$

where $TV(P, P')$ denotes the total variation distance Pollard (2002). Further, from Pinsker's inequality Tsybakov (2008), we have

$$TV(P, P') \leq \sqrt{\frac{1}{2} KL(P, P')}\ .$$

Combining the two results completes the proof. □

B.1 Proofs for Section 2.2

Similar results have appeared in (Li et al., 2020; Bun et al., 2018), we give a simple proof.

Lemma 4. *Let Q, Q', R be any three distributions such that Q, Q' are both absolutely continuous w.r.t. R. Then, for $s \in (0, 1)$*

$$KL\left(sQ + (1-s)R \,\middle\|\, sQ' + (1-s)R\right) \leq \frac{s^2}{1-s} \int_{\mathbf{w}} \frac{(Q(\mathbf{w}) - Q'(\mathbf{w}))^2}{R(\mathbf{w})} d\mathbf{w}\ . \tag{12}$$

Proof. Let $U = sQ' + (1-s)R$. Then, with $F(x) = Q(x) - Q'(x)$, we have

$$\begin{aligned}
KL(sQ + (1-s)R \| sQ' + (1-s)R) &= KL(U + s(Q - Q') \| U) \\
&= \int (U(x) + sF(x)) \log \left(\frac{U(x) + sF(x)}{U(x)} \right) dx \\
&= \int (U(x) + sF(x)) \log \left(1 + \frac{sF(x)}{U(x)} \right) dx
\end{aligned}$$

$$= \int sF(x) + \frac{s^2F^2(x)}{2U(x)} - \frac{s^3F^3(x)}{6U^2(x)} + \frac{s^4F^4(x)}{12U^3(x)} - \cdots dx$$
$$= s^2 \int \frac{F^2(x)}{U(x)} \left(\frac{1}{2} - \frac{sF(x)}{6U(x)} + \frac{s^2F^2(x)}{12U^2(x)} - \cdots \right) dx \ ,$$

where the first term vanishes since $\int F(x)dx = \int (Q(x) - Q'(x))dx = 0$. This is the reason the dependency is on s^2, not s. With $W(x) = sF(x)/U(x)$, noting that $W(x) > -1$, and

$$\left(\frac{1}{2} - \frac{W(x)}{6} + \frac{W^2(x)}{12} - \cdots \right) = \frac{(1+W(x))\log(1+W(x)) - W(x)}{W^2(x)} \leq 1 \ , \tag{13}$$

we have

$$\begin{aligned} KL(sP + (1-s)R \| sQ + (1-s)R) &= s^2 \int \frac{F^2(x)}{U(x)} \left(\frac{(1+W(x))\log(1+W(x)) - W(x)}{W^2(x)} \right) dx \\ &\leq s^2 \int \frac{F^2(x)}{U(x)} dx \\ &\leq \frac{s^2}{(1-s)} \int \frac{(Q(x) - Q'(x))^2}{R(x)} dx \ . \end{aligned}$$

That completes the proof. □

Proposition 3. *Consider the mixture models* $Q_{t|} = \frac{1}{|G_1|} \sum_{B_t \in G_1} P_{B_t, \boldsymbol{\xi}_t}$, $Q'_{t|} = \frac{1}{|G_1|} \sum_{B_t \in G_1} P'_{B_t, \boldsymbol{\xi}_t}$, $R_t = \frac{1}{|G_0|} \sum_{A_t \in G_0} P_{A_t, \boldsymbol{\xi}_t}$. *Then, with* $s = \frac{|G_1|}{|G|} = \frac{\binom{n-1}{b-1}}{\binom{n}{b}} = \frac{b}{n}$, *we have*

$$P_{t|} = sQ_{t|} + (1-s)R_{t|} \ , \qquad \textit{and} \qquad P'_{t|} = sQ'_{t|} + (1-s)R_{t|} \ .$$

Proof. The proof follows the argument in the proof of Lemma 21 in Li et al. (2020). □

Lemma 5. *Consider a general NSI algorithm with updates of the form* (3) *with mini-batch size* $B_t = b$. *Then, for any trajectory* $\mathbf{w}_{0:t-1}$, *we have*

$$KL\left(P_{t|} \| P_{t|}\right) \leq \frac{b^2}{n^2} \frac{n}{n-b} \mathbb{E}_{B_t \in G_1} \mathbb{E}_{A_t \in G_0} \left[\int_{\boldsymbol{\xi}_t} \frac{\left(dP_{B_t, \boldsymbol{\xi}_t} - dP'_{B_t, \boldsymbol{\xi}_t} \right)^2}{dP_{A_t, \boldsymbol{\xi}_t}} d\boldsymbol{\xi}_t \right] \ . \tag{14}$$

Proof. By definition and Proposition 3,

$$\begin{aligned} KL\left(dP_{t|} \| P'_{t|}\right) &\overset{(a)}{\leq} \left(\frac{b}{n}\right)^2 \left(\frac{n}{n-b}\right) \int_{\mathbf{w}} \frac{(Q-Q')^2}{R} d\mathbf{w} \\ &= \frac{b^2}{n^2} \left(1 + \frac{b}{n-b}\right) \int_{\mathbf{w}} \frac{\left(\frac{1}{|G_1|} \left[\sum_{B \in G_1} dP_{B,\boldsymbol{\xi}}(\mathbf{w}) - \sum_{B \in G_1} P'_{B,\boldsymbol{\xi}}(\mathbf{w}) \right] \right)^2}{\frac{1}{|G_0|} \sum_{A \in G_0} dP_{A,\boldsymbol{\xi}}(\mathbf{w})} d\mathbf{w} \\ &\overset{(b)}{\leq} \frac{b^2}{n^2} \left(1 + \frac{b}{n-b}\right) \int_{\mathbf{w}} \frac{1}{|G_1|} \sum_{B \in G_1} \frac{\left(dP_{B,\boldsymbol{\xi}}(\mathbf{w}) - P'_{B,\boldsymbol{\xi}}(\mathbf{w}) \right)^2}{\frac{1}{|G_0|} \sum_{A \in G_0} dP_{A,\boldsymbol{\xi}}(\mathbf{w})} d\mathbf{w} \\ &\overset{(c)}{\leq} \frac{b^2}{n^2} \left(1 + \frac{b}{n-b}\right) \int_{\mathbf{w}} \frac{1}{|G_1|} \sum_{B \in G_1} \left\{ \left(dP_{B,\boldsymbol{\xi}}(\mathbf{w}) - P'_{B,\boldsymbol{\xi}}(\mathbf{w}) \right)^2 \frac{1}{|G_0|} \sum_{A \in G_0} \frac{1}{dP_{A,\boldsymbol{\xi}}(\mathbf{w})} \right\} d\mathbf{w} \\ &= \frac{b^2}{n^2} \left(1 + \frac{b}{n-b}\right) \frac{1}{|G_1||G_0|} \sum_{B \in G_1} \sum_{A \in G_0} \int_{\mathbf{w}} \frac{\left(dP_{B,\boldsymbol{\xi}}(\mathbf{w}) - P'_{B,\boldsymbol{\xi}}(\mathbf{w}) \right)^2}{dP_{A,\boldsymbol{\xi}}(\mathbf{w})} d\mathbf{w} \ , \end{aligned}$$

where (a) is from Lemma 4, (b)(c) is from Jensen's inequality since function $f(x) = x^2$ is convex and $f(x) = \frac{1}{x}$ is convex on $(0, \infty)$. That completes the proof. □

Lemma 1. *Consider a noisy stochastic iterative algorithms of the form* (3) *with mini-batch size* $b \leq n/2$*. Then, with* $c_1 = \sqrt{2}c_0$ *(with* c_0 *as in Proposition 1), we have*

$$|\mathbb{E}_{S_n}[L_D(A(S_n)) - L_{S_n}(A(S_n))]| \leq c_1 \frac{b}{n} \mathbb{E}_{S_n} \mathbb{E}_{z'_n} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{W_{0:(t-1)}} \mathop{\mathbb{E}}_{B_t \in G_1} \mathop{\mathbb{E}}_{A_t \in G_0} \left[\int_{\boldsymbol{\xi}_t} \frac{\left(dP_{B_t,\boldsymbol{\xi}_t} - dP'_{B_t,\boldsymbol{\xi}_t} \right)^2}{dP_{A_t,\boldsymbol{\xi}_t}} d\boldsymbol{\xi}_t \right]} . \tag{4}$$

Proof. Based on Proposition 1 and 2, we have

$$\begin{aligned} \mathbb{E}_{S \sim D^n}[L_D(A(S)) - L_S(A(S))] &\leq \sqrt{2KL\left(P_T \| P'_T\right)} \\ &\leq \sqrt{2 \sum_{t=1}^{T} P_{0:(t-1)} \left[KL\left(P_{t|} \| P'_{t|} \right) \right]} \end{aligned}$$

Bringing the result of Lemma 5 into the above inequality completes the proof. □

B.2 High Probability Generalization Bounds

Let $S \sim D^n$ with $S = (Z_1 \ldots, Z_n)$ corresponding to the training data. Consider the random variable $Y(S)$, the Scaled (by n) Generalization Error (SGE), defined as

$$Y(S) \triangleq \sum_{i=1}^{n} \mathbb{E}_A \big[\mathbb{E}_{Z \sim D}[\ell(A(S), Z)] - \ell(A(S), Z_i) \big] = n \big(L_D(A(S)) - L_S(A(S)) \big) . \tag{15}$$

Theorem 1 establishes a bound on $|\frac{1}{n}\mathbb{E}_S[Y(S)]|$. We now focus on establishing a high probability bound on $\frac{1}{n}(Y(S) - \mathbb{E}_S[Y(S)])$. Let $S' = (Z'_1, \ldots, Z'_n)$ be such that Z'_i is an independent copy of Z_i. Further, let $S'_i = (Z_1, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_n)$. Then, the change in SGE

$$Y(S) - Y(S'_i) = n \big[\big(L_D(A(S)) - L_D(A(S'_i)) \big) - \big(L_S(A(S)) - L_{S'_i}(A(S'_i)) \big) \big] \tag{16}$$

is a symmetric random variable, and is identically distributed for $i = 1, \ldots, n$. Our analysis is based on the following assumption:

Assumption 3. *The random variable* $(Y(S) - Y(S'_i))^2$ *is sub-Gaussian with* ψ_2*-norm* κ_A^2*, i.e.,* $\|(Y - Y'_i)^2\|_{\psi_2} = \sup_{q \geq 1} (\mathbb{E}_{S.S'} |Y(S) - Y(S'_i)|^{2q})^{1/q} / \sqrt{q} = \kappa_A^2$.

Assumption 3 implies that $\|Y - Y'_i\|_{\psi_2} \leq \kappa_A$. Note that since $Y(S) - Y(S'_i)$ is identically distributed for all i, κ_A is the same for all i. Further, κ_A is a property of the algorithm A, and can be viewed as a measure of stability. If $\kappa_A = O(1)$, i.e., swapping one point effectively leads to $O(\frac{1}{n})$ change in the generalization error, then A can be considered stable; on the other hand, if $\kappa_A = O(\sqrt{n})$, then A is not stable since the effective change in generalization error $\frac{1}{n}(Y(S) - Y(S'_i))$ is $O(1/\sqrt{n})$, the same order as typical generalization error $\frac{1}{n}Y(S)$ itself. The sharpness of the high-probability bound we present meaningfully depends on κ_A, with smaller values implying sharper bounds.

Theorem 5. *Under Assumption 3, with probability at least* $(1 - \delta)$ *over the draw* $S \sim D^n$*, we have*

$$L_D(A(S)) \leq L_S(A(S)) + \mathbb{E}_S \Big[L_D(A(S) - L_S(A(S)) \Big] + \max\left(\frac{1}{n}, \frac{c_1 \kappa_A}{\sqrt{n}} \right) \log\left(\frac{16}{\delta} \right) . \tag{17}$$

If $\kappa_A = O(1)$, i.e., swapping one point effectively leads to $O(\frac{1}{n})$ change in the generalization error, then A can be considered very stable; if $\kappa_A = O(\sqrt[4]{n})$, then A can be considered somewhat stable; and if $\kappa_A^2 = O(\sqrt{n})$, then A is not stable since the effective change in generalization error $\frac{1}{n}(Y(S) - Y(S'_i))$ is $O(1/\sqrt{n})$, the same order as the generalization error $\frac{1}{n}Y(S)$ itself.

