Towards Understanding Generalization in Deep Learning by Revisiting the Bias-Variance Decomposition

Brady Neal

Outline

Part 1: Contradiction between traditional complexity measures and over-parameterization

Part 2: Bias-variance decomposition

Part 3: Over-parameterization and variance

Part 4: Zhang et al. (2017) via bias-variance decomposition

Outline

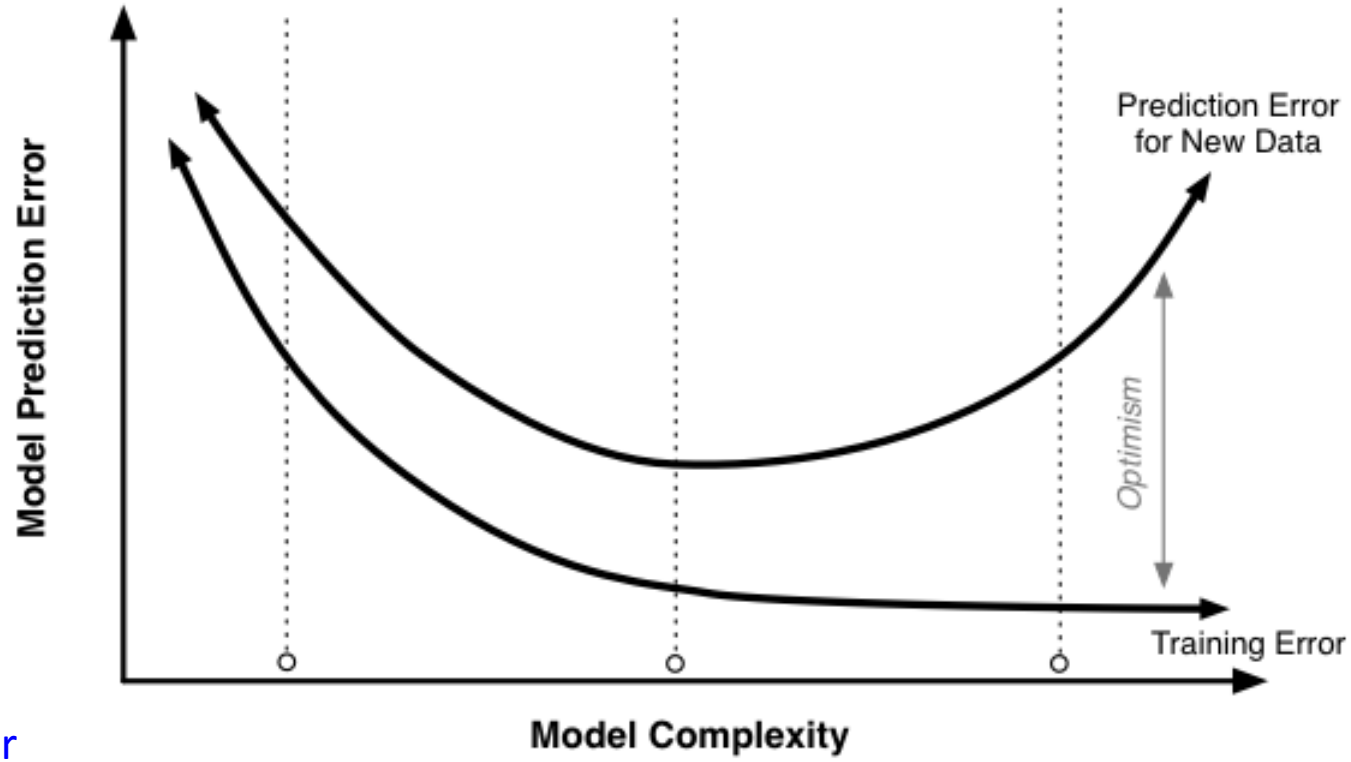
Part 1: Contradiction between traditional complexity measures and over-parameterization

Part 2: Bias-variance decomposition

Part 3: Over-parameterization and variance

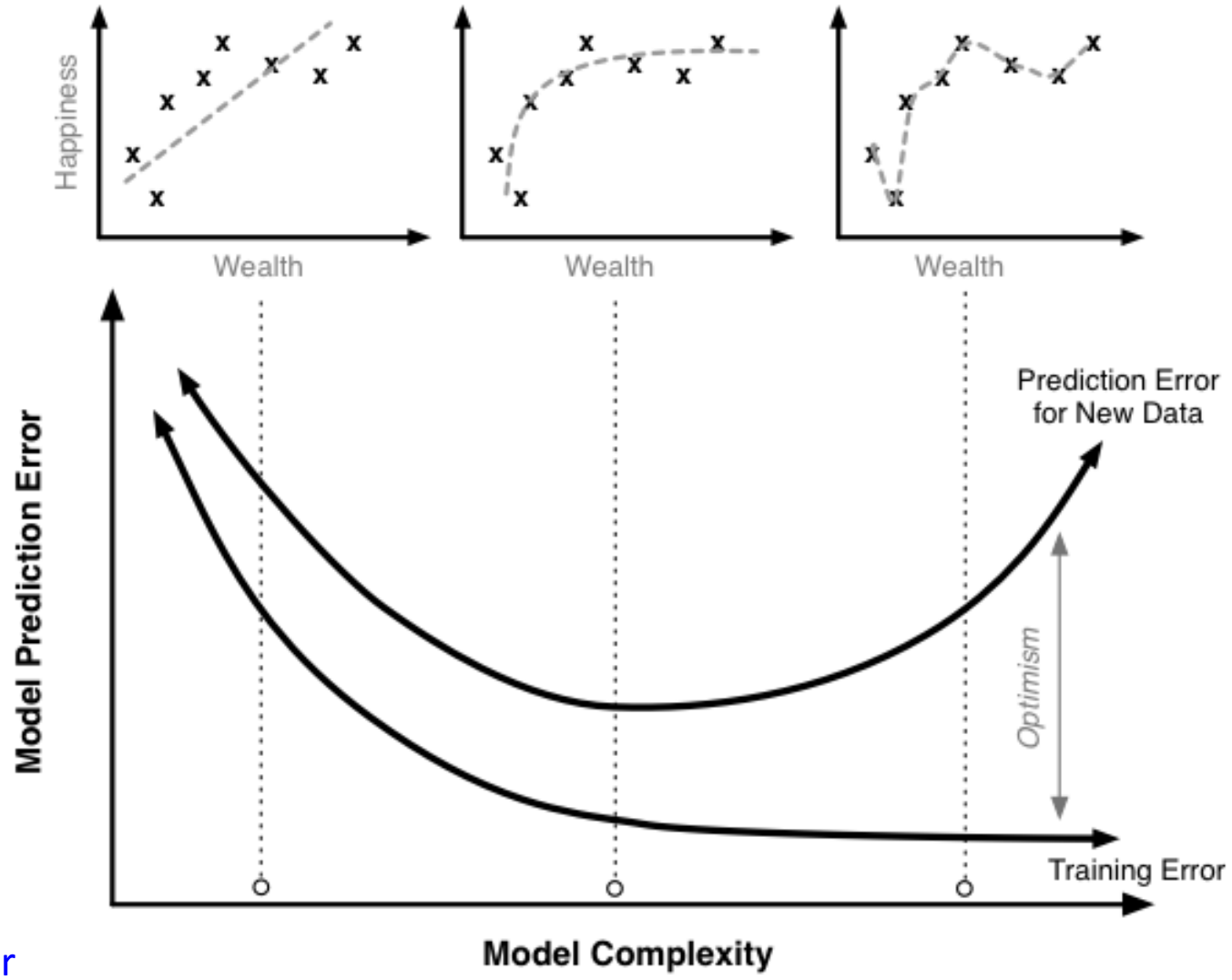
Part 4: Zhang et al. (2017) via bias-variance decomposition

The learning problem and generalization



Quora answer

The learning problem and generalization



[Quora answer]

Main goal: minimize expected risk

Expected risk: $R(h) = \mathop{\mathbb{E}}_{(x,y)\sim p(x,y)} \ell(h(x), y)$

where $p(x, y)$ is the data distribution and

ℓ is the loss on a particular example

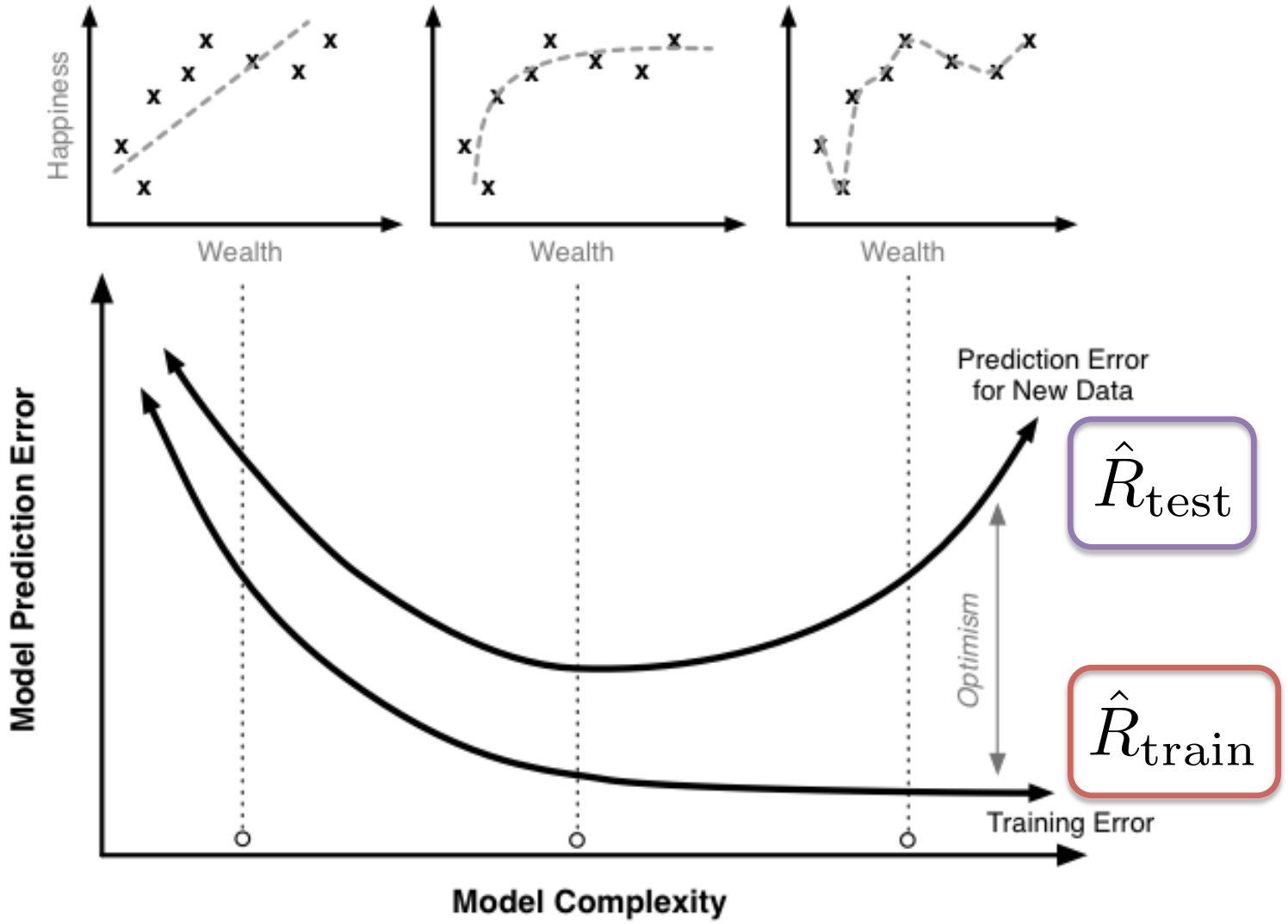
approximate with empirical risk on test set:

$$R(h) \approx \hat{R}_{\text{test}}(h) = \frac{1}{m_{\text{test}}} \sum_{(x,y)\in \text{test set}} \ell(h(x), y)$$

attempt to learn by minimizing training error:

$$\hat{R}_{\text{train}}(h) = \frac{1}{m_{\text{train}}} \sum_{(x,y)\in \text{train set}} \ell(h(x), y)$$

