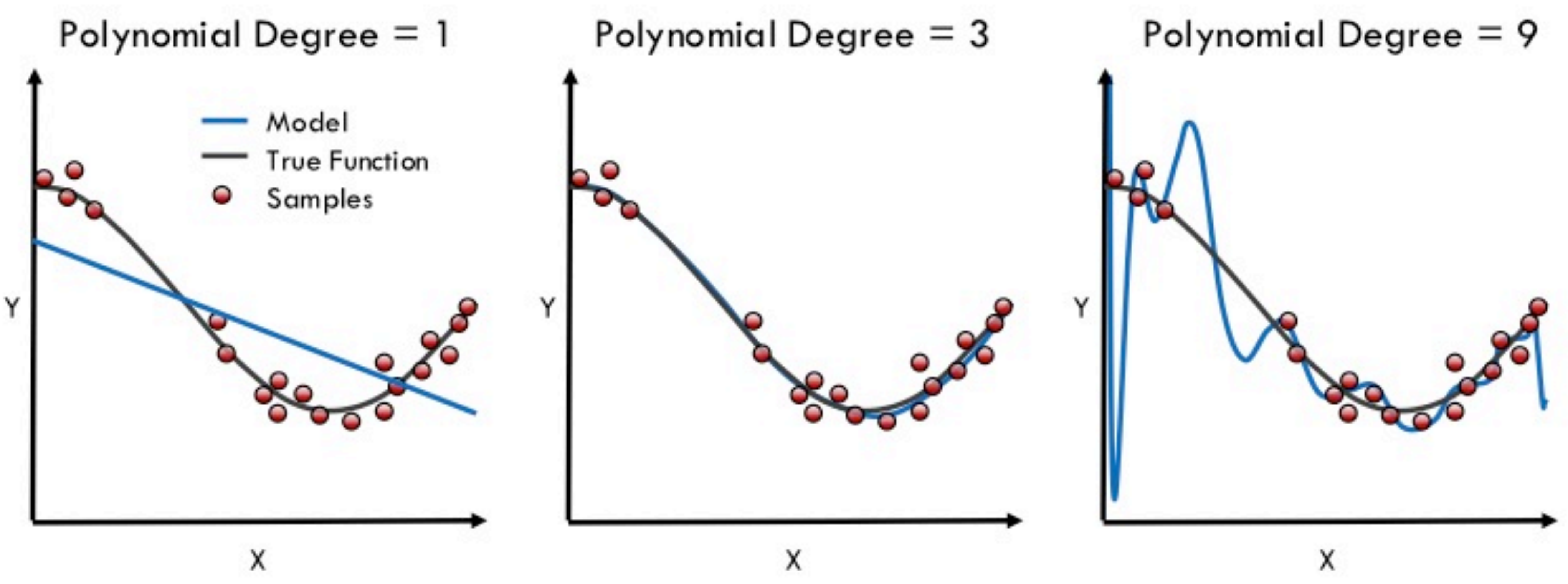
Regularization

Kookjin Lee

(kookjin.Lee@asu.edu)

How to prevent overfitting?