Proof. Let $S = (Z_1, \ldots, Z_n)$ and let $S' = (Z'_1, \ldots, Z'_n)$ be such that Z'_i is an independent copy of Z_i. With $(x)_+ = max(x, 0)$, let

$$V^+ = V^+(S) = \sum_{i=1}^{n} \mathbb{E}_{S'}[(Y - Y'_i)_+^2] \,, \tag{18}$$

Our proof uses the following exponential version of the Efron-Stein inequality [Theorem 6.16 in Boucheron et al. (2013)], whose proof is based on a combination of the symmetric modified log-Sobolev inequality [Theorem 6.15 in Boucheron et al. (2013)] with the change of measure:

Theorem 6. *Let $Y = f(Z_1, \ldots, Z_n)$, where $Z_i, i = 1, \ldots, n$ are independent. Let $\theta, \lambda > 0$ be such that $\theta\lambda < 1$ and $\mathbb{E}[\exp(\lambda V^+/\theta)] < \infty$. Then,*

$$\log \mathbb{E}\left[e^{\lambda(Y - \mathbb{E}Y)}\right] \leq \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbb{E}\left[e^{\lambda V^+/\theta}\right] \,. \tag{19}$$

For the proof of Theorem 5, note that V^+ is sub-Gaussian with $\|V^+\|_{\psi_2} \leq n\kappa_A^2$. Let $\mu_+ = \mathbb{E}[V^+]$ and $\kappa_+ = \|V^+\|_{\psi_2}$. Then from Theorem 6, we have

$$\psi(\lambda) \triangleq \log \mathbb{E}\left[e^{\lambda(Y - \mathbb{E}Y)}\right] \leq \frac{\lambda\theta}{1 - \lambda\theta}\left[\frac{\lambda\mu_+}{\theta} + c\frac{\lambda^2\kappa_+^2}{\theta^2}\right].$$

Then, by Markov's inequality we have

$$\begin{aligned}\log \mathbb{P}(Y - \mathbb{E}Y > t) &\leq \psi(\lambda) - \lambda t \\ &\leq \frac{1}{1 - \lambda\theta}\left[\lambda^2\mu_+ + c\lambda^3\kappa_+^2/\theta - (1 - \lambda\theta)\lambda t\right] \,.\end{aligned}$$

Choosing $\theta = 1/(2\lambda)$, we have

$$\log \mathbb{P}(Y - \mathbb{E}Y > t) \leq 2\lambda^2\mu_+ + 4c\lambda^4\kappa_+^2 - \lambda t \,.$$

Consider choosing $\lambda = \min\left\{1, \frac{1}{c_0\sqrt{2\kappa_+}}\right\}$, where $c_0 = \max(1, \sqrt[4]{c})$. Since $\mu_+ \leq \kappa_+$, we have

$$\begin{aligned}&\log \mathbb{P}(Y - \mathbb{E}Y > t) \leq 2 - \min\left\{1, \frac{1}{c_0\sqrt{2\kappa_+}}\right\} t \\ \Rightarrow \quad &\mathbb{P}(Y - \mathbb{E}Y > t) \leq 8\exp\left(-\min\left\{1, \frac{1}{c_0\sqrt{2\kappa_+}}\right\} t\right) \,.\end{aligned}$$

With $t = n\epsilon$, and noting that $\kappa_+ \leq n\kappa_A^2$, with $c_1 = c_0\sqrt{2}$, we have

$$\mathbb{P}\left(Y - \mathbb{E}Y > n\epsilon\right) \leq 8\exp\left(-\min\left\{n, \frac{\sqrt{n}}{c_1\kappa_A}\right\}\epsilon\right) \,.$$

Now, we choose δ such that

$$8\exp\left(-\min\left\{n, \frac{\sqrt{n}}{c_1\kappa_A}\right\}\epsilon\right) \leq 8\exp\left(-n\epsilon\right) + 8\exp\left(-\frac{\sqrt{n}}{c_1\kappa_A}\epsilon\right) \leq \delta \,.$$

It suffices to choose ϵ such that

$$\begin{aligned}8\exp\left(-n\epsilon\right) \leq \delta/2 \qquad &\Rightarrow \qquad \epsilon \geq \frac{1}{n}\log\left(\frac{16}{\delta}\right) \\ 8\exp\left(-\frac{\sqrt{n}}{c_1\kappa_A}\epsilon\right) \leq \delta/2 \qquad &\Rightarrow \qquad \epsilon \geq \frac{c_1\kappa_A}{\sqrt{n}}\log\left(\frac{16}{\delta}\right) \,.\end{aligned}$$

As a result,

$$\mathbb{P}\left(\frac{1}{n}(Y - \mathbb{E}Y) > \max\left(\frac{1}{n}, \frac{c_1\kappa_A}{\sqrt{n}}\right)\log\left(\frac{16}{\delta}\right)\right) \leq \delta \,.$$

That completes the proof. □

C ANALYSIS AND PROOFS FOR EFLD (SECTION 3)

In this section, we provide the proofs for Section 3. We first review and show details of a few examples of exponential family.

C.1 EXAMPLES OF EXPONENTIAL FAMILY

Example C.1 (Gaussian)**.** For univariate Gaussians with fixed variance σ^2, the scaling $\alpha = \sigma$, the sufficient statistic $\xi = x/\sigma$, the log-partition function $\psi(\theta) = \theta^2/2$, and the base measure $\pi_0(\xi) = \frac{1}{\sqrt{2\pi}}\exp(-\xi^2/2)$ or equivalently $\pi_0(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-x^2/2\sigma^2)$ using $\xi = x/\sigma$ and $d\xi = dx/\sigma$. The scaled parameter $\theta_\alpha = \theta_\sigma = \theta/\sigma$. Then, the distribution from the natural parameter form is:

$$p_{\theta/\sigma}(x) = \exp(x\theta/\sigma^2 - \theta^2/(2\sigma^2)) \times \frac{1}{\sqrt{2\pi}\sigma}\exp(-x^2/2\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-(x-\mu)^2/2\sigma^2)\,. \quad (20)$$

The expectation parameter $\mu = \nabla\psi(\theta) = \theta$. The scaled expectation parameter $\mu_\alpha = \nabla\psi(\theta_\alpha) = \theta_\alpha = \theta/\sigma = \mu/\sigma$. Since $\xi = x/\sigma$ and the Bregman divergence $d_\phi(\xi,\mu) = \frac{1}{2}(\xi-\mu)^2$, the expectation parameter form

$$p_{\mu_\alpha}(\xi) = \exp(-(x/\sigma - \mu/\sigma)^2/2)\frac{1}{\sqrt{2\pi}\sigma} = \frac{1}{\sqrt{2\pi}\sigma}\exp(-(x-\mu)^2/2\sigma^2)\,. \quad \square \quad (21)$$

Example C.2 (Bernoulli over $\{0,1\}$)**.** For Bernoulli over $\{0,1\}$, the sufficient statistic $\xi = x \in \{0,1\}$, base measure is 1 on $\{0,1\}$, and the log-partition function $\psi(\theta) = \log(1+\exp(\theta))$ for natural parameter $\theta \in \mathbb{R}$. The expectation parameter $\mu = \nabla\psi(\theta) = \frac{exp(\theta)}{1+\exp(\theta)} = \frac{1}{1+\exp(-\theta)}$, the sigmoid function of θ. The Bregman divergence is the Bernoulli KL-divergence given by: $d_\phi(\xi,\mu) = x\log\frac{x}{\mu} + (1-x)\log\frac{1-x}{1-\mu}$. For scaled parameters $\theta_\alpha = \theta/\alpha$, the corresponding expectation parameter $\mu_\alpha = \frac{1}{1+\exp(-\theta_\alpha)}$. The mean parameter form distribution is given by

$$p_{\mu_\alpha}(\xi) = \exp(-d_\phi(\xi,\mu_\alpha) = \mu_\alpha^\xi(1-\mu_\alpha)^{1-\xi}\,. \quad (22)$$

$\square$

Example C.3 (Bernoulli over $\{-1,1\}$)**.** For Bernoulli over $\{-1,1\}$, the sufficient statistic $\xi = x \in \{-1,1\}$, base measure is 1 on $\{-1,1\}$, and the log-partition function $\psi(\theta) = \log(\exp(-\theta) + \exp(\theta))$ for natural parameter $\theta \in \mathbb{R}$. The expectation parameter $\mu = \nabla\psi(\theta) = \frac{\exp(\theta)-\exp(-\theta)}{\exp(-\theta)+\exp(\theta)} = \frac{\exp(2\theta)-1}{\exp(2\theta)+1}$, then its inverse function $\theta = \nabla\phi(\mu) = \frac{1}{2}\log\left(\frac{1+\mu}{1-\mu}\right)$, by integration we have $\phi(\mu) = \frac{1+\mu}{2}\log\frac{1+\mu}{2} + \frac{1-\mu}{2}\log\frac{1-\mu}{2}$. The Bregman divergence is given by: $d_\phi(\xi,\mu) = \frac{1+\xi}{2}\log\frac{1+\xi}{1+\mu} + \frac{1-\xi}{2}\log\frac{1-\xi}{1-\mu}$. For scaled parameters $\theta_\alpha = \theta/\alpha$, the corresponding expectation parameter $\mu_\alpha = \frac{\exp(2\theta_\alpha)-1}{\exp(2\theta_\alpha)+1}$. Noting that $\frac{1+\xi}{2}\log\frac{1+\xi}{2} = \frac{1-\xi}{2}\log\frac{1-\xi}{2} = 0$ for the base measure, the distribution is given by

$$p_{\mu_\alpha}(\xi) = \exp(-d_\phi(\xi,\mu_\alpha)) = \left(\frac{1+\mu_\alpha}{2}\right)^{\frac{1+\xi}{2}}\left(\frac{1-\mu_\alpha}{2}\right)^{\frac{1-\xi}{2}}\,, \quad (23)$$

which is $\xi = 1$ with probability $\frac{1+\mu_\alpha}{2}$ and $\xi = -1$ with probability $\frac{1-\mu_\alpha}{2}$. $\square$

C.2 PROOF OF THEOREM 1

Theorem 1. *For a given set $\bar{S} \sim D^{n+1}$ and $\mathbf{w}_{t-1}$ at iteration $(t-1)$, let $\Delta_{t|\mathbf{w}_{t-1}}(\bar{S}) = \max_{z,z'\in\bar{S}}\|\nabla\ell(\mathbf{w}_{t-1},z) - \nabla\ell(\mathbf{w}_{t-1},z')\|_2$. Further, for a c_2-smooth log-partition function ψ, let the scaling $\alpha_{t|\mathbf{w}_{t-1}}$ be data-dependent such that $\alpha^2_{t|\mathbf{w}_{t-1}} \geq 8c_2\Delta^2_{t|\mathbf{w}_{t-1}}(S_{n+1})$. Then, we have*

$$I_{A_t,B_t} \leq 5c_2\|\boldsymbol{\theta}_{B_t,\alpha_t} - \boldsymbol{\theta}_{B'_t,\alpha_t}\|_2^2 = \frac{5c_2}{2\alpha^2_{t|\mathbf{w}_{t-1}}}\left[\left\|\nabla\ell(\mathbf{w}_{t-1},S_{B_t}) - \nabla\ell\left(\mathbf{w}_{t-1},S'_{B_t}\right)\right\|_2^2\right]\,, \quad (7)$$

To avoid clutter, we drop the subscript t for the analysis and note that the analysis holds for any step t. When the density $dP_{B,\boldsymbol{\xi}} = \mathbf{p}_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B,\alpha})$, by mean-value theorem, for each $\boldsymbol{\xi}$, we have

$$p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B,\alpha}) - p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B',\alpha}) = \langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \nabla_{\tilde{\boldsymbol{\theta}}_{B,\alpha}} p_\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle \ , \tag{24}$$

for some $\tilde{\boldsymbol{\theta}}_{B,\alpha} = \gamma_{\boldsymbol{\xi}}\boldsymbol{\theta}_{B,\alpha} + (1-\gamma_{\boldsymbol{\xi}})\boldsymbol{\theta}'_{B,\alpha}$ where $\gamma_{\boldsymbol{\xi}} \in [0,1]$. Then,

$$\begin{aligned} I_{A,B} &= \int_{\boldsymbol{\xi}} \frac{\left(p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B,\alpha}) - p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B',\alpha})\right)^2}{p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi} = \int_{\boldsymbol{\xi}} \frac{\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \nabla_{\tilde{\boldsymbol{\theta}}_{B,\alpha}} p_\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle^2}{p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi} \\ &= \int_{\boldsymbol{\xi}} \frac{\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}, \boldsymbol{\xi} - \nabla_{\tilde{\boldsymbol{\theta}}_{B',\alpha}} \psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle^2 p_\psi^2(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})}{p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi} \ , \end{aligned} \tag{25}$$

since $p_\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha}) = \exp(\langle \boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}_{B,\alpha}\rangle - \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}))\pi_0(\boldsymbol{\xi})$.

