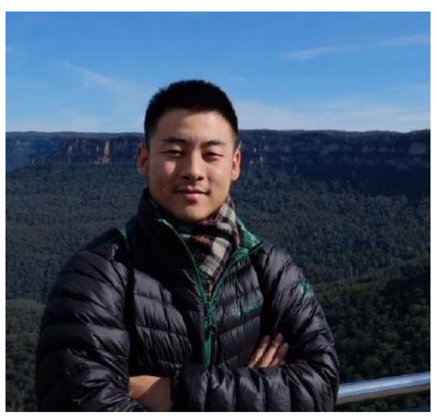
Bias-Variance Tradeoff?

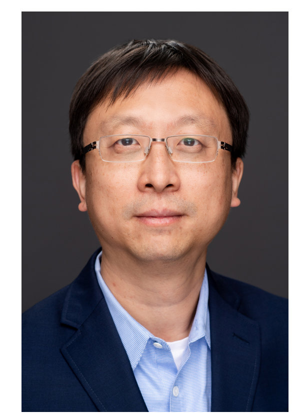
Speaker: Zitong Yang*



*Yaodong Yu**

Chong You

Jacob Steinhardt



Yi Ma

ICML 2020

*equal contribution

https://arxiv.org/pdf/2002.11328.pdf

Outline:

3 minutes overview

1. Bias Variance Tradeoff v.s. Double Descent

2. Our Proposal: Unimodal Variance

3. Measurement: Experimental Set-up

4. Theory: Analysis of Two-Layer Network

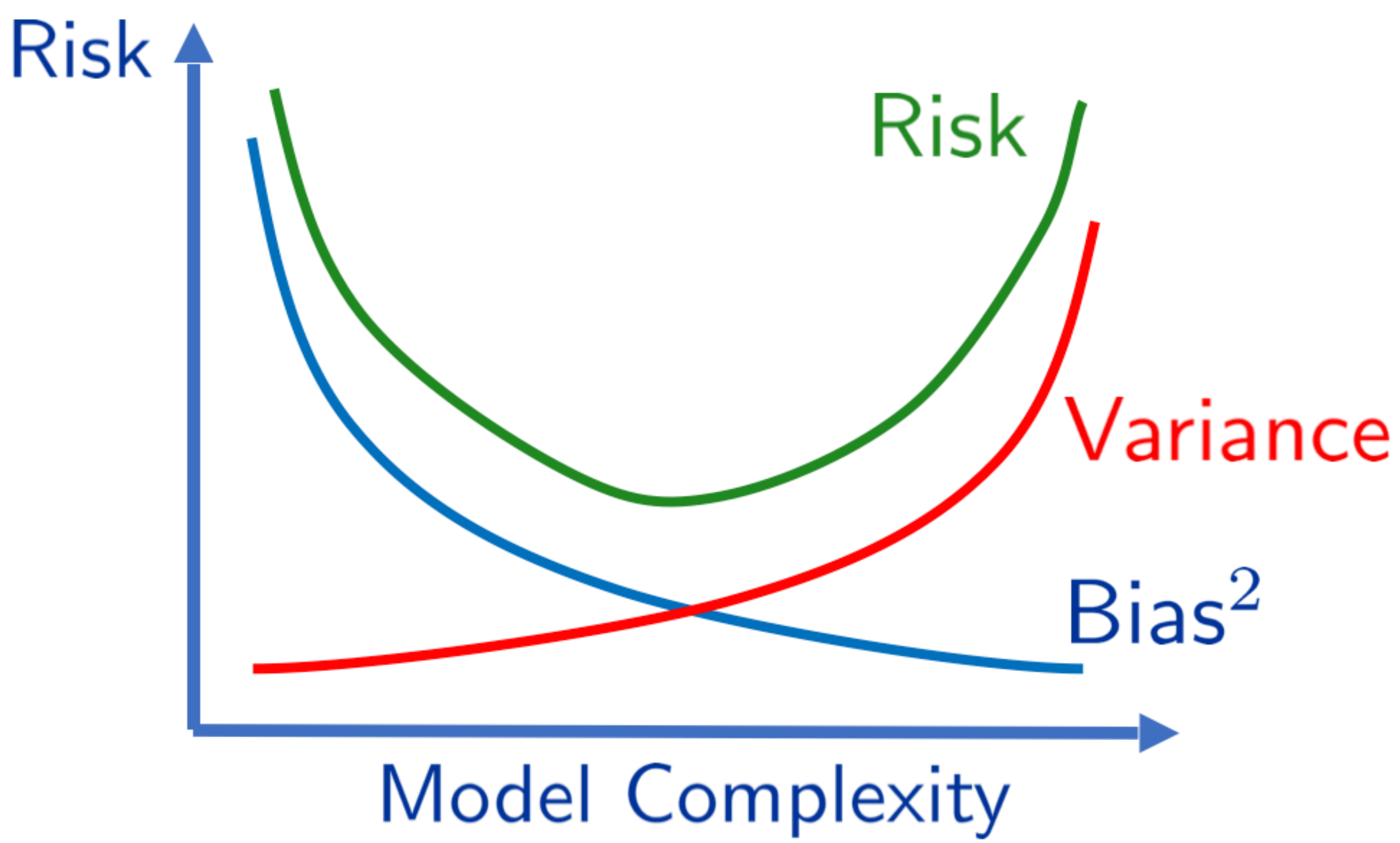
Outline:

1. Bias Variance Tradeoff v.s. Double Descent
2. Our Proposal: Unimodal Variance
3. Measurement: Experimental Set-up
4. Theory: Analysis of Two-Layer Network

Bias Variance Tradeoff v.s. Double Descent

- Recall the classical principle:

$$\text{Risk} = \text{Bias}^2 + \text{Variance}$$



- Decreasing bias, increasing variance
- Minimize Risk by obtaining a balance

Prior Explanation: Double Descent Risk Curve

- Practice of Modern Neural Networks:

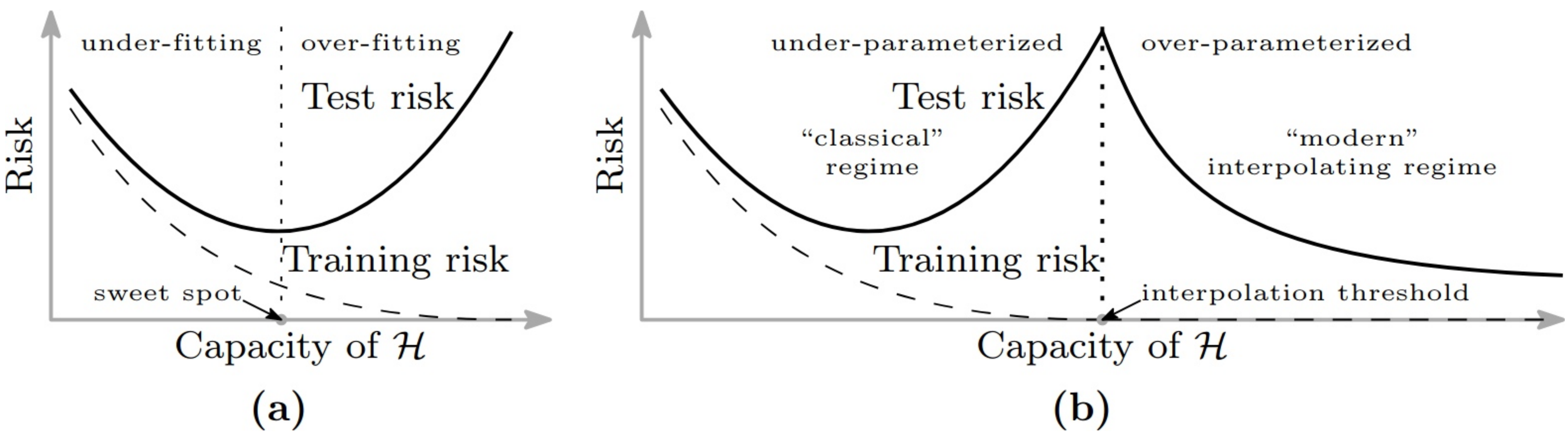
Bigger Models Generalize Better!

