More Than a Toy: Random Matrix Models Predict How Real-World Neural Representations Generalize

Alexander Wei

ICML 2022

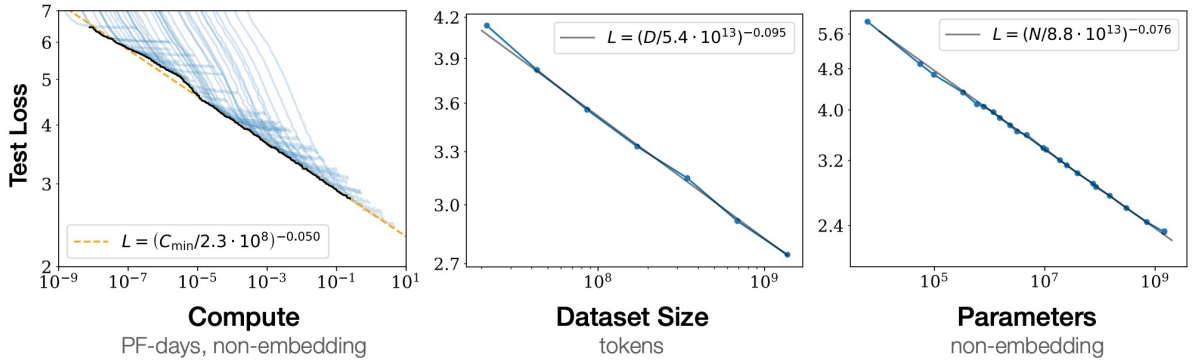
Wei Hu

Jacob Steinhardt

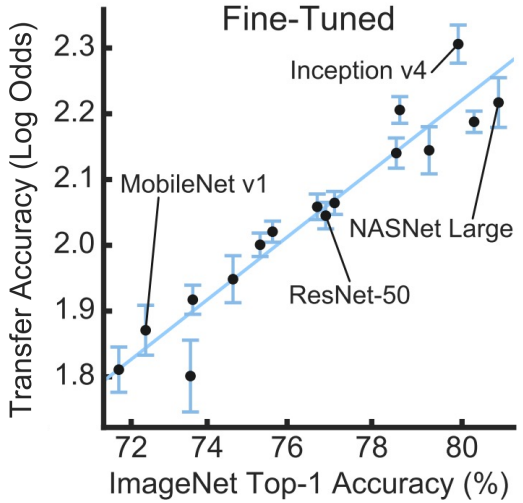
Berkeley
UNIVERSITY OF CALIFORNIA

A compelling theory of generalization should…

1. Accurately predict **empirical phenomena**
   - E.g., **scaling laws**, **pretraining** >> random initialization



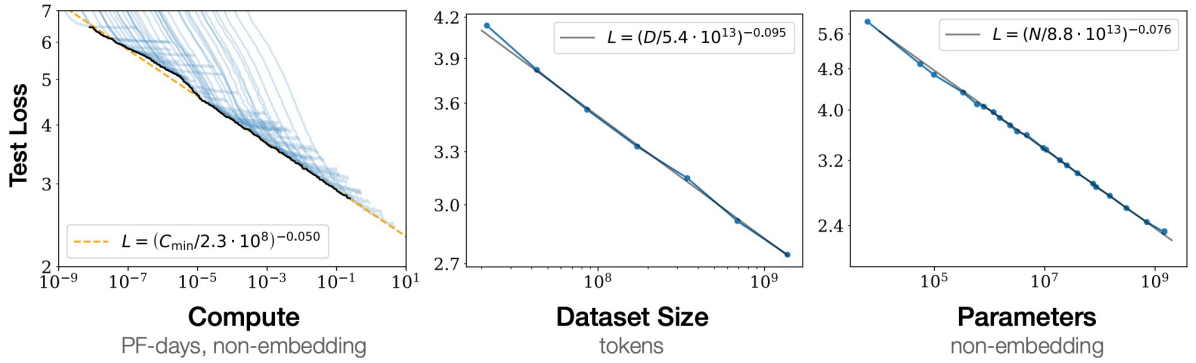
[Kaplan et al., 2020]



