What Happens After SGD Reaches Zero Loss? --A Mathematical Framework

Zhiyuan Li
Princeton University

Tianhao Wang
Yale University

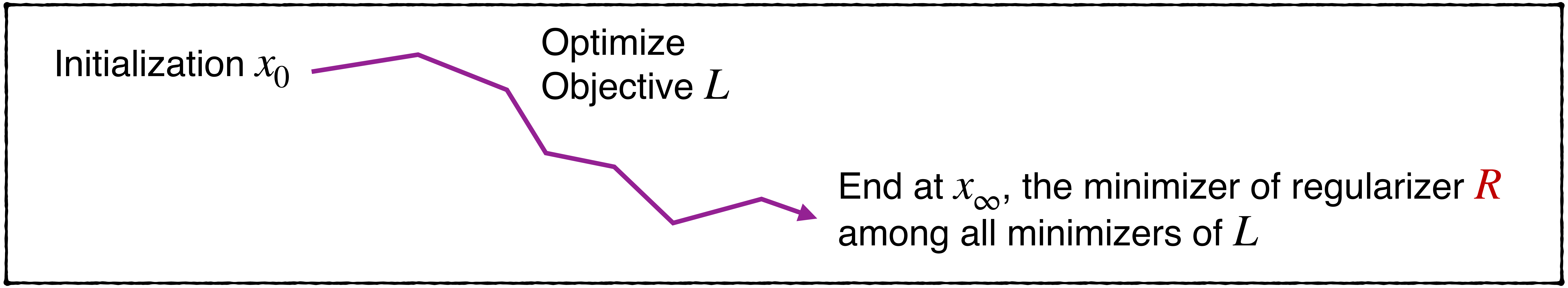
Sanjeev Arora
Princeton Univeristy

ICLR, 2022

Background

- Modern deep nets are vastly over-parametrized: able to fit random labels. (Zhang et al.,2017)
- Yet they perform well on proper labels $\Longrightarrow$generalization bound based on uniform convergence fails.

- An alternative explanation: Implicit regularization of training algorithm



- **Linear Model**: GD on $L(x) = \|Ax - b\|_2^2 \Longrightarrow R(x) = \|x - x_0\|_2^2$ (Including nets in NTK regime.)

Implicit Regularization for Non-linear Model

A brief survey:

- Matrix Factorization:
  Gunasekar et al., 2017; Du et al., 2018; Li et al., 2018; Arora et al., 2019; Gidel et al., 2019; Mulayoff & Michaeli, 2020; Blanc et al., 2020; Gissin et al., 2020; Razin & Cohen, 2020; Chou et al., 2020; Eftekhari & Zygalakis, 2021; Yun et al., 2021; Min et al., 2021; Li et al., 2021a; Razin et al., 2021; Milanesi et al., 2021; Ge et al., 2021
- Polynomially Overparametrized Linear Models with a Single Output:
  Ji & Telgarsky, 2019a; Woodworth et al., 2020; Moroshko et al., 2020; Azulay et al., 2021; Vardi et al., 2021
- Shallow Nonlinear Neural Nets:
  Vardi & Shamir, 2021; Hu et al., 2020; Sarussi et al., 2021; Mulayoff et al., 2021; Lyu et al., 2021

