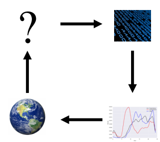
The Bias Variance Tradeoff and Regularization

Slides by:

Joseph E. Gonzalez

jegonzal@cs.berkeley.edu



Recap: *Least Squares Regression*

- Least squares regression:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \overbrace{(y_i - \underbrace{f_\theta(x)}_{\text{Parametric Model}})^2}^{\text{Loss Function}} \qquad \mathbf{L}(y, \hat{y}) = (y - \hat{y})^2$$

  - Generic framework
- Linear models:

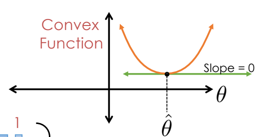
$$f_\theta(x) = \sum_{j=1}^{p} \theta_j x_j$$

  - Linear in the parameters

- Linear models:

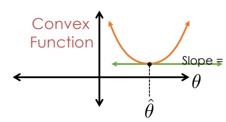
$$f_\theta(x) = \sum_{j=1}^{p} \theta_j x_j$$

  - Linear in the parameters
- Finding Optimal Parameters
  - Normal Equations:



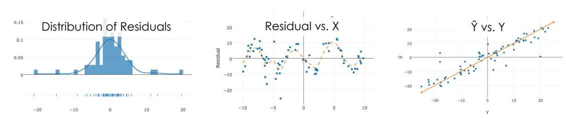
$$\hat{\theta} = \left(X^T X\right)^{-1}\left(X^T Y\right)$$

- Finding Optimal Parameters
  - Normal Equations:



$$\hat{\theta} = \left(X^T X\right)^{-1}\left(X^T Y\right)$$

  - Using software packages (e.g., linalg.solve, sklearn.linear_models)
- Diagnostics:

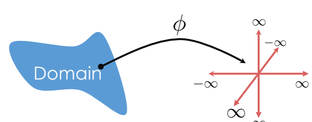


Recap: *Feature Engineering*

- Linear models with feature functions:

$$f_\theta(x) = \sum_{j=1}^{p} \theta_j \phi_j(x)$$

- Feature Functions: $\phi : \mathcal{X} \rightarrow \mathbb{R}^p$



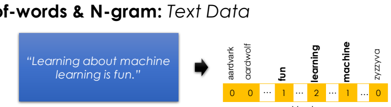
- **One-hot encoding:** *Categorical Data*

state
NY
WA
CA

➡

AL	...	CA	...	NY	...	WA	...	WY
0	...	0	...	1	...	0	...	0
0	...	0	...	0	...	1	...	0
0	...	1	...	0	...	0	...	0

- **Bag-of-words & N-gram:** *Text Data*



- **Custom Features:** *Domain Knowledge*

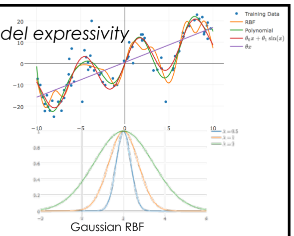
$$\phi(\text{lat}, \text{lon}, \text{amount}) = \frac{\text{amount}}{\textbf{Stores}[\textbf{ZipCode}[\text{lat}, \text{lon}]]}$$

- **Generic Features:** *increase model expressivity*
  - **Polynomial and trigonometric:**

$$\phi(x) = \left[x, x^2, \ldots, x^p\right]$$

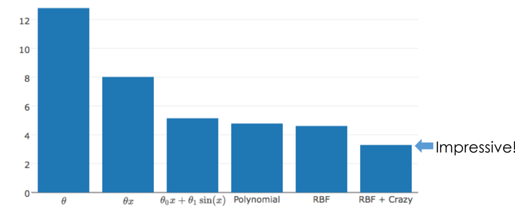
  - **Gaussian Radial Basis Functions:**

$$\phi_{\lambda_i,\mu_i}(x) = \exp\left(-\frac{||x-\mu_i||_2^2}{\lambda_i}\right)$$