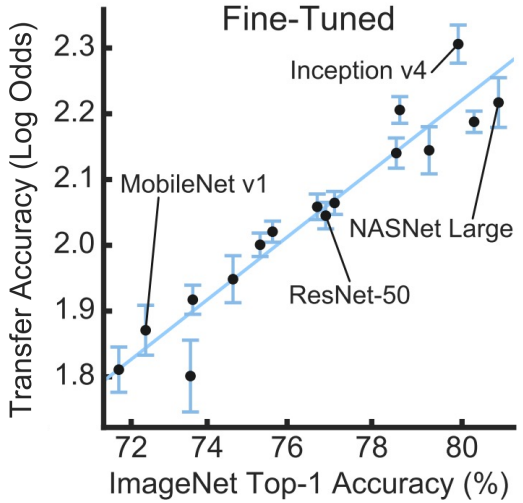
[Kornblith et al., CVPR 2019]

A compelling theory of generalization should…

1. Accurately predict **empirical phenomena**
   - E.g., **scaling laws**, **pretraining** >> random initialization



[Kaplan et al., 2020]

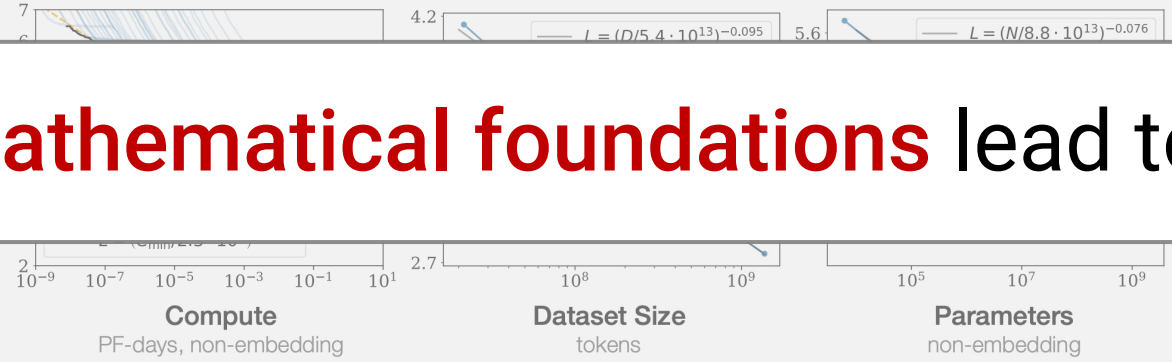


[Kornblith et al., CVPR 2019]

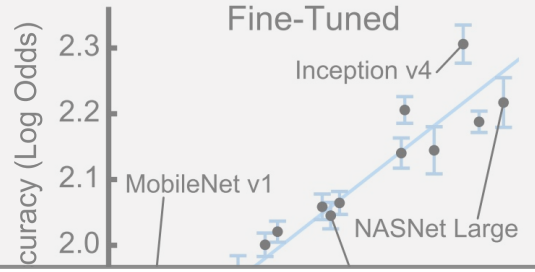
2. Precisely explain why these phenomena arise
   - Test error has no explanatory power—not a theory!

A compelling theory of generalization should…

1. Accurately predict **empirical phenomena**
   - E.g., **scaling laws**, **pretraining** >> random initialization



[Kaplan et al., 2020]



[Kornblith et al., CVPR 2019]

What **mathematical foundations** lead to such a theory?

2. Precisely explain why these phenomena arise
   - Test error has no explanatory power—not a theory!

NTK regression as a testbed for theories

NN-induced linear regression (ResNet NTK features on image data):

- Achieves comparable performance to NNs

Configuration	*Finetuning*	eNTK	Last layer
CIFAR-10 / ResNet-18	4.3%	6.7%	14.0%
CIFAR-100 / ResNet-34	15.9%	19.0%	33.9%
Flowers-102 / ResNet-50	5.6%	7.0%	9.7%
Food-101 / ResNet-101	15.3%	21.3%	33.7%

NTK regression as a testbed for theories

NN-induced linear regression (ResNet NTK features on image data):

- Achieves comparable performance to NNs

Configuration	*Finetuning*	eNTK	Last layer
CIFAR-10 / ResNet-18	4.3%	6.7%	14.0%
CIFAR-100 / ResNet-34	15.9%	19.0%	33.9%
Flowers-102 / ResNet-50	5.6%	7.0%	9.7%
Food-101 / ResNet-101	15.3%	21.3%	33.7%

