

Data Mining

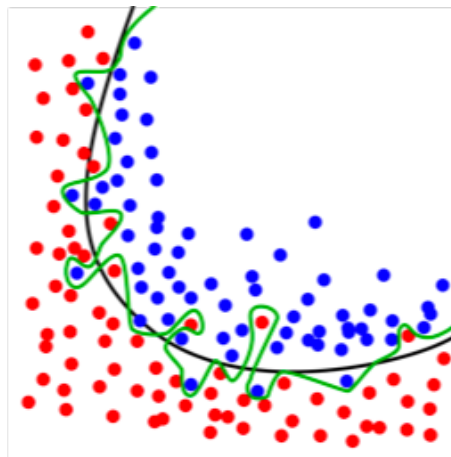
CS57300
Purdue University

Bruno Ribeiro

January 30, 2018

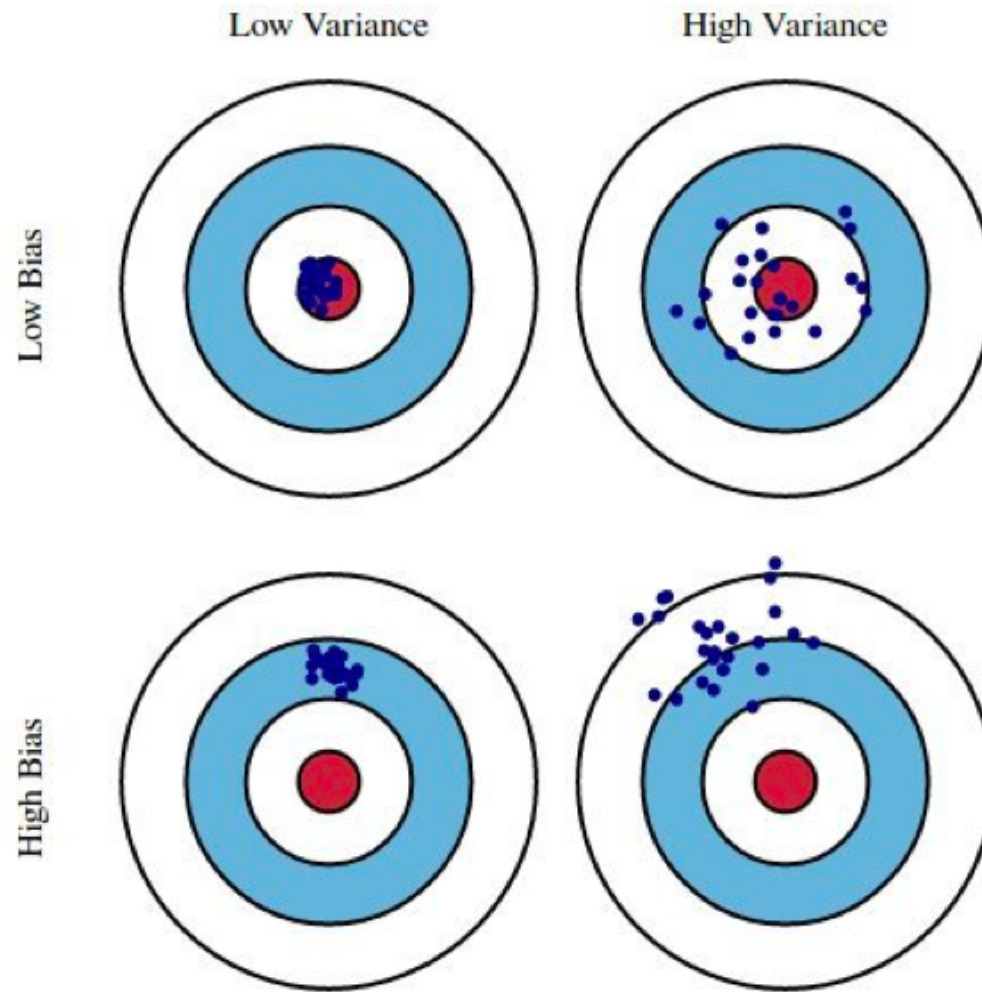
Overfitting

- **Definition:** Overfitting is when the model captures the noise in the training data rather than the underlying structure in the data
- Overfitting can occur even in models such as Logistic regression, linear regression, and SVMs
 - In these simpler linear boundary models, overfitting often happens if feature space is very high dimensional
 - E.g., 2 years of web browsing activity of 10 users
- Can also happen with fewer features and very complex models:



Wikipedia

A Cartoon View of Bias and Variance



Bias-Variance Tradeoff

- To understand overfitting we must understand the space of models
- Assume the linear model

$$\underbrace{y}_{\text{scalar}} = \underbrace{\mathbf{w}^T}_{1 \times p} \underbrace{\mathbf{x}}_{p \times 1} + \underbrace{b}_{\text{scalar}} + \underbrace{\epsilon}_{\text{scalar}}$$

where w and b are parameters and ϵ is some random noise

- In regression analysis, our major goal is to come up with some good regression function

$$\hat{y}(\mathbf{x}) = \hat{\mathbf{w}}^T \mathbf{x} + \hat{b}$$

where the variables with \wedge are estimated values

- To see if the estimated parameters of w and b are good candidates, we can ask two questions:
 - Are $\hat{\mathbf{w}}$ and \hat{b} close to the true values of \mathbf{w} and b ?
 - Will $\hat{y}(\mathbf{x})$ accurately predict unseen data?