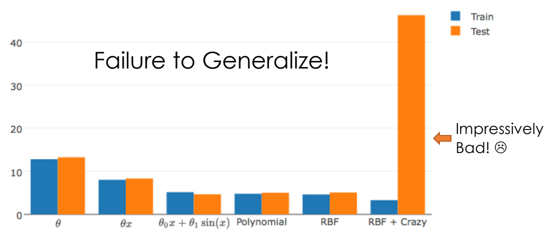


- **Hyper-parameters:** parameters that are not "learned" but instead selected using prior knowledge or through *cross-validation*.
  - Example: Poly. deg., μ_i, and σ_i

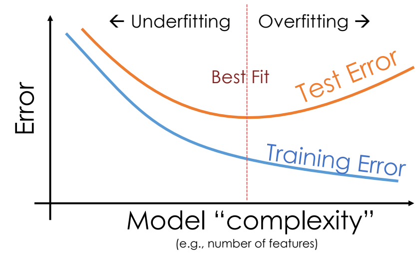
Training Error



Training vs **Test** Error

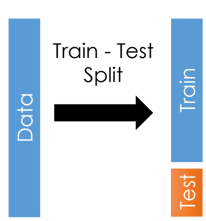


Training vs **Test** Error



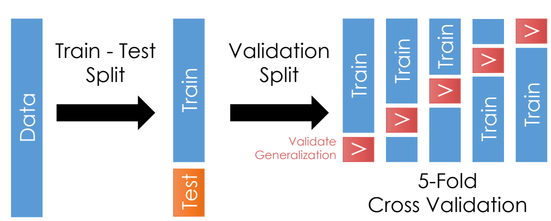
Generalization: *The Train-Test Split*

- **Training Data:** used to fit model
- **Test Data:** check generalization error
- How to split?
  - Randomly, Temporally, Geo…
  - Depends on application (usually randomly)
- What size? (90%-10%)
  - Larger training set → more complex models
  - Larger test set → better estimate of generalization error
  - Typically between 75%-25% and 90%-10%



You can only use the test dataset once after deciding on the model.

Generalization: *Validation Split*



Cross validation simulates multiple train test-splits on the training data.

Python Demo!

Regularization Notebook Part 1

Fundamental Challenges of Prediction

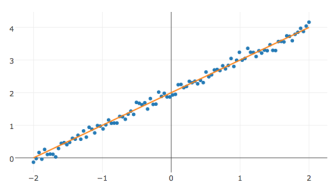
- **Noise:** the intrinsic variability in the process we are trying to model
- **Bias:** the expected deviation between the predicted value and the true value
- **Variance:** variability between the estimated value and the true value across different training datasets

Noise

The intrinsic variability in the process we are trying to predict

- measurement variability
- stochasticity
- missing information

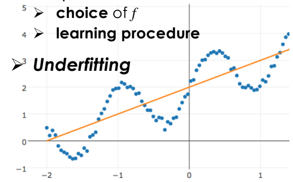
Beyond our control (usually)

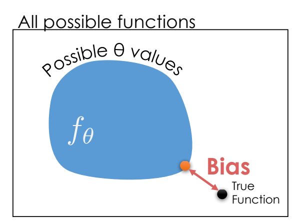


Bias

The expected deviation between the predicted value and the true value

- Depends on both the:
  - **choice** of f
  - **learning procedure**
- ***Underfitting***

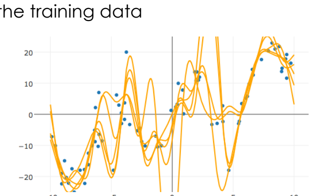




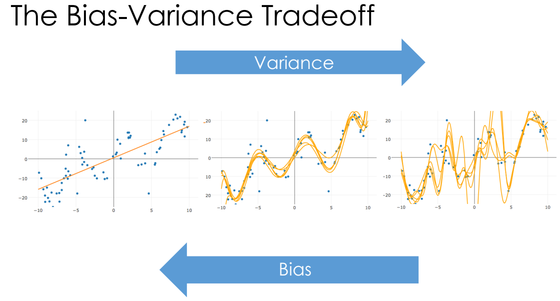
Variance

variability between the estimated value and the true value across different training datasets

- Sensitivity to variation in the training data
- Poor generalization
- **Overfitting**