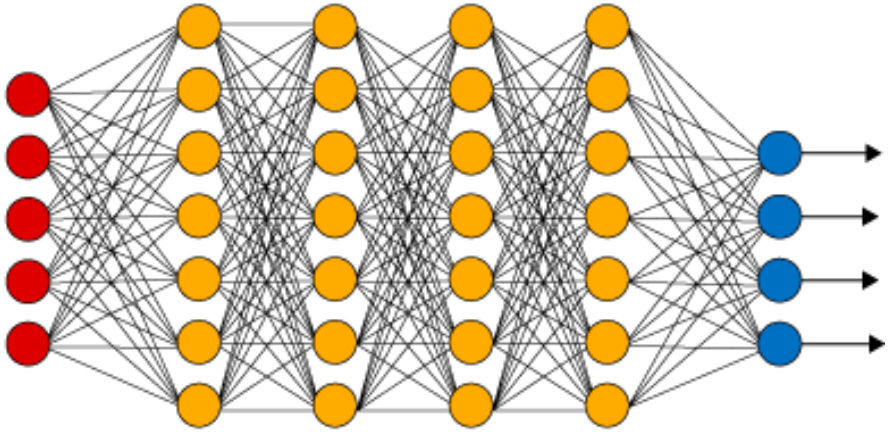
The learning problem and generalization



Traditional Measures of Complexity

$$h \in \mathcal{H}$$

e.g. class of neural networks that can be represented by neural network with fixed architecture

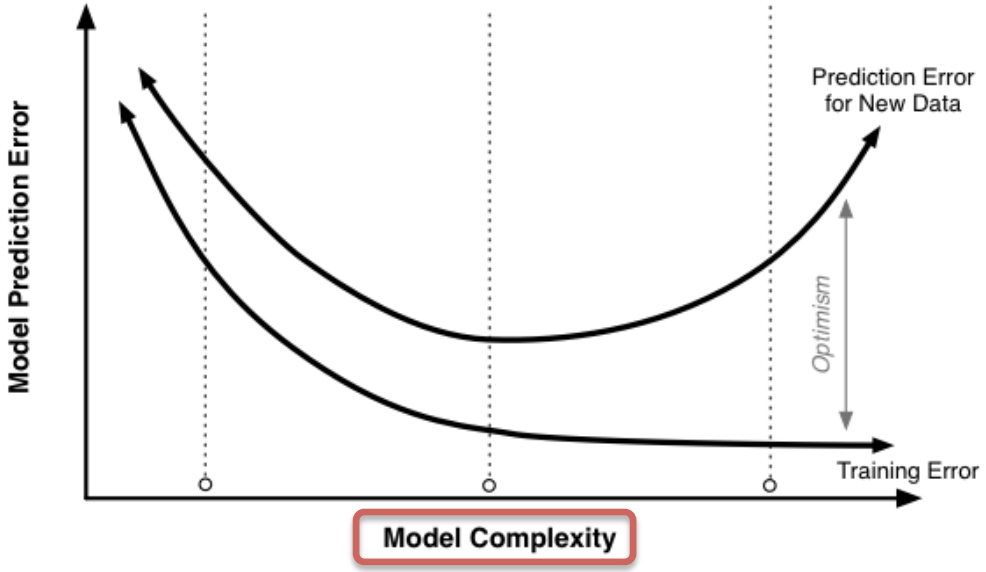


$$R(h) \leq \hat{R}_{\text{train}} + \sqrt{\frac{\text{VC}(\mathcal{H}) + \ln \frac{1}{\delta}}{m}}$$

$$R(h) \leq \hat{R}_{\text{train}} + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{m}}$$

both depend on size of network

Traditional Measures of Complexity

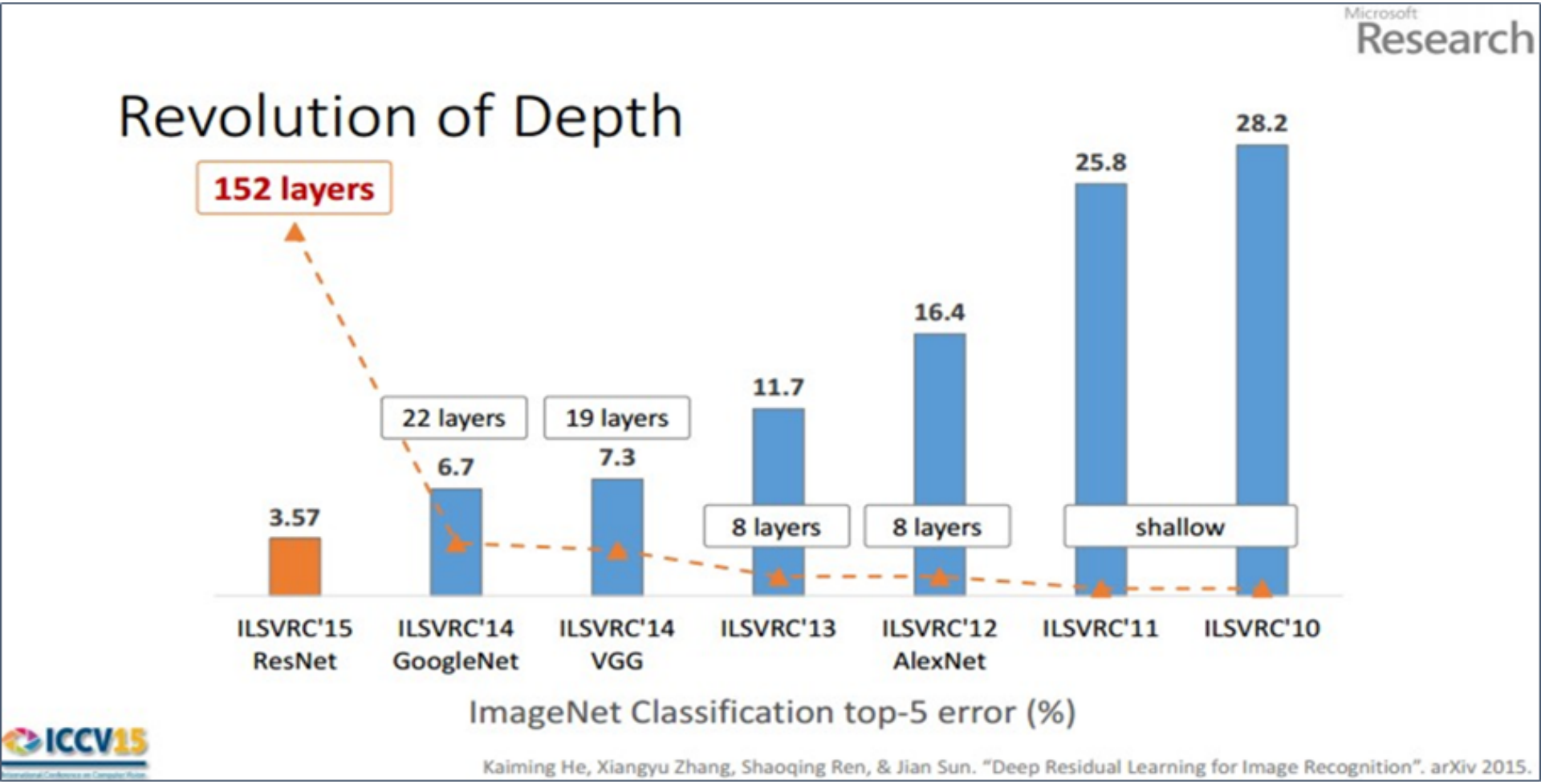


$$R(h) \leq \hat{R}_{\text{train}} + \sqrt{\frac{\text{VC}(\mathcal{H}) + \ln \frac{1}{\delta}}{m}}$$

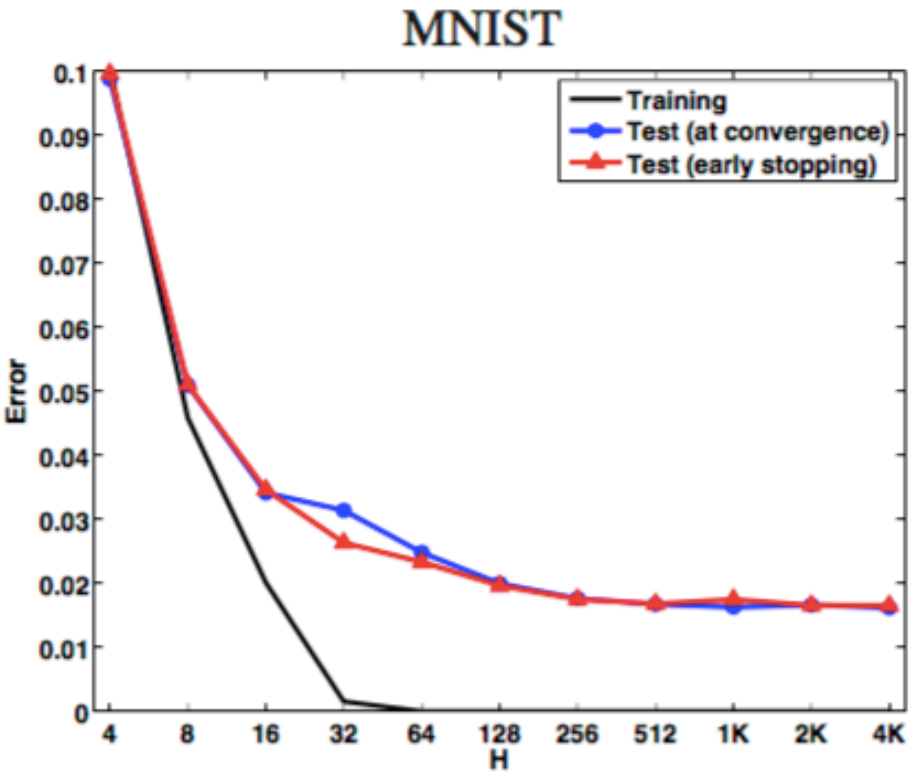
$$R(h) \leq \hat{R}_{\text{train}} + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{m}}$$