C.2.1 Handling Distributional Dependency of $\tilde{\boldsymbol{\theta}}_B$

Note that we cannot proceed with the analysis with the density term depending on $\tilde{\boldsymbol{\theta}}_B$ since $\tilde{\boldsymbol{\theta}}_B$ depends on $\boldsymbol{\xi}$. In this step we focus on bounding the density term depending on $\tilde{\boldsymbol{\theta}}_B$ in terms of exponential family densities with parameters $\boldsymbol{\theta}_B$ and $\boldsymbol{\theta}'_B$.

Lemma 2. *For some $\gamma_{\boldsymbol{\xi}} \in [0,1]$, $\tilde{\boldsymbol{\theta}}_{B,\alpha} = \gamma_{\boldsymbol{\xi}}\boldsymbol{\theta}_{B,\alpha} + (1-\gamma_{\boldsymbol{\xi}})\boldsymbol{\theta}'_{B,\alpha}$, we have*

$$\frac{\exp\left[\langle \boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}_{B,\alpha}\rangle - \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha})\right]}{\max\left(\exp\left[\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}_{B,\alpha})\right], \exp\left[\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B',\alpha}\rangle - \psi(\boldsymbol{\theta}_{B',\alpha})\right]\right)} \leq \exp\left[c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2\right] \ .$$

Proof. Denoting $\gamma_{\boldsymbol{\xi}}$ as γ for convenience (the dependence on $\boldsymbol{\xi}$ does not play a role in the analysis), we have

$$\begin{aligned} \langle \boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}_{B,\alpha}\rangle - \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) &= \langle \boldsymbol{\xi}, \gamma\boldsymbol{\theta}_{B,\alpha} + (1-\gamma)\boldsymbol{\theta}'_{B,\alpha}\rangle - \psi\left(\gamma\boldsymbol{\theta}_{B,\alpha} + (1-\gamma)\boldsymbol{\theta}'_{B,\alpha}\right) \\ &= \gamma\Big[\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}_{B,\alpha})\Big] + (1-\gamma)\Big[\langle \boldsymbol{\xi}, \boldsymbol{\theta}'_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}'_{B,\alpha})\Big] \\ &\qquad + \Big[\gamma\psi(\boldsymbol{\theta}_{B,\alpha}) + (1-\gamma)\psi(\boldsymbol{\theta}'_{B,\alpha}) - \psi(\gamma\boldsymbol{\theta}_{B,\alpha} + (1-\gamma)\boldsymbol{\theta}'_{B,\alpha})\Big] \\ &\overset{(a)}{\leq} \max\Big(\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}_{B,\alpha}), \langle \boldsymbol{\xi}, \boldsymbol{\theta}'_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}'_{B,\alpha})\Big) \\ &\qquad + \Big[\gamma d_\psi(\boldsymbol{\theta}_{B,\alpha}, \tilde{\boldsymbol{\theta}}_{B,\alpha}) + (1-\gamma) d_\psi(\boldsymbol{\theta}'_{B,\alpha}, \tilde{\boldsymbol{\theta}}_{B,\alpha})\Big] \ , \end{aligned}$$

where the second term in (a) follows since difference between two sides of Jensen's inequality is given by the Bregman information, i.e., expected Bregman divergence to the expectation (See section 3.1.1 in Banerjee et al. (2005)):

$$\gamma\psi(\boldsymbol{\theta}_{B,\alpha}) + (1-\gamma)\psi(\boldsymbol{\theta}'_{B,\alpha}) - \psi(\gamma\boldsymbol{\theta}_{B,\alpha} + (1-\gamma)\boldsymbol{\theta}'_{B,\alpha}) = \gamma d_\psi(\boldsymbol{\theta}_{B,\alpha}, \tilde{\boldsymbol{\theta}}_{B,\alpha}) + (1-\gamma) d_\psi(\boldsymbol{\theta}'_{B,\alpha}, \tilde{\boldsymbol{\theta}}_{B,\alpha}) \ .$$

Now, note that

$$\begin{aligned} \gamma d_\psi(\boldsymbol{\theta}_{B,\alpha}, \tilde{\boldsymbol{\theta}}_{B,\alpha}) + (1-\gamma) d_\psi(\boldsymbol{\theta}'_{B,\alpha}, \tilde{\boldsymbol{\theta}}_{B,\alpha}) &\leq \gamma d_\psi(\boldsymbol{\theta}_{B,\alpha}, \boldsymbol{\theta}'_{B,\alpha}) + (1-\gamma) d_\psi(\boldsymbol{\theta}'_{B,\alpha}, \boldsymbol{\theta}_{B,\alpha}) \\ &\leq \gamma c_2\|\boldsymbol{\theta}'_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}\|_2^2 + (1-\gamma) c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}\|_2^2 \\ &\leq c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ . \end{aligned}$$

As a result,

$$\frac{\exp\left[\langle \boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}_{B,\alpha}\rangle - \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha})\right]}{\max\left(\exp\left[\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}_{B,\alpha})\right], \exp\left[\langle \boldsymbol{\xi}, \boldsymbol{\theta}'_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}'_{B,\alpha})\right]\right)}$$

$$\leq \exp\left[c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B',\alpha}\|_2^2\right]\frac{\exp\left[\max\left(\langle\boldsymbol{\xi},\boldsymbol{\theta}_{B,\alpha}\rangle-\psi(\boldsymbol{\theta}_{B,\alpha}),\langle\boldsymbol{\xi},\boldsymbol{\theta}'_{B,\alpha}\rangle-\psi(\boldsymbol{\theta}'_{B,\alpha})\right)\right]}{\max\left(\exp\left[\langle\boldsymbol{\xi},\boldsymbol{\theta}_{B,\alpha}\rangle-\psi(\boldsymbol{\theta}_{B,\alpha})\right],\exp\left[\langle\boldsymbol{\xi},\boldsymbol{\theta}'_{B,\alpha}\rangle-\psi(\boldsymbol{\theta}'_{B,\alpha})\right]\right)}$$

$$= \exp\left[c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B',\alpha}\|_2^2\right] \ .$$

That completes the proof. □

C.2.2 Bounding the Density Ratio

Focusing on density ratio, we have

Lemma 3. *For any $\boldsymbol{\xi}$, we have*

$$\frac{\exp\left[\langle\boldsymbol{\xi},2\boldsymbol{\theta}_{B,\alpha}\rangle-2\psi(\boldsymbol{\theta}_{B,\alpha})\right]}{\exp\left[\langle\boldsymbol{\xi},\boldsymbol{\theta}_{A,\alpha}\rangle-\psi(\boldsymbol{\theta}_{A,\alpha})\right]}\leq\exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2^2\right]\exp\left[\langle\boldsymbol{\xi},(2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\rangle-\psi(2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})\right].$$

Proof. Note that

$$\begin{aligned}\frac{\exp\left[\langle\boldsymbol{\xi},2\boldsymbol{\theta}_{B,\alpha}\rangle-2\psi(\boldsymbol{\theta}_{B,\alpha})\right]}{\exp\left[\langle\boldsymbol{\xi},\boldsymbol{\theta}_{A,\alpha}\rangle-\psi(\boldsymbol{\theta}_{A,\alpha})\right]}&=\exp\left[\langle\boldsymbol{\xi},(2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})\rangle-\left(2\psi(\boldsymbol{\theta}_{B,\alpha})-\psi(\boldsymbol{\theta}_{A,\alpha})\right)\right]\\&=\beta_{B,A,\alpha}\exp\left[\langle\boldsymbol{\xi},(2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\rangle-\psi((2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})\right] \ ,\end{aligned}$$

where

$$\beta_{B,A,\alpha}=\exp\left[\psi((2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})-\left(2\psi(\boldsymbol{\theta}_{B,\alpha})-\psi(\boldsymbol{\theta}_{A,\alpha})\right)\right] \ .$$

Note that

$$\begin{aligned}\log\beta_{B,A,\alpha}&=\psi((2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})-\left(2\psi(\boldsymbol{\theta}_{B,\alpha})-\psi(\boldsymbol{\theta}_{A,\alpha})\right)\\&=-\Big(\psi(\boldsymbol{\theta}_{B,\alpha})-\psi((2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})\Big)-\Big(\psi(\boldsymbol{\theta}_{B,\alpha})-\psi(\boldsymbol{\theta}_{A,\alpha})\Big)\\&\overset{(a)}{\leq}\Big\langle\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha},\nabla\psi(2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})\Big\rangle+\Big\langle\boldsymbol{\theta}_{A,\alpha}-\boldsymbol{\theta}_{B,\alpha},\nabla\psi(\boldsymbol{\theta}_{A,\alpha}\Big\rangle\\&=\Big\langle\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha},\nabla\psi((2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})-\nabla\psi(\boldsymbol{\theta}_{A,\alpha})\Big\rangle\\&\overset{(b)}{\leq}\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2\|\nabla\psi((2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})-\nabla\psi(\boldsymbol{\theta}_{A,\alpha})\|_2\\&\overset{(c)}{\leq}c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2\|(2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})-\boldsymbol{\theta}_{A,\alpha}\|_2\\&=2c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2^2 \ ,\end{aligned}$$

where (a) follows from the convexity of ψ, (b) follows from Cauchy-Schwartz, and (c) follows by smoothness of ψ.

As a result, we have

$$\frac{\exp\left[\langle\boldsymbol{\xi},2\boldsymbol{\theta}_{B,\alpha}\rangle-2\psi(\boldsymbol{\theta}_{B,\alpha})\right]}{\exp\left[\langle\boldsymbol{\xi},\boldsymbol{\theta}_{A,\alpha}\rangle-\psi(\boldsymbol{\theta}_{A,\alpha})\right]}\leq\exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2^2\right]\exp\left[\langle\boldsymbol{\xi},(2\boldsymbol{\theta}_B-\boldsymbol{\theta}_A)/\alpha\rangle-\psi((2\boldsymbol{\theta}_B-\boldsymbol{\theta}_A)/\alpha)\right] \ .$$

That completes the proof. □

The analysis for the term involving $\boldsymbol{\theta}'_B$ is exactly the same.

