# Disentangling Trainability and Generalization In Deep Neural Networks

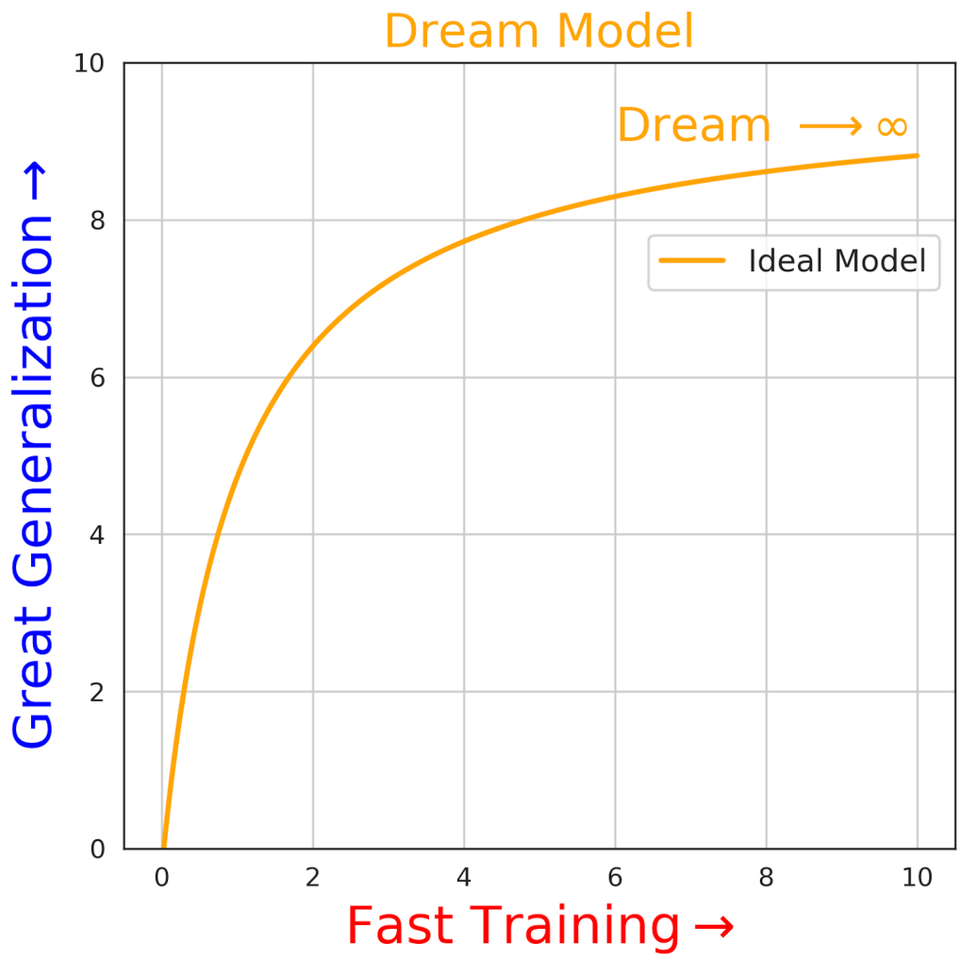
Lechao Xiao, Jeffrey Pennington and Samuel S. Schoenholz

Google Brain Team, Google Research

Colab Tutorial

# Two Fundamental Theoretical Questions in Deep Learning

- Trainability / Optimization
  - Efficient algorithm to reach global minima
- Generalization
  - performant on unseen data

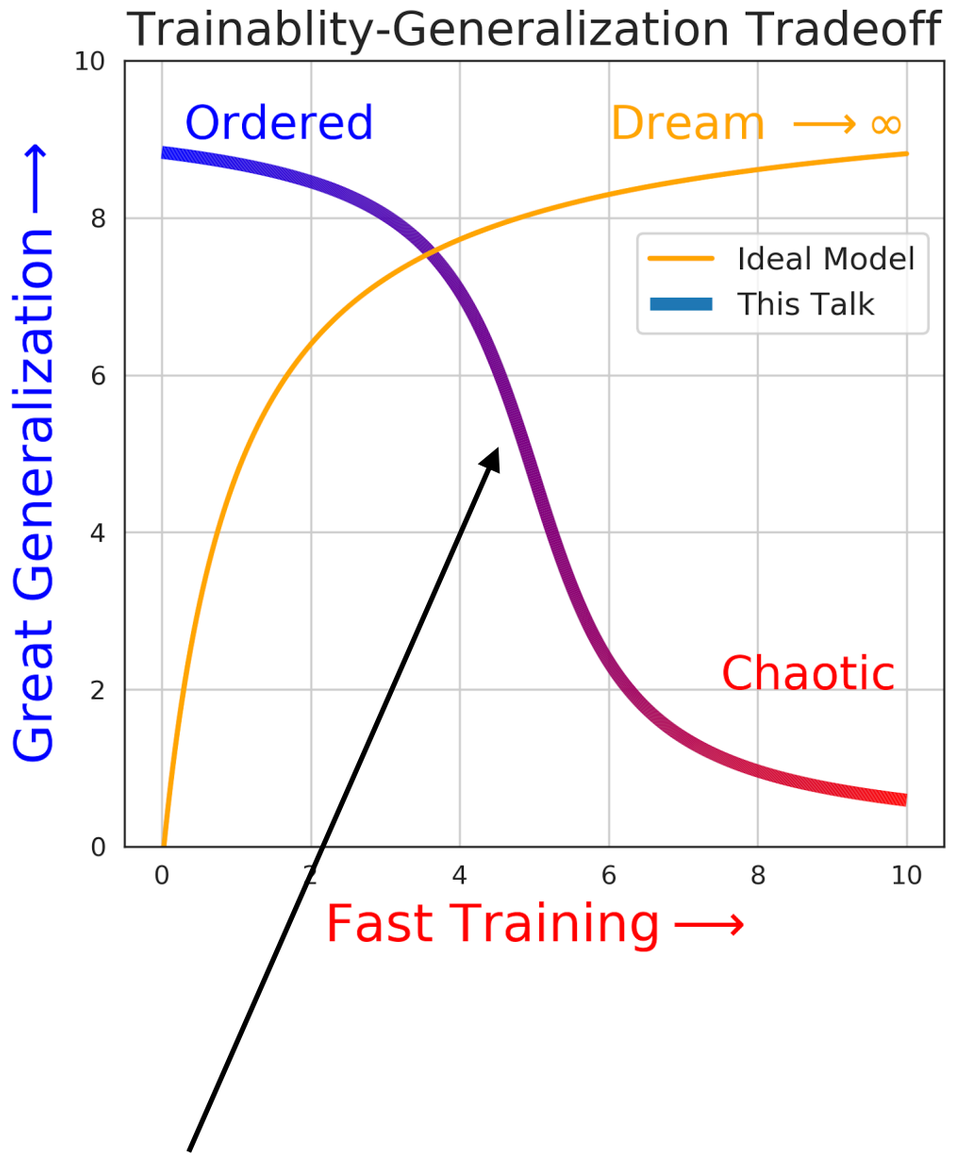



- **Dream**
  - (model, algorithm): Fast Training + Fantastic Generalization
  - Solves AGI