Prior Explanation: Double Descent Risk Curve

- Practice of Modern Neural Networks:

Bigger Models Generalize Better!

- Proposed Solution:

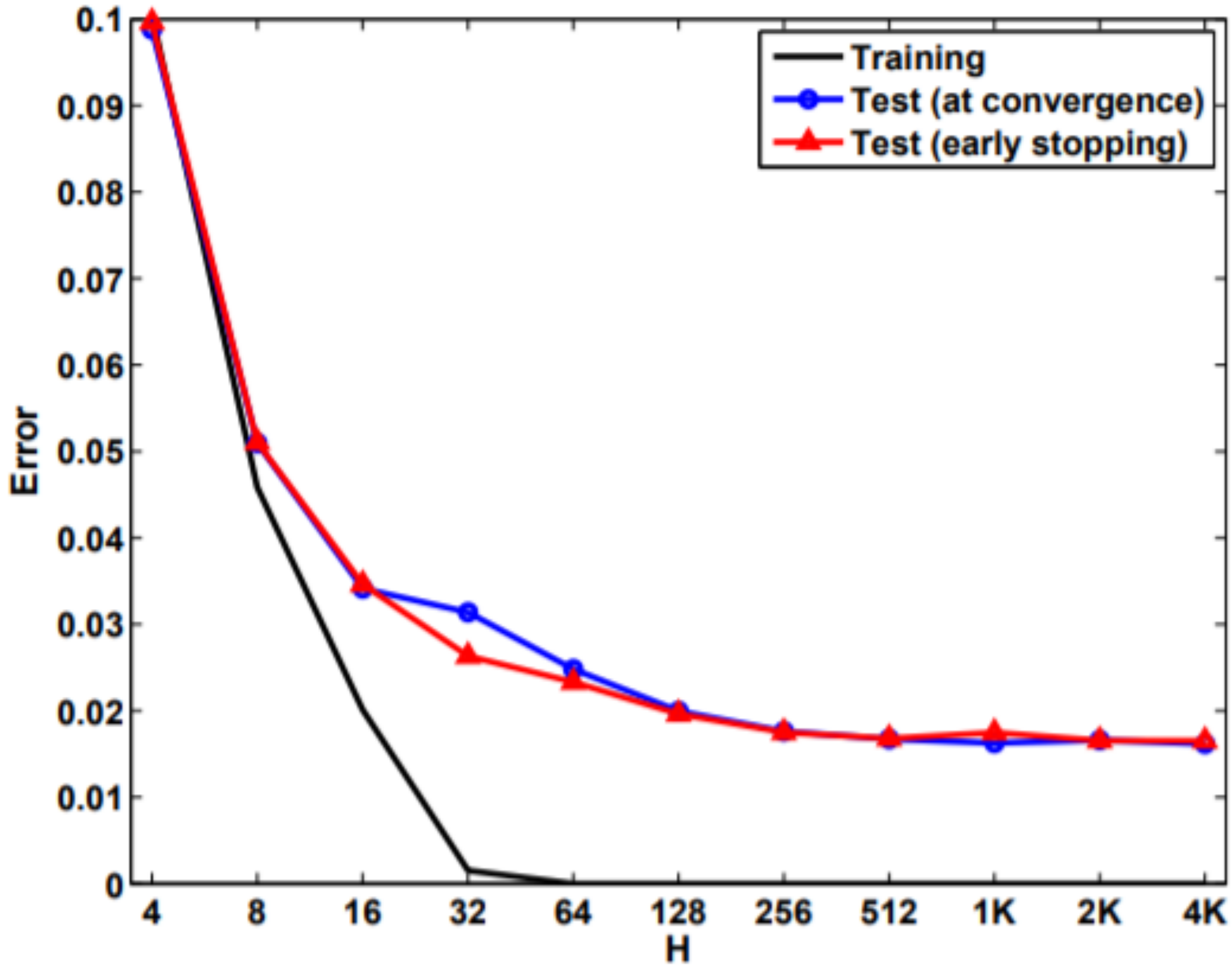


M. Belkin, D. Hsu, S. Ma, S. Mandal (2018)

Prior Explanation: Double Descent Risk Curve

Mysteries:

- More often get monotonically decreasing risk in practice.



B. Neyshabur, R. Tomioka, N. Srebro (2014)

Prior Explanation: Double Descent Risk Curve

Mysteries:

- More often get monotonically decreasing risk in practice.

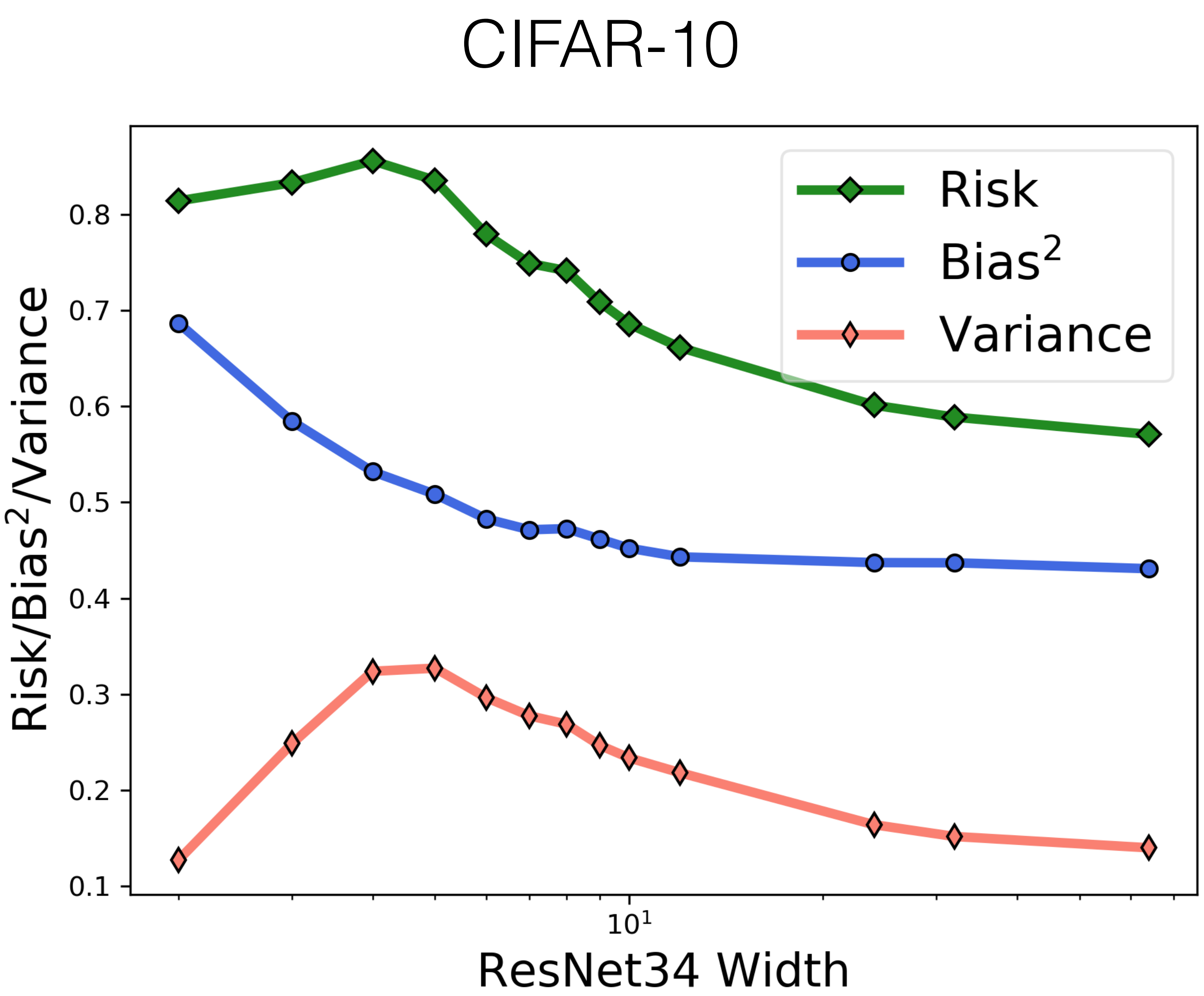
Is there a simpler underlying phenomenon?

