Stability Based Generalization Bounds
for Exponential Family Langevin Dynamics

Arindam Banerjee [1]   **Tiancong Chen** [2]   **Xinyan Li** [2]   **Yingxue Zhou** [2]

Abstract

Recent years have seen advances in generalization bounds for noisy stochastic algorithms, especially stochastic gradient Langevin dynamics (SGLD) based on stability (Mou et al., 2018; Li et al., 2020) and information theoretic approaches (Xu & Raginsky, 2017; Negrea et al., 2019; Steinke & Zakynthinou, 2020). In this paper, we unify and substantially generalize stability based generalization bounds and make three technical contributions. First, we bound the generalization error in terms of expected (not uniform) stability which arguably leads to quantitatively sharper bounds. Second, as our main contribution, we introduce Exponential Family Langevin Dynamics (EFLD), a substantial generalization of SGLD, which includes noisy versions of Sign-SGD and quantized SGD as special cases. We establish data-dependent expected stability based generalization bounds for any EFLD algorithm with a $O(1/n)$ sample dependence and dependence on gradient discrepancy rather than the norm of gradients, yielding significantly sharper bounds. Third, we establish optimization guarantees for special cases of EFLD. Further, empirical results on benchmarks illustrate that our bounds are non-vacuous, quantitatively sharper than existing bounds, and behave correctly under noisy labels.

1. Introduction

Recent years have seen renewed interest 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

---

[1]Department of Computer Science, University of Illinois Urbana-Champaign [2]Department of Computer Science, University of Minnesota, Twin Cities. Correspondence to: Tiancong Chen <chen6271@umn.edu>.

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 (Bousquet & Elisseeff, 2002), although the analysis needed rather small step sizes $\eta_t = 1/t$ which is arguably not useful in practice. While improving the stability 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. 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; Wang et al., 2021b).

While these developments have led to advances in analyzing generalization of noisy SGD algorithms, and we elaborate on these developments in Appendix A, they each have certain limitations, e.g., dependence on global Lipschitz constant L (Mou et al., 2018), tiny $O(1/L)$ step sizes (Li et al., 2020), $O(1/\sqrt{n})$ sample dependence (Mou et al., 2018; Negrea et al., 2019), dependence on gradient norms (Li et al., 2020), restrictions on nature of mini-batching (Wang et al., 2021b), etc. Further, most prior work primarily focuses on SGLD and cannot be readily extended to popular variants such as noisy versions of Sign-SGD or quantized SGD (Bernstein et al., 2018a; Alistarh et al., 2017; Jiang & Agrawal, 2018). In this paper, we build on the core strengths of such existing approaches, most notably (a) the $O(1/n)$ sample dependence of stability based bounds (Mou et al., 2018; Li et al., 2020), (b) the dependence on some measures of gradient discrepancy rather than the norm of gradients (Negrea et al., 2019; Haghifam et al., 2020), and (c) no dependence on the global Lipschitz constant L, and develop a framework (Section 2) for establishing generalization bounds for a general family of noisy stochastic iterative algorithms which includes SGLD as a special case. 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; Farghly & Rebeschini, 2021), and yields distribution dependent generalization bounds which avoid

the worst-case setting of uniform stability. Recall that for any data domain $\mathcal{Z}$ and a distribution D over the domain, uniform stability considers the worst case difference in loss over two datasets $S_n, S_n' \in \mathcal{Z}^n$ of size n which differ by one point, i.e., over $\sup_{S_n, S_n', |S_n \Delta S_n'|=1} \cdots$ (Bousquet & Elisseeff, 2002; Hardt et al., 2016). In Section 2, we show that one gets a valid generalization bound for stochastic algorithms by replacing the supremum $\sup$ by an expectation $\mathbb{E}_{S_n, S_n'}$, where $S_n \sim D^n$ and, without loss of generality, S_n' shares the first $(n-1)$ samples with S_n with the n-th sample $z_n' \sim D$. Replacing $\sup$ by $\mathbb{E}$ makes the bound distribution dependent and arguably leads to quantitatively sharper and computable bounds with less assumptions. Further, we show that expected stability of general noisy stochastic iterative algorithms can be bounded by the expectation of a Le Cam Style Divergence (LSD) between distributions over parameters obtained from S_n and S_n'. Thus, getting an expected stability based generalization bound for a specific stochastic algorithm reduces to that of bounding the expected LSD.

In Section 3, we introduce Exponential Family Langevin Dynamics (EFLD), a family of noisy stochastic gradient descent algorithms based on exponential family noise. Special cases of EFLD include SGLD and noisy versions of Sign-SGD or quantized SGD algorithms (Bernstein et al., 2018a;b; Jin et al., 2020; Alistarh et al., 2017). Our main result provides an expected stability based generalization bound for *any* EFLD algorithm with several aforementioned desirable properties: (a) a $O(1/n)$ sample dependence, (b) a dependence on the gradient discrepancy, a variant of gradient incoherence (Negrea et al., 2019), rather than a dependence on the norm of gradients, (c) no dependence on the global Lipschitz constant L, and (d) step sizes η_t need not be tiny, i.e., $\eta_t = O(1/L)$ is not needed. Existing generalization bounds for SGLD (Li et al., 2020; Negrea et al., 2019) usually use properties of the Gaussian distribution, and do not generalize to EFLD. Our proof technique is new, and uses properties of exponential family distributions. We also provide optimization guarantees for EFLD, i.e., convergence results for noisy Sign-SGD and SGLD.

In Section 4, we present experimental results on benchmark datasets. We illustrate that our bounds for SGLD are non-vacuous and quantitatively tighter than existing bounds (Li et al., 2020; Negrea et al., 2019) due to the desirable dependence on sample size and gradient discrepancy norms, which are empirically shown to be orders of magnitude smaller than gradient norms. We also report results on random labels (Zhang et al., 2017) where as the training error goes to zero, our bound correctly captures the increase in generalization error due to increasing fraction of random labels. We also present results for Noisy Sign-SGD and illustrate that our bounds give a quantitatively tight upper bound on the empirical test error across epochs.

2. Expected Stability based Generalization

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 $S_n \sim D^n$ 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 $L_D(P) \triangleq \mathbb{E}_{z \sim D} \mathbb{E}_{\mathbf{w} \sim P}[\ell(\mathbf{w}, z)]$, and $L_S(P) \triangleq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbf{w} \sim P}[\ell(\mathbf{w}, z_i)]$. We consider a randomized algorithm A which works with $S_n = \{z_1, \ldots, z_n\} \sim D^n$ and creates a distribution over the hypothesis space $\mathcal{W}$. For convenience, we will denote the distribution as $A(S_n)$. The focus of our analysis is to bound the generalization error of A defined as: $\text{gen}(A(S_n)) \triangleq L_D(A(S_n)) - L_S(A(S_n))$. We will assume A is permutation invariant, i.e., the ordering of samples in S_n do not modify $A(S_n)$, an assumption satisfied by most learning algorithms. All technical proofs for results in this section are in 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 $H(P\|P')$ between two distributions given by (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 all $S_n \in \mathcal{Z}^n, z \in \mathcal{Z}$, $\mathbb{E}_{W \sim A(S_n)}[\ell^2(W, z)] \leq c_0^2/4$ for some constant $c_0 > 0$. With $H(\cdot, \cdot)$ denoting the Hellinger divergence, we have*

$$
\begin{aligned}
&|\mathbb{E}_{S_n \sim D^n}[L_D(A(S_n)) - L_S(A(S_n))]| \\
&\leq c_0 \mathbb{E}_{S_n \sim D^n} \mathbb{E}_{z_n' \sim D} \sqrt{2H^2\big(A(S_n), A(S_n')\big)} .
\end{aligned}
\tag{1}
$$

Remark 2.1. Proposition 1 does not need bounded losses. The *second moment* of $\ell(W, z), W \sim A(S_n), \forall S_n, z$ need to be bounded. In comparison, recent information theoretic bounds (Haghifam et al., 2020; Xu & Raginsky, 2017) assume $\ell(\mathbf{w}, Z), Z \sim D, \forall \mathbf{w} \in \mathcal{W}$ to be *sub-Gaussian*. □

Remark 2.2. The bound in Proposition 1 is in terms of *expected stability* where we consider $\mathbb{E}_{S_n \sim D^n} \mathbb{E}_{z_n' \sim D}[\cdots]$, an important departure from bounds based on *uniform stability* (Elisseeff et al., 2005; Bousquet & Elisseeff, 2002; Mou et al., 2018; Bousquet et al., 2020) where one considers $\sup_{S_n, S_n' \in \mathcal{Z}^n, |S_n \Delta S_n'|=1}[\cdots]$. Replacing $\sup$ by $\mathbb{E}$ makes the bounds distribution dependent, avoids the worst

case analysis associated with uniform stability, and arguably leads to quantitatively tighter bounds. □

2.2. Expected Stability of Noisy Iterative Algorithms

We consider a general family of noisy stochastic iterative (NSI) algorithms. Given $S_n \sim D^n$, such iterative algorithms have two (additional) sources of randomness in each iteration t: (a) a stochastic mini-batch of samples S_{B_t}, with $|S_{B_t}| = b \leq n$, drawn uniformly at random with replacement from S_n; and (b) noise $\boldsymbol{\xi}_t$ suitably included in the iterative update. Given a trajectory (realization) 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 $\mathcal{W}$:

$$W_t \sim P_{B_t, \boldsymbol{\xi}_t | \mathbf{w}_{0:(t-1)}}(W) . \quad (2)$$

We will often drop conditioning $\mathbf{w}_{0:(t-1)}$ to avoid clutter.

Let $\bar{S}_{n+1} \sim D^{n+1}$ with $\bar{S}_{n+1} = \{z_1, \ldots, z_{n+1}\}$. Let $S_0 = \{z_1, \ldots, z_{n-1}\}$. S_n, S'_n are size n subsets of $\bar{S}_{n+1}$ with $S_n = S_0 \cup \{z_n\}$ and $S'_n = S_0 \cup \{z'_n\}$, where $z'_n = z_{n+1}$. The algorithms we consider use a mini-batch of size b in each iteration uniformly sampled from S_n or S'_n. Let G, G_0, G_1 respectively denote the set of all mini-batches of size b drawn from $\bar{S}_{n+1}$, the set only drawn from S_0, and the set which includes the last sample, viz. z_n for S_n and z'_n for S'_n. Note that $|G_0| = \binom{n-1}{b}$, $|G_1| = \binom{n-1}{b-1}$, and $|G_0| + |G_1| = \binom{n-1}{b} + \binom{n-1}{b-1} = \binom{n}{b} = |G|$.

Based on (2), let $P_{0:(t-1)}$ denote the joint distribution over $W_{0:(t-1)} = (W_0, \ldots, W_{t-1})$, and let $P_{t|} := P_{B_t, \boldsymbol{\xi}_t | \mathbf{w}_{0:(t-1)}}$ compactly denote the conditional distribution on W_t. 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. For randomized algorithms of the form (2), from Proposition 1 we first bound the Hellinger divergence with KL-divergence, i.e., $2H^2(P_T, P'_T) \leq KL(P_T, P'_T)$ (Proposition 2 in Appendix B), and then use the following chain rule (Pensia et al., 2018; Negrea et al., 2019; Haghifam et al., 2020) to bound the KL-divergence between P_T and P'_T: $KL(P_T \| P'_T) \leq KL(P_{0:T} \| P'_{0:T}) = \sum_{t=1}^{T} \mathbb{E}_{P_{0:(t-1)}} \left[KL(P_{t|} \| P'_{t|}) \right]$. We can bound the per-step 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 $LSD(P \| P') \triangleq \frac{1}{2} \int \frac{(dP - dP')^2}{dP + dP'}$ (where dP denotes the density), our bounds are in terms of

$$LSD(P_{t|} \| \| P'_{t|}) := \mathbb{E}_{B_t \in G_1} \mathbb{E}_{A_t \in G_0} [\Lambda(B_t, A_t)] ,$$
$$\Lambda(B_t, A_t) := \int_{\boldsymbol{\xi}_t} \frac{(dP_{B_t, \boldsymbol{\xi}_t} - dP'_{B_t, \boldsymbol{\xi}_t})^2}{dP_{A_t, \boldsymbol{\xi}_t}} d\boldsymbol{\xi}_t . \quad (3)$$

Note that $P_{B_t, \boldsymbol{\xi}_t}$ and $P'_{B_t, \boldsymbol{\xi}_t}$ represent the conditional 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. Then, we have the following LSD based generalization bound.

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

$$|\mathbb{E}_{S_n}[L_D(A(S_n)) - L_S(A(S_n))]|$$
$$\leq c_1 \frac{b}{n} \mathbb{E}_{S_n} \mathbb{E}_{z'_n} \sqrt{\sum_{t=1}^{T} \mathbb{E}_{W_{0:(t-1)}} \mathbb{E}_{B_t \in G_1} \mathbb{E}_{A_t \in G_0} [\Lambda(B_t, A_t)]} .$$
$$(4)$$

Remark 2.3. Though not stated explicitly, Li et al. (2020) essentially has this result for SGLD and inspired our work. Our proofs are significantly simpler, does not make any additional assumptions, and illustrates applicability to general noisy iterative algorithms of the form (2). □

Remark 2.4. The bound depends on expectations over samples S_n, z'_n, trajectories $W_{0:(t-1)}$, and mini-batches B_t, A_t. Unlike uniform stability and other worst case analysis, there is no sup over samples, trajectories, or mini-batches. □

Remark 2.5. The bound seems to worsen with b, the size of the mini-batch. As we show in Section 3, the LSD terms have a $\frac{1}{b^2}$ dependence for SGLD and its generalizations we introduce, so the leading b is neutralized. □

Remark 2.6. A high probability version of the result based on an exponential version of the Efron-Stein inequality (Boucheron et al., 2013) is presented in Appendix B.2.

3. Exponential Family Langevin Dynamics

Recent years have seen advances in establishing generalization bounds for SGLD (Li et al., 2020; Pensia et al., 2018; Negrea et al., 2019; Haghifam et al., 2020) which adds isotropic Gaussian noise at every step of SGD:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, S_{B_t}) + \mathcal{N}(0, \sigma_t^2 \mathbb{I}) , \quad (5)$$

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. 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 (2). In addition to being a mathematical generalization of the popular SGLD, the proposed EFLD provides flexibility to use noisy gradient algorithms with different representation of the gradient, e.g., skewed Rademacher 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). All technical proofs for results in this section are in Appendix C.

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.,

$$p_\psi(\boldsymbol{\xi}, \boldsymbol{\theta}_\alpha) = \exp(\langle \boldsymbol{\xi}, \boldsymbol{\theta}_\alpha \rangle - \psi(\boldsymbol{\theta}_\alpha))\boldsymbol{\pi}_{0,\alpha}(\boldsymbol{\xi})$$
$$= \prod_{j=1}^{p} \exp(\xi_j \theta_{j\alpha} - \psi_j(\theta_{j\alpha}))\pi_{0,\alpha}(\xi_j) \,,$$

where $\boldsymbol{\xi} \in \mathbb{R}^p$ is the sufficient statistic, $\psi(\boldsymbol{\theta}_\alpha) = \sum_{j=1}^{p} \psi_j(\theta_{j\alpha})$ is the log-partition function, and $\boldsymbol{\pi}_{0,\alpha}(\xi) = \prod_{j=1}^{p} \pi_{0,\alpha}(\xi_j)$ is the base measure. ψ is a smooth convex function by construction (Barndorff-Nielsen, 2014; Banerjee et al., 2005; Wainwright & Jordan, 2008) which implies $\nabla^2_{\boldsymbol{\theta}_\alpha} \psi(\boldsymbol{\theta}_\alpha) \le c_2 \mathbb{I}$ for some constant $c_2 > 0$.