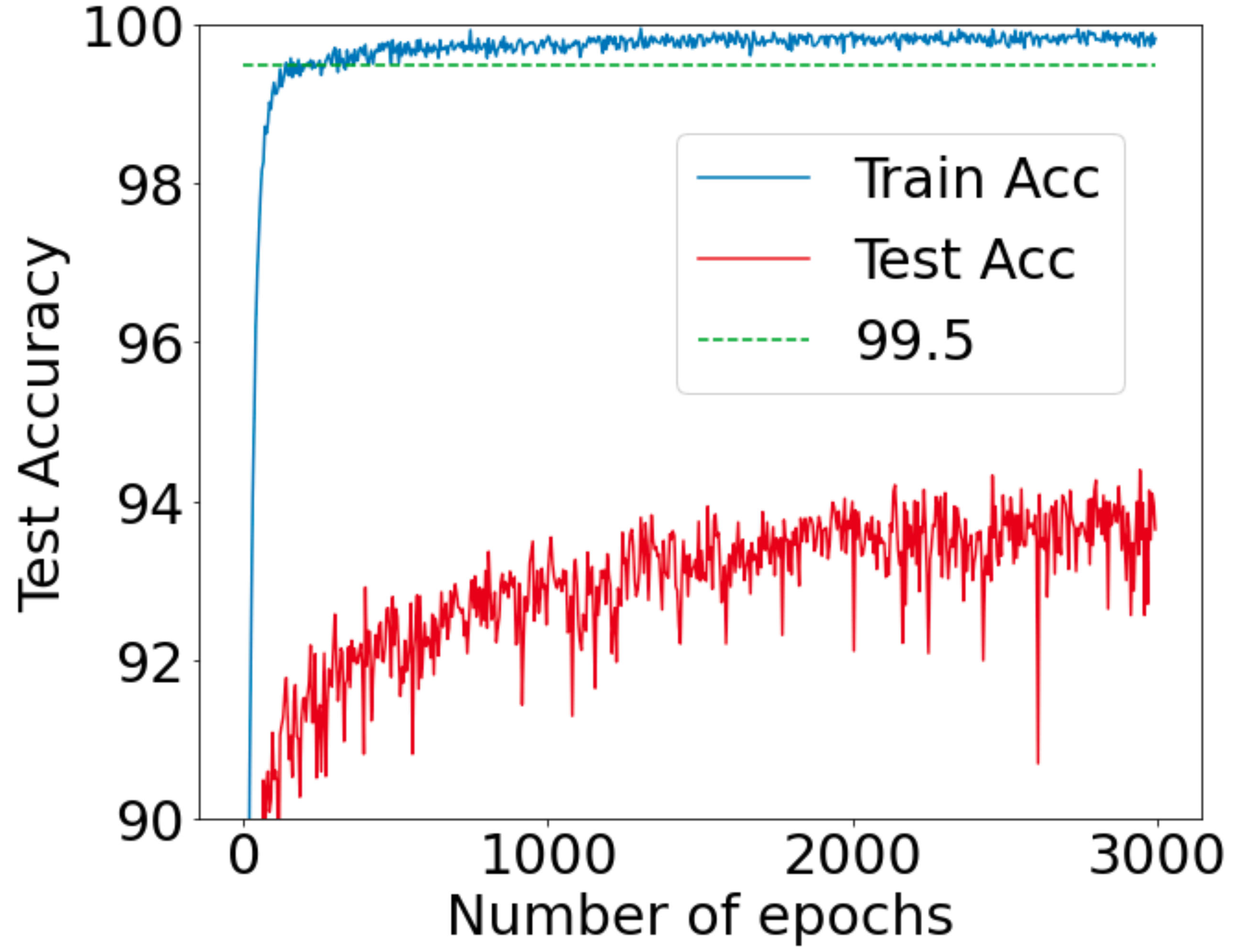
All above are essentially for deterministic GD. Cannot explain generalization benefit of Stochasticity.

Question:

What is the role of stochastic gradient noise in implicit regularization?

- Popular Belief:
  - Larger noise/LR →Flatter minima→Better generalization.

- Experimental Observation [**Li**, Lyu & **Arora**, 20]:
  - Small LR generalizes equally well, if trained longer.



