

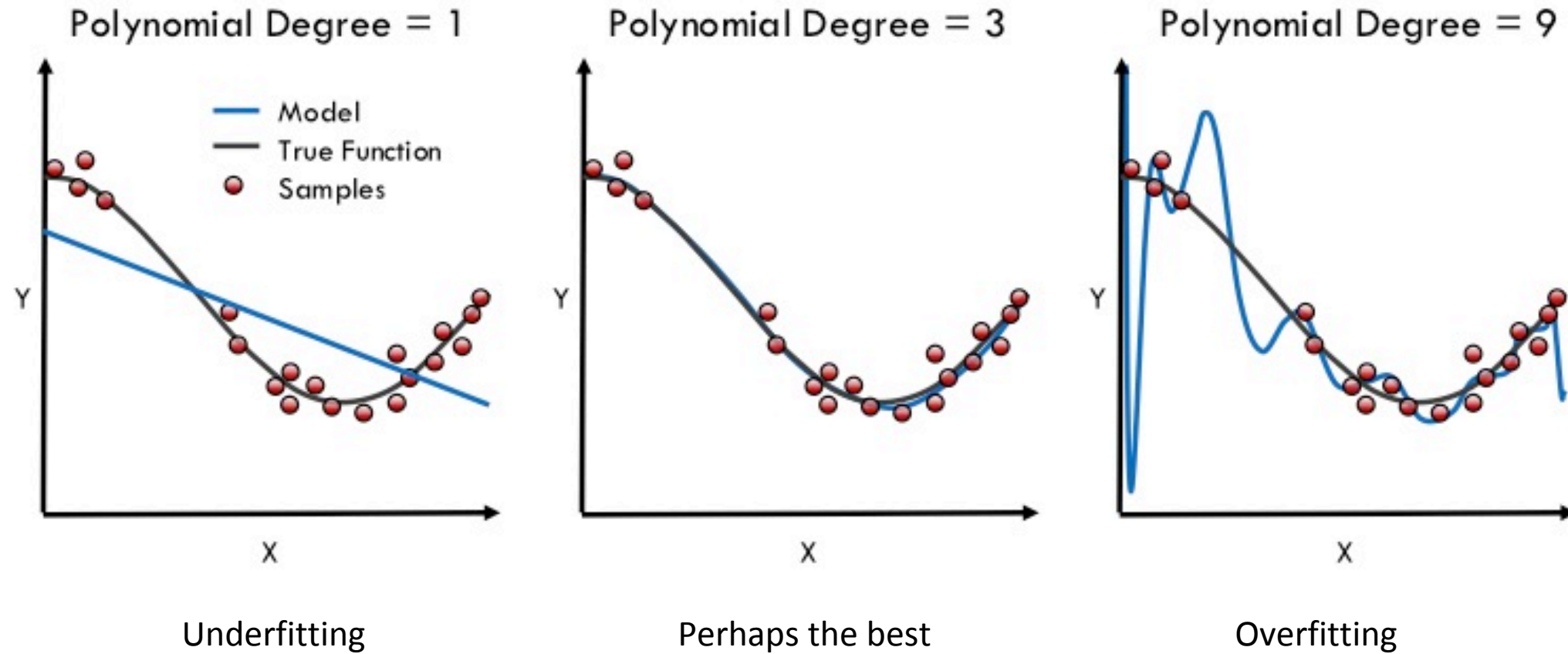
# Regularization

Kookjin Lee

([kookjin.Lee@asu.edu](mailto:kookjin.Lee@asu.edu))