Concrete Example

- Consider the following scenario

$$y = \mathbf{w}^T \mathbf{x} + b + \epsilon$$

with:

- $\mathbf{x} \in \{0, 1\}^4$
- $\epsilon \sim \text{Normal}(0, 1)$
- $\mathbf{w} = \{1, 1, 1, 1\}$
- $b = 0$

- Consider the dataset:  = Wrong parameters that “perfectly” fit a datapoint

1. $\mathbf{x}_1 = (0, 0, 1, 1), y = 2.1$



$$\hat{\mathbf{w}} \in \{\{10, -10, 2, 0.1\}, \{10.9, -9, -10, 12.1\}, \dots\}$$

2. $\mathbf{x}_2 = (1, 1, 0, 0), y = 1.9$

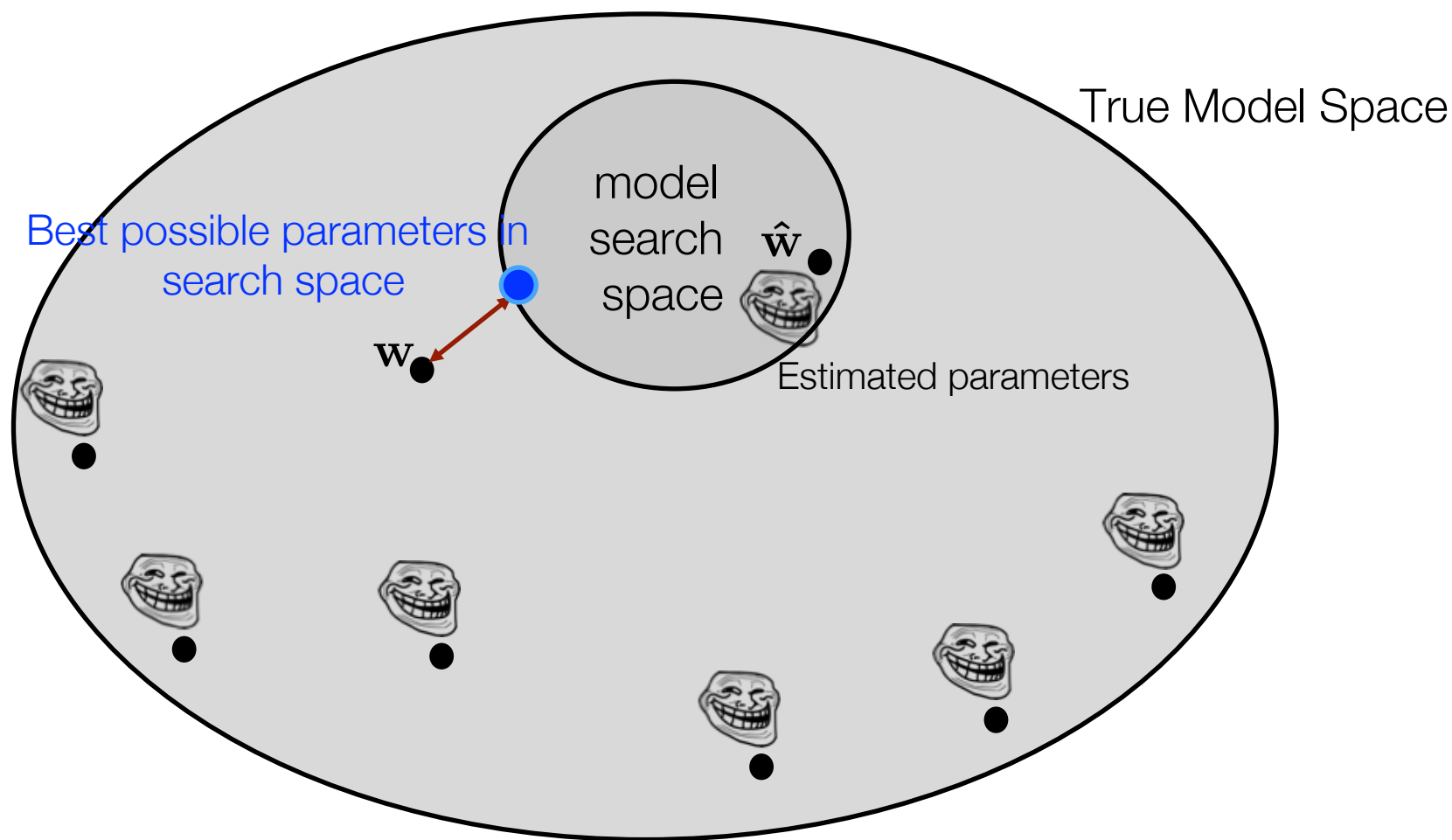
Note that as more datapoints arrive, the intersection
between the troll sets reduce




$$\hat{\mathbf{w}} \in \{\{2.9, -1, 10, -10\}, \{10.9, -9, -10, 12.1\}, \dots\}$$