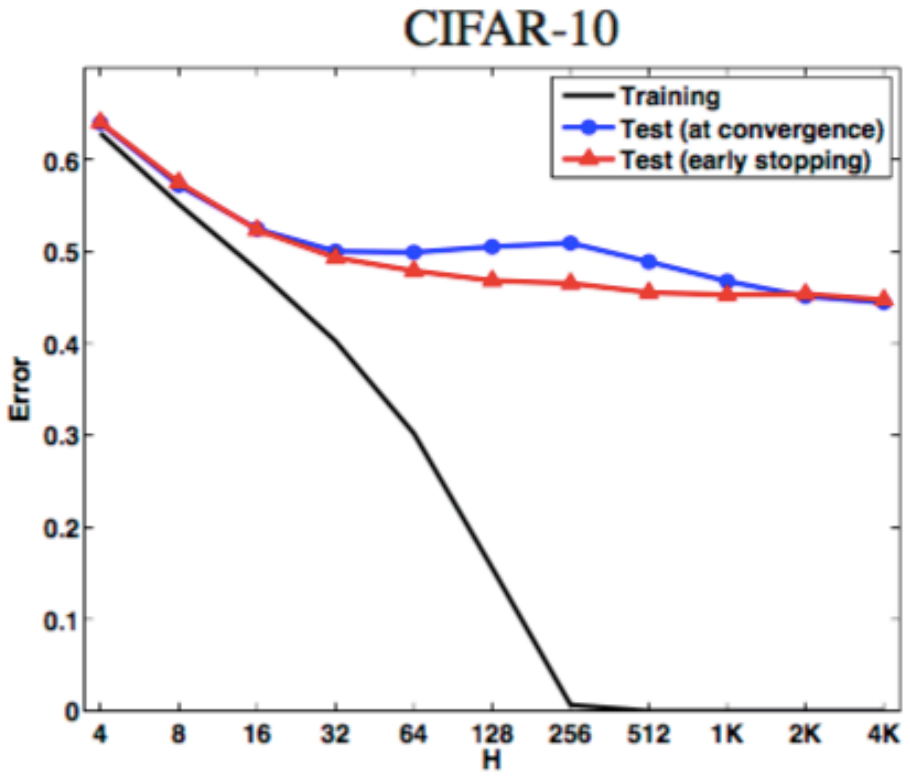
both depend on size of network

ImageNet Performance



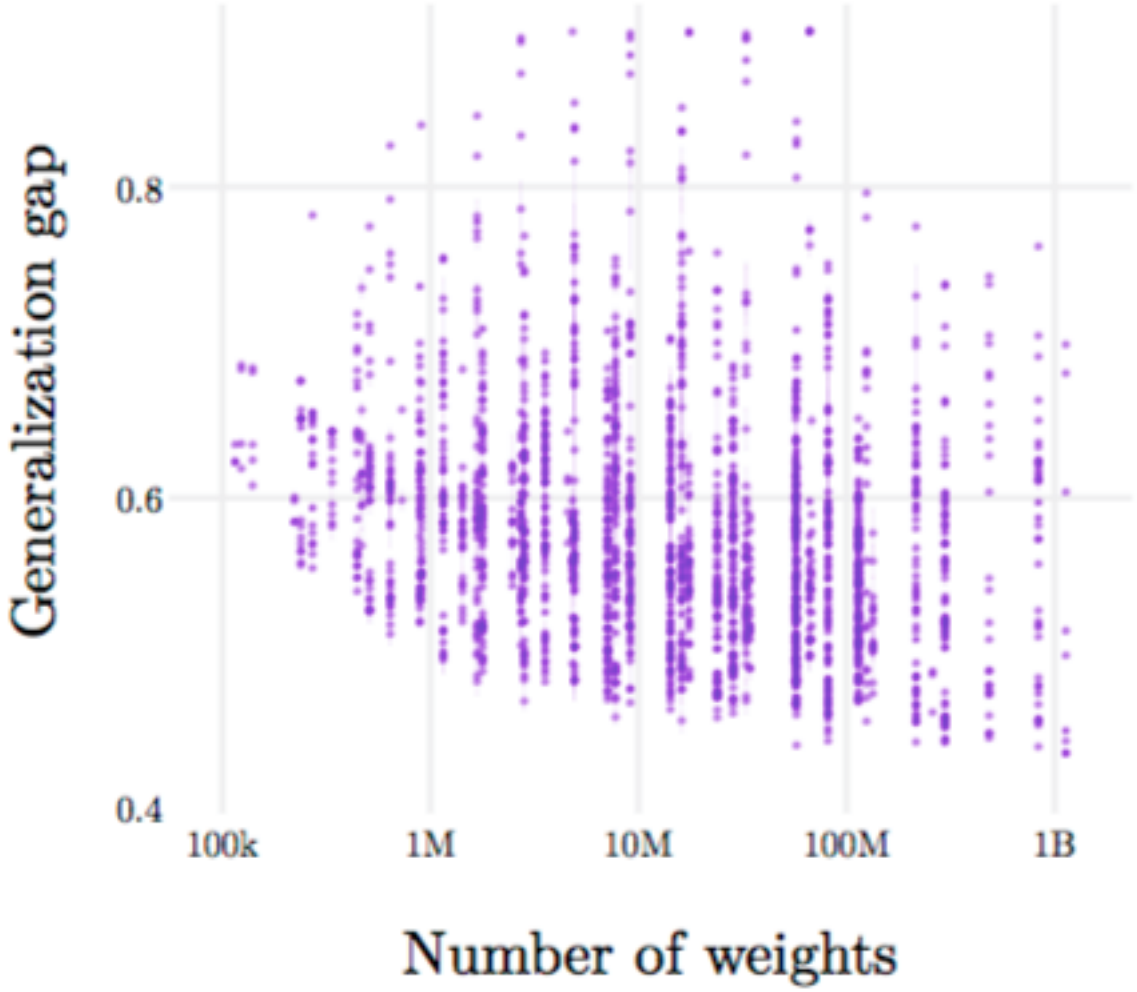
Number of Hidden Units: Bad Metric for Model Complexity





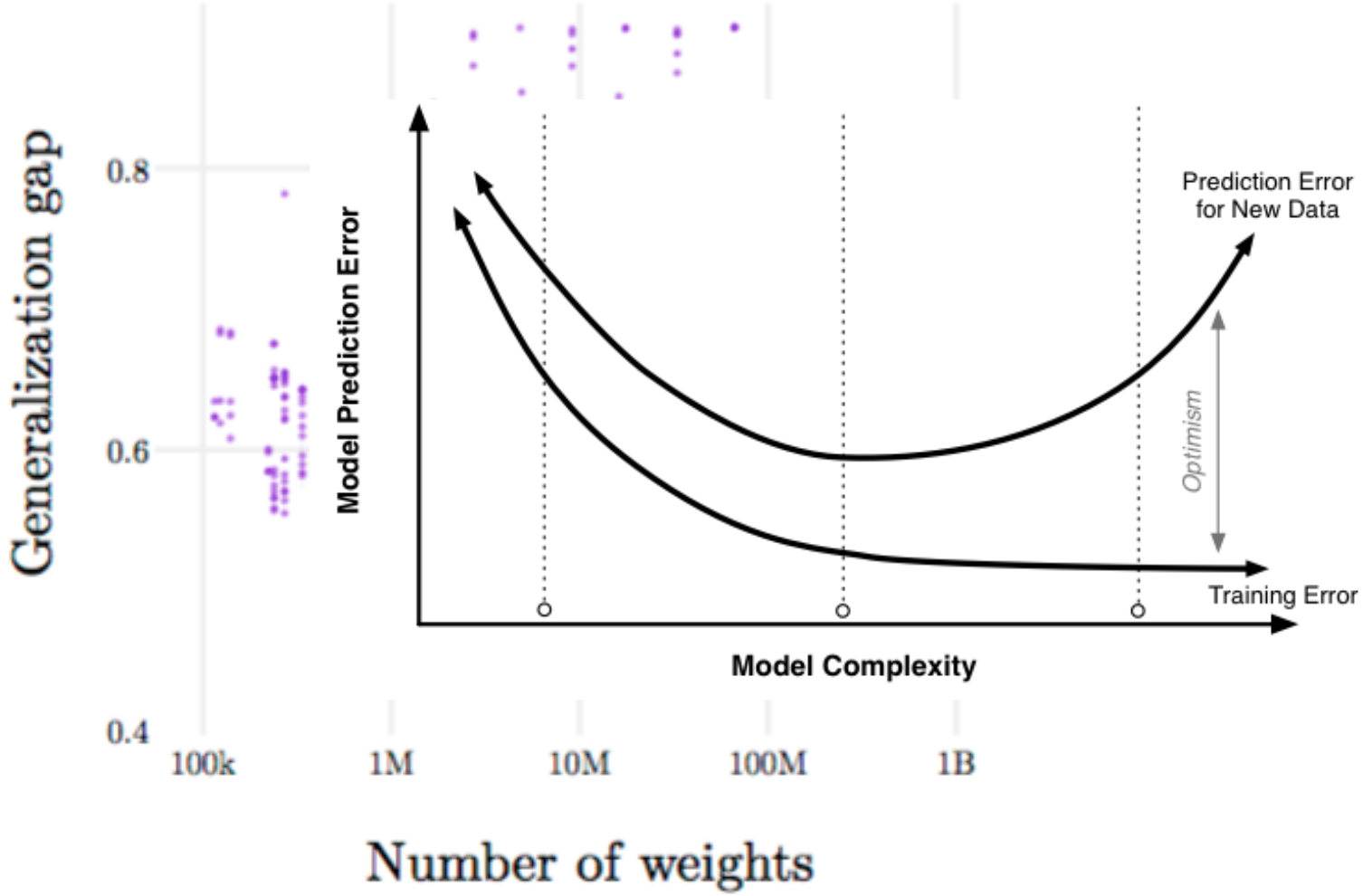
Neyshabur et al. (2015)

Number of Weights: Bad Metric for Model Complexity



Novak et al. (2018)

Number of Weights: Bad Metric for Model Complexity



Novak et al. (2018)

Outline

Part 1: Contradiction between traditional complexity measures and over-parameterization

Part 2: Bias-variance decomposition

Part 3: Over-parameterization and variance

Part 4: Zhang et al. (2017) via bias-variance decomposition

Reducible and Irreducible Error

$$y = f(x) + \epsilon \qquad f : \mathcal{X} \to \mathcal{Y} \text{ (true mapping)}$$

ϵ: noise with mean 0 and independent from S

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$

$$h_S : \mathcal{X} \to \mathcal{Y} \text{ (learned hypothesis)}$$

$$\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\epsilon}[R(h_S)] = \mathop{\mathbb{E}}_{(x,y)}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\epsilon}[(h_S(x) - y)^2]$$

$$= \underbrace{\mathop{\mathbb{E}}_{(x,y)}\mathop{\mathbb{E}}_{S}[(h_S(x) - f(x))^2]}_{\text{reducible error}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible error}}$$

Geman et al. (1992)

Bias-Variance Decomposition

$$\text{Reducible error:} \quad \mathop{\mathbb{E}}_{(x,y)} \mathop{\mathbb{E}}_{S}[(h_S(x) - f(x))^2]$$

$$= \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$= \mathop{\mathbb{E}}_{(x,y)} \left[\left(\mathop{\mathbb{E}}_{S}[h_S(x)] - f(x) \right)^2 + \text{Var}(h_S) \right]$$

$$= \mathop{\mathbb{E}}_{(x,y)} [\text{Bias}^2(h_S(x)) + \text{Var}(h_S(x))]$$

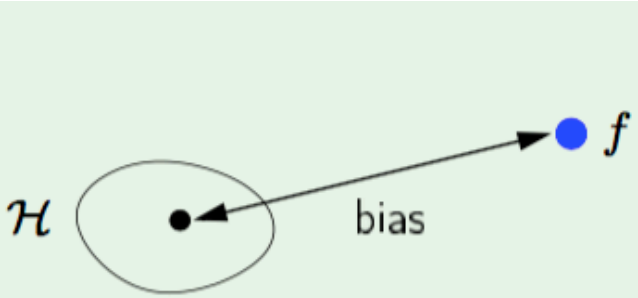
$$\mathbb{E}[R(h_S)] = \mathop{\mathbb{E}}_{(x,y)} [\text{Bias}^2(h_S(x)) + \text{Var}(h_S(x))] + \text{Var}(\epsilon)$$

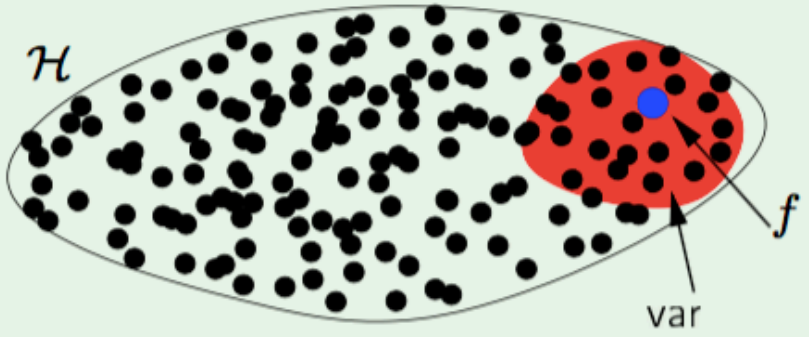
Geman et al. (1992)