# A trade-off between Trainability and Generalization for very *deep* and very *wide* NNs

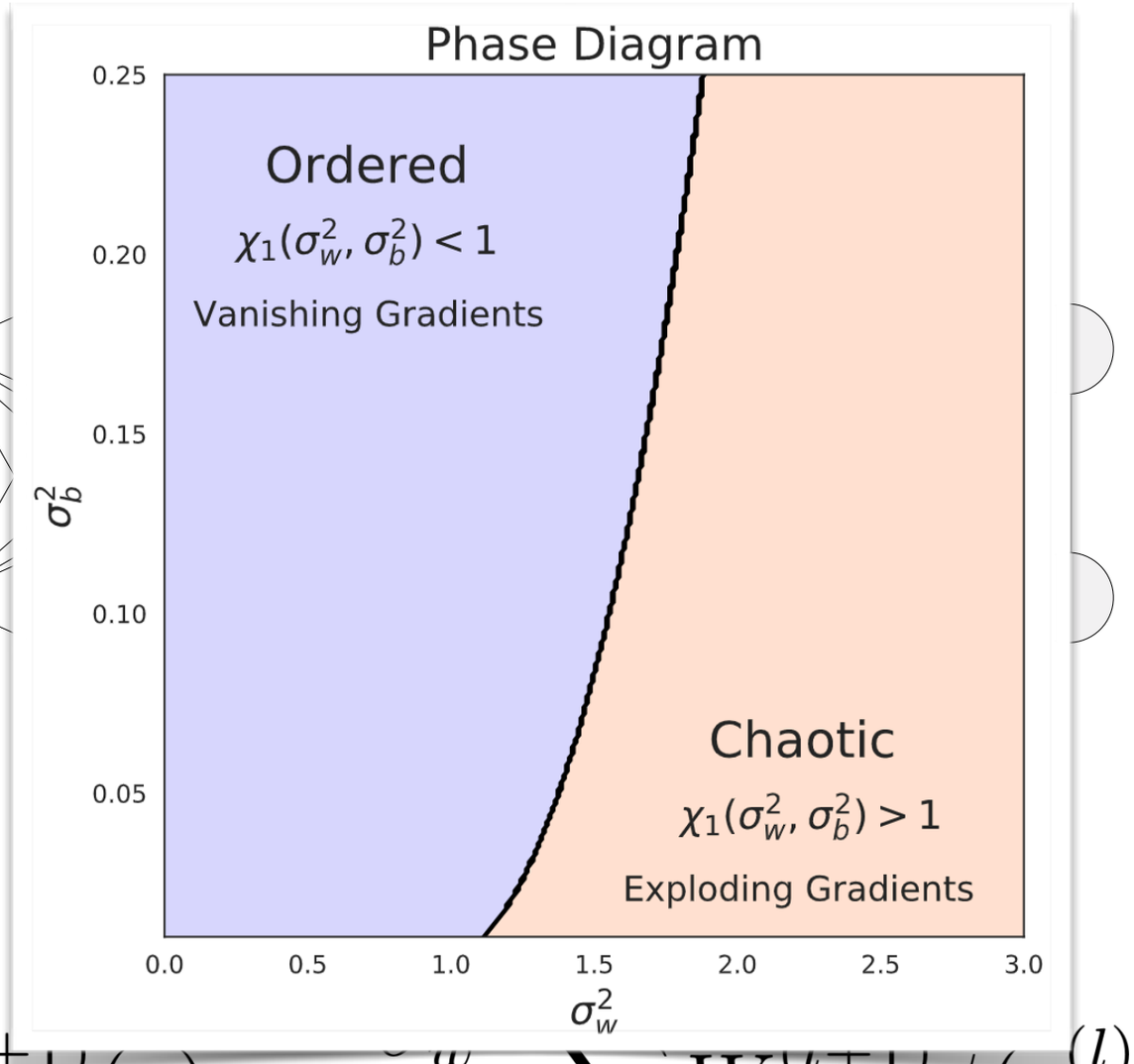
- Trained Fast, but NOT generalizable
  - **Large** Weight Initialization (Chaotic Phase)

- Trained Slowly, able to generalize
  - **Small** Weight Initialization (Ordered Phase)




**Deep Neural Networks**

# Neural Networks Initialization




$$W_{ij}^{(l)}, b_i^{(l)} \sim \mathcal{N}(0,1)$$

$\sigma_\omega^2$ : Weights Variance

$\sigma_b^2$ : Biases Variance

$$z_i^{(l+1)}(x) = \frac{\omega}{\sqrt{n^{(l)}}} \sum_{j=1} W_{ij}^{(l+1)} \phi(z_j^{(l)}(x)) + \sigma_b b_i^{(l+1)}$$

# Training Dynamics and NTK

Gradient descent dynamics with Mean Squared Error

**Function Space**

$$\frac{df_\theta(X_{\text{train}})}{dt} = -\eta\Theta_{\text{train, train}}\left(f_\theta(X_{\text{train}}) - Y_{\text{train}}\right)$$

**Neural Tangent Kernel (NTK)**

$$\Theta_{\text{train, train}} = \nabla f_\theta(X_{\text{train}})\nabla f_\theta(X_{\text{train}})^T$$

- In the infinite width setting, the NTK is deterministic and remains a constant through training (NTK Jacot et al., 2018)

- The above ODE has a closed form solution.

