Two Facets of SDE Under an Information-Theoretic Lens: Generalization of SGD via Training Trajectories and via Terminal States

Ziqiao Wang [1] Yongyi Mao [1]

[1]School of Electrical Engineering and Computer Science, University of Ottawa

uOttawa

Motivation

- Prevalent method of analyzing the generalization error of SGD via information-theoretic (IT) generalization bounds [Neu et al., 2021, Wang and Mao, 2022]:

$$\text{Gen. Err.(SGD)} = \text{Gen. Err.(SGD)} + \text{Gen. Err.(NGD)} - \text{Gen. Err.(NGD)}$$
$$\leq \text{ITBound(NGD)} + |\text{Gen. Err.(SGD)} - \text{Gen. Err.(NGD)}|,$$

  where NGD is some noisy (stochastic) gradient descent.
- Empirical evidences [Wu et al., 2020, Li et al., 2021] show that $|\text{Gen. Err.(SGD)} - \text{Gen. Err.(SDE)}|$ is small: let NGD=SDE!
- Steady-state estimation of SDE enable us to analyze its terminal state.

Background

- Learning algorithm $\mathcal{A}: S \to W$ i.e. mapping a training sample (with size n) to a hypothesis; $\text{Gen. Err.}(\mathcal{A}) = \mathbb{E}\left[\text{Test Err.} - \text{Train Err.}\right]$
- SGD: $w_t = w_{t-1} - \eta \widetilde{G}_t$, where η is step size and $\widetilde{G}_t$ is the mini-batch gradient with batch size b.
- SDE: $w_t = w_{t-1} - \eta G_t + \eta C_t^{1/2} N_t$, where G_t is the full-batch gradient, $N_t \sim \mathcal{N}(0, \mathrm{I}_d)$ and C_t is gradient noise covariance (GNC):

$$C_t \triangleq \frac{n-b}{b(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i \nabla \ell_i^{\mathrm{T}} - G_t G_t^{\mathrm{T}}\right)$$

- Information-theoretic generalization bounds:

  Lemma 1. For a subGaussian loss, $\text{Gen. Err.} \leq \mathcal{O}\left(\sqrt{\frac{I(W;S)}{n}}\right)$.

  Lemma 2. For a bounded loss, $\text{Gen. Err.} \leq \mathcal{O}\left(\sqrt{\mathrm{D_{KL}}\left(Q_{W|S} || P_{W|S_J}\right)}\right)$, where S_J is a random subset of S, $Q_{W|S}$ is the posterior induced by $\mathcal{A}$ and $P_{W|S_J}$ is a data-dependent prior.

Generalization Bounds Via Full Trajectories

Recall $I(X;Y) \leq \mathbb{E}_X\left[\mathrm{D_{KL}}\left(Q_{Y|X} || P_Y\right)\right]$, P_Y is some arbitrary prior.

- Using an isotropic Gaussian as prior, we have

  Theorem 1. Let $\Sigma_t^{\mu} \triangleq \mathbb{E}\left[\nabla \ell \nabla \ell^{\mathrm{T}}\right] - \mathbb{E}\left[\nabla \ell\right]\mathbb{E}\left[\nabla \ell\right]^{\mathrm{T}}$ be the population GNC. Assume $\Sigma_t^{\mu}, C_t \succ 0$,

$$\text{Gen. Err.} \precsim \sqrt{\frac{1}{n}\sum_{t=1}^{T} \mathbb{E}\left[d \log \frac{tr\{\Sigma_t^{\mu}\}}{bd} - \mathbb{E}\left[tr \log C_t\right]\right]}.$$

  Remark. $tr\{\Sigma_t^{\mu}\} = \mathbb{E}\left[||G_t - \mathbb{E}\left[\nabla \ell\right]||^2 + tr\left\{C_t\right\}\right] \Longrightarrow$
  - First term: the sensitivity of G_t to some variation of the training set S.
  - Second term: the gradient noise magnitude induced by mini-batch.
- By-product: recovering a bound for Gradient Langevin dynamics