Outline:

1. Bias Variance Tradeoff v.s. Double Descent
2. Our Proposal: Unimodal Variance
3. Measurement: Experimental Set-up
4. Theory: Analysis of Two-Layer Network

Our Proposal

- Solution: Revisiting Bias-Variance Tradeoff

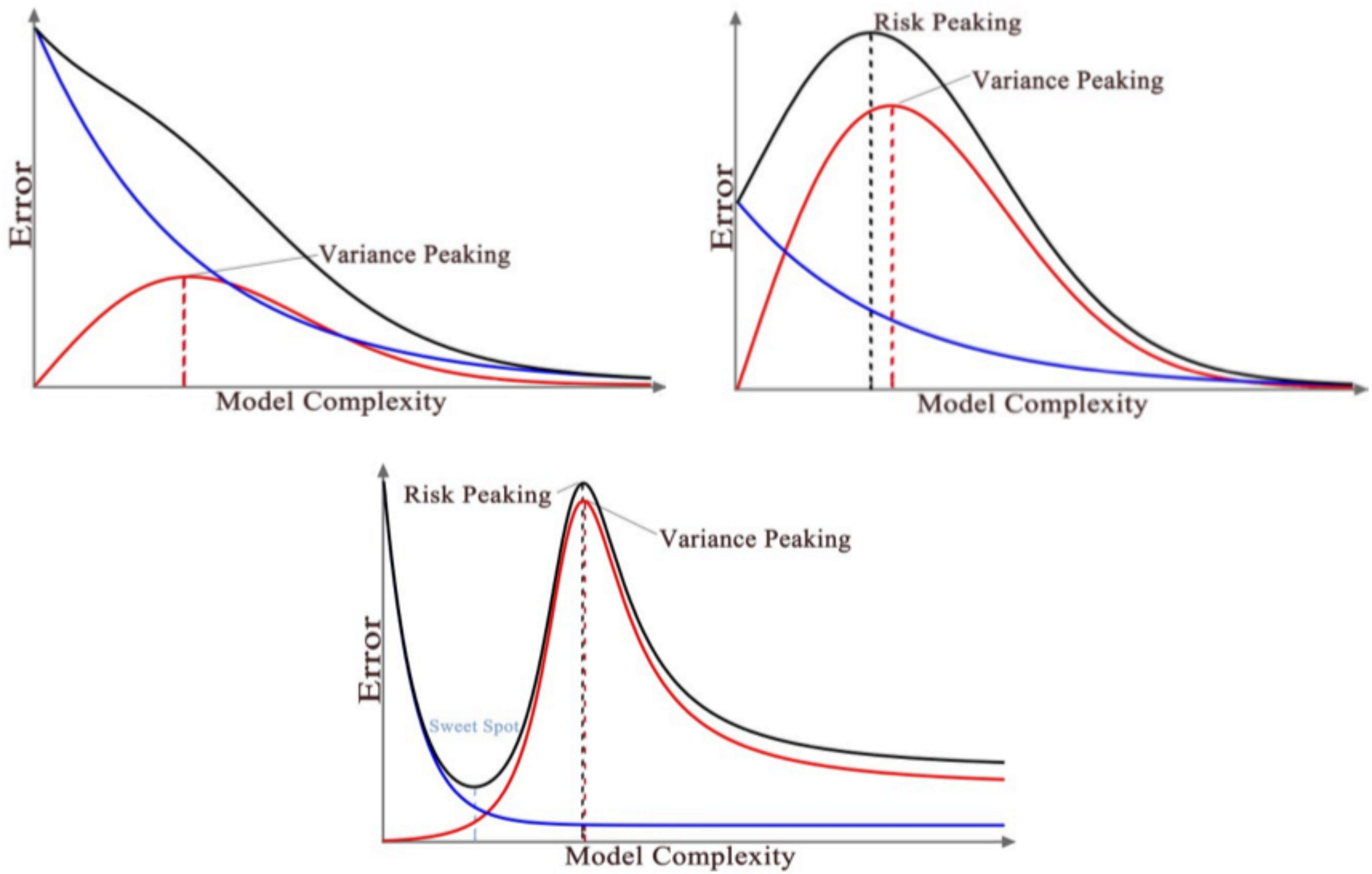


Phenomenon: **monotonic** bias + **unimodal** variance

Z. Yang, Y. Yu, C. You, J. Steinhardt, Y. Ma (2020)

Experiment: Unimodal Variance Curve

- Three Possible Patterns



Z. Yang, Y. Yu, C. You, J. Steinhardt, Y. Ma (2020)

Outline:

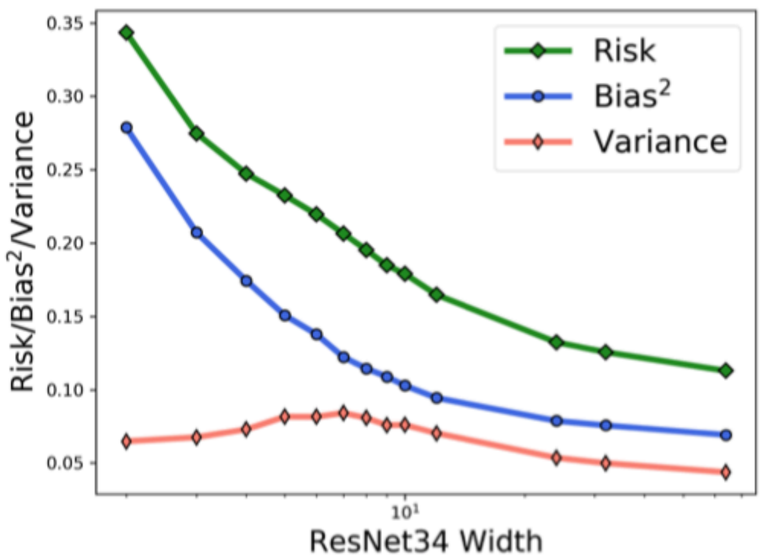
3 minutes overview

1. Bias Variance Tradeoff v.s. Double Descent
2. Our Proposal: Unimodal Variance
3. Measurement: Experimental Set-up
4. Theory: Analysis of Two-Layer Network