Model Search Space

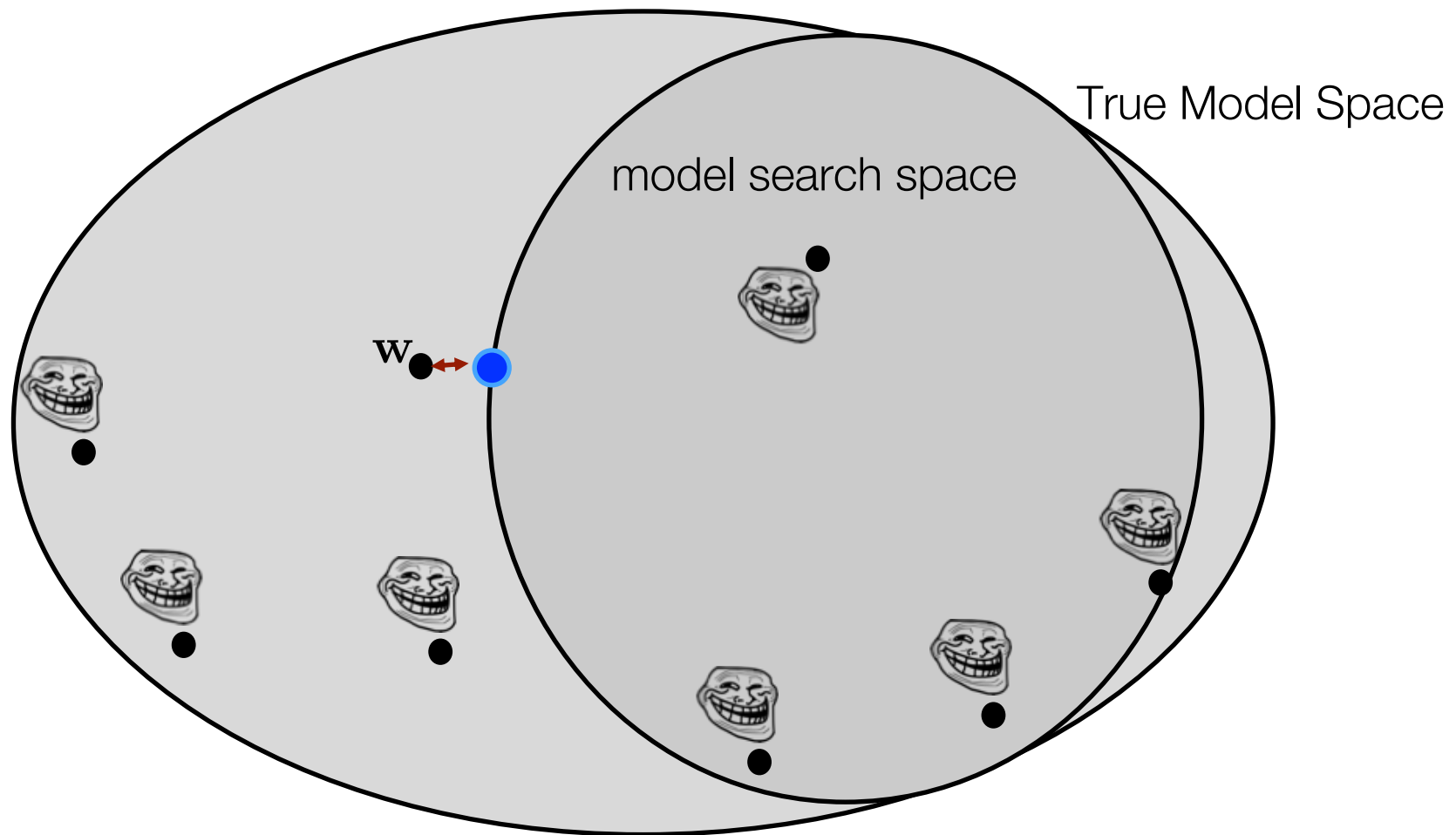
Are $\hat{\mathbf{w}}$ and \hat{b} close to the true values of \mathbf{w} and b ?



