Open Problems about **Sharpness**, **Implicit Bias**, and **Generalization**

Zhiyuan Li
TTIC

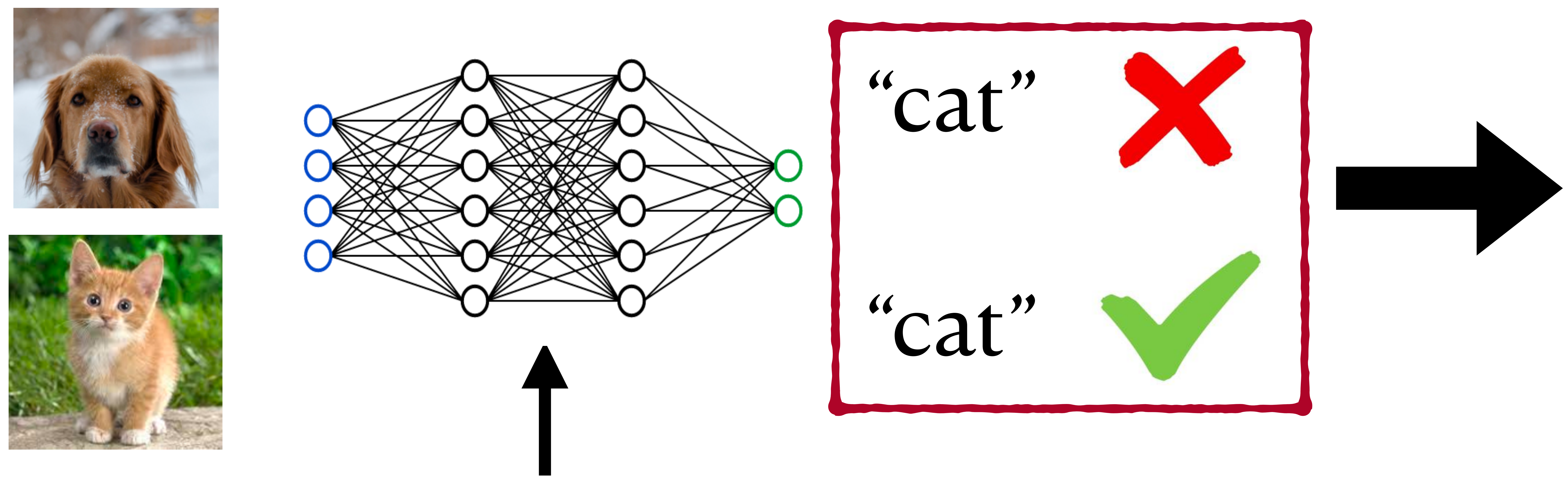
Uconn
July 18, 2024

Based on Joint work with Tianhao Wang, Kaiyue Wen, Tengyu Ma, Kaifeng Lyu, Khashayar Gatmiry, and Sanjeev Arora, …

Basic Paradigm of Deep Learning

Given ***training dataset*** and ***architecture***, find parameter with small ***objective***.