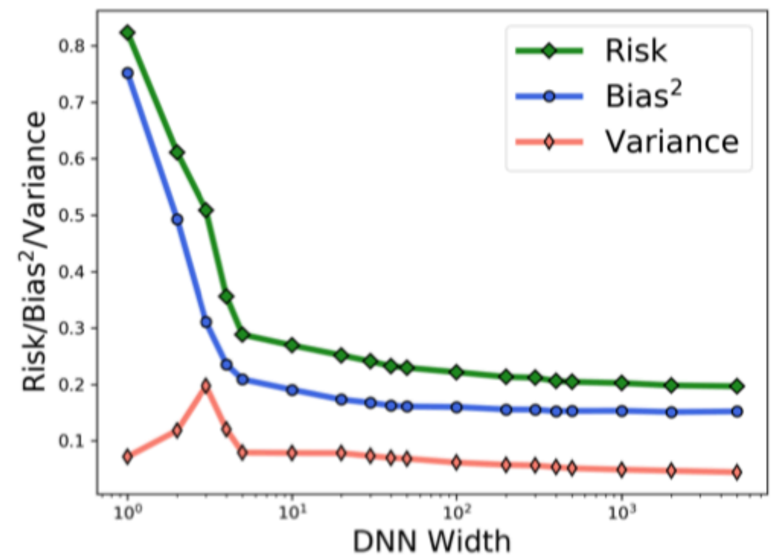
Experiment: Unimodal Variance Curve

- Robustness of the phenomenon

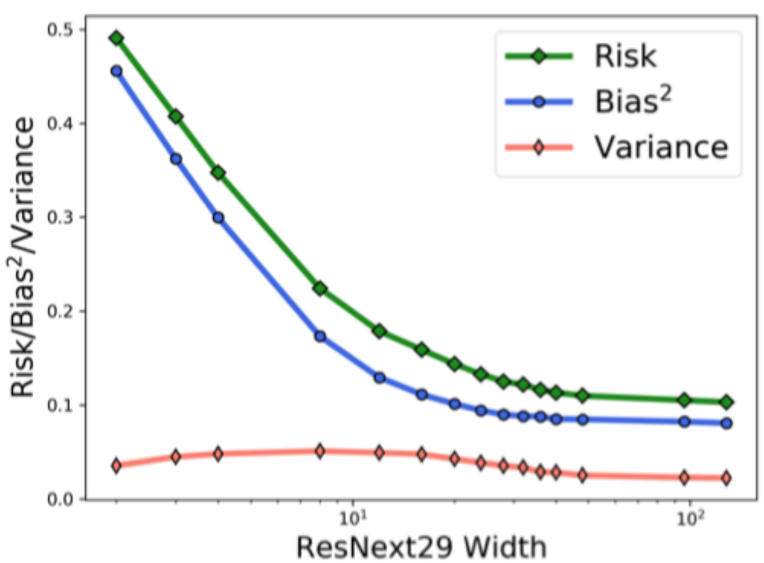
CIFAR-10



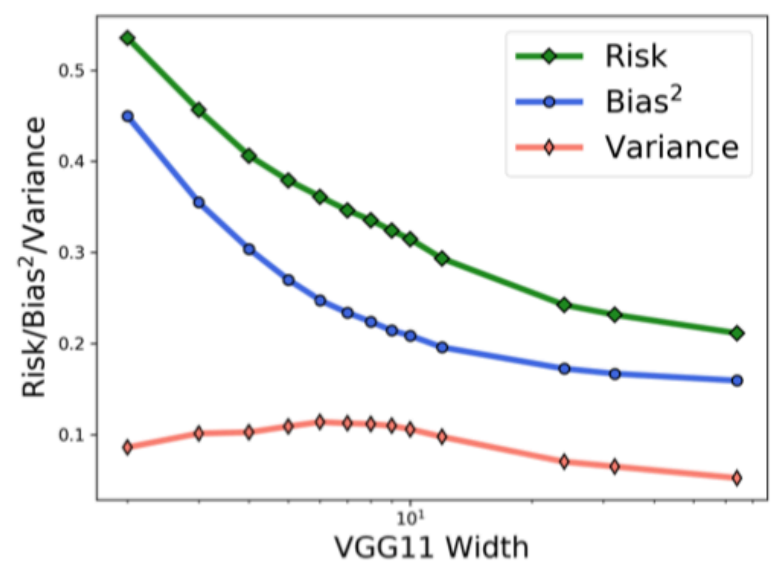
Fashion-MNIST



ResNext29



VGG11



Z. Yang, Y. Yu, C. You, J. Steinhardt, Y. Ma (2020)

Experiment: Unimodal Variance Curve

Computing Bias and Variance:

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y

Experiment: Unimodal Variance Curve

Computing Bias and Variance:

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y

- Recall bias-variance decomposition for mean-squared error:

$$\underbrace{\mathbb{E}_{\mathcal{D}}[(y - f(x))^2]}_{\text{MSE}} = \underbrace{(y - \mathbb{E}_{\mathcal{D}}[f(x)])^2}_{\text{Bias}^2} + \underbrace{\mathrm{Var}_{\mathcal{D}}[f(x)]}_{\text{Variance}}$$

Experiment: Unimodal Variance Curve

Computing Bias and Variance:

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y

- Recall bias-variance decomposition for mean-squared error:

$$\underbrace{\mathbb{E}_{\mathcal{D}}[(y - f(x))^2]}_{\text{MSE}} = \underbrace{(y - \mathbb{E}_{\mathcal{D}}[f(x)])^2}_{\text{Bias}^2} + \underbrace{\text{Var}_{\mathcal{D}}[f(x)]}_{\text{Variance}}$$

- Expectation taken over randomness in training data $\mathcal{D}$

Experiment: Unimodal Variance Curve

Computing Bias and Variance:

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y

- Recall bias-variance decomposition for mean-squared error:

$$\underbrace{\mathbb{E}_{\mathcal{D}}[(y - f(x))^2]}_{\text{MSE}} = \underbrace{(y - \mathbb{E}_{\mathcal{D}}[f(x)])^2}_{\text{Bias}^2} + \underbrace{\mathrm{Var}_{\mathcal{D}}[f(x)]}_{\text{Variance}}$$

- Expectation taken over randomness in training data $\mathcal{D}$

- Will consider average bias/variance over test dist., i.e.

$$\text{Bias}^2 := \mathbb{E}_{x,y}[(y - \mathbb{E}_{\mathcal{D}}[f(x)])^2]$$

Experiment: Unimodal Variance Curve

How to compute from data? (Only one dataset $\mathcal{D}$)

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y

Experiment: Unimodal Variance Curve

How to compute from data? (Only one dataset $\mathcal{D}$)

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y

- Split data into two halves $\mathcal{D}_1, \mathcal{D}_2$

Experiment: Unimodal Variance Curve

How to compute from data? (Only one dataset $\mathcal{D}$)

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y

- Split data into two halves $\mathcal{D}_1, \mathcal{D}_2$

- Train classifiers f_1, f_2

Experiment: Unimodal Variance Curve

How to compute from data? (Only one dataset $\mathcal{D}$)

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y

- Split data into two halves $\mathcal{D}_1, \mathcal{D}_2$

- Train classifiers f_1, f_2

- Unbiased estimate of variance: $\frac{1}{2}(f_1(x) - f_2(x))^2$

Experiment: Unimodal Variance Curve

How to compute from data? (Only one dataset $\mathcal{D}$)