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 $\rho_t > 0$

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \rho_t \boldsymbol{\xi}_t \,, \qquad \boldsymbol{\xi}_t \sim p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B_t, \alpha_t}) \,, \quad (6)$$

where

$$p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B_t, \alpha_t}) = \exp(\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B_t, \alpha_t} \rangle - \psi(\boldsymbol{\theta}_{B_t, \alpha_t}))\boldsymbol{\pi}_{0,\alpha}(\boldsymbol{\xi}) \,,$$
$$\boldsymbol{\theta}_{B_t, \alpha_t} \triangleq \frac{\boldsymbol{\theta}_{B_t}}{\alpha_t} = \frac{\nabla\ell(\mathbf{w}_{t-1}, S_{B_t})}{\alpha_t} \,. \quad (7)$$

For EFLD, the natural parameter $\boldsymbol{\theta}_{B_t, \alpha_t}$ at step t is simply a scaled version of the mini-batch gradient $\nabla\ell(\mathbf{w}_{t-1}, S_{B_t})$. EFLD becomes SGLD when the exponential family is *Gaussian* with $\boldsymbol{\theta}_{B_t} = \boldsymbol{\mu}_{B_t} = \nabla\ell(\mathbf{w}_{t-1}, S_{B_t})$, i.e., the (unscaled) mean $\boldsymbol{\mu}_{B_t}$ is the mini-batch gradient, since the natural and expectation parameters are equal for Gaussians (Barndorff-Nielsen, 2014; Banerjee et al., 2005). EFLD becomes noisy sign-SGD (Bernstein et al., 2018a;b) when the exponential family is a *skewed Rademacher distribution* over $\{-1, +1\}$ with $P(\xi_j = +1) = \frac{\exp(\overline{\nabla}\ell_j)}{\exp(-\overline{\nabla}\ell_j)+\exp(\overline{\nabla}\ell_j)}, P(\xi_j = -1) = 1 - P(\xi_j = +1)$ where $\overline{\nabla}\ell_j = [\nabla\ell(\mathbf{w}_{t-1}, S_{B_t})/\alpha_t]_j$ which becomes Sign-SGD as $\alpha_t \to 0$ (Appendix C.1).

3.2. Expected Stability of EFLD

From Lemma 1, conditioned on a trajectory $W_{0:(t-1)} = \mathbf{w}_{0:(t-1)}$, mini-batches S_{B_t}, S_{A_t}, we can get an expected stability based generalization bound by suitably bounding the expected LSD $\mathbb{E}[\Lambda(B_t, A_t)]$ as in (3). For EFLD, we have the following bound on the per step LSD $\Lambda(B_t, A_t)$.

Theorem 1. *For a given set* $\bar{S}_{n+1} \sim D^{n+1}$ *and* $\mathbf{w}_{t-1}$ *at iteration* $(t - 1)$*, let* $\Delta_{t|\mathbf{w}_{t-1}}(\bar{S}_{n+1}) = \max_{z,z' \in \bar{S}_{n+1}} \|\nabla\ell(\mathbf{w}_{t-1}, z) - \nabla\ell(\mathbf{w}_{t-1}, z')\|_2$. *Further, for a* c_2*-smooth log-partition function* ψ*, let the scaling* $\alpha_{t|\mathbf{w}_{t-1}}$ *be data-dependent such that* $\alpha^2_{t|\mathbf{w}_{t-1}} \ge 8c_2 \Delta^2_{t|\mathbf{w}_{t-1}}(\bar{S}_{n+1})$. *Then, for* $\Lambda(A_t, B_t)$ *as in* (3) *we have*

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

Remark 3.1. Theorem 1 shows that per step LSD $\Lambda(B_t, A_t)$ can be bounded by the (scaled) mini-batch gradient discrepancy. The result holds for *all* EFLD algorithms of the form (6)-(7). ☐

Remark 3.2. Since S_{B_t} and S'_{B_t} only differ at samples z_n and $z'_n = z_{n+1}$, $\nabla\ell(\mathbf{w}_{t-1}, S_{B_t}) - \nabla\ell(\mathbf{w}_{t-1}, S'_{B_t}) = \frac{1}{b}(\nabla\ell(\mathbf{w}_{t-1}, z_n) - \nabla\ell(\mathbf{w}_{t-1}, z'_n))$. The $1/b$ scale factor neutralizes the leading b term in Lemma 1. ☐

Theorem 1 can now be directly applied to Lemma 1 to get expected stability based generalization bounds for EFLD.

Theorem 2. *In the setting of Proposition 1 consider an exponential family Langevin dynamics (EFLD) algorithm of the form* (6)-(7) *with a* c_2*-smooth log-partition function* ψ*. Then, for mini-batch size* $b \le n/2$*, with* $c = c_0\sqrt{5c_2}$ *(with* c_0 *as in Proposition 1) and* $\alpha^2_{t|} \ge 8c_2 \Delta^2_t(\bar{S}_{n+1})$ *(as in Theorem 1, with the conditioning on* $\mathbf{w}_{t-1}$ *hidden to avoid clutter), we have*

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

Remark 3.3. The key term in the bound is the expected gradient discrepancy only on the sample z_n, z'_n where S_n, S'_n differ. Further, the only dependence on the specific exponential family is through the smoothness constant c_2. ☐

Remark 3.4. Since SGLD is a special case of EFLD, Theorem 2 gives a generalization bound for SGLD. The bound has effectively the same dependence on n and T as the bound in Li et al. (2020). However, the bound is quantitatively much sharper since the *gradient norm* term $\frac{1}{n}\sum_{z \in S} \|\nabla\ell(\mathbf{w}_t, z)\|^2$ in Li et al. (2020) gets replaced by the *gradient discrepancy* term $\|\nabla\ell(\mathbf{w}_t, z) - \nabla\ell(\mathbf{w}_t, z')\|^2$. As illustrated in our experiments (Section 4), the gradient discrepancy is orders of magnitude smaller than the gradient norm. The bound in Negrea et al. (2019) depends on a related *gradient incoherence* which we found to be empirically smaller than gradient discrepancy in our experiments (Section 4). However, their bound has a $1/\sqrt{n}$ sample dependence, which is worse than the $1/n$ dependence in our

bound. Wang et al. (2021b) also obtained a $1/n$ rate in their bound depending on the sum of gradient variances. However, their bound scales inversely with b, since gradient variance increases as b decreases. In contrast, our bound is suitable for small batch size as well. Lei & Ying (2020) considered "on average stability" based generalization bounds, but has a dependence either on L or on some form of convexity. Our bound does not depend on L, and works for any non-convex and non-smooth loss. $\square$

Remark 3.5. Our bounds hold for non-convex and/or non-smooth loss functions. Hardt et al. (2016) developed uniform stability based generalization bounds for SGD for smooth losses. To compare with Hardt et al. (2016) for the non-convex case, note that by construction, $\gamma_t^2 := \frac{\|\nabla\ell(w,z_n) - \nabla\ell(w,z_n')\|_2^2}{\alpha^2} \le \frac{\Delta^2(\bar{S}_{n+1})}{\alpha^2} \le \frac{1}{8c_2}$, so that our bound in Theorem 2 can be upper bounded by $\frac{c}{n} \mathbb{E}_{\bar{S}_{n+1}} \sqrt{\sum_{t=1}^{T} \mathbb{E}_{W_{0:(t-1)}} \left[\frac{1}{8c_2}\right]} \le \frac{\sqrt{5}c_0}{2\sqrt{2}} \frac{\sqrt{T}}{n}$, since $c = c_0\sqrt{5c_2}$. For a $\beta = 1$ smooth loss, with step size $\eta_t = 1/t$, Hardt et al. (2016) gets a $O(\sqrt{T}/n)$ bound by their Theorem 3.12. However, *we do not need the loss to be smooth* and *we work with constant step sizes*. Their results do not extend to non-smooth losses or constant step sizes. $\square$

Remark 3.6. 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 (7) 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., $\|\mathbf{g} - \mathbf{g}'\|_{2,\boldsymbol{\alpha}}^2 = \sum_j (g_j - g_j')^2/\alpha_j^2$. $\square$

Remark 3.7. The lower bound on α_t in Theorem 2 is a data-dependent quantity $\Delta_{t|}(\bar{S}_{n+1})$. For SGLD in (5), the condition implies $\eta_t \le \sigma_t/c_3\Delta_{t|}(\bar{S}_{n+1})$, a much more benign (and computable) condition on the step size compared to those in the related work (Mou et al., 2018; Li et al., 2020; Hardt et al., 2016) which require step size to be bounded by σ_t/L, where L is the global Lipschitz constant for the loss ℓ. Note that $\Delta_{t|}(\bar{S}_{n+1}) \ll L$ because L is a uniform bound. Further, $\Delta_{t|}(\bar{S}_{n+1})$ is expected to decrease over iterations, i.e., as t increases, and gradients get smaller. $\square$

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 densities $dP_{B,\boldsymbol{\xi}} = p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B,\alpha})$ and $dP_{B,\boldsymbol{\xi}}' = p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B,\alpha}')$, i.e., densities in the same exponential family but with different parameters $\boldsymbol{\theta}_{B,\alpha}$ and $\boldsymbol{\theta}_{B,\alpha}'$ because of the difference in the mini-batches, 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 $\gamma_{\boldsymbol{\xi}} \in [0,1]$ with the subscript $\boldsymbol{\xi}$ illustrating dependence on $\boldsymbol{\xi}$. Then,

$$
\begin{aligned}
\Lambda_{A,B} &= \int_{\boldsymbol{\xi}} \frac{(p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B,\alpha}) - p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B,\alpha}'))^2}{p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi} \\
&= \int_{\boldsymbol{\xi}} \frac{\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}', \nabla_{\tilde{\boldsymbol{\theta}}_{B,\alpha}} p_\psi(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha}) \rangle^2}{p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi} \\
&= \int_{\boldsymbol{\xi}} \frac{\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}', \boldsymbol{\xi} - \nabla_{\tilde{\boldsymbol{\theta}}_{B,\alpha}} \psi(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha}) \rangle^2 p_\psi^2(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha})}{p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi},
\end{aligned}
\tag{9}
$$

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

Handling Distributional Dependence of $\tilde{\boldsymbol{\theta}}_{B,\alpha}$. Note that it is difficult to proceed with the analysis with the density term depending on parameter $\tilde{\boldsymbol{\theta}}_{B,\alpha}$ since $\tilde{\boldsymbol{\theta}}_{B,\alpha}$ depends on $\boldsymbol{\xi}$ and there is an outside integral over $\boldsymbol{\xi}$ in (9). 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 $\boldsymbol{\theta}_{B,\alpha}'$ essentially using c_2-smoothness of ψ.

Lemma 2. 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

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

In other words, for any $\boldsymbol{\xi}$ we have

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

Since the parameters $\boldsymbol{\theta}_{B,\alpha}, \boldsymbol{\theta}_{B,\alpha}'$ in the right-hand-side depend on $\boldsymbol{\xi}$, the outside integral over $\boldsymbol{\xi}$ in (9) will not pose any unusual challenges.

Bounding the Density Ratio. Next we focus on the density ratio $p_\psi^2(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha})/p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{A,\alpha})$ in (9). 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 distribution in the same exponential family p_ψ with parameters $(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})$.

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

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

In other words, for any $\boldsymbol{\xi}$ we have

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

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 $p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}'_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})$. 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 $p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})$, not $p_\psi(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha})$. We handle this discrepancy by using

$$\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}, \boldsymbol{\xi} - \nabla\psi(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle^2$$
$$= \langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}, (\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]) + (\mathbb{E}[\boldsymbol{\xi}] - \nabla\psi(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha}))\rangle^2$$
$$\le 2\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}, \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]\rangle^2$$
$$+ 2\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}, \mathbb{E}[\boldsymbol{\xi}] - \nabla\psi(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle^2 ,$$

where the expectation $\mathbb{E}[\boldsymbol{\xi}]$ is with respect to $p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})$. Quadratic form for the first term yields the covariance $\mathbb{E}[(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^T] = \nabla^2\psi(\boldsymbol{\theta}_{2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}}) \le c_2\mathbb{I}$, by smoothness and since the covariance matrix of an exponential family is the Hessian of the log-partition function (Wainwright & Jordan, 2008). Since $\mathbb{E}[\boldsymbol{\xi}] = \nabla\psi(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})$, 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 skewed Rademacher noise over $\{-1, +1\}$ and SGLD with Gaussian noise. The details and the proof for results in this subsection are relegated in Appendix D.

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

$$\boldsymbol{\xi}_{t,j} \sim p_{\boldsymbol{\theta}_{B_t,\alpha_t,j}}(\xi_j) = \frac{\exp(\xi_j\theta_{B_t,\alpha_t,j})}{\exp(-\theta_{B_t,\alpha_t,j}) + \exp(\theta_{B_t,\alpha_t,j})},$$

where $\xi_j \in \{-1, +1\}$ and $\boldsymbol{\theta}_{B_t,\alpha_t} = \nabla\ell(\mathbf{w}_{t-1}, S_{B_t})/\alpha_t$ is the scaled mini-batch gradient. The full-batch version uses parameters $\boldsymbol{\theta}_{B_t,\alpha_t} = \nabla L_S(\mathbf{w}_{t-1})/\alpha_t$. For full batch gradient descent, we assume that the loss is smooth.

Assumption 1. *The loss function $L_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}, z_i)$ satisfies: $\forall \mathbf{w}, \mathbf{w}'$, for some non-negative constants $\vec{K} := [K_1, \ldots, K_p]$, we have $L_S(\mathbf{w}) \le L_S(\mathbf{w}') + \nabla L_S(\mathbf{w}')^T(\mathbf{w} - \mathbf{w}') + \frac{1}{2}\sum_i K_i(w_i - w'_i)^2$.*

For mini-batch analysis, we assume the mini-batch gradients are unbiased, symmetric, and sub-Gaussian.

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 $p_{B_t|\mathbf{w}_{t-1}}(\boldsymbol{\xi})$ of $\boldsymbol{\xi} \equiv \nabla\ell(\mathbf{w}_{t-1}, S_{B_t})$ is symmetric 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 $\kappa_t > 0$.*

The smoothness assumption in Assumption 1 is standard in non-convex optimization especially for sign-SGD literature (Bernstein et al., 2018a;b). Assumption 2 for the mini-batch setting helps the theoretical analysis, where (a) is satisfied when the batches S_{B_t} are taken uniformly from samples S as the standard training does; (b) assumes symmetry of the mini-batch gradients; and (c) is similar and stronger assumption compared to Assumption 3 in Bernstein et al. (2018a), where they assume bounded variance for stochastic gradient and our assumption implies suitably bounded higher moments of $\nabla\ell(\mathbf{w}_{t-1}, S_{B_t}) - \nabla L_S(\mathbf{w}_{t-1})$, which is referred as the minibatch noise in recent noisy SGD literature e.g. Damian et al. (2021). Similar to such literature, if we consider mini-batch stochastic gradient be modeled as the average of $|B_t|$ calls to the full-batch gradient, from concentration property κ_t is scaled by $1/\sqrt{|B_t|}$.

