What Happens after SGD Reaches Zero Loss? – A Mathematical Framework

Zhiyuan Li† Tianhao Wang‡ Sanjeev Arora†
†Princeton University ‡Yale University

Introduction

Stochastic gradient descent (SGD) is widely used in training of modern machine learning models such as deep neural networks, and the implicit bias of SGD underlies the generalization ability of the trained models. While it still remains unclear how to mathematically characterize such bias.

Formulation of SGD: Given training loss $L : \mathbb{R}^D \to \mathbb{R}$,

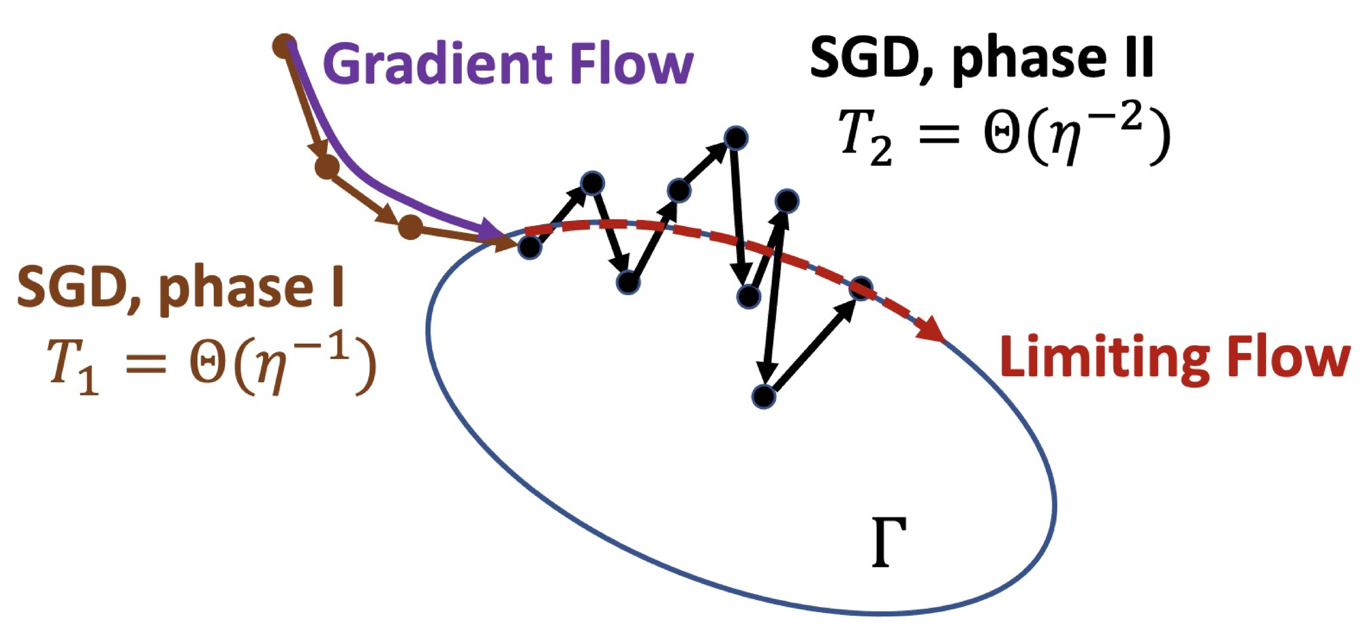
$$x_\eta(k+1) = x_\eta(k) - \eta(\nabla L(x_\eta(k)) + \sqrt{\Xi}\sigma_{\xi_k}(x_\eta(k))) \quad (1)$$

- η is the learning rate (LR)
- $\sigma(x) = [\sigma_1(x), \sigma_2(x), \ldots, \sigma_\Xi(x)] \in \mathbb{R}^{D\times\Xi}$ is the noise function
- ξ_k is sampled uniformly from $\{1, 2, \ldots, \Xi\}$ and $\mathbb{E}_{\xi_k}[\sigma_{\xi_k}(x)] = 0$

Main Contributions of This Work:

#1. A mathematical framework to study implicit bias of SGD with small LR

#2. Provable generalization benefit of stochasticity: Minimax optimal rate for learning sparse quadratically overparametrized linear models.



Problem Setting

Manifold of Local Minimizers: Γ is a $(D-M)$-dimensional submanifold of $\mathbb{R}^D$ such that for all $x \in \Gamma$, x is a local minimizer of L and $\mathrm{rank}(\nabla^2 L(x)) = M$.

When Does Such A Manifold Exist? Overparametrization!