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y

- Split data into two halves $\mathcal{D}_1, \mathcal{D}_2$

- Train classifiers f_1, f_2

- Unbiased estimate of variance: $\frac{1}{2}(f_1(x) - f_2(x))^2$

- Average over multiple splits to get better estimate

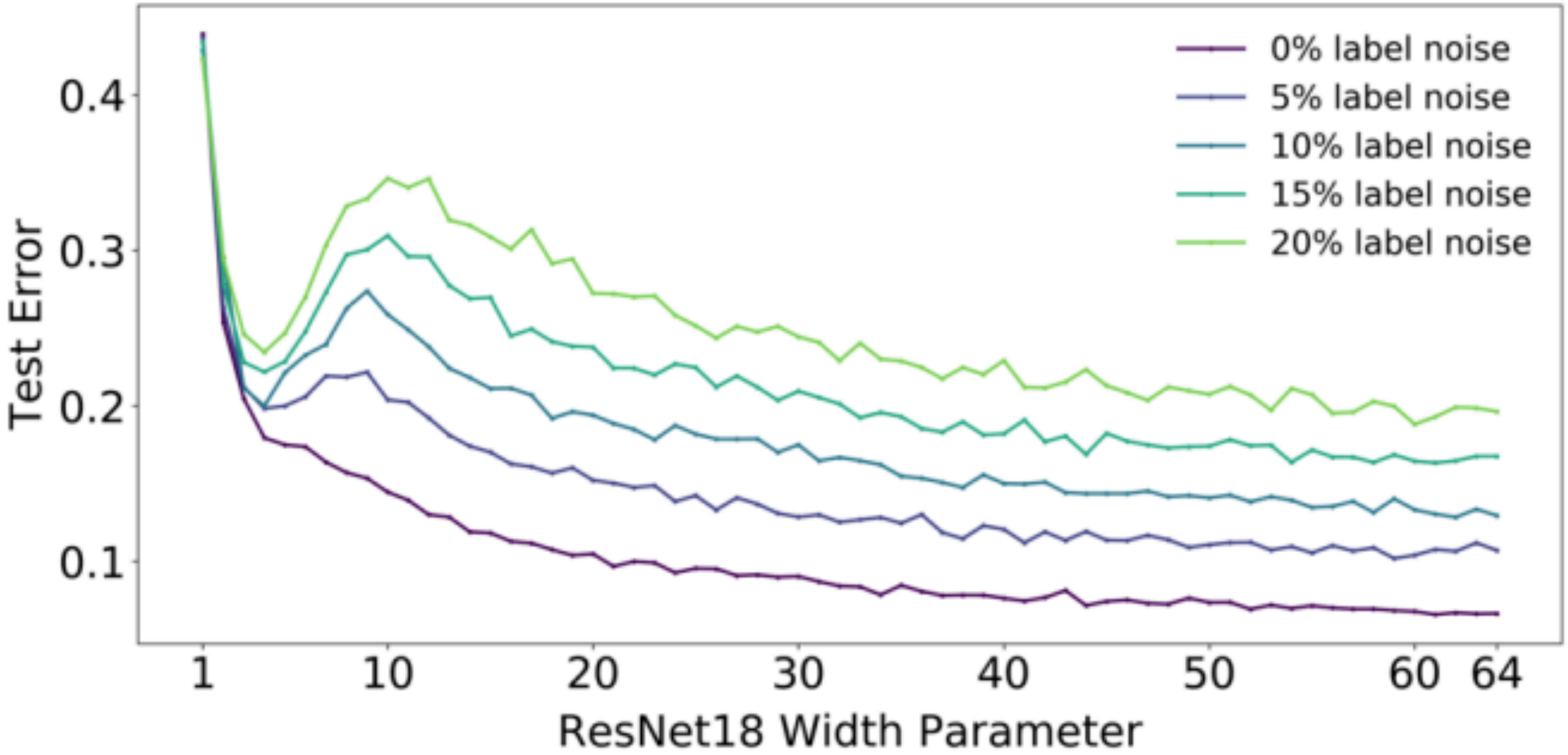
Experiment: Unimodal Variance Curve

How to compute from data? (Only one dataset $\mathcal{D}$)

- Learned classifier $f(x)$ (depends on dataset $\mathcal{D}$), predict y
- Split data into two halves $\mathcal{D}_1, \mathcal{D}_2$
- Train classifiers f_1, f_2
- Unbiased estimate of variance: $\frac{1}{2}(f_1(x) - f_2(x))^2$
- Average over multiple splits to get better estimate
- Compute Bias via $\text{Bias}^2 = \text{MSE} - \text{Variance}$

Experiment: Unimodal Variance Curve

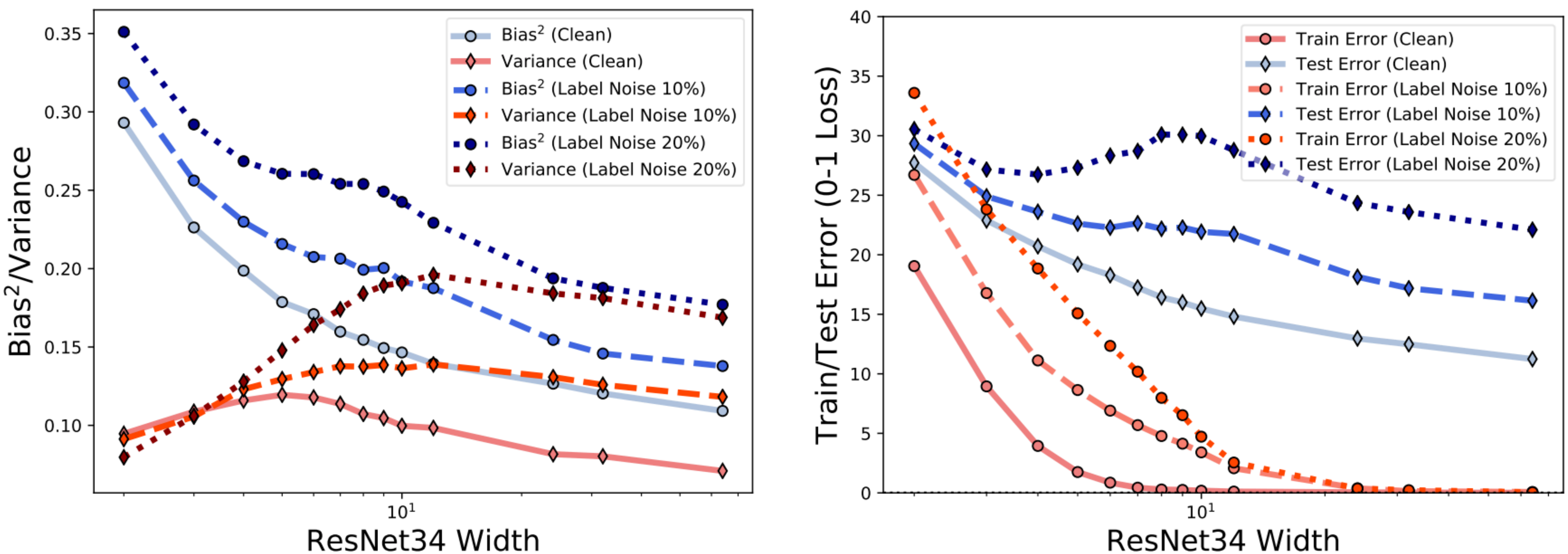
- Label Noise: Direct Connection to Double Descent



