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

Motivation

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

  \[
  \text{Gen. Err. (SGD)} = \text{Gen. Err. (NGD)} + \text{Gen. Err. (NDG)} \leq \text{ITBound (NGD)} + \text{Gen. Err. (SGD)} \]

  where (NGD) is some noisy (stochastic) gradient descent.

• Empirical evidences [Wu et al., 2020, Li et al., 2021] show that [Gen. Err. (SGD) — Gen. Err. (SDE)] is small: let NGD=SDE!

• Steady-state estimation of SDE enable us to analyze its terminal state.

Background

• Learning algorithm \( A : S \to W \) i.e. mapping a training sample (with size \( n \)) to a hypothesis; Gen. Err. \( A \equiv \text{[Test Err.] - [Train Err.]} \)

• SGD: \( w_t = w_{t-1} - \eta G_t \), where \( \eta \) is step size and \( G_t \) is the mini-batch gradient with batch size \( b \).

• SDE: \( w_t = w_{t-1} - \eta G_t + \eta \sqrt{1/2} N_t \), where \( G_t \) is the full-batch gradient, \( N_t \sim N(0, I_d) \) and \( G_t \) is gradient noise (covariance):

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

  Information-theoretic generalization bounds:

  \[
  \text{Lemma 1. For a subGaussian loss, Gen. Err.} \leq \mathcal{O} \left( \sqrt{\frac{n||W||^2}{d}} \right)
  \]

  where \( S \) is a random subset of \( S \), \( Q_{WS} \) is the posterior induced by \( A \) and \( P_{WS} \) is a data-dependent prior.

Generalization Bounds Via Terminal State

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

• Using an isotropic Gaussian as prior, we have \text{Theorem 1.} Let \( \Sigma_0^t \equiv \text{[\nabla \ell Cycl]} - \mathbb{E} \left[ \nabla \ell \right] \left[ \nabla \ell \right]^T \) be the population GNC. Assume \( \Sigma_0^t \succ 0 \),

  \[
  \text{Gen. Err.} \lesssim \sqrt{\frac{1}{n} \sum_{t=1}^{T} \mathbb{E} \left[ d \log \left( \text{tr}(\Sigma_0^t) \right) \right]} - \mathbb{E} \left[ \text{tr log} \left( C_t \right) \right]
  \]

  \text{Remark.} \text{tr}(\Sigma_0^t) = \mathbb{E} \left[ ||G_t - \mathbb{E} || \left[ \nabla \ell \right] ||^2 + \text{tr} \left[ C_t \right] \right) \implies

  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

  \text{Corollary 1. If} \ C_t = \frac{b}{d}, \text{then}

  \[
  \text{Gen. Err.} \lesssim \frac{1}{n} \sum_{t=1}^{T} \mathbb{E} \left( \text{log} \left( \frac{\mathbb{E} ||G_t - \mathbb{E} ||}{d + 1} \right) \right).
  \]

  \text{Remark. Not necessarily depends on} \ d \ \text{by log}(x + 1) \leq x.

  Using an anisotropic Gaussian as prior, we have \text{Theorem 2.} Under the same conditions in \text{Theorem 1.},

  \[
  \text{Gen. Err.} \lesssim \frac{1}{n} \sum_{t=1}^{T} \mathbb{E} \left( \text{log} \left( \frac{\text{tr log}(\Sigma_0^t)}{b} \right) \right).
  \]

  \text{Remark. Theorem 2. is tighter than Theorem 1.}

  Let \( \Sigma_0^t = bG_t \), then \( \Sigma_0^t \Sigma_0^{-1} \) is small \iff SGD is insensitive to the randomness of \( S \). Same intuition with \( I(W; S) \) in Lemma 1.

Empirical Results

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

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

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

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

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

Figure 3: Scaled terminal-state based bound.

Reference