Combining Lemma 2 and 3, we have the following result:

Lemma 6. *We have*

$$\begin{aligned}\frac{p_\psi^2(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})}{p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})}\leq&\exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B',\alpha}\|_2^2\right]\\&\times\max\Big(\exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2^2\right]p_\psi(\boldsymbol{\xi};2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}) \ ,\\&\qquad\exp\left[2c_2\|\boldsymbol{\theta}_{B',\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2^2\right]p_\psi(\boldsymbol{\xi};2\boldsymbol{\theta}_{B',\alpha}-\boldsymbol{\theta}_{A,\alpha})\Big) \ .\end{aligned}\tag{26}$$

Proof. From Lemma 2, we have

$$p_\psi^2(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha}) = \exp\left[2\left\langle \boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}_{B,\alpha}\right\rangle - 2\psi\left(\tilde{\boldsymbol{\theta}}_{B,\alpha}\right)\right]$$
$$\leq \exp\left[c_2\left\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B',\alpha}\right\|_2^2\right]\cdot \max\left(\exp\left[\langle\boldsymbol{\xi},\boldsymbol{\theta}_{B,\alpha}\rangle - \psi\left(\boldsymbol{\theta}_{B,\alpha}\right)\right], \exp\left[\langle\boldsymbol{\xi},\boldsymbol{\theta}_{B',\alpha}\rangle - \psi\left(\boldsymbol{\theta}_{B',\alpha}\right)\right]\right)$$

Thus,

$$\frac{p_\psi^2\left(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha}\right)}{p_\psi\left(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha}\right)} \leq \exp\left[2c_2\left\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B',\alpha}\right\|_2^2\right]\cdot\frac{\max\left(\exp\left[2\left\langle\boldsymbol{\xi},\boldsymbol{\theta}_{B,\alpha}\right\rangle - 2\psi\left(\boldsymbol{\theta}_{B,\alpha}\right)\right], \exp\left[2\left\langle\boldsymbol{\xi},\boldsymbol{\theta}_{B',\alpha}\right\rangle - 2\psi\left(\boldsymbol{\theta}_{B',\alpha}\right)\right]\right)}{\exp\left[\langle\boldsymbol{\xi},\boldsymbol{\theta}_{A,\alpha}\rangle - \psi\left(\boldsymbol{\theta}_{A,\alpha}\right)\right]}$$

Based on Lemma 3, we have

$$\frac{\exp\left[\langle\boldsymbol{\xi},2\boldsymbol{\theta}_{B,\alpha}\rangle - 2\psi\left(\boldsymbol{\theta}_{B,\alpha}\right)\right]}{\exp\left[\langle\boldsymbol{\xi},\boldsymbol{\theta}_{A,\alpha}\rangle - \psi\left(\boldsymbol{\theta}_{A,\alpha}\right)\right]} \leq \exp\left[2c_2\left\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\right\|_2^2\right]\exp\left[\langle\boldsymbol{\xi},\left(2\boldsymbol{\theta}_B-\boldsymbol{\theta}_A\right)/\alpha\rangle - \psi\left(\left(2\boldsymbol{\theta}_B-\boldsymbol{\theta}_A\right)/\alpha\right)\right]$$

Combining above inequalities we have

$$\begin{aligned}\frac{p_\psi^2(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})}{p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})} \leq{}& \exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B',\alpha}\|_2^2\right]\\ &\times\max\Big(\exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2^2\right]p_\psi(\boldsymbol{\xi};2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}),\\ &\qquad\exp\left[2c_2\|\boldsymbol{\theta}_{B',\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2^2\right]p_\psi(\boldsymbol{\xi};2\boldsymbol{\theta}_{B',\alpha}-\boldsymbol{\theta}_{A,\alpha})\Big)\ .\end{aligned} \tag{27}$$

That completes the proof. □

C.2.3 Bounding the Integral

Ignoring multiplicative terms which do not depend on $\boldsymbol{\xi}$ for the moment, the analysis needs to bound an integral term of the form

$$\int_{\boldsymbol{\xi}}\langle\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}'_{B,\alpha},\boldsymbol{\xi}-\nabla\psi(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle^2\,p_\psi(\boldsymbol{\xi};2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})d\boldsymbol{\xi}\ , \tag{28}$$

and a similar term with $p_\psi^2(\boldsymbol{\xi};2\boldsymbol{\theta}_{B',\alpha}-\boldsymbol{\theta}_{A,\alpha})$. The proof of Theorem 1 can be done by suitably bounding the integral.

Proof of Theorem 1. From Lemma 6, we have

$$I_{A,B} \leq \int_{\boldsymbol{\xi}}\frac{\left\langle\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}'_{B,\alpha},\boldsymbol{\xi}-\nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha})\right\rangle^2 p_\psi^2(\boldsymbol{\xi};\tilde{\boldsymbol{\theta}}_{B,\alpha})}{p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{A,\alpha})}d\boldsymbol{\xi}$$
$$\begin{aligned}&\leq \exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B',\alpha}\|_2^2\right]\\ &\qquad\times\max\Bigg(\exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2^2\right]\int_{\boldsymbol{\xi}}\left\langle\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B,\alpha},\boldsymbol{\xi}-\nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha})\right\rangle^2 p_\psi(\boldsymbol{\xi};2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})d\boldsymbol{\xi}\ ,\\ &\qquad\qquad\exp\left[2c_2\|\boldsymbol{\theta}_{B',\alpha}-\boldsymbol{\theta}_{A,\alpha}\|_2^2\right]\int_{\boldsymbol{\xi}}\left\langle\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B',\alpha},\boldsymbol{\xi}-\nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha})\right\rangle^2 p_\psi(\boldsymbol{\xi};2\boldsymbol{\theta}_{B',\alpha}-\boldsymbol{\theta}_{A,\alpha})d\boldsymbol{\xi}\Bigg)\ .\end{aligned}$$

Focusing on the integral in the first term (the analysis for the second term is essentially the same), we have

$$\int_{\boldsymbol{\xi}}\left\langle\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{B',\alpha},\boldsymbol{\xi}-\nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha})\right\rangle^2 p_\psi(\boldsymbol{\xi};2\boldsymbol{\theta}_{B,\alpha}-\boldsymbol{\theta}_{A,\alpha})d\boldsymbol{\xi}$$

$$
\begin{aligned}
&= \int_{\boldsymbol{\xi}} \left\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, (\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]) - \left(\nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) - \mathbb{E}[\boldsymbol{\xi}]\right) \right\rangle^2 p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) d\boldsymbol{\xi} \\
&\leq 2 \underbrace{\int_{\boldsymbol{\xi}} \langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}] \rangle^2 p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) d\boldsymbol{\xi}}_{T_1} \\
&\quad + 2 \underbrace{\int_{\boldsymbol{\xi}} \left\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}, \nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) - \mathbb{E}[\boldsymbol{\xi}] \right\rangle^2 p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) d\boldsymbol{\xi}}_{T_2} \ .
\end{aligned}
$$

For T_1, note that

$$
\begin{aligned}
T_1 &= \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}] \rangle^2 \right] \\
&= (\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha})^T \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^T \right] (\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}) \\
&= (\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha})^T \nabla^2\psi(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})(\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}) \\
&\leq c_2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}\|_2^2 \ ,
\end{aligned}
$$

since, by smoothness, the spectral norm of $\nabla^2\psi$ is bounded by c_2.

For T_2, first note that

$$
\mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})}[\boldsymbol{\xi}] = \nabla\psi(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})) \ .
$$

Hence, with $\tilde{\boldsymbol{\theta}}_{B,\alpha} = \gamma_{\boldsymbol{\xi}} \boldsymbol{\theta}_{B,\alpha} + (1 - \gamma_{\boldsymbol{\xi}}) \boldsymbol{\theta}_{B',\alpha}$ for some $\gamma_{\boldsymbol{\xi}} \in [0, 1]$, we have

$$
\begin{aligned}
T_2 &= \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}, \nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) - \mathbb{E}[\boldsymbol{\xi}] \right\rangle^2 \right] \\
&= \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) - \nabla\psi(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\rangle^2 \right] \\
&\leq \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}\|_2^2 \ \ \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\| \nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) - \nabla\psi((2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\|_2^2 \right] \\
&= c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \ \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\| \tilde{\boldsymbol{\theta}}_{B,\alpha} - (2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\|_2^2 \right] \\
&= c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \ \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\| \gamma_{\boldsymbol{\xi}} \boldsymbol{\theta}_{B,\alpha} + (1 - \gamma_{\boldsymbol{\xi}}) \boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{B,\alpha} - (\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\|_2^2 \right] \\
&= c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \ \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\| (1 - \gamma_{\boldsymbol{\xi}})(\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{B,\alpha}) - (\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\|_2^2 \right] \\
&\leq 2c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \ \left(\mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[(1 - \gamma_{\boldsymbol{\xi}})^2 \left\| \boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{B,\alpha} \right\|_2^2 \right] + \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\| \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha} \right\|_2^2 \right] \right) \\
&= 2c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \ \left(\left\| \boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{B,\alpha} \right\|_2^2 \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[(1 - \gamma_{\boldsymbol{\xi}})^2 \right] + \left\| \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha} \right\|_2^2 \right) \\
&\leq 2c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \ \left(\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 + \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2 \right) \ .
\end{aligned}
$$

Putting everything back together

$$
\begin{aligned}
&\int_{\boldsymbol{\xi}} \left\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \boldsymbol{\xi} - \nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) \right\rangle^2 p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) d\boldsymbol{\xi} \\
&\qquad\qquad = c_2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \ \left(1 + 2c_2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 + 2c_2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2 \right) \ .
\end{aligned}
$$

Similarly

$$
\int_{\boldsymbol{\xi}} \left\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \boldsymbol{\xi} - \nabla\psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) \right\rangle^2 p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha}) d\boldsymbol{\xi}
$$

$$= c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \left(1 + 2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 + 2c_2\|\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right) \ .$$

Then, plugging into bound on $I_{A,B}$, we have

$$\begin{aligned} I_{A,B} \leq\ & c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \times \exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2\right] \\ & \times \max\Big(\exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right] \times \left(1 + 2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 + 2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right) \ , \\ & \quad \exp\left[2c_2\|\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right] \times \left(1 + 2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 + 2c_2\|\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right)\Big) \ . \end{aligned}$$

Since $\alpha^2 \geq 8c_2\Delta^2(S_{n+1})$ where $\Delta(S_{n+1}) = \max_{z,z' \in S_{n+1}} \|\nabla\ell(\mathbf{w}, z) - \nabla\ell(\mathbf{w}, z')\|_2$, recalling that $\boldsymbol{\theta}_{B,\alpha} = \nabla\ell(\mathbf{w}, B)/\alpha$, we have

$$2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \leq \frac{1}{4} \ , \quad 2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2 \leq \frac{1}{4} \ , \quad 2c_2\|\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2 \leq \frac{1}{4} \ .$$

As a result, we have

$$\begin{aligned} I_{A,B} &\leq c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \times \exp\left(\frac{1}{4}\right) \times \exp\left(\frac{1}{4}\right) \times \left(1 + \frac{1}{4} + \frac{1}{4}\right) \\ &\leq \frac{5c_2}{2\alpha^2}\|\nabla\ell(\mathbf{w}, S_B) - \nabla\ell(\mathbf{w}, S'_B)\|_2^2 \ . \end{aligned}$$

That completes the proof. □

C.3 Expected Stability of EFLD

Theorem 2. *Consider an exponential family Langevin dynamics (EFLD) algorithm of the form* (5)-(6) *with a c_2-smooth log-partition function ψ. Then, for mini-batch size $b \leq n/2$, with $c = c_0\sqrt{5c_2}$ (with c_0 as in Lemma 1) and $\alpha_{t|}^2 \geq 8c_2\Delta_{t|}^2(S_{n+1})$ (as in Theorem 1, with the conditioning on $\mathbf{w}_{t-1}$ hidden to avoid clutter), we have*

$$|\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]| \leq c\frac{1}{n}\mathop{\mathbb{E}}_{S_{n+1}}\sqrt{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{W_{0:(t-1)}}\frac{1}{\alpha_{t|}^2}\left[\|\nabla\ell\left(\mathbf{w}_{t-1}, z_n\right) - \nabla\ell\left(\mathbf{w}_{t-1}, z'_n\right)\|_2^2\right]} \ . \tag{8}$$

Proof. Based on Lemma 1, we have

$$|\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]| \leq c_1\frac{b}{n}\mathbb{E}_S\mathbb{E}_{z'_n}\sqrt{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{W_{0:(t-1)}}\mathop{\mathbb{E}}_{B_t\in G_1}\mathop{\mathbb{E}}_{A_t\in G_0}\left[I_{A_t,B_t}\right]} \ . \tag{29}$$

with $I_{A_t,B_t} = \int_{\boldsymbol{\xi}_t}\frac{\left(dP_{B_t,\boldsymbol{\xi}_t} - dP'_{B_t,\boldsymbol{\xi}_t}\right)^2}{dP_{A_t,\boldsymbol{\xi}_t}}d\boldsymbol{\xi}_t$.