NTK regression as a testbed for theories

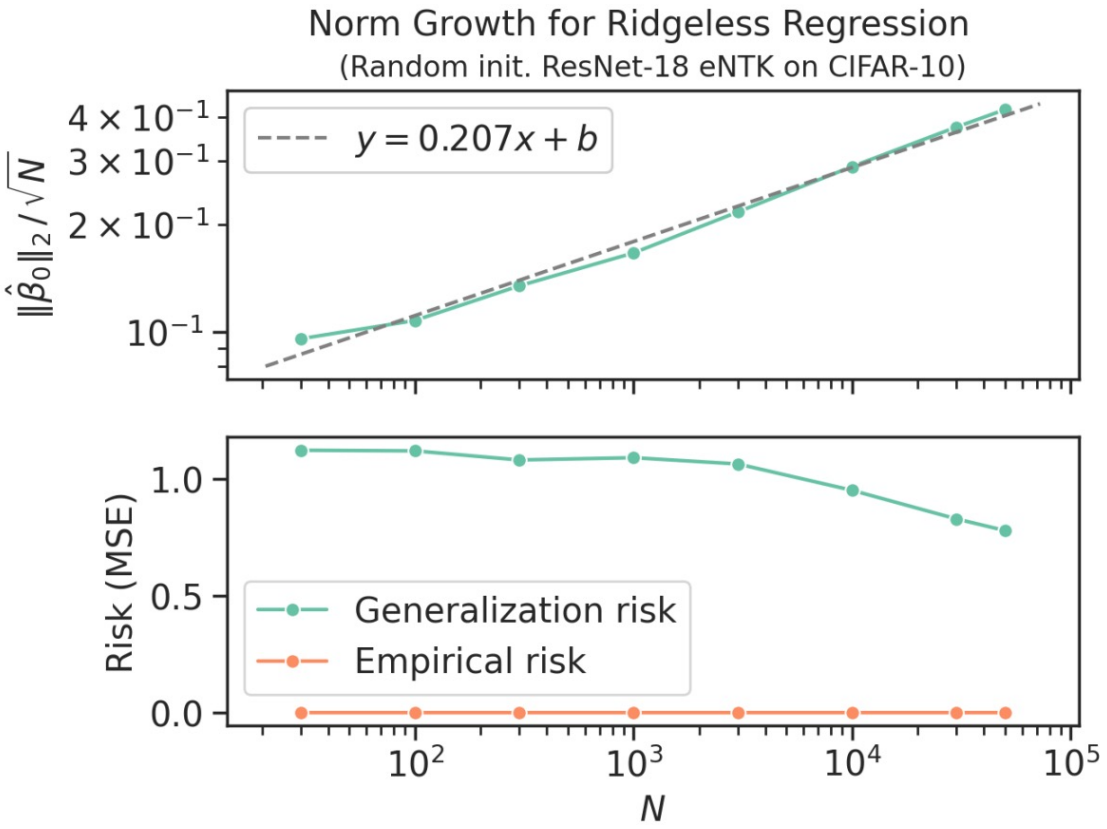
NN-induced linear regression (ResNet NTK features on image data):

- Achieves comparable performance to NNs

Configuration	*Finetuning*	eNTK	Last layer
CIFAR-10 / ResNet-18	4.3%	6.7%	14.0%
CIFAR-100 / ResNet-34	15.9%	19.0%	33.9%
Flowers-102 / ResNet-50	5.6%	7.0%	9.7%
Food-101 / ResNet-101	15.3%	21.3%	33.7%