P. Nakkiran et al. (2019)

Experiment: Unimodal Variance Curve

- Label Noise: Direct Connection to Double Descent



Z. Yang, Y. Yu, C. You, J. Steinhardt, Y. Ma (2020)

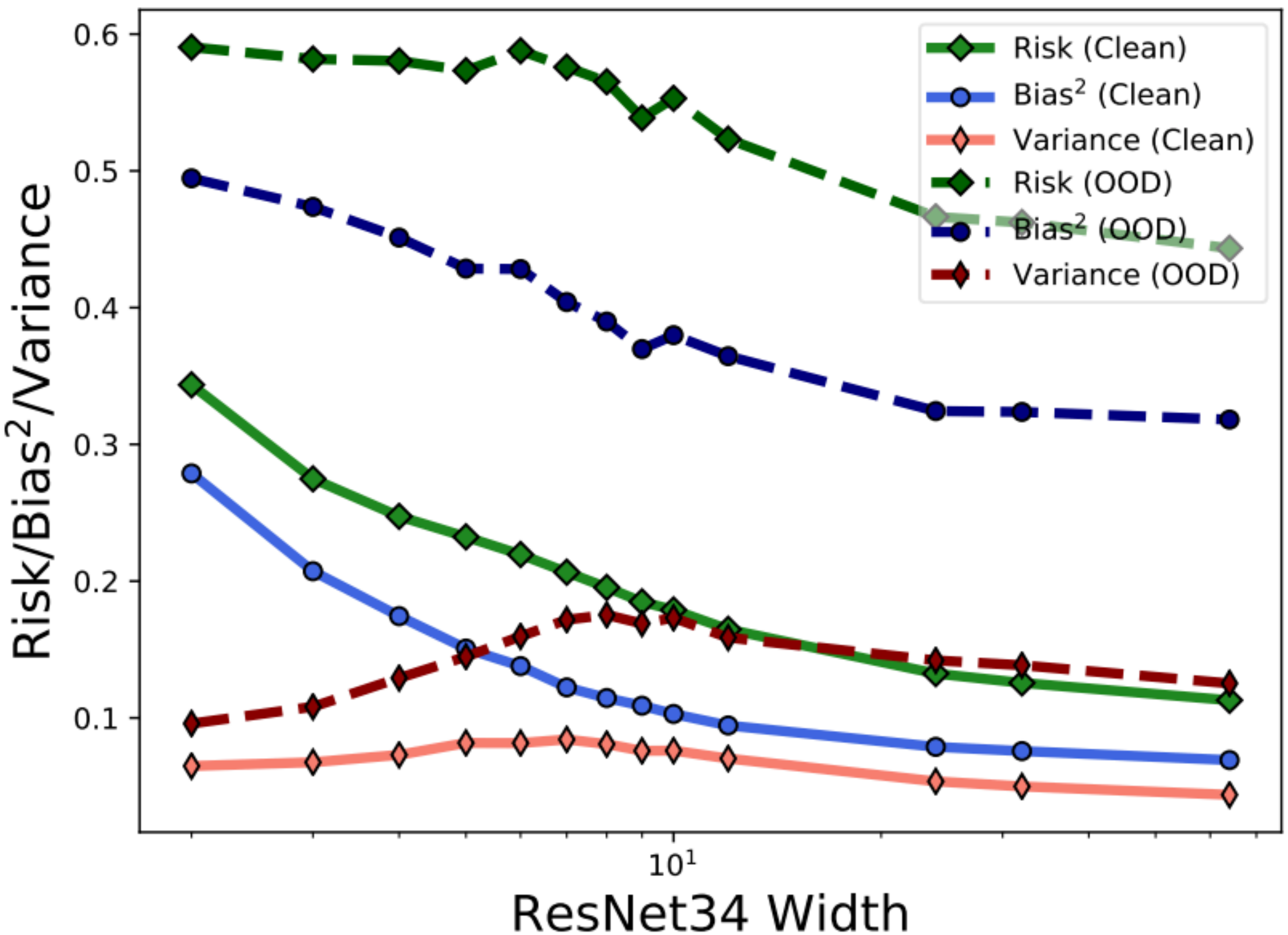
Experiment: Unimodal Variance Curve

- Bonus: Increased bias explains drop in out-of-distribution accuracy

D. Hendrycls, T. Dietterich (2020)

Experiment: Unimodal Variance Curve

- Bonus: Increased bias explains drop in out-of-distribution accuracy



Z. Yang, Y. Yu, C. You, J. Steinhardt, Y. Ma (2020)

Outline:

1. Bias Variance Tradeoff v.s. Double Descent
2. Our Proposal: Unimodal Variance
3. Measurement: Experimental Set-up
4. Theory: Analysis of Two-Layer Network

Theory: Analysis of Two-Layer Network

- Statistical Assumption on Data:

$$x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d); \quad y_i \,|\, x_i = \langle \beta, x_i \rangle; \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} = \{X \in \mathbb{R}^{d \times n}, y \in \mathbb{R}^{n}\}$$

Theory: Analysis of Two-Layer Network

- Statistical Assumption on Data:

$$x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d); \quad y_i \mid x_i = \langle \beta, x_i \rangle; \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} = \{X \in \mathbb{R}^{d \times n}, y \in \mathbb{R}^{n}\}$$

- Assumptions on Two-Layer Network:

First Layer: $W \in \mathbb{R}^{p \times d}, W_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/d)$

Second Layer: $\beta_\lambda(\mathcal{D}, W) = \text{argmin}_\beta \|(WX)^T \beta - y\|_2^2 + \lambda \|\beta\|_2^2$

Theory: Analysis of Two-Layer Network

- Statistical Assumption on Data:

$$x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d); \quad y_i \,|\, x_i = \langle \beta, x_i \rangle; \mathscr{D} = \{(x_i, y_i)\}_{i=1}^n = \{X \in \mathbb{R}^{d \times n}, y \in \mathbb{R}^n\}$$

- Assumptions on Two-Layer Network:

First Layer: $W \in \mathbb{R}^{p \times d}, W_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/d)$

Second Layer: $\beta_\lambda(\mathscr{D}, W) = \text{argmin}_\beta \|(WX)^T \beta - y\|_2^2 + \lambda \|\beta\|_2^2$

- Asymptotic Assumption: $\lim_{d \to \infty} \frac{p(d)}{d} = \gamma \in \mathbb{R}; \lim_{n \to \infty} \frac{n(d)}{d} = \eta \in \mathbb{R};$

Theory: Analysis of Two-Layer Network

- Statistical Assumption on Data:

$$x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d); \quad y_i \mid x_i = \langle \beta, x_i \rangle; \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} = \{X \in \mathbb{R}^{d \times n}, y \in \mathbb{R}^n\}$$