The Bias-Variance Tradeoff



Analysis of Squared Error

- For the test point $\boldsymbol{x}$ the expected error:
  - *Random variables are **red***

Assume noisy observations → y is a random variable

True Function

Noise term: $\mathbf{E}[\epsilon] = 0$

$$y = h(x) + \epsilon$$

$$\mathbf{E}\left[\left(y - f_\theta(x)\right)^2\right]$$

Assume ***training data*** is random → θ is a random variable

Analysis of Squared Error

Goal:

$$\mathbf{E}\left[\left(y - f_\theta(x)\right)^2\right] =$$

Noise + (**Bias**)^2 + **Variance**

$$\mathbf{E}\left[\left(y - f_\theta(x)\right)^2\right] = \mathbf{E}\left[\left(y \underbrace{- h(x) + h(x)}_{\text{Subtracting and adding } h(x)} - f_\theta(x)\right)^2\right]$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$\mathbf{E}\left[\left(y - f_\theta(x)\right)^2\right] = \mathbf{E}\left[\left(\underbrace{y - h(x)}_{a} + \underbrace{h(x) - f_\theta(x)}_{b}\right)^2\right]$$

Expanding in terms of *a* and *b*: $(a + b)^2 = a^2 + b^2 + 2ab$

$$= \mathbf{E}\left[\left(y - h(x)\right)^2\right] + \mathbf{E}\left[\left(h(x) - f_\theta(x)\right)^2\right]$$

a^2 b^2

$$+2\mathbf{E}\left[\left(y - h(x)\right)\left(h(x) - f_\theta(x)\right)\right]$$

$2ab$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

Expanding Terms

$$+2\mathbf{E}\left[yh(x) - yf_\theta(x) - h(x)h(x) + h(x)f_\theta(x)\right]$$

$$\mathbf{E}\left[\left(y - f_\theta(x)\right)^2\right] = \mathbf{E}\left[\left(y - h(x) + h(x) - f_\theta(x)\right)^2\right]$$

$$= \mathbf{E}\left[\left(y - h(x)\right)^2\right] + \mathbf{E}\left[\left(h(x) - f_\theta(x)\right)^2\right]$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$+2\mathbf{E}\left[\underbrace{yh(x) - yf_\theta(x)}_{y = h(x) + \epsilon} - h(x)h(x) + h(x)f_\theta(x)\right]$$

$$\left(h(x) + \epsilon\right)h(x) - \left(h(x) + \epsilon\right)f_\theta(x)$$

$$= h(x)h(x) + \epsilon h(x) - h(x)f_\theta(x) - \epsilon f_\theta(x)$$

$$\mathbf{E}\left[\left(y - f_\theta(x)\right)^2\right] = \mathbf{E}\left[\left(y - h(x) + h(x) - f_\theta(x)\right)^2\right]$$

$$= \mathbf{E}\left[\left(y - h(x)\right)^2\right] + \mathbf{E}\left[\left(h(x) - f_\theta(x)\right)^2\right]$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$+2\mathbf{E}\left[\underbrace{yh(x) - yf_\theta(x)}_{y = h(x) + \epsilon} - \cancel{h(x)h(x)} + \cancel{h(x)f_\theta(x)}\right]$$

$$= \cancel{h(x)h(x)} + \boxed{\epsilon h(x)} - \cancel{h(x)f_\theta(x)} - \boxed{\epsilon f_\theta(x)}$$

$$+2\mathbf{E}\left[\epsilon h(x) - \epsilon f_\theta(x)\right]$$

$$\mathbf{E}\left[(y - f_\theta(x))^2\right] = \mathbf{E}\left[(y - h(x) + h(x) - f_\theta(x))^2\right]$$

$$= \mathbf{E}\left[(y - h(x))^2\right] + \mathbf{E}\left[(h(x) - f_\theta(x))^2\right]$$

$$\underbrace{+2\mathbf{E}\left[\epsilon h(x) - \epsilon f_\theta(x)\right]}$$

$$\mathbf{E}\left[\epsilon h(x) - \epsilon f_\theta(x)\right] = \mathbf{E}\left[\epsilon\,(h(x) - f_\theta(x))\right]$$