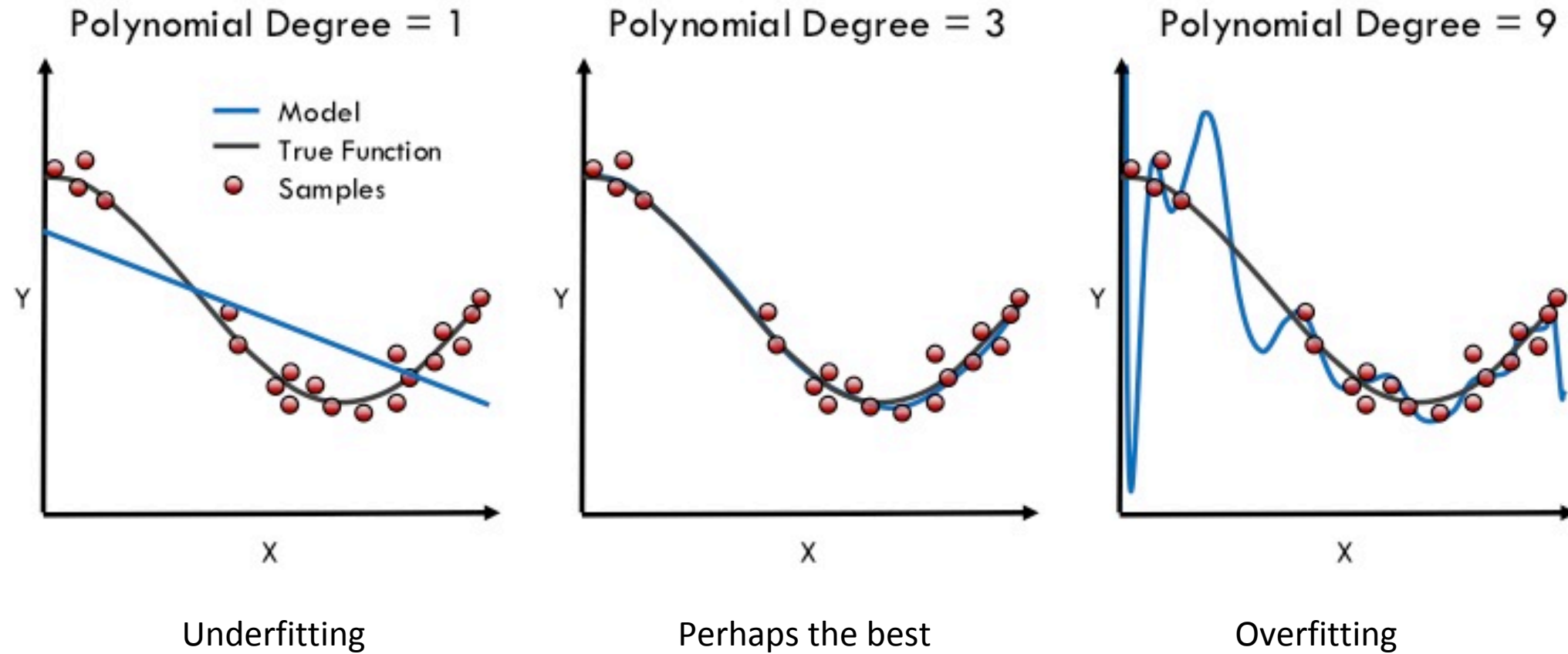
The contents of this course, including lectures and other instructional materials, are copyrighted materials. Students may not share outside the class, including uploading, selling or distributing course content or notes taken during the conduct of the course. Any recording of class sessions is authorized only for the use of students enrolled in this course during their enrollment in this course. Recordings and excerpts of recordings may not be distributed to others. (see ACD 304 – 06 , “ Commercial Note Taking Services ” and ABOR Policy 5 - 308 F.14 for more information).