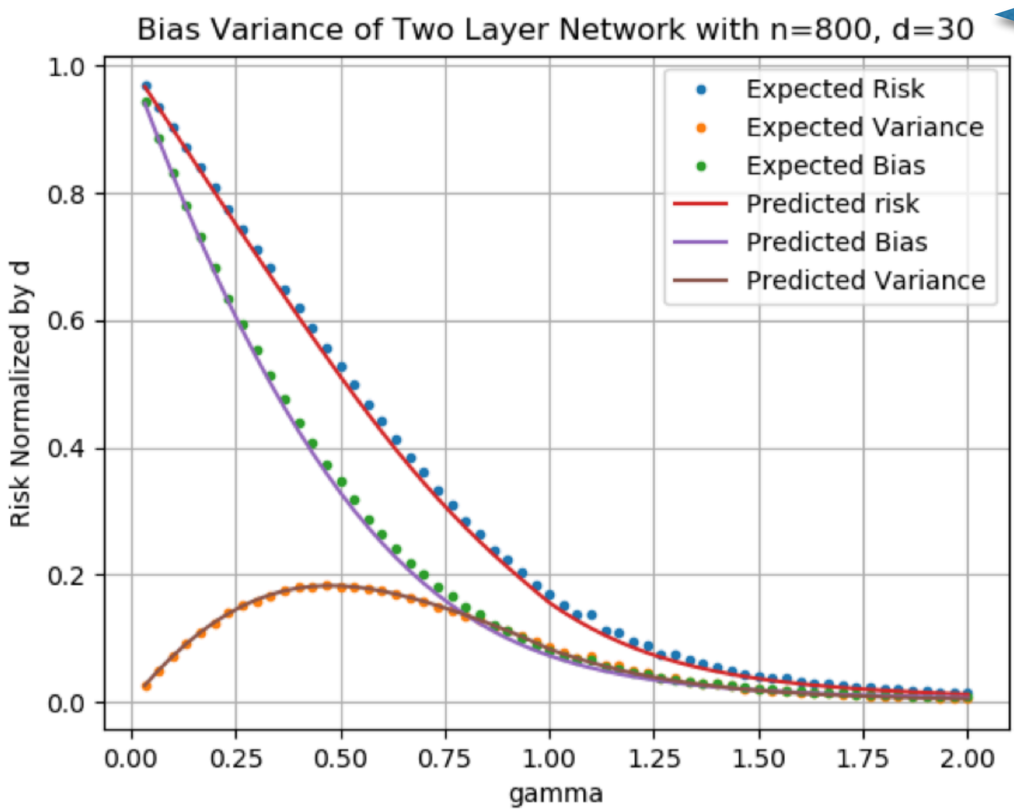
- Assumptions on Two-Layer Network:

First Layer: $W \in \mathbb{R}^{p \times d}, W_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/d)$

Second Layer: $\beta_\lambda(\mathcal{D}, W) = \text{argmin}_\beta \|(WX)^T\beta - y\|_2^2 + \lambda\|\beta\|_2^2$

- Asymptotic Assumption: $\lim_{d \to \infty} \frac{p(d)}{d} = \gamma \in \mathbb{R}; \lim_{n \to \infty} \frac{n(d)}{d} = \eta \in \mathbb{R};$

- Main Results: Analytical Expression for Bias/Variance/Risk in λ, η, γ



The dots represents the simulated risk;
The solid line represents analytically derived results;

Theory: Analysis of Two-Layer Network

- Statistical Assumption on Data:

$$x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d); \quad y_i | x_i = \langle \beta, x_i \rangle; \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} = \{X \in \mathbb{R}^{d \times n}, y \in \mathbb{R}^n\}$$

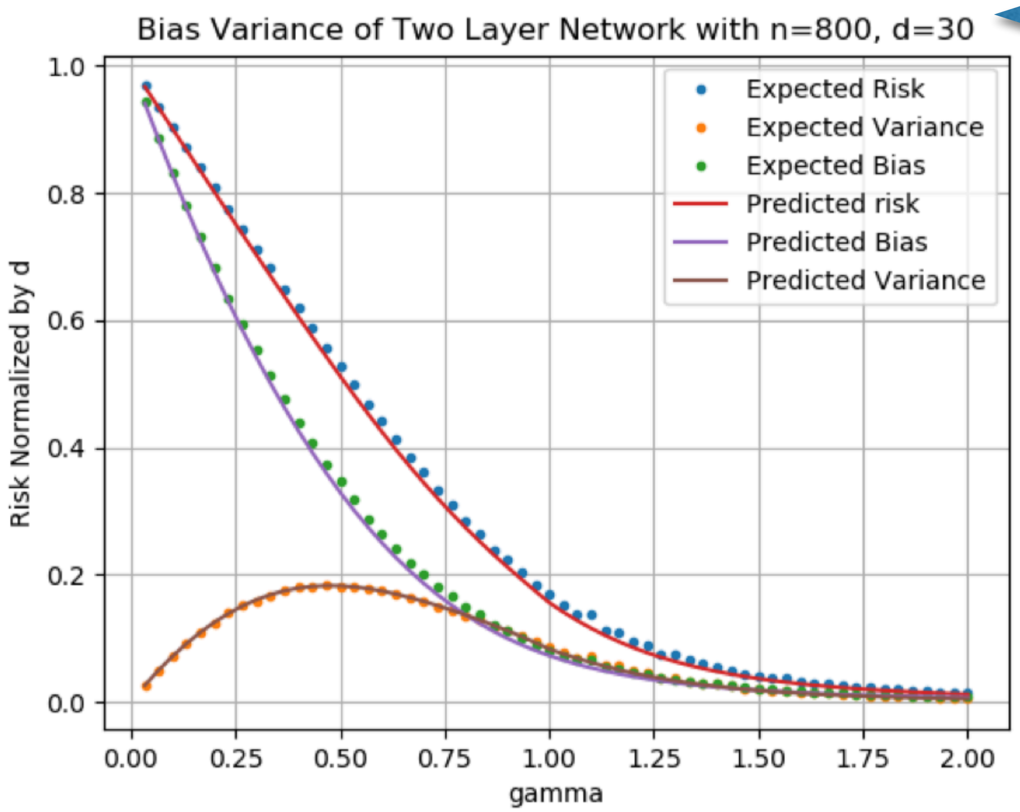
- Assumptions on Two-Layer Network:

First Layer: $W \in \mathbb{R}^{p \times d}, W_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/d)$

Second Layer: $\beta_\lambda(\mathcal{D}, W) = \text{argmin}_\beta \|(WX)^T \beta - y\|_2^2 + \lambda \|\beta\|_2^2$

- Asymptotic Assumption: $\lim_{d \to \infty} \frac{p(d)}{d} = \gamma \in \mathbb{R}; \lim_{n \to \infty} \frac{n(d)}{d} = \eta \in \mathbb{R};$

- Main Results: Analytical Expression for Bias/Variance/Risk in λ, η, γ



The dots represents the simulated risk;
The solid line represents analytically derived results;

We can see that:

1. Formula obtained using RMT is well aligned with simulation;
2. This model displays unimodal variance pattern

Theory: Analysis of Two-Layer Network

- Statistical Assumption on Data:

$$x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d); \quad y_i \,|\, x_i = \langle \beta, x_i \rangle; \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} = \{X \in \mathbb{R}^{d \times n}, y \in \mathbb{R}^n\}$$