Based on the assumptions, we have following optimization guarantee for mini-batch noisy Sign-SGD, the full-batch version can be found in Appendix D.

Theorem 3. *The following holds for any S, any initialization $\mathbf{w}_0$, and the expectation is taken over the randomness of algorithm: if Assumption 1 and 2 hold, 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*

$$\mathbb{E}\|\nabla L_S(\mathbf{w}_R)\|_2^2 \le O\left(\frac{1}{\sqrt{T}}\right) + O\left(\frac{\|\vec{K}\|_1}{\sqrt{T}}\right) ,$$

where $\mathbf{w}_R$ is uniformly randomly sampled from $\{\mathbf{w}_t\}_{t=1}^T$.

Stochastic Gradient Langevin Dynamics (SGLD). We acknowledge that following optimization result for SGLD exists in various forms, as noisy gradient descent algorithms with Gaussian noise 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).
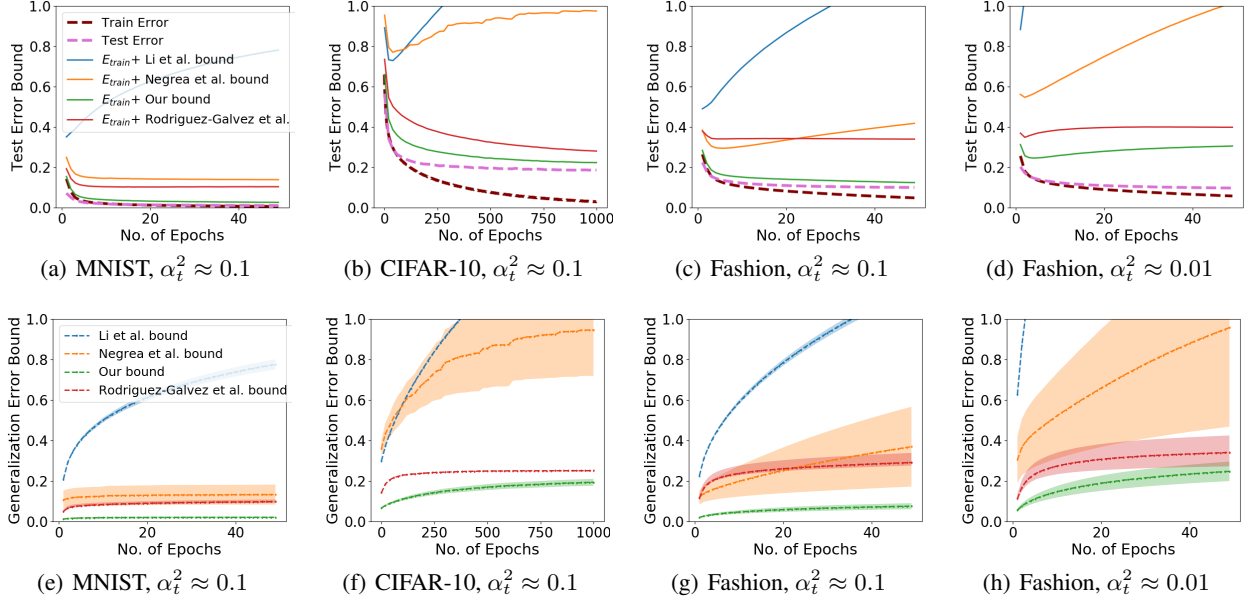
Figure 1. Results for training CNN using SGLD on MNIST, Fashion-MNIST and CIFAR-10. (a)-(d) shows our bound is non-vacuous and can be used to bound the empirical test error. (e)-(h) compare our bound with the existing bounds and show the effect on α_t^2. Our bounds are numerically sharper than existing bounds, and larger α_t^2 leads to tighter generalization bounds which is consistent with Theorem 2.

Theorem 4. *The following holds for any S, any initialization $\mathbf{w}_0$, and the expectation is taken over the randomness of algorithm: under Assumptions 1 and 2 with $K_i = K, \forall i \in [p]$, SGLD, i.e., EFLD with Gaussian noise, $\rho_t = \eta_t$, $\alpha_t = \sigma_t/\eta_t$, and step size $\eta_t = \frac{1}{\sqrt{T}}$ satisfies*

$$\mathbb{E}\|\nabla L_S(\mathbf{w}_R)\|^2 \le O\left(\frac{1}{\sqrt{T}}\right) + O\left(\frac{\frac{K}{2T}\sum_{t=1}^{T}(p\alpha_t^2 + \kappa_t^2)}{\sqrt{T}}\right) ,$$

where $\mathbf{w}_R$ is uniformly randomly sampled from $\{\mathbf{w}_t\}_{t=1}^{T}$.

The error rate of SGLD depends on the noise level α_t and the sub-Gaussian parameter κ_t. The bound has a $O(1/\sqrt{T})$ rate as long as the average noise level and sub-Gaussian parameter are bounded by a constant. Similar to differentially private 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 align with the gradient structure can improve 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 experiments to evaluate our generalization error bounds. For SGLD, we compare our bound in Theorem 2 with existing bounds in Li et al. (2020); Negrea et al. (2019); Rodríguez-Gálvez et al. (2021) for various datasets. Note that the bound presented in Rodríguez-Gálvez et al. (2021) is an extension of that in Haghifam

et al. (2020) from full-batch setting to mini-batch setting. We also evaluate the optimization performance of proposed Noisy Sign-SGD, comparing it with the original Sign-SGD (Bernstein et al., 2018a) and present the corresponding generalization bound in Theorem 2.

The details of model architectures, learning rate schedules, hyper-parameter selections, and additional experimental results can be found in Appendix E. Evaluation of the expectation in Theorem 2 is done based on (re)sampling (Appendix E). We emphasize that the goal for the experiments is to do a comparative study relative to existing bounds. We note that the empirical performance of the methods can potentially be improved with better architectures and training strategies, e.g., deeper/wider networks, data augmentation, batch/layer normalization, etc.

4.1. Stochastic Gradient Langevin Dynamics

Comparison with existing work. We have derived generalization error bounds that depend on the data-dependent *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. All these generalization error bounds can be added to the empirical training error to get bounds on the empirical test error. As shown in Figure 1 (a)-(d), our bound is able to generate a much tighter upper bound on the test error. The improvements is mainly due to the fact that we replace the squared *gradient norm* in
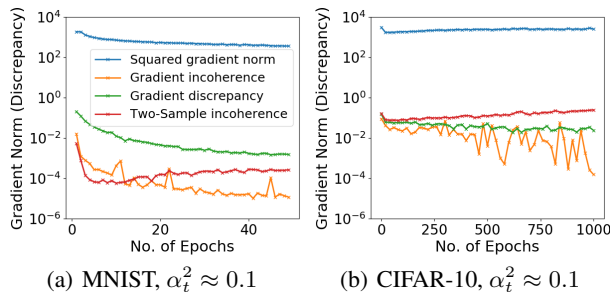
(a) MNIST, $\alpha_t^2 \approx 0.1$ (b) CIFAR-10, $\alpha_t^2 \approx 0.1$

Figure 2. Comparison of the squared gradient norm (Li et al., 2020), the gradient incoherence (Negrea et al., 2019), the two-sample incoherence (Rodríguez-Gálvez et al., 2021), and the gradient discrepancy in our bound. Incoherence or discrepancy based quantities are orders of magnitude smaller than the gradient norm.

Li et al. (2020), the squared norm of *gradient incoherence* in Negrea et al. (2019), and that of *two-sample incoherence* in Rodríguez-Gálvez et al. (2021) with the gradient discrepancy while maintaining a $1/n$ sample dependence. Figure 1 (e)-(h) shows that our bounds are much sharper than those of Li et al. (2020) because our gradient discrepancy norms (Figure 2) are usually 2-4 order of magnitude smaller than the squared gradient norms in Li et al. (2020). Our bounds are also sharper than those of Negrea et al. (2019) and Rodríguez-Gálvez et al. (2021) due to our $O(1/n)$ sample dependence compared to their $O(1/\sqrt{n})$ dependence. Although the gradient incoherence in Negrea et al. (2019) is can be about 1 to 2 order of magnitude smaller than the gradient discrepancy for simple problems such as MNIST (Figure 2(a)), the difference between the gradient incoherence and our gradient discrepancy reduces as the problem becomes harder (see results for CIFAR-10 in Figure 2(b)).

Effect of Random Labels. Motivated by Zhang et al. (2017), we train CNN with SGLD on a smaller subset of MNIST dataset ($n = 10,000$) with randomly corrupted labels. The corruption fraction varies from 0% (no label corruption) to 60%. As shown in Figure 3 (a), for long enough training time, all experiments with different levels of label randomness can achieve almost zero training error. However, increase in random labels leads to increase in gradient discrepancy (Figure 3(b)) which in turn leads to increase in the generalization bound (Figure 3(c)). As a result, as the empirical error rate increases with increase in label randomness (Figure 3(d) dashed lines), we get the correct increase in the test error bound (solid lines).

4.2. Noisy Sign-SGD

In this section, we present numerical results for Noisy Sign-SGD proposed in section 3.1. Since none of the existing bounds can give a valid generalization bound for Noisy Sign-SGD, we only present our bound here.
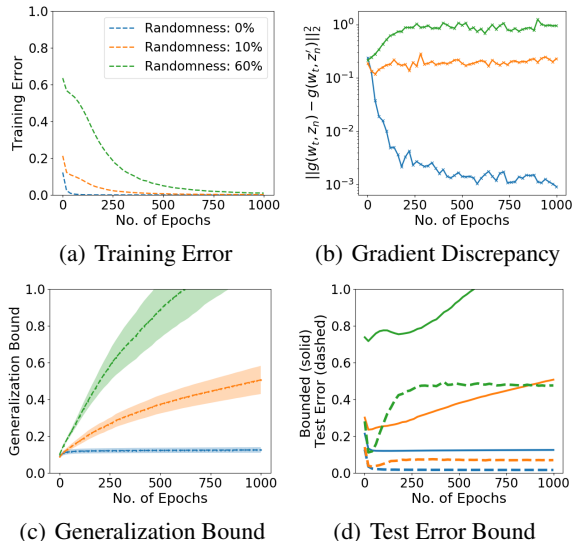

(a) Training Error   (b) Gradient Discrepancy

(c) Generalization Bound   (d) Test Error Bound

Figure 3. Results for training CNN using SGLD on a subset of MNIST ($n = 10000$) with different label randomness.(a) Training error: While training takes more epochs with random labels, training error does go to zero. (b) Gradient discrepancy: As the randomness increases, so does the gradient discrepancy, leading to the increase in the generalization bound. (c) Generalization bound: As the randomness increases, so does the generalization bound in Theorem 2. (d) Test error and bound: As the randomness increases, the empirical test error (dashed lines) increases and so does the test error bound (solid lines); overall, the bounds stay valid.

Optimization. Figure 4 (a)-(b) show the training dynamics of Noisy Sign-SGD under various choices of α_t. For small α_t, Noisy Sign-SGD matches both the optimization trajectory as well as the test accuracy of the original Sign-SGD (Bernstein et al., 2018a). However, as α_t increases, the distribution over $\{-1, +1\}$ is more spread out and the corresponding Noisy Sign-SGD seems to converge but to a sub-optimal value.

Generalization Bound. Figure 4(c)-(d) show that our bound successfully bounds the empirical test error. Larger α_t leads to sharper generalization bounds. However, larger α_t adversely affects the optimization, e.g., Figure 4 (a)-(b) blue and orange lines. The results illustrate the trade-off between the empirical optimization and the generalization bound. 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 algorithms. We have also in-
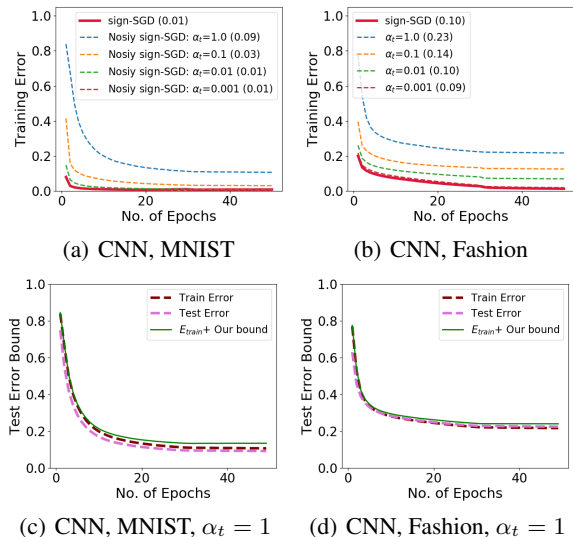
(a) CNN, MNIST

(b) CNN, Fashion

(c) CNN, MNIST, $\alpha_t = 1$

(d) CNN, Fashion, $\alpha_t = 1$

Figure 4. (a)-(b) show the training dynamics of CNN on MNIST and Fashion-MNIST 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). (c)-(d) show that empirical test error can be bounded by our bound and the corresponding training error.

troduced Exponential Family Langevin Dynamics (EFLD), a large 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 discrepancy, rather than gradient norms. Further, we have provided optimization guarantees for special cases of EFLD, viz. 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

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

Asi, H., Duchi, J., Fallah, A., Javidbakht, O., and Talwar, K. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pp. 383–392. PMLR, 2021.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

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

Bassily, R., Smith, A., and Thakurta, A. 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.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 2019.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. signsgd: Compressed optimisation for nonconvex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018a.

Bernstein, J., Zhao, J., Azizzadenesheli, K., and Anandkumar, A. signsgd with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2018b.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

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

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

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

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

Canonne, C. L., Kamath, G., and Steinke, T. The discrete gaussian for differential privacy. In *NeurIPS*, 2020.

Chen, X., Chen, T., Sun, H., Wu, Z. S., and Hong, M. Distributed training with heterogeneous data: Bridging median-and mean-based algorithms. *arXiv preprint arXiv:1906.01736*, 2019.

Damian, A., Ma, T., and Lee, J. Label noise sgd provably prefers flat global minimizers. *arXiv preprint arXiv:2106.06530*, 2021.

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

Elisseeff, A., Evgeniou, T., Pontil, M., and Kaelbing, L. P. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

Farghly, T. and Rebeschini, P. Time-independent generalization bounds for SGLD in non-convex settings. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=tNT4APQ0Wgj.

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

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

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

Haghifam, M., Negrea, J., Khisti, A., Roy, D. M., and Dziugaite, G. K. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 2020.

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

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

Hellström, F. and Durisi, G. 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.

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

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pp. 1724–1732, 2017.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *arXiv preprint arXiv:1902.04811*, 2019.

Jin, R., Huang, Y., He, X., Wu, T., and Dai, H. Stochastic-sign sgd for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

Kairouz, P., Ribero, M., Rush, K., and Thakurta, A. Dimension independence in unconstrained private erm via adaptive preconditioning. *arXiv preprint arXiv:2008.06570*, 2020.