# Training and Learning Dynamics

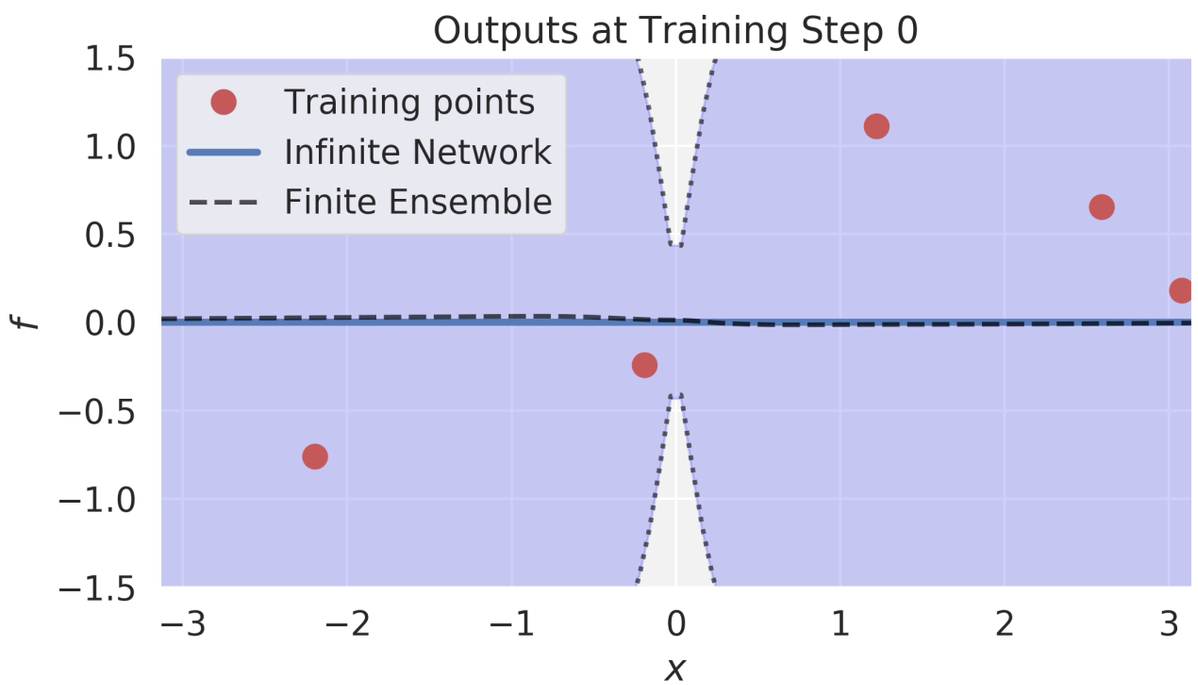
**Training Dynamics:**

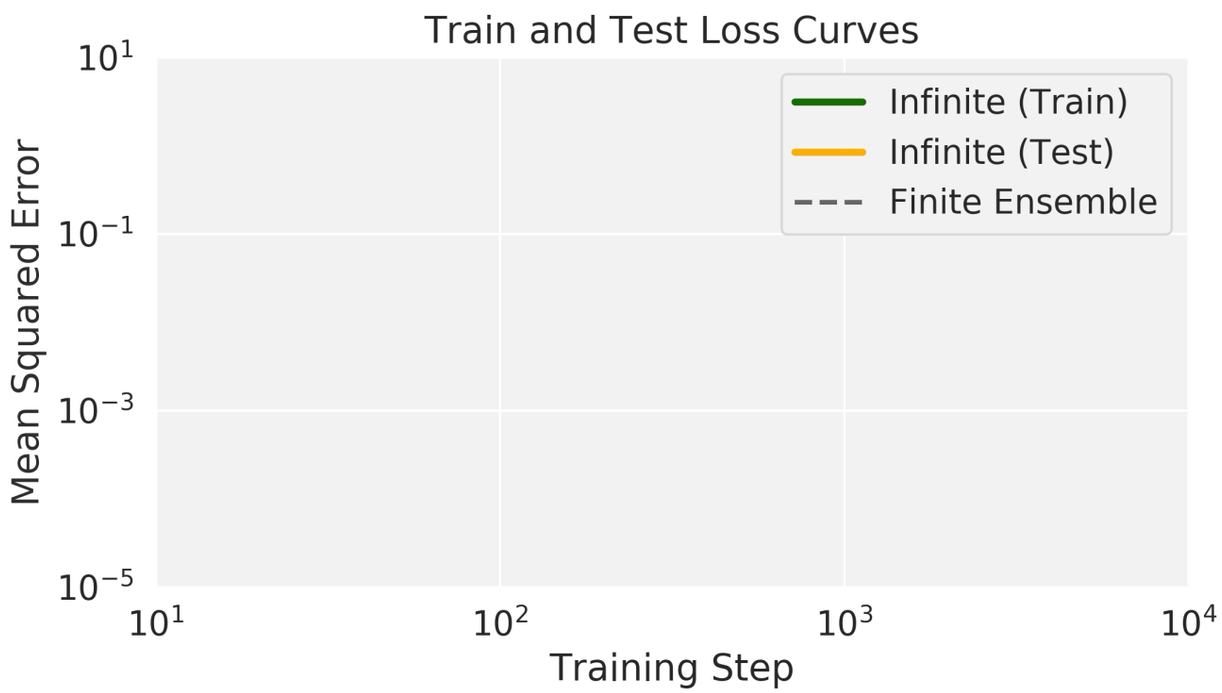
$$\mu_t(X_{\text{train}}) = (\mathbf{Id} - e^{-\eta \Theta_{\text{train, train}} t}) Y_{\text{train}}$$

**Learning Dynamics:**

$$\mu_t(X_{\text{test}}) = \Theta_{\text{test, train}} \Theta_{\text{train, train}}^{-1} (\mathbf{Id} - e^{-\eta \Theta_{\text{train, train}} t}) Y_{\text{train}}$$

**Agreement between finite- and infinite-width networks**




Credit: Roman Novak

# Metric for Trainability: Condition Number

**Training Dynamics:**

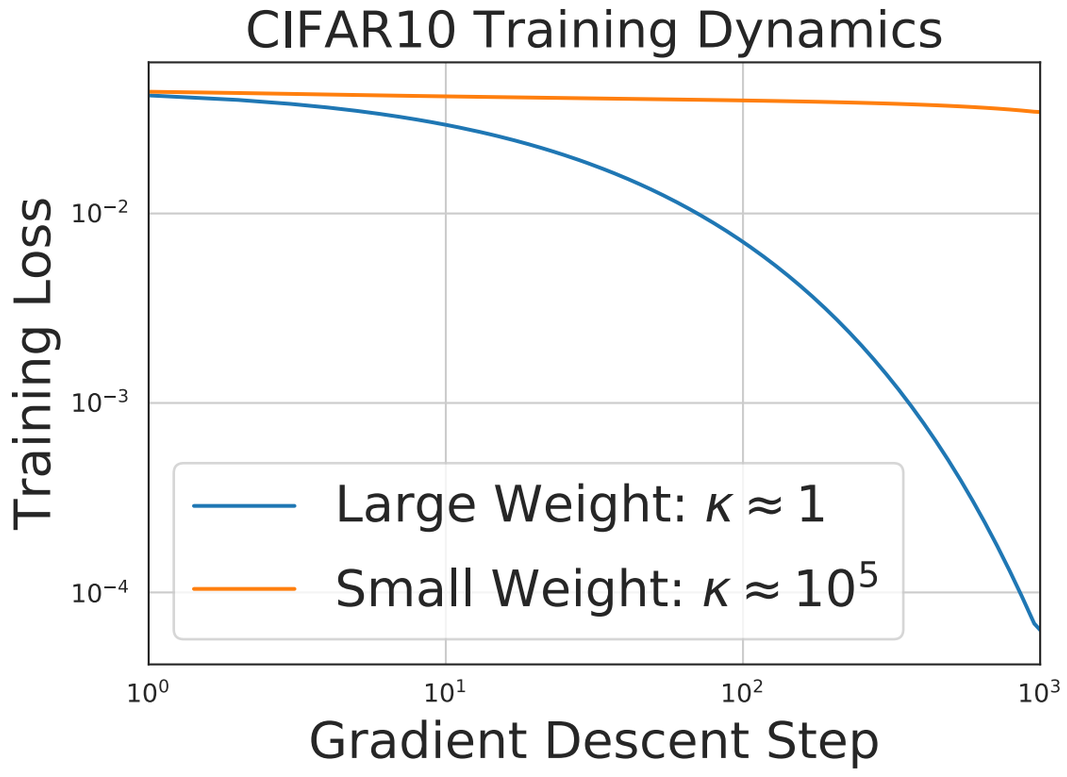
$$\mu_t(X_{\text{train}}) = (\mathbf{Id} - e^{-\eta \Theta_{\text{train, train}} t}) Y_{\text{train}}$$