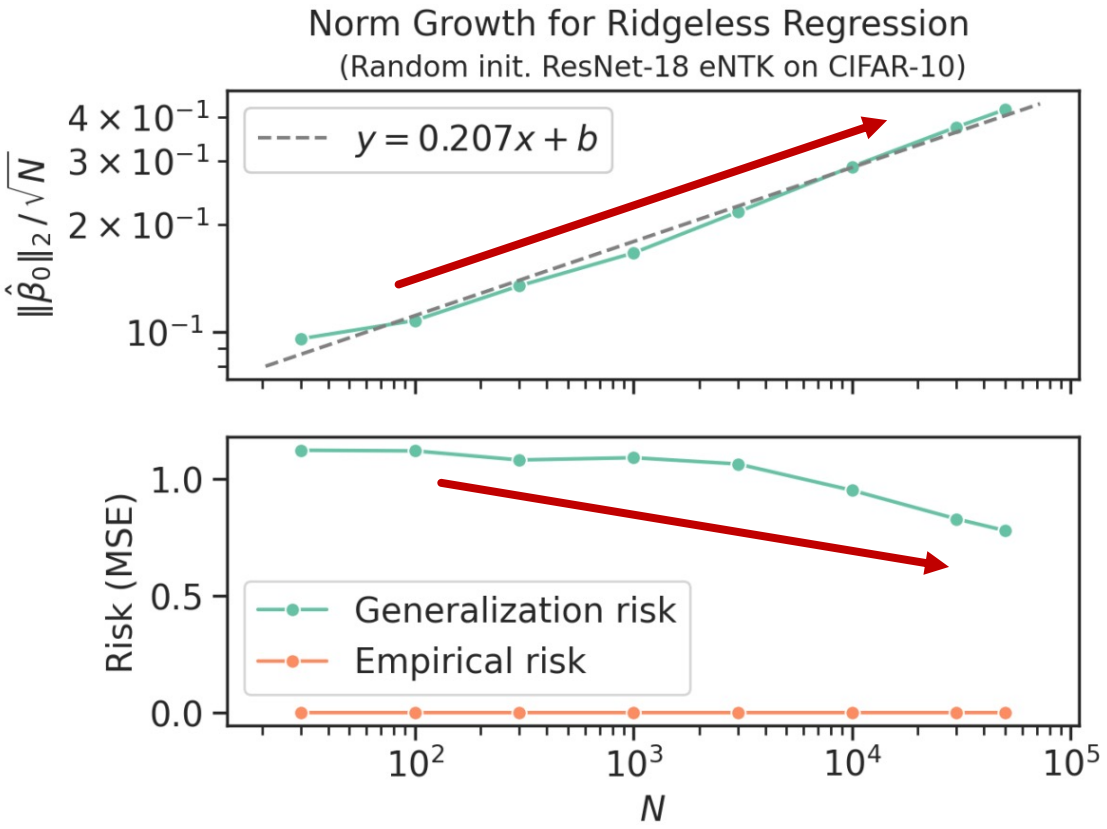
- Exhibits many of the **same empirical phenomena**
  - Power-law scaling; effect of pretraining

Empirical obstructions for classical theories



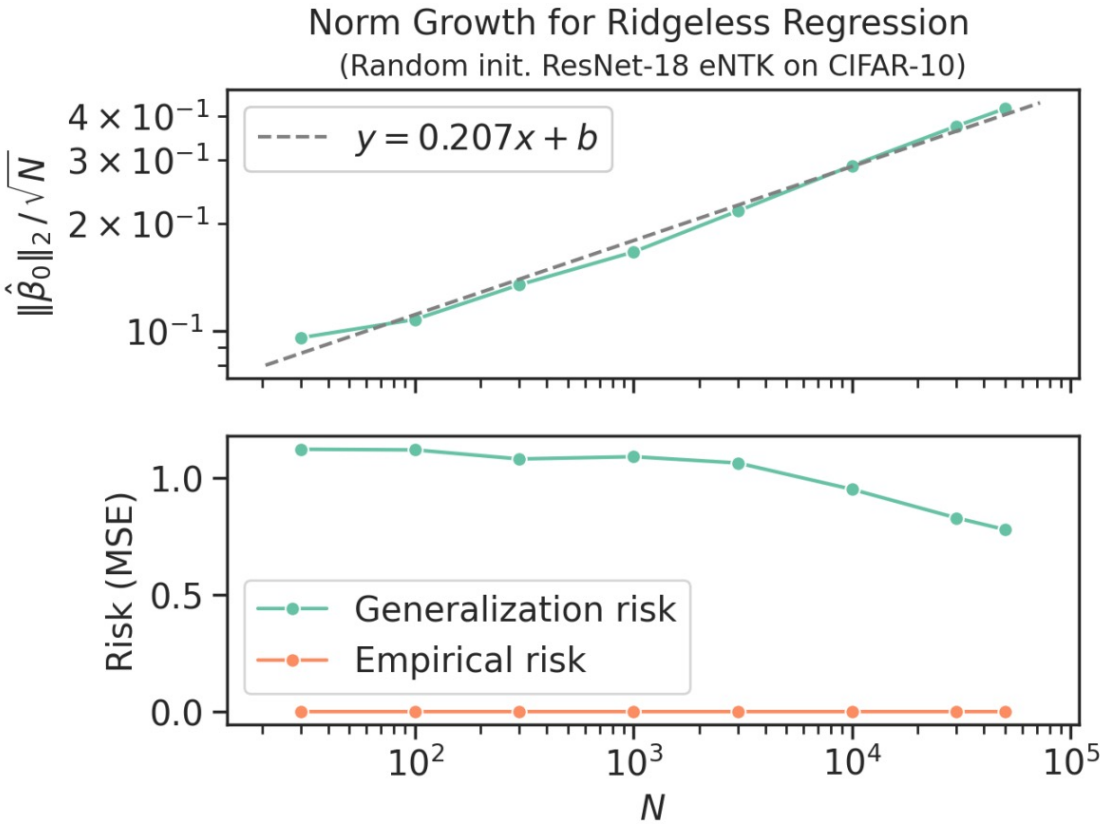
Norm-based generalization bounds: **predict wrong sign** as N increases!