Independence $= \mathbf{E}[\epsilon]\,\mathbf{E}\left[(h(x) - f_\theta(x))\right]$

Noise definition (Useful eqns.) $= 0$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$\mathbf{E}\left[(y - f_\theta(x))^2\right] =$$

$$\mathbf{E}\left[(\underbrace{y}_{\text{Obs. Value}} - \underbrace{h(x)}_{\text{True Value}})^2\right] +$$

Noise Term

$$\mathbf{E}\left[(\underbrace{h(x)}_{\text{True Value}} - \underbrace{f_\theta(x)}_{\text{Pred. Value}})^2\right]$$

Model Estimation Error

$$\mathbf{E}\left[(h(x) - f_\theta(x))^2\right] =$$ Next we will show….

$$\mathbf{E}\left[(h(x) - \mathbf{E}_\theta\left[f_\theta(x)\right])^2\right] + \mathbf{E}\left[(\mathbf{E}_\theta\left[f_\theta(x)\right] - f_\theta(x))^2\right]$$

- How?
- Adding and Subtracting what?

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$\mathbf{E}\left[(h(x) - f_\theta(x))^2\right] =$$

$$\mathbf{E}\left[(h(x) \underbrace{- \mathbf{E}_\theta\left[f_\theta(x)\right] + \mathbf{E}_\theta\left[f_\theta(x)\right]}_{} - f_\theta(x))^2\right]$$

Adding and subtracting $\mathbf{E}_\theta\left[f_\theta(x)\right]$
- Note expectation is over θ
- Random variable in x

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$\mathbf{E}\left[(h(x) - f_\theta(x))^2\right] =$$

$$\mathbf{E}\left[(\underbrace{h(x) - \mathbf{E}_\theta\left[f_\theta(x)\right]}_{a} + \underbrace{\mathbf{E}_\theta\left[f_\theta(x)\right] - f_\theta(x)}_{b})^2\right]$$

Expanding in terms of a and b: $(a+b)^2 = a^2 + b^2 + 2ab$

$$\overbrace{\mathbf{E}\left[(h(x) - \mathbf{E}_\theta\left[f_\theta(x)\right])^2\right]}^{a^2} + \overbrace{\mathbf{E}\left[(\mathbf{E}_\theta\left[f_\theta(x)\right] - f_\theta(x))^2\right]}^{b^2}$$

$$\underbrace{+2\mathbf{E}\left[(h(x) - \mathbf{E}_\theta\left[f_\theta(x)\right])\,(\mathbf{E}_\theta\left[f_\theta(x)\right] - f_\theta(x))\right]}_{2ab}$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$\mathbf{E}\left[(h(x) - f_\theta(x))^2\right] =$$

$$\mathbf{E}\left[(h(x) - \mathbf{E}_\theta\left[f_\theta(x)\right])^2\right] + \mathbf{E}\left[(\mathbf{E}_\theta\left[f_\theta(x)\right] - f_\theta(x))^2\right]$$

$$+2\mathbf{E}\left[(h(x) - \mathbf{E}_\theta\left[f_\theta(x)\right])\,(\mathbf{E}_\theta\left[f_\theta(x)\right] - f_\theta(x))\right]$$

Expanding Terms and applying linearity of expectation

$$+2\mathbf{E}\left[h(x)\mathbf{E}_\theta\left[f_\theta(x)\right]\right] - 2\mathbf{E}\left[h(x)f_\theta(x)\right]$$

$$-2\mathbf{E}\left[\mathbf{E}_\theta\left[f_\theta(x)\right]\mathbf{E}_\theta\left[f_\theta(x)\right]\right] + 2\mathbf{E}\left[\mathbf{E}_\theta\left[f_\theta(x)\right]f_\theta(x)\right]$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$\mathbf{E}\left[(h(x) - f_\theta(x))^2\right] =$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$\mathbf{E}\left[(h(x) - \mathbf{E}_\theta[f_\theta(x)])^2\right] + \mathbf{E}\left[(\mathbf{E}_\theta[f_\theta(x)] - f_\theta(x))^2\right]$$
$$+2\mathbf{E}[h(x)\mathbf{E}_\theta[f_\theta(x)]] - 2\mathbf{E}[h(x)f_\theta(x)]$$
$$-2\mathbf{E}[\mathbf{E}_\theta[f_\theta(x)]\,\mathbf{E}_\theta[f_\theta(x)]] + 2\mathbf{E}[\mathbf{E}_\theta[f_\theta(x)]\,f_\theta(x)]$$