  Corollary 1. If $C_t = \mathrm{I}_d$, then

$$\text{Gen. Err.} \precsim \sqrt{\frac{d}{n}\sum_{t=1}^{T} \mathbb{E} \log\left(\mathbb{E}\left[||G_t - \mathbb{E}\left[\nabla \ell\right]||^2\right]/d + 1\right)}.$$

  Remark. Not necessarily depends on d (by $\log(x+1) \leq x$).
- Using an anisotropic Gaussian as prior, we have

  Theorem 2. Under the same conditions in **Theorem 1.**,

$$\text{Gen. Err.} \precsim \sqrt{\sum_{t=1}^{T} \frac{\mathbb{E}\left[tr \log\left(\Sigma_t^{\mu} C_t^{-1}/b\right)\right]}{n}}.$$

  Remark. Theorem 2. is tighter than **Theorem 1.** Let $\Sigma_t = bC_t$, then $\Sigma_t^{\mu}\Sigma_t^{-1}$ is small $\iff$ SGD is insensitive to the randomness of S. Same intuition with $I(W;S)$ in **Lemma 1**.

Take-Home Messages

1. Trajectories-based bounds need less assumptions but are time-dependent.
2. Terminal-state-based bounds are time-independent but require additional assumptions and approximations.

Generalization Bounds Via Terminal State

Quadratic loss: $w \to$ local minimum w^*, let H_{w^*} be Hessian at w^*,

$$\text{Loss of } w = \text{Loss of } w^* + \frac{1}{2}(w - w^*)^{\mathrm{T}} H_{w^*}(w - w^*).$$

$\overset{T\to\infty}{\Longrightarrow}$ given a s and its local minimum w_s^*, $W_T \sim \mathcal{N}(w_s^*, \Lambda_{w_s^*})$.

$\overset{W_s^* \sim Q_{W_S^*|s}}{\Longrightarrow}$ $Q_{W_T|s} = \mathbb{E}^s_{W_s^*}\left[\mathcal{N}(W_s^*, \Lambda_{W_s^*})\right]$ is a mixture of Gaussian.

- **Lemma 3.** $\Lambda_{w^*} H_{w^*} + H_{w^*}\Lambda_{w^*} - \eta H_{w^*}\Lambda_{w^*}H_{w^*} = \eta C_T$.
- Hessian-based Result

  Theorem 3. Let $\Lambda_{w_\mu^*} \triangleq \mathbb{E}\left[\left(W - \mathbb{E}\left[W_S^*\right]\right)\left(W - \mathbb{E}\left[W_S^*\right]\right)^{\mathrm{T}}\right]$. Under some mild assumptions,

$$\text{Gen. Err.} \precsim \sqrt{\frac{1}{n}\mathbb{E}\left[tr\log\left(\left[H_{w^*}C_T^{-1}\right]\Lambda_{w_\mu^*}\right)\right]}.$$

  Remark. Alignment between a population and a sample stationary dist.
- Norm-based Result

  Theorem 4. Let $\hat{w}$ be a reference vector. Under some mild assumptions,

$$\text{Gen. Err.} \precsim \sqrt{\frac{d}{n}\log\left(\frac{b}{\eta d}\mathbb{E}||W_S^* - \hat{w}||^2 + 1\right)}.$$