Eigen-decomposition

$$\Theta_{\text{train, train}} = U^T D U$$

$$D = \text{diag}(\lambda_0, \lambda_1, \cdots, \lambda_m)$$

$$\tilde{\mu}_t(X_{\text{train}})_i = (\text{Id} - e^{-\eta \lambda_i t}) \tilde{Y}_{\text{train},i}$$




The smallest eigenvector converges at rate $1/\kappa$

**Trainability Metric:** $\kappa = \dfrac{\lambda_0}{\lambda_m}$

8-layers finite width FCN on CIFAR10

- Blue $\sigma_w^2 = 25$
- Orange $\sigma_w^2 = 0.5$

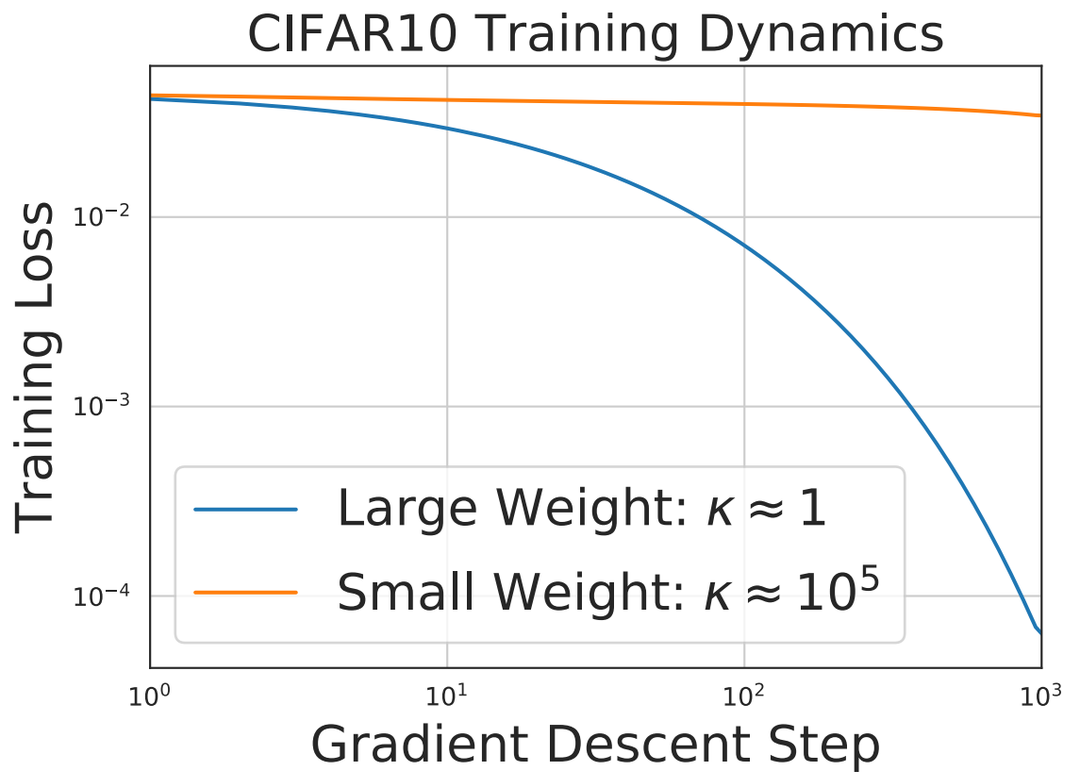
# Metric for Generalization: Mean Prediction

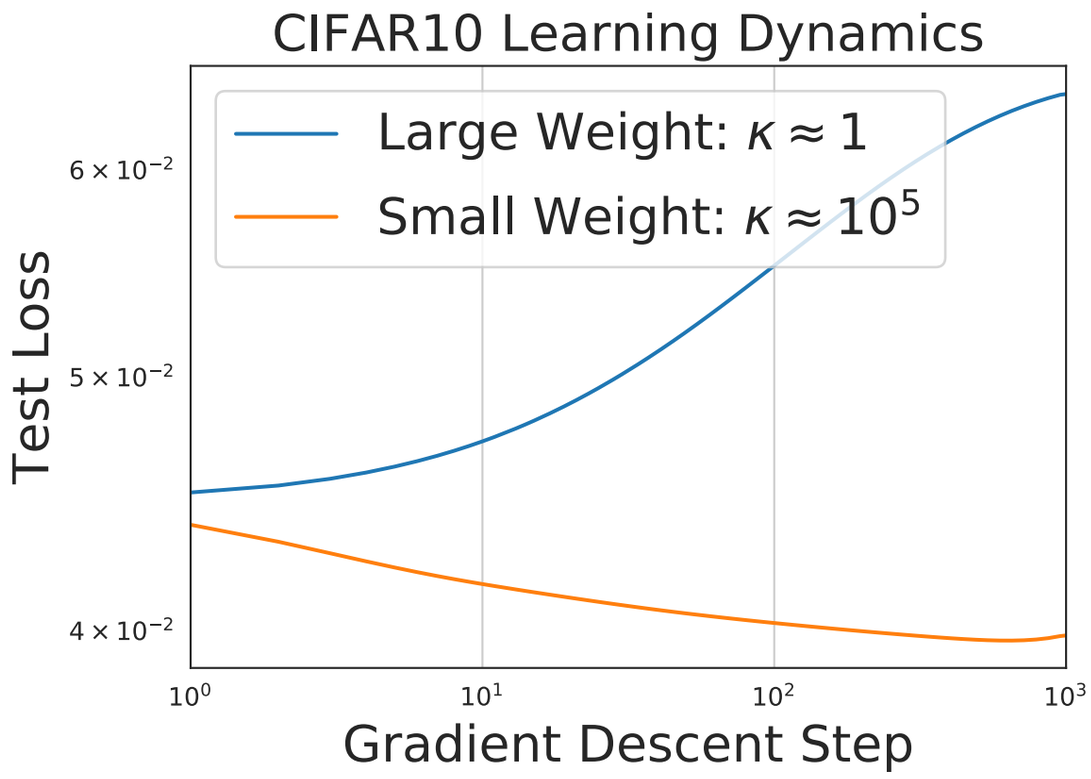
**Learning Dynamics:**

$$\mu_t(X_{\text{test}}) = \Theta_{\text{test, train}}\Theta_{\text{train, train}}^{-1}\left(\mathbf{Id} - e^{-\eta\Theta_{\text{train, train}}t}\right)Y_{\text{train}}$$

**Mean**

**Gene**




**Cannot generalize if $P(\Theta)Y_{train}$ becomes completely independent of the inputs.**

# Evolution of the Metrics with depth

**Neural Networks** $\cdots \to f_\theta^{(l)}(X) \to f_\theta^{(l+1)}(X) \longrightarrow \cdots$

## Analyzing Induced Dynamical Systems

**NTK** $\cdots \longrightarrow \Theta^{(l)} \longrightarrow \Theta^{(l+1)} \longrightarrow \cdots \longrightarrow \Theta^*$