Empirical obstructions for classical theories

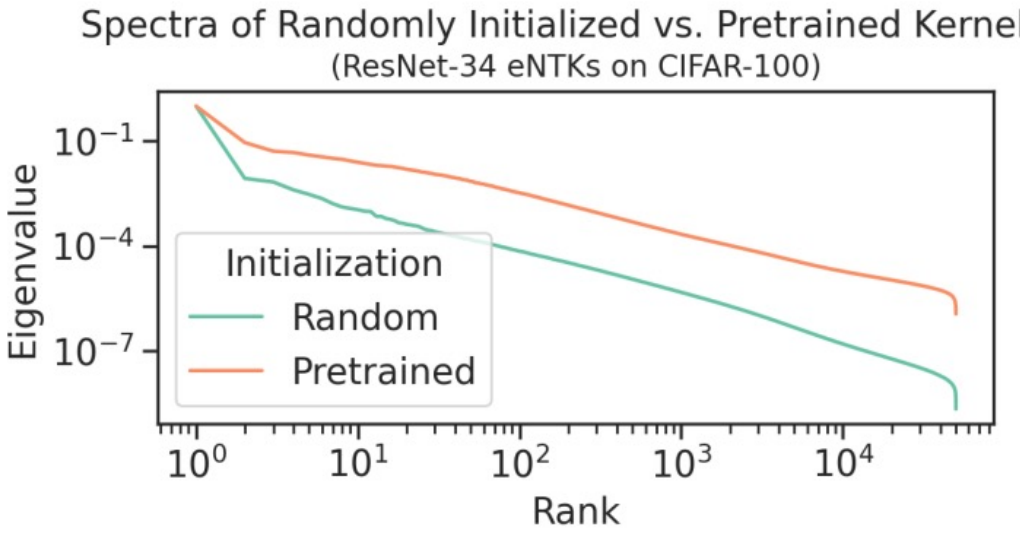


Norm-based generalization bounds: **predict wrong sign** as N increases!