- Bias related to arrow  = parameter values look better than \mathbf{w} in the training data
- Variance related to number of "trolls"

Enlarging Model Search Space

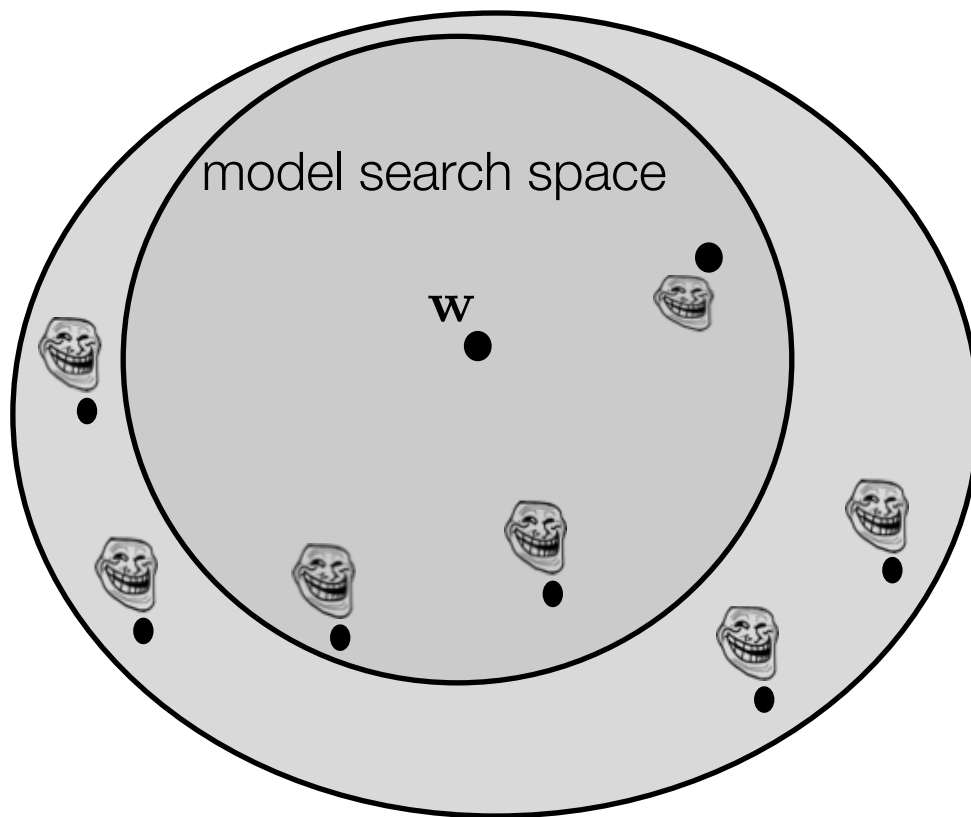
Are $\hat{\mathbf{w}}$ and \hat{b} close to the true values of \mathbf{w} and b ?



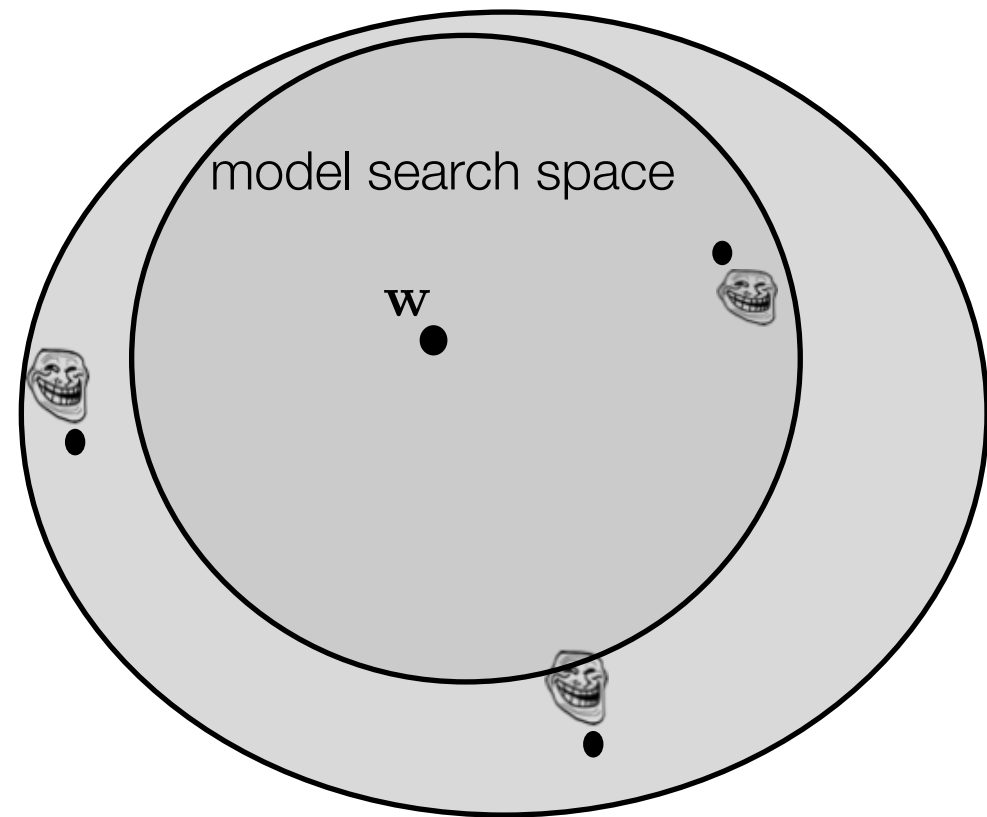
- Enlarging model search space allow us to be closer to the "true" model (reduces bias) but will also increase the changes we will make more outrageous mistakes (variance)

