Introduction to Double Descent

CS 229
Summer 2022

Overview

- Background
- Double Descent Phenomenon
- Possible Explanations
- Related Research

Data generating process

Let the training dataset be:

$$\mathcal{S} = \{x(i), y(i)\}_{i=1}^{n}$$

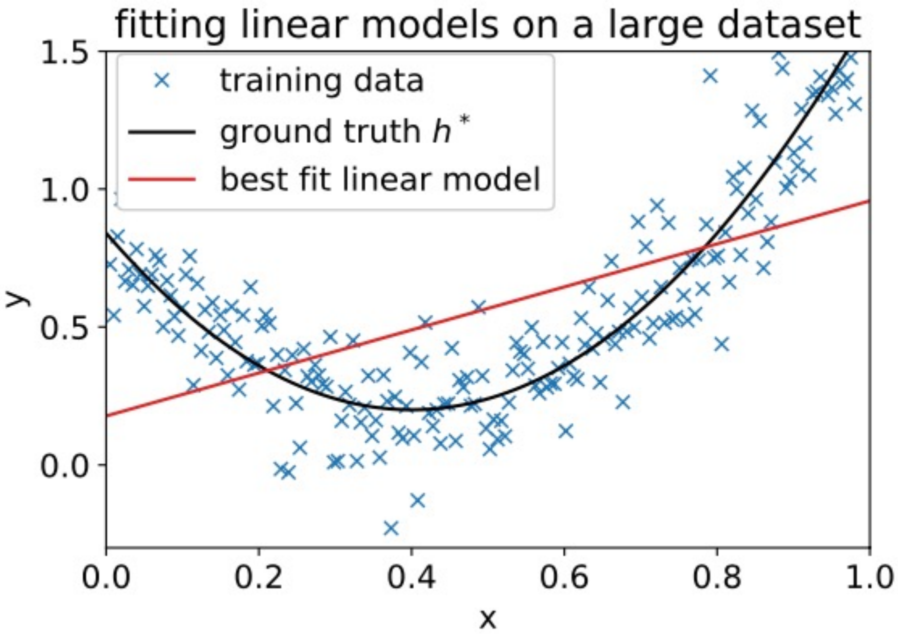
Such that:

$$y(i) = h^*(x(i)) + \xi(i)$$
$$\xi(i) \sim \mathcal{N}(0, \sigma^2)$$

Denote the model trained on the dataset as: $\hat{h}_{\mathcal{S}}$

Recap of Bias

- Data generated by a quadratic h^* function:



Recap of Bias

Bias = Test error when model is given infinite training data

Equivalently, let h_{avg} be average of models trained on many different datasets *S*:

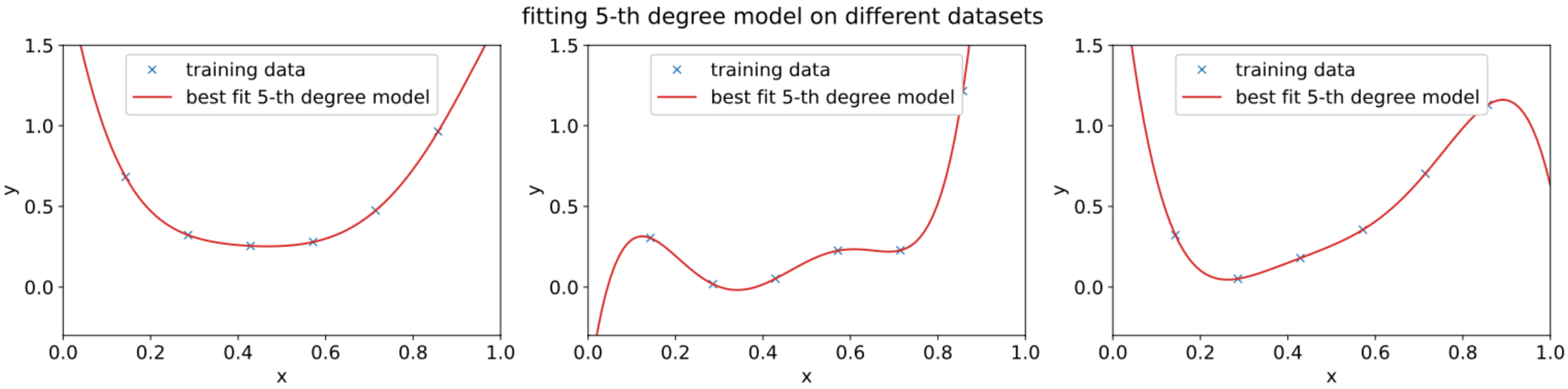
$$h_{avg}(x) = \mathbb{E}_{\mathcal{S}}[h_{\mathcal{S}}(x)]$$

Then,

$$\mathrm{Bias}^2 = (h^*(x) - h_{avg}(x))^2$$

Recap of Variance

Depending on the training dataset S we draw, our learned model can look very different



Recap of Variance

Variance = Measure of deviation of a prediction function varies from average

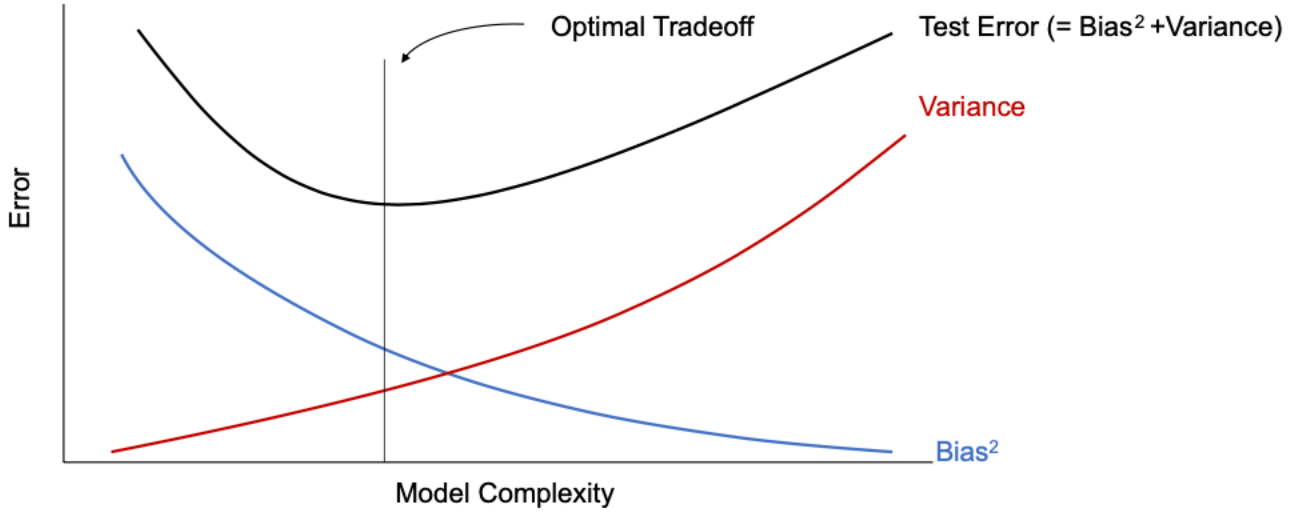
The term captures sensitivity of model to randomness in training data:

$$\text{Variance} = \mathbb{E}[(h_{\mathcal{S}}(x) - h_{avg}(x))^2]$$

Bias-Variance Trade-off

$$\mathrm{MSE}(x) = \sigma^2 + \mathbb{E}[(h^\star(x) - h_S(x))^2]$$

$$= \sigma^2 + \mathrm{Bias}^2 + \mathrm{Variance}$$



What is model complexity?

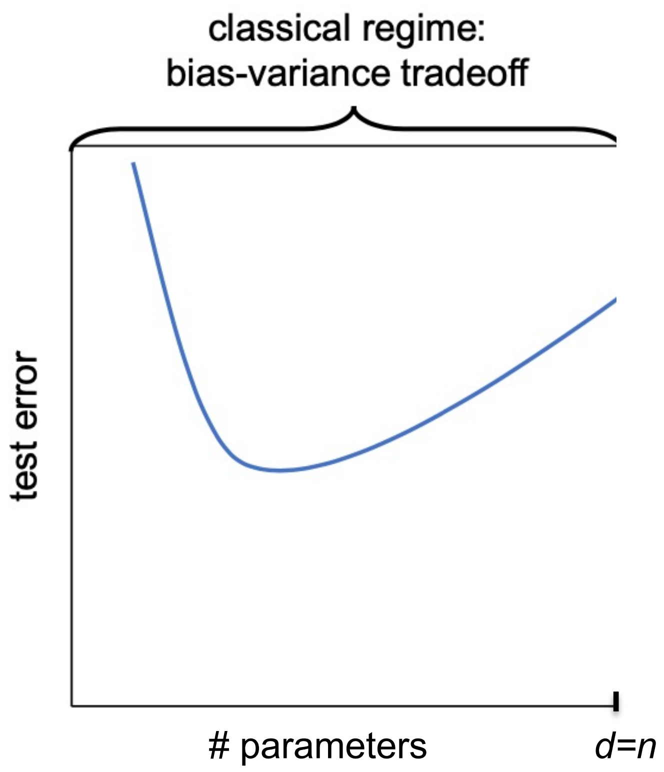
This can be ambiguous:

- Number of parameters? $y = \theta_0 + \theta_1 \cdot x$ vs. $y = \theta_0 + \theta_1 \cdot x + \theta_2 \cdot x^2$
- Structure of model? $y = \theta_0 + \theta_1 \cdot x + \theta_2 \cdot x^2$ $y = \theta_0 * \sin(\theta_1 \cdot x)$

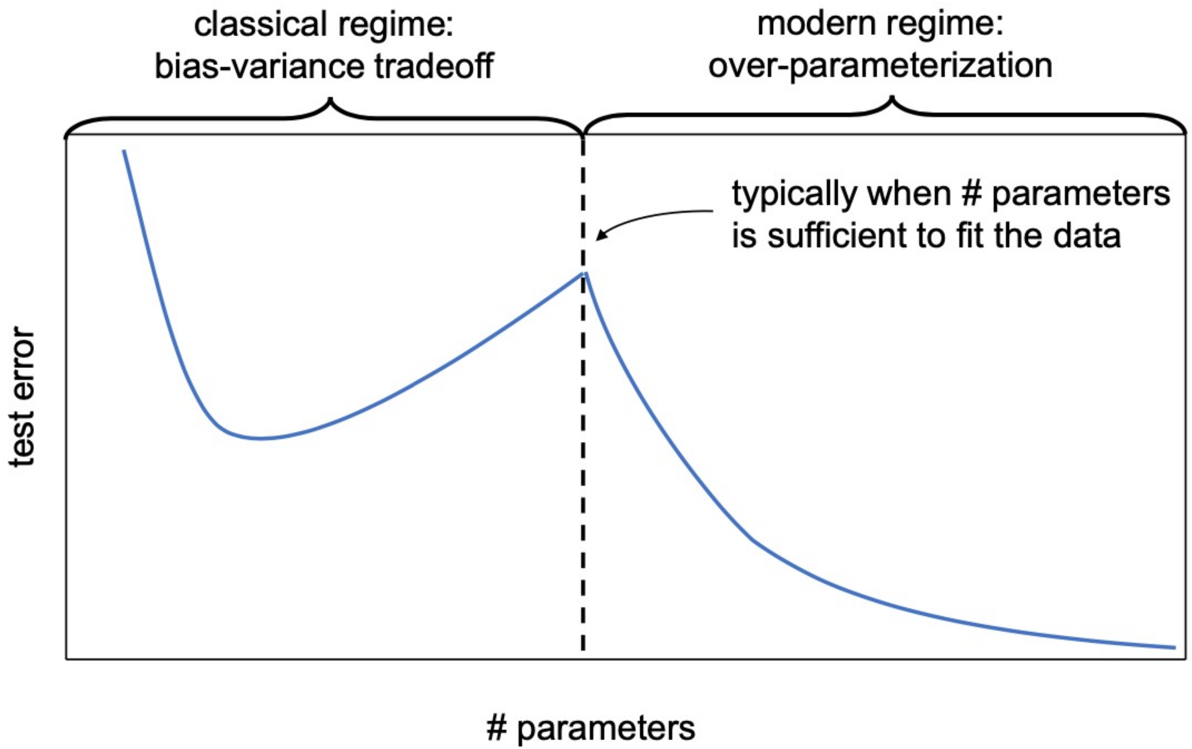
Complexity = Number of parameters

Let *d* represent the number of parameters

Let *n* represent the training dataset size



The test error starts decreasing again!



It gets even more interesting…

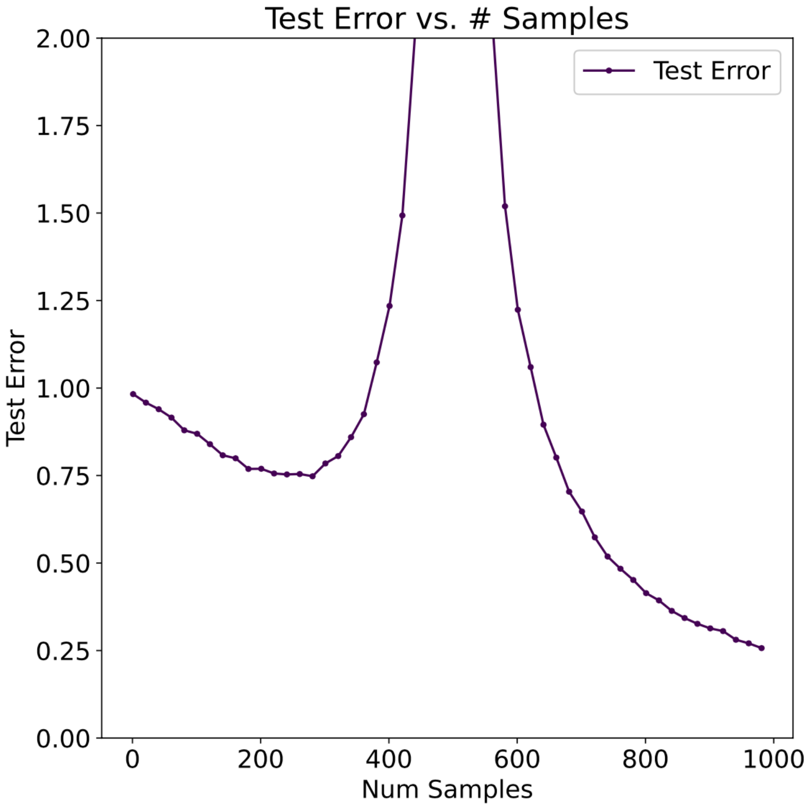
Data distribution (x, y) is:

$$x \sim \mathcal{N}(0, I_d)$$

$$y \sim x^T \beta + \mathcal{N}(0, \sigma^2)$$

- $\sigma = 0.5$
- $d = 500$
- $\| \beta \|_2 = 1$

Data is isotropic gaussian $(\Sigma \propto I)$



A Statistical Perspective

This phenomenon is interesting for a slightly different reason:

Example: Sample mean in statistics:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

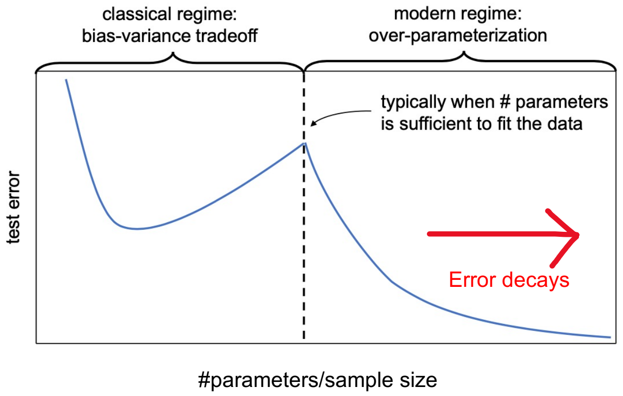
$$\overline{X} \xrightarrow{p} \mu$$

Why is this happening?

Nobody knows for sure…

One idea is gradient descent is implicitly regularizing the learning process (e.g., gradient descent acts like L2-norm regularization)

The test error seems to get smaller in the over-parameterized regime



So, why the peak? A case in linear regression

Long story short. For the data generating process shown below:

$$x \sim \mathcal{N}(0, I_d) \qquad\qquad y \sim x^T\beta + \mathcal{N}(0, \sigma^2) \quad ,$$

$$\hat{\beta} = X^{\dagger}y = \begin{cases} \underset{\beta : X\beta = y}{\operatorname{argmin}} ||\beta||^2 & \text{when } n \leq d \quad \textbf{(“Overparameterized”)} \\ \underset{\beta}{\operatorname{argmin}} ||X\beta - y||^2 & \text{when } n > d \quad \textbf{(“Underparameterized”)} \end{cases}$$

Let’s break down the overparameterized case, into two sub-cases:

- $n \ll d$
- $n \approx d$

Overparameterized case in linear regression

- For $n \ll d$, there are many solutions to beta: $X\hat{\beta} = y$. So gradient descent can find the minimum-norm solution.

- For $n \approx d$ (interpolation threshold), there is only one solution. However, since data is noisy, β must fit to the noise:

$$\hat{\beta} = X^{\dagger}y = X^{\dagger}(X\beta + \eta) = \underbrace{X^{\dagger}X\beta}_{\text{signal}} + \underbrace{X^{\dagger}\eta}_{\text{noise}}$$

X becomes singular near $n \approx d$, causing noise (and thus, error) to have very high norm.

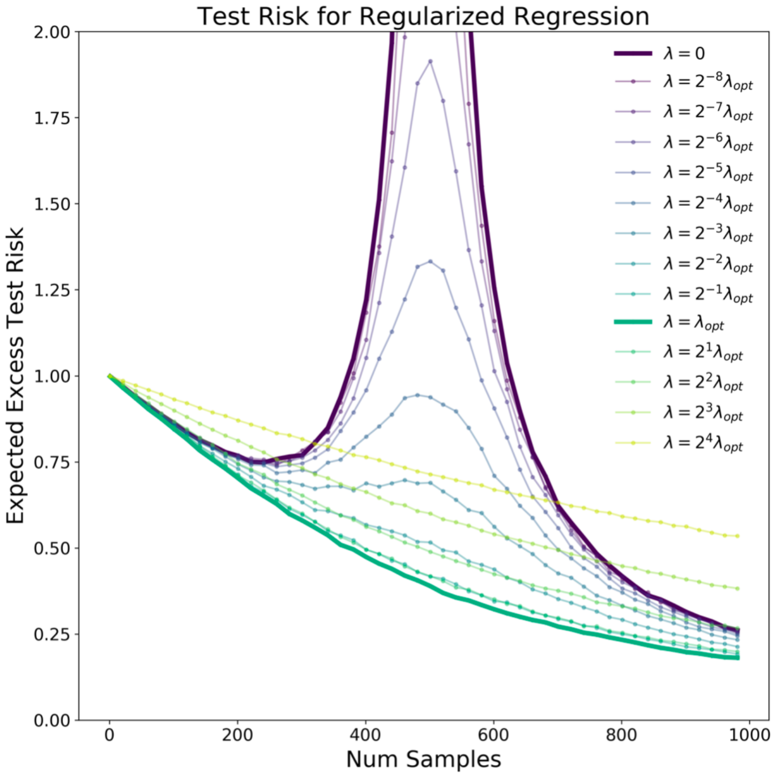
Read [1] for more details!

Optimal regularization mitigates double descent

Adding a penalty to the loss function forces a solution that does not blow up the test error (ridge regression) [2]:

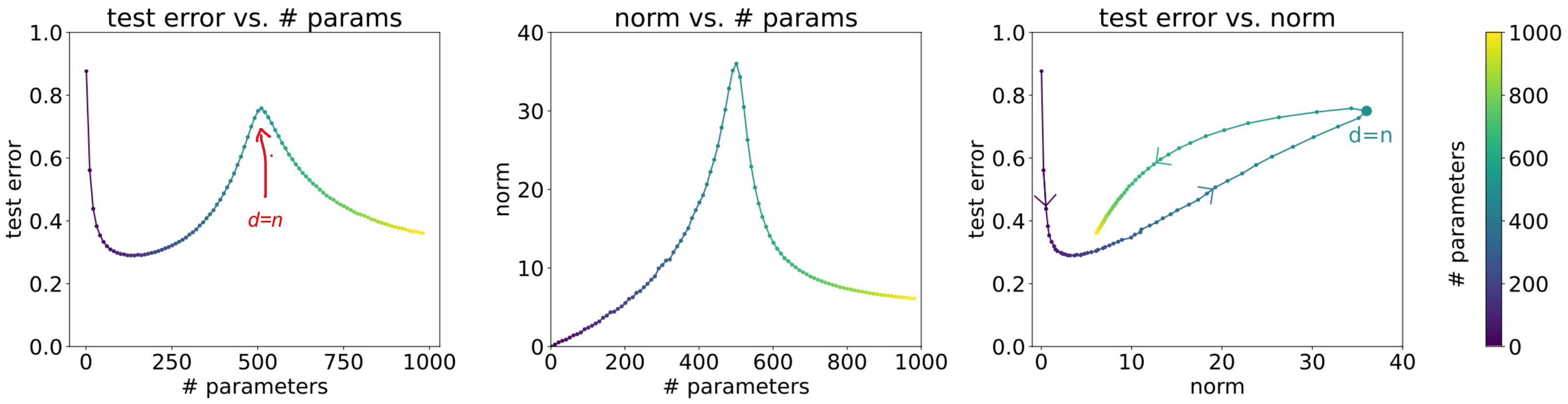
$$\hat{\beta}_\lambda := \text{argmin}_\beta \, ||X\beta - \vec{y}||_2^2 + \lambda||\beta||_2^2$$

λ is tuned for each sample size *n*. Under certain assumptions, λ_{opt} is independent of *n*!



Double Descent is *not* a universal phenomenon

- What if we measure model complexity by the “norm” of the parameters?



Model shown here is linear regression on n = 500 examples, from the Fashion-MNIST dataset [2].

This is still an open research area!

- What happens if we remove isotropic restriction from data?
- Properties for non-linear regression case – can we mathematically show what's happening for more complex models?

References

1. Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent. 2019. https://arxiv.org/pdf/1912.07242.pdf
2. Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. 2020. https://arxiv.org/pdf/2003.01897.pdf
3. Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. Communications on Pure and Applied Mathematics, 75(4):667–766, 2022.