Empirical obstructions for classical theories

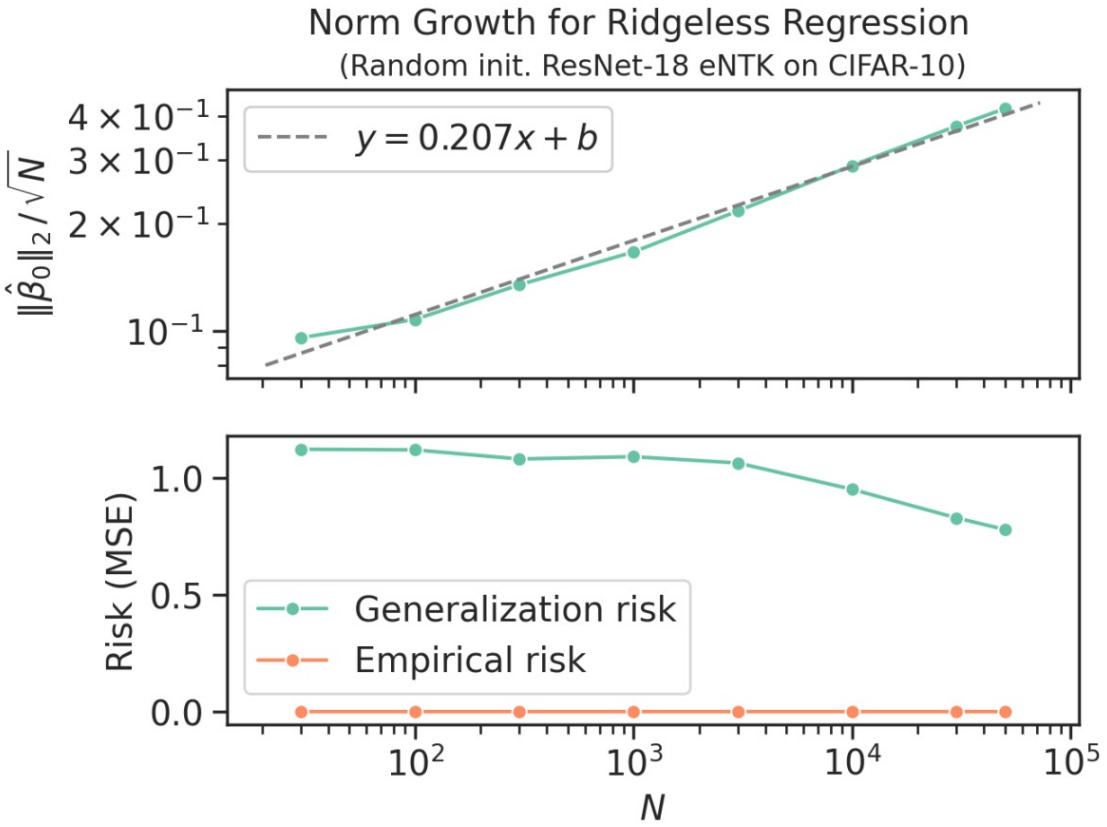


Norm-based generalization bounds: **predict wrong sign** as N increases!

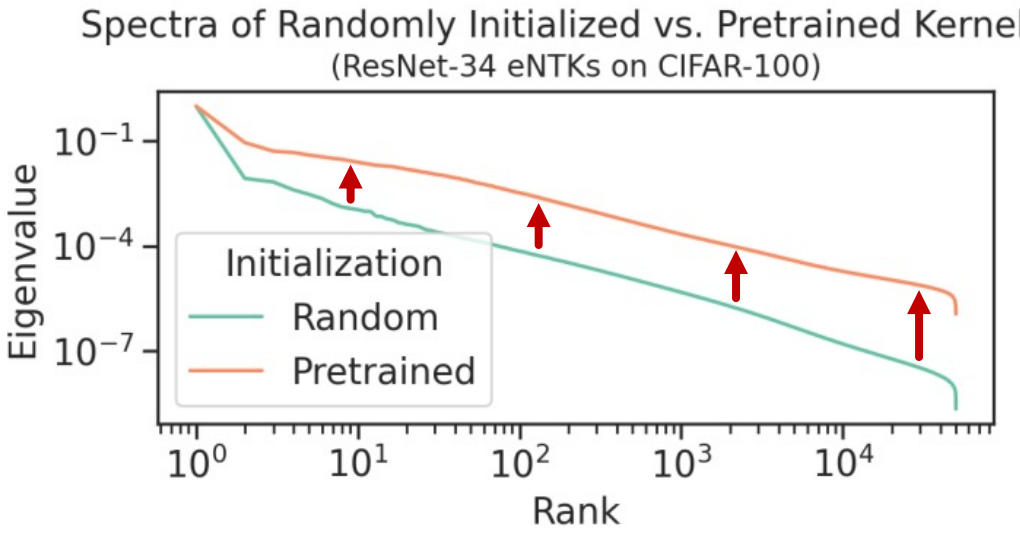


Spectral generalization bounds: randomly initialized to pretrained representations → **increased effective dimension**!

Empirical obstructions for classical theories



Norm-based generalization bounds: **predict wrong sign** as N increases!



Spectral generalization bounds: randomly initialized to pretrained representations → **increased effective dimension**!