With More Training Data

- Fewer data points



- More data points



Probabilistic interpretation:

Each of these points in the model search is a hypothesis... model search seeks a likely hypothesis given the data

Bias-variance decomposition (revisited)

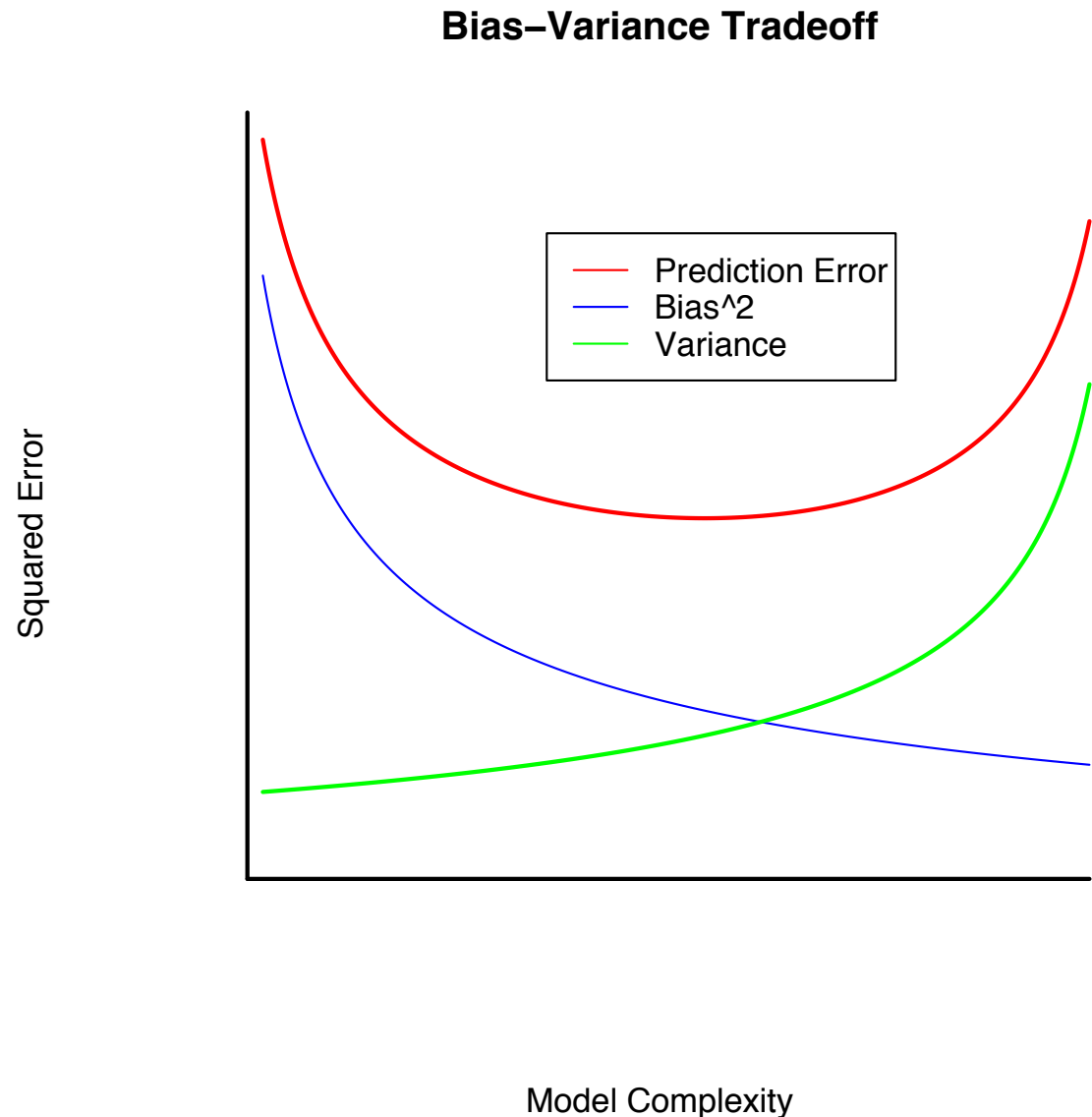
- The mean-squared error (MSE) of $\hat{\mathbf{w}}$ is:

$$\begin{aligned} E[(\hat{\mathbf{w}} - \mathbf{w})^2] &= E[(\hat{\mathbf{w}} - E[\hat{\mathbf{w}}] + E[\hat{\mathbf{w}}] - \mathbf{w})^2] \\ &= \underbrace{(E[\hat{\mathbf{w}}] - \mathbf{w})^2}_{\text{bias}} + \underbrace{E[(\hat{\mathbf{w}} - E[\hat{\mathbf{w}}])^2]}_{\text{variance}} \end{aligned}$$

