Stability based Generalization Bound for Exponential Family Langevin Dynamics

Arindam Banerjee(UIUC) Tiancong Chen (UMN) Xinyan Li (UMN) Yingxue Zhou (UMN)

International Conference on Machine Learning (ICML) July, 2022

Problem Setting

Consider the statistical learning setting:

- i.i.d samples $S_n = \{z_1, \ldots, z_n\} \sim D^n$;
- A randomized algorithm A works with S_n creating a distribution over hypotheses: $A(S_n)$;
- For distribution P over hypothesis, expected population and empirical loss:

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

- The generalization error:

$$\text{gen}(A(S_n)) \triangleq L_D(A(S_n)) - L_S(A(S_n))$$

Noisy Iterative Algorithms

Given S_n and realization of past iterates $W_{0:t-1} = w_{0:t-1}$, a noisy iterative algorithm updates parameter by

$$W_t \sim P_{B_t,\xi_t|w_{0:(t-1)}}(W)$$

where the distribution has two sources of randomness:

1) Stochastic mini-batch of samples S_{B_t} with batch size b, drawn uniformly at random with replacement;
2) Noise ξ_t suitably added in the iterations.

Generalization Bound based on Le Cam Style Divergence

Under mild assumptions, for noisy iterative algorithm we have

$$|\mathbb{E}\mathrm{gen}(A(S_n))| \leq c\frac{b}{n}\mathop{\mathbb{E}}_{S_n, z_n'}\sqrt{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{W_{0:(t-1)}} LSD(P_{t|}|||P'_{t|})}$$

where Le Cam Style Divergence

$$LSD(P_{t|}|||P'_{t|}) := \mathbb{E}\int_{\xi_t}\frac{(dP_{B_t,\xi_t} - dP'_{B_t,\xi_t})^2}{dP_{A_t,\xi_t}}d\xi_t \ ,$$

measures the distance of output $P_{B_t,\xi_t}, P'_{B_t,\xi_t}$ run on S_n, S_n'.

- The bound is based on expected stability instead of uniform stability;
- The bound can be extended to high probability bound.

Exponential Family Langevin Dynamics

We propose EFLD using exponential family noise: for smooth convex function ψ

$$w_t = w_{t-1} - \rho_t \xi_t \ , \ \xi_t \sim p_\psi(\xi; \theta_{B_t,\alpha_t}),$$

where $p_\psi(\xi;\theta_{B_t,\alpha_t}) = \exp(\langle \xi, \theta_{B_t,\alpha_t}\rangle - \psi(\theta_{B_t,\alpha_t}))\pi_{0,\alpha}(\xi), \ \theta_{B_t,\alpha_t} \triangleq \frac{\theta_{B_t}}{\alpha_t} = \frac{\nabla \ell(w_{t-1}, S_{B_t})}{\alpha_t}$.

EFLD becomes 1) SGLD when exponential family is Gaussian

$$w_t = w_{t-1} - \eta_t \nabla \ell(w_{t-1}, S_{B_t}) + \mathcal{N}\left(0, \sigma_t^2 \mathbb{I}\right);$$

2) Noisy Sign-SGD when exponential family is Skewed Rademacher distribution

$$w_t = w_{t-1} - \eta_t \xi_t$$

where

$$\xi_{t,j} = \begin{cases} 1 & \text{with prob. } \frac{1}{2}(1 + \tanh(\theta_{B_t,\alpha_t,j})) \\ -1 & \text{with prob. } \frac{1}{2}(1 - \tanh(\theta_{B_t,\alpha_t,j})) \end{cases}$$

Generalization Bound for EFLD

With data dependent scale parameter $\alpha_{t|}$, under some mild assumptions:

$$|\mathbb{E}\mathrm{gen}(A(S_n))| \leq \frac{c}{n} \mathop{\mathbb{E}}_{S_n, z_n'} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{W_{0:(t-1)}} \frac{\|\nabla \ell\left(w_{t-1}, z_n\right) - \nabla \ell\left(w_{t-1}, z_n'\right)\|_2^2}{\alpha_{t|}^2}}.$$

Comparison to some previous work for SGLD:

- Gradient discrepancy < gradient norm in Li et al.(2020).
- Sample dependence $1/n$ is sharper than $1/\sqrt{n}$ in Negrea et al. (2019)

EFLD and its bound can be extended to anisotropic noise.

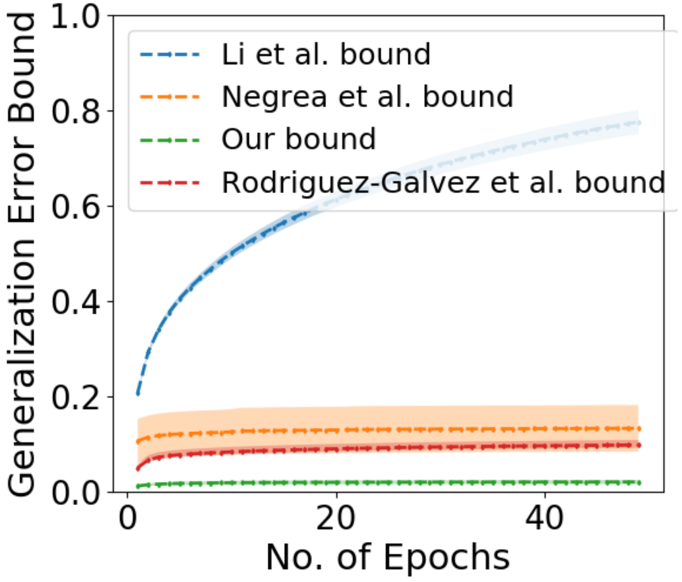
Optimization Guarantees for EFLD

We provide optimization guarantees for two variants of EFLD:

- For SGLD, result is similar to previous work [Bassily et al.,2014; Wang & Xu, 2019];
- For Noisy Sign-SGD, we provide novel optimization guarantee: under assumptions, the full/mini-batch noisy sign-SGD satisfies $O(1/\sqrt{T})$ convergence rate.