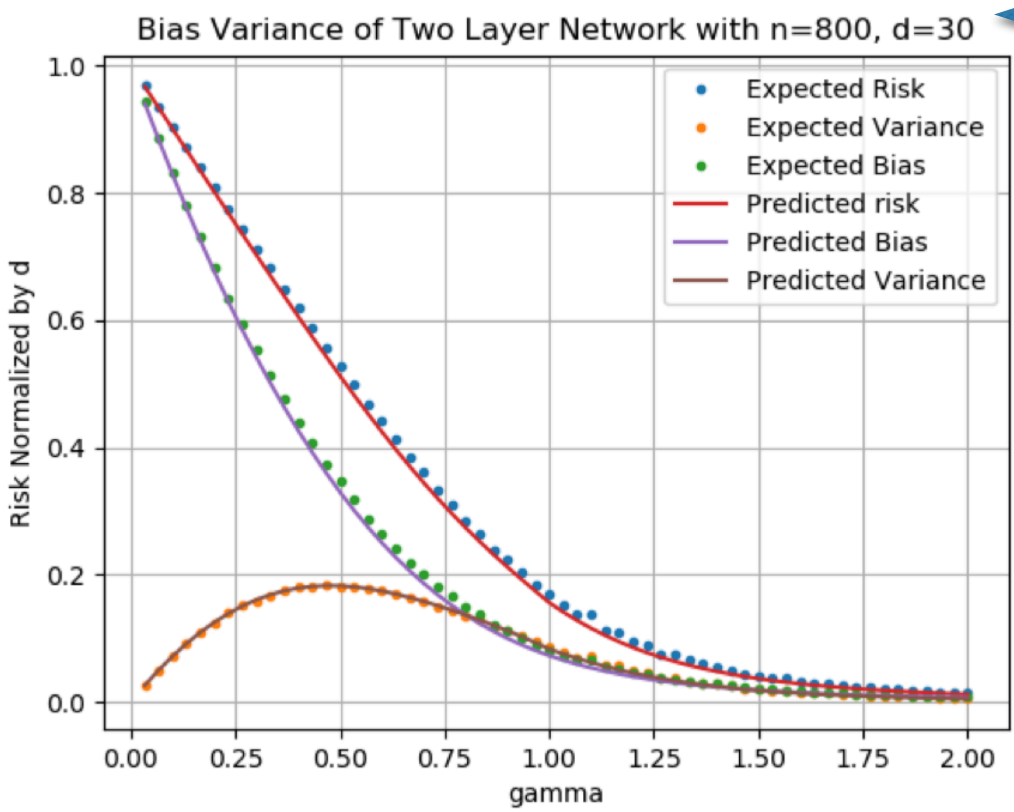
- Assumptions on Two-Layer Network:

First Layer: $W \in \mathbb{R}^{p \times d}, W_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/d)$

Second Layer: $\beta_\lambda(\mathcal{D}, W) = \text{argmin}_\beta \|(WX)^T \beta - y\|_2^2 + \lambda \|\beta\|_2^2$

- Asymptotic Assumption: $\lim_{d \to \infty} \frac{p(d)}{d} = \gamma \in \mathbb{R}; \lim_{n \to \infty} \frac{n(d)}{d} = \eta \in \mathbb{R};$

- Main Results: Analytical Expression for Bias/Variance/Risk in λ, η, γ



The dots represents the simulated risk;
The solid line represents analytically derived results;

We can see that:

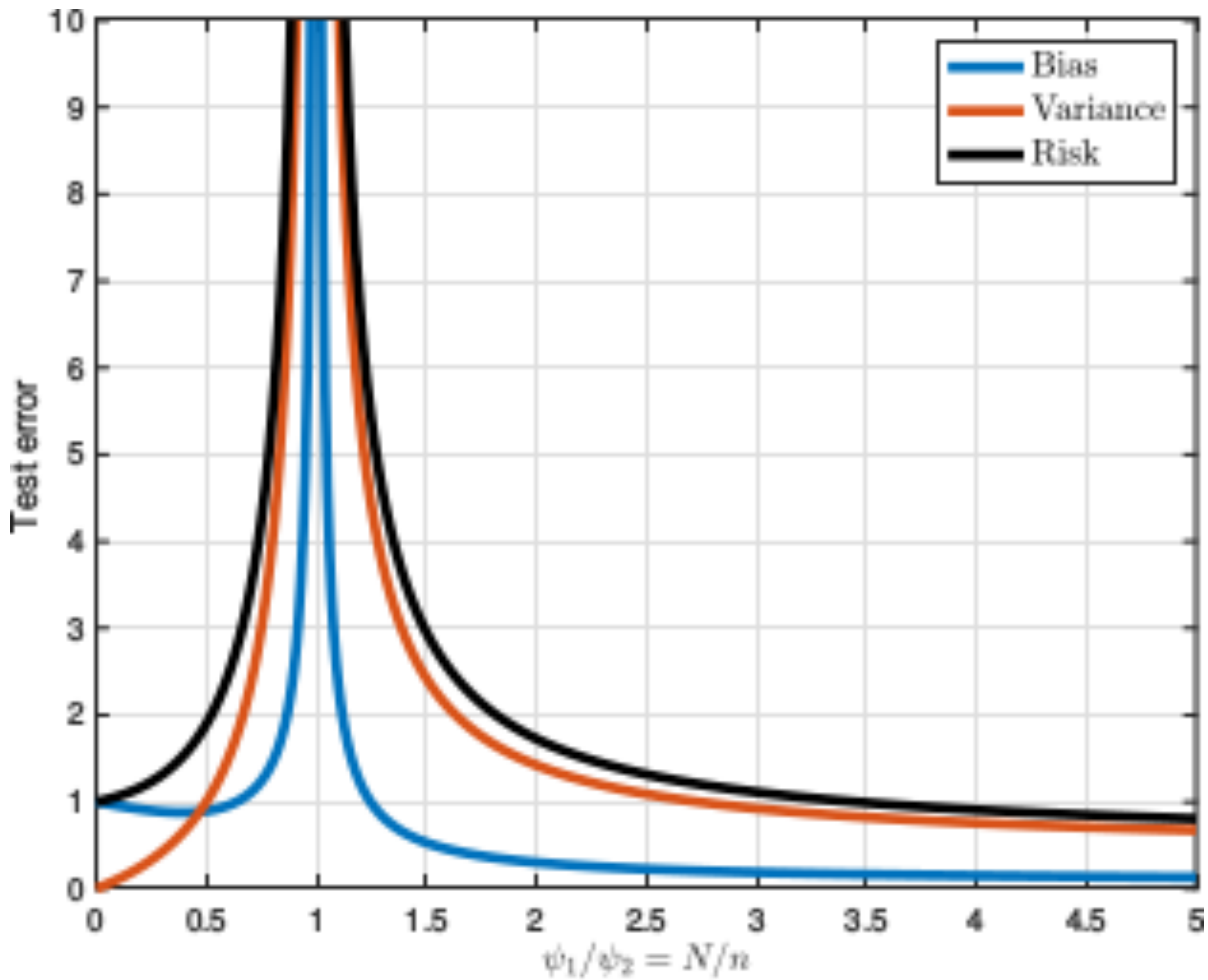
1. Formula obtained using RMT is well aligned with simulation;
2. This model displays unimodal variance pattern

Techniques of Proof:

1. Random Matrix Theory (Spectral Theorems);
2. Combinatorics of Non-crossing partitions.

Theory: Analysis of Two-Layer Network

- Comparison: Random design v.s. Fixed Design



S. Mei, A. Montanari (2019)

Take-aways

- **Monotonic** bias + **unimodal** variance demystifies double descent

Take-aways

- **Monotonic** bias + **unimodal** variance demystifies double descent

- Need to get **details** right (estimator? random or fixed design?)

Take-aways

- **Monotonic** bias + **unimodal** variance demystifies double descent

- Need to get **details** right (estimator? random or fixed design?)

- **Robustness** of phenomenon: suggests "fundamentalness", provides target for explanation

Take-aways

- **Monotonic** bias + **unimodal** variance demystifies double descent
- Need to get **details** right (estimator? random or fixed design?)
- **Robustness** of phenomenon: suggests “fundamentalness”, provides target for explanation
- **Open question**: Why is variance unimodal?

Thanks!