The random matrix theory perspective

We prove that the **GCV estimator** [Craven and Wahba, 1978] predicts linear regression generalization under a random matrix hypothesis …

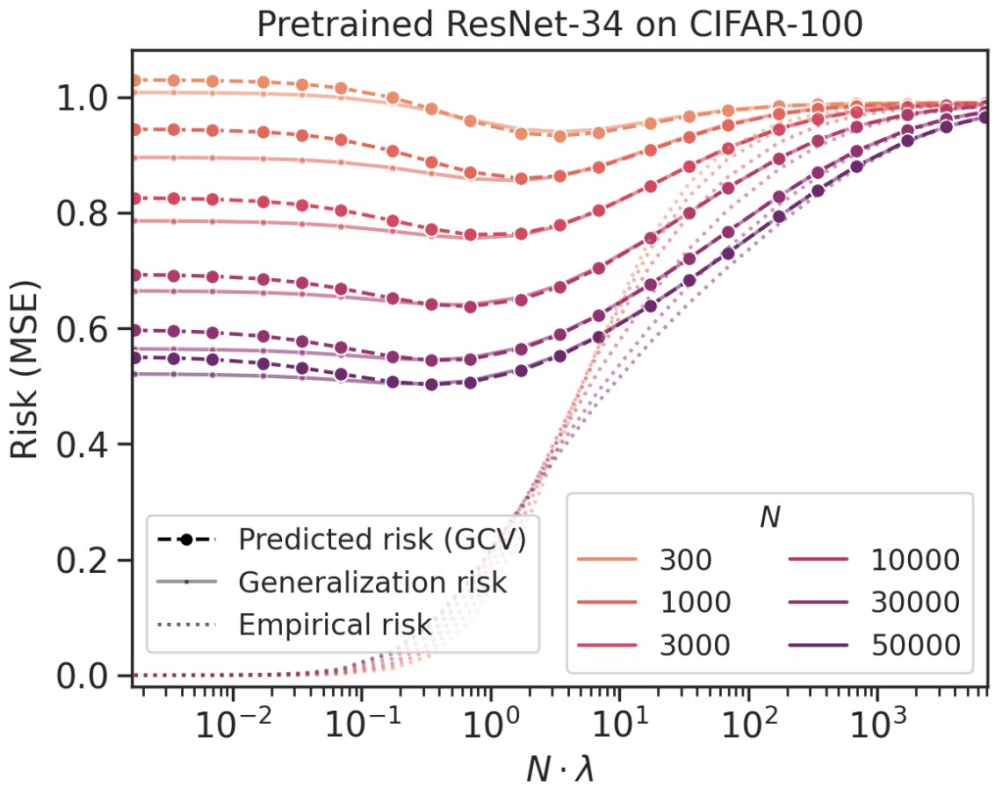
$$\mathrm{GCV}_\lambda := \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\lambda}{\lambda + \hat{\lambda}_i} \right)^{-2} \mathcal{R}_{\mathrm{empirical}}(\hat{\beta}_\lambda)$$

The random matrix theory perspective

We prove that the **GCV estimator** [Craven and Wahba, 1978] predicts linear regression generalization under a random matrix hypothesis …

$$\mathrm{GCV}_\lambda := \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\lambda}{\lambda + \hat{\lambda}_i} \right)^{-2} \mathcal{R}_{\mathrm{empirical}}(\hat{\beta}_\lambda)$$

… and find it to be empirically accurate 🎉



Empirical phenomena via random matrices

We predict **scaling law rates** …

- Verify **eigendecay** exponent + **alignment** exponent ≈ scaling exponent

[Cui et al., NeurIPS 2021]

between eigenvectors and ground truth

Empirical phenomena via random matrices

We predict **scaling law rates** …

- Verify **eigendecay** exponent + **alignment** exponent ≈ scaling exponent

[Cui et al., NeurIPS 2021]

between eigenvectors and ground truth

… and investigate **the role of pretraining** in generalization

- Better **alignment** prevails over slower **eigendecay** / high effective dim

Summary

What makes a compelling scientific theory of high-dimensional models?

- Accuracy in qualitative phenomena, precision in explanations

We find random matrix theory **predicts empirical phenomena** (even when more classical approaches fail)

- Setting: linear regression + ResNet NTK features + image data

Apply toward understanding **scaling laws**, **role of pretraining**