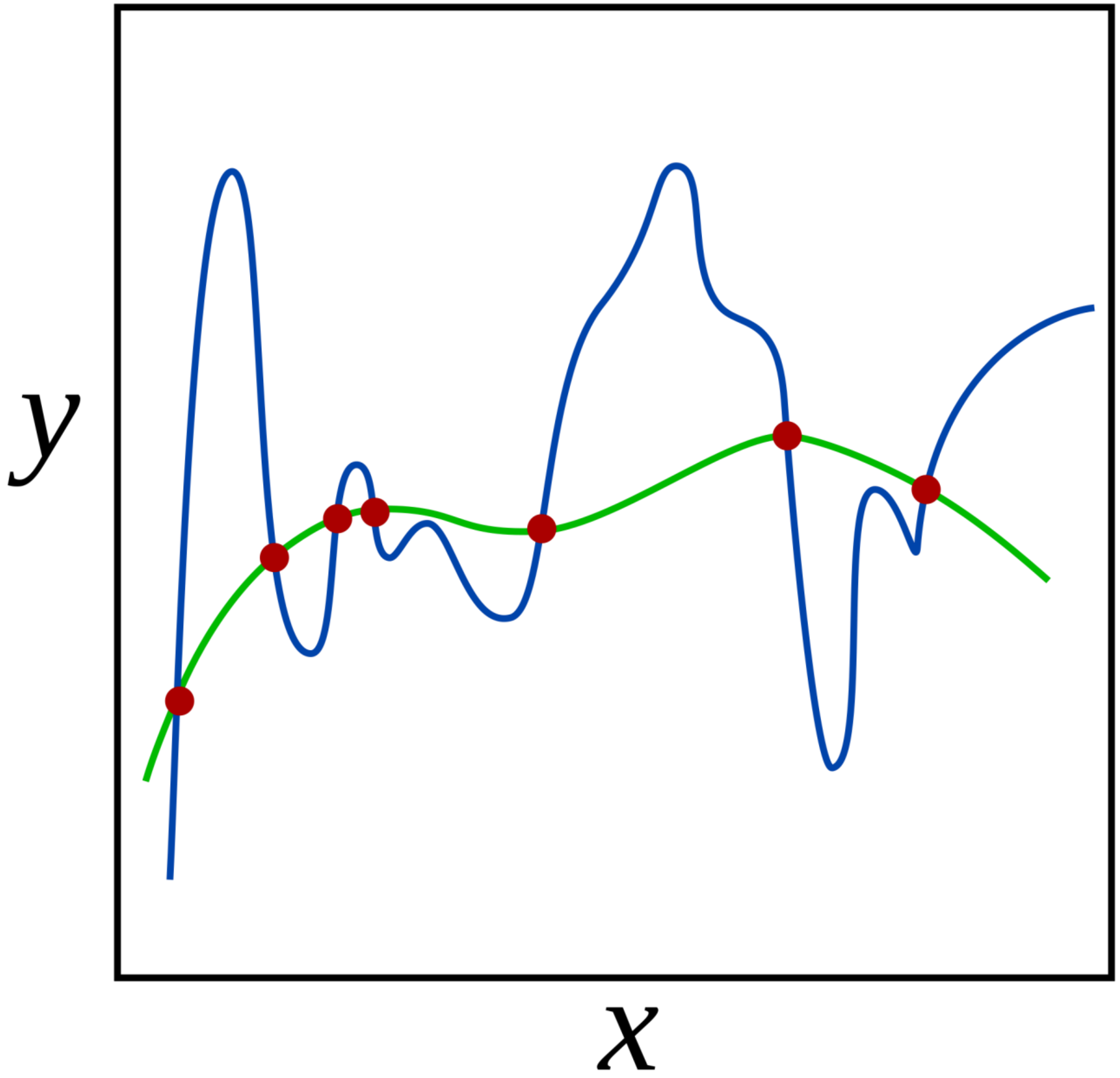
Regression Model Comparison and Overfitting