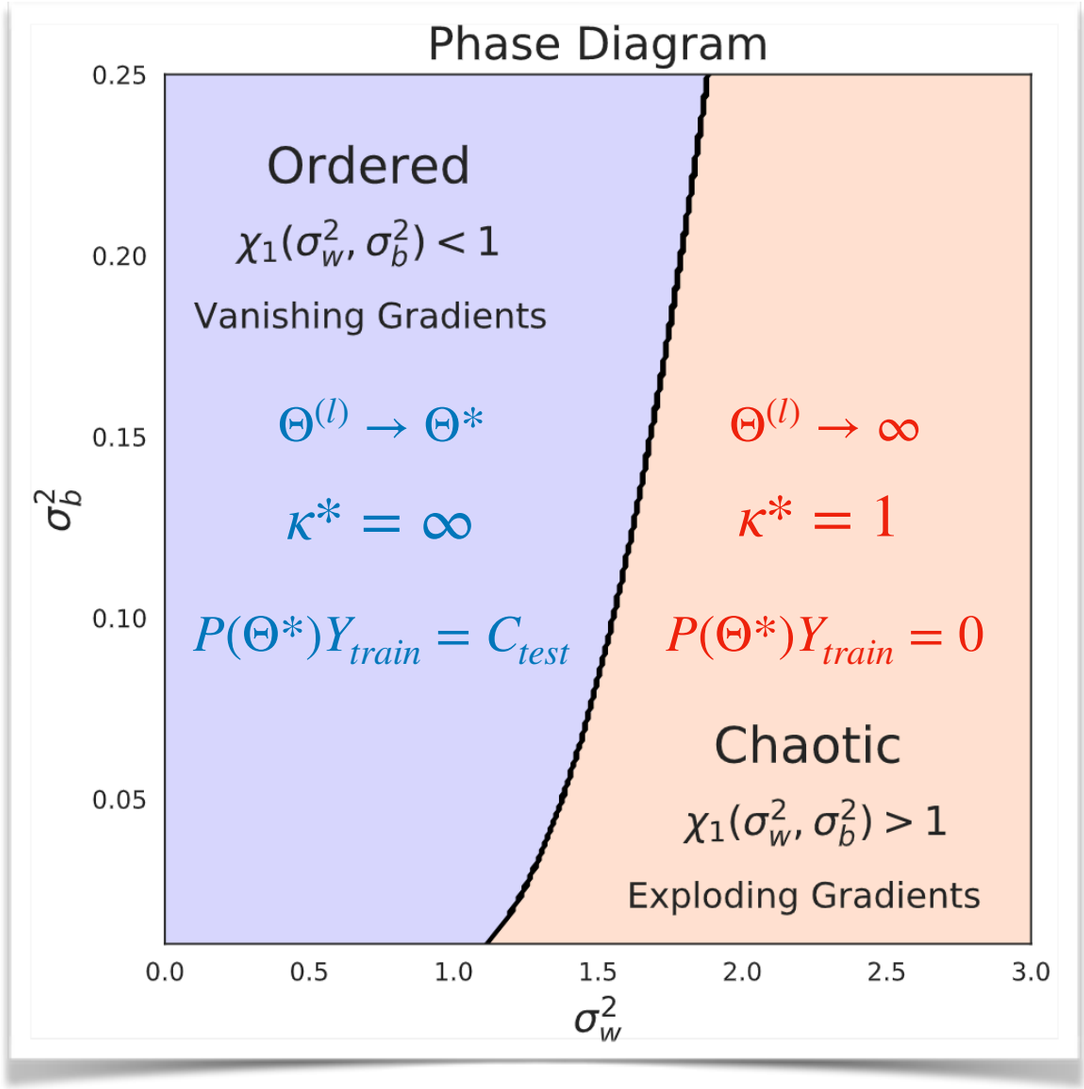
**Condition Number** $\cdots \longrightarrow \kappa^{(l)} \longrightarrow \kappa^{(l+1)} \longrightarrow \cdots \longrightarrow \kappa^*$

**Mean Prediction** $\cdots \to P(\Theta^{(l)})Y_{\text{train}} \to P(\Theta^{(l+1)})Y_{\text{train}} \to \cdots \to P(\Theta^*)Y_{\text{train}}$

# Convergence of NTK and Phase Diagram

Convergence of $\Theta^{(l)}$ is determined by a bivariate function $\chi_1$ defined on the $(\sigma_w^2, \sigma_b^2)$-plane

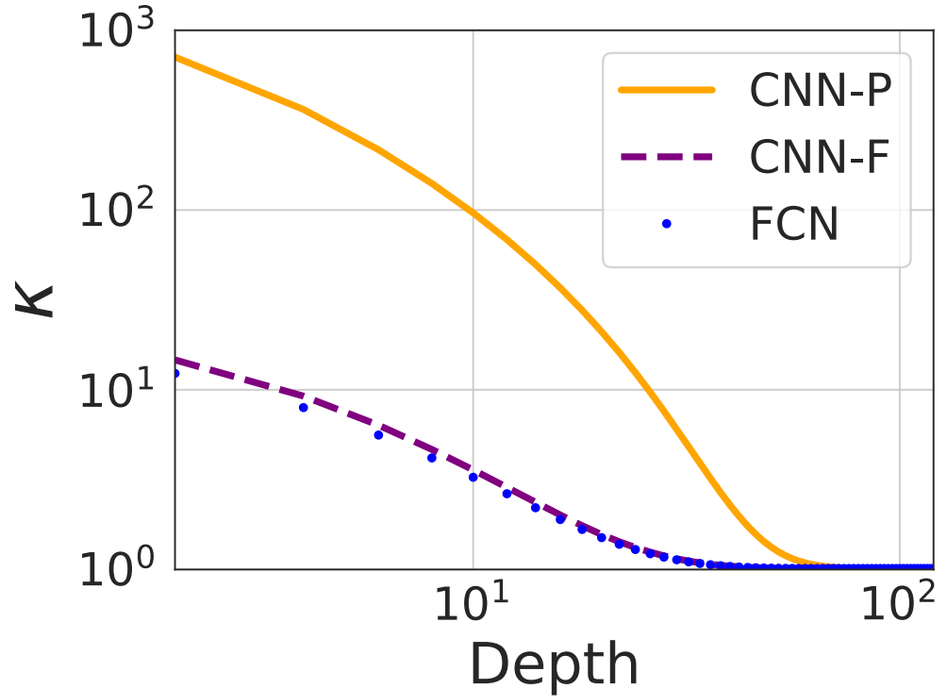
- Ordered Phase $\chi_1 < 1$:
  - $\Theta^{(l)} \to \Theta^* = C\mathbf{1}\mathbf{1}^T$
  - $\kappa^{(l)} \to \infty$
  - $P(\Theta^{(l)})Y_{train} \to C_{test}$
- Chaotic Phase $\chi_1 > 1$:
  - $\Theta^{(l)} \to \infty$
  - $\kappa^{(l)} \to 1$
  - $P(\Theta^{(l)})Y_{train} \to 0$