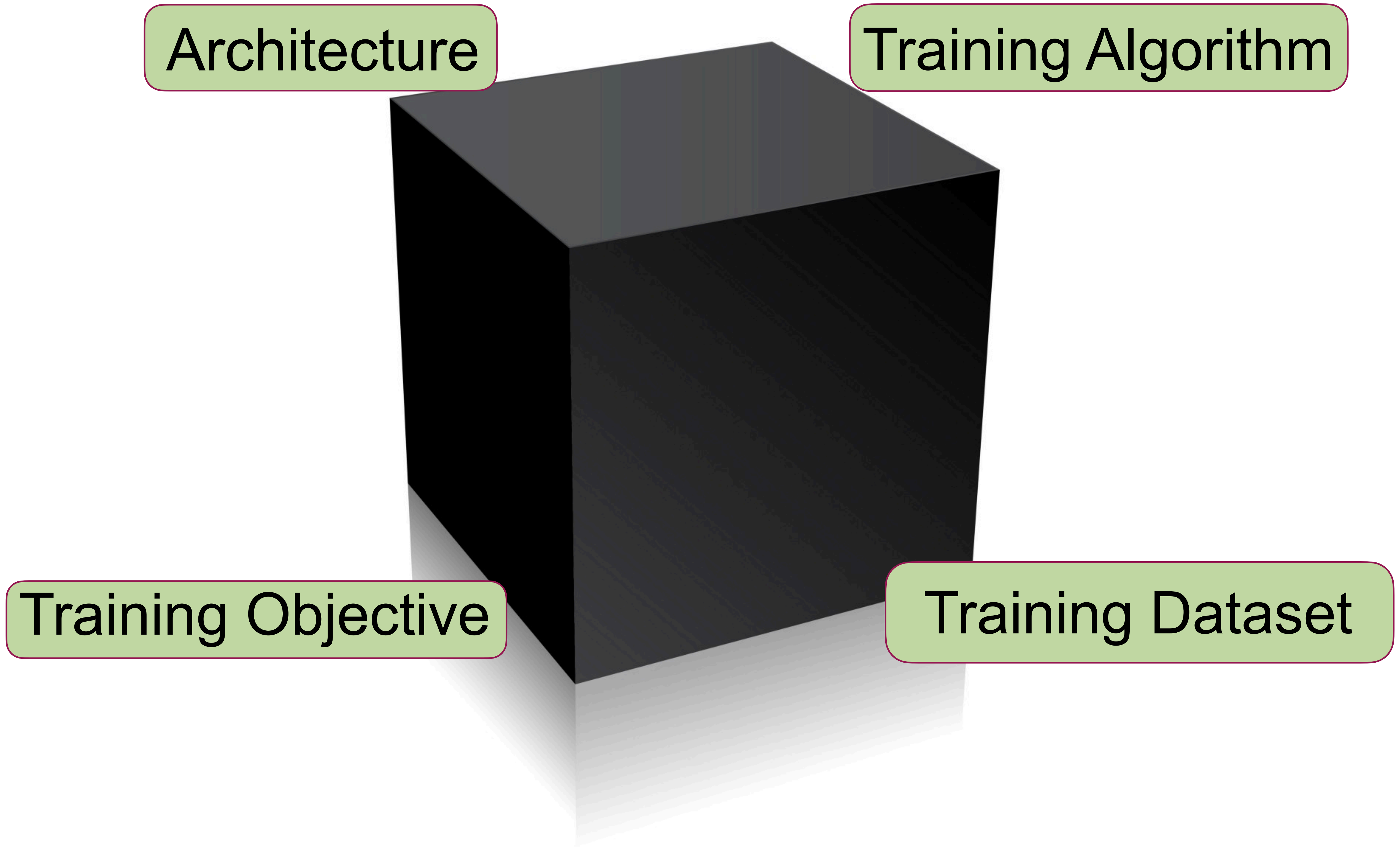
training dataset ***architecture***



parameter x

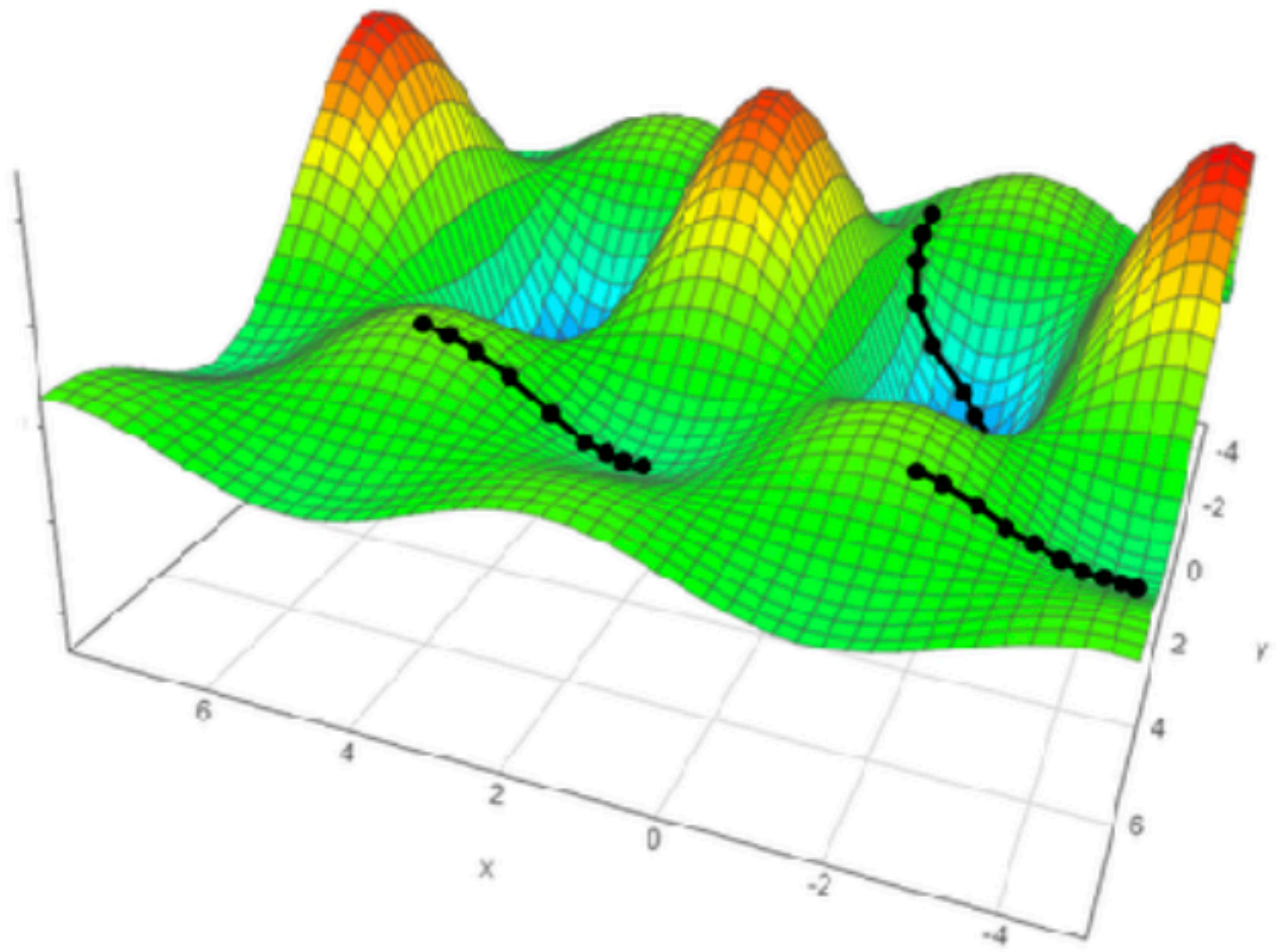
Objective $L(x)$

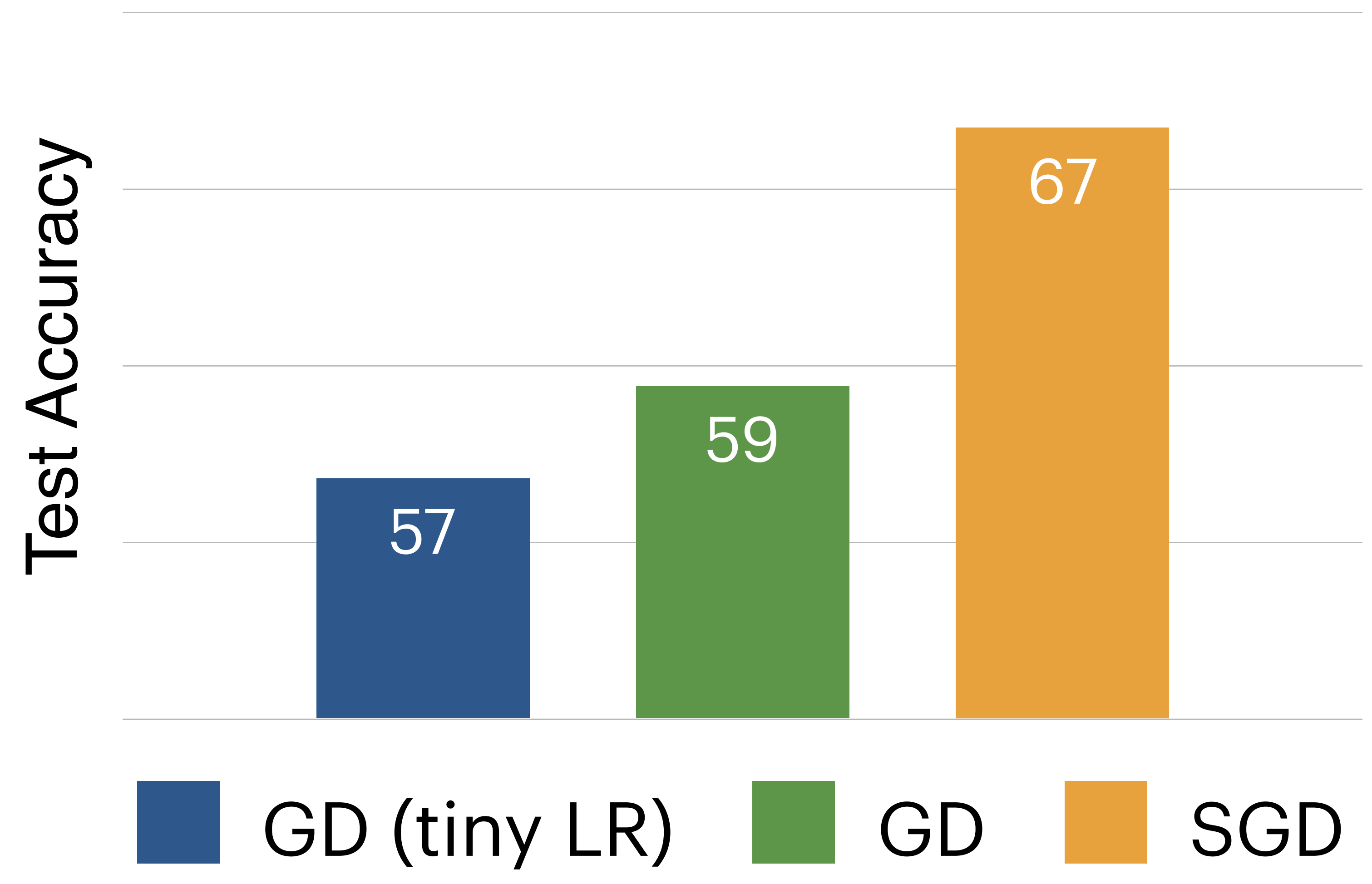
Deep Learning in Practice is a Black Box