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Technical Report Vol. 1. No. 4., University of Toronto, 2009.

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

Lei, Y. and Ying, Y. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, 2020.

Li, B., Chen, C., Liu, H., and Carin, L. On connecting stochastic gradient mcmc and differential privacy. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 557–566. PMLR, 2019.

Li, J., Luo, X., and Qiao, M. 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.

Mou, W., Wang, L., Zhai, X., and Zheng, K. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pp. 605–638. PMLR, 2018.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. Information-theoretic generalization bounds for sgld via data-dependent estimates. In *Advances in Neural Information Processing Systems*, 2019.

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. Information-theoretic generalization bounds for stochastic gradient descent. In *COLT*, 2021.

Pensia, A., Jog, V., and Loh, P.-L. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 546–550. IEEE, 2018.

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

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703. PMLR, 2017.

Rodríguez-Gálvez, B., Bassi, G., Thobaben, R., and Skoglund, M. 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.

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

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

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

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Stochastic convex optimization. In *COLT*, 2009.

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

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

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

Wainwright, M. J. and Jordan, M. I. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Wang, B., Zhang, H., Zhang, J., Meng, Q., Chen, W., and Liu, T.-Y. Optimizing information-theoretical generalization bound via anisotropic noise of SGLD. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021a. URL https://openreview.net/forum?id=lN2Uqm-ScC.

Wang, D. and Xu, J. 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.

Wang, H., Huang, Y., Gao, R., and Calmon, F. Analyzing the generalization capability of sgld using properties of gaussian channels. *Advances in Neural Information Processing Systems*, 34, 2021b.

Wang, Y.-X., Fienberg, S., and Smola, A. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pp. 2493–2502. PMLR, 2015.

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

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

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

Yang, G., Zhang, T., Kirichenko, P., Bai, J., Wilson, A. G., and De Sa, C. Swalp: Stochastic weight averaging in low-precision training. *36th International Conference on Machine Learning (ICML)*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. 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.

Zhang, H., Mironov, I., and Hejazinia, M. Wide network learning with differential privacy. *arXiv preprint arXiv:2103.01294*, 2021.

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

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

A. Related Work

Uniform stability. Uniform stability is a classical approach for bounding generalization error (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Bousquet et al., 2020; Shalev-Shwartz et al., 2009; Feldman & Vondrak, 2018; 2019), pioneered by Rogers & Wagner (1978); Devroye & Wagner (1979). Recently, uniform stability has been used in analyzing the stability of stochastic gradient descent (SGD) (Hardt et al., 2016). Mou et al. (2018) prove the uniform stability of SGLD (Welling & Teh, 2011; Raginsky et al., 2017) by showing that 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 uniform stability based generalization bounds for SGLD as $\frac{L}{n}\sqrt{\sum_t \eta_t^2/\sigma_t^2}$ which depends on L, the global Lipschitz constant for gradients, and the step size $\eta_t \le \frac{\sigma_t \ln 2}{L}$ (Mou et al., 2018; Li et al., 2020). Recently, Li et al. (2020) followed up on Mou et al. (2018) and derived a data-dependent bound 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 L to the expected gradient norm square. Recently, Farghly & Rebeschini (2021) provided a time-independent bound for SGLD of the form $O(1/n + \eta^{1/2}d^{1/2})$ which requires the step size scales as $O(1/(n^2 d))$ to obtain an $O(1/n)$ bound. 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 bounds. 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 $O(\sqrt{I(S;W)/n})$, where $I(S;W)$ is the mutual information between the algorithm input S and the algorithm output W. Recent work following this approach focus on bounding the mutual information for a broad class of iterative algorithms, including SGLD to obtain 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, which was extended by Rodríguez-Gálvez et al. (2021) from full-batch gradient to stochastic setting. Recently, Wang et al. (2021b) provided a bound for SGLD of the from $O(\frac{b}{n}\sum_{j=1}^{m}\sqrt{\sum_{t\in\mathcal{T}_j}\eta_t^2 Var(\mathbf{g}_j(t))/\sigma_t^2})$, where b is the mini-batch size, m is the number of mini-batches, $\mathcal{T}_j$ contains the indices of iterations for mini-batch S_j and $Var(\mathbf{g}_j(t))$ is the variance of mini-batch gradient on S_j evaluated on the training set. However, Wang et al. (2021b) requires splitting training samples into disjoint m mini-batches before training and obtained a bound depending on the sum of gradient variance. Their bound also scales inversely to the batch size since the number of mini-batches m and gradient variance increase as the batch size decreases. Neu et al. (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), which also provides a unified view of some of the above results.

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. There has been some work (Wang et al., 2015; Li et al., 2019) connecting SGLD with differentially private SGD algorithm (DP-SGD) (Bassily et al., 2020; Wang & Xu, 2019) which usually adds noise with constant variance to the stochastic gradient. 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 has gained popularity as it reduces communication cost in distributed learning. Existing versions of noisy sign-SGD first adds symmetric noise 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. Recently, Wang et al. (2021a) study the anisotropic noise for SGLD, where they optimize the information-theoretical generalization bound by manipulating the noise structure in SGLD. They prove that with constraint to guarantee low

empirical risk, the optimal noise covariance is the square root of the expected gradient covariance. In recent work, Lei & Ying (2020) considered "on average stability" based generalization bounds, which is related to our work. Their bound has a dependence either on L or on some form of convexity.

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 all $S_n \in \mathcal{Z}^n, z \in \mathcal{Z}$, $\mathbb{E}_{W \sim A(S_n)}[\ell^2(W, z)] \le c_0^2/4$ for some constant $c_0 > 0$. With $H(\cdot, \cdot)$ denoting the Hellinger divergence, we have*

$$
|\mathbb{E}_{S_n \sim D^n}[L_D(A(S_n)) - L_S(A(S_n))]|
$$
$$
\le c_0 \mathbb{E}_{S_n \sim D^n} \mathbb{E}_{z'_n \sim D} \sqrt{2H^2(A(S_n), A(S'_n))} . \tag{1}
$$

Proof. Let $S_n = (z_1, \ldots, z_n)$ and $\tilde{S}'_n = (z'_1, \ldots, z'_n)$ two independent random samples and let $S_n^{(i)} = (z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n)$ be the sample that is identical to S except in the i-th example where we replace z_i with z'_i. Note that $S^{(n)} = S'_n$ where S'_n is the dataset obtained by replacing $z_n \in S_n$ with z'_n as in the Proposition statement. Now, by definition we have

$$
|\mathbb{E}_{S_n \sim D^n}[L_D(A(S_n)) - L_S(A(S_n))]|
$$
$$
= |\mathbb{E}_{S_n \sim D^n}[L_S(A(S_n)) - L_D(A(S_n))]|
$$
$$
= \left| \mathbb{E}_{S_n} \mathbb{E}_{\tilde{S}'_n} \mathbb{E}_A \left[\frac{1}{n} \sum_{i=1}^n \left(\ell(A(S_n^{(i)}), z'_i) - \ell(A(S_n), z'_i) \right) \right] \right|
$$
$$
= \left| \mathbb{E}_{S_n} \mathbb{E}_{\tilde{S}'_n} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_A \left[\ell(A(S_n^{(i)}), z'_i) - \ell(A(S_n), z'_i) \right] \right|
$$
$$
\overset{(a)}{=} \left| \mathbb{E}_{S_n} \mathbb{E}_{\tilde{S}'_n} \frac{1}{n} \sum_{i=1}^n \left(\int_{\mathbb{R}^d} \ell(\mathbf{w}; z'_i) p^{(i)}(\mathbf{w}) d\mathbf{w} - \int_{\mathbb{R}^d} \ell(\mathbf{w}; z'_i) p(\mathbf{w}) d\mathbf{w} \right) \right|
$$
$$
= \left| \mathbb{E}_{S_n} \mathbb{E}_{\tilde{S}'_n} \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} \ell(\mathbf{w}; z'_i) \left(p^{(i)}(\mathbf{w}) - p(\mathbf{w}) \right) d\mathbf{w} \right|
$$
$$
= \left| \mathbb{E}_{S_n} \mathbb{E}_{\tilde{S}'_n} \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} \ell(\mathbf{w}; z'_i) \left(\sqrt{p^{(i)}(\mathbf{w})} + \sqrt{p(\mathbf{w})} \right) \left(\sqrt{p^{(i)}(\mathbf{w})} - \sqrt{p(\mathbf{w})} \right) d\mathbf{w} \right|
$$
$$
\overset{(b)}{\le} \left| \mathbb{E}_{S_n} \mathbb{E}_{\tilde{S}'_n} \frac{1}{n} \sum_{i=1}^n \left\{ \left(\int_{\mathbb{R}^d} \ell^2(\mathbf{w}; z'_i) \left(\sqrt{p^{(i)}(\mathbf{w})} + \sqrt{p(\mathbf{w})} \right)^2 d\mathbf{w} \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \left(\sqrt{p^{(i)}(\mathbf{w})} - \sqrt{p(\mathbf{w})} \right)^2 d\mathbf{w} \right)^{\frac{1}{2}} \right\} \right|
$$

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

$$
\left(\int_{\mathbb{R}^d} \ell^2(\mathbf{w}; z'_i) \left(\sqrt{p^{(i)}(\mathbf{w})} + \sqrt{p(\mathbf{w})} \right)^2 d\mathbf{w} \right)^{1/2} \le \left(2 \int_{\mathbb{R}^d} \ell^2(\mathbf{w}; z'_i) p^{(i)}(\mathbf{w}) d\mathbf{w} + 2 \int_{\mathbb{R}^d} \ell^2(\mathbf{w}; z'_i) p(\mathbf{w}) d\mathbf{w} \right)^{1/2}
$$
$$
\le \left(c_0^2/2 + c_0^2/2 \right)^{1/2} = c_0 .
$$

Hence, by definition of the Hellinger divergence, we have

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

$$\overset{(a)}{=} \mathbb{E}_{S_n}\mathbb{E}_{\tilde{S}_n'}\frac{1}{n}\sum_{i=1}^{n}c_0\sqrt{2H^2\left(A(S_n), A(S_n')\right)}$$

$$= c_0\mathbb{E}_{S_n \sim D^n}\mathbb{E}_{z_n' \sim D}\sqrt{2H^2\left(A(S_n), A(S_n')\right)},$$

where (a) follows since the samples are drawn i.i.d. and the randomized algorithm $A(\cdot)$ is permutation invariant. That completes the proof. $\square$

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:

$$KL(P, P') = \int\left(\log\frac{p(\mathbf{w})}{p'(\mathbf{w})}\right)p(\mathbf{w})\,d\mathbf{w}$$

$$= 2\int\left(-\log\sqrt{\frac{p'(\mathbf{w})}{p(\mathbf{w})}}\right)p(\mathbf{w})\,d\mathbf{w}$$

$$\overset{(a)}{\geq} 2\int\left(1 - \sqrt{\frac{p'(\mathbf{w})}{p(\mathbf{w})}}\right)p(\mathbf{w})\,d\mathbf{w}$$

$$= 2\int\left(p(\mathbf{w}) - \sqrt{p(\mathbf{w})p'(\mathbf{w}))}\right)d\mathbf{w}$$

$$= \int\left(p(\mathbf{w}) + p'(\mathbf{w}) - 2\sqrt{p(\mathbf{w})p'(\mathbf{w})}\right)d\mathbf{w}$$

$$= \int\left(\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})}\right)^2 d\mathbf{w}$$

$$= 2H^2(P, P'),$$

where (a) follows since for $z > -1$, $\log(1 + z) \leq z$, using $x = 1 + z$ and changing signs gives $-\log x \geq 1 - x$.

For the second part, note that:

$$2H^2(P, P') = \int_{\mathbf{w}}(\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})})^2 d\mathbf{w}$$

$$= \int_{\mathbf{w}}|\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})}| \times |\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})}|\,d\mathbf{w}$$

$$\leq \int_{\mathbf{w}}|\sqrt{p(\mathbf{w})} - \sqrt{p'(\mathbf{w})}| \times |\sqrt{p(\mathbf{w})} + \sqrt{p'(\mathbf{w})}|\,d\mathbf{w}$$

$$= \int_{\mathbf{w}}|p(\mathbf{w}) - p'(\mathbf{w})|\,d\mathbf{w}$$

$$= TV(P, P'),$$

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. $\square$

B.1. Proofs for Section 2.2

Our first result establishes a bound on the KL-divergence between two component mixture models in terms of the mixing weight of the unique components. Similar results have appeared in (Li et al., 2020; Bun et al., 2018) in related contexts. Our proof is different, simple, and self-contained.

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

$$KL\left(sQ + (1-s)R \,\|\, sQ' + (1-s)R\right) \leq \frac{s^2}{1-s} \int_{\mathbf{w}} \frac{(Q(\mathbf{w}) - Q'(\mathbf{w}))^2}{R(\mathbf{w})} d\mathbf{w} \,. \tag{10}$$

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

$$KL(sQ + (1-s)R\|sQ' + (1-s)R) = KL(U + s(Q - Q')\|U)$$

$$= \int (U(x) + sF(x)) \log\left(\frac{U(x) + sF(x)}{U(x)}\right) dx$$

$$= \int (U(x) + sF(x)) \log\left(1 + \frac{sF(x)}{U(x)}\right) dx$$

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

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

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

$$\left(\frac{1}{2} - \frac{W(x)}{6} + \frac{W^2(x)}{12} - \cdots\right) = \frac{(1 + W(x))\log(1 + W(x)) - W(x)}{W^2(x)} \leq 1 \,, \tag{11}$$

we have

$$KL(sP + (1-s)R\|sQ + (1-s)R) = s^2 \int \frac{F^2(x)}{U(x)}\left(\frac{(1 + W(x))\log(1 + W(x)) - W(x)}{W^2(x)}\right) dx$$

$$\leq s^2 \int \frac{F^2(x)}{U(x)} dx$$

$$\leq \frac{s^2}{(1-s)} \int \frac{(Q(x) - Q'(x))^2}{R(x)} dx \,.$$

That completes the proof. $\qquad\square$

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*

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

Proof. The proof follows the argument in the proof of Lemma 21 in (Li et al., 2020). $\qquad\square$

Lemma 5. *Consider a general noisy stochastic iterative algorithm with updates of the form (2) with mini-batch size $B_t = b$. Then, conditioned on any trajectory $\mathbf{w}_{0:t-1}$, we have*

$$KL\left(P_{t|}\|P'_{t|}\right) \leq \frac{b^2}{n^2}\frac{n}{n-b}\mathbb{E}_{B_t \in G_1}\mathbb{E}_{A_t \in G_0}\left[\int_{\boldsymbol{\xi}_t} \frac{\left(dP_{B_t, \boldsymbol{\xi}_t} - dP'_{B_t, \boldsymbol{\xi}_t}\right)^2}{dP_{A_t, \boldsymbol{\xi}_t}} d\boldsymbol{\xi}_t\right] \,. \tag{12}$$