Bias-Variance Intuition

$$\mathop{\mathbb{E}}_{(x,y)}\left[\left(\mathop{\mathbb{E}}_{S}[h_S(x)] - f(x)\right)^2 + \mathrm{Var}(h_S)\right]$$

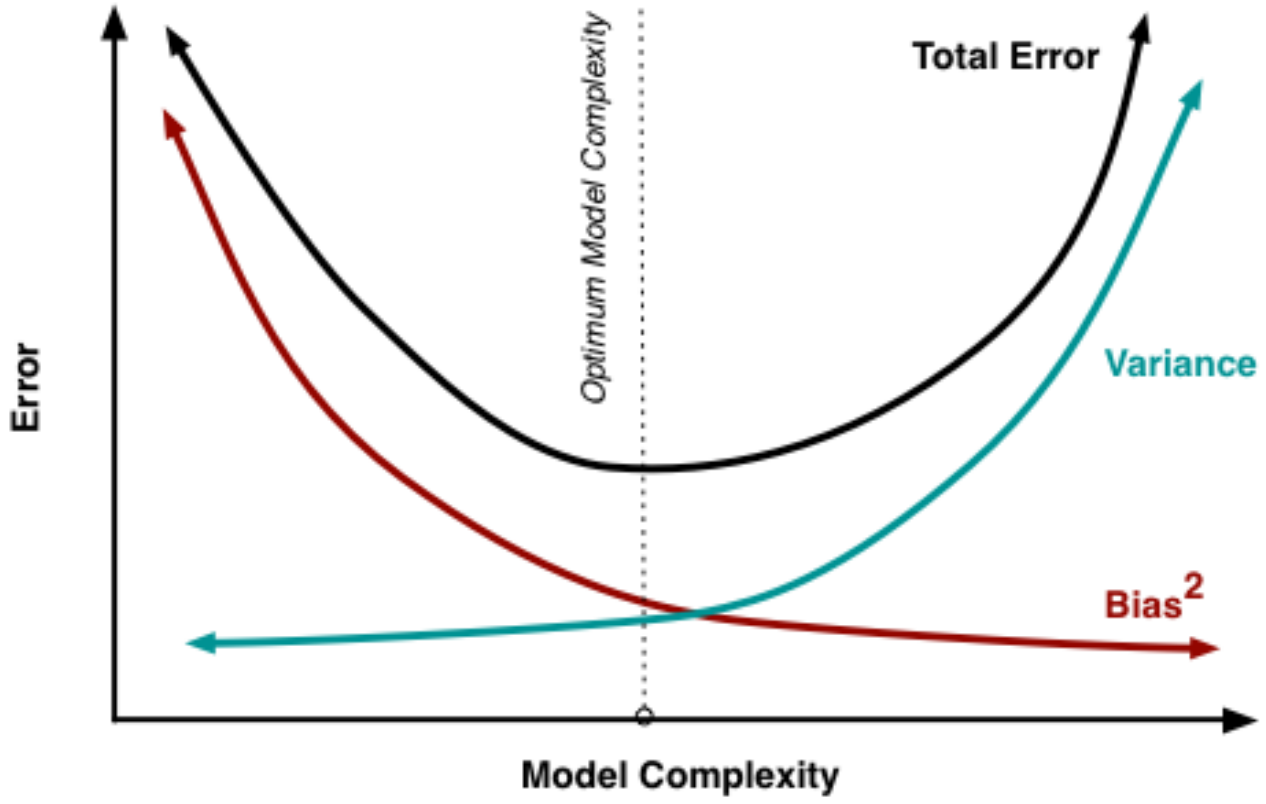
$$= \mathop{\mathbb{E}}_{(x,y)}[\mathrm{Bias}^2(h_S(x)) + \mathrm{Var}(h_S(x))]$$





Abu-Mostafa (2012)

Interpretation from Ben Recht



Bias-Variance vs. Complexity Measures

- tight! (equality)
- inherently depends on everything
  - distribution
  - learning algorithm
  - hypothesis class
- empirical expected risk
- in expectation
- no explicit dependence on size of network

- extremely general
  - distribution free
  - learning algorithm free
- only depends on training loss and hypothesis class
- analytical generalization gap
- complexity of hypothesis class grows with size of network
- loose inequality

Original paper from 1992

Title: Neural Networks and the Bias/Variance Dilemma

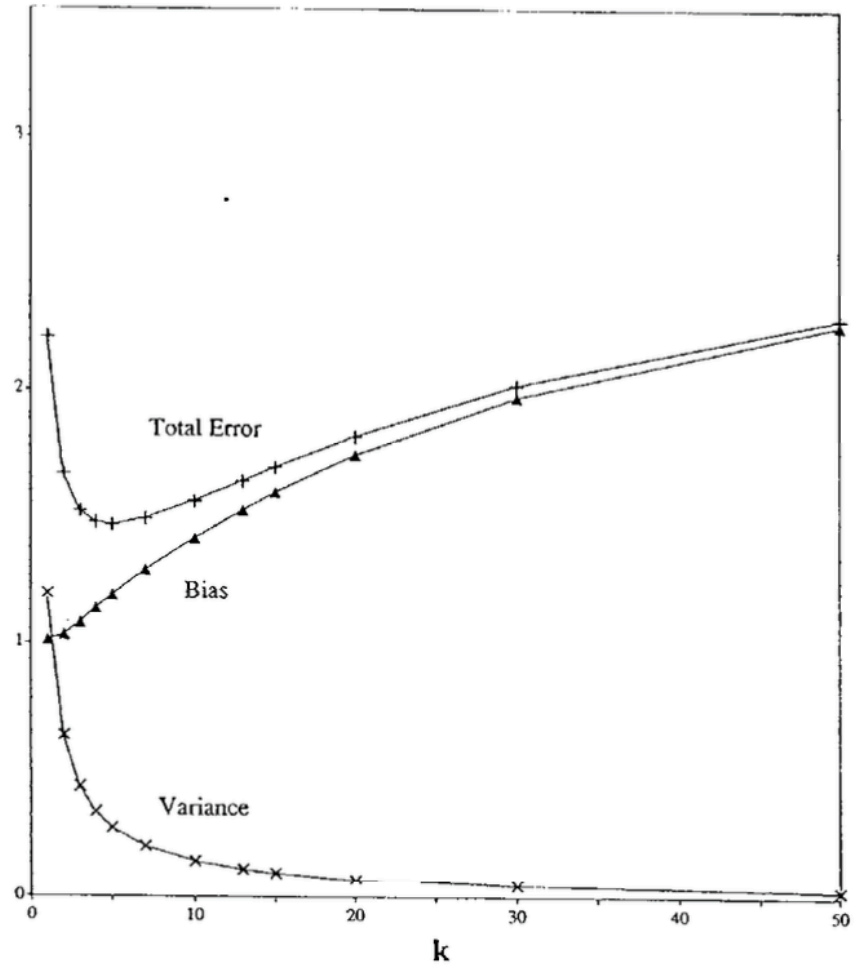
Neural networks are “like” nonparametric models

Experiments to show similarities in bias/variance

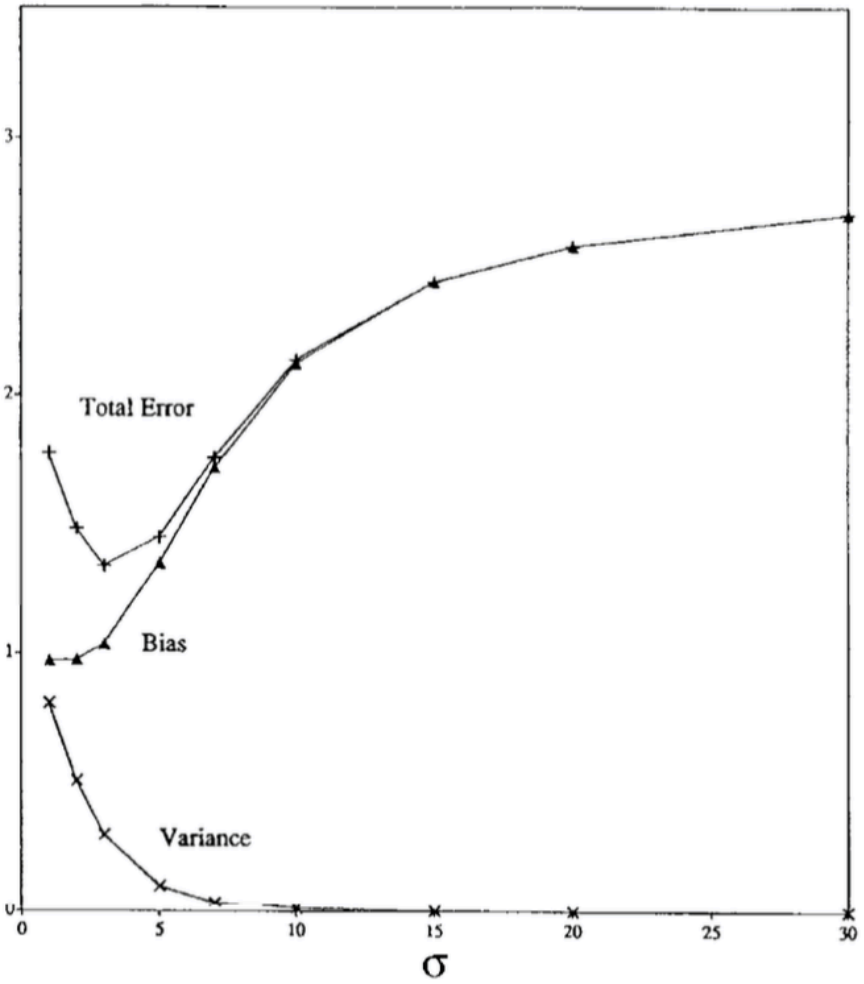
Geman et al. (1992)