Tuned empirically using ‘holdout’ data

Black Box View is Insufficient for Generalization





VGG16 on CIFAR10 (5k samples)

Multiple local minimizers
$\Longrightarrow$different generalization property

GD(tiny LR) < GD < SGD
(wrt generalization)

Implicit Regularization

Implicit Capacity Control

Initialization x_0

Optimize Objective L

End at x_∞, the minimizer of regularizer R among all minimizers of L

Classic Example: GD on Least Square, $L(x) = \|Ax - b\|_2^2$.

$$x_\infty = (A^\top A)^{-1} A^\top (b - Ax_0) + x_0 \implies R(x) = \|x - x_0\|_2^2$$

Implicit Regularization for Non-linear Model

Thm[GWBNS'17]: For matrix factorization loss $L(U, V) = \sum_{i=1}^{n} (\langle UV^\top, A_i \rangle - y_i)^2$, GD from tiny initialization finds min nuclear norm solution.

A brief survey of follow-up works (till 2022):

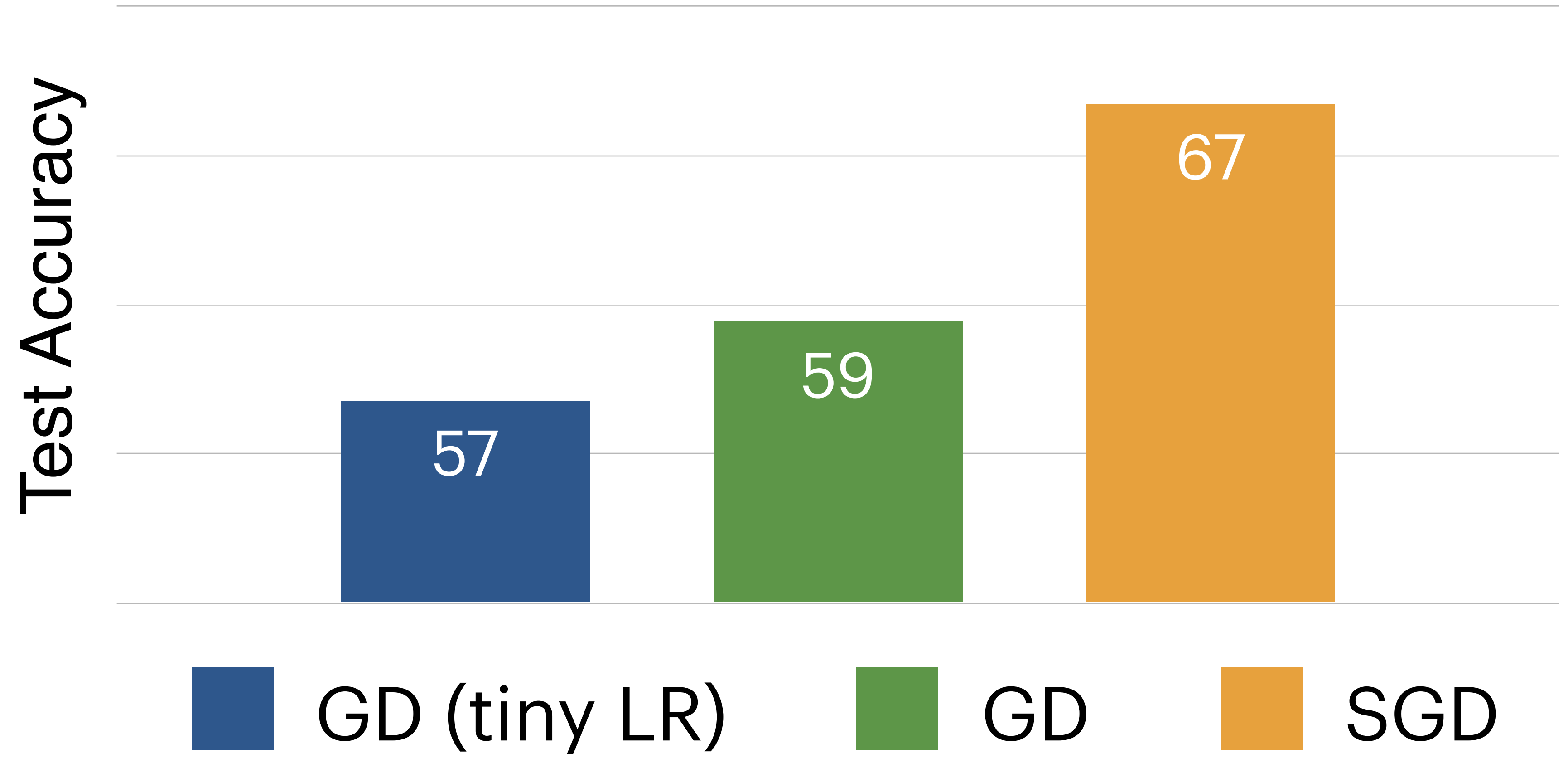
- Matrix Factorization:
  Du et al., 2018; Li et al., 2018; Arora et al., 2019; Gidel et al., 2019; Mulayoff & Michaeli, 2020; Blanc et al., 2020; Gissin et al., 2020; Razin & Cohen, 2020; Chou et al., 2020; Eftekhari & Zygalakis, 2021; Yun et al., 2021; Min et al., 2021; Li et al., 2021a; Razin et al., 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 the results are essentially for deterministic Gradient Flow (GD with infinitesimal LR).
(though some analysis can be discretized)

Gunasekar, Suriya, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. "Implicit regularization in matrix factorization." NeurIPS'17

Question:

What is the role of large learning rate and stochastic gradient noise in implicit regularization?