# How to prevent overfitting?



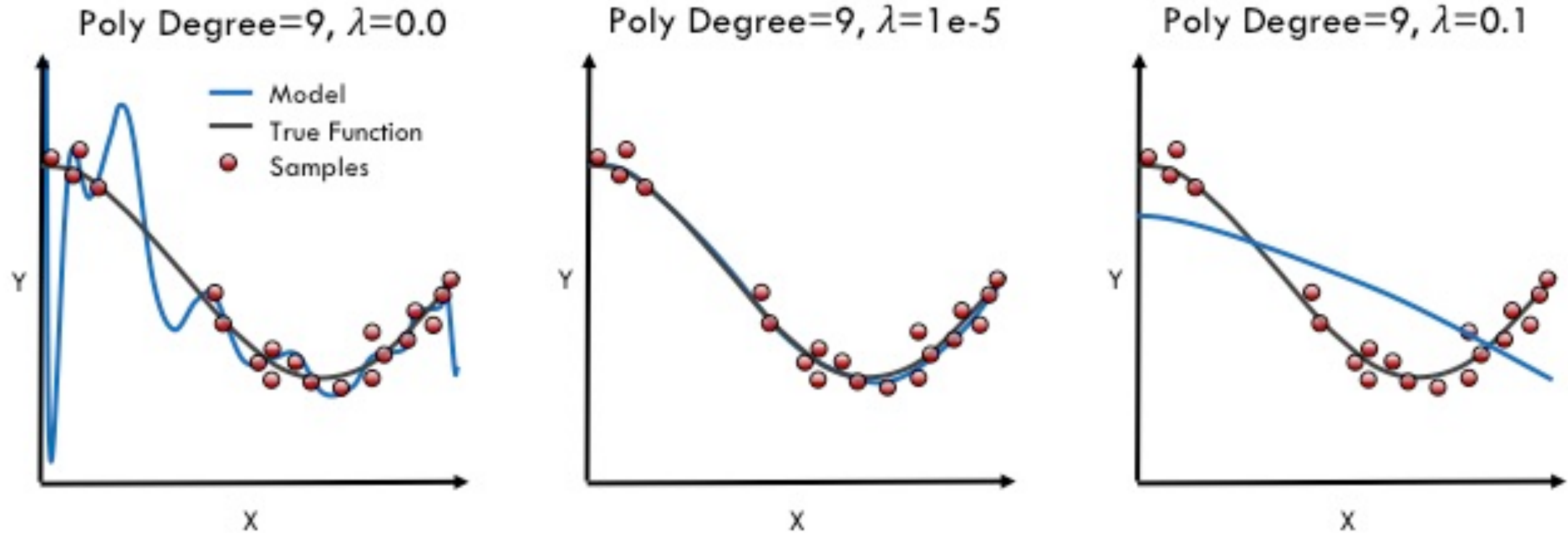
$$L = \frac{1}{2} \sum_{i=1}^{n_{\text{train}}} \left( y^{(i)} - h_{\theta}(x^{(i)}) \right)^2$$

# Regularization



$$L = \frac{1}{2} \sum_{i=1}^{n_{\text{train}}} \left( y^{(i)} - h_{\theta}(x^{(i)}) \right)^2$$

# Regularization



$$L = \frac{1}{2} \sum_{i=1}^{n_{\text{train}}} \left( y^{(i)} - h_{\theta}(x^{(i)}) \right)^2 + \lambda \sum_{i=1}^{d_{\text{max}}} \theta_i^2$$