ResNet trained on CIFAR10 with small LR

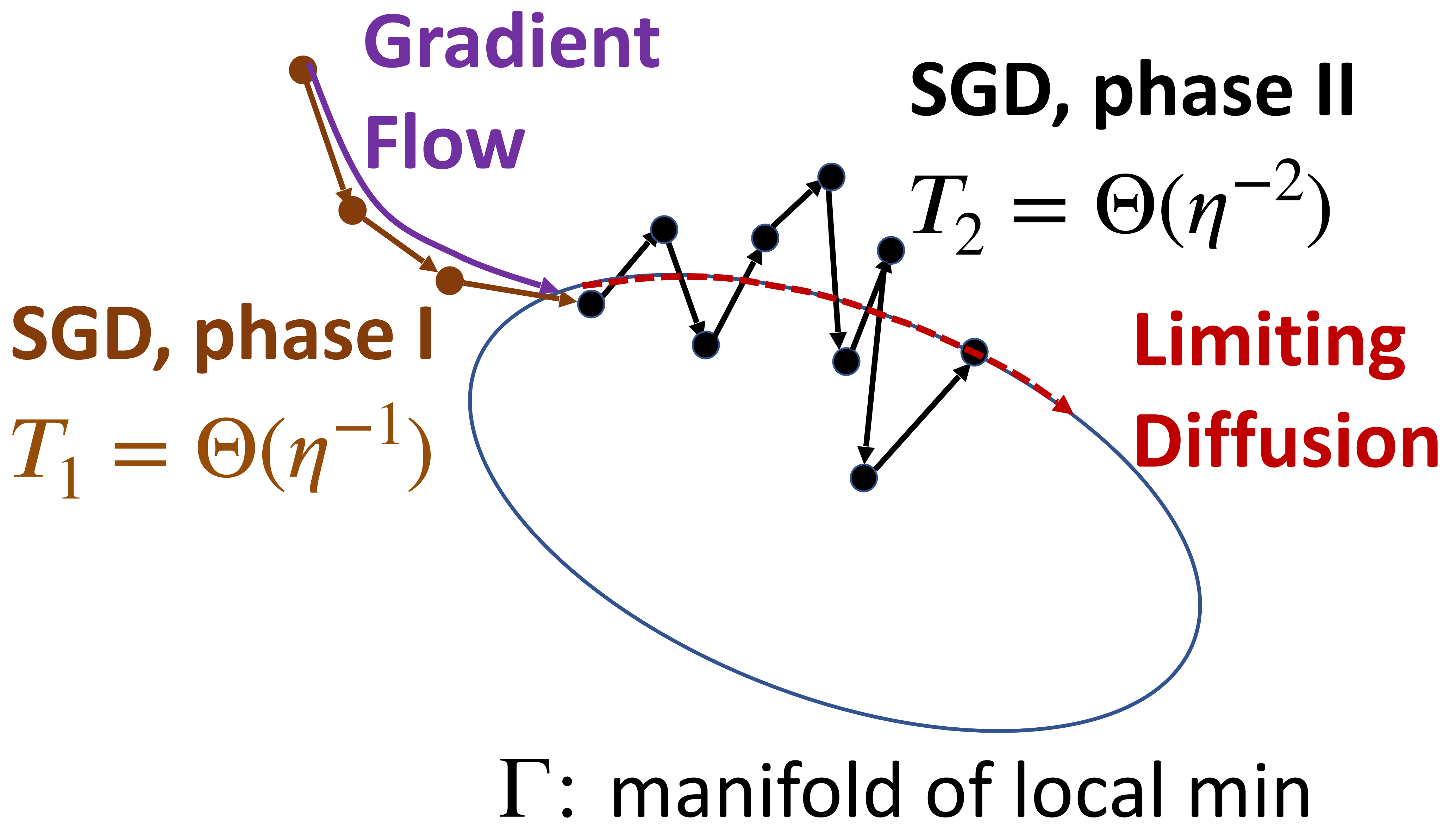
This paper: A complete* characterization for the regularization effect of SGD (with small LR) around manifold of minimizers, using Stochastic Differential Equation (SDE).