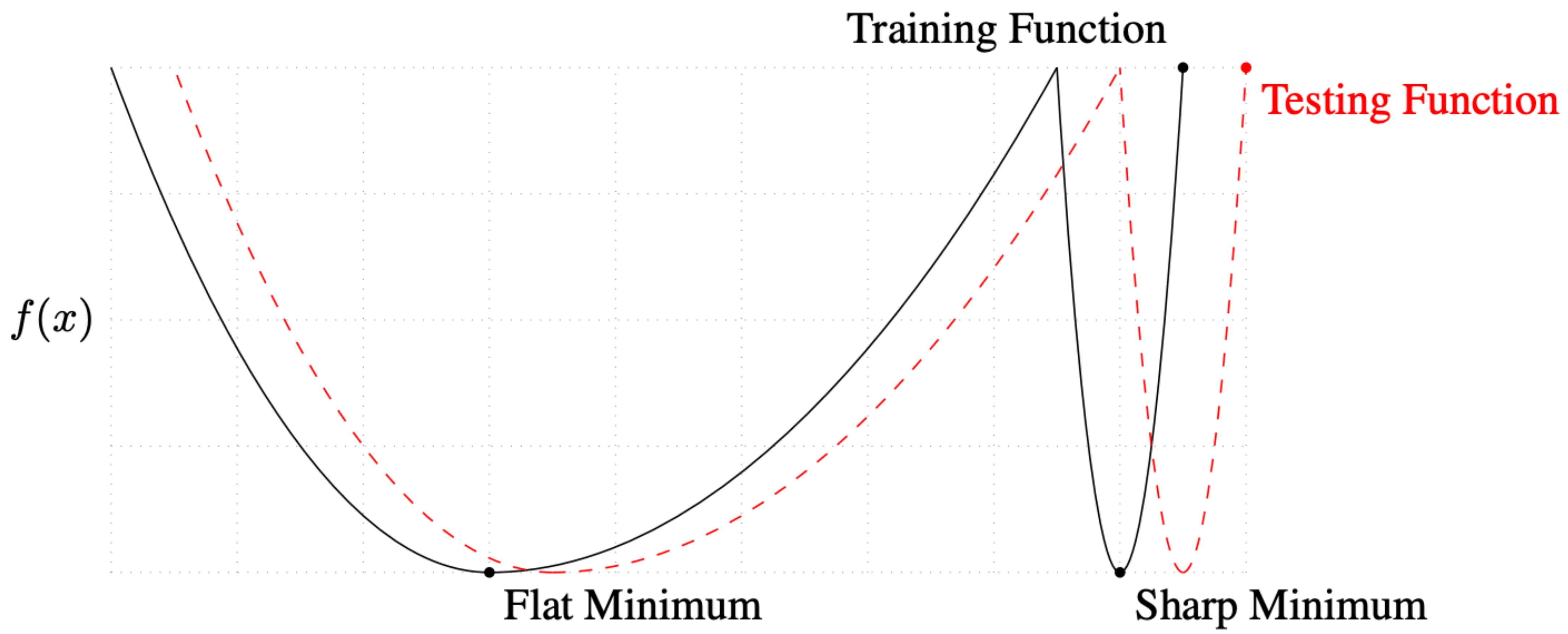
VGG16 on CIFAR10 (5k samples)

A plausible explanation: They reduce sharpness(flatness) of final solution.

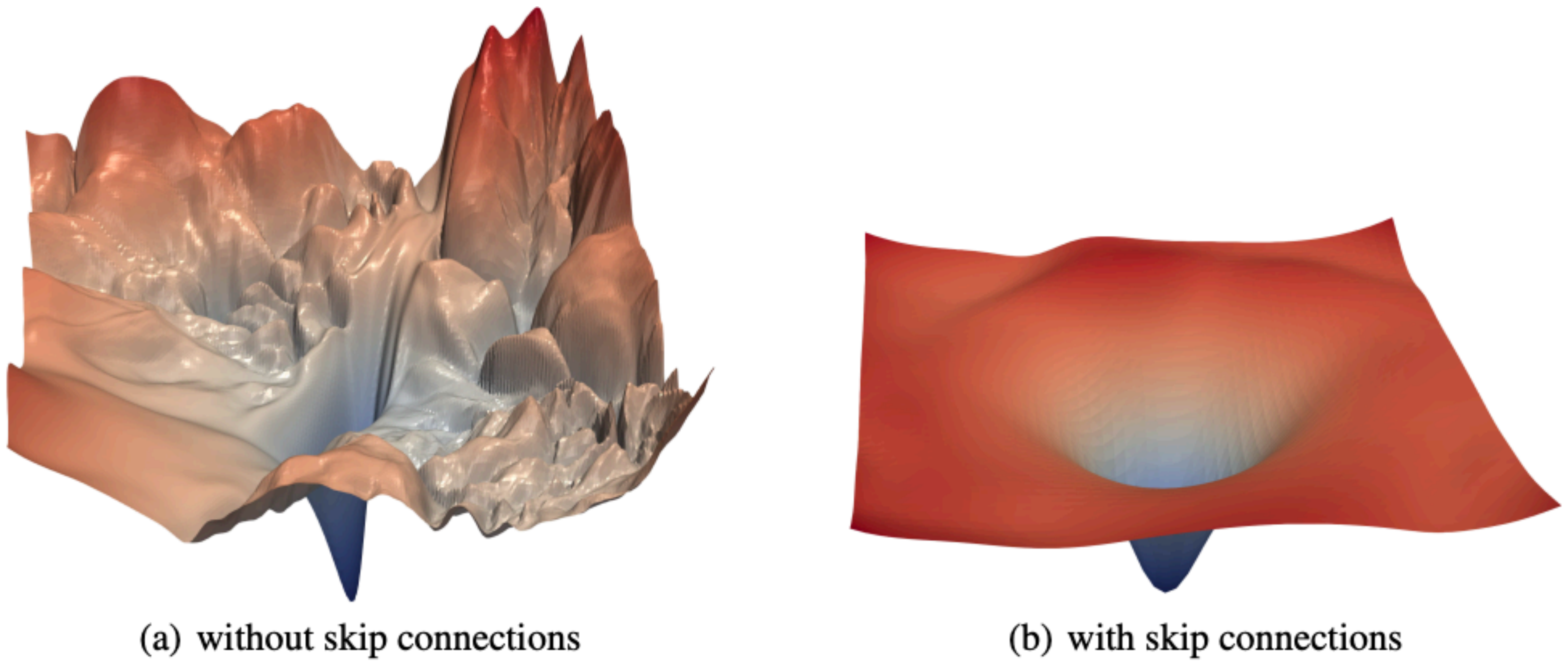
Large LR/Stochasticity **Flat Minima** **Good Generalization**

History and Intuition of Sharpness and Generalization

- Flatter minimizer $\Longrightarrow$ shorter description $\Longrightarrow$ better generalization [Hochreiter&Schmidhuber,97]



[Keskar et al, 16]



Visualization for ResNet-56 [Li et al, 19]

- This talk: sharpness = some function of hessian, e.g., $\lambda_1(\nabla^2 L)$, $\mathrm{Tr}(\nabla^2 \mathrm{L})$, $\det(\nabla^2 \mathrm{L})$.