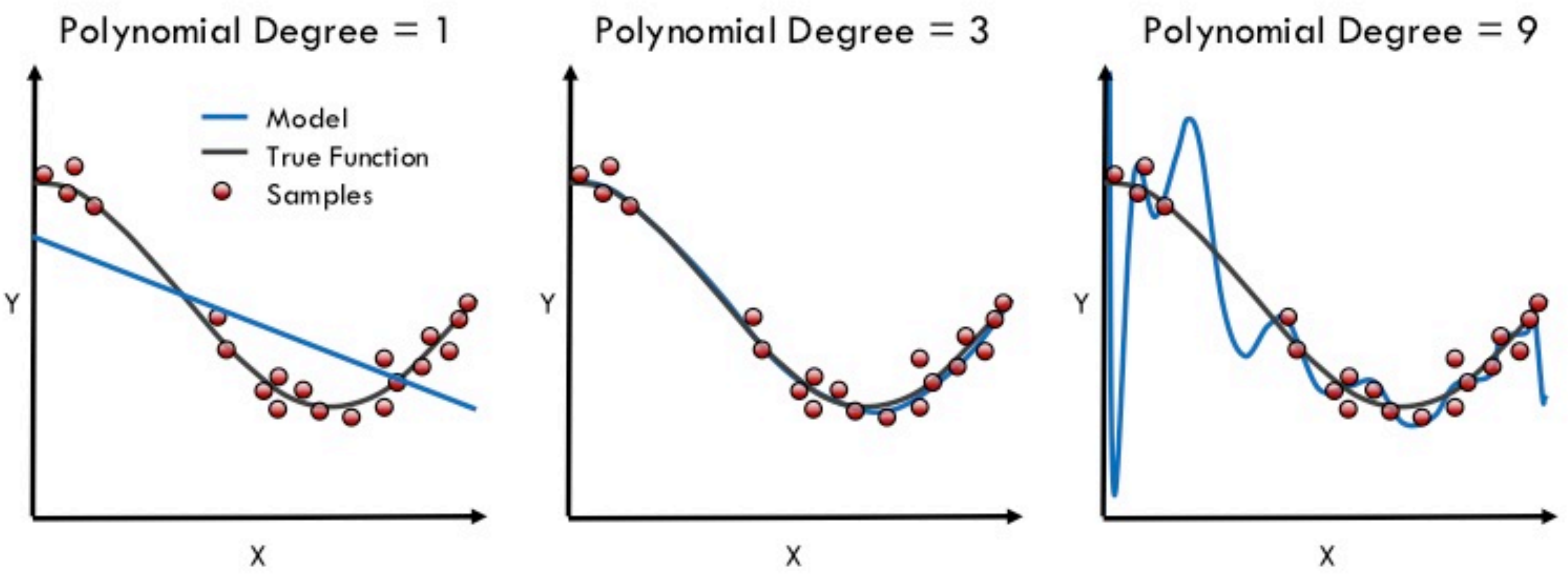
Underfitting

Perhaps the best

Overfitting

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

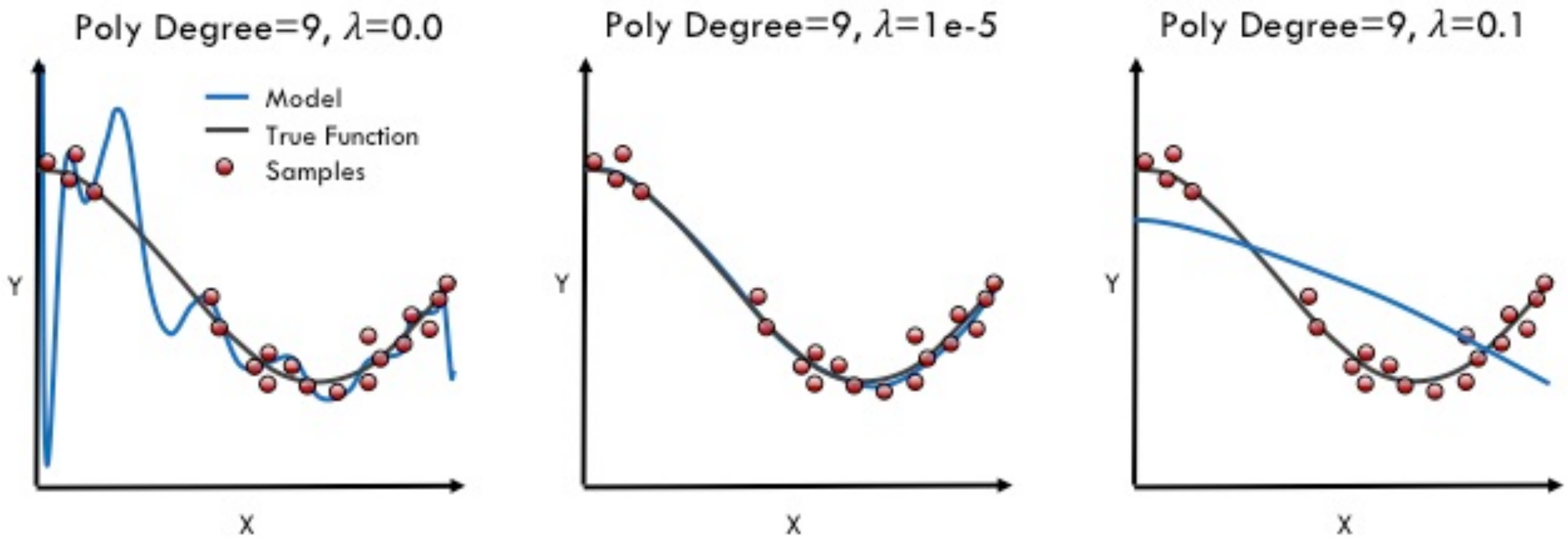
Regularization



Underfitting Perhaps the best Overfitting

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