*: complete = any position-dependent noise with bounded covariance $\Sigma(x)$, improves over [Blanc et al,19], [Damian'21]

Li, Zhiyuan, Kaifeng Lyu, and Sanjeev Arora. "Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate." *NeurIPS,20*
Blanc, Guy, Neha Gupta, Gregory Valiant, and Paul Valiant. "Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process." *COLT'20*.
Damian, Alex, Tengyu Ma, and Jason Lee. "Label Noise SGD Provably Prefers Flat Global Minimizers." NeurIPS, 21

Main Result

Thm: When $\eta \to 0$, SGD on loss $L(x)$ has two phases:
1. **Gradient Flow phase** ($\Theta(1/\eta)$ steps): $x_{\frac{T}{\eta}} \to$ Gradient Flow solution at time T;
2. **Limiting Diffusion phase**($\Theta(1/\eta^2)$ steps): $x_{\frac{T}{\eta^2}} \to Y_T$, where $Y_t \in \Gamma$ is the solution of some SDE related to $\nabla^2 L, \nabla^3 L$ and covariance of gradient noise Σ.



Implications of Main Result

General Form of SDE on manifold: $dY_t/dt =$ diffusion term - drift term

- $\Sigma \equiv I_D$ on manifold, e.g., isotropic gaussian noise.
  - Diffusion term = White Noise in Tangent space;
  - Drift term = riemannian gradient of log of pseudo-determinant of $\nabla^2 L(X_t)$;

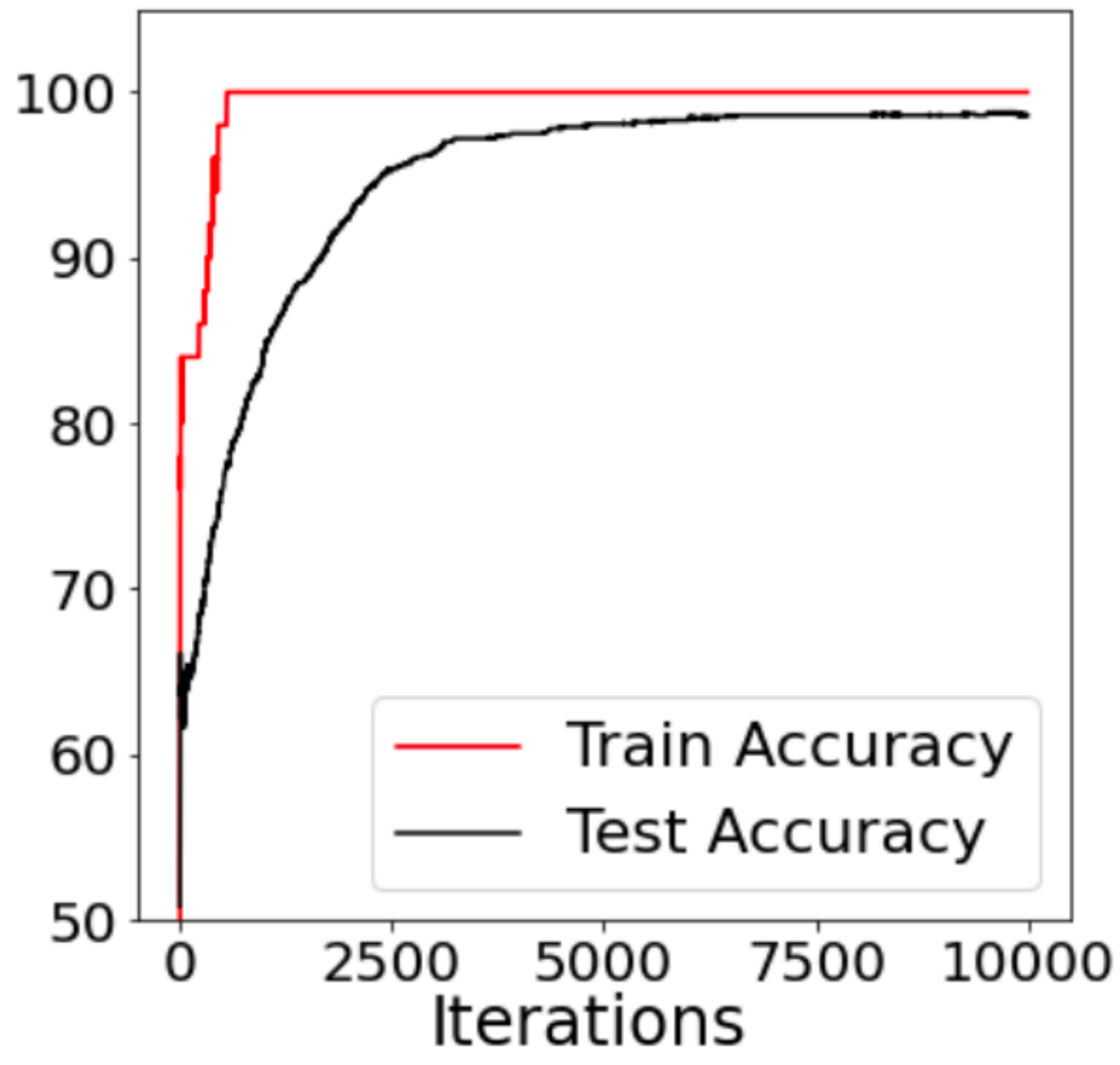
- $\Sigma \equiv \nabla^2 L$ on manifold, e.g., Label Noise ($x_{t+1} = x_t - \eta \nabla_x (f_{z_{i_t}}(x_t) - y_{i_t} - \delta_{i_t})^2$, where $\delta_{i_t} \overset{iid}{\sim}$ Unif$\{-\delta, \delta\}$)
  - No Diffusion term
  - Drift term = riemannian gradient of tr[$\nabla^2 L(X_t)$];

