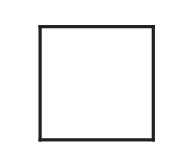


The Double Descent Phenomenon in Machine Learning

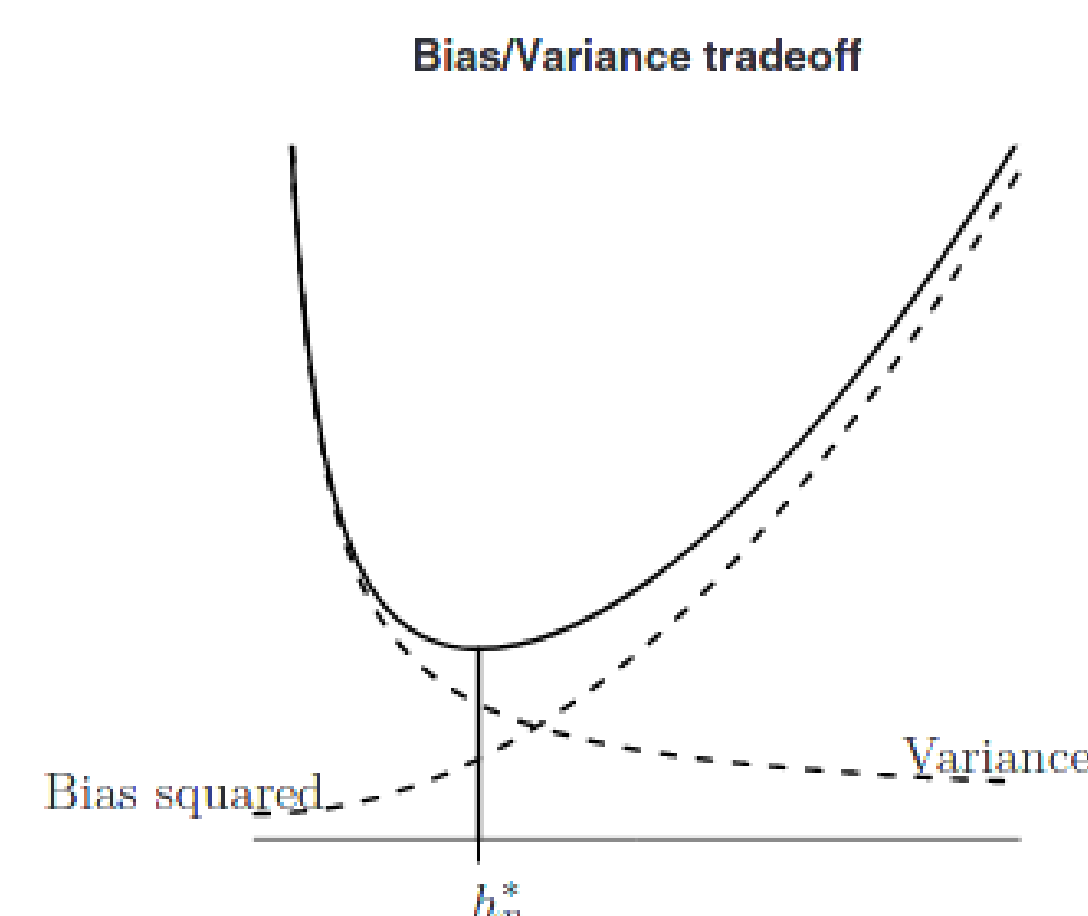
Josh Gammage (REU Student), Mike Jeong (REU Student), Wenjing Liao (Mentor)



Motivation

Standard Bias-Variance Tradeoff

- The standard bias-variance tradeoff suggests test risk will decrease with increasing model complexity count down to a minimum after which it will begin increasing (Tsybakov, 2009)



Surprises in Machine Learning Models

- However, overparameterized machine learning models can have similar or better performance to optimal underparameterized models, contradicting the bias-variance tradeoff

Neural Network Simulation

Data

- MNIST handwritten digits 0-9
- $n = 4,000$ training samples (28 by 28 grayscale images for $d = 784$) in $K = 10$ classes

Architecture

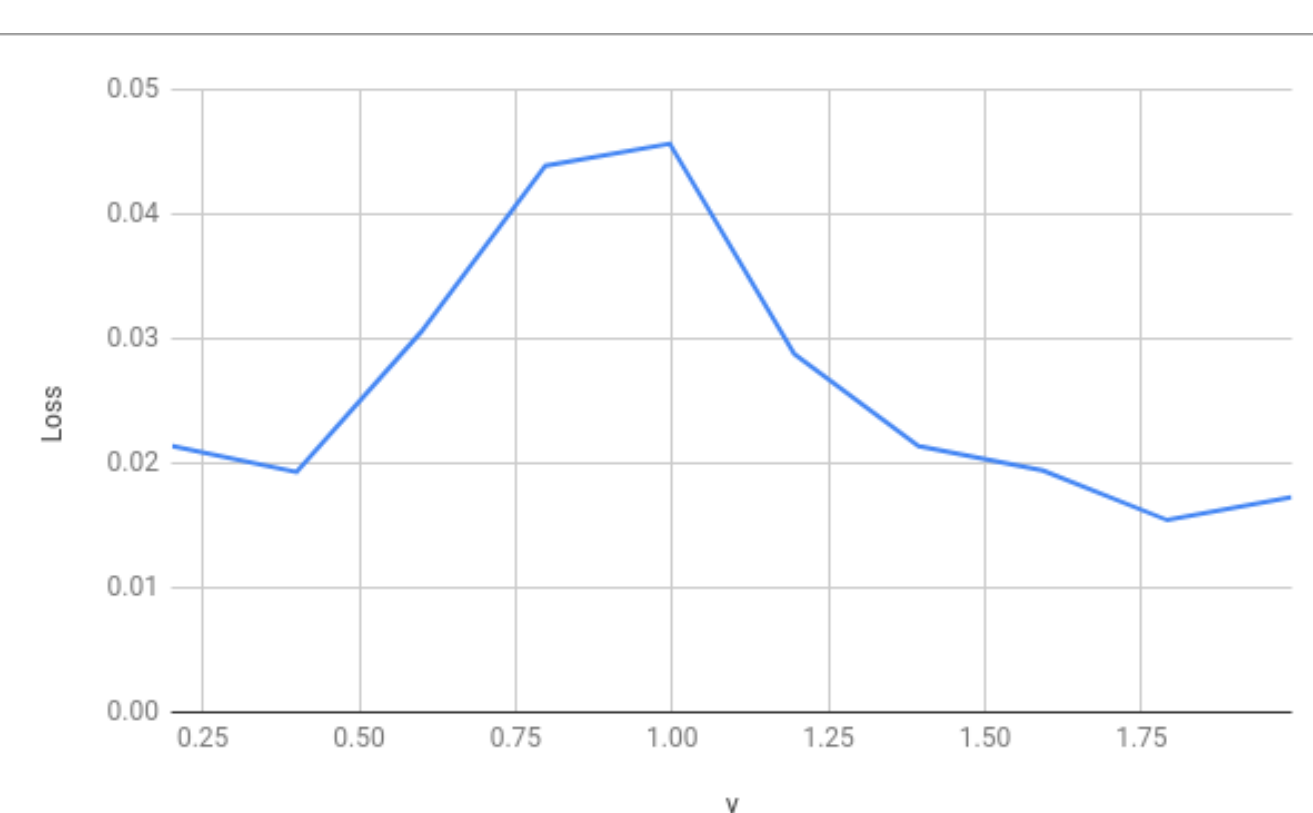
- One hidden layer of H neurons
- ReLU activation function
- Number of parameters $p = (d + 1)H + (H + 1)K$
- Overparameterization ratio $\gamma = p/n$

Training

- $n = 4,000$ training samples
- Trained until 6,000 epochs or 0 training loss for networks below interpolation threshold
- SGD with 0.95 momentum and 0.001 initial learning rate
- Learning rate decreased by 10% every 500 epochs

Results

- Test loss averaged over 3 randomly initialized, fully trained networks



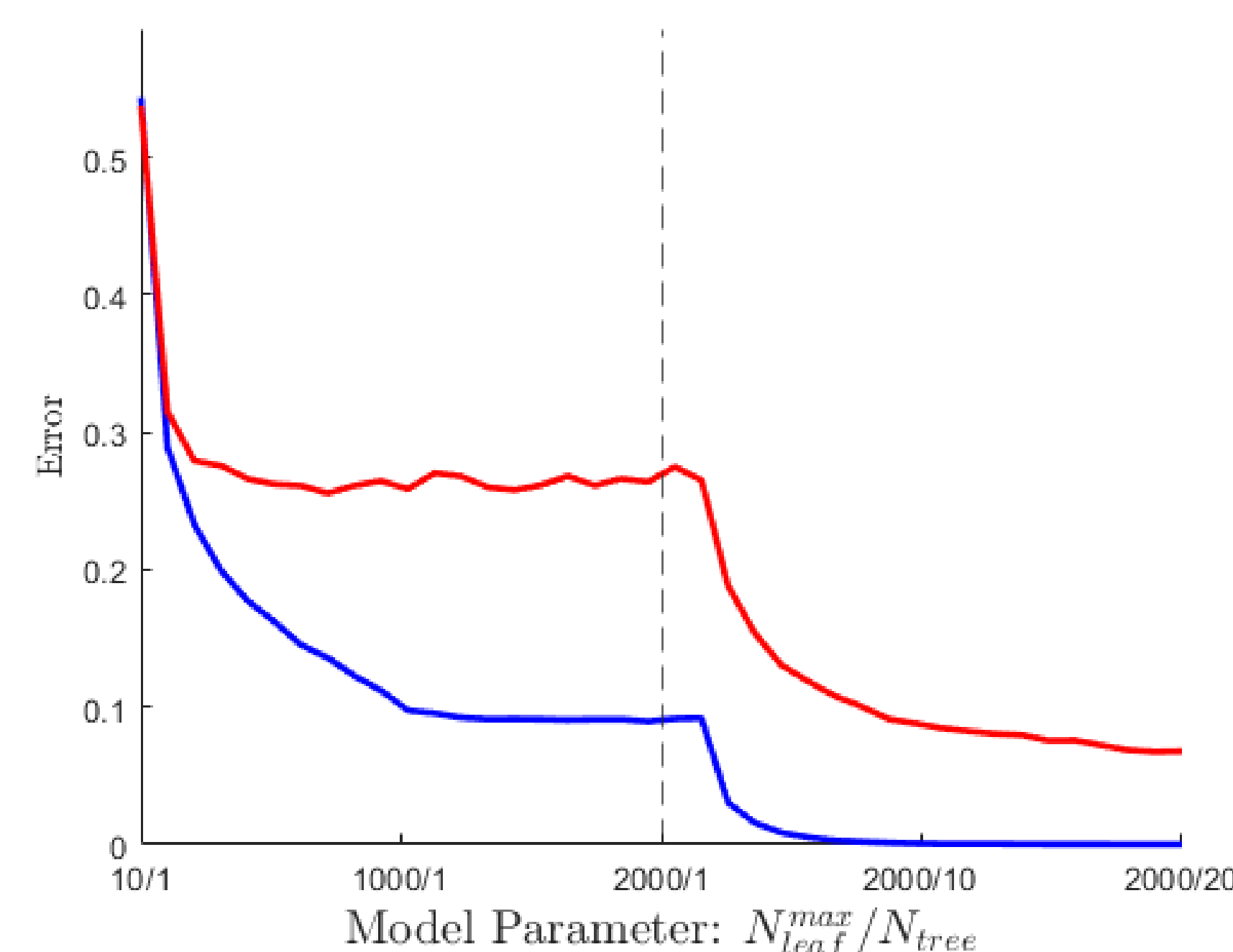
Random Forests

Data

- MNIST handwritten digits 0-9, $n = 10000$ training samples, 1000 test samples

Complexity

- Complexity is controlled by number of trees and maximum number of leaves allowed for each tree [1]
- $p = N_{leaf}^{max} / N_{tree}$



Training error (blue) and testing error (red) for different model parameter p , averaged over 10 trials.

Neural Networks and Least Squares

Connection Between Neural Networks and Linear Regression

Consider the neural network $f(\cdot, \theta) : \mathbb{R}^d \rightarrow \mathbb{R}$ with weights $\theta \in \mathbb{R}^p$ such that $z \mapsto f(z, \theta)$.

- For networks with a large number of parameters p , training changes θ by only a small amount relative to some initialization θ_0
- Therefore f can be expressed as $z \mapsto \nabla_{\theta} f(z, \theta_0)^T \beta$
- This linearization can be accurate if $p > n$ and f remains a neural network when scaled
- It converges to the least squares solution for β

Model

- n i.i.d. $x_i \in \mathbb{R}^p$ sampled from P_x
- $\mathbb{E} x_i = 0$ and $\text{Cov}(x_i) = \Sigma$
- n i.i.d. $\varepsilon_i \sim P_{\varepsilon}$ where $\mathbb{E} \varepsilon_i = 0$ and $\text{Var} \varepsilon_i = \sigma^2$

$$y_i = x_i^T \beta + \varepsilon_i$$

Problem Estimate β given (x_i, y_i)

- Risk of the estimator $\hat{\beta}$ is

$$R_X(\hat{\beta}, \beta) = \mathbb{E}_{\varepsilon, x_0} [(x_0^T \hat{\beta} - x_0^T \beta)^2 | X]$$

- Overparameterization ratio $\gamma = p/n$
- Signal to noise ratio $SNR = \|\beta\|_2^2 / \sigma^2$

Main Theorems

Least Squares Estimator

$$\hat{\beta} = (X^T X)^{\dagger} X^T y$$

where A^{\dagger} is the Moore Penrose inverse (pseudoinverse) of the matrix A , $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times p}$, $y = \{y_i\}_{i=1}^n \in \mathbb{R}^n$, and $\varepsilon = \{\varepsilon_i\}_{i=1}^n \in \mathbb{R}^n$

Proposition 1 (Gradient descent and least-squares estimator). Initialize $\beta^{(0)} = 0$, and consider running gradient descent on the least square loss, yielding iterates

$$\beta^{(k)} = \beta^{(k-1)} + t X^T (y - X \beta^{(k-1)})$$

where we take $0 < t \leq 1/\lambda_{max}(X^T X)$. Then $\lim_{k \rightarrow \infty} \beta^{(k)} = \hat{\beta}$, the min-norm least squares estimator.

Bias Variance Decomposition

$$R_X(\hat{\beta}, \beta) = \underbrace{\|\mathbb{E}_{\varepsilon}[\hat{\beta} | X] - \beta\|_{\Sigma}^2}_{B_X(\hat{\beta}, \beta)} + \underbrace{\text{Tr}(\text{Cov}(\hat{\beta} | X) \Sigma)}_{V_X(\hat{\beta}, \beta)}$$

where $\|x\|_{\Sigma}^2 = x^T \Sigma x$

Underparameterized case ($\gamma < 1$) Bias is always zero so risk is entirely determined by the variance.

Theorem 2 (Asymptotic risk). Assume $x \sim P_x$ has i.i.d. entries with zero mean, unit variance, and a finite fourth moment of order $4 + \eta$, for some $\eta > 0$. Also assume $\|\beta\|_2^2 = r^2$ for all n, p . Then, for the min-norm least-squares estimator $\hat{\beta}$, as $n, p \rightarrow \infty$ such that $p/n \rightarrow \gamma \in (0, \infty)$, it holds almost surely that

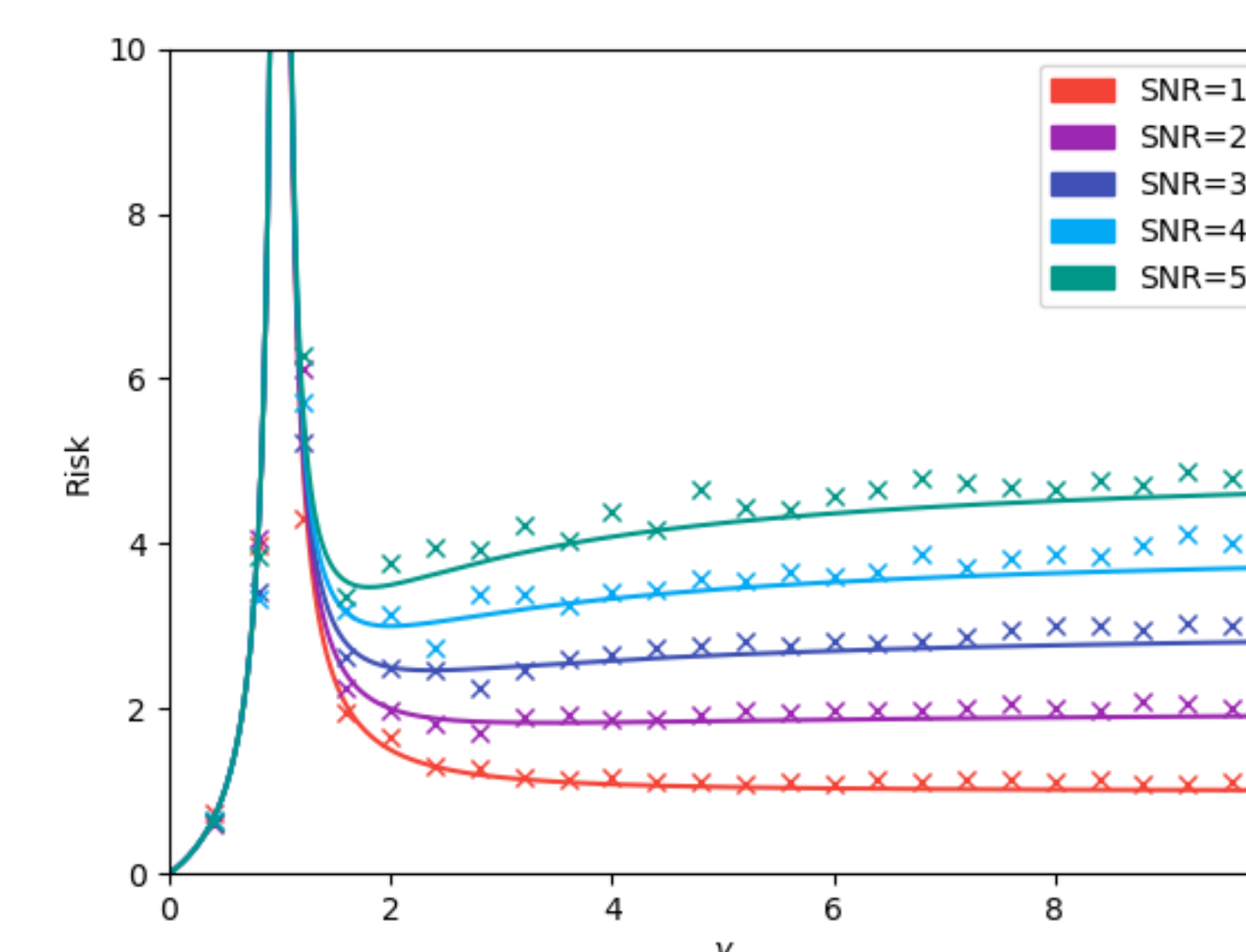
$$B_X(\hat{\beta}, \beta) \rightarrow r^2 \left(1 - \frac{1}{\gamma}\right)$$

$$V_X(\hat{\beta}, \beta) \rightarrow \sigma^2 \frac{\gamma}{\gamma - 1}$$

Hence,

$$R_X(\hat{\beta}, \beta) \rightarrow \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1 \\ r^2 \left(1 - \frac{1}{\gamma}\right) + \sigma^2 \frac{\gamma}{\gamma-1} & \text{for } \gamma > 1 \end{cases}$$

Comparison of Theoretical and Actual Risk



Theoretical (lines) and actual (points) risk of least squares estimator averaged over 5 trials with $n = 200$ and $p = \lfloor \gamma n \rfloor$

Conclusion

Overparameterized models can and often do perform better than underparameterized models

- The bias-variance tradeoff is still a good rubric for very large datasets where overparameterization is prohibitively expensive
- Computational resources are constantly increasing, which has recently made overparameterization on datasets like ImageNet (10^6 examples and 10^3 classes) practical, where the current best model is overparameterized
- If possible, overparameterized models are often the best approach, but for larger datasets the bias-variance tradeoff is the best rubric

Proof ingredients

- Marchenko-Pastur Law
- Spectral measure
- Bai-Yin theorem
- Stieltjes transform

Future Work

Where does double descent appear?

- Linear regression
- Linear Discriminant Analysis
- Logistic regression

Open problems It is not known whether or not the following models exhibit double descent

- Quadratic Discriminant Analysis (work on Linear Discriminant Analysis may be extended)
- Random forests
- Support Vector Machines
- Neural networks with nonlinear activation

References

References

- [1] Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal "Reconciling Modern Machine-Learning Practice and the Classical Bias-Variance Trade-Off." Proceedings of the National Academy of Sciences, vol. 116, no. 32, 2019, pp. 15849-15854.
- [2] Trevor Hastie, Andrea Montanari, Saharon Rosset, Ryan J. Tibshirani "Surprises in High-Dimensional Ridgeless Least Squares Interpolation." (2020 preprint archive).
- [3] Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009. URL <https://doi.org/10.1007/b13794>. Revised and extended from the, 9:10, 2004.