Experiments with nonparametric models

KNN



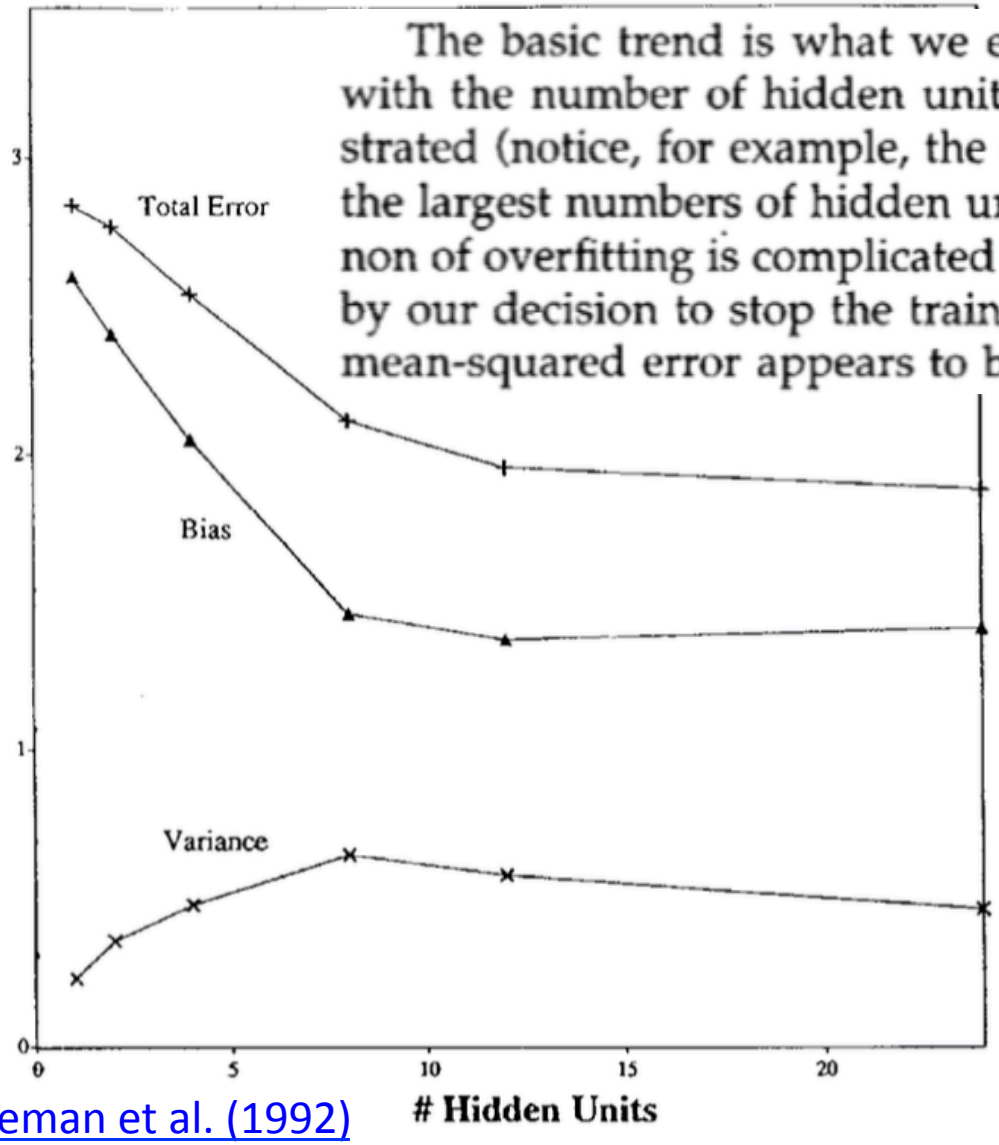
Kernel Regression

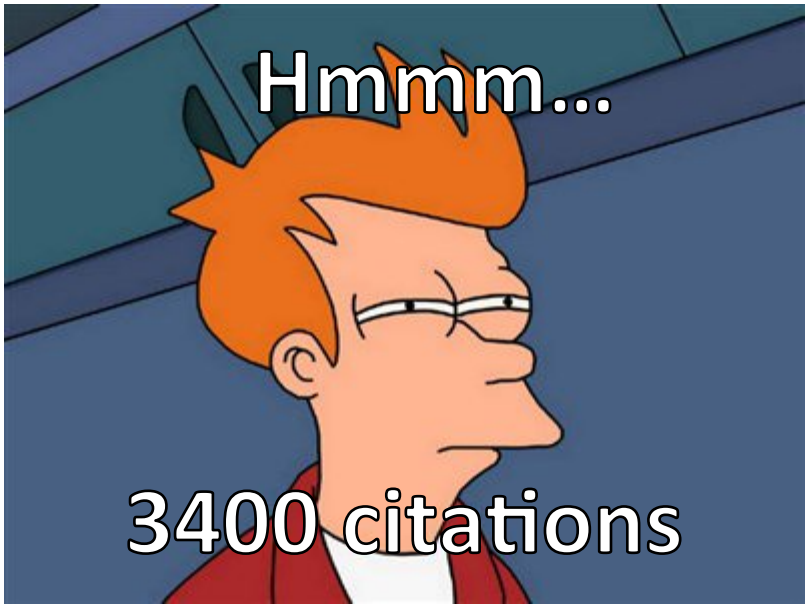


Geman et al. (1992)

Experiments with neural network

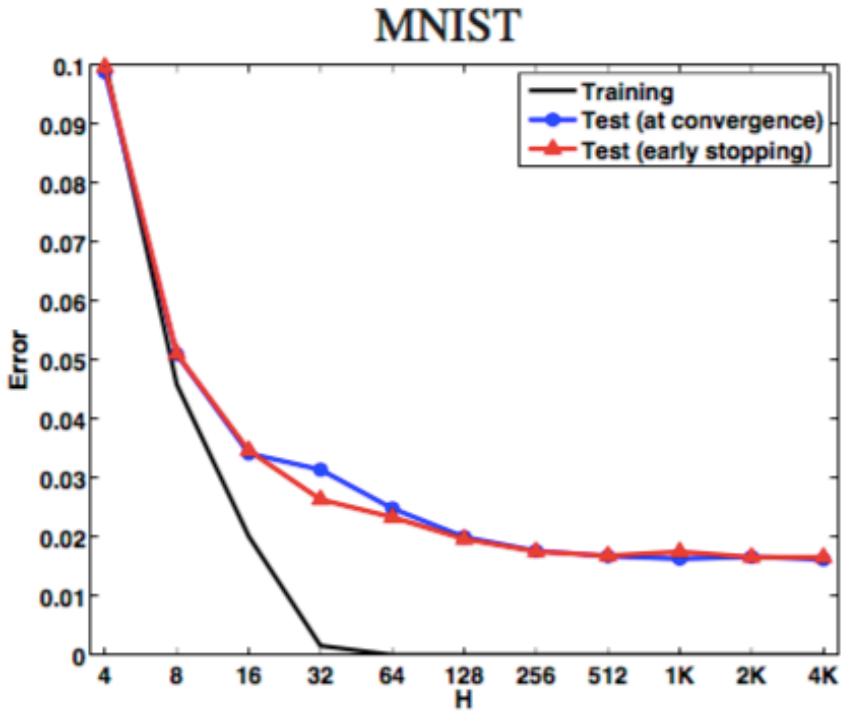
The basic trend is what we expect: bias falls and variance increases with the number of hidden units. The effects are not perfectly demonstrated (notice, for example, the dip in variance in the experiments with the largest numbers of hidden units), presumably because the phenomenon of overfitting is complicated by convergence issues and perhaps also by our decision to stop the training prematurely. The lowest achievable mean-squared error appears to be about 2.

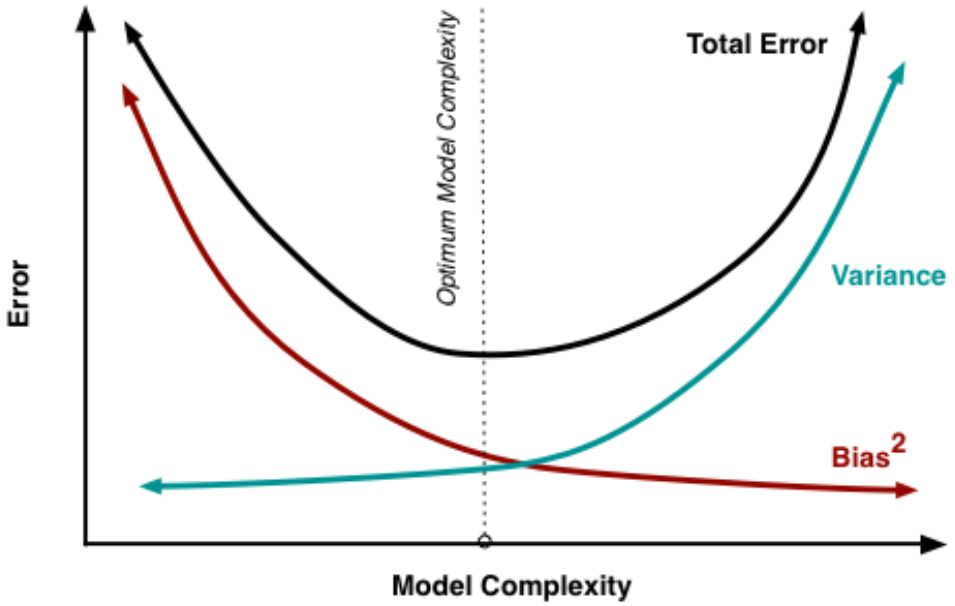




Geman et al. (1992)

Something wrong with this picture





[Neyshabur et al. (2015)]

Outline

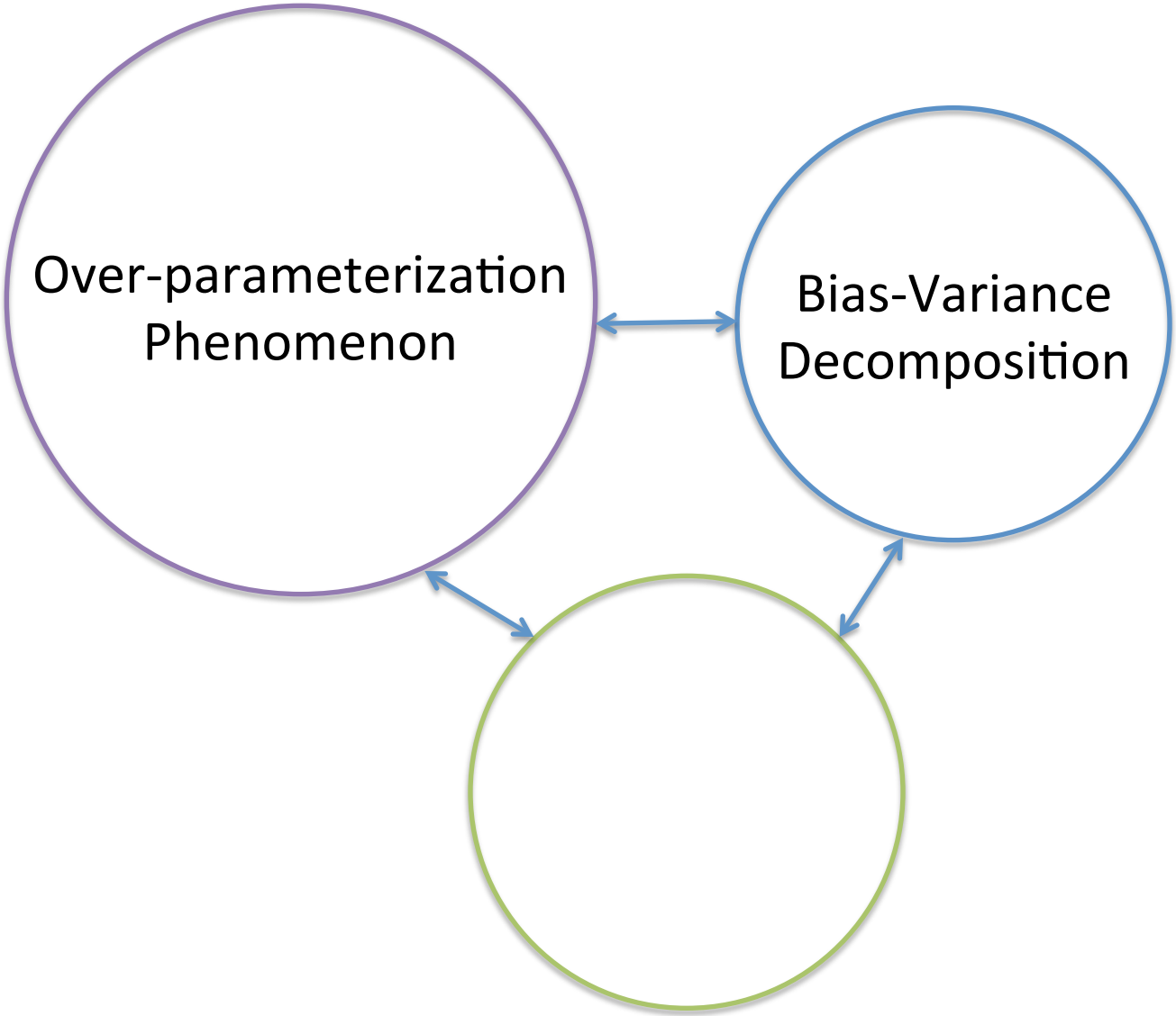
Part 1: Contradiction between traditional complexity measures and over-parameterization