Blanc, Guy, Neha Gupta, Gregory Valiant, and Paul Valiant. "Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process." *COLT'20*.
Damian, Alex, Tengyu Ma, and Jason Lee. "Label Noise SGD Provably Prefers Flat Global Minimizers." NeurIPS, 21

Provable Generalization Benefit of SGD in Two-layer Net

Thm: *Two-layer diagonal network* + label noise SGD (any initialization)
is statistically optimal for learning sparse linear function.

k-sparse linear function in $\mathbb{R}^d$,
$O(k \ln d)$ samples.



large init = NTK regime and needs $O(d)$ samples.
SGD escapes NTK regime after reaching manifold.

Woodworth, Blake, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. "Kernel and rich regimes in overparametrized models."COLT'20

Future directions

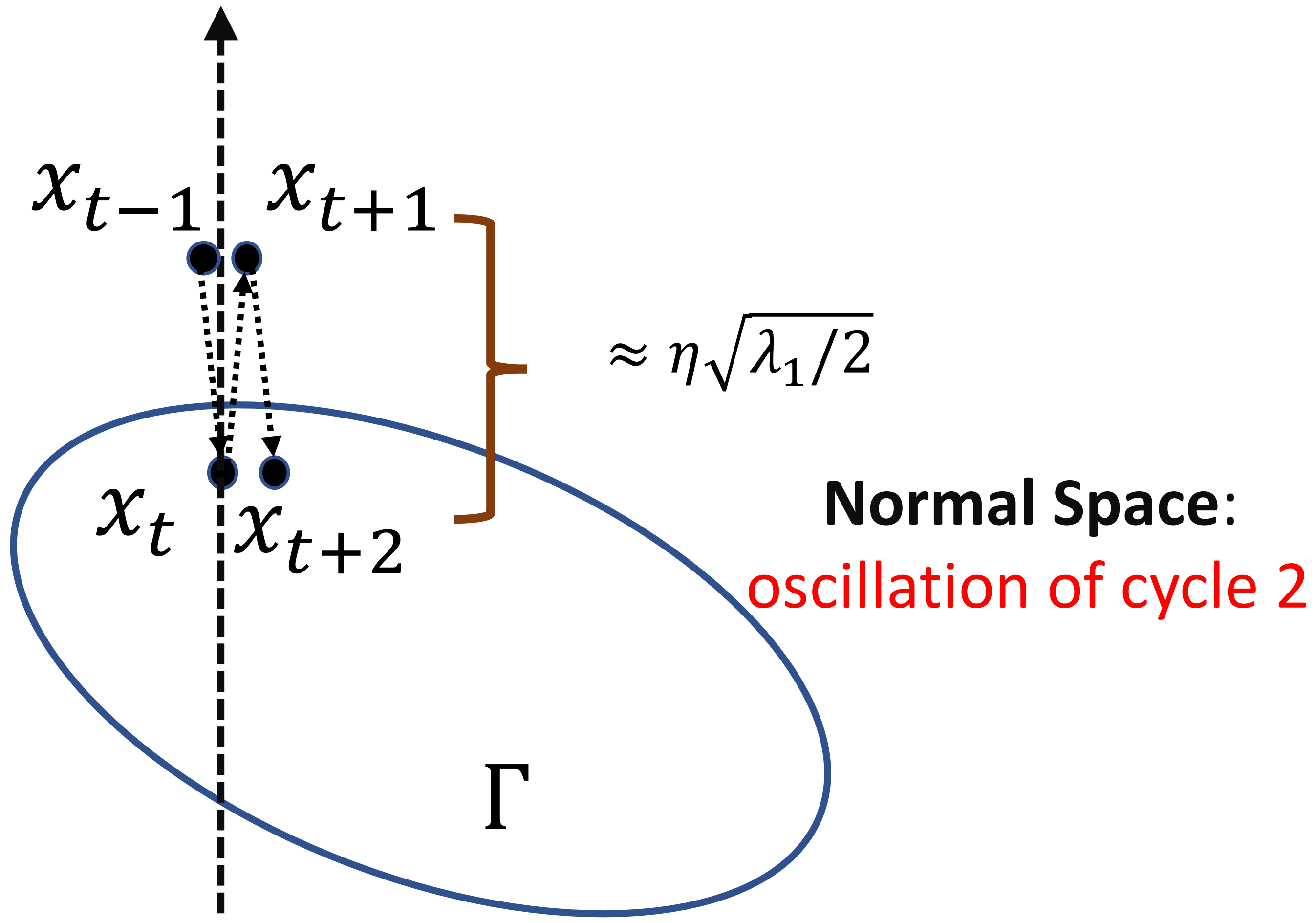
- Implicit regularization of SGD before reaching manifold of minimizers
  - so far only analysis for simple diagonal linear nets [Pesme et al, 21].

- Limiting diffusion for adaptive gradient methods, like momentum-SGD, ADAM

Pesme, Scott, Loucas Pillaud-Vivien, and Nicolas Flammarion. "Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity." *NeurIPS,* 2021.