From Theorem 1, we have

$$\begin{aligned} I_{A_t,B_t} &\leq \frac{5c_2}{2\alpha_{t|\mathbf{w}_{t-1}}^2}\left[\left\|\nabla\ell\left(\mathbf{w}_{t-1}, S_{B_t}\right) - \nabla\ell\left(\mathbf{w}_{t-1}, S'_{B_t}\right)\right\|_2^2\right] \\ &= \frac{5c_2}{2\alpha_{t|\mathbf{w}_{t-1}}^2}\left[\left\|\frac{1}{b}\sum_{z\in S'_{B_t}}\nabla\ell\left(\mathbf{w}_{t-1}, z\right) - \frac{1}{b}\sum_{z\in S_{B_t}}\nabla\ell\left(\mathbf{w}_{t-1}, z\right)\right\|_2^2\right] \\ &= \frac{5c_2}{2b\alpha_{t|\mathbf{w}_{t-1}}^2}\left[\|\nabla\ell\left(\mathbf{w}_{t-1}, z_n\right) - \nabla\ell\left(\mathbf{w}_{t-1}, z'_n\right)\|_2^2\right] \ , \end{aligned}$$

where the last equation holds because S_{B_t} and S'_{B_t} only differ at z_n and z'_n.

Combining the above two inequalities, we have

$$|\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]| \leq c_1\frac{b}{n}\mathbb{E}_S\mathbb{E}_{z'_n}\sqrt{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{W_{0:(t-1)}}\left[\|\nabla\ell\left(\mathbf{w}_{t-1}, z_n\right) - \nabla\ell\left(\mathbf{w}_{t-1}, z'_n\right)\|_2^2\right]} \ .$$

That completes the proof. □

D NOISY SIGN-SGD OPTIMIZATION BOUND

The "density" for a mini-batch B at scale α is:

$$p_\psi(\boldsymbol{\xi};\boldsymbol{\theta}_{B,\alpha}) = \exp(\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}_{B,\alpha}))\boldsymbol{\pi}_0(\boldsymbol{\xi}) \ , \qquad \boldsymbol{\theta}_{B,\alpha} \triangleq \frac{\boldsymbol{\theta}_B}{\alpha} = \frac{\nabla \ell(\mathbf{w}_t, S_B)}{\alpha} \ . \tag{30}$$

Note that the corresponding expectation parameter

$$\boldsymbol{\mu}_{B,\alpha} = \nabla_{\boldsymbol{\theta}_{B,\alpha}} \psi(\theta_{B,\alpha}) \ . \tag{31}$$

The full-batch Noisy Sign-SGD update the parameters as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{\xi}_t \ , \qquad \boldsymbol{\xi}_{t,i} \sim B\left(\frac{1}{1+\exp\left(-2\nabla \mathcal{L}_S(\mathbf{w}_t)/\alpha_t\right)}\right) \ , \ \forall i \in [d] \ , \tag{32}$$

where $B(x)$ is a pdf with density x at 1 and density $1-x$ at -1. For mini-batch B_t and scaling α_t, mini-batch Noisy Sign-SGD updates the parameters as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{\xi}_t \ , \qquad \boldsymbol{\xi}_{t,i} \sim B\left(\frac{1}{1+\exp\left(-2\boldsymbol{\theta}_{B_t,\alpha_t}\right)}\right) \ , \ \forall i \in [d] \ . \tag{33}$$

We make the following smoothness assumptions of the empirical loss function $\mathcal{L}_S(\mathbf{w})$:

Assumption 1. *The loss function L_S satisfies: for all $\mathbf{w}$ and $\mathbf{w}'$, for some non-negative constant $\vec{K} := [K_1, \ldots, K_d]$, we have $L_S(\mathbf{w}) \le L_S(\mathbf{w}') + \nabla L_S(\mathbf{w}')^T(\mathbf{w}-\mathbf{w}') + \frac{1}{2}\sum_i K_i(\mathbf{w}_i - \mathbf{w}_i')^2$.*

The assumption on on the empirical loss $\mathcal{L}_S(\mathbf{w})$ is common in optimization analysis, besides that, we also assume some natural statistical properties of the batch gradient of the loss $\nabla \ell(\mathbf{w}_t, S_{B_t})$, where the randomness comes from batches conditioned on $\mathbf{w}_t$, satisfies the following assumptions:

Assumption 2. *Given $\mathbf{w}_{t-1}$, the mini-batch gradient $\nabla \ell(\mathbf{w}_{t-1}, S_{B_t})$ is (a) unbiased, i.e., $\mathbb{E}_{B_t|\mathbf{w}_{t-1}} \nabla \ell(\mathbf{w}_{t-1}, S_{B_t}) = \nabla L_S(\mathbf{w}_{t-1})$; (b) symmetric, i.e., the density $p(x)$ of $x \equiv \nabla \ell(\mathbf{w}_{t-1}, S_{B_t})$ is symmetric around its expectation $L_S(\mathbf{w}_{t-1})$: $p(x) = p(2\nabla L_S(\mathbf{w}_{t-1}) - x)$ and (c) sub-Gaussian, i.e., for any $\lambda > 0$, any $\mathbf{v}$ s.t. $\|\mathbf{v}\|_2 = 1$, $\mathbb{E}_{B_t|\mathbf{w}_{t-1}} \exp \lambda \langle \mathbf{v}, \nabla \ell(\mathbf{w}_{t-1}, S_{B_t}) - \nabla L_S(\mathbf{w}_{t-1})\rangle \le \exp(\lambda^2 \kappa_t^2/2)$ for some constant $\kappa_t > 0$.*

With Assumption 1 and Assumption 2, we have the following guarantee for convergence of noisy signSGD under full batch and mini batch settings. The following is a restate theorem from the main paper for the mini-batch noisy sign-SGD

Theorem 3. *Under Assumption 1 and 2, for mini-batch noisy Sign-SGD with step size $\eta_t = 1/\sqrt{T}$, α_t satisfying $c \ge \alpha_t \ge \max[\sqrt{2}\kappa_t, 4\|\nabla L_S(\mathbf{w}_{t-1})\|_\infty]$, we have for any S and any initialization $\mathbf{w}_0$*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla L_S(\mathbf{w}_t)\|_2^2\right] \le \frac{4c}{\sqrt{T}}\left(L_S(\mathbf{w}_0) - L_S(\mathbf{w}^*) + \frac{1}{2}\|\vec{K}\|_1\right), \tag{10}$$

where the expectation is taken over the randomness of algorithm.

Here we state and prove the result for both full-batch and mini-batch settings:

Theorem 7. *If the loss satisfies Assumption 1, for full-batch noisy Sign-SGD with step size $\rho_t = 1/\sqrt{T}$ and α_t satisfying $c \ge \alpha_t \ge \|\nabla L_S(\mathbf{w}_t)\|_\infty$, we have*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla L_S(\mathbf{w}_t)\|_2^2\right] \le \frac{5c}{3\sqrt{T}}\left(L_S(\mathbf{w}_0) - L_S(\mathbf{w}^*) + \frac{1}{2}\|\vec{K}\|_1\right). \tag{34}$$

Further, if Assumption 2 holds, for mini-batch noisy Sign-SGD with step size $\eta_t = 1/\sqrt{T}$, and α_t satisfying $c \ge \alpha_t \ge \max[\sqrt{2}\kappa_t, 4\|\nabla L_S(\mathbf{w}_t)\|_\infty]$, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla L_S(\mathbf{w}_t)\|_2^2\right] \le \frac{4c}{\sqrt{T}}\left(L_S(\mathbf{w}_0) - L_S(\mathbf{w}^*) + \frac{1}{2}\|\vec{K}\|_1\right), \tag{35}$$

where equation (34)(35) *holds for any S, any initialization $\mathbf{w}_0$, and the expectation is taken over the randomness of algorithm.*

Proof for full-batch: Conditioned at t-th iteration, with Assumption 1, we have

$$\begin{aligned}\mathcal{L}_S(\mathbf{w}_{t+1}) - \mathcal{L}_S(\mathbf{w}_t) &\leq \nabla\mathcal{L}_S(\mathbf{w}_t)^T(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{1}{2}\sum_{i=1}^{d} L_i(\mathbf{w}_{t+1,i} - \mathbf{w}_{t,i})^2 \\ &= -\eta_t \nabla\mathcal{L}_S(\mathbf{w}_t)^T \boldsymbol{\xi}_t + \eta_t^2 \sum_{i=1}^{d} \frac{L_i}{2}\end{aligned}$$

Then taking conditional expectation on both side we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1}) - \mathcal{L}_S(\mathbf{w}_t)|\mathbf{w}_t] &\leq -\eta_t \nabla\mathcal{L}_S(\mathbf{w}_t)^T \mathbb{E}_{\boldsymbol{\xi}_t|\mathbf{w}_t}\boldsymbol{\xi}_t + \frac{\eta_t^2}{2}\|\vec{L}\|_1 \\ &= -\eta_t \sum_i \nabla\mathcal{L}_S(\mathbf{w}_t)_i \left(\frac{2}{1+\exp\left(-2\nabla\mathcal{L}_S(\mathbf{w}_t)_i/\alpha_t\right)} - 1\right) + \frac{\eta_t^2}{2}\|\vec{L}\|_1 \\ &= -\eta_t \sum_i \nabla\mathcal{L}_S(\mathbf{w}_t)_i \left(\frac{\exp\left(2\nabla\mathcal{L}_S(\mathbf{w}_t)_i/\alpha_t\right) - 1}{\exp\left(2\nabla\mathcal{L}_S(\mathbf{w}_t)_i/\alpha_t\right) + 1}\right) + \frac{\eta_t^2}{2}\|\vec{L}\|_1 \\ &= -\eta_t \sum_i \nabla\mathcal{L}_S(\mathbf{w}_t)_i \tanh(\nabla\mathcal{L}_S(\mathbf{w}_t)_i/\alpha_t) + \frac{\eta_t^2}{2}\|\vec{L}\|_1 \\ &= -\eta_t \sum_i |\nabla\mathcal{L}_S(\mathbf{w}_t)_i||\tanh(\nabla\mathcal{L}_S(\mathbf{w}_t)_i/\alpha_t)| + \frac{\eta_t^2}{2}\|\vec{L}\|_1\end{aligned}$$

By taking $c \geq \alpha_t \geq \|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_\infty$, and $\eta_t = 1/\sqrt{T}$, we have $\nabla\mathcal{L}_S(\mathbf{w}_t)_i/\alpha_t \leq 1$ so we can apply Lemma 7 to have

$$\begin{aligned}\mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1}) - \mathcal{L}_S(\mathbf{w}_t)|\mathbf{w}_t] &\leq -\frac{e^2-1}{(e^2+1)\alpha_t\sqrt{T}}\|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2T}\|\vec{L}\|_1 \\ &\leq -\frac{3}{5c\sqrt{T}}\|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2T}\|\vec{L}\|_1\end{aligned}$$

by telescope sum we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_2^2\right] \leq \frac{5c}{3\sqrt{T}}\left(\mathcal{L}_S(\mathbf{w}_0) - \mathcal{L}_S(\mathbf{w}^*) + \frac{1}{2}\|\vec{L}\|_1\right), \tag{36}$$

which completes the proof of full-batch updates.