Part 2: Bias-variance decomposition

Part 3: Over-parameterization and variance

Part 4: Zhang et al. (2017) via bias-variance decomposition

Triad of Observations

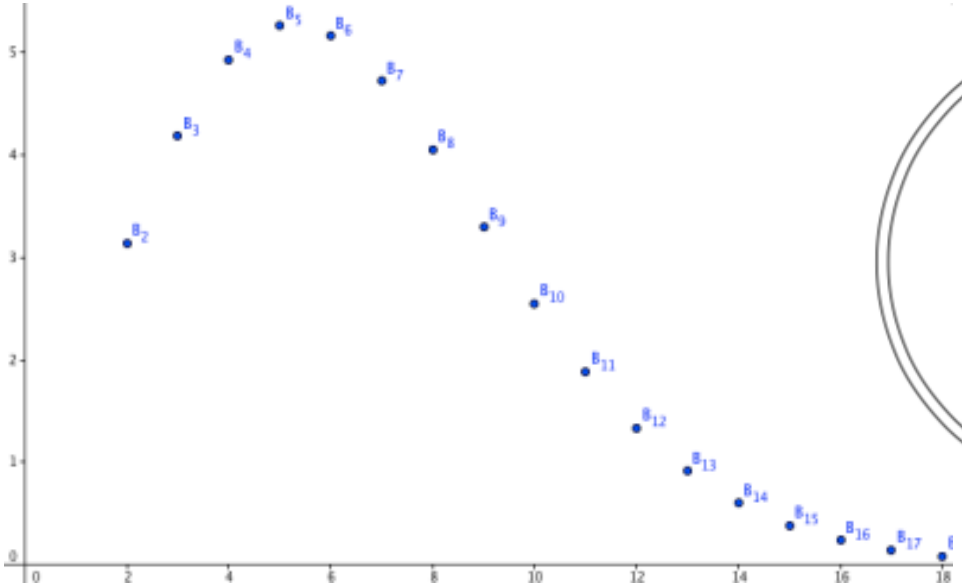


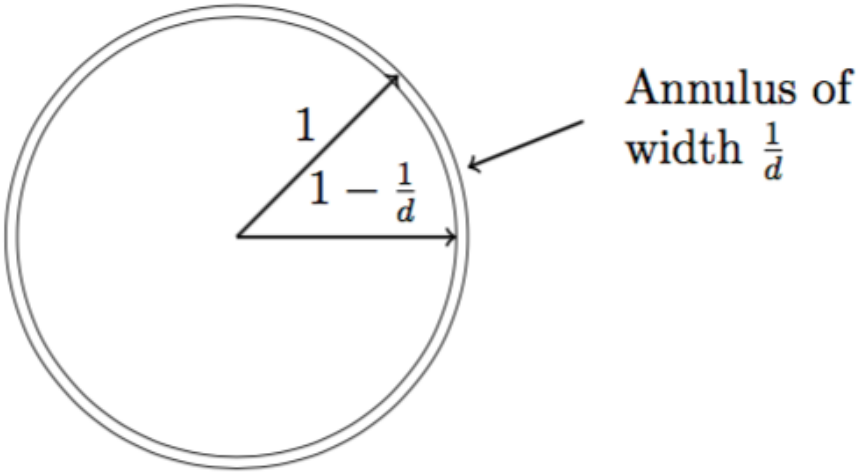
Intuitions fail in high dimensions: High dimensional ball example

area of unit circle $(d = 2)$: $\pi r^2 = \pi$

volume of unit ball $(d = 3)$: $\frac{4}{3}\pi r^3 = \frac{4}{3}\pi$

volume of d dimensional unit ball: $\frac{2\pi^{\frac{d}{2}}}{d\,\Gamma(\frac{d}{2})}$





Guruswami and Kannan (2012); Blum et al. (2018)

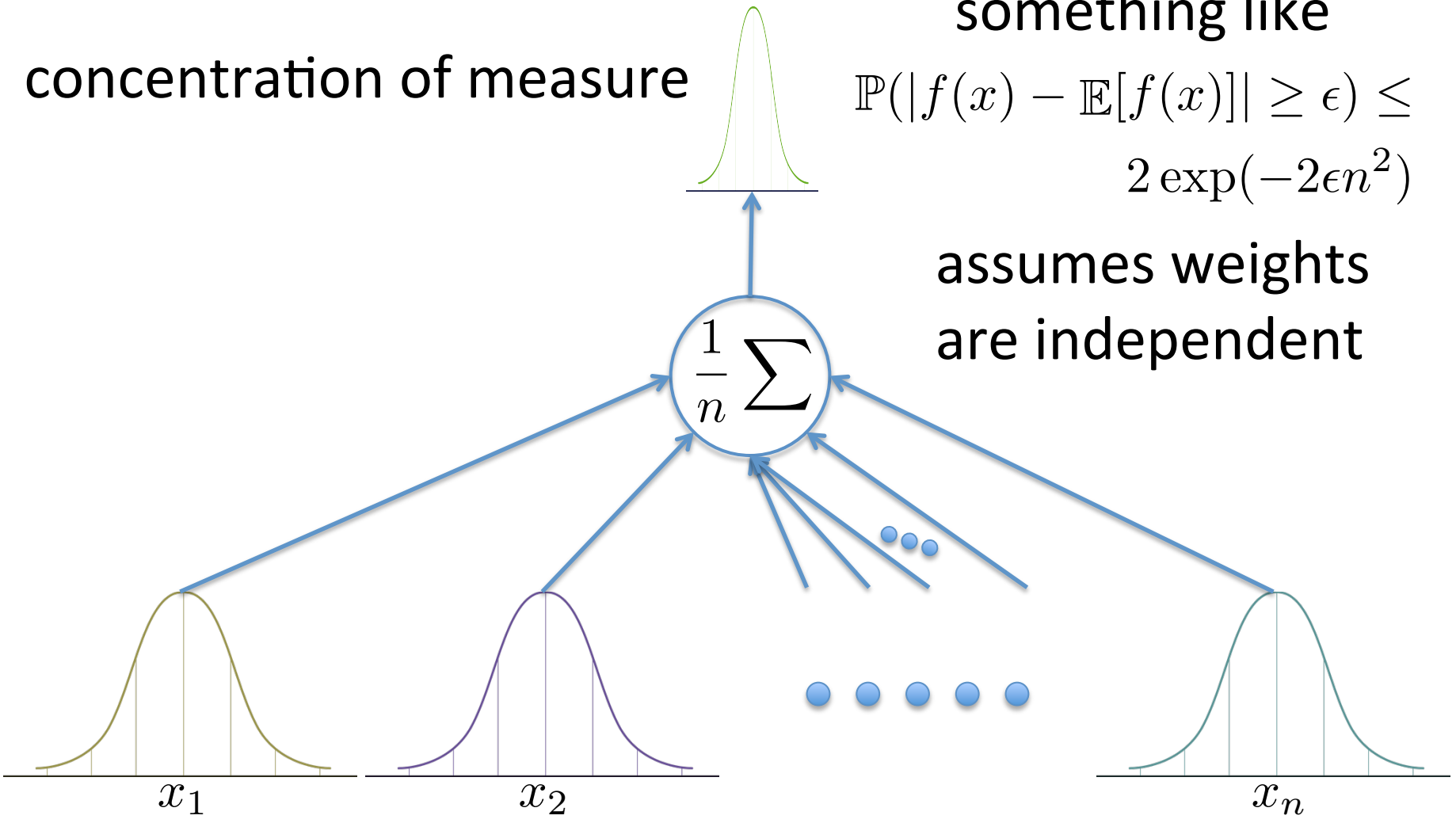
Blessing of Dimensionality

concentration of measure

something like

$$\mathbb{P}(|f(x) - \mathbb{E}[f(x)]| \geq \epsilon) \leq 2\exp(-2\epsilon n^2)$$

assumes weights are independent



Donoho (2000)

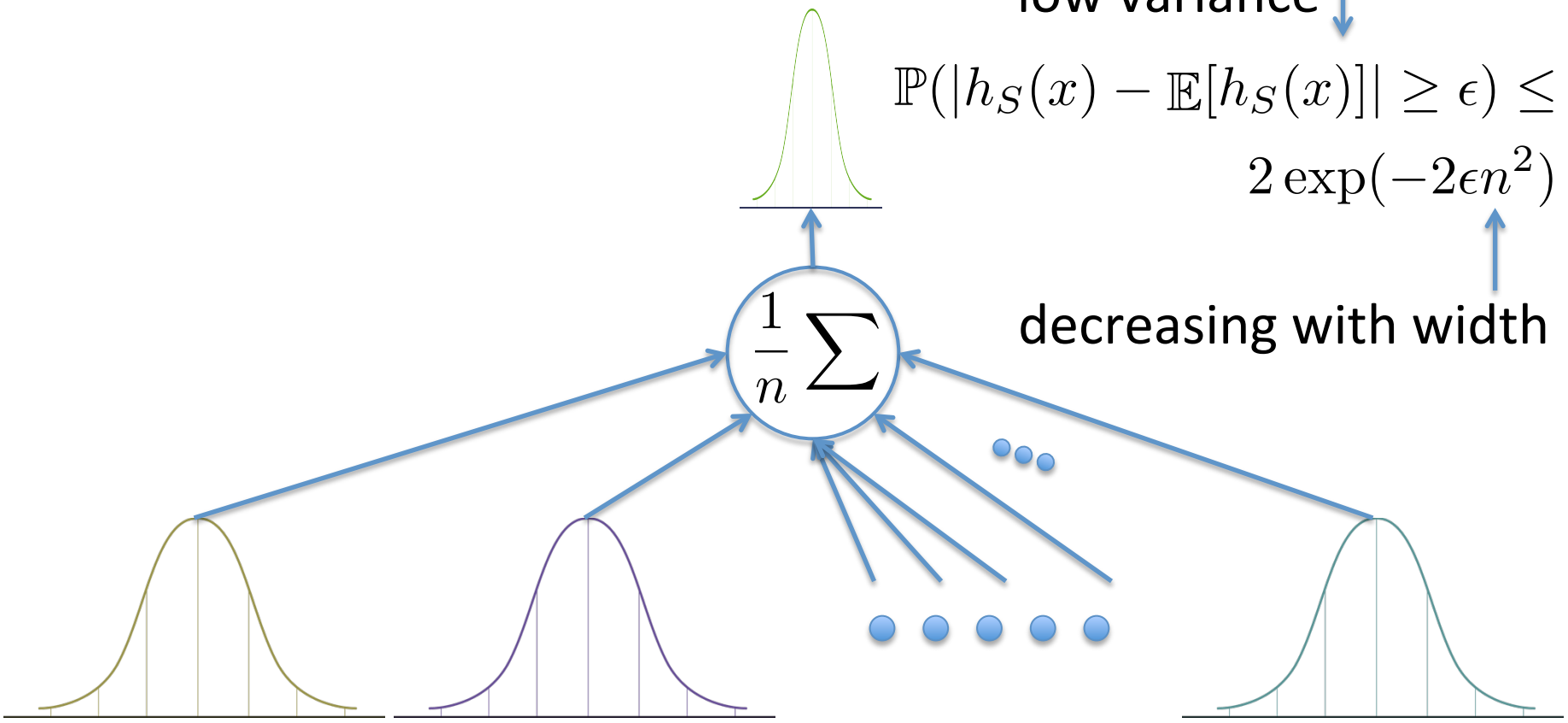
Random neural networks have low variance