Large LR $\Longrightarrow$ Flatness: A Discrete Dynamical System View

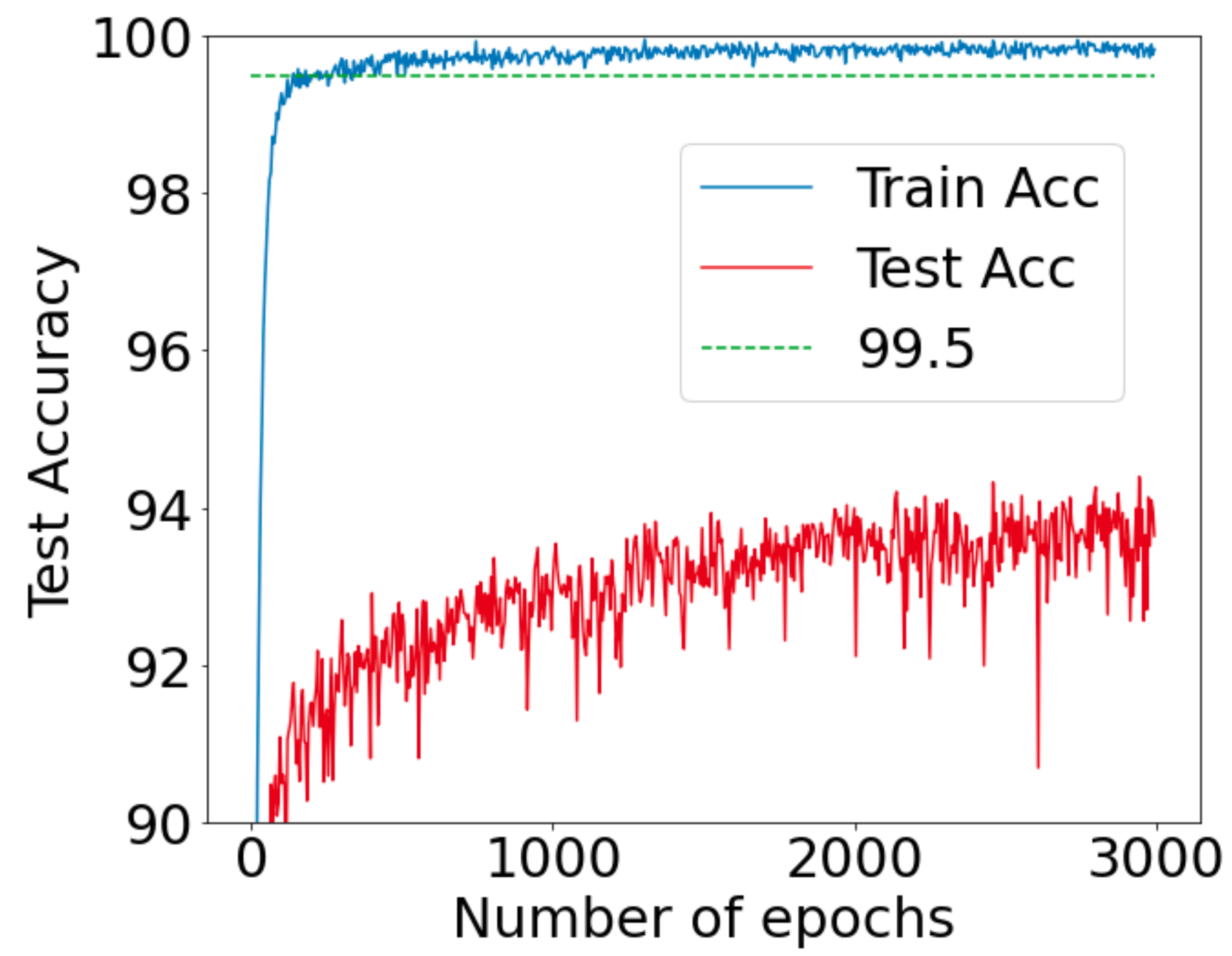
- **Gradient Descent** (GD): $x_{t+1} = x_t - \eta \nabla L(x_t)$
- **Linear Stability**(Folklore): For all x and almost all η,

$$\lambda_1(\nabla^2 L(x)) > 2/\eta \implies \{x_0 \mid \text{GD starting from } x_0 \text{ converges to } x\} \text{ is a 0-measure set.}$$

- **Interpretation**: GD with large LR finds flat minimizers, if it converges.
- Extension to Stochastic GD. [Wu et al.,18, Ma et al., 21]

Gradient Noise $\Longrightarrow$ Flatness: A Continuous Dynamical System View

- Experimental Observation [**L**, Lyu & Arora, 20]:
  - Small LR generalizes equally well, if trained longer.
  - Same phenomena happens for continuous limit of SGD (SDE)
  - Fundemantally different from stability based arguments.



ResNet trained on CIFAR10 with small LR

- Label Noise SGD: $x_{t+1} = x_t - \eta \nabla_x (f_{z_{i_t}}(x_t) - y_{i_t} - \delta_t)^2$, where $\delta_t \overset{iid}{\sim}$ Unif$\{-\delta, \delta\}$.
  - Same phenomena to minibatch SGD, easier to show implicit sharpness regularization.

Li, Zhiyuan, Kaifeng Lyu, and Sanjeev Arora. "Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate." *NeurIPS,20*

Assumptions

(Manifold of Minimizers and Hessian of Maximal Rank)

1. A $(D-M)$-dimensional smooth manifold, $\Gamma \subset \mathbb{R}^D$, consists only of minimizers of loss L
   - Our results hold in the attraction set of Γ under gradient flow.
   - Empirical evidence: Mode Connectivity [Garipov et al., 18; Draxler et al., 18]

2. Hessian has maximal rank at every point on Γ, i.e., $\mathrm{rank}(\nabla^2 L(x)) = M$.

Same assumptions made by [Fehrman et al.,20; Li et al., 21; Arora et al., 22]

Why manifold and maximal Hessian rank? Overparametrization!

Assumptions

(Manifold of Minimizers and Hessian of Maximal Rank)

1. A $(D-M)$-dimensional smooth manifold, $\Gamma \subset \mathbb{R}^D$, consists only of minimizers of loss L
2. Hessian has maximal rank at every point on Γ, i.e., $\mathrm{rank}(\nabla^2 L(x)) = M$.