$$\mathbf{E}\left[(h(x) - f_\theta(x))^2\right] =$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$\mathbf{E}\left[(h(x) - \mathbf{E}_\theta[f_\theta(x)])^2\right] + \mathbf{E}\left[(\mathbf{E}_\theta[f_\theta(x)] - f_\theta(x))^2\right]$$
$$\cancel{+2\mathbf{E}[h(x)\mathbf{E}_\theta[f_\theta(x)]]} \underbrace{-2\mathbf{E}[h(x)f_\theta(x)]}_{\cancel{-2\mathbf{E}[h(x)\mathbf{E}_\theta[f_\theta(x)]]}}$$

$$\boxed{\int_x\int_\theta h(x)f_\theta(x)p(\theta,x)\mathrm{d}\theta\mathrm{d}x = \int_x h(x)\int_\theta f_\theta(x)p(\theta,x)\mathrm{d}\theta\mathrm{d}x}$$

h(x) does not depend on the θ

$$-2\mathbf{E}[\mathbf{E}_\theta[f_\theta(x)]\,\mathbf{E}_\theta[f_\theta(x)]] + 2\mathbf{E}[\mathbf{E}_\theta[f_\theta(x)]\,f_\theta(x)]$$

$$\mathbf{E}\left[(h(x) - f_\theta(x))^2\right] =$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$

$$\mathbf{E}\left[(h(x) - \mathbf{E}_\theta[f_\theta(x)])^2\right] + \mathbf{E}\left[(\mathbf{E}_\theta[f_\theta(x)] - f_\theta(x))^2\right]$$
$$\cancel{-2\mathbf{E}[\mathbf{E}_\theta[f_\theta(x)]\,\mathbf{E}_\theta[f_\theta(x)]]} \underbrace{+2\mathbf{E}[\mathbf{E}_\theta[f_\theta(x)]\,f_\theta(x)]}_{\cancel{+2\mathbf{E}[\mathbf{E}_\theta[f_\theta(x)]\,\mathbf{E}_\theta[f_\theta(x)]]}}$$

$\mathbf{E}_\theta[f_\theta(x)]$ does not depend on the θ

$$\mathbf{E}\left[(h(x) - f_\theta(x))^2\right] =$$

$\mathbf{E}\left[(h(x) - \mathbf{E}_\theta[f_\theta(x)])^2\right] +$ **(Bias)²** Term

$\mathbf{E}\left[(\mathbf{E}_\theta[f_\theta(x)] - f_\theta(x))^2\right]$ **Variance** Term

---

$$\mathbf{E}\left[(y - f_\theta(x))^2\right] =$$

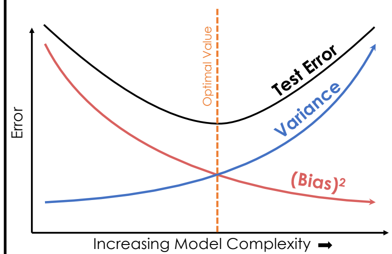
$\mathbf{E}\left[(y - h(x))^2\right] +$ **Noise** Term

$\mathbf{E}\left[(h(x) - \mathbf{E}_\theta[f_\theta(x)])^2\right] +$ **(Bias)²** Term

$\mathbf{E}\left[(\mathbf{E}_\theta[f_\theta(x)] - f_\theta(x))^2\right]$ **Variance** Term

---

Bias Variance Plot



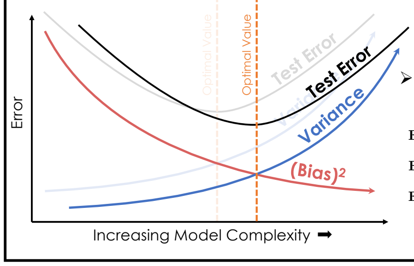
How do we compute test error?
- Use test set to pick best model?
  - No no no no noooo!
- Cross validation?
  - Yes!!!!
- What happens with more data?