Recall: $\mathbb{E}[R(h_S)] = \mathop{\mathbb{E}}_{(x,y)}[\mathrm{Bias}^2(h_S(x)) + \mathrm{Var}(h_S(x))]$

low variance

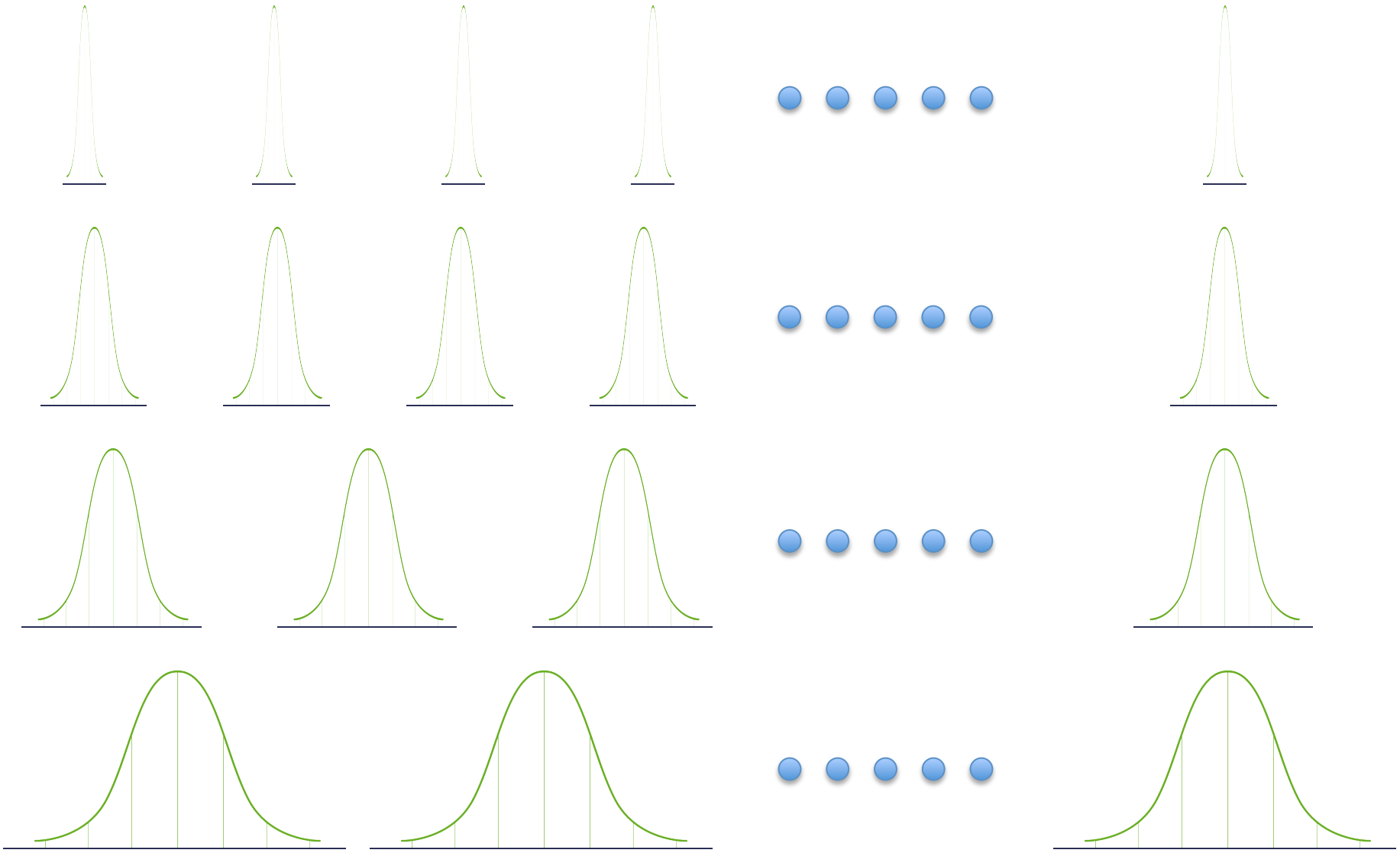
$$\mathbb{P}(|h_S(x) - \mathbb{E}[h_S(x)]| \geq \epsilon) \leq 2\exp(-2\epsilon n^2)$$

decreasing with width

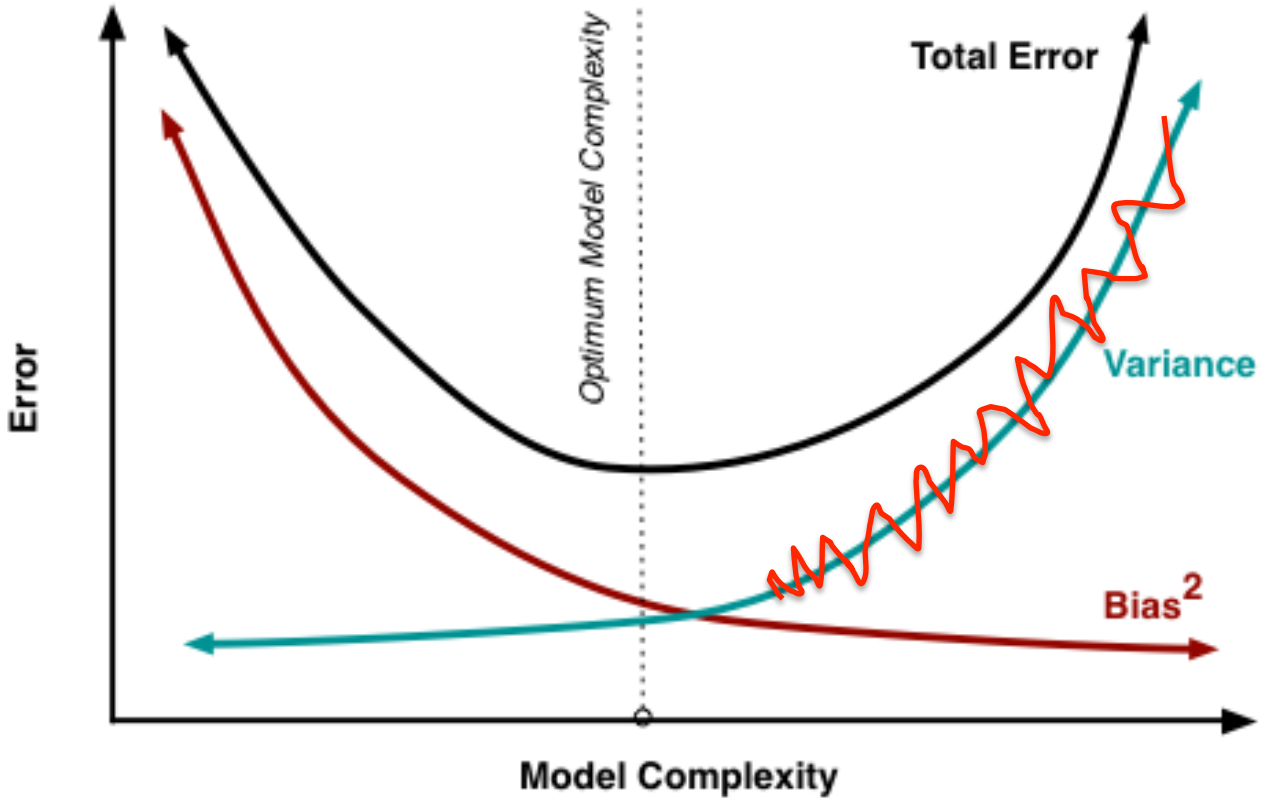


last hidden layer of a neural network

Over-parameterization: Width vs. Depth



Revisiting Intuitions



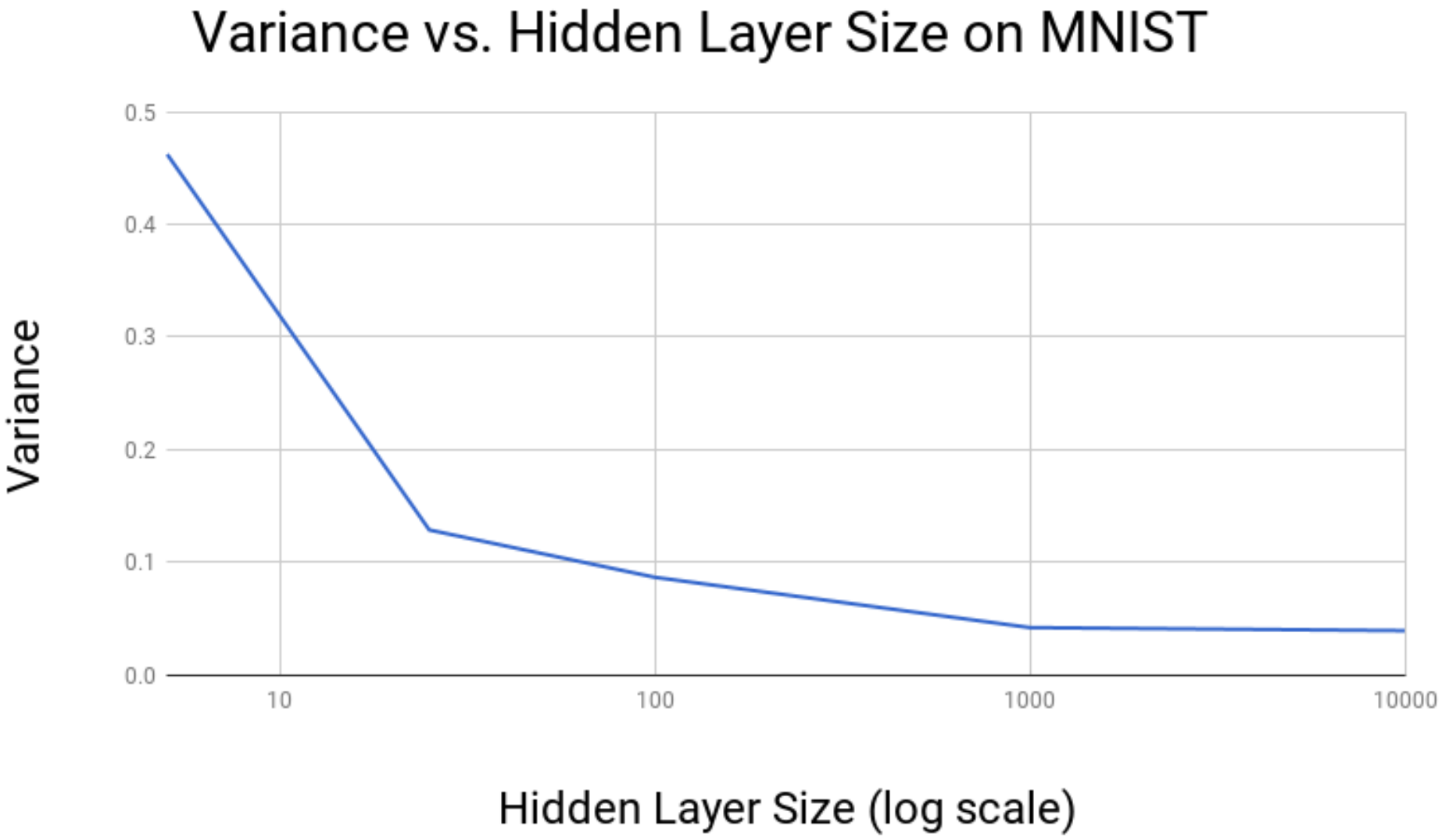
Previous slide suggests variance should be decreasing with increasing over-parameterization