# Chaotic Phase $\chi_1 > 1$

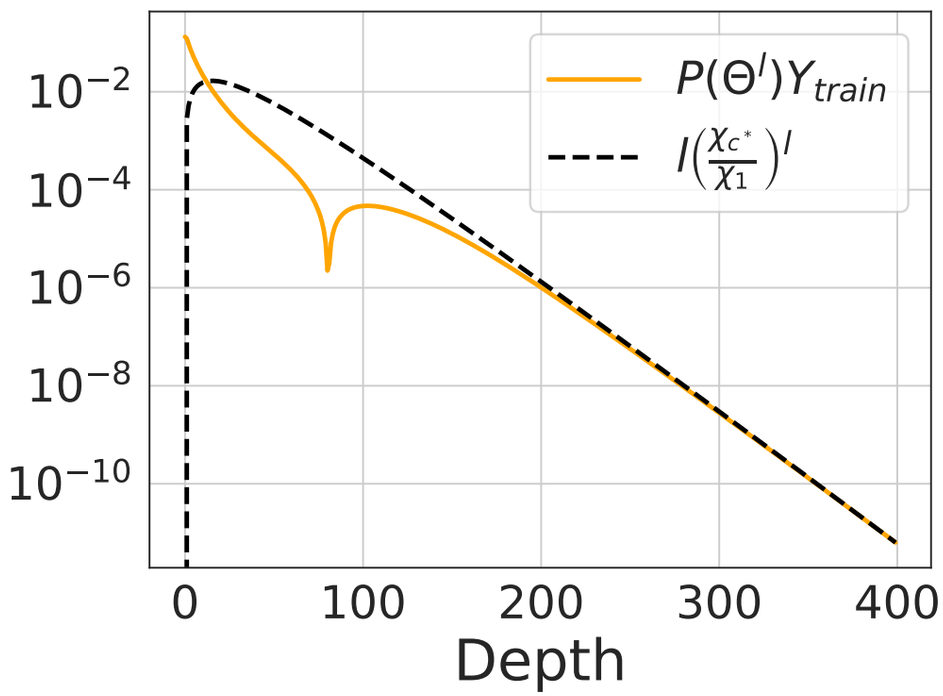
- Entries dynamics of NTK

$$\Theta^{(l)}(x,x) \propto \chi_1^l \to \infty$$
$$\Theta^{(l)}(x,x') \to p^*_{ab} < \infty$$
$$\Theta^{(l)} \approx \chi_1^l \mathrm{Id}$$




- Trainability / Generalization metrics

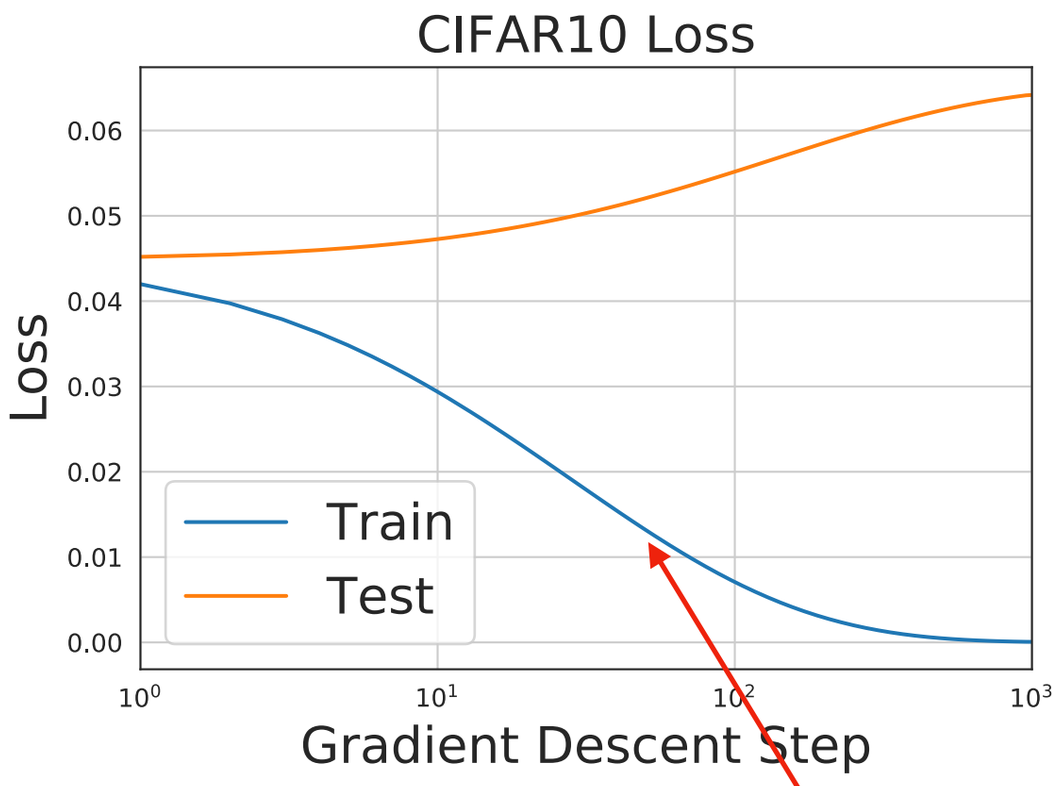
$$\kappa^{(l)} = 1 + \mathcal{O}\left(\chi_1^{-l}\right)$$
$$P(\Theta^{(l)})Y_{\text{train}} = \mathcal{O}\left(l\left(\frac{\chi_{c^*}}{\chi_1}\right)^l\right)$$