Lemma 7. *For any $-1 \leq x \leq 1$, the following holds:*

$$|\tanh x| \geq \frac{e^2-1}{e^2+1}|x| \tag{37}$$

Proof. Without loss of generality, we focus on $0 < x \leq 1$. we prove $\tanh x/x$ is decreasing function on $\mathbb{R}^+$, which is equivalent to

$$(\frac{\tanh x}{x})' = \frac{x(1-\tanh^2 x) - \tanh x}{x^2} \leq 0,$$

and is equivalent to

$$x \leq \frac{\tanh x}{1-\tanh^2 x},$$

where the right hand side is $\sinh x$ and use the fact that $x \leq \sinh x$ for $x > 0$ implies $\tanh x/x$ is decreasing function, so

$$\tanh x/x \geq \tanh 1 = \frac{e^2-1}{e^2+1},$$

which completes the proof. □

Proof for mini-batch: From smoothness condition Assumption 1, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1}) - \mathcal{L}_S(\mathbf{w}_t)|\mathbf{w}_t] &\leq \mathbb{E}_{\mathbf{w}_{t+1}|\mathbf{w}_t} \nabla \mathcal{L}_S(\mathbf{w}_t)^T(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{1}{2}\mathbb{E}_{\mathbf{w}_{t+1}|\mathbf{w}_t} \sum_{i=1}^d L_i(\mathbf{w}_{t+1,i} - \mathbf{w}_{t,i})^2 \\
&= \mathbb{E}_{\boldsymbol{\xi}_t|\mathbf{w}_t} \nabla \mathcal{L}_S(\mathbf{w}_t)^T(-\eta_t \boldsymbol{\xi}_t) + \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}_t|\mathbf{w}_t} \sum_{i=1}^d L_i(-\eta_t \boldsymbol{\xi}_t)^2 \\
&= -\eta_t \nabla \mathcal{L}_S(\mathbf{w}_t)^T \mathbb{E}_{\boldsymbol{\xi}_t|\mathbf{w}_t} \boldsymbol{\xi}_t + \frac{\eta_t^2}{2}\|\vec{L}\|_1 \\
&= -\eta_t \sum_i \nabla \mathcal{L}_S(\mathbf{w}_t)_i \mathbb{E}_{B_t|\mathbf{w}_t}\left[\frac{2}{1+\exp\left(-2\nabla_i \ell(\mathbf{w}_t, S_{B_t}/\alpha_t)\right)} - 1\right] + \frac{\eta_t^2}{2}\|\vec{L}\|_1 \\
&= -\eta_t \sum_i \nabla \mathcal{L}_S(\mathbf{w}_t)_i \mathbb{E}_{B_t|\mathbf{w}_t}\left[\frac{\exp\left(2\nabla_i \ell(\mathbf{w}_t, S_{B_t})/\alpha_t\right) - 1}{\exp\left(2\nabla_i \ell(\mathbf{w}_t, S_{B_t})/\alpha_t\right) + 1}\right] + \frac{\eta_t^2}{2}\|\vec{L}\|_1 \\
&= -\eta_t \sum_i \mathbb{E}_{B_t|\mathbf{w}_t} \theta_{B_t,i} \mathbb{E}\left[\frac{\exp\left(2\theta_{B_t,i}/\alpha_t\right) - 1}{\exp\left(2\theta_{B_t,i}/\alpha_t\right) + 1}\right] + \frac{\eta_t^2}{2}\|\vec{L}\|_1.
\end{aligned}$$

Focus on each individual term in the sum, we have

$$\begin{aligned}
&\mathbb{E}_{B_t|\mathbf{w}_t} \theta_{B,i} \mathbb{E}_{B_t|\mathbf{w}_t}\left[\frac{\exp\left(2\theta_{B_t,i}/\alpha_t\right) - 1}{\exp\left(2\theta_{B_t,i}/\alpha_t\right) + 1}\right] \\
=&\mathbb{E}_{B_t|\mathbf{w}_t} \theta_{B_t,i} \mathbb{E}_{B_t|\mathbf{w}_t}\left[\frac{\exp\left(2\theta_{B_t,i}/\alpha_t - 2\mathbb{E}_{B_t|\mathbf{w}_t}\theta_{B_t,i}/\alpha_t\right) - \exp -2\mathbb{E}_{B_t|\mathbf{w}_t}\theta_{B_t,i}/\alpha_t}{\exp\left(2\theta_{B_t,i}/\alpha_t - 2\mathbb{E}_{B_t|\mathbf{w}_t}\theta_{B_t,i}/\alpha_t\right) + \exp -2\mathbb{E}_{B_t|\mathbf{w}_t}\theta_{B_t,i}/\alpha_t}\right]
\end{aligned}$$

For ease of notation, denote $2\theta_{B,i}/\alpha_t - 2\mathbb{E}_{B_t|\mathbf{w}_t}\theta_{B,i}/\alpha_t = \theta,\ \mathbb{E}_{B_t|\mathbf{w}_t} 2\theta_{B,i}/\alpha_t = \mu$, and the pdf of θ is p_θ for the moment, then from Assumption 2, we have θ is mean zero, symmetric around zero, and subgaussian with ψ_2 norm $2\kappa_t/\alpha_t$ by taking $\lambda = 1,\ v = \mathbb{1}_i$ in Sub-Gaussian assumption: $\mathbb{E}_\theta \exp\theta \leq \exp(2\kappa_t^2/\alpha_t^2)$.

Therefore, by changing notation we have

$$\mathbb{E}_{B_t|\mathbf{w}_t} \theta_{B,i} \mathbb{E}_{B_t|\mathbf{w}_t}\left[\frac{\exp\left(2\theta_{B_t,i}/\alpha_t\right) - 1}{\exp\left(2\theta_{B_t,i}/\alpha_t\right) + 1}\right] = \frac{\alpha_t \mu}{2}\mathbb{E}_\theta\left[\frac{\exp\theta - \exp(-\mu)}{\exp\theta + \exp(-\mu)}\right].$$

By symmetricity of θ, we have

$$\begin{aligned}
&\frac{\alpha_t \mu}{2}\mu\mathbb{E}_\theta\left[\frac{\exp\theta - \exp(-\mu)}{\exp\theta + \exp(-\mu)}\right] \\
&\quad= \frac{\alpha_t \mu}{2}\int_{-\infty}^{\infty} p_\theta(x)\left[\frac{\exp x - \exp(-\mu)}{\exp x + \exp(-\mu)}\right]dx \\
&\quad= \frac{\alpha_t \mu}{2}\int_{0}^{\infty} p_\theta(x)\left[\frac{\exp x - \exp(-\mu)}{\exp x + \exp(-\mu)} + \frac{\exp(-x) - \exp(-\mu)}{\exp(-x) + \exp(-\mu)}\right]dx \\
&\quad= \frac{\alpha_t \mu}{2}\int_{0}^{\infty} p_\theta(x)\frac{(\exp x - \exp(-\mu))(\exp(-x) + \exp(-\mu)) + (\exp(-x) - \exp(-\mu))(\exp x + \exp(-\mu))}{(\exp x + \exp(-\mu))(\exp(-x) + \exp(-\mu))}dx \\
&\quad= \alpha_t/2\int_{0}^{\infty} p_\theta(x)\frac{2\mu(1 - \exp(-2\mu))}{(\exp x + \exp(-\mu))(\exp(-x) + \exp(-\mu))}dx
\end{aligned}$$

By symmetricity, we have $p_\theta(x) = p_\theta(-x)$, and therefore

$$\begin{aligned}
&\alpha_t/2\int_{0}^{\infty} p_\theta(x)\frac{2\mu(1 - \exp(-2\mu))}{(\exp x + \exp(-\mu))(\exp(-x) + \exp(-\mu))}dx \\
&= \alpha_t/2\int_{-\infty}^{\infty} p_\theta(x)\frac{\mu(1 - \exp(-2\mu))}{(\exp x + \exp(-\mu))(\exp(-x) + \exp(-\mu))}dx \\
&= \frac{\alpha_t}{2}(\mu(1 - \exp(-2\mu)))\mathbb{E}_\theta\frac{1}{(\exp\theta + \exp(-\mu))(\exp(-\theta) + \exp(-\mu))}
\end{aligned}$$

Since $\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu))) \geq 0$, and $\frac{1}{x}$ is convex on $\mathbb{R}^+$, we have

$$\begin{aligned}&\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\mathbb{E}_\theta \frac{1}{(\exp\theta+\exp(-\mu))(\exp(-\theta)+\exp(-\mu))}\\ \geq &\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{\mathbb{E}_\theta(\exp\theta+\exp(-\mu))(\exp(-\theta)+\exp(-\mu))}.\end{aligned}$$

Using Sub-Gaussian property of θ: $\mathbb{E}_\theta \exp\theta \leq \exp(2\kappa_t^2/\alpha_t^2)$, and symmetricity so $\mathbb{E}_\theta \exp -\theta \leq \exp(2\kappa_t^2/\alpha_t^2)$, we have

$$\begin{aligned}&\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{\mathbb{E}_\theta(\exp\theta+\exp(-\mu))(\exp(-\theta)+\exp(-\mu))}\\ =&\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{1+\exp(-2\mu)+\exp(-\mu)(\mathbb{E}\exp\theta+\mathbb{E}\exp(-\theta))}\\ \geq&\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{1+\exp(-2\mu)+2\exp(-\mu)\exp(2\kappa_t^2/\alpha_t^2)}\\ \geq&\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{2+2\exp(2\kappa_t^2/\alpha_t^2)}.\end{aligned}$$

Switching back to our previous notation:

$$\begin{aligned}\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{2+2\exp(2\kappa_t^2/\alpha_t^2)} &= \mathbb{E}_{B_t|\mathbf{w}_t}\theta_{B,i}(1-\exp(-4\mathbb{E}_{B_t|\mathbf{w}_t}\theta_{B,i}/\alpha_t))\frac{1}{2+2\exp(2\kappa_t^2/\alpha_t^2)}\\ &= \nabla\mathcal{L}_S(\mathbf{w}_t)_i(1-\exp(-4\nabla\mathcal{L}_S(\mathbf{w}_t)_i/\alpha_t))\frac{1}{2+2\exp(2\kappa_t^2/\alpha_t^2)},\end{aligned}$$

which implies we have

$$\mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1})-\mathcal{L}_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{\eta_t}{2+2\exp(2\kappa_t^2/\alpha_t^2)}\sum_i \nabla\mathcal{L}_S(\mathbf{w}_t)_i(1-\exp(-4\nabla\mathcal{L}_S(\mathbf{w}_t)_i/\alpha_t)) + \frac{\eta_t^2}{2}\|\vec{L}\|_1.$$

Using Lemma 8, we have

$$\mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1})-\mathcal{L}_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{\eta_t}{2(1+\exp(2\kappa_t^2/\alpha_t^2))}\sum_i |\nabla\mathcal{L}_S(\mathbf{w}_t)_i|\min\left[|2\nabla\mathcal{L}_S(\mathbf{w}_t)_i/\alpha_t|, 0.5\right]+\frac{\eta_t^2}{2}\|\vec{L}\|_1. \tag{38}$$

We choose α_t such that $\alpha_t \geq 4\|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_\infty$, then we have

$$\mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1})-\mathcal{L}_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{\eta_t}{\alpha_t(1+\exp(2\kappa_t^2/\alpha_t^2))}\|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_2^2 + \frac{\eta_t^2}{2}\|\vec{L}\|_1. \tag{39}$$

We choose α_t such that $\alpha_t \geq \max[\sqrt{2}\kappa_t, 4\|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_\infty]$, then we have

$$\mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1})-\mathcal{L}_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{\eta_t}{c(1+e)}\|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_2^2 + \frac{\eta_t^2}{2}\|\vec{L}\|_1. \tag{40}$$

Therefore, if we choose $\eta_t = 1/\sqrt{T}$, we have

$$\mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1})-\mathcal{L}_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{1}{c(1+e)\sqrt{T}}\|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2T}\|\vec{L}\|_1. \tag{41}$$

By telescope sum, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla\mathcal{L}_S(\mathbf{w}_t)\|_2^2\right] \leq \frac{(1+e)c}{\sqrt{T}}\left(\mathcal{L}_S(\mathbf{w}_0)-\mathcal{L}_S(\mathbf{w}^*)+\frac{1}{2}\|\vec{L}\|_1\right). \tag{42}$$

and note that $1+e<4$ which completes the proof.