Randomness Modeling and Independence Assumptions

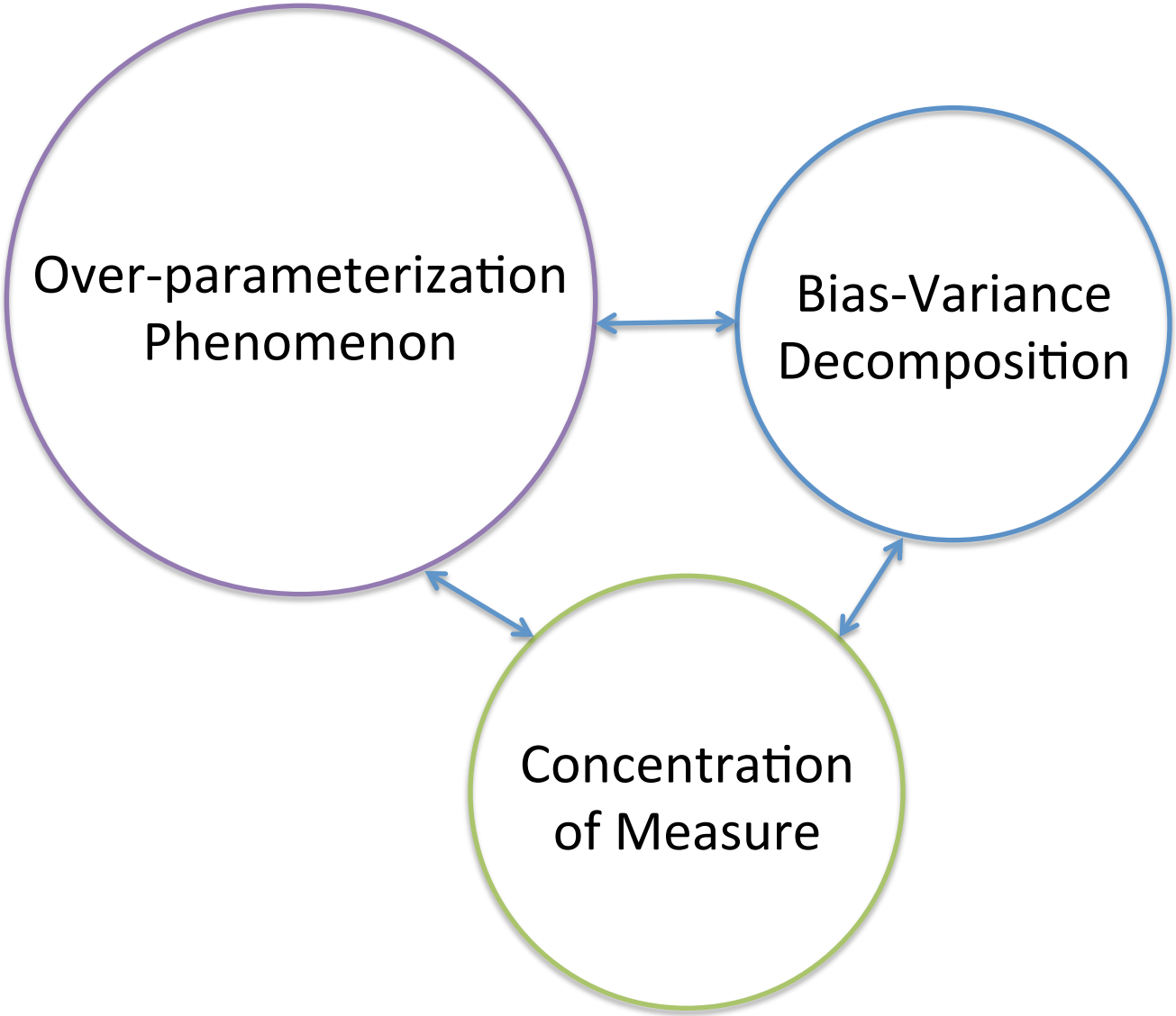
- Weights are random!
  - randomness in data sampling
  - randomness in gradient sampling if mini-batching
  - randomness in initialization
- Results with these kinds of assumptions have surprising degree of generality in mean field theory
- Correlations between variables diminishes with increasing dimensionality

Related work: Choromanska et al. (2015); Schoenholz et al. (2017); Pennington and Worah (2017); Pennington and Bahri (2017)

Preliminary Empirical Results



Triad of Observations



Outline

Part 1: Contradiction between traditional complexity measures and over-parameterization

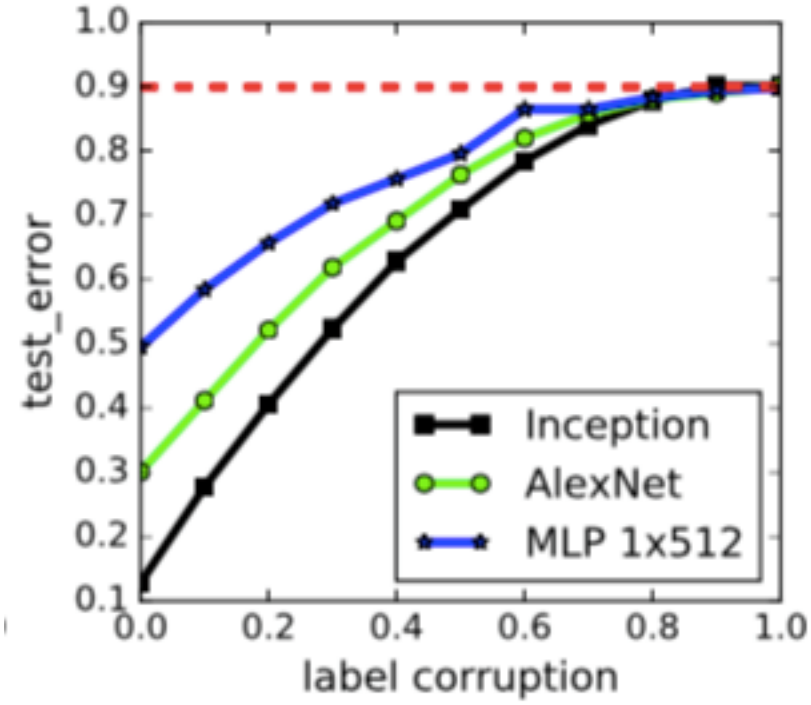
Part 2: Bias-variance decomposition

Part 3: Over-parameterization and variance

Part 4: Zhang et al. (2017) via bias-variance decomposition

Zhang et al. (2017) Recap

- attacked generalization bounds that don't depend on data
- can arbitrarily increase test error by only changing the data
- bias-variance decomposition doesn't depend on data



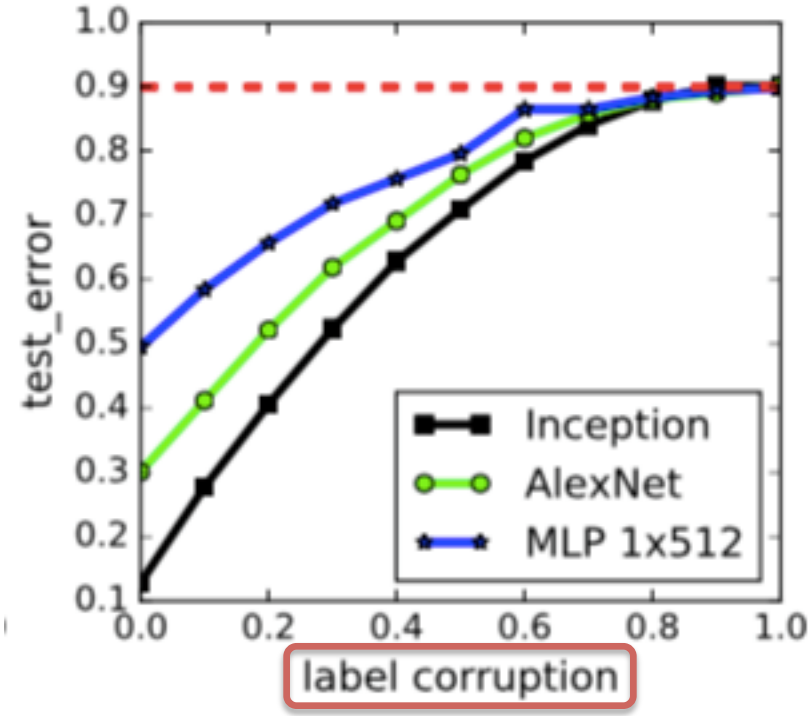
Zhang et al. (2017): Via Bias-Variance Decomposition

$$\mathbb{E}[R(h_S)] = \mathop{\mathbb{E}}_{(x,y)}[\text{Bias}^2(h_S(x)) + \text{Var}(h_S(x))] + \text{Var}(\epsilon)$$

Over-parameterization Phenomenon

Bias-Variance Decomposition

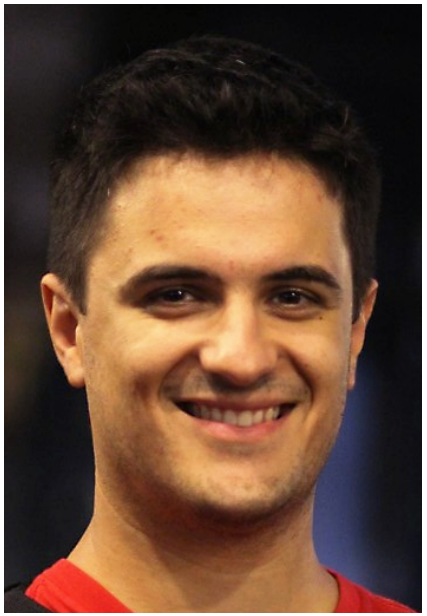
Concentration of Measure



Future Work and Connections

- tighter random matrix/network bounds
- experiment showing decreasing correlation with over-parameterization
- derive analog in analytical learning theory framework
- connection to stability
- think about if any PAC-Bayes relation
- DL Theory Reading Group on Mondays at 1

Thanks for coming!



Appendix

Reducible and Irreducible Error

$$
\begin{aligned}
\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\epsilon}[R(h_S)] &= \mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\epsilon}\mathop{\mathbb{E}}_{(x,y)}[(h_S(x)-y)^2] && \text{(squared loss)} \\
&= \mathop{\mathbb{E}}_{(x,y)}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\epsilon}[(h_S(x)-y)^2] && \text{(Fubini's theorem)} \\
&= \mathop{\mathbb{E}}_{(x,y)}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\epsilon}[(h_S(x)-(f(x)+\epsilon))^2] \\
&= \mathop{\mathbb{E}}_{(x,y)}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\epsilon}[((h_S(x)-f(x))-\epsilon)^2] \\
&= \mathop{\mathbb{E}}_{(x,y)}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\epsilon}[(h_S(x)-f(x))^2+\epsilon^2-2(h_S(x)-f(x))\epsilon] \\
&= \mathop{\mathbb{E}}_{(x,y)}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\epsilon}[(h_S(x)-f(x))^2+\epsilon^2] \\
& && \text{(ϵ independent of S and $\mathbb{E}[\epsilon]=0$)} \\
&= \mathop{\mathbb{E}}_{(x,y)}\mathop{\mathbb{E}}_{S}[(h_S(x)-f(x))^2]+\mathrm{Var}(\epsilon) && (\mathbb{E}[\epsilon]=0)
\end{aligned}
$$

Bias-Variance Decomposition

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)} \mathop{\mathbb{E}}_{S}[(h_S(x) - f(x))^2] \\
&\quad = \mathop{\mathbb{E}}_{(x,y)} \mathop{\mathbb{E}}_{S}[h_S(x)^2 - 2h_S(x)f(x) + f(x)^2] \\
&\quad = \mathop{\mathbb{E}}_{(x,y)} \mathop{\mathbb{E}}_{S}\left[h_S(x)^2 - 2h_S(x)f(x) + f(x)^2 + \left(\mathop{\mathbb{E}}_{S}[h_S(x)]^2 - \mathop{\mathbb{E}}_{S}[h_S(x)]^2\right)\right] \\
&\quad = \mathop{\mathbb{E}}_{(x,y)} \mathop{\mathbb{E}}_{S}\left[\left(\mathop{\mathbb{E}}_{S}[h_S(x)]^2 - 2h_S(x)f(x) + f(x)^2\right) + \left(h_S(x)^2 - \mathop{\mathbb{E}}_{S}[h_S(x)]^2\right)\right] \\
&\quad = \mathop{\mathbb{E}}_{(x,y)}\left[\left(\mathop{\mathbb{E}}_{S}[h_S(x)]^2 - 2\mathop{\mathbb{E}}_{S}[h_S(x)]f(x) + f(x)^2\right) + \left(\mathop{\mathbb{E}}_{S}[h_S(x)^2] - \mathop{\mathbb{E}}_{S}[h_S(x)]^2\right)\right] \\
&\quad = \mathop{\mathbb{E}}_{(x,y)}\left[\left(\mathop{\mathbb{E}}_{S}[h_S(x)] - f(x)\right)^2 + \mathrm{Var}(h_S)\right] \\
&\quad = \mathop{\mathbb{E}}_{(x,y)}[\mathrm{Bias}^2(h_S(x)) + \mathrm{Var}(h_S(x))]
\end{aligned}
$$