Regression Model Comparison

Regression Model	Pros	Cons
Linear Regression	Works on any size of dataset, gives information about relevance of features	The Linear Regression Assumptions
Polynomial Regression	Works on any size of dataset, works very well on non-linear problems	Need to choose the right polynomial degree for a good bias/variance tradeoff
SVR	Easily adaptable, works very well on non-linear problems, not biased by outliers	Require feature scaling (though R does it automatically), low interpretability
Decision Tree Regression	Interpretability, no need for feature scaling, works on both linear & non-linear problems	Poor results on too small datasets, overfitting can easily occur
Random Forest Regression	Powerful and accurate, good performance on many problems, including non-linear, unlikely to overfit	Low interpretability, need to choose the number of trees

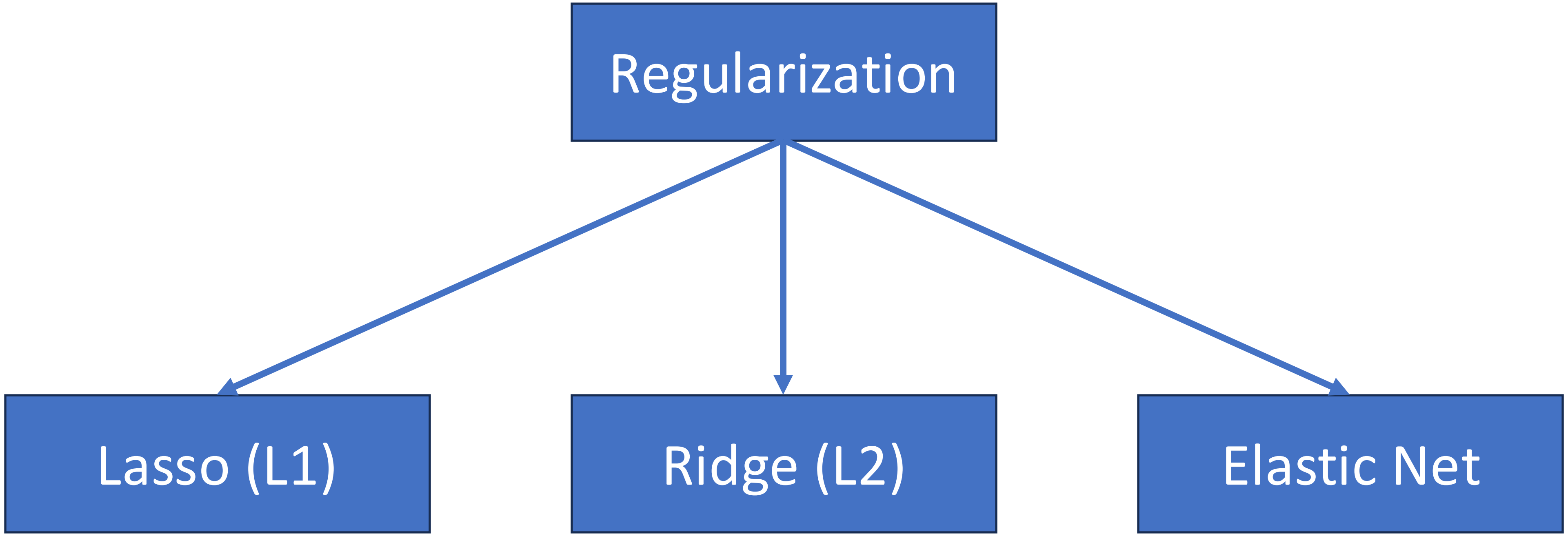
The Problem Of Overfitting



Solution

Regularization

Examples of Regularization



No Regularization

$$\text{Minimize} \sum_{i=1}^{n} \left(y^i - (b_0 + b_1 x_1^i + \ldots + b_m x_m^i)\right)^2$$

Lasso (L1 Penalty)

$$\text{Minimize} \sum_{i=1}^{n} \left(y^i - (b_0 + b_1 x_1^i + ... + b_m x_m^i)\right)^2 + \lambda(|b_1| + ... + |b_m|)$$

Ridge (L2 Penalty)

$$\text{Minimize } \sum_{i=1}^{n} \left(y^i - (b_0 + b_1 x_1^i + \ldots + b_m x_m^i)\right)^2 + \lambda(b_1^2 + \ldots + b_m^2)$$

Elastic Net

$$\text{Minimize} \sum_{i=1}^{n} \left(y^i - (b_0 + b_1 x_1^i + \ldots + b_m x_m^i)\right)^2 + \lambda_1(|b_1| + \ldots + |b_m|) + \lambda_2(b_1^2 + \ldots + b_m^2)$$