UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Lesson 5.

Overfitting and regularization

DATA SCIENCE AND AUTOMATION COURSE

MASTER DEGREE SMART TECHNOLOGY ENGINEERING

TEACHER
Mirko Mazzoleni

PLACE
University of Bergamo

Outline

1. Overfitting
2. Regularization: the technique
3. Regularization: types
4. Regularization: bias and variance

Outline

1. **Overfitting**

2. Regularization: the technique

3. Regularization: types

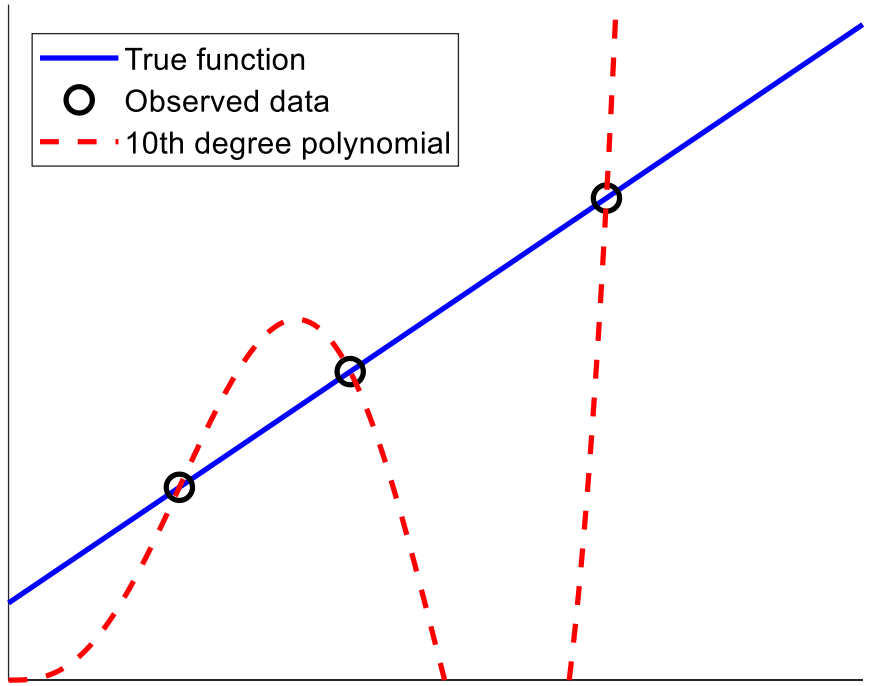
4. Regularization: bias and variance

Overfitting

Overfitting

We encountered the overfitting phenomenon when we talked about the **approximation-generalization tradeoff**.

- We saw how we have to use **simpler models if we have few data**, independently of the complexity of the true function



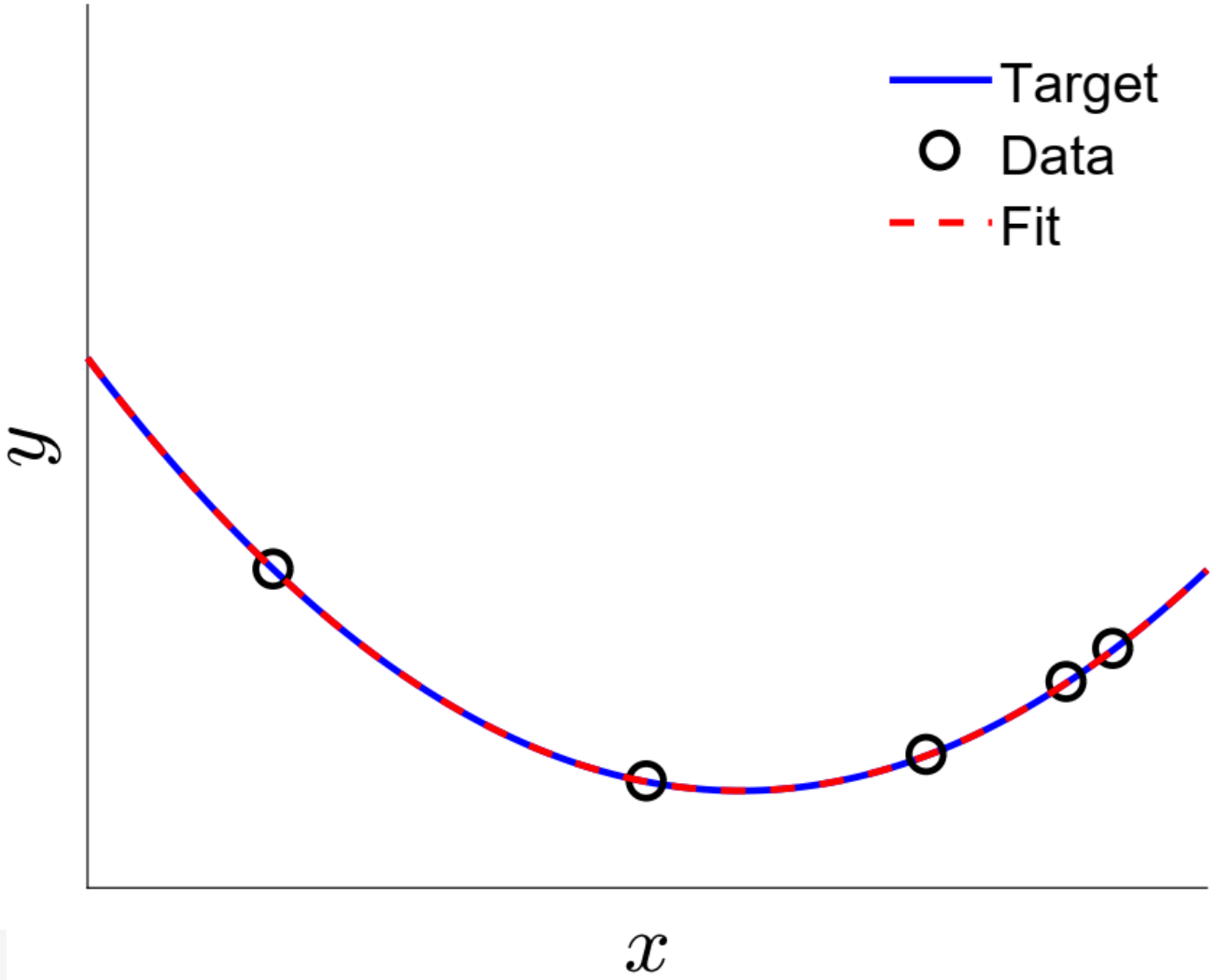
Match the **"model complexity"** to the **data resources**, not to the **target complexity**

We now introduce another cause for the overfitting: the **stochastic noise** on output data y

Overfitting example