- MSE measures systematic bias and random variance between estimate and population value

Illustration of Impact of Bias-Variance Trade-off

- More complex models have larger model search spaces
- Impact over prediction error (red) of bias (blue) and variance (green) with distinct model complexities



Understanding Bias-Variance in the Context of Multiple Hypotheses

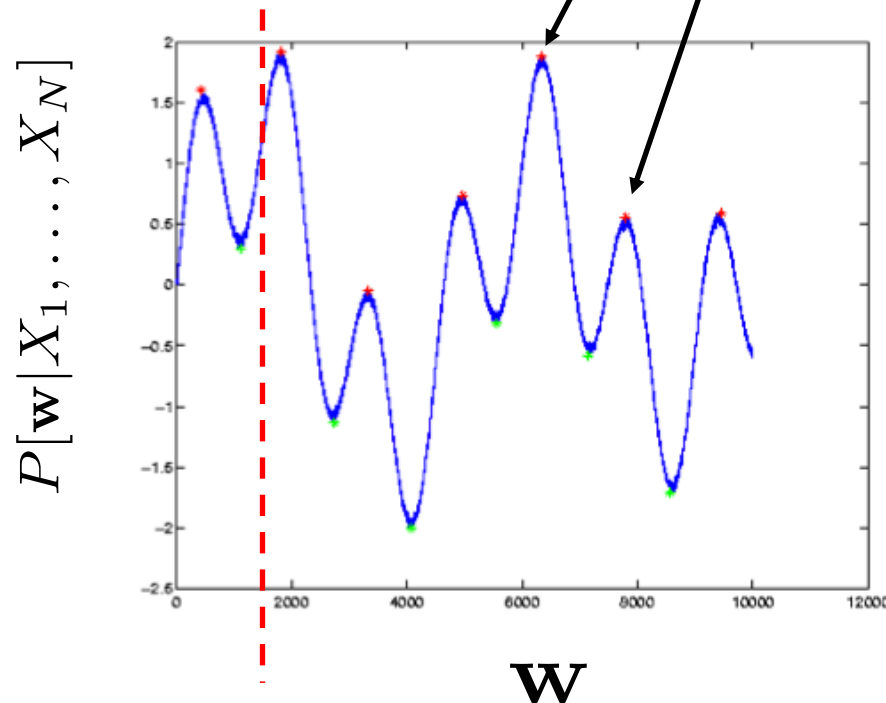
- Model search is a procedure to find a likely hypothesis for the training data
 1. \mathcal{H}_1 : model with parameters w_1 is the correct model
 2. \mathcal{H}_2 : model with parameters w_2 is the correct model
 3. ...
- Less data means more good-looking (likely) hypotheses
 - One extreme: no data means any hypothesis is likely



- The other extreme: infinite data means only correct hypothesis is likely
 - In probabilistic models, Maximum Likelihood Estimator (MLE) converges to true parameter value (Fisher) [if model not misspecified]
 - If model misspecified, MLE it will converge to the model that best explains the entire population data

Effect of Finite Data in Hypothesis Test

- Assume an uninformative prior (very weak prior): $P[\mathbf{w}]$
 - Data: X_1, \dots, X_N
 - Consider posterior $P[\mathbf{w}|X_1, \dots, X_N]$
 - Say, \mathbf{w} is 1-Dimensional:
- Global maximum
($w=6020$ is a very likely hypothesis given data)
- Local maximum
($w=7300$ is a likely hypothesis given data)



True \mathbf{w}

Findings

- Bias
 - Often related to size of model space
 - More complex models tend to have lower bias
- Variance
 - Often related to size of dataset
(number of trolls in your model search space)
 - When data is large enough to estimate parameters well then models have lower variance
- Simple models can perform surprisingly well due to lower variance

“Reducing” Model (Search) Space without Changing the Model

Reducing the Model Search Space (without changing the model)

$$\underbrace{y}_{\text{scalar}} = \underbrace{\mathbf{w}^T}_{1 \times p} \underbrace{\mathbf{x}}_{p \times 1} + \underbrace{b}_{\text{scalar}} + \underbrace{\epsilon}_{\text{scalar}}$$

- What if we limit \mathbf{w} to the surface of a hypersphere or radius $r = 1$?