Canonical SDE Approximation of SGD:

$$\mathrm{d}\widetilde{X}_\eta(t) = -\eta\nabla L(\widetilde{X}_\eta(t))\mathrm{d}t + \eta\cdot\sigma(\widetilde{X}_\eta(t))\mathrm{d}W(t) \quad (2)$$

where $\{W(t)\}_{t\geq 0}$ is a Ξ-dimensional Wiener Process and $\Sigma(x) = \sigma(x)\sigma(x)^\top$ is the covariance matrix of gradient noise.

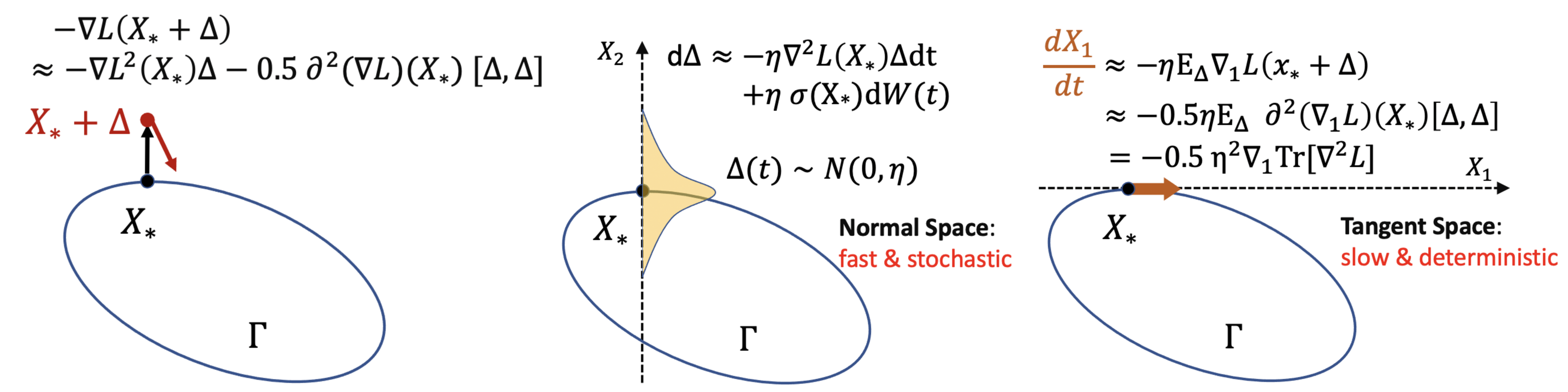
Intuitive Explanation of The Implicit Bias

Blanc et. al, (2020) showed that around some manifold of local minimizers SGD decreases $\mathrm{tr}[\nabla^2 L]$ if gradient covariance is equal to trace of hessian, $\Sigma \equiv \nabla^2 L$, on the manifold. (e.g., SGD with label noise)

Taylor Expansion Around A Local Minimizer: Let $\Delta(t) = \widetilde{X}_\eta(t) - X^*$,

$$\mathrm{d}\Delta(t) \approx -\eta\nabla^2 L(X^*)\Delta(t)\mathrm{d}t + \eta\sigma(X^*)\mathrm{d}W(t)$$

which behaves like an Ornstein-Uhlenbeck process in the normal space



The fast dynamics in the normal space activates the second order Taylor expansion in the tangent space, creating a $\Theta(\eta^2)$ velocity, which is slow but deterministic and acumulates over time.

Issue: The above local analysis only holds for $O(\eta^{-1.6})$ time. How to do global analysis?

Our Approach: Separating Slow from Fast

Time Rescaling for SDE: Let $X_\eta(t) = \widetilde{X}_\eta(t/\eta^2)$, then

$$\mathrm{d}X_\eta(t) = \underbrace{-\eta^{-1}\nabla L(X_\eta(t))\mathrm{d}t}_{\text{Fast}} + \underbrace{\sigma(X_\eta(t))\mathrm{d}W(t)}_{\text{Slow}}$$

As $\eta \to 0$, the *Fast* part rapidly drives $X_\eta(t)$ towards Γ via the projection induced by the gradient flow, denoted by $\Phi(X_\eta(t))$.

Lemma 1. $\partial\Phi(x)\nabla L(x) \equiv 0$.

Applying Lemma 1 and Ito's lemma, we get

$$\mathrm{d}\Phi(X_\eta(t)) = \partial\Phi(X_\eta)\sigma(X_\eta)\mathrm{d}W(t) + \frac{1}{2}\partial^2\Phi(X_\eta)[\sigma(X_\eta)\sigma(X_\eta)^\top]\mathrm{d}t.$$

Since $\Phi(X_\eta(t)) \approx X_\eta(t)$ near Γ, the *Slow* part survives:

$$\mathrm{d}X_\eta(t) \approx \underbrace{\partial\Phi(X_\eta)\sigma(X_\eta)\mathrm{d}W(t)}_{\text{Tangent noise}} + \frac{1}{2}\underbrace{\partial^2\Phi(X_\eta)[\sigma(X_\eta)\sigma(X_\eta)^\top]\mathrm{d}t}_{\text{Compensation and regularization}}.$$

The above analysis can be made rigorous and extended to SGD by viewing SGD as an asymptotically continuous stochastic process and further applying the classic results by Katzenberger (1991).