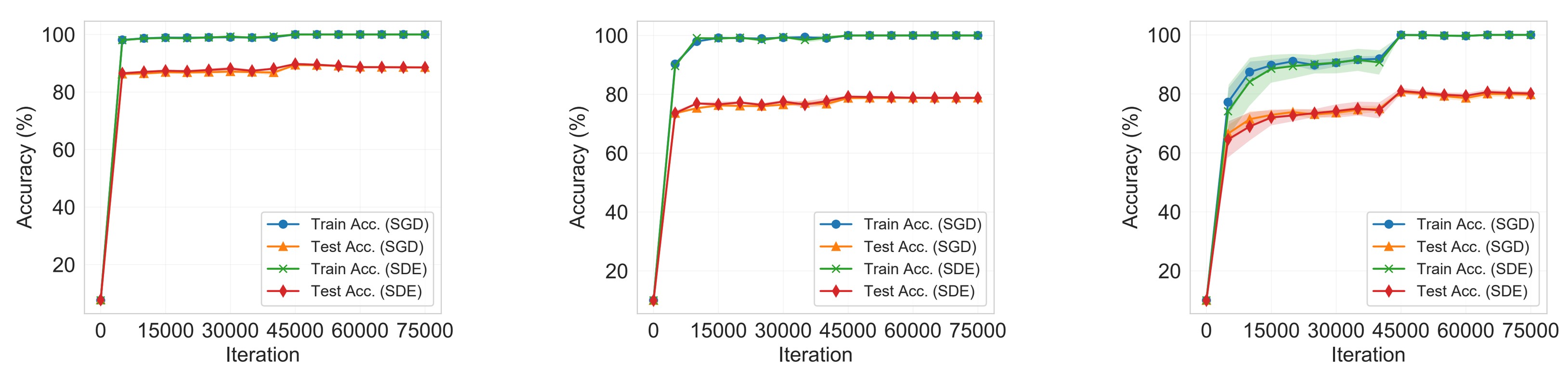
  Remark. i) $\hat{w} = \mathbb{E}\left[W_S^*\right] \Rightarrow$ Optimal; ii) $\hat{w} = w_0 \Rightarrow$ "Distance from initialization"; iii) $\hat{w} = 0 \Rightarrow$ Weight Decay.
- Stability-based Result

  Theorem 5. Recall **Lemma 2.** and under some mild assumptions,

$$\text{Gen. Err.} \precsim \sqrt{\frac{b}{\eta}\mathbb{E}||W_S^* - W_{S_J}^*||^2}.$$

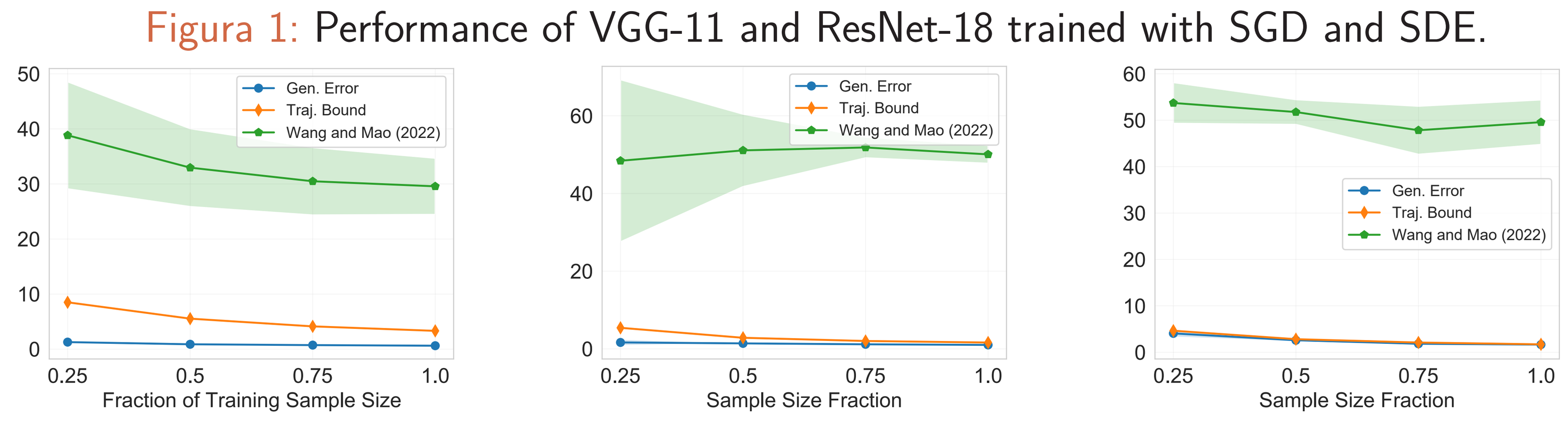
  Remark. No Lipschitz constant contained; Fast-rate in some cases.