Smith, Samuel L., Benoit Dherin, David Barrett, and Soham De. "On the Origin of Implicit Regularization in Stochastic Gradient Descent." ICLR'20.
Liu,Yucong, Tong Lin, "Regularizing Deep Neural Networks with Stochastic Estimators of Hessian Trace", Open Review'22

Similar Implicit Bias for GD + finite LR

- Γ: a smooth manifold of minimizers of smooth loss L, where $L_{min} = 0$.
- GD on **non-smooth** loss $\sqrt{L}$, $x_{t+1} - x_t = -\eta \nabla \sqrt{L}(x_t) = -\eta \frac{\nabla L(x_t)}{2\sqrt{L(x_t)}}$
- $\Phi(X)$ is 'landing point' of GF for L on manifold starting from X.



[**ALP**'21]: When $\eta \to 0$, GD on $\sqrt{L}$ dynamic contains two phases:

1. **Gradient Flow phase** ($\Theta(1/\eta)$ steps): $x_{\frac{T}{\eta}} \approx \phi(x_0, T)$.
2. **Limit flow phase**($\Theta(1/\eta^2)$ steps): $x_{\frac{T}{\eta^2}} \approx Y_T$,

where $Y_0 = \Phi(x_0)$, and $Y_t \in \Gamma$ is the Riemannian Gradient Flow minimizing sharpness of L, $\lambda_1(\nabla^2 L(Y_t))$ on manifold.

(Same implicit bias for Normalized GD on L)