Regularization



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

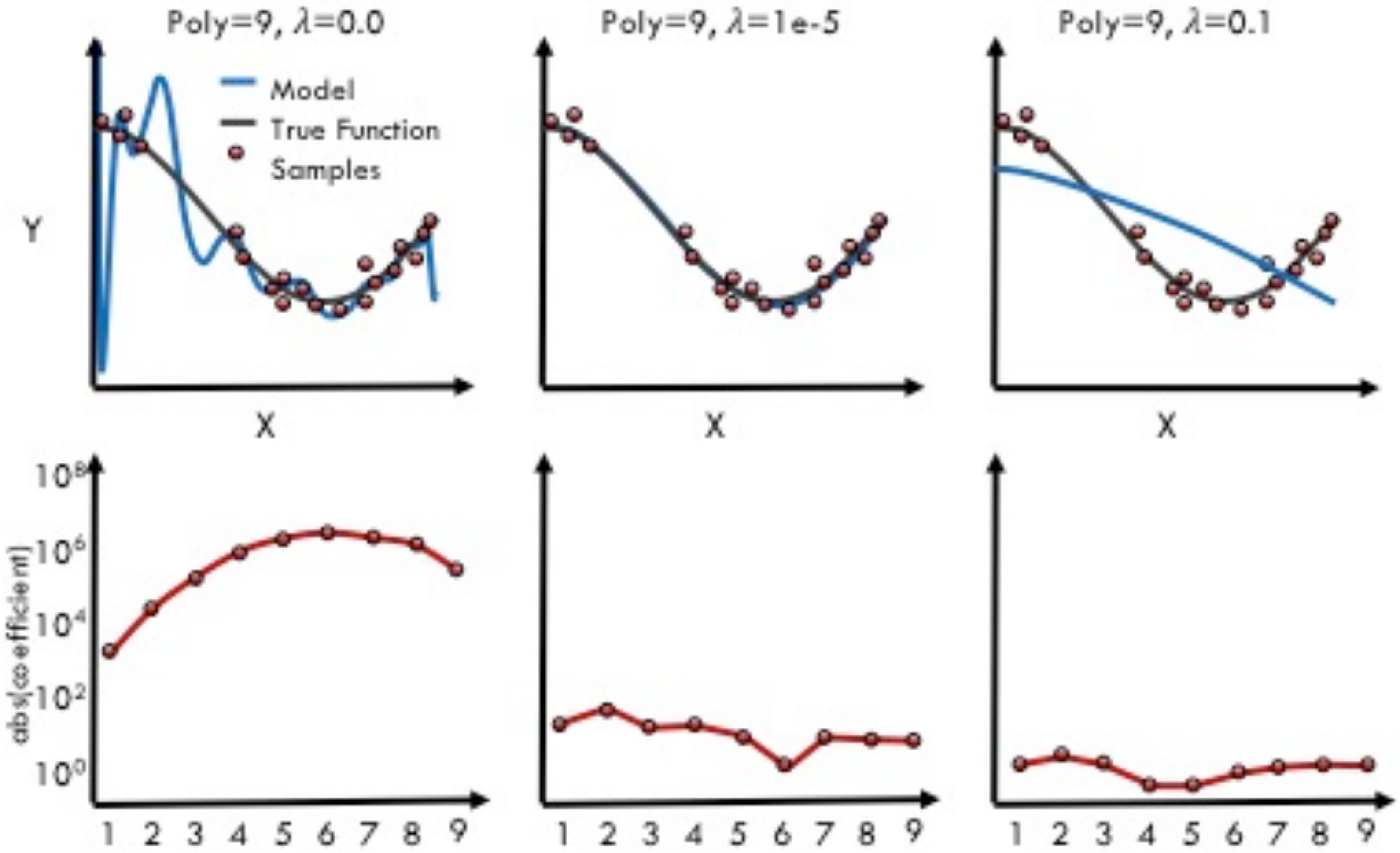
Ridge regression

- L2-penalty

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

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