Empirical Results



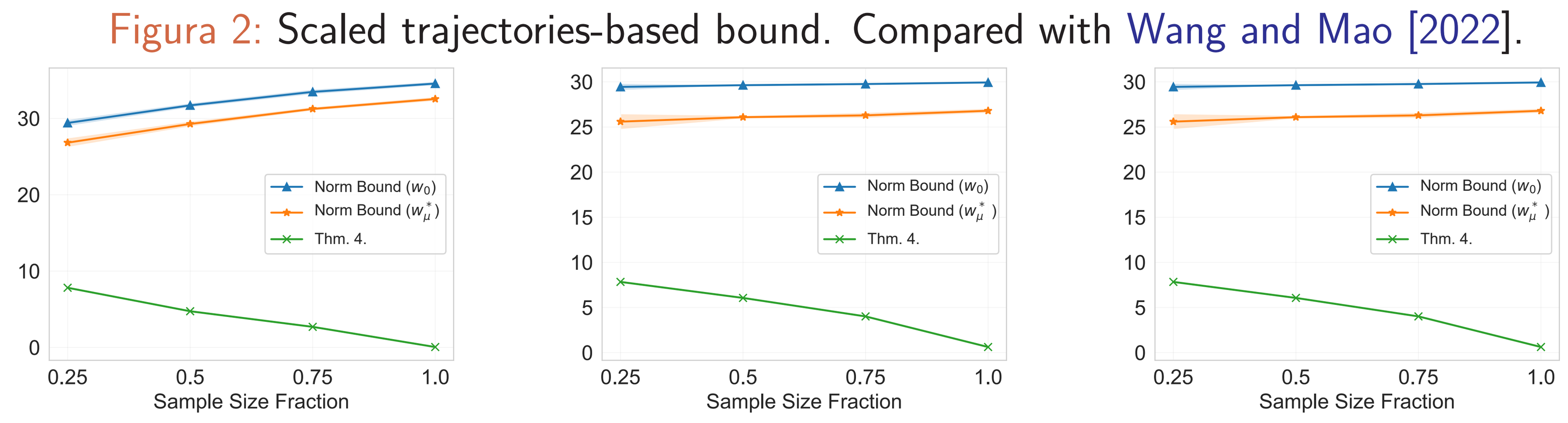
(a) VGG on (small) SVHN (b) VGG on CIFAR10 (c) ResNet on CIFAR10

Figura 1: Performance of VGG-11 and ResNet-18 trained with SGD and SDE.



(a) VGG on (small) SVHN (b) VGG on CIFAR10 (c) ResNet on CIFAR10

Figura 2: Scaled trajectories-based bound. Compared with Wang and Mao [2022].



(a) VGG on (small) SVHN (b) VGG on CIFAR10 (c) ResNet on CIFAR10

Figura 3: Scaled terminal-state based bound.

Reference

Zhiyuan Li et al. On the validity of modeling sgd with stochastic differential equations (sdes). *NeurIPS*, 2021.

Gergely Neu et al. Information-theoretic generalization bounds for stochastic gradient descent. In *COLT*, 2021.

Ziqiao Wang and Yongyi Mao. On the generalization of models trained with SGD: Information-theoretic bounds and implications. In *ICLR*, 2022.

Jingfeng Wu et al. On the noisy gradient descent that generalizes as sgd. In *ICML*, 2020.

ziqiaowanggeothe.github.io zwang286@uottawa.ca