□

Lemma 8. *For any x,*

$$|1-\exp(-2x))| \geq \min(|x|, \frac{1}{2}).$$

Proof. Since $\exp(-x) \geq 1-x$, so we have for $x<0$:

$$1-\exp(-2x) \leq 2x \leq x.$$

For $x>0$, since $\exp x \geq 1+x$, we have $\exp -x \leq \frac{1}{1+x}$, so

$$1-\exp(-2x) \geq \frac{2x}{1+2x}.$$

Then when $\frac{1}{2} > x > 0$,

$$\frac{2x}{1+2x} > x,$$

and when $x \geq \frac{1}{2}$,

$$\frac{2x}{1+2x} \geq \frac{1}{2},$$

which completes the proof. □

Theorem 4. *Under Assumption 1 and 2, with $K_i = K, \forall i \in [d]$, for any S, SGLD (EFLD with step size $\rho_t = \sqrt{\eta_t \sigma_t}$, $\alpha_t = \sqrt{\sigma_t/\eta_t}$), $|B_t| = b$, and $\eta_t = \frac{1}{\sqrt{T}}$, can achieve*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L_S(\mathbf{w}_t)\|^2 \leq O\left(\frac{1}{\sqrt{T}}\right) + O\left(K\frac{p\sum_{t=1}^{T}\alpha_t^4 + \log T}{\sqrt{T}}\right) , \tag{11}$$

where the expectation is over the randomness of the algorithm.

Proof. Recall the update of SGLD:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell\left(\mathbf{w}_t, S_{B_t}\right) + \frac{\sigma_t}{\sqrt{2}}\mathcal{N}\left(0, \mathbb{I}_d\right) . \tag{43}$$

$$\begin{aligned}\mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1})] &\leq \mathcal{L}_S(\mathbf{w}_t) + \mathbb{E}[\langle \nabla \mathcal{L}_S(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle] + \frac{L}{2}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2] \\ &= \mathcal{L}_S(\mathbf{w}_t) - \eta_t\|\nabla \mathcal{L}_S(\mathbf{w}_t)\|^2 + \frac{L\eta_t^2}{2}\left(\kappa^2 + p\sigma_t^2/\eta_t^2\right) ,\end{aligned}$$

where κ^2 is the upper bound of the variance of $\nabla \ell(\mathbf{w}_t, S_{B_t})$.

Rearrange the above inequality, with $\alpha_t = \sqrt{\sigma_t/\eta_t}$ we have

$$\eta_t \mathbb{E}\|\nabla \mathcal{L}_S(\mathbf{w}_t)\|^2 \leq \mathcal{L}_S(\mathbf{w}_t) - \mathbb{E}[\mathcal{L}_S(\mathbf{w}_{t+1})] + \frac{L\eta_k^2}{2}\left(\kappa^2 + p\alpha_t^4\right) . \tag{44}$$

Sum over $t=1$ to $t=T$ and apply expectation to each step, we have

$$\sum_{t=1}^{T}\eta_t \mathbb{E}\|\nabla \mathcal{L}_S(\mathbf{w}_t)\|^2 \leq \mathcal{L}_S(\mathbf{w}_t) - \mathcal{L}_S(\mathbf{w}_n^\star) + \sum_{t=1}^{T}\frac{L\eta_k^2}{2}(p\alpha_t^4 + \kappa^2) .$$

Take $\eta_t = \frac{1}{\sqrt{T}}$ for all $k \in [T]$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla \mathcal{L}_S(\mathbf{w}_t)\|^2 \leq O\left(\frac{1}{\sqrt{T}}\right) + O\left(L\frac{p\sum_{t=1}^{T}\alpha_t^4 + \kappa^2\log T}{\sqrt{T}}\right).$$

□

E EXPERIMENT DETAILS

Datasets. We use MNIST LeCun et al. (1998), Fashion-MNIST Xiao et al. (2017), CIFAR-10 Krizhevsky (2009) and CIFAR-100 Krizhevsky (2009) in our experiments.

MNIST dataset: 60,000 black and white training images, including handwritten digits 0 to 9. We use a subset of MNIST with $n = 10,000$ data points where 1,000 samples from each class are randomly selected. Each image of size 28×28 is first re-scaled into [0,1] by dividing each pixel value by 255, then z-scored by subtracting the mean and dividing the standard deviation of the training set.

Fashion-MNIST dataset: 60,000 gray-scale training images and 10,000 test images, including 10 clothing categories such as shirts, dresses, sandals, etc. Each image of size 28×28 is first re-scaled into [0,1] by dividing each pixel value by 255, then z-scored by subtracting the mean and dividing the standard deviation of the training set.

CIFAR-10/-100 dataset: 60,000 color images consisting of 10/100 categories, e.g., airplane, cat, dog etc. The training set includes 50,000 images while the test set contains the rest 10,000 images. Each image of size 32×32 has 3 color channels. We first re-scale each image into [0, 1] by dividing each pixel value by 255, then each image is normalized by subtracting the mean and dividing the standard deviation of the training set for each color channel. We also use *RandomCrop* and *RandomHorizontalFlip* for data augmentation.

Network Architectures. For experiments on both MNIST and Fashion-MNIST, we use a convolutional neural network with two convolutional layers followed by two fully connected layers with ReLU activations. For experiments use the CIFAR-10 dataset, we consider CNN architecture with two convolutional layers and three fully connected layers. The detail of each CNN architecture can be found in Table 1 and Table 2.

Table 1: CNN architecture for MNIST and Fashion MNIST.

Layer	Parameters
Convolution	32 filters of 5×5
Max-Pooling	2×2
Convolution	64 filters of 5×5
Max-Pooling	2×2
Fully connected	1024 units
Softmax	10 units

Table 2: CNN architecture for CIFAR-10.

Layer	Parameters
Convolution	64 filters of 5×5
Max-Pooling	2×2
Convolution	192 filters of 5×5
Max-Pooling	2×2
Fully connected	384 units
Fully connected	192 units
Softmax	10 units

Experimental setup. We are interested in stochastic gradient Langevin dynamics, whose iterative updates are given by $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \boldsymbol{\xi}_t$, where $\boldsymbol{\xi}_t \sim \mathcal{N}\left(\nabla \ell\left(\mathbf{w}_{t-1}, S_{B_t}\right), \sigma_t^2/\eta_t^2 \mathbb{I}_d\right)$. In particular, we consider $\sigma_t = \sqrt{2\eta_t/\beta_t}$, where β_t is the the inverse temperature at time t. For MNIST and Fashion-MNIST, the initial learning rate is $\eta_0 = 0.004$ and it decays by 0.96 after every 5 epochs. For CIFAR-10,the initial learning rate is $\eta_0 = 0.005$ and it decays by 0.995 after every 5 epochs. We use batch size $|B_t| = 100$ for MNIST and Fashion-MNIST, and $|B_t| = 200$ for CIFAR-10.

Motivated by Zhang et al. (2017), we train CNN with SGLD on a smaller subset of MNIST dataset ($n = 10000$) with randomly corrupted labels. The corruption fraction varies from 0% (without label corruption) to 60%. For different level of randomness, we use the same training setting with batch

size $|B_t| = 100$, initial step size $\eta_0 = 0.005$, noise variance $\sigma_t = 0.2 \cdot \eta_t$, and we decay η_t by 0.995 for every 30 epochs.

We are also interested in Noisy Sign-SGD whose iterative updates are given by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{\xi}_t$, where $\boldsymbol{\xi}_{t,j} \sim p_{\boldsymbol{\theta}_{B_t,\alpha_t,j}}(\xi_j) = \frac{\exp(\xi_j \boldsymbol{\theta}_{B_t,\alpha_t,j})}{\exp(-\boldsymbol{\theta}_{B_t,\alpha_t,j})+\exp(\boldsymbol{\theta}_{B_t,\alpha_t,j})}$. The initial learning rate is $\eta_0 = 10^{-4}$ and it decays by 0.1 after every 30 epochs. We use batch size $|B_t| = 100$ for all benchmarks.

All experiments minimize cross-entropy loss for a fixed number of epochs and have been run on NVIDIA Tesla K40m GPUs. For CNN, we repeat each experiment 30 times, and for ResNet-18, we repeat 5 times.

Table 3: Details of Experiments reported in Figure 1 and 4 for MNIST and Fashion-MNIST with CNN

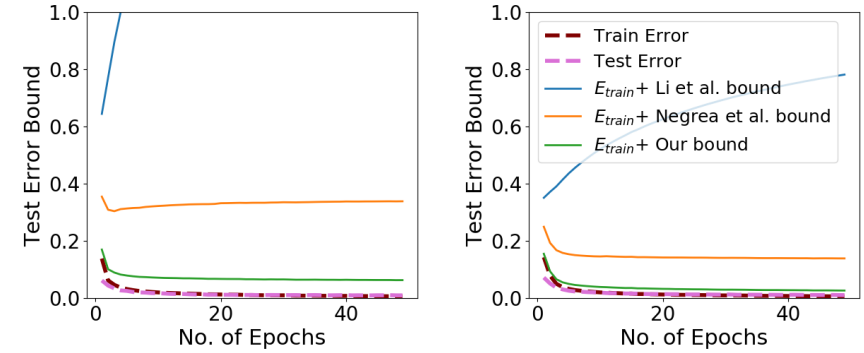
Parameter	Values
Dataset	MNIST/ Fashion-MNIST
Architecture	CNN with 2 conv. layers and 2 linear layers
Batch Size	100
Learning Rate	$\eta_0 = 4 \times 10^{-3}$, decay epochs=5, decay rate=0.96
Inverse Temperature	$\beta \in [5000, 55000]$
Number of Epochs	50
No. of training examples	55000
Number of Repeated Runs	30

Table 4: Details of Experiments reported in Figure 1 and 5 for CIFAR-10 with CNN

Parameter	Values
Dataset	CIFAR-10
Architecture	CNN with 2 conv. layers and 3 linear layers
Batch Size	200
Learning Rate	$\eta_0 = 5 \times 10^{-3}$, decay epochs=5, decay rate=0.995
Inverse Temperature	$\beta \in [5000, 55000]$
Number of Epochs	1000
No. of training examples	55000
Number of Repeated Runs	20

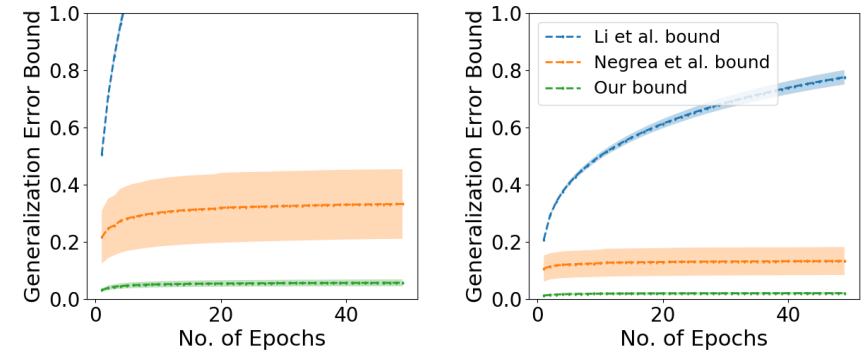
Table 5: Details of Experiments reported in Figure 2 for MNIST with CNN ($\sigma_t = 0.2 \cdot \eta_t$)

Parameter	Values
Dataset	MNIST
Architecture	CNN with 2 conv. layers and 2 linear layers
Batch Size	100
Learning Rate	$\eta_0 = 5 \times 10^{-3}$, decay epochs=30, decay rate=0.995
Number of Epochs	1000
No. of training examples	10000
Number of Repeated Runs	30



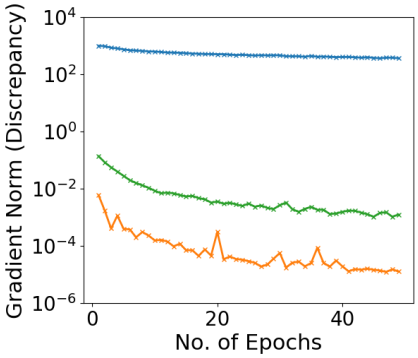
(a) MNIST, $\alpha_t^2 \approx 0.1$, $\beta_t = 55000$