A concrete example satisfying assumptions:

- $L_i(x) = \ell(f_i(x), y_i)$, where $f_i(x)$ is output on ith data and y_i is the ith label.
- $\ell(y, y') = 0.5(y - y')^2$. (can be other losses)
- $L(x) = (1/n)\sum_{i=1}^{n} L_i(x)$ and assume $\min_x L(x) = 0$
- $\Gamma \triangleq \{x \mid L(x) = 0 \wedge \{\nabla f_i(x)\}_{i=1}^{n}$ are linearly independent$\}$ = global minimizers whose NTK is full-rank

Thm: Above defined Γ and L satisfy Assumptions 1 and 2.

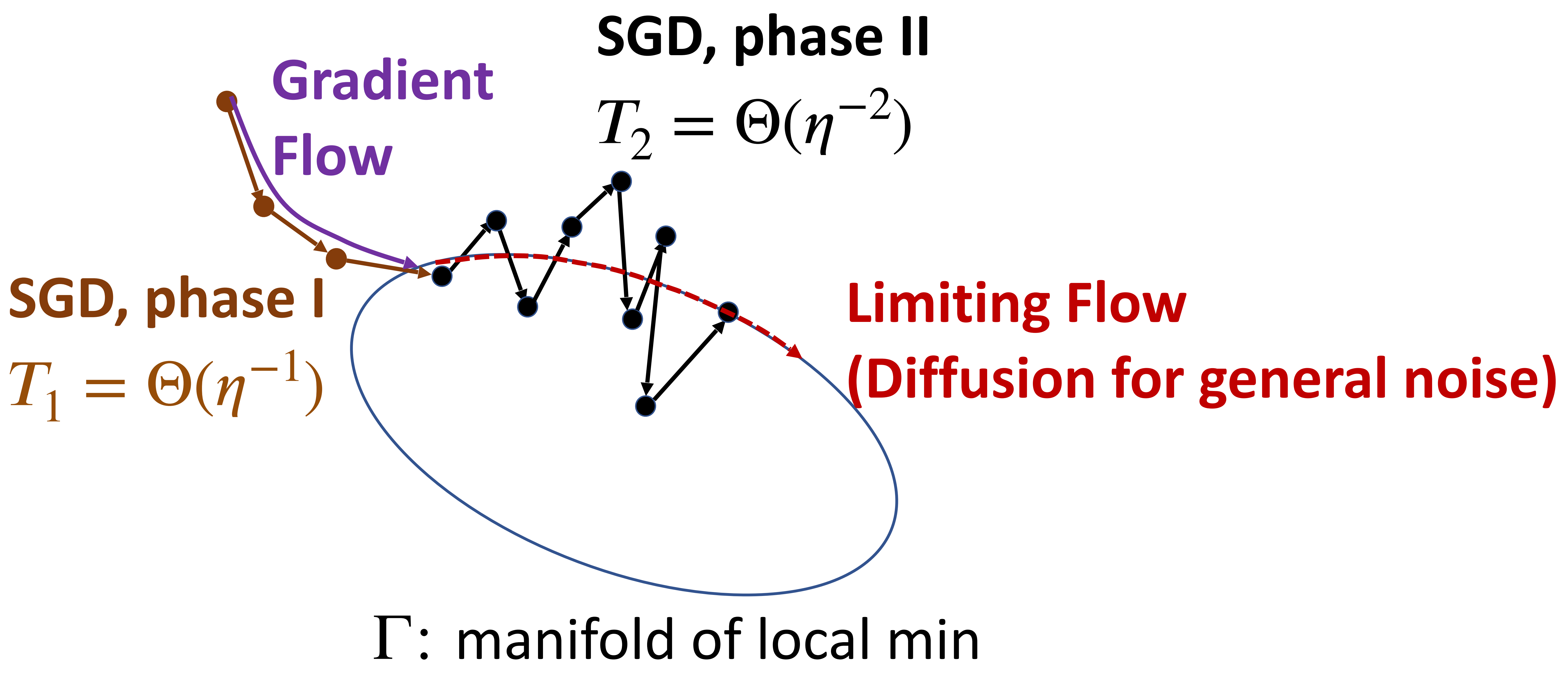
Thm[Cooper, 18]**:** For almost all $\{y_i\}_{i=1}^{n}$, $\Gamma = \{x \mid L(x) = 0\}$. (Use Sard's theorem)

Sharpness-reduction Flow

Assumption: Minimizers of loss $L(x)$ form a smooth manifold. (Why? Overparametrization)

Thm[L,Wang&Arora,21]: When $\eta \to 0$, label noise 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 Flow phase**($\Theta(1/\eta^2)$ steps): $x_{\frac{T}{\eta^2}} \to Y_T$, where $\frac{dY_\tau}{d\tau} = -\frac{\delta^2}{4}\nabla_\Gamma \mathrm{Tr}(\nabla^2 L(Y_\tau))$



Li, Zhiyuan, Tianhao Wang, and Sanjeev Arora. "What Happens after SGD Reaches Zero Loss?--A Mathematical Framework." *ICLR'22 spotlight*

Sharpness as a Generalization Bound

- PAC-Bayesian bound [McAllester, 99; Dziugaite&Roy,17; Neyshabur et al.,17]

$$\text{Generalization Gap} \leq \underbrace{\mathbb{E}_{v\sim\mathcal{N}(0,\sigma^2 I_d)}[L_n(\tilde{x}+v)] - L_n(\tilde{x})}_{PAC-Bayesian-sharpness} + \frac{\|\tilde{x}\|_2/2\sigma}{\sqrt{2m}}$$

- PAC-Bayesian-Sharpness $\approx$ Trace of Loss Hessian for small σ

$$\mathbb{E}_{v\sim\mathcal{N}(0,\sigma^2 I_d)}[L(\tilde{x}+v)] - L(\tilde{x}) \approx \frac{\sigma^2}{2}\mathrm{Tr}[\nabla^2 L(\tilde{x})].$$

Sharpness as a Generalization Predictor

- ε-sharpness [Keskar et al.,16]: (simplified version)

$$\phi_\varepsilon(x, L) = \sup_{\|x'-x\|\le\varepsilon} L(x') - L(x)$$

- If L is C^2 and $\nabla L(x) = 0$, then $2\phi_\varepsilon(x, L)/\varepsilon^2 \approx \lambda_1(\nabla^2 L(x))$ for small ε.

$$\begin{tikzcd}
 & \textbf{Batch Size}\downarrow \arrow[dl] \arrow[dr] & \\
\textbf{Test Error}\downarrow & & \varepsilon\textbf{-sharpness}\downarrow
\end{tikzcd}$$

- ε- and PAC-Bayesian-sharpness correlates well with generalization error. [Jiang et al.,19]
- Good predictor should be aware of architectural-symmetry.[Dinh et al.,17]
  - It is trivial to use homogeneity to construct networks with good generalization and arbitrarily large sharpness.
  - But cannot construct nets w. small sharpness