Main Results

Lemma 2. $\partial\Phi(x)$ is the projection matrix of the tangent space of Γ at x.

Notation: . Define $\Sigma_\parallel(x) = \partial\Phi(x)\Sigma(x)\Phi(x)$ (noise covariance in the tangent space), $\Sigma_\perp(x) = (I - \partial\Phi(x))\Sigma(x)(I - \partial\Phi(x))$ (noise covariance in the normal space), and $\Sigma_{\parallel,\perp} = \Sigma_{\perp,\parallel}^\top = \partial\Phi(x)\Sigma(x)(I - \partial\Phi(x))$ (covariance across the tangent and normal space).

Lyapunov Operator: For a symmetric matrix H, define $W_H = \{\Sigma \mid \Sigma = \Sigma^\top, HH^\dagger\Sigma = \Sigma = \Sigma HH^\dagger\}$. The *Lyapunov operator* $\mathcal{L}_H : W_H \to W_H$ is defined as $\mathcal{L}_H(\Sigma) = H\Sigma + \Sigma H$.

Main Theorem. For SGD (1) and any $T > 0$, $x_\eta(\lfloor T/\eta^2\rfloor)$ converges in distribution to $Y(T)$ as $\eta \to 0$, where $Y(T)$ is the solution to the following SDE at time T when the global solution exists:

$$\mathrm{d}Y(t) = \underbrace{\Sigma_\parallel^{1/2}(Y)\mathrm{d}W(t)}_{\text{Tangent noise}} - \underbrace{\frac{1}{2}\nabla^2 L(Y)^\dagger\partial^2(\nabla L)(Y)[\Sigma_\parallel(Y)]\mathrm{d}t}_{\text{Tangent noise compensation}}$$
$$-\underbrace{\frac{1}{2}\partial\Phi(Y)\partial^2(\nabla L)(Y)[\nabla^2 L(Y)^\dagger\Sigma_{\perp,\parallel}(Y)]\mathrm{d}t}_{\text{Mixed regularization}}$$
$$-\underbrace{\frac{1}{2}\partial\Phi(Y)\partial^2(\nabla L)(Y)[\mathcal{L}_{\nabla^2 L}^{-1}(\Sigma_\perp(Y))]\mathrm{d}t}_{\text{Normal regularization}}.$$

Provable Generalization Benefit of Stochasticity

Setting: Data $\{(z_i, y_i)\}_{i=1}^n$ where $z_1, \ldots, z_n \overset{i.i.d}{\sim} \mathrm{Unif}(\{\pm 1\}^d)$ or $\mathcal{N}(0, I_d)$ and each $y_i = \langle z_i, w^*\rangle$ for some unknown κ-sparse $w^* \in \mathbb{R}^d$. Denote $x = \binom{u}{v} \in \mathbb{R}^D = \mathbb{R}^{2d}$. For each $i \in [n]$, define $f_i(x) = \langle z_i, u^{\odot 2} - v^{\odot 2}\rangle$. Consider the ℓ_2 loss $L(x) = \frac{1}{n}\sum_{i=1}^n (f_i(x) - y_i)^2$.

Label Noise SGD: At iteration k, replace the true label y_{i_k} by a perturbed label $y_{i_k} + \delta_k$ where $\delta_k \sim \mathrm{Unif}(\{\pm 1\})$ and run SGD on the perturbed label.

Regularizer: $R(x) = \mathrm{tr}[\nabla^2 L(x)] = \frac{4}{n}\sum_{j=1}^D \left(\sum_{i=1}^n z_{i,j}^2\right)(u_j^2 + v_j^2)$.

Limiting Dynamics = Riemannian gradient flow of R on Γ:

$$\mathrm{d}x(t) = -\partial\Phi(x(t))\nabla R(x(t))\mathrm{d}t.$$

Optimal Sparse Recovery $\Longleftarrow$ Constrained minimization of R on Γ.

Theorem. Under the above setting with $n \geq \Omega(\kappa\ln d)$ data, for any generic initialization x_0 and any $\epsilon > 0$, there exist $\eta_0, T > 0$ such that for any $\eta < \eta_0$, label noise SGD with LR η returns an ϵ-optimal solution in $\lfloor T/\eta^2\rfloor$ steps with high probability.