(b) MNIST, $\alpha_t^2 \approx 0.1$, $\beta_t = 5000$

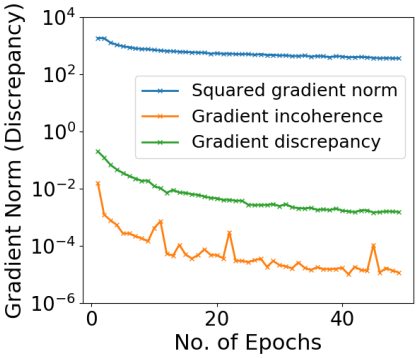


(c) MNIST, $\alpha_t^2 \approx 0.1$, $\beta_t = 55000$

(d) MNIST, $\alpha_t^2 \approx 0.1$, $\beta_t = 5000$

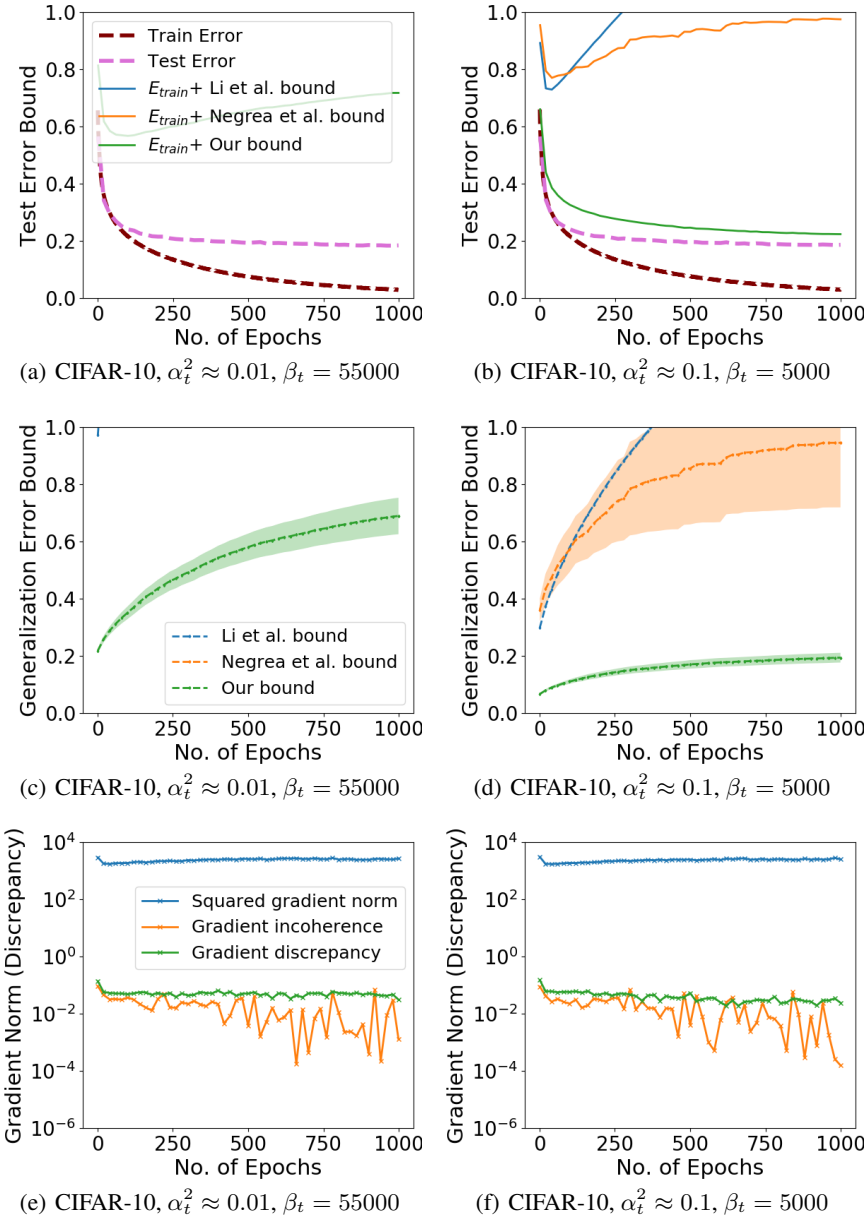


(e) MNIST, $\alpha_t^2 \approx 0.1$, $\beta_t = 55000$



(f) MNIST, $\alpha_t^2 \approx 0.1$, $\beta_t = 5000$

Figure 4: Numerical results for CNN trained on MNIST using SGLD ($\sigma_t = \sqrt{2\eta_t/\beta_t}$). $\beta = 55000$ which is equivalent a small noise variance $\sigma_t \approx 10^{-4}$.



(a) CIFAR-10, $\alpha_t^2 \approx 0.01$, $\beta_t = 55000$

(b) CIFAR-10, $\alpha_t^2 \approx 0.1$, $\beta_t = 5000$

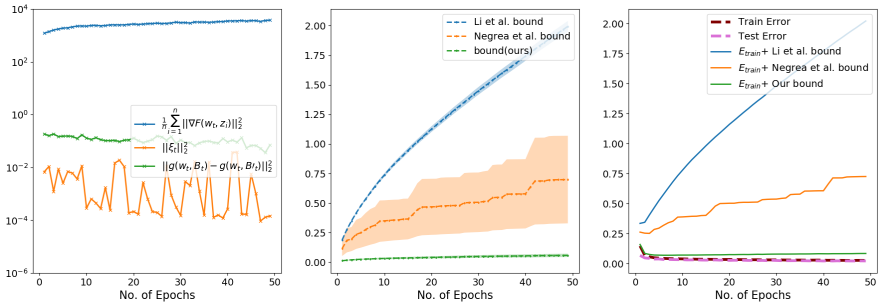
(c) CIFAR-10, $\alpha_t^2 \approx 0.01$, $\beta_t = 55000$

(d) CIFAR-10, $\alpha_t^2 \approx 0.1$, $\beta_t = 5000$

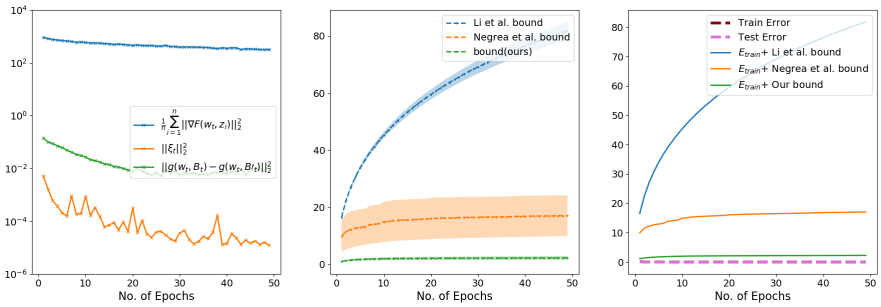
(e) CIFAR-10, $\alpha_t^2 \approx 0.01$, $\beta_t = 55000$

(f) CIFAR-10, $\alpha_t^2 \approx 0.1$, $\beta_t = 5000$

Figure 5: Numerical results for training CNN using SGLD with $\sigma_t = \sqrt{2\eta_t/\beta_t}$ on CIFAR-10.

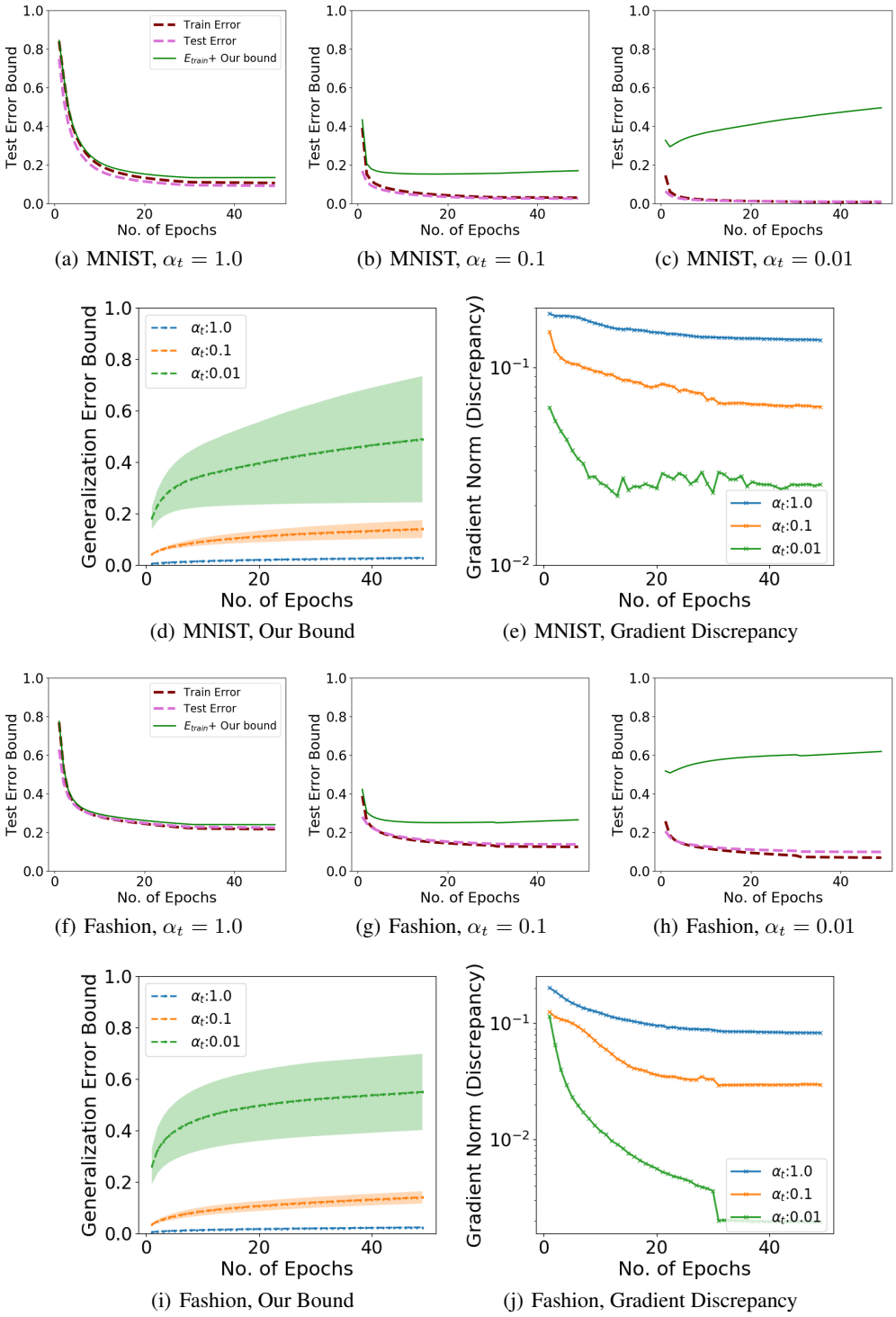


(a) CNN, MNIST, Noisy-SGD, $\alpha_t = 0.2$, $\sigma_t \approx 10^{-4}$



(b) CNN, MNIST, Noisy-SGD, $\alpha_t = 0.002$, $\sigma_t \approx 10^{-6}$

Figure 6: Numerical results for CNN trained on MNIST using Noisy SGD, where $\sigma_t = \alpha_t \eta_t$. We follow the setting described in Li et al. (2020). The initial η_t is 4×10^{-3}, and decays by 0.95 ($\alpha_t = 0.2$) or 0.995 ($\alpha_t = 0.002$) for every 5 epochs.



(a) MNIST, $\alpha_t = 1.0$ (b) MNIST, $\alpha_t = 0.1$ (c) MNIST, $\alpha_t = 0.01$

(d) MNIST, Our Bound (e) MNIST, Gradient Discrepancy

(f) Fashion, $\alpha_t = 1.0$ (g) Fashion, $\alpha_t = 0.1$ (h) Fashion, $\alpha_t = 0.01$

(i) Fashion, Our Bound (j) Fashion, Gradient Discrepancy

Figure 7: Numerical results for training CNN on MNIST and Fashion-MNIST using Noisy Sign-SGD. Increasing α_t leads to a tighter generalization bound even though it leads to a slightly larger gradient discrepancy.