- This will limit the model search space
 - A bit too harsh... may give very bad estimates...
 - What if we look for the smallest radius that still make good predictions?
- Original regression problem: data $\{(\mathbf{x}_i, y_i)\}_{i=1, \dots, N}$, where $\mathbf{x}_i = (\text{<data>, 1})$ and $\beta_0 = b$

$$\beta^{\text{reg}} = \arg \min_{\beta} \sum_{i=1}^N (y_i - \beta \mathbf{x}_i)^2$$

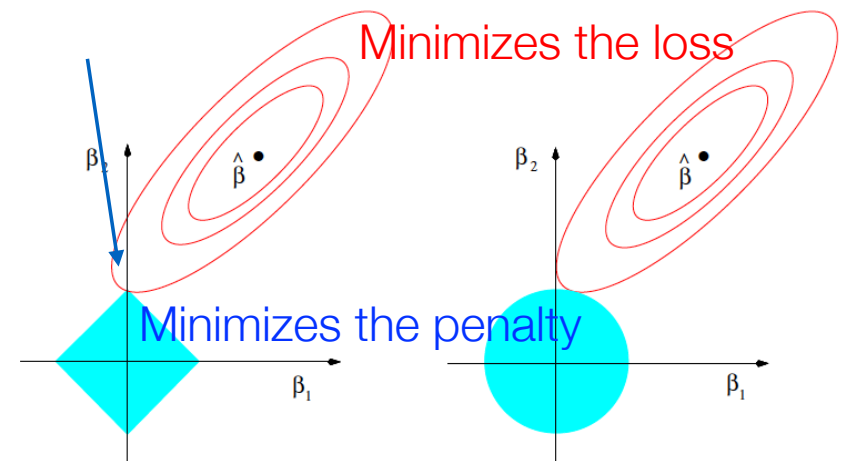
Ridge Regression & Lasso

- Ridge regression (parameters $\beta = (\mathbf{w}, b)$ with $x_i = (\langle \text{data} \rangle, 1)$)

$$\begin{aligned}\hat{\beta}^{\text{ridge}} &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \\ &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss (score)}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}\end{aligned}$$

- Lasso regression
(Lasso = Least Absolute Selection and Shrinkage Operator)

$$\begin{aligned}\hat{\beta}^{\text{lasso}} &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^p |\beta_j| \\ &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss (score)}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}\end{aligned}$$



L1 norm more strongly force parameters to zero

Gaussian Priors

- Consider again the model $y = \beta^T \mathbf{x} + \epsilon$
- Assume $\beta \sim \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{I})$
- Assume $\epsilon \sim \text{Normal}(0, 1)$
- Maximum a posteriori estimate (MAP)

$$\hat{\beta} = \arg \max_{\beta} \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \beta^T \mathbf{x}_i)^2\right) \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\beta^T \beta}{2\sigma^2}\right)$$

$$= \arg \max_{\beta} \log \prod_{i=1}^N \exp\left(-\frac{1}{2}(y_i - \beta^T \mathbf{x}_i)^2\right) \cdot \exp\left(-\frac{\beta^T \beta}{2\sigma^2}\right)$$

$$= \arg \max_{\beta} \sum_{i=1}^N -(y_i - \beta^T \mathbf{x}_i)^2 - \frac{1}{2\sigma^2} \sum_{j=0}^p \beta_j^2$$

$$= \arg \min_{\beta} \sum_{i=1}^N (y_i - \beta^T \mathbf{x}_i)^2 + \frac{1}{2\sigma^2} \sum_{j=0}^p \beta_j^2$$

Compare to ridge-regression equation... only difference is regularizing b

Variance of Prior over Parameters

- In HW0 we have seen that very high-dimensional Gaussians lie in the surface of a hypersphere



- What happens when we increase variance over prior parameters?
 - We are saying that, without data, we believe β lives on the surface of a very large hypersphere, a much larger model space
 - As more data is gathered, weaker prior gives in to the data but strong prior forces spherical parameters