$\mathbf{E}\left[(y - h(x))^2\right]$ **Noise**

$\mathbf{E}\left[(h(x) - \mathbf{E}[f_\theta(x)])^2\right]$ **(Bias)²**

$\mathbf{E}\left[(\mathbf{E}[f_\theta(x)] - f_\theta(x))^2\right]$ **Variance**

Bias Variance Plot



How do we compute test error?

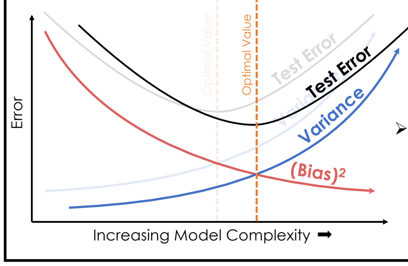
- Use test set to pick best model?
  - No no no no noooo!
- Cross validation?
  - Yes!!!!
- What happens with more data?

$\mathbf{E}\left[(y - h(x))^2\right]$ **Noise**

$\mathbf{E}\left[(h(x) - \mathbf{E}\left[f_\theta(x)\right])^2\right]$ **(Bias)²**

$\mathbf{E}\left[(\mathbf{E}\left[f_\theta(x)\right] - f_\theta(x))^2\right]$ **Variance**

Bias Variance Plot



- Cross validation?
  - Yes!!!!
- What happens with more data?

$\mathbf{E}\left[(y - h(x))^2\right]$ **Noise**

$\mathbf{E}\left[(h(x) - \mathbf{E}\left[f_\theta(x)\right])^2\right]$ **(Bias)²**

$\mathbf{E}\left[(\mathbf{E}\left[f_\theta(x)\right] - f_\theta(x))^2\right]$ **Variance**

- How do we tune the model complexity?
  - Number of features
  - Types of features
  - Different models
  - **Regularization**

Regularization

Parametrically Controlling the *Model Complexity*

Basic Idea of Regularization

$$\hat{\theta} = \arg\min_{\theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbf{Loss}\left(y_i, f_\theta(x_i)\right)}_{\text{Fit the Data}} + \underbrace{\lambda}_{\text{Regularization Parameter}} \underbrace{\mathbf{R}(\theta)}_{\text{Penalize Complex Models}}$$

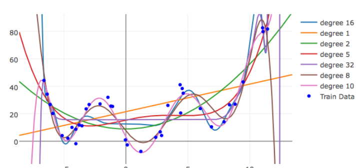
- How should we define **R(θ)**?
- How do we determine λ?

The Regularization Function **R(θ)**

Goal: *Penalize model complexity*

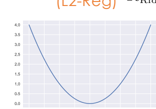
Recall earlier: $\phi(x) = \left[x, x^2, x^3, \ldots, x^p\right]$

- More features → overfitting …
- How can we control overfitting through **θ**
- **Proposal:** set weights = 0 to remove features



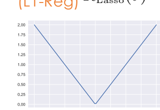
Common Regularization Functions

Ridge Regression (L2-Reg) $R_{\text{Ridge}}(\theta) = \sum_{i=1}^{d} \theta_i^2$



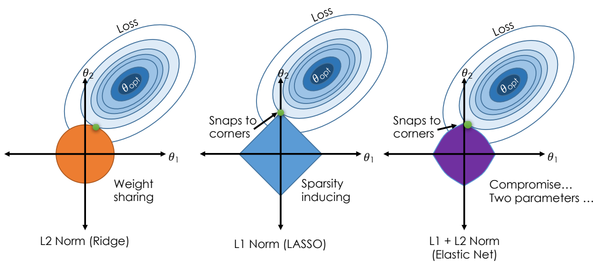
- Distributes weight across related features (robust)
- Analytic solution (easy to compute)
- Does not encourage sparsity → small but non-zero weights.