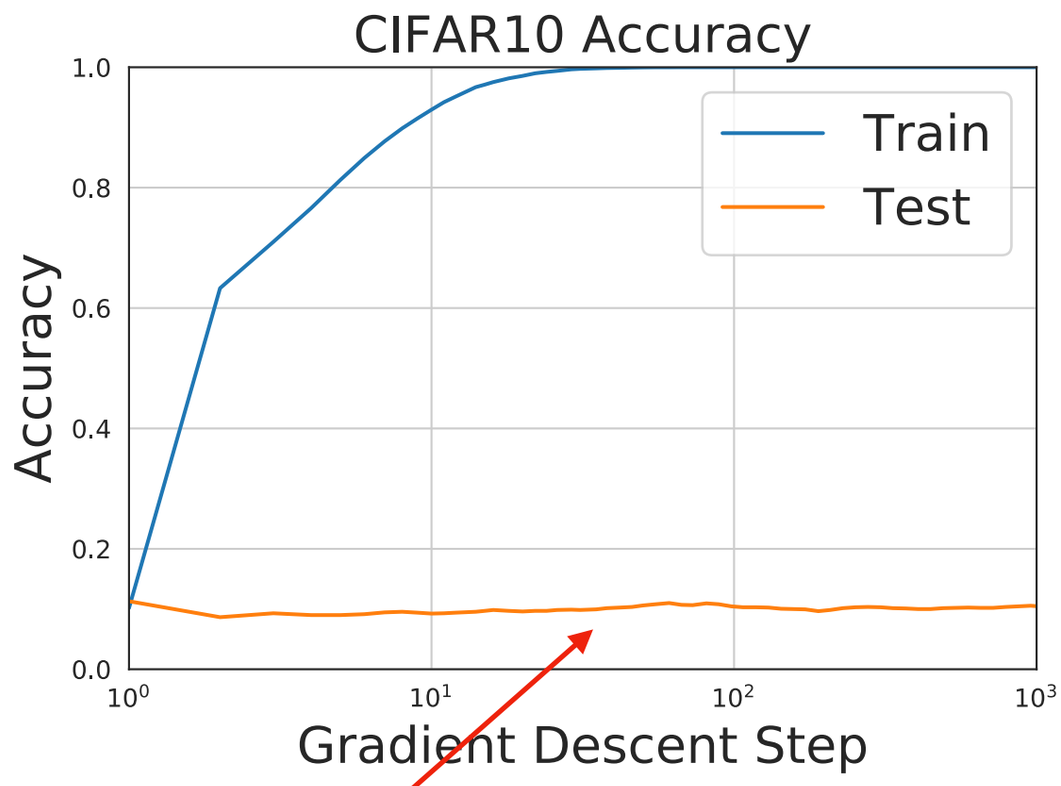



**Easy to Train, but not Generalizable**

# Chaotic Phase / Memorization

- 10k/2k Training / Test, CIFAR10 (10 classes)
- Full Batch + Gradient Descent
- $\sigma_w^2 = 25,\ \sigma_b^2 = 0,\ l = 8$




**Easy to Train, but Not Generalize**

# Ordered Phase $\chi_1 < 1$

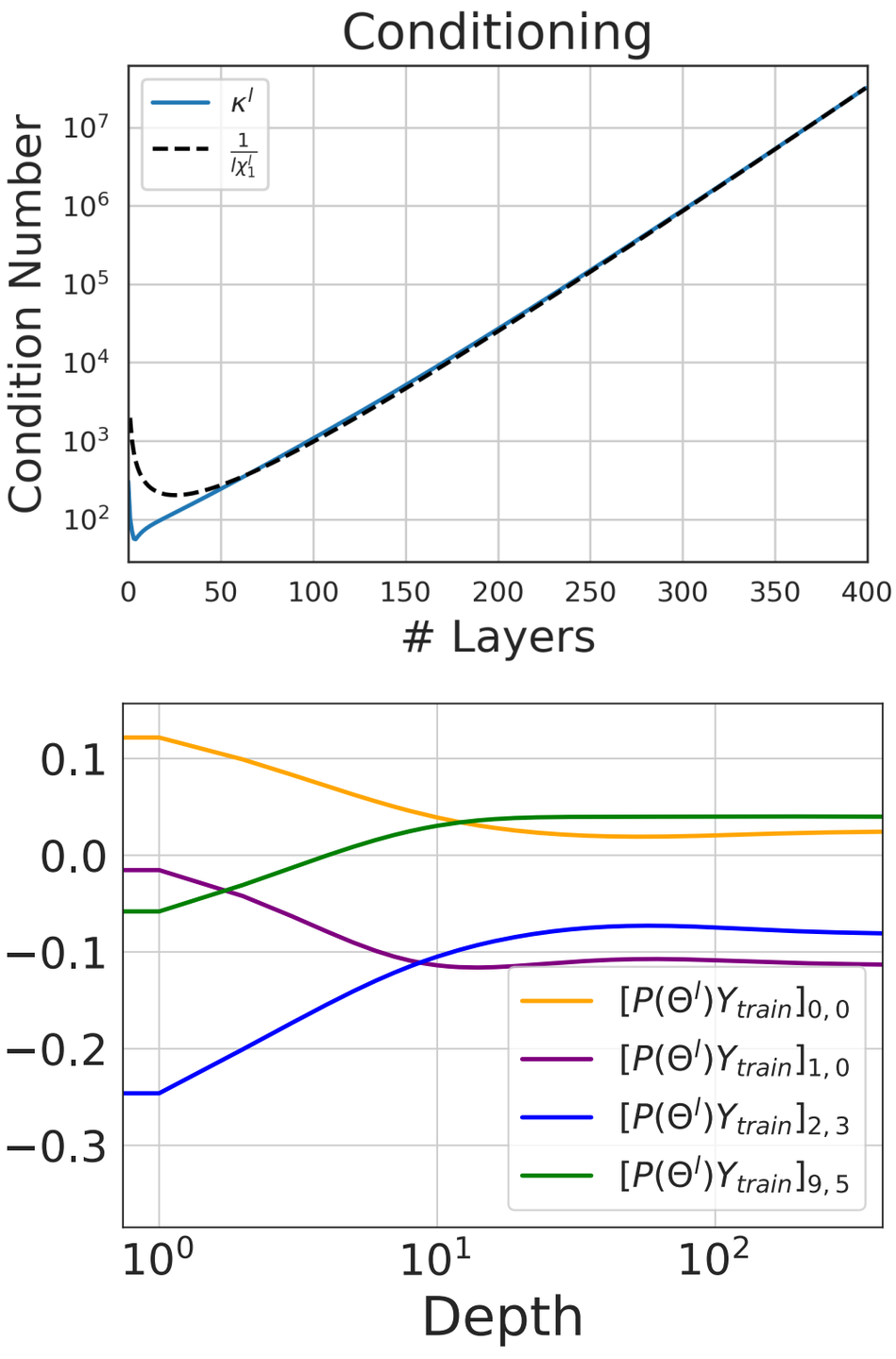
- Entries of the NTK

$$|\Theta^{(l)}(x,x) - p^*| \propto \chi_1^l$$
$$|\Theta^{(l)}(x,x') - p^*| \propto l\chi_1^l$$
$$\Theta^{(l)} = p^* \mathbf{1}\mathbf{1}^T + \mathcal{O}(l\chi_1^l)$$

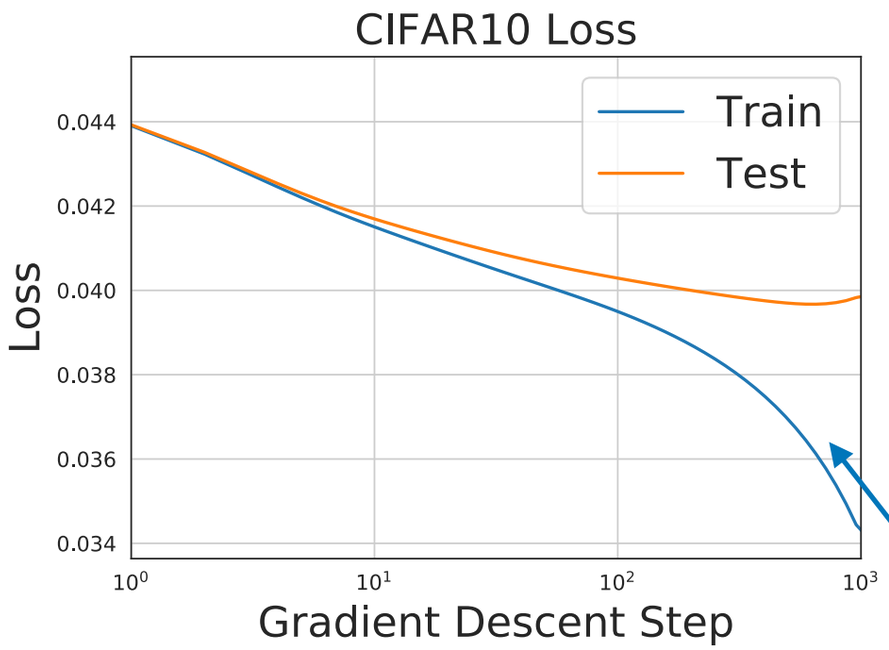
- Trainability / Generalization metrics

$$\kappa^{(l)} \gtrsim \chi_1^{-l}/l$$
$$P(\Theta^{(l)})Y_{\text{train}} \to C_{test}$$