Probability mass
of prior near any
small area
is still large

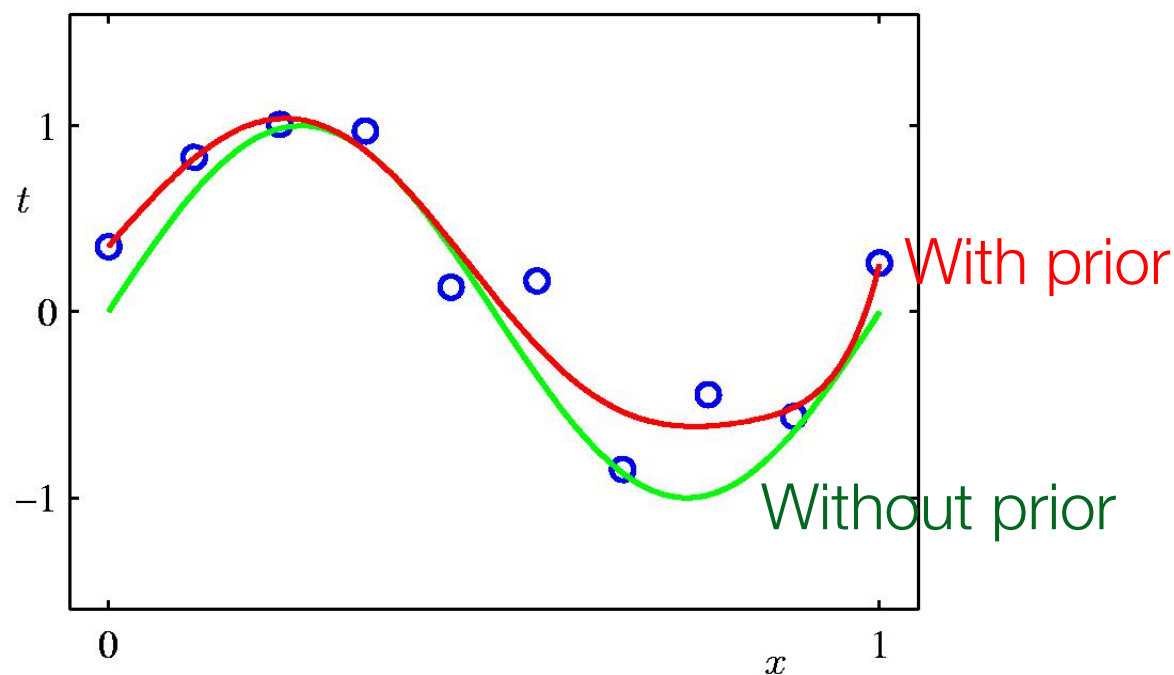


Probability mass
of prior near any
small area
is very small

Weak vs Strong Priors

- Simple 1D example: weak Gaussian prior over 3rd degree polynomial curve parameters

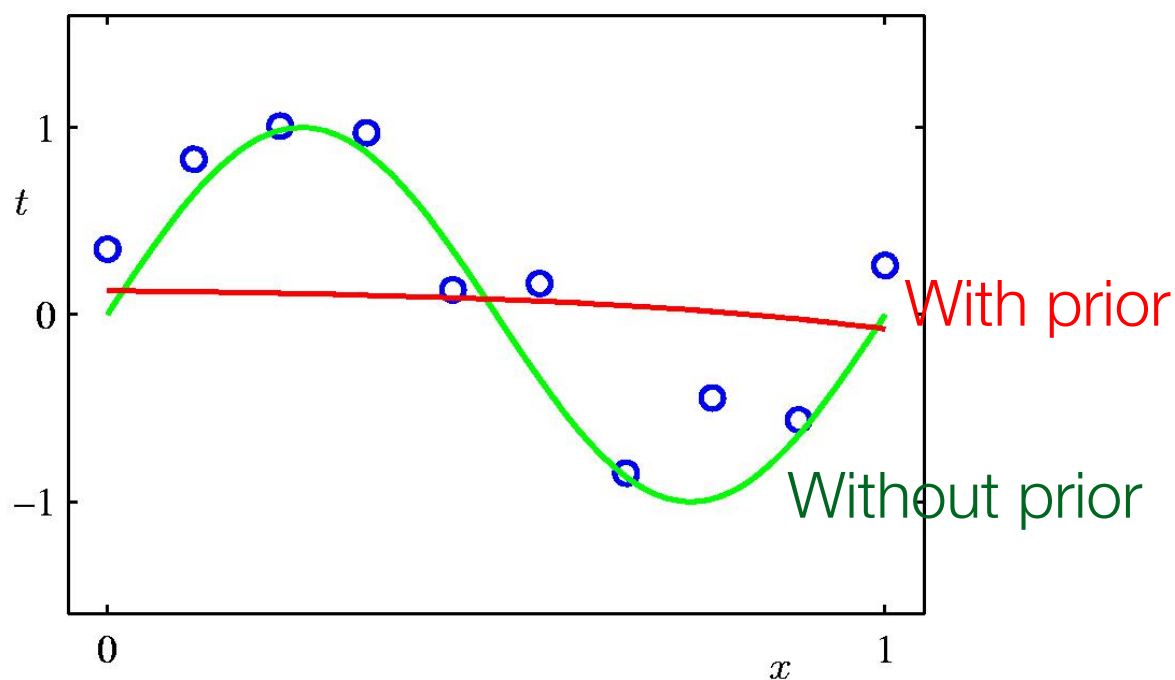
$$\lambda = \frac{1}{\sigma^2} = 10^{-8}$$



Weak vs Strong Priors

- Simple 1D example: weak Gaussian prior over 3rd degree polynomial curve parameters

$$\lambda = \frac{1}{\sigma^2} = 1$$



Practical Advice

- Don't Regularize the bias (Intercept) parameter b
Regularizers always avoid penalizing this bias / intercept parameter
 - Why? Because otherwise the learning algorithms wouldn't be invariant to a shift in the y-values
- Whitening Data
 - It's common to whiten each feature by subtracting its mean and dividing by its variance
 - For regularization, this helps all the features be penalized in the same units, that is, we are assuming they have the same variance σ^2