Effect of Ridge regression on parameters



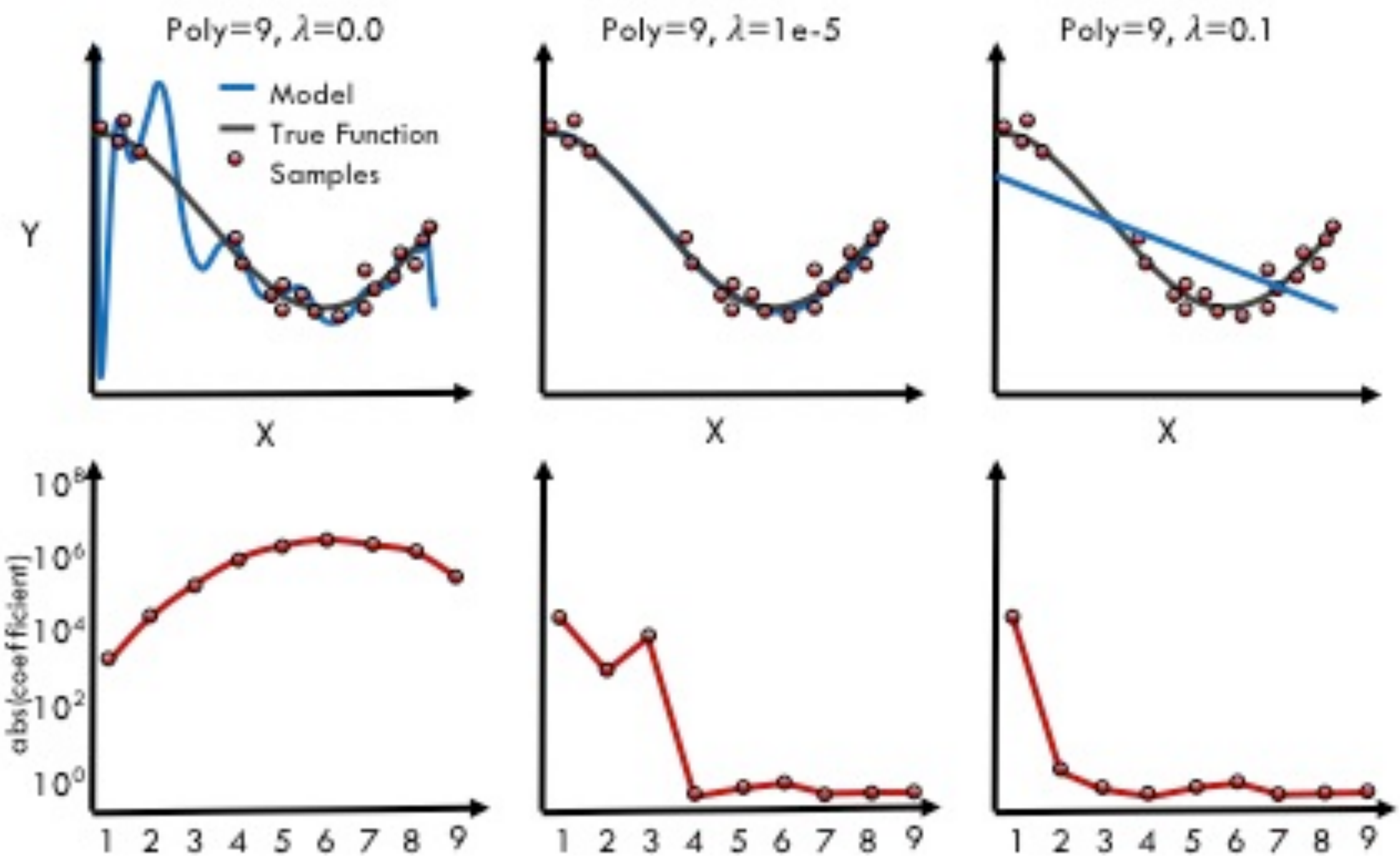
Lasso regression

- L1-penalty

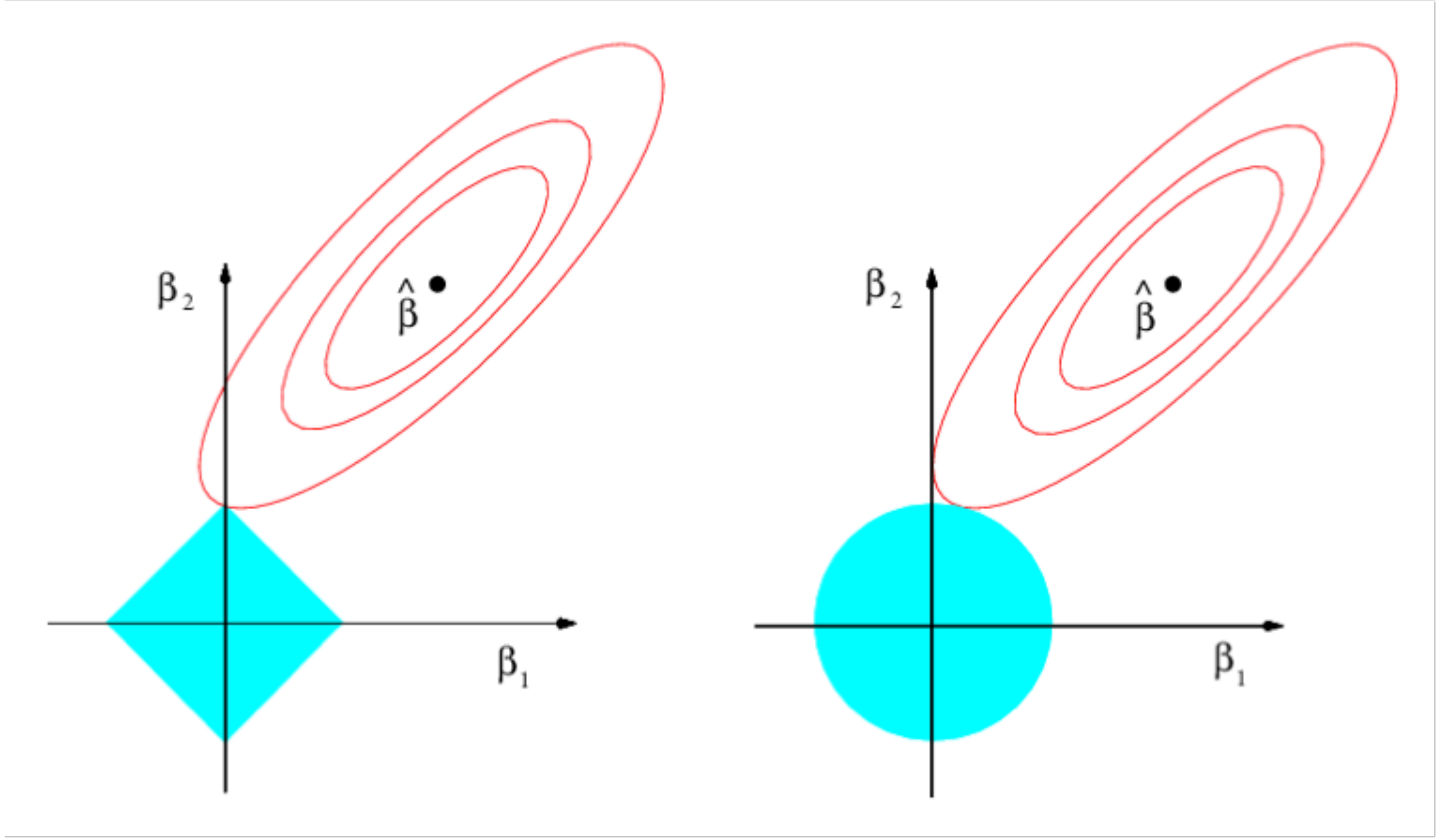
$$L = \frac{1}{2}\sum_{i=1}^{n_{\mathrm{train}}}\left(y^{(i)} - h_\theta(x^{(i)})\right)^2 + \lambda \sum_{i=1}^{d_{\mathrm{max}}} |\theta_i|$$

  - 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



Hastie, et al, Book

Elastic Net regularization

- L1 + L2

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

  - 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