Sharpness as A Regularizer

- Regularizing ε-sharpness explicitly using Sharpness Aware Minimization method (SAM) improves generalization of ResNet. [Foret et al.,20]

- SAM improves performance of ViT-B and MLP-Mixer as well. [Chen et al.,22]

Sharpness-Aware Minimization

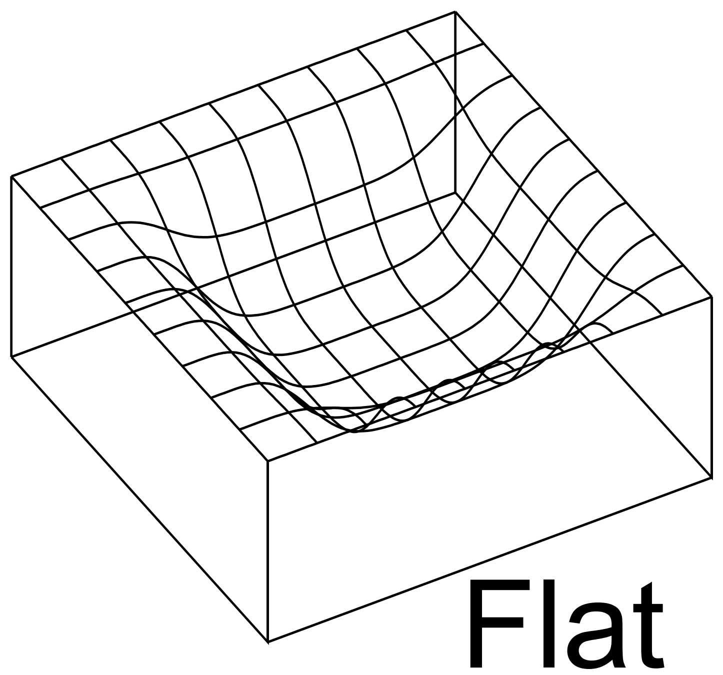
- Sharpness-Aware Minimization (SAM)[Foret et al.,21;Zheng et al.,21;Norton&Royset,21]:

$$x(t+1) = x(t) - \eta \nabla L\left(x(t) + \rho \boxed{\frac{\nabla L(x(t))}{\|\nabla L(x(t))\|}}\right)$$

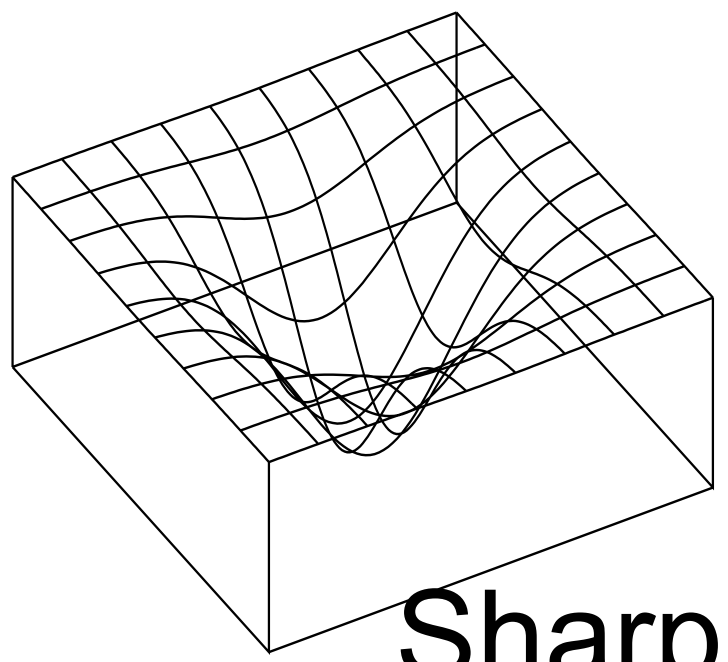
where $x \in \mathbb{R}^D$ is parameters, η is learning rate, ρ is perturbation radius.

Stop gradient on 2nd&3rd occurrence of x!

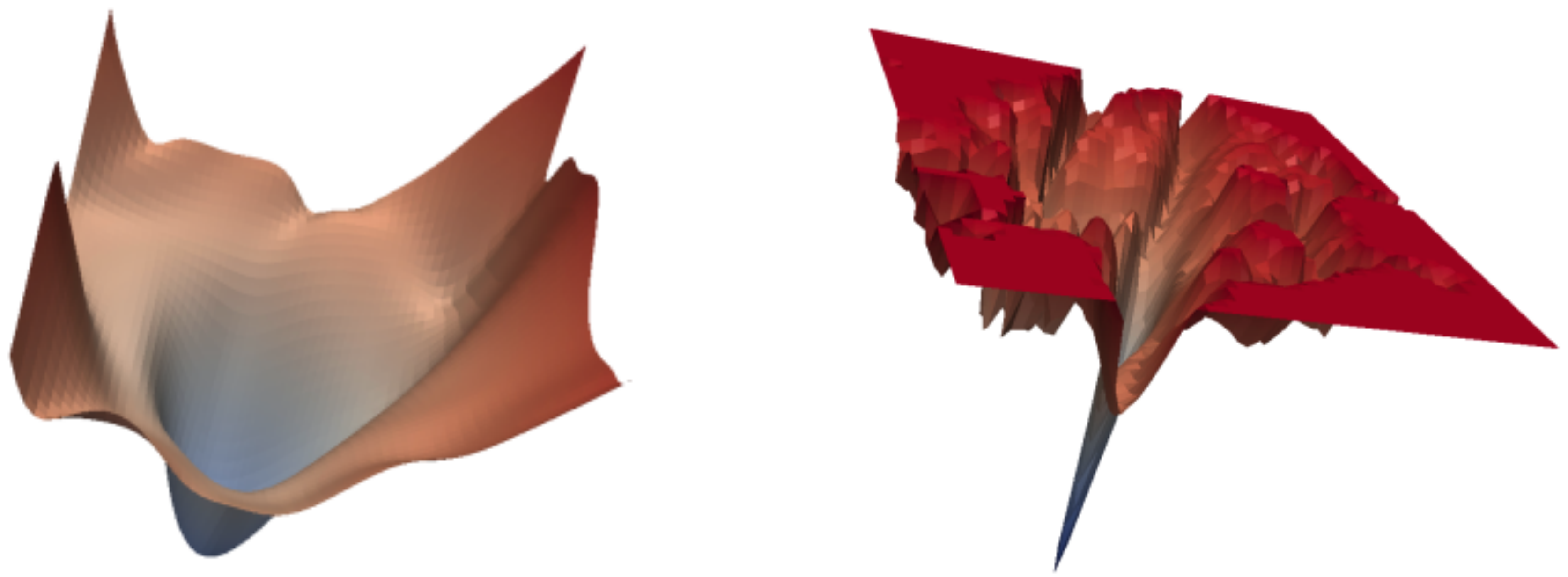
- **Intuition**: Improve generalization by finding a flat solution.