Comparison to Existing Bounds

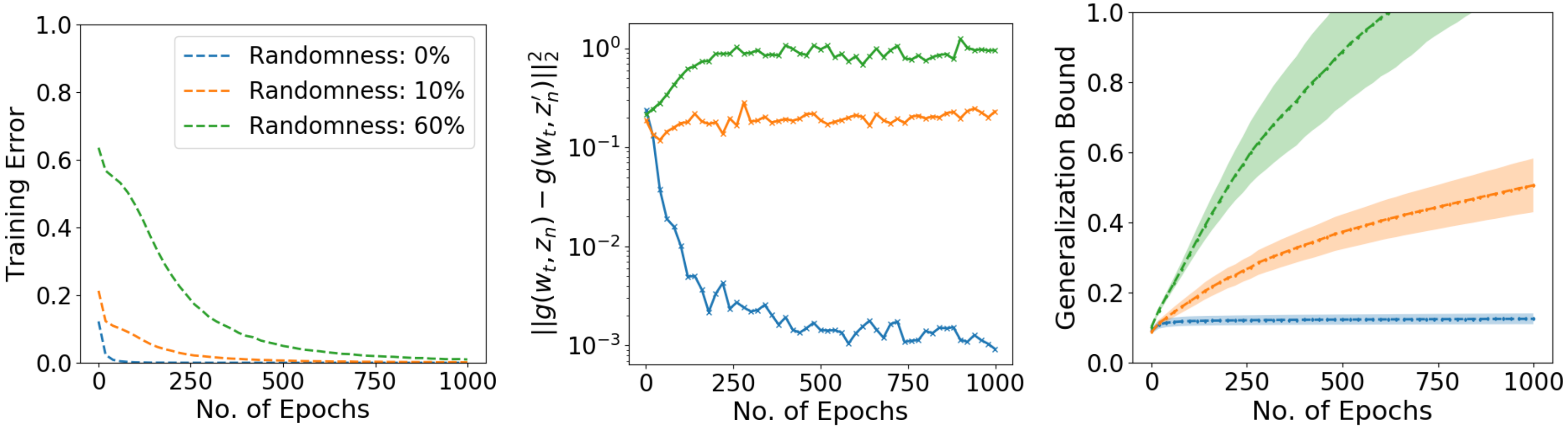
Our bound is sharper compared to existing bounds across dataset and settings:



MNIST, $\alpha_t^2 \approx 0.1$

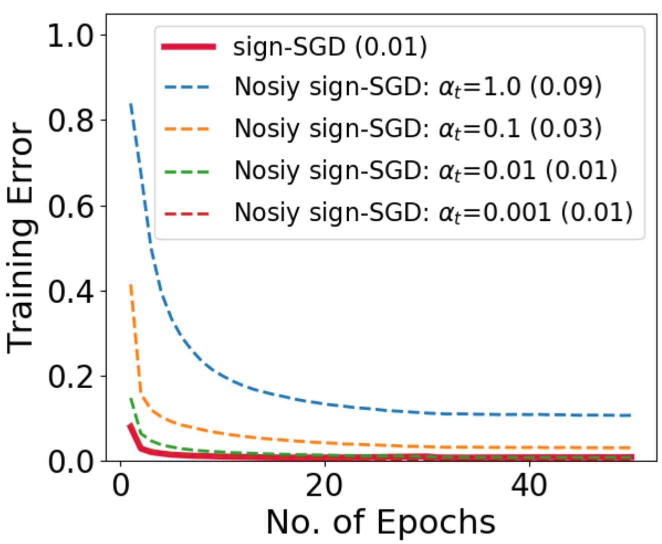
Random Label Experiment

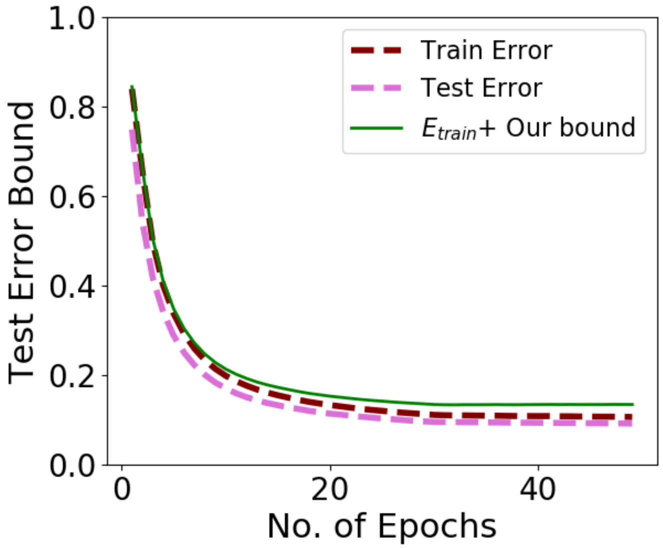
We consider the effect of random label motivated by Zhang et al. (2017). We show increase random labels ⇨ increase gradient discrepancy ⇨ increase bound.



Convergence and Generalization of Noisy Sign-SGD

Our result shows Noisy Sign-SGD matches performance of vanilla Sign-SGD when α_t is suitably small. And our bound successfully bounds the empirical test error.





Thanks for watching!

The research was supported by NSF grants IIS 21-31335, OAC 21-30835, DBI 20-21898, and a C3.ai research award. We would like to thank the reviewers for valuable comments and the Minnesota Supercomputing Institute (MSI) for computational resources and support.