Proof. By definition, Proposition 3, and Lemma 4 with $s = b/n$, we have

$$KL\left(P_{t|}\|P'_{t|}\right) \overset{(a)}{\leq} \left(\frac{b}{n}\right)^2 \left(\frac{n}{n-b}\right) \int_{\mathbf{w}} \frac{(Q-Q')^2}{R} d\mathbf{w}$$

$$= \frac{b^2}{n^2}\left(1 + \frac{b}{n-b}\right) \int_{\mathbf{w}} \frac{\left(\frac{1}{|G_1|}\left[\sum_{B \in G_1} dP_{B,\boldsymbol{\xi}}(\mathbf{w}) - \sum_{B \in G_1} P'_{B,\boldsymbol{\xi}}(\mathbf{w})\right]\right)^2}{\frac{1}{|G_0|}\sum_{A \in G_0} dP_{A,\boldsymbol{\xi}}(\mathbf{w})} d\mathbf{w}$$

$$\overset{(b)}{\leq} \frac{b^2}{n^2}\left(1 + \frac{b}{n-b}\right) \int_{\mathbf{w}} \frac{1}{|G_1|}\sum_{B \in G_1} \frac{\left(dP_{B,\boldsymbol{\xi}}(\mathbf{w}) - P'_{B,\boldsymbol{\xi}}(\mathbf{w})\right)^2}{\frac{1}{|G_0|}\sum_{A \in G_0} dP_{A,\boldsymbol{\xi}}(\mathbf{w})} d\mathbf{w}$$

$$\overset{(c)}{\leq} \frac{b^2}{n^2}\left(1 + \frac{b}{n-b}\right) \int_{\mathbf{w}} \frac{1}{|G_1|}\sum_{B \in G_1} \left\{\left(dP_{B,\boldsymbol{\xi}}(\mathbf{w}) - P'_{B,\boldsymbol{\xi}}(\mathbf{w})\right)^2 \frac{1}{|G_0|}\sum_{A \in G_0} \frac{1}{dP_{A,\boldsymbol{\xi}}(\mathbf{w})}\right\} d\mathbf{w}$$

$$= \frac{b^2}{n^2}\left(1 + \frac{b}{n-b}\right) \frac{1}{|G_1||G_0|}\sum_{B \in G_1}\sum_{A \in G_0} \int_{\mathbf{w}} \frac{\left(dP_{B,\boldsymbol{\xi}}(\mathbf{w}) - P'_{B,\boldsymbol{\xi}}(\mathbf{w})\right)^2}{dP_{A,\boldsymbol{\xi}}(\mathbf{w})} d\mathbf{w} \,,$$

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. $\square$

Now we have all the pieces to prove the following result:

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

$$|\mathbb{E}_{S_n}[L_D(A(S_n)) - L_S(A(S_n))]|$$

$$\leq c_1 \frac{b}{n}\mathbb{E}_{S_n}\mathbb{E}_{z'_n}\sqrt{\sum_{t=1}^{T} \underset{W_{0:(t-1)}}{\mathbb{E}}\underset{B_t \in G_1}{\mathbb{E}}\underset{A_t \in G_0}{\mathbb{E}} [\Lambda(B_t, A_t)]} \,. \tag{4}$$

Proof. Based on Propositions 1 and 2, we have

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

Applying Lemma 5 to bound $KL\left(P_{t|}\|P'_{t|}\right)$ and noting that $b/(n-b) \leq 1$ for $b \leq n/2$ completes the proof. $\square$

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

$$Y(S) \triangleq \sum_{i=1}^{n}\mathbb{E}_A\left[\mathbb{E}_{Z \sim D}[\ell(A(S), Z)] - \ell(A(S), Z_i)\right] = n\left(L_D(A(S)) - L_S(A(S))\right) \,. \tag{13}$$

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

$$Y(S) - Y(S'_i) = n\left[\left(L_D(A(S)) - L_D(A(S'_i))\right) - \left(L_S(A(S)) - L_{S'_i}(A(S'_i))\right)\right] \tag{14}$$

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*

$$L_D(A(S)) \leq L_S(A(S)) + \mathbb{E}_S\left[L_D(A(S) - L_S(A(S))\right] + \max\left(\frac{1}{n}, \frac{c_1\kappa_A}{\sqrt{n}}\right) \log\left(\frac{16}{\delta}\right) . \tag{15}$$

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

$$V^+ = V^+(S) = \sum_{i=1}^n \mathbb{E}_{S'}[(Y - Y_i')_+^2] , \tag{16}$$

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,*

$$\log \mathbb{E}\left[e^{\lambda(Y - \mathbb{E}Y)}\right] \leq \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbb{E}\left[e^{\lambda V^+/\theta}\right] . \tag{17}$$

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

$$\log \mathbb{P}(Y - \mathbb{E}Y > t) \leq \psi(\lambda) - \lambda t$$

$$\leq \frac{1}{1 - \lambda\theta}\left[\lambda^2\mu_+ + c\lambda^3\kappa_+^2/\theta - (1 - \lambda\theta)\lambda t\right] .$$

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

$$\log \mathbb{P}(Y - \mathbb{E}Y > t) \leq 2 - \min\left\{1, \frac{1}{c_0\sqrt{2\kappa_+}}\right\}t$$

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

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

$$8\exp\left(-n\epsilon\right) \leq \delta/2 \qquad \Rightarrow \qquad \epsilon \geq \frac{1}{n}\log\left(\frac{16}{\delta}\right)$$

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

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. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ $\square$

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

We 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 skewed over $\{-1, +1\}$.

Example C.1 (Gaussian). SGLD uses scaled Gaussian noise with $\psi(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2/2, \alpha_t = \sigma_t/\eta_t, \pi_{0,\alpha_t}(\boldsymbol{\xi}) = \frac{1}{\sqrt{(2\pi)^p\alpha_t^p}}\exp(-\|\boldsymbol{\xi}\|_2^2/2\alpha_t^2)$ so that $p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B_t,\alpha_t}) = \mathcal{N}(\boldsymbol{\theta}_{B_t}, \alpha_t^2\mathbb{I}_d)$. Then, the distribution from the natural parameter form is:

$$p_{\boldsymbol{\theta}/\sigma}(\xi) = \exp(\langle\xi, \boldsymbol{\theta}\rangle/\sigma^2 - \|\theta\|_2^2/(2\sigma^2)) \times \frac{1}{\sqrt{2\pi}\sigma}\exp(-\|\xi\|_2^2/2\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\|\boldsymbol{\xi} - \boldsymbol{\mu}\|_2^2/2\sigma^2) . \tag{18}$$

The expectation parameter $\boldsymbol{\mu} = \nabla\psi(\boldsymbol{\theta}) = \boldsymbol{\theta}$. The scaled expectation parameter $\boldsymbol{\mu}_\alpha = \nabla\psi(\boldsymbol{\theta}_\alpha) = \boldsymbol{\theta}_\alpha = \boldsymbol{\theta}/\sigma = \boldsymbol{\mu}/\sigma$. Since $\xi = x/\sigma$ and the Bregman divergence $d_\phi(\boldsymbol{\xi}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{\mu}\|_2^2$, the expectation parameter form

$$p_{\boldsymbol{\mu}_\alpha}(\boldsymbol{\xi}) = \exp(-\|\boldsymbol{\xi}/\sigma - \boldsymbol{\mu}/\sigma\|_2^2/2)\frac{1}{\sqrt{2\pi}\sigma} = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\|\boldsymbol{\xi} - \boldsymbol{\mu}\|_2^2/2\sigma^2) . \tag{19}$$

By taking $\rho_t = \eta_t$, the update (6) based on $\rho_t\boldsymbol{\xi}_t$ is distributed as $\mathcal{N}(\eta_t\boldsymbol{\theta}_{B_t}, \eta_t^2\alpha_t^2\mathbb{I}_d) = \mathcal{N}(\eta_t\nabla\ell(\mathbf{w}_{t-1}, S_{B_t}), \sigma_t^2\mathbb{I}_d)$. Thus the EFLD update reduces to the SGLD update: $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t\nabla\ell(\mathbf{w}_{t-1}, S_{B_t}) + \mathcal{N}\left(0, \sigma_t^2\mathbb{I}_d\right) .$ $\qquad$ $\square$

Example C.2 (Skewed Rademacher). For skewed Rademacher 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 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: $d_\phi(\xi, \mu) = \xi\log\frac{\xi}{\mu} + (1 - \xi)\log\frac{1-\xi}{1-\mu}$. 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

$$p_{\mu_\alpha}(\xi) = \exp(-d_\phi(\xi, \mu_\alpha)) = \mu_\alpha^\xi(1 - \mu_\alpha)^{1-\xi} . \tag{20}$$

Noisy Sign-SGD takes $\rho_t = \eta_t$ and componentwise $\xi_j \in \{-1, 1\}$, $\pi_{0,\alpha_t}(\xi_j) = 1$, $\psi(\theta) = \log(\exp(-\theta) + \exp(\theta))$ in exponential family update equation (6), the j-th component of exponential family distribution $p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B_t,\alpha_t})$ becomes $p_\psi(\xi_j; \boldsymbol{\theta}_{B_t,\alpha_t,j}) = \frac{\exp(\xi_j \boldsymbol{\theta}_{B_t,\alpha_t,j})}{\exp(-\theta_{B_t,\alpha_t,j}) + \exp(\theta_{B_t,\alpha_t,j})}$. Thus, the EFLD update reduces to a noisy version of Sign-SGD: $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \boldsymbol{\xi}_t, \xi_{t,j} \sim p_\psi(\xi_j, \theta_{B_t,\alpha_t,j}), j \in [p]$, where $\boldsymbol{\theta}_{B_t,\alpha_t} = \nabla \ell(\mathbf{w}_{t-1}, S_{B_t})/\alpha_t$ is the scaled mini-batch gradient. $\qquad \square$

Example C.3 (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

$$p_{\mu_\alpha}(\xi) = \exp(-d_\phi(\xi, \mu_\alpha)) = \mu_\alpha^\xi (1 - \mu_\alpha)^{1-\xi} . \tag{21}$$

By taking the j-th component of exponential family distribution $p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B_t,\alpha_t})$ becomes $p_\psi(\xi_j; \boldsymbol{\theta}_{B_t,\alpha_t,j}) = \frac{\exp(\xi_j \theta_{B_t,\alpha_t,j})}{1+\exp(\theta_{B_t,\alpha_t,j})}$. Thus, the EFLD update reduces to : $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \boldsymbol{\xi}_t, \xi_{t,j} \sim p_\psi(\xi_j, \boldsymbol{\theta}_{B_t,\alpha_t,j}), j \in [p]$, where $\boldsymbol{\theta}_{B_t,\alpha_t} = \nabla \ell(\mathbf{w}_{t-1}, S_{B_t})/\alpha_t$ is the scaled mini-batch gradient. $\qquad \square$

C.2. Proof of Theorem 1

Theorem 1. *For a given set $\bar{S}_{n+1} \sim D^{n+1}$ and $\mathbf{w}_{t-1}$ at iteration $(t-1)$, let $\Delta_{t|\mathbf{w}_{t-1}}(\bar{S}_{n+1}) = \max_{z,z' \in \bar{S}_{n+1}} \|\nabla \ell(\mathbf{w}_{t-1}, z) - \nabla \ell(\mathbf{w}_{t-1}, z')\|_2$. Further, for a c_2-smooth log-partition function ψ, let the scaling $\alpha_{t|\mathbf{w}_{t-1}}$ be data-dependent such that $\alpha_{t|\mathbf{w}_{t-1}}^2 \geq 8c_2 \Delta_{t|\mathbf{w}_{t-1}}^2(\bar{S}_{n+1})$. Then, for $\Lambda(A_t, B_t)$ as in (3) we have*

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

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

$$p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B,\alpha}) - p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B',\alpha}) = \langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \nabla_{\tilde{\boldsymbol{\theta}}_{B,\alpha}} p_\psi(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha}) \rangle , \tag{22}$$

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,

$$I_{A,B} = \int_{\boldsymbol{\xi}} \frac{\left(p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B,\alpha}) - p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B',\alpha})\right)^2}{p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi} = \int_{\boldsymbol{\xi}} \frac{\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \nabla_{\tilde{\boldsymbol{\theta}}_{B,\alpha}} p_\psi(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha}) \rangle^2}{p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi}$$
$$= \int_{\boldsymbol{\xi}} \frac{\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}, \boldsymbol{\xi} - \nabla_{\tilde{\boldsymbol{\theta}}_{B',\alpha}} \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) \rangle^2 p_\psi^2(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha})}{p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{A,\alpha})} d\boldsymbol{\xi} , \tag{23}$$

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. *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*

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

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

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

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. $\qquad\square$

C.2.2. BOUNDING THE DENSITY RATIO

Focusing on density ratio, we have

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

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

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

Proof. Note that

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

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

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

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. $\qquad\square$

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*

$$
\begin{aligned}
\frac{p_\psi^2(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha})}{p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{A,\alpha})} &\leq \exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2\right] \\
&\quad \times \max\Big(\exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right] p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) , \\
&\qquad\qquad \exp\left[2c_2\|\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right] p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha})\Big) \ .
\end{aligned}
\tag{24}
$$

Proof. From Lemma 2, we have

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

Thus,

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

Based on Lemma 3, we have

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

Combining above inequalities we have

$$
\frac{p_\psi^2(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha})}{p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{A,\alpha})} \le \exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2\right]
$$
$$
\times \max\big(\exp\left[2c_2\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right] p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}),
$$
$$
\exp\left[2c_2\|\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2\right] p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{A,\alpha})\big) . \tag{25}
$$

That completes the proof. $\qquad\square$

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

$$
\int_{\boldsymbol{\xi}} \langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}'_{B,\alpha}, \boldsymbol{\xi} - \nabla\psi(\boldsymbol{\xi}; \tilde{\boldsymbol{\theta}}_{B,\alpha})\rangle^2\, p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})d\boldsymbol{\xi} , \tag{26}
$$

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

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

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

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

For T_1, note that

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

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

$$T_2 = \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}, \nabla \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) - \mathbb{E}[\boldsymbol{\xi}] \right\rangle^2 \right]$$

$$= \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\langle \boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}, \nabla \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) - \nabla \psi(2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\rangle^2 \right]$$

$$\leq \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B,\alpha}\|_2^2 \ \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\| \nabla \psi(\tilde{\boldsymbol{\theta}}_{B,\alpha}) - \nabla \psi((2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\|_2^2 \right]$$

$$= c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\| \tilde{\boldsymbol{\theta}}_{B,\alpha} - (2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\|_2^2 \right]$$

$$= c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\| \gamma_{\boldsymbol{\xi}} \boldsymbol{\theta}_{B,\alpha} + (1 - \gamma_{\boldsymbol{\xi}}) \boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{B,\alpha} - (\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\|_2^2 \right]$$

$$= c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\left\| (1 - \gamma_{\boldsymbol{\xi}})(\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{B,\alpha}) - (\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}) \right\|_2^2 \right]$$

$$\leq 2c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \left(\mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[(1 - \gamma_{\boldsymbol{\xi}})^2 \|\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{B,\alpha}\|_2^2 \right] \right.$$

$$\left. + \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[\|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2 \right] \right)$$

$$= 2c_2^2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \ \left(\|\boldsymbol{\theta}_{B',\alpha} - \boldsymbol{\theta}_{B,\alpha}\|_2^2 \mathbb{E}_{\boldsymbol{\xi} \sim p_\psi(\boldsymbol{\xi}; 2\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha})} \left[(1 - \gamma_{\boldsymbol{\xi}})^2 \right] + \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{A,\alpha}\|_2^2 \right)$$