Flat



Sharp



ResNet found by SAM ResNet found by SGD

[Foret et al., 21]

Sharpness as A Regularizer

- Regularizing ε-sharpness explicitly using Sharpness Aware Minimization method (SAM) improves generalization of ResNet. [Foret et al.,20]

- SAM improves performance of ViT-B and MLP-Mixer as well. [Chen et al.,22]

- Regularizing $\mathrm{Tr}[\nabla^2 L]$ escapes from minimizer with poor generalization. [Damian et al.,21]

- **Open Question 1**: Does min $\mathrm{Tr}[\nabla^2 L]$ interpolating solution have good generalization for most network architectures and datasets?

Generalization Bounds for Flattest Interpolating Solution

- Regression settings where flattest (min $\mathrm{Tr}(\nabla^2 L)$)interpolating solution provably generalizes:
  - Quadratically Overparametrized Linear Model [**Li**,Wang,Arora,22]:
    $x = (u, v), u, v \in \mathbb{R}^d. \quad f_z(x) = \langle u^{\odot 2} - v^{\odot 2}, z \rangle, z_i \overset{iid}{\sim} \mathrm{Unif}\{-1,1\}^d$
    Ground truth = sparse linear function
  - Matrix Facorization [Ding et al., 2024]:
    $x = (U, V), U, V \in \mathbb{R}^{d\times d}. \quad f_A(x) = \langle UV^\top, A \rangle, (A_i)_{jk} \overset{iid}{\sim} N(0,1)$
    Ground truth = low-rank matrix
  - Deep Matrix Facorization [Gatmiry,**Li**,Chuang,Reddi,Ma,Jegelka, 2023]:
    $x = (W_1, \ldots, W_L), W_i \in \mathbb{R}^{d\times d}. \quad f_A(x) = \langle W_1 W_2 \ldots W_L, A \rangle, (A_i)_{jk} \overset{iid}{\sim} N(0,1)$
    Ground truth = low-rank matrix

Gatmiry, K., Li, Z., Chuang, C.Y., Reddi, S., Ma, T. and Jegelka, S., 2023. The inductive bias of flatness regularization for deep matrix factorization. *arXiv preprint arXiv:2306.13*

Ding, L., Drusvyatskiy, D., Fazel, M. and Harchaoui, Z., 2024. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*,

Provable Generalization Benefit of Label Noise SGD

- $x = (u, v), u, v \in \mathbb{R}^d$. $f_z(x) = \langle u^{\odot 2} - v^{\odot 2}, z \rangle, z_i \overset{iid}{\sim}$ Unif$\{-1,1\}^d$
- $L(X) = \frac{1}{2n} \sum_{i=1}^{n} (f_{z_i}(x) - y_i)^2, \Gamma = \{x \mid f_{z_i}(x) = y_i, u_i^2 + v_i^2 > 0\}$. $(n \gg d)$
- limiting flow of Label Noise SGD: $dY_t = -c \nabla_\Gamma \mathrm{tr}[\nabla^2 L(Y_t)] dt$

- $\nabla f_{z_i}(u, v) = 2 \begin{pmatrix} z_i \odot u \\ -z_i \odot v \end{pmatrix}$, so the regularizer is

$$\mathrm{tr}[\nabla^2 L(u, v)] = \mathrm{tr}\left(\frac{1}{n} \sum_{i=1}^{n} \nabla f_i(u, v) \nabla f_i(u, v)^\top\right) = \frac{4}{n} \sum_{i=1}^{n} \left\| \nabla f_{z_i}(u, v) \right\|_2^2 = 4(\|u\|_2^2 + \|v\|_2^2)$$

Provable Generalization Benefit of SGD: Details

- $x = (u, v), u, v \in \mathbb{R}^d$. $\quad f_z(x) = \langle u^{\odot 2} - v^{\odot 2}, z \rangle,$
- $\Gamma = \{x \mid f_{z_i}(x) = y_i, u_i^2 + v_i^2 > 0\}$. $(n \gg d)$
- Regularizer $\frac{1}{4}\mathrm{tr}[\nabla^2 L(u, v)] = \|u\|_2^2 + \|v\|_2^2 \geq \|u^{\odot 2} - v^{\odot 2}\|_1$

- Moreover, $\mathrm{argmin}_{u,v \in \overline{\Gamma}} \|u\|_2^2 + \|v\|_2^2 = \mathrm{argmin}_{u,v \in \overline{\Gamma}} \|u^{\odot 2} - v^{\odot 2}\|_1$
  - u, v must have disjoint support, implying $\|u\|_2^2 + \|v\|_2^2 = \|u^{\odot 2} - v^{\odot 2}\|_1$

- Suppose $y_i = \langle z_i, w^* \rangle$, w^* is k-sparse. Optimal sparse recovery like LASSO!

Generalization Bounds for Flattest Interpolating Solution

- Regression settings where flattest (min $\mathrm{Tr}(\nabla^2 L)$)interpolating solution provably generalizes:
  - Two-layer network with ReLU activation: [Wu&Su,23][Wen,**Li**,Ma,23]
    $x = (W_2, W_1), W_2 \in \mathbb{R}^{1\times h}, W_1 \in \mathbb{R}^{h\times d}, f_z(x) = W_2\mathrm{ReLU}(W_1 x)\,.\, z_i \sim \mathrm{Unif}(S^{d-1})$
    Ground truth = two-layer relu network with small Barron norm.

    *Barron norm = $\displaystyle\sum_{i=1}^{h} |(W_2)_i| \|(W_1)_{i:}\|_2$

- Open Question 1.1: Can we prove generalization benefit for deep (depth ≥ 3) relu networks?

Wu, L. and Su, W.J., 2023, July. The implicit regularization of dynamical stability in stochastic gradient descent. In *International Conference on Machine Learning*

Wen, K., Li, Z. and Ma, T., 2024. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *NeurIPS*

Open Questions on Computational Efficiency

- Assuming flattest interpolating solution has good generalization…
- **Open Question 2:** How do we efficiently find the flat/flattest minimizer?
  - **Q 2.1:** Does the Riemmanian gradient flow of trace of hessian even converge to the global minimizer on manifold?

$$dY_t = -c\nabla_\Gamma \mathrm{tr}[\nabla^2 L(Y_t)]dt$$

  - **Q 2.2:** Our current analysis [Li,Wang,Arora,21] is essentially asymptotic for learning rate $\eta \to 0$. Need non-asymptomatic version to decide what the largest η is which still tracks the flow on manifold.
  - **Q 2.3:** Analysis for SGD with more realistic noise structure which also exhibits flatness implicit bias.