LASSO (L1-Reg) $R_{\text{Lasso}}(\theta) = \sum_{i=1}^{d} |\theta_i|$



- **Encourages sparsity** by setting weights = 0
  - Used to select informative features
- Does not have an analytic solution → numerical methods

Regularization and Norm Balls



Standardization and the **Intercept** Term

Height = θ_1 age_in_seconds + θ_2 weight_in_tons

Small Large

- Regularization penalized dimensions equally
- **Standardization**
  - Ensure that each dimensions has the same scale
  - centered around zero
- **Intercept Terms**
  - Center *y* values (e.g., subtract mean)
  - Don't regularize intercept term ← **Suggested**

Standardization

For each dimension *k*:

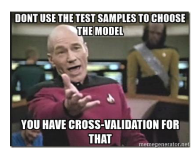
$$z_k = \frac{x_k - \mu_k}{\sigma_k}$$

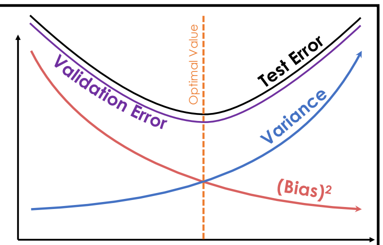
Determining the Optimal λ

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \mathbf{Loss}\left(y_i, f_\theta(x_i)\right) + \lambda \mathbf{R}(\theta)$$

- Value of λ determines bias-variance tradeoff
  - Larger values → more regularization → more bias → less variance

Determining the Optimal λ



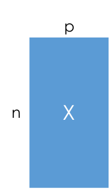


- Value of λ determines bias-variance tradeoff
  - Larger values → more regularization → more bias → less variance
- Determined through cross validation

Python Demo!

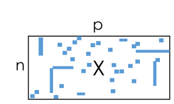
Regularization and **High-Dimensional Data**

Regularization is often used with high-dimensional data



Tall Skinny Matrix

- n >> p
- typically dense
- Regularization can help with complex feature transformations



High-dimensional sparse matrix

- p > n
- **requires** regularization
- **Goal:** to determine **informative dimensions**
  - Consider L1 Regularization. Why?
- **Goal:** is to make **robust predictions**
  - Consider L2 (+L1) Regularization

Connection to Bayesian Priors

- Ridge Regression:

Lik.: $y \sim \mathbf{N}(x^T\theta, \sigma^2_{\text{noise}})$
Prior: $\theta \sim \mathbf{N}(0, 1/\lambda)$

$$\xrightarrow[\text{(proportional)}]{\text{Posterior (Assume IID)}} \prod_{I=1}^{n} \exp\left(-\frac{\left(y - x^T\theta\right)^2}{2\sigma^2_{\text{noise}}} - \lambda\frac{\theta^2}{2}\right)$$

- LASSO:

Lik.: $y \sim \mathbf{N}(x^T\theta, \sigma^2_{\text{noise}})$
Prior: $\theta \sim \mathbf{Laplace}(0, p/\lambda)$

$$\xrightarrow[\text{(proportional)}]{\text{Posterior (Assume IID)}} \prod_{I=1}^{n} \exp\left(-\frac{\left(y - x^T\theta\right)^2}{2\sigma^2_{\text{noise}}} - \lambda\sum_{k=1}^{p}|\theta_k|\right)$$

- Regularization is often seen as applying a prior.

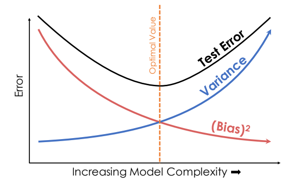
Summary

$$\mathbf{E}\left[(y - f_\theta(x))^2\right] =$$

$\mathbf{E}\left[(y - h(x))^2\right] +$ **Noise** Term

$\mathbf{E}\left[(h(x) - \mathbf{E}_\theta\left[f_\theta(x)\right])^2\right] +$ **(Bias)²** Term

$\mathbf{E}\left[(\mathbf{E}_\theta\left[f_\theta(x)\right] - f_\theta(x))^2\right]$ **Variance** Term



Regularization

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \mathbf{Loss}\left(y_i, f_\theta(x_i)\right) + \lambda\mathbf{R}(\theta)$$