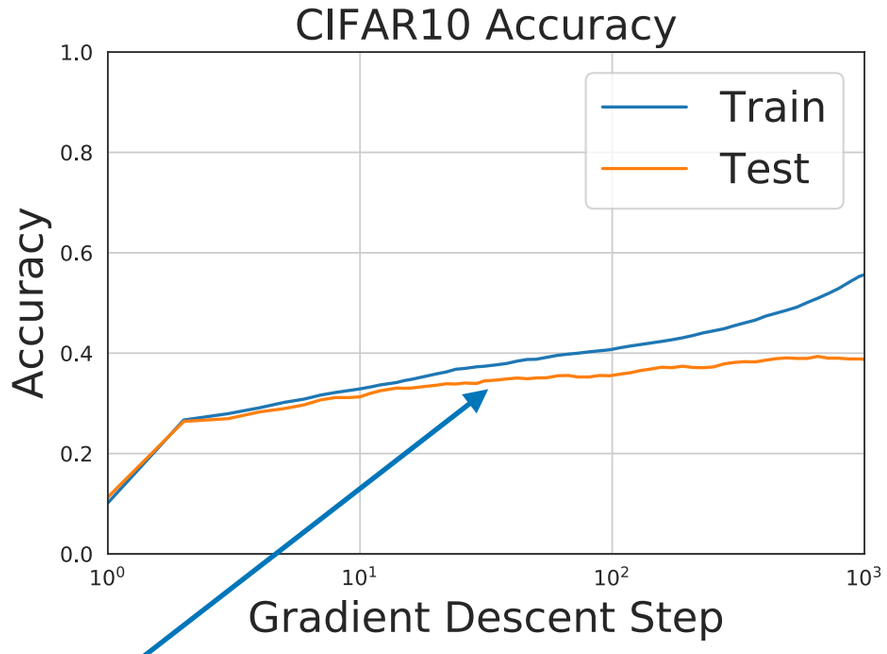
**Difficult to Train, Generalizable**



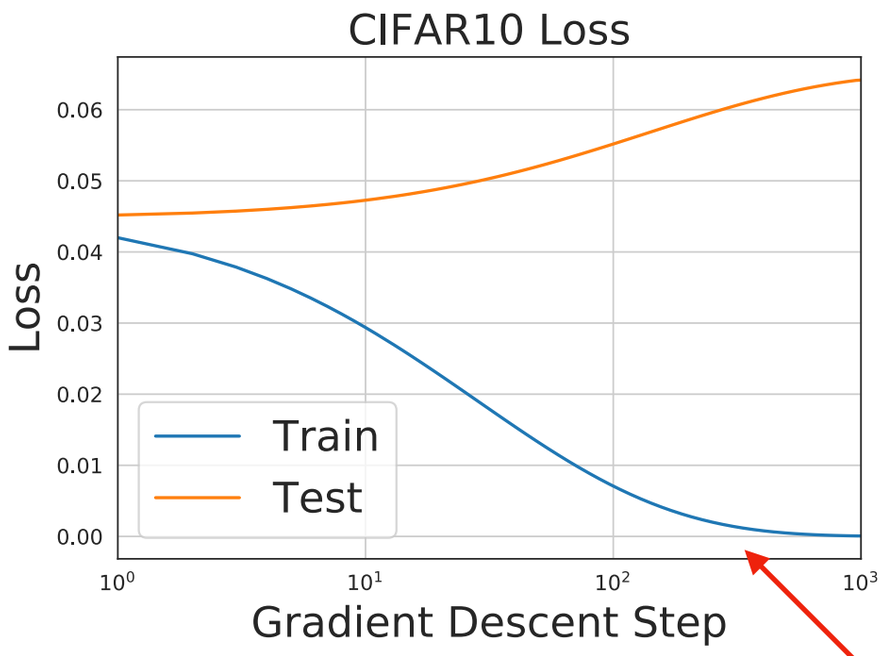
# Ordered Phase / Generalization



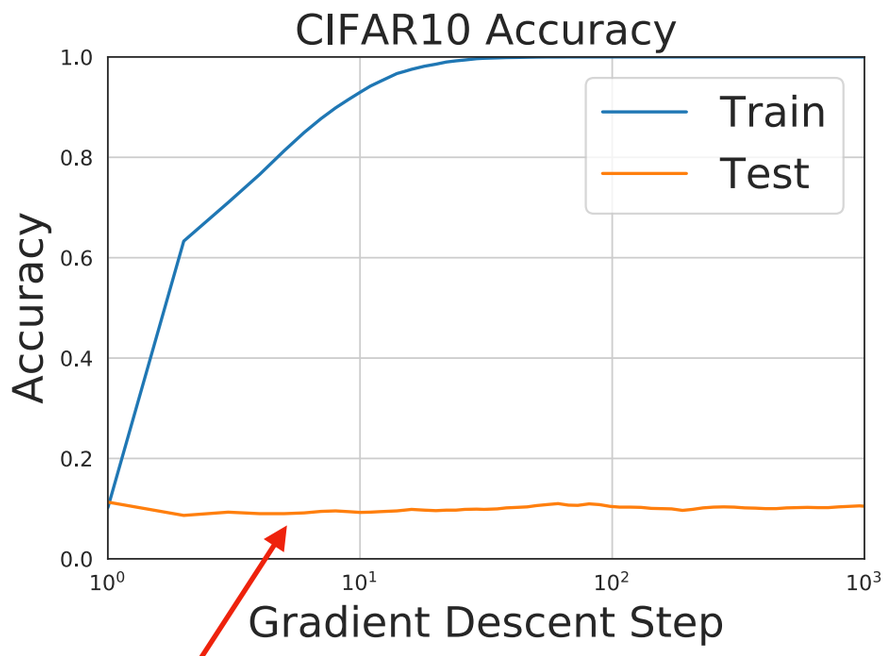

• $\sigma_w^2 = 0.5$




## Difficult to Train, Generalizable




• $\sigma_w^2 = 25$




## Easy to Train, but Not Generalize

# Summary

- A tradeoff between trainability and generalization for deep and wide networks
  - Fast training + memorization (e.g. Chaotic Phase)
  - Slow training + generalizable (e.g. Ordered Phase)
- More results
  - Pooling, Dropout, Skip Connection, LayerNorm, etc.
  - Conjugate Kernels

Colab Tutorial