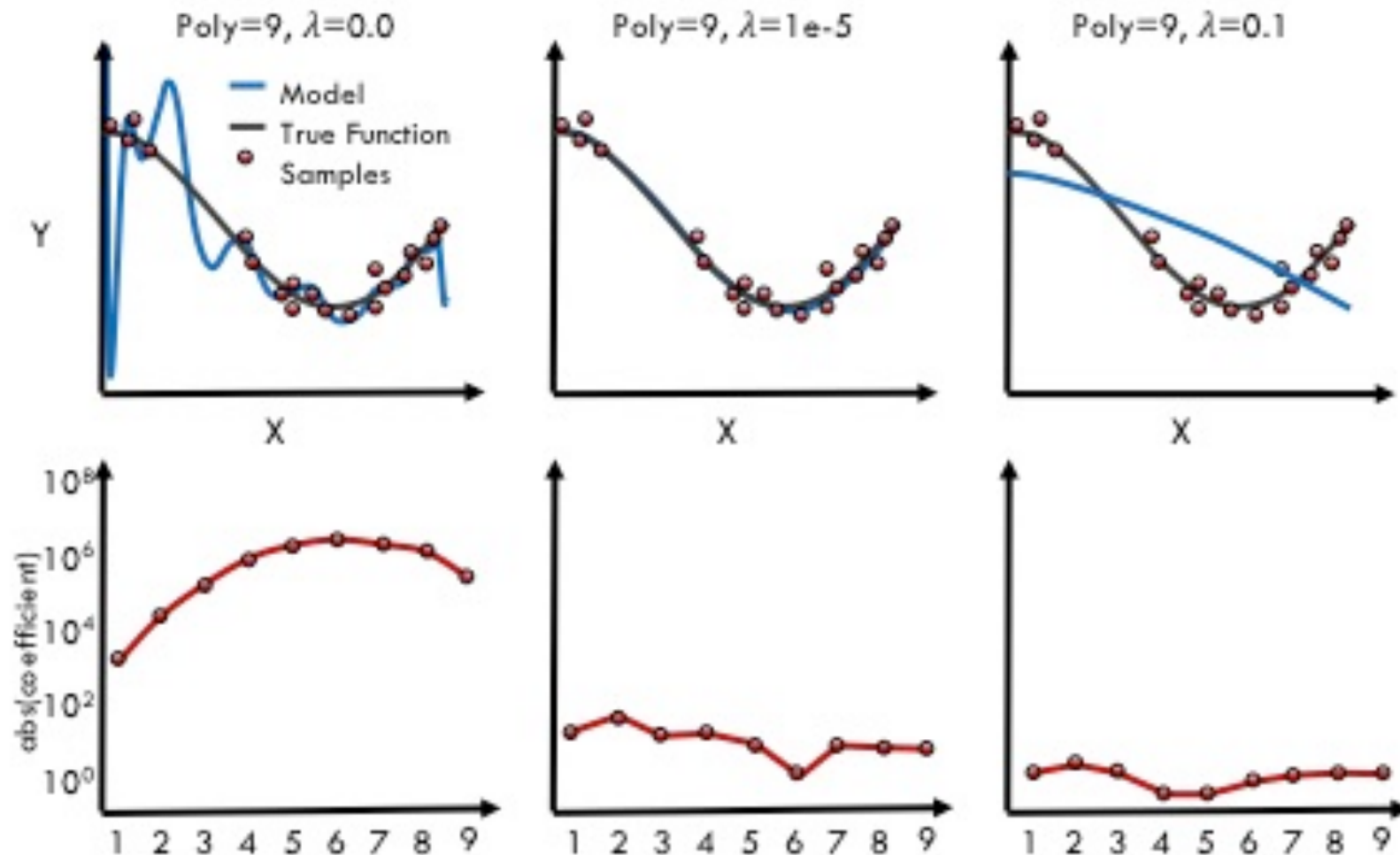
# Ridge regression

- L2-penalty

$$L = \frac{1}{2} \sum_{i=1}^{n_{\text{train}}} \left( y^{(i)} - h_{\theta}(x^{(i)}) \right)^2 + \lambda \sum_{i=1}^{d_{\text{max}}} \theta_i^2$$

- Shrinks magnitude of all coefficients
- Larger coefficients strongly penalized because of the squaring

# Effect of Ridge regression on parameters



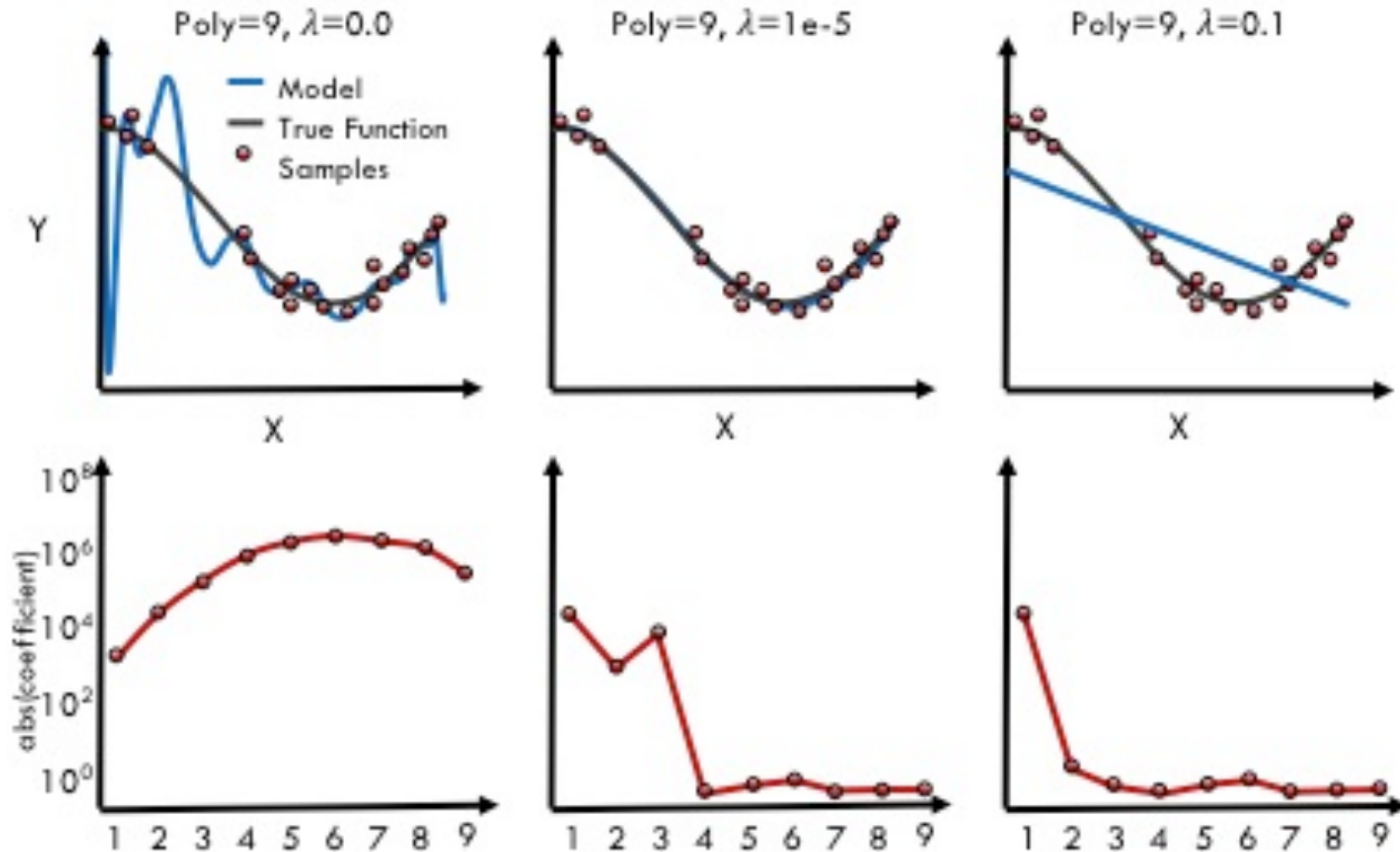
# Lasso regression

- L1-penalty

$$L = \frac{1}{2} \sum_{i=1}^{n_{\text{train}}} \left( y^{(i)} - h_{\theta}(x^{(i)}) \right)^2 + \lambda \sum_{i=1}^{d_{\text{max}}} |\theta_i|$$

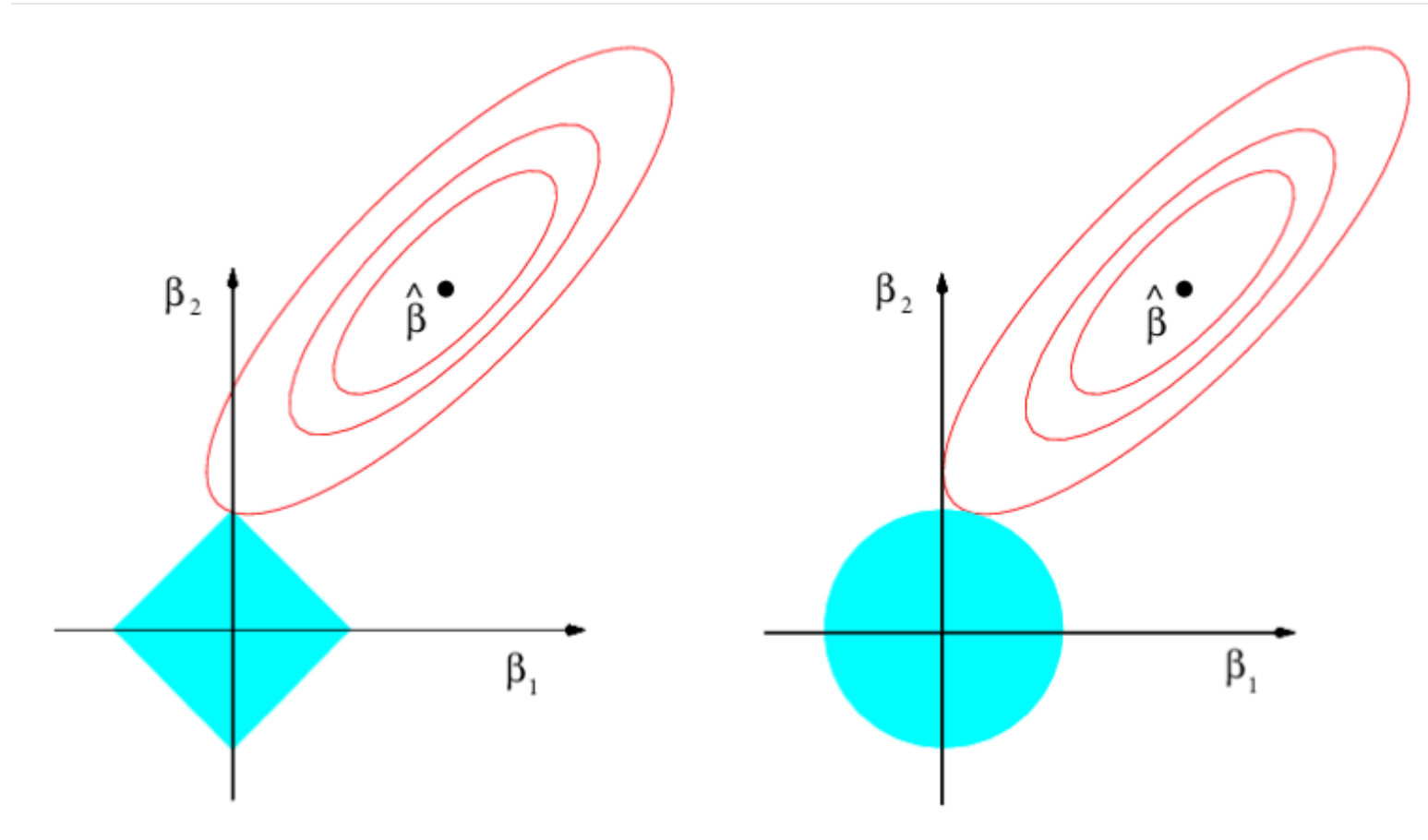
- Penalty selectively shrinks some coefficients
- Can be used for feature selection
- Slower convergence than Ridge regression

# Effect of Lasso Regression on Parameters





# L2 and L1 comparison



# Elastic Net regularization

- L1 + L2

$$L = \frac{1}{2} \sum_{i=1}^{n_{\text{train}}} \left( y^{(i)} - h_{\theta}(x^{(i)}) \right)^2 + \lambda_1 \sum_{i=1}^{d_{\text{max}}} \theta_i^2 + \lambda_2 \sum_{i=1}^{d_{\text{max}}} |\theta_i|$$

- Compromise of both Ridge and Lasso regression
- Requires tuning of additional parameter that distributes regularization penalty between L1 and L2

# Effect of Elastic Net Regression on Parameters