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

Putting everything back together

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

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

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

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

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

$$I_{A,B} \leq c_2 \|\boldsymbol{\theta}_{B,\alpha} - \boldsymbol{\theta}_{B',\alpha}\|_2^2 \times \exp\left(\frac{1}{4}\right) \times \exp\left(\frac{1}{4}\right) \times \left(1 + \frac{1}{4} + \frac{1}{4}\right)$$

$$\leq \frac{5c_2}{2\alpha^2} \|\nabla\ell(\mathbf{w}, S_B) - \nabla\ell(\mathbf{w}, S'_B)\|_2^2 .$$

That completes the proof. □

C.3. Expected Stability of EFLD

Theorem 2. *In the setting of Proposition 1 consider an exponential family Langevin dynamics (EFLD) algorithm of the form (6)-(7) 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 Proposition 1) and $\alpha_{t|}^2 \geq 8c_2\Delta_{t|}^2(\bar{S}_{n+1})$ (as in Theorem 1, with the conditioning on $\mathbf{w}_{t-1}$ hidden to avoid clutter), we have*

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

Proof. Based on Lemma 1, we have

$$|\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]| \leq c_1 \frac{b}{n} \mathbb{E}_S \mathbb{E}_{z'_n} \sqrt{\sum_{t=1}^{T} \underset{W_{0:(t-1)}}{\mathbb{E}} \underset{B_t \in G_1}{\mathbb{E}} \underset{A_t \in G_0}{\mathbb{E}} [I_{A_t,B_t}]} . \tag{27}$$

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

From Theorem 1, we have

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

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

$$= \frac{5c_2}{2b\alpha_{t|\mathbf{w}_{t-1}}^2} \left[\|\nabla\ell\left(\mathbf{w}_{t-1}, z_n\right) - \nabla\ell\left(\mathbf{w}_{t-1}, z'_n\right)\|_2^2\right] ,$$

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

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

That completes the proof. □

D. Optimization Guarantees for EFLD

D.1. Optimization Guarantees for Noisy Sign-SGD

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

$$p_\psi(\boldsymbol{\xi}; \boldsymbol{\theta}_{B,\alpha}) = \exp(\langle \boldsymbol{\xi}, \boldsymbol{\theta}_{B,\alpha}\rangle - \psi(\boldsymbol{\theta}_{B,\alpha}))\boldsymbol{\pi}_0(\boldsymbol{\xi}) , \qquad \boldsymbol{\theta}_{B,\alpha} \triangleq \frac{\boldsymbol{\theta}_B}{\alpha} = \frac{\nabla\ell(\mathbf{w}_t, S_B)}{\alpha} . \tag{28}$$

Note that the corresponding expectation parameter

$$\boldsymbol{\mu}_{B,\alpha} = \nabla_{\boldsymbol{\theta}_{B,\alpha}} \psi(\theta_{B,\alpha}) \,. \tag{29}$$

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

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{\xi}_t \,, \qquad \boldsymbol{\xi}_{t,i} \sim \text{Rad}\left(\frac{1}{1 + \exp\left(-2\nabla L_S(\mathbf{w}_t)/\alpha_t\right)}\right) \,, \forall i \in [d] \,, \tag{30}$$

where $\text{Rad}(x)$ is the parametric Rademacher distribution 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

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{\xi}_t \,, \qquad \boldsymbol{\xi}_{t,i} \sim \text{Rad}\left(\frac{1}{1 + \exp\left(-2\boldsymbol{\theta}_{B_t,\alpha_t}\right)}\right) \,, \forall i \in [d] \,. \tag{31}$$

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

Assumption 1. *The loss function $L_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}, z_i)$ satisfies: $\forall \mathbf{w}, \mathbf{w}'$, for some non-negative constants $\vec{K} := [K_1, \ldots, K_p]$, we have $L_S(\mathbf{w}) \leq L_S(\mathbf{w}') + \nabla L_S(\mathbf{w}')^T(\mathbf{w} - \mathbf{w}') + \frac{1}{2}\sum_i K_i(w_i - w_i')^2$.*

The assumption on on the empirical loss $L_S(\mathbf{w})$ 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 $p_{B_t|\mathbf{w}_{t-1}}(\boldsymbol{\xi})$ of $\boldsymbol{\xi} \equiv \nabla \ell(\mathbf{w}_{t-1}, S_{B_t})$ is symmetric 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 $\kappa_t > 0$.*

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 7. *The following holds for any S, any initialization $\mathbf{w}_0$, and the expectation is taken over the randomness of algorithm: if the loss satisfies Assumption 1, for full-batch noisy Sign-SGD with step size $\eta_t = 1/\sqrt{T}$ and α_t satisfying $c \geq \alpha_t \geq \|\nabla L_S(\mathbf{w}_t)\|_\infty$, we have*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\nabla L_S(\mathbf{w}_t)\|_2^2\right] \leq \frac{5c}{3\sqrt{T}}\left(L_S(\mathbf{w}_0) - L_S(\mathbf{w}^*) + \frac{1}{2}\|\vec{K}\|_1\right). \tag{32}$$

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

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\nabla L_S(\mathbf{w}_t)\|_2^2\right] \leq \frac{4c}{\sqrt{T}}\left(L_S(\mathbf{w}_0) - L_S(\mathbf{w}^*) + \frac{1}{2}\|\vec{K}\|_1\right). \tag{33}$$

Proof. First we prove Equation (32) for full-batched settings. Conditioned at t-th iteration, with Assumption 1, we have

$$L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) \leq \nabla L_S(\mathbf{w}_t)^T(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{1}{2}\sum_{i=1}^{d} K_i(\mathbf{w}_{t+1,i} - \mathbf{w}_{t,i})^2$$

$$= -\eta_t \nabla L_S(\mathbf{w}_t)^T \boldsymbol{\xi}_t + \eta_t^2 \sum_{i=1}^{d} \frac{K_i}{2}.$$

Then taking conditional expectation on both side for above equation we have

$$\mathbb{E}[L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\eta_t \nabla L_S(\mathbf{w}_t)^T \mathbb{E}_{\boldsymbol{\xi}_t|\mathbf{w}_t} \boldsymbol{\xi}_t + \frac{\eta_t^2}{2} \|\vec{K}\|_1$$

$$= -\eta_t \sum_i \nabla L_S(\mathbf{w}_t)_i \left(\frac{2}{1 + \exp\left(-2\nabla L_S(\mathbf{w}_t)_i/\alpha_t\right)} - 1 \right) + \frac{\eta_t^2}{2} \|\vec{K}\|_1$$

$$= -\eta_t \sum_i \nabla L_S(\mathbf{w}_t)_i \left(\frac{\exp\left(2\nabla L_S(\mathbf{w}_t)_i/\alpha_t\right) - 1}{\exp\left(2\nabla L_S(\mathbf{w}_t)_i/\alpha_t\right) + 1} \right) + \frac{\eta_t^2}{2} \|\vec{K}\|_1$$

$$= -\eta_t \sum_i \nabla L_S(\mathbf{w}_t)_i \tanh(\nabla L_S(\mathbf{w}_t)_i/\alpha_t) + \frac{\eta_t^2}{2} \|\vec{K}\|_1$$

$$= -\eta_t \sum_i |\nabla L_S(\mathbf{w}_t)_i|| \tanh(\nabla L_S(\mathbf{w}_t)_i/\alpha_t)| + \frac{\eta_t^2}{2} \|\vec{K}\|_1.$$

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

$$\mathbb{E}[L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{e^2 - 1}{(e^2 + 1)\alpha_t \sqrt{T}} \|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2T} \|\vec{K}\|_1$$

$$\leq -\frac{3}{5c\sqrt{T}} \|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2T} \|\vec{K}\|_1.$$

By telescope sum we have

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

which completes the proof of full-batch updates.

Then we turn to prove Equation (33) for mini-batch settings. From smoothness condition Assumption 1, we have

$$\mathbb{E}[L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t)|\mathbf{w}_t] \leq \mathbb{E}_{\mathbf{w}_{t+1}|\mathbf{w}_t} \nabla L_S(\mathbf{w}_t)^T(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{1}{2} \mathbb{E}_{\mathbf{w}_{t+1}|\mathbf{w}_t} \sum_{i=1}^d L_i(\mathbf{w}_{t+1,i} - \mathbf{w}_{t,i})^2$$

$$= \mathbb{E}_{\boldsymbol{\xi}_t|\mathbf{w}_t} \nabla L_S(\mathbf{w}_t)^T(-\eta_t \boldsymbol{\xi}_t) + \frac{1}{2} \mathbb{E}_{\boldsymbol{\xi}_t|\mathbf{w}_t} \sum_{i=1}^d L_i(-\eta_t \boldsymbol{\xi}_t)^2$$

$$= -\eta_t \nabla L_S(\mathbf{w}_t)^T \mathbb{E}_{\boldsymbol{\xi}_t|\mathbf{w}_t} \boldsymbol{\xi}_t + \frac{\eta_t^2}{2} \|\vec{K}\|_1$$

$$= -\eta_t \sum_i \nabla L_S(\mathbf{w}_t)_i \mathbb{E}_{B_t|\mathbf{w}_t} \left[\frac{2}{1 + \exp\left(-2\nabla_i \ell(\mathbf{w}_t, S_{B_t}/\alpha_t)\right)} - 1 \right] + \frac{\eta_t^2}{2} \|\vec{K}\|_1$$

$$= -\eta_t \sum_i \nabla L_S(\mathbf{w}_t)_i \mathbb{E}_{B_t|\mathbf{w}_t} \left[\frac{\exp\left(2\nabla_i \ell(\mathbf{w}_t, S_{B_t})/\alpha_t\right) - 1}{\exp\left(2\nabla_i \ell(\mathbf{w}_t, S_{B_t})/\alpha_t\right) + 1} \right] + \frac{\eta_t^2}{2} \|\vec{K}\|_1$$

$$= -\eta_t \sum_i \mathbb{E}_{B_t|\mathbf{w}_t}[\theta_{B_t,i}] \mathbb{E} \left[\frac{\exp\left(2\theta_{B_t,i}/\alpha_t\right) - 1}{\exp\left(2\theta_{B_t,i}/\alpha_t\right) + 1} \right] + \frac{\eta_t^2}{2} \|\vec{K}\|_1.$$

Focus on each individual term in the sum, we have

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

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

For ease of notation, denote $2\theta_{B,i}/\alpha_t - \mathbb{E}_{B_t|\mathbf{w}_t}[2\theta_{B,i}/\alpha_t] = \theta$, $\mathbb{E}_{B_t|\mathbf{w}_t}[2\theta_{B,i}/\alpha_t] = \mu$, 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 the sub-Gaussian assumption: $\mathbb{E}_\theta[\exp\theta] \leq \exp(2\kappa_t^2/\alpha_t^2)$.

Therefore, by changing notation we have

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

By symmetry of the distribution of θ, we have

$$\frac{\alpha_t\mu}{2}\mu\mathbb{E}_\theta\left[\frac{\exp\theta-\exp(-\mu)}{\exp\theta+\exp(-\mu)}\right]$$

$$= \frac{\alpha_t\mu}{2}\int_{-\infty}^\infty p_\theta(x)\left[\frac{\exp x-\exp(-\mu)}{\exp x+\exp(-\mu)}\right]dx$$

$$= \frac{\alpha_t\mu}{2}\int_0^\infty p_\theta(x)\left[\frac{\exp x-\exp(-\mu)}{\exp x+\exp(-\mu)} + \frac{\exp(-x)-\exp(-\mu)}{\exp(-x)+\exp(-\mu)}\right]dx$$

$$= \frac{\alpha_t\mu}{2}\int_0^\infty p_\theta(x)\frac{(\exp x-\exp(-\mu))(\exp(-x)+\exp(-\mu))+(\exp(-x)-\exp(-\mu))(\exp x+\exp(-\mu))}{(\exp x+\exp(-\mu))(\exp(-x)+\exp(-\mu))}dx$$

$$= \alpha_t/2\int_0^\infty p_\theta(x)\frac{2\mu(1-\exp(-2\mu))}{(\exp x+\exp(-\mu))(\exp(-x)+\exp(-\mu))}dx$$

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

$$\alpha_t/2\int_0^\infty p_\theta(x)\frac{2\mu(1-\exp(-2\mu))}{(\exp x+\exp(-\mu))(\exp(-x)+\exp(-\mu))}dx$$

$$= \alpha_t/2\int_{-\infty}^\infty p_\theta(x)\frac{\mu(1-\exp(-2\mu))}{(\exp x+\exp(-\mu))(\exp(-x)+\exp(-\mu))}dx$$

$$= \frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\mathbb{E}_\theta\frac{1}{(\exp\theta+\exp(-\mu))(\exp(-\theta)+\exp(-\mu))}$$

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

$$\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\mathbb{E}_\theta\frac{1}{(\exp\theta+\exp(-\mu))(\exp(-\theta)+\exp(-\mu))}$$

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

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

$$\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{\mathbb{E}_\theta(\exp\theta+\exp(-\mu))(\exp(-\theta)+\exp(-\mu))}$$

$$= \frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{1+\exp(-2\mu)+\exp(-\mu)(\mathbb{E}\exp\theta+\mathbb{E}\exp(-\theta))}$$

$$\geq \frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{1+\exp(-2\mu)+2\exp(-\mu)\exp(2\kappa_t^2/\alpha_t^2)}$$

$$\geq \frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{2+2\exp(2\kappa_t^2/\alpha_t^2)}.$$

Switching back to our previous notation:

$$\frac{\alpha_t}{2}(\mu(1-\exp(-2\mu)))\frac{1}{2+2\exp(2\kappa_t^2/\alpha_t^2)} = \mathbb{E}_{B_t|\mathbf{w}_t}\theta_{B,i}(1-\exp(-4\mathbb{E}_{B_t|\mathbf{w}_t}\theta_{B,i}/\alpha_t))\frac{1}{2+2\exp(2\kappa_t^2/\alpha_t^2)}$$

$$= \nabla L_S(\mathbf{w}_t)_i(1-\exp(-4\nabla L_S(\mathbf{w}_t)_i/\alpha_t))\frac{1}{2+2\exp(2\kappa_t^2/\alpha_t^2)},$$

which implies we have

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

Using Lemma 8, we have

$$\mathbb{E}[L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{\eta_t}{2(1 + \exp(2\kappa_t^2/\alpha_t^2))} \sum_i |\nabla L_S(\mathbf{w}_t)_i| \min\left[|2\nabla L_S(\mathbf{w}_t)_i/\alpha_t|, 0.5\right] + \frac{\eta_t^2}{2}\|\vec{K}\|_1. \quad (34)$$

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

$$\mathbb{E}[L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{\eta_t}{\alpha_t(1 + \exp(2\kappa_t^2/\alpha_t^2))}\|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{\eta_t^2}{2}\|\vec{K}\|_1. \quad (35)$$

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

$$\mathbb{E}[L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{\eta_t}{c(1 + e)}\|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{\eta_t^2}{2}\|\vec{K}\|_1. \quad (36)$$

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

$$\mathbb{E}[L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t)|\mathbf{w}_t] \leq -\frac{1}{c(1 + e)\sqrt{T}}\|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2T}\|\vec{K}\|_1. \quad (37)$$

By telescope sum, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla L_S(\mathbf{w}_t)\|_2^2\right] \leq \frac{(1+e)c}{\sqrt{T}}\left(L_S(\mathbf{w}_0) - L_S(\mathbf{w}^*) + \frac{1}{2}\|\vec{K}\|_1\right). \quad (38)$$

With $\mathbf{w}_R$ to be uniformly randomly sampled from $\{\mathbf{w}_1, .., \mathbf{w}_T\}$, we have

$$\mathbb{E}\|\nabla L_S(\mathbf{w}_R)\|_2^2 \leq \frac{(1+e)c}{\sqrt{T}}\left(L_S(\mathbf{w}_0) - L_S(\mathbf{w}^*) + \frac{1}{2}\|\vec{K}\|_1\right), \quad (39)$$

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

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

$$|\tanh x| \geq \frac{e^2 - 1}{e^2 + 1}|x| \quad (40)$$

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. □

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. $\qquad\square$

Theorem 4. *The following holds for any S, any initialization $\mathbf{w}_0$, and the expectation is taken over the randomness of algorithm: under Assumptions 1 and 2 with $K_i = K, \forall i \in [p]$, SGLD, i.e., EFLD with Gaussian noise, $\rho_t = \eta_t$, $\alpha_t = \sigma_t/\eta_t$, and step size $\eta_t = \frac{1}{\sqrt{T}}$ satisfies*

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

where $\mathbf{w}_R$ is uniformly randomly sampled from $\{\mathbf{w}_t\}_{t=1}^{T}$.

Proof. Recall that the update of SGLD is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, S_{B_t}) + \sigma_t \mathcal{N}(0, \mathbb{I}_d). \tag{41}$$

By smoothness of the loss, taking expectation w.r.t. the randomness of the mini-batch B_t and the Gaussian draw $\mathbf{g}_t \sim \mathcal{N}(0, \mathbb{I}_d)$ conditioned on $\mathbf{w}_{1:t}$, we have

$$\mathbb{E}[L_S(\mathbf{w}_{t+1})] \leq L_S(\mathbf{w}_t) + \mathbb{E}[\langle \nabla L_S(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle] + \frac{K}{2}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2]$$

$$= L_S(\mathbf{w}_t) + \mathbb{E}[\langle \nabla L_S(\mathbf{w}_t), -\eta_t \nabla \ell(\mathbf{w}_t, S_{B_t}) + \sigma_t \mathbf{g}_t \rangle] + \frac{K}{2}\mathbb{E}[\| - \eta_t \nabla \ell(\mathbf{w}_t, S_{B_t}) + \sigma_t \mathbf{g}_t\|^2]$$

$$\overset{(a)}{\leq} L_S(\mathbf{w}_t) - \eta_t \|\nabla L_S(\mathbf{w}_t)\|^2 + \frac{K\eta_t^2}{2}\left(c_3\kappa_t^2 + p\sigma_t^2/\eta_t^2\right),$$

where (a) follows since $\mathbb{E}[\nabla \ell(\mathbf{w}_t, S_{B_t})] = 0, \mathbb{E}[\mathbf{g}] = 0, \mathbb{E}[\|\mathbf{g}\|^2] = p$, and $\mathbb{E}[\|\nabla \ell(\mathbf{w}_t, S_{B_t})\|^2] \leq c_3\kappa_t^2$ for some absolute constant c_3 since $\nabla \ell(\mathbf{w}_t, S_{B_t})$ is sub-Gaussian with ψ_2-norm κ_t.

Rearranging the above inequality and using $\alpha_t = \sigma_t/\eta_t$ we have

$$\eta_t \|\nabla L_S(\mathbf{w}_t)\|^2 \leq L_S(\mathbf{w}_t) - \mathbb{E}[L_S(\mathbf{w}_{t+1})] + \frac{K\eta_k^2}{2}\left(c_3\kappa_t^2 + p\alpha_t^2\right). \tag{42}$$

Summing over $t = 1$ to $t = T$ and apply expectation over the trajectory at each step, we have

$$\sum_{t=1}^{T} \eta_t \mathbb{E}\|\nabla L_S(\mathbf{w}_t)\|^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}^\star) + \sum_{t=1}^{T} \frac{K\eta_t^2}{2}(p\alpha_t^2 + \kappa_t^2),$$

where $\mathbf{w}^*$ is a minima of $L_S(\mathbf{w})$. With $\eta_t = \frac{1}{\sqrt{T}}$ for all $t \in [T]$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L_S(\mathbf{w}_t)\|^2 \leq \frac{L_S(\mathbf{w}_1) - L_S(\mathbf{w}^\star)}{\sqrt{T}} + \frac{\frac{K}{2T}\sum_{t=1}^{T}(p\alpha_t^2 + c_3\kappa_t^2)}{\sqrt{T}} \ .$$

With $\mathbf{w}_R$ to be uniformly randomly sampled from $\{\mathbf{w}_1, .., \mathbf{w}_T\}$, we have

$$\mathbb{E}\|\nabla L_S(\mathbf{w}_R)\|^2 \leq \frac{L_S(\mathbf{w}_1) - L_S(\mathbf{w}^\star)}{\sqrt{T}} + \frac{\frac{K}{2T}\sum_{t=1}^{T}(p\alpha_t^2 + c_3\kappa_t^2)}{\sqrt{T}} \ .$$

That completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

E. Experiment Details

E.1. 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.

E.2. 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

E.3. Experimental Setup

We are interested in stochastic gradient Langevin dynamics, whose iterative updates are given by $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \nabla \ell(\mathbf{w}_{t-1}, S_{B_t}) + \mathcal{N}\left(0, \sigma_t^2 \mathbb{I}_d\right)$. We also denote $\beta_t = 2\eta_t/\sigma_t^2$ as the inverse temperature at time t. For MNIST and

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

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 $\boldsymbol{\xi}_{t,j} \sim p_{\boldsymbol{\theta}_{B_t, \alpha_t, j}}(\xi_j) = \frac{\exp(\xi_j \boldsymbol{\theta}_{B_t, \alpha_t})}{\exp(-\boldsymbol{\theta}_{B_t, \alpha_t}) + \exp(\boldsymbol{\theta}_{B_t, \alpha_t})}$. 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.

E.4. Evaluation of the bound in Theorem 2

Estimation of c_0 in Proposition 1. Something similar to c_0 exists in all prior bounds, e.g., see Lemma 1 of Xu & Raginsky (2017), Theorem 9 in Li et al. (2020), etc. If the loss is bounded, then one can get c_0, but in general, it is difficult. Empirically, we chose c_0 based on maximum observed training loss for our bound as well as baseline approaches.

Computation of $\Delta_t^2(\bar{S}_{n+1})$ and the expectation over S_{n+1} in Theorem 2. We start with the uniform analysis. Let L be the Lipschitz constant so $\|\nabla \ell(w, z)\|_2 \le L$. Then $\Delta_{t|}(\bar{S}_{n+1}) \le 2L$. So, suffices to have $\alpha_t = 4\sqrt{2}c_2 L$. This will always work if L is known or ensured using gradient clipping.

More generally, for any given $\bar{S}_{n+1}$, $\Delta_{t|}(\bar{S}_{n+1})$ can be computed by definition from Theorem 1 by taking the maximum discrepancy over all $\binom{n+1}{2}$ pairs of points in $\bar{S}_{n+1}$. The argmax for the discrepancy will arguably have $z \in S_n$ and $z' = z_{n+1}$, which simplifies the argmax to just consider n pairs rather than $\binom{n+1}{2}$. We verified and used this simpler calculation for the experiments.

The expectation over $\bar{S}_{n+1}$ is hard to compute but (re)sampling based estimates can be used. The term entails sampling $\bar{S}_{n+1}$, running the training with n samples S_n, using z_{n+1} to compute the discrepancy. To get Monte Carlo estimates, the analysis has to be repeated for different z_{n+1} (done in the main paper) or by resampling (S_n, z_{n+1}), and computing the average which we show in Figure 5 (a). The estimates are shown to have rather low variance.

Dependence of the bound on batchsize b. Our bound in Theorem 2 does not have a dependence on b, because of what Remark 3.2 shows, i.e., the $1/b$ scale factor neutralizes the leading b term in Lemma 1. The gradient discrepancy term itself may have a mild dependence on b, with smaller batch sizes having mildly smaller empirical gradient discrepancy as shown in Figure 5 (b).
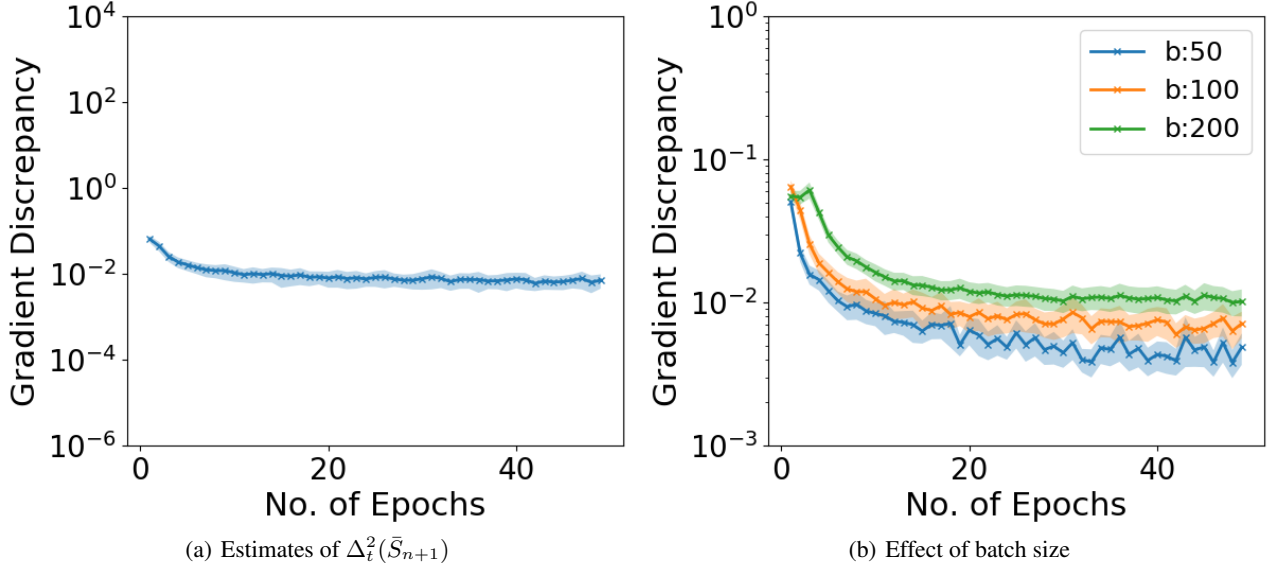
(a) Estimates of $\Delta_t^2(\bar{S}_{n+1})$

(b) Effect of batch size

Figure 5. Results for training CNN using SGLD on MNIST. (a) Estimation of $\Delta_t^2(\bar{S}_{n+1})$ using re-sampled (S_n, z_{n+1}). The variance of such estimation is small. (b) The gradient discrepancy for different batch sizes. Overall, the differences are not significant.

Table 3. Details of Experiments reported in Figure 1 and 6 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 7 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 3 for MNIST with CNN ($\sigma_t = 0.2 \cdot \eta_t$)

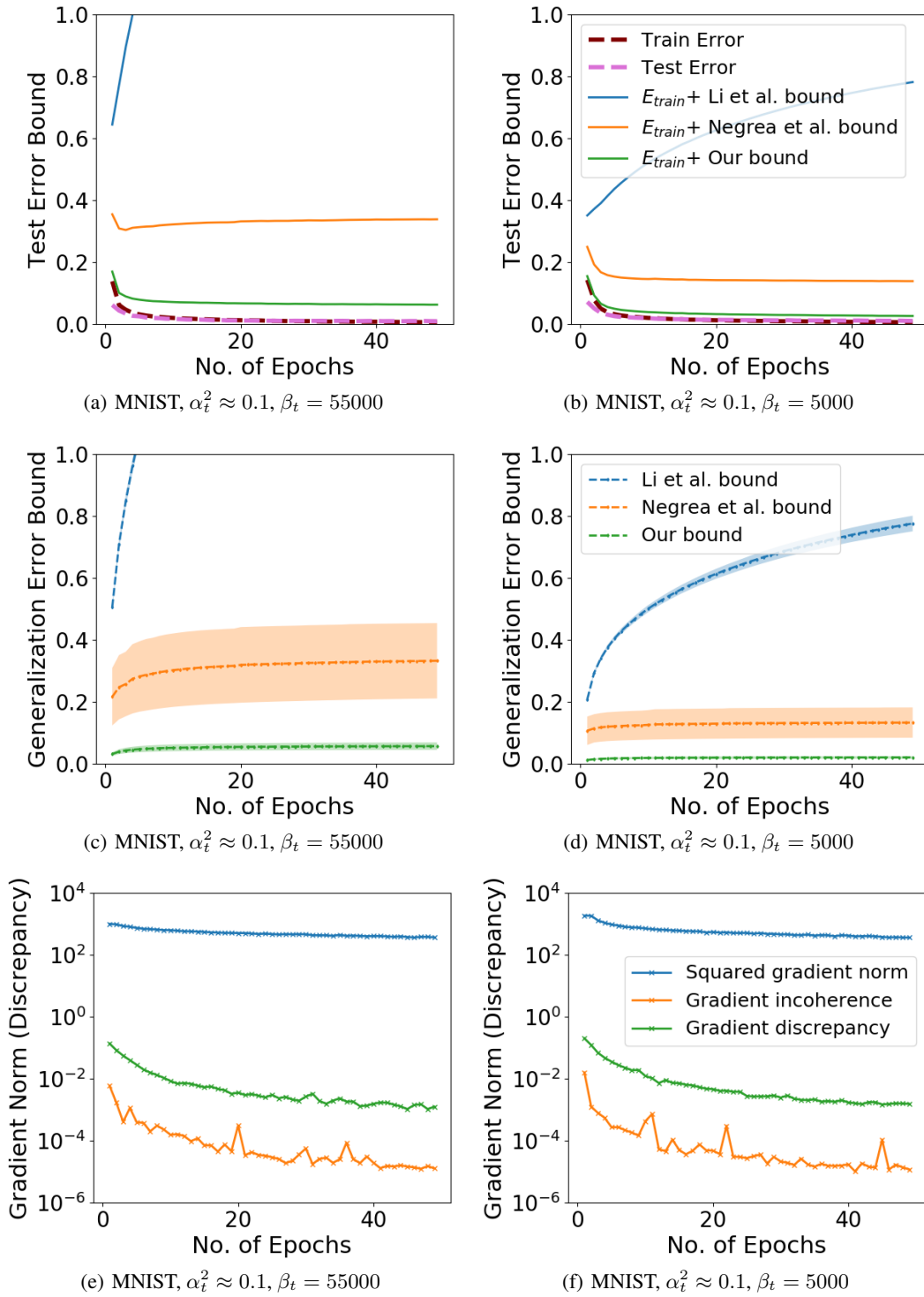| Parameter	Values
Dataset	MNIST
Architecture	CNN with 2 conv. layers and 2 linear layers
Batch Size	100
Learning Rate	$\eta_0 = 5 \times 10^{-3}$, decay epochs=30, decay rate=0.995
Number of Epochs	1000
No. of training examples	10000
Number of Repeated Runs	30

(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 6. Numerical results for CNN trained on MNIST using SGLD with a small noise variance $\sigma_t \approx 10^{-4}$.

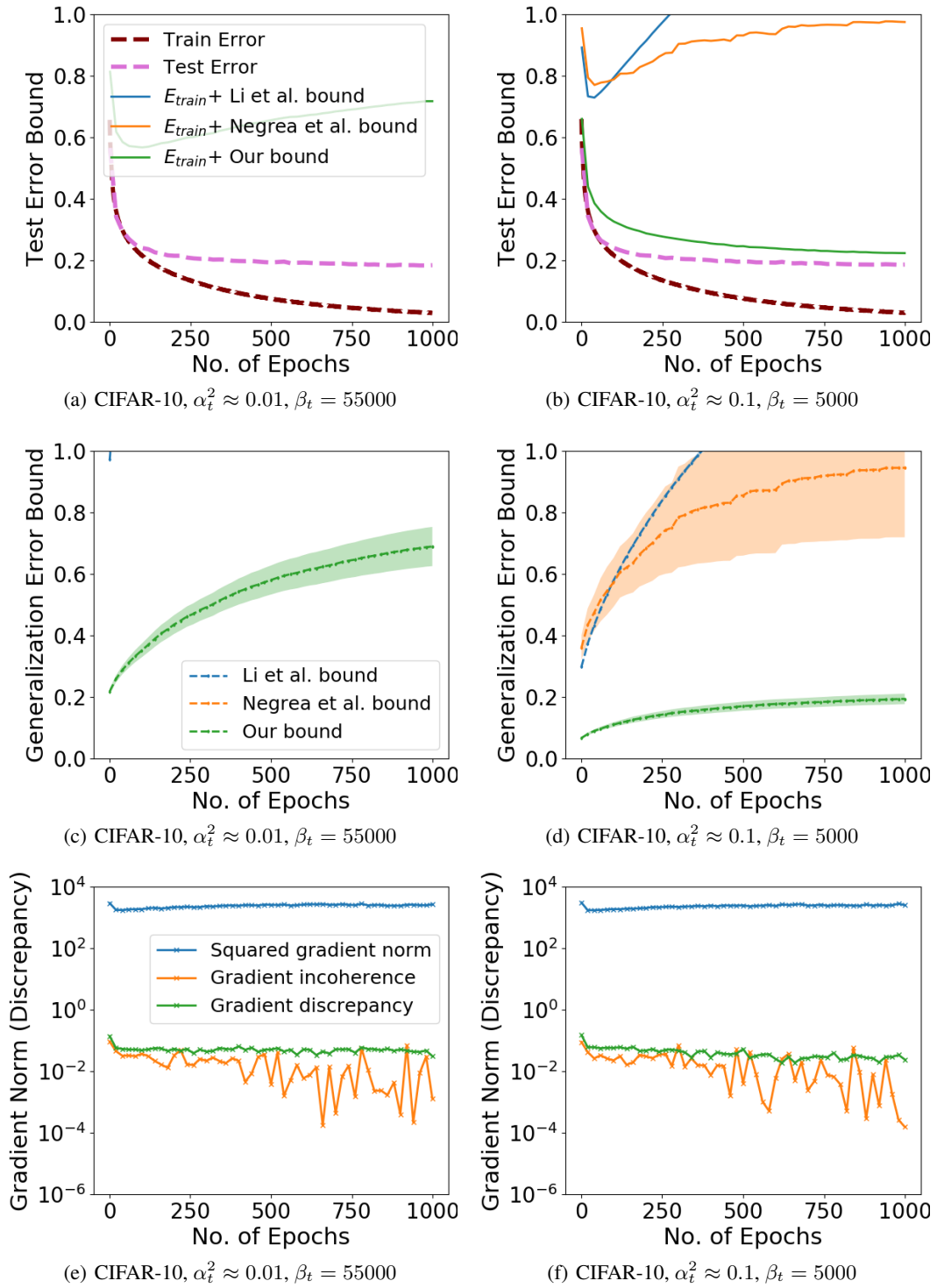(a) CIFAR-10, $\alpha_t^2 \approx 0.01$, $\beta_t = 55000$

(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 7. Numerical results for training CNN using SGLD on CIFAR-10.

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

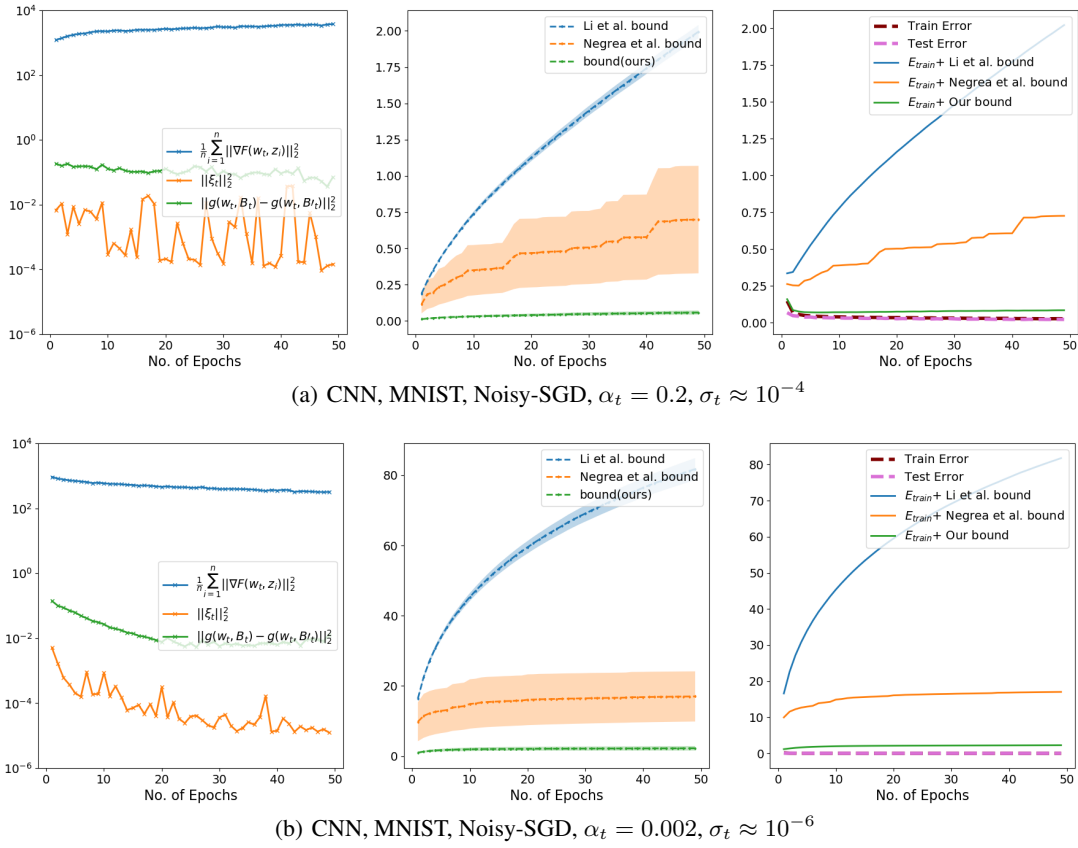(b) CNN, MNIST, Noisy-SGD, $\alpha_t = 0.002$, $\sigma_t \approx 10^{-6}$

Figure 8. Numerical results for CNN trained on MNIST using Noisy SGD. 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.

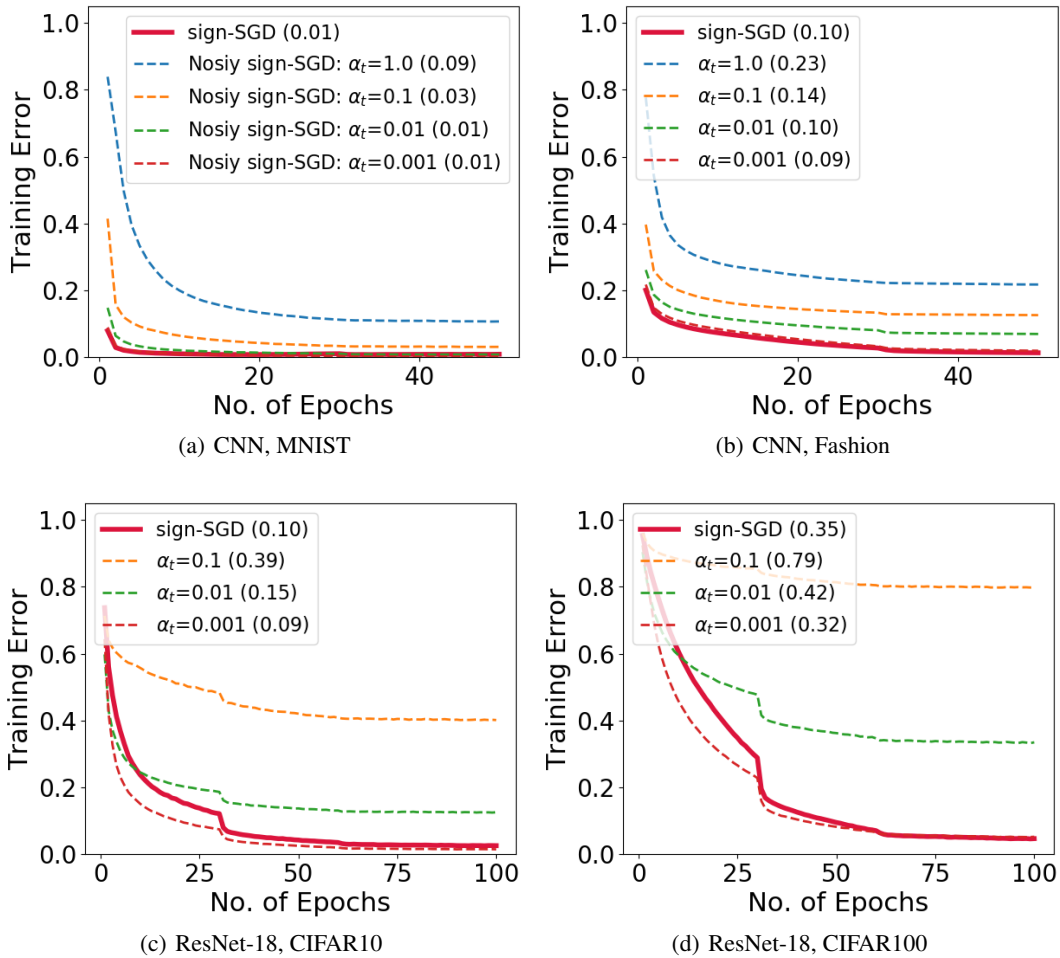*Figure 9.* Training dynamics of CNN on MNIST and Fashion-MNIST, and ResNet-18 on CIFAR-10 and CIFAR-100 using noisy sign-SGD with different scaling α_t. Legends indicate the choice of α_t and the numbers in brackets are test errors at convergence. As $\alpha_t \to 0$, Nosiy sign-SGD matches both the optimization trajectory as well as the final test accuracy of the original sign-SGD (Bernstein et al., 2018a).
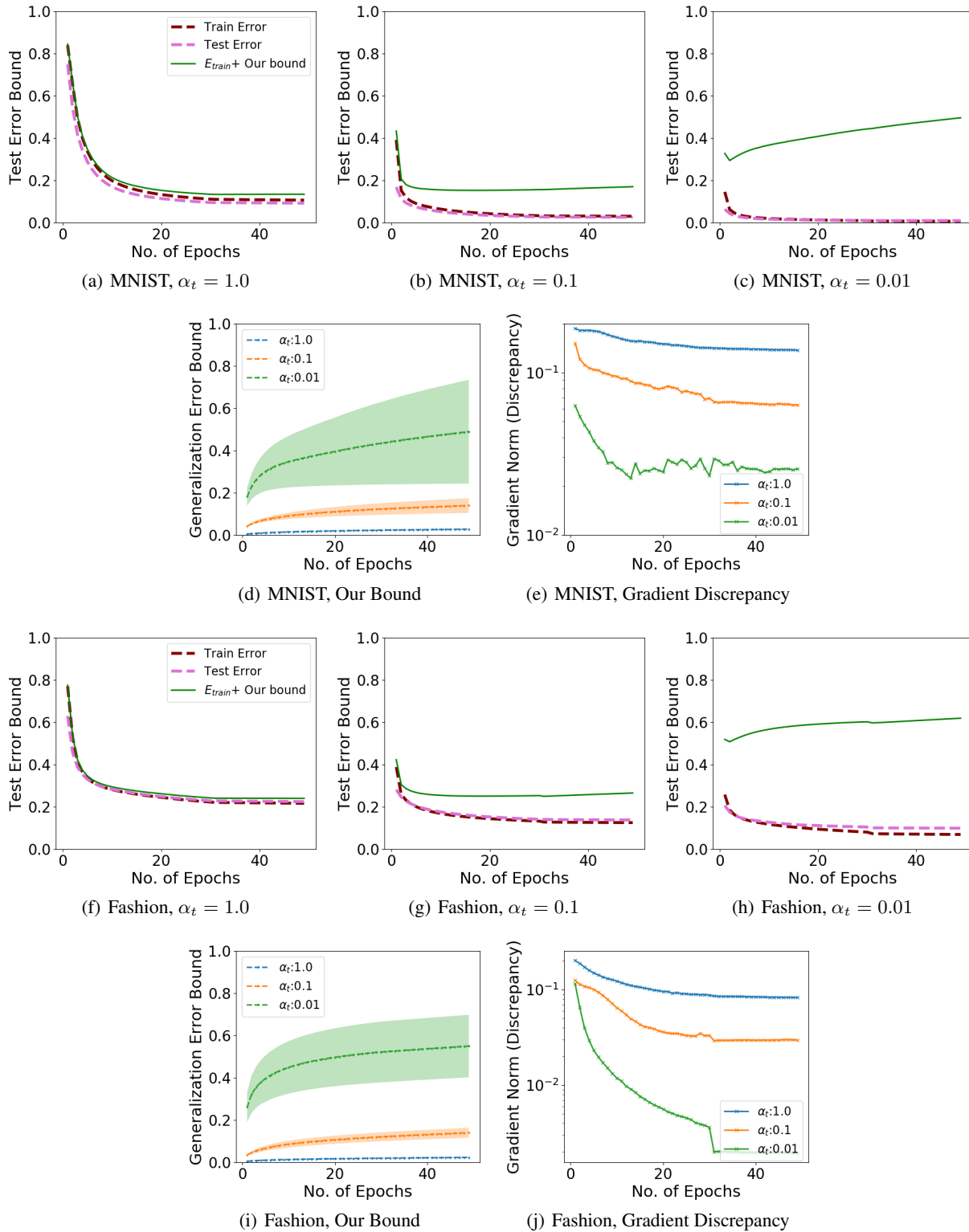
(a) MNIST, $\alpha_t = 1.0$

(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 10. 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.