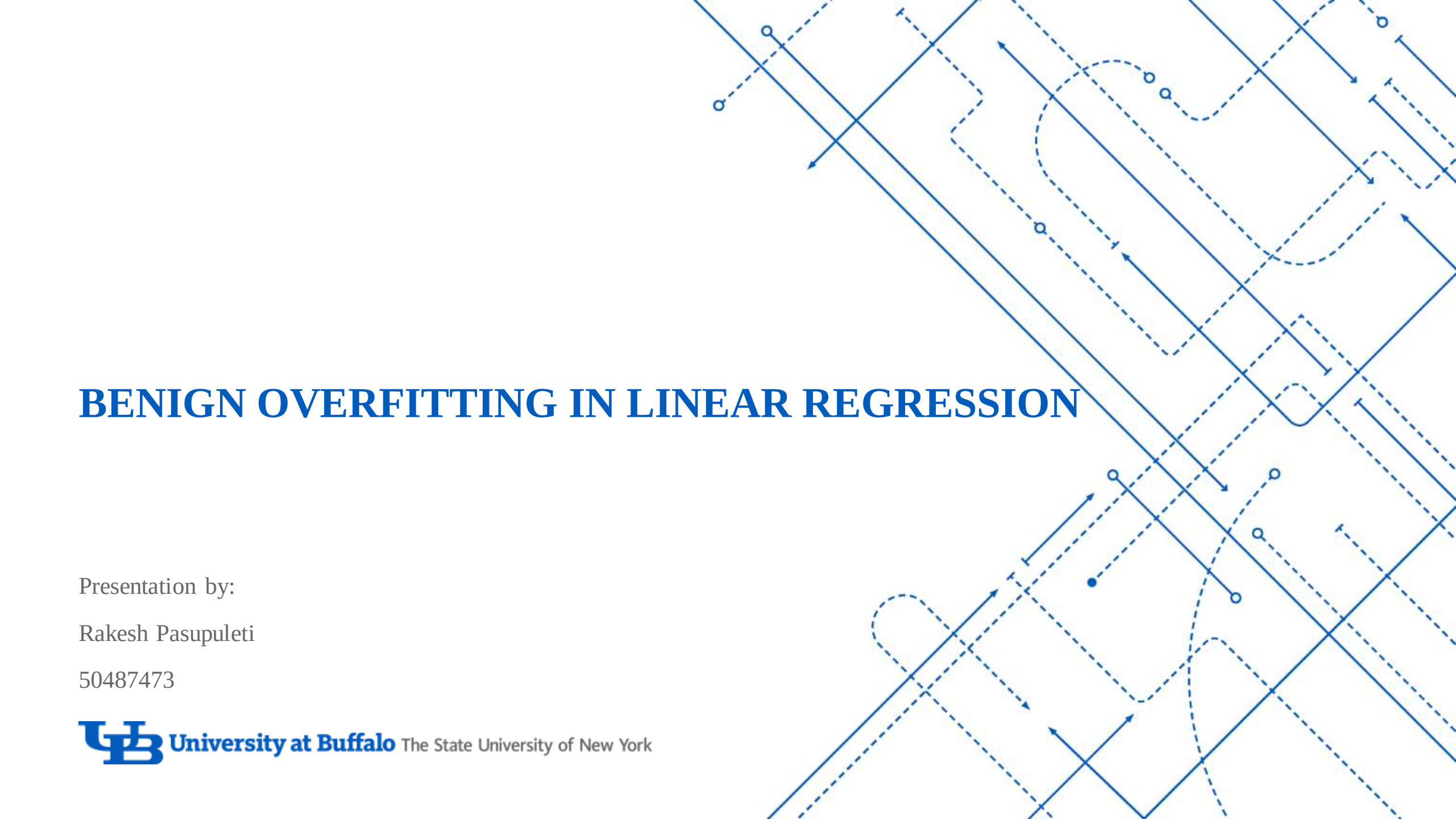


BENIGN OVERFITTING IN LINEAR REGRESSION

Presentation by:

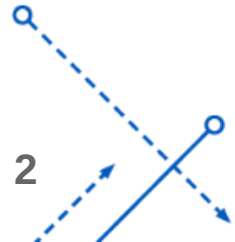
Rakesh Pasupuleti

50487473

University at Buffalo The State University of New York

OUTLINE:

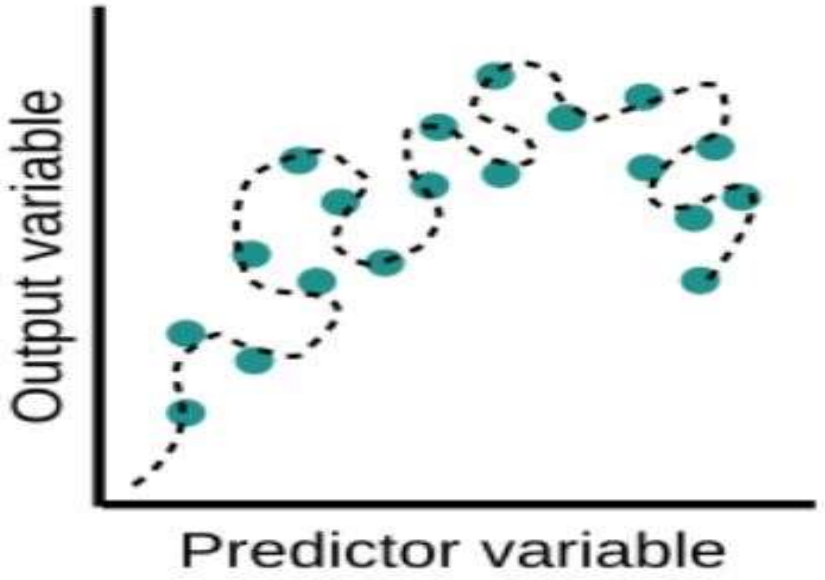
- Overfitting
- Benign Overfitting
- Benign Overfitting In Linear Regression
- Characterizing Benign Overfitting
- Limitations
- Conclusion



OVERFITTING

Overfitting is a common problem in machine learning that occurs when a model is trained too well on the training data and as a result, it becomes too specialized to the training data and loses its ability to generalize to new, unseen data.

- Poor Generalization
- Wasted Resources



Statistical Wisdom and Overfitting:

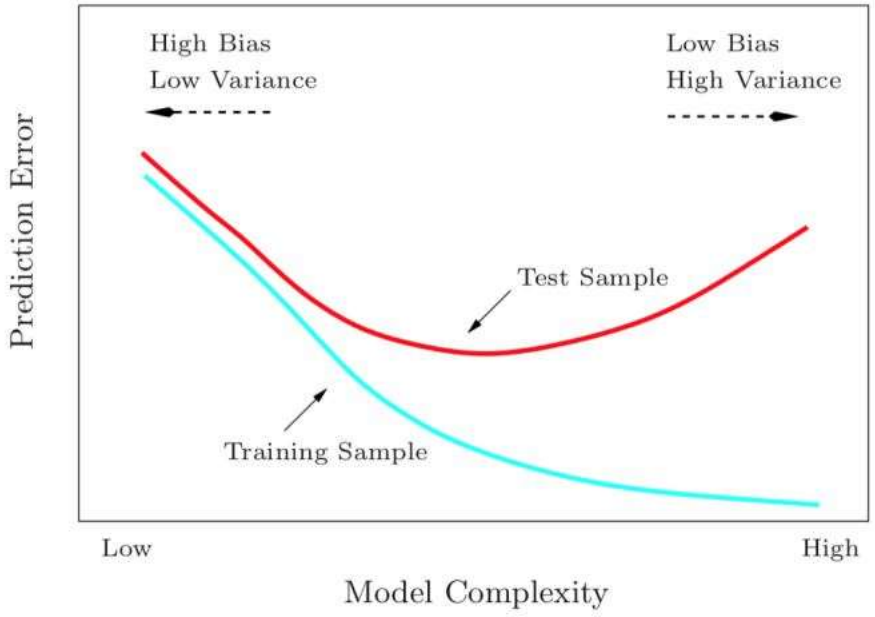
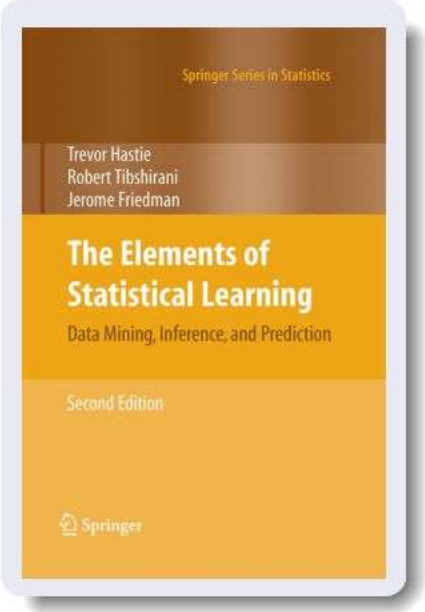


FIGURE 2.11. *Test and training error as a function of model complexity.*

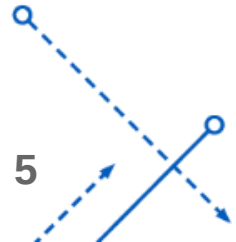
Figure 2.11 shows the typical behavior of the test and training error, as model complexity is varied. The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder. However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error). In

"... interpolating fits... [are] unlikely to predict future data well at all."



A new statistical phenomenon:
good prediction with zero training error for regression loss

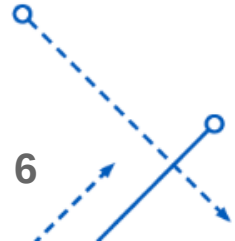
- Statistical wisdom says a prediction rule should not fit too well.
- But deep networks are trained to fit noisy data perfectly, and they predict well.
- Deep networks can be trained to zero training error (for regression loss) with near state-of-the-art performance even for noisy problems.
- Benign overfitting.



BENIGN OVERFITTING

overfitting may not always be harmful. In some cases, models that overfit the training data can still generalize well to new data, a phenomenon known as benign overfitting. This paper investigates the conditions under which benign overfitting can occur in linear regression models.

✓ Better generalization performance



BENIGN OVERFITTING IN LINEAR REGRESSION:

- Covariate $x \in H$ (Hilbert space); response $y \in R$.
- (x, y) Gaussian, mean zero.
- Define:

$$\Sigma := \mathbb{E} x x^\top = \sum_i \lambda_i v_i v_i^\top, \qquad (\text{assume } \lambda_1 \geq \lambda_2 \geq \cdots)$$

$$\theta^* := \arg\min_\theta \mathbb{E}\left(y - x^\top \theta\right)^2,$$

$$\sigma^2 := \mathbb{E}(y - x^\top \theta^*)^2.$$

MINIMUM NORM ESTIMATOR :

- Given data $X \in \mathrm{H}^n$ and $y \in \mathrm{R}^n$, the minimum norm estimator $\hat{\theta}$ solves the optimization problem

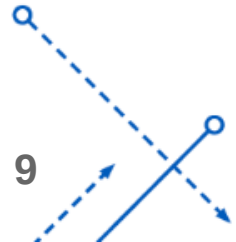
$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ such\ that \quad & \|X\theta - \boldsymbol{y}\|^2 = \min_{\beta} \|X\beta - \boldsymbol{y}\|^2 \end{aligned}$$

By the projection theorem, parameter vectors that solve the least squares problem solve the normal equations, so we can equivalently write $\hat{\theta}$ as the minimum norm solution to the normal equations

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta} \left\{ \|\theta\|^2 : X^\top X\theta = X^\top \boldsymbol{y} \right\} \\ &= \left(X^\top X\right)^\dagger X^\top \boldsymbol{y} \\ &= X^\top \left(XX^\top\right)^\dagger \boldsymbol{y}, \end{aligned}$$

Overfitting Regime:

- We consider situations where $\min_{\beta}\|X\beta - Y\|^2 = 0$.
- Hence, $Y_1 = X_1^T\hat{\theta}, \ldots\ldots, Y_n = X_n^T\hat{\theta}$.
- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?



MAIN RESULT:

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$,

1. With high probability,

$$R(\hat{\theta}) \leq c\left(\|\theta^*\|^2\|\Sigma\|\sqrt{\frac{r_0(\Sigma)}{n}} + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right),$$

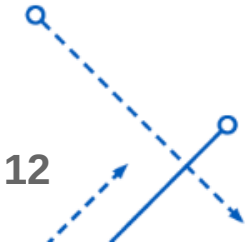
2. $\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\min\left\{\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1\right\}.$

Notions of Effective Rank:

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}$$

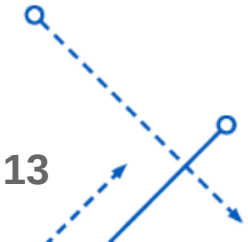
$$R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}$$

- The mix of eigenvalues of Σ determines:
- How the label noise is distributed in $\hat{\theta}$, and
- How errors in $\hat{\theta}$,affect prediction accuracy.
- To avoid harming prediction accuracy, the noise energy must be distributed across many unimportant directions.
- Overparameterization is essential for benign overfitting
  - Number of non-zero eigenvalues: large compared to n,
  - Their sum: small compared to n
  - Number of ‘small’ eigenvalues: large compared to n,
  - Small eigenvalues: roughly equal



LIMITATIONS:

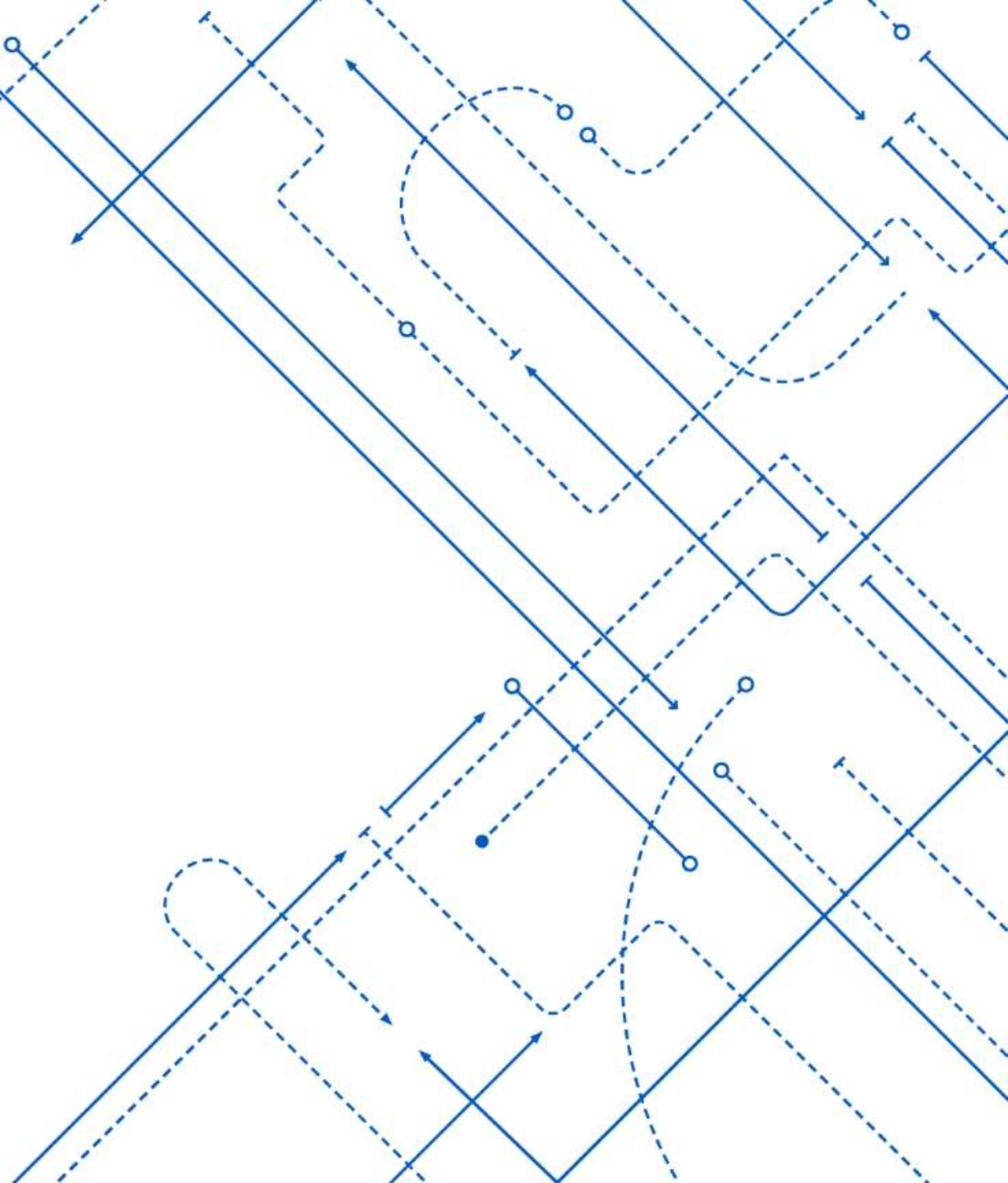
- First, the authors only investigate linear regression models and it is unclear whether their results generalize to other types of models.
- It leads to huge sensitivity to (adversarial) perturbations.
- The authors do not investigate the computational cost of building models that overfit the training data.



Conclusion:

- In linear regression, a long, flat tail of the covariance eigenvalues is necessary and sufficient for the minimum norm interpolant to predict well:
  - The noise is hidden in many unimportant directions

THANK YOU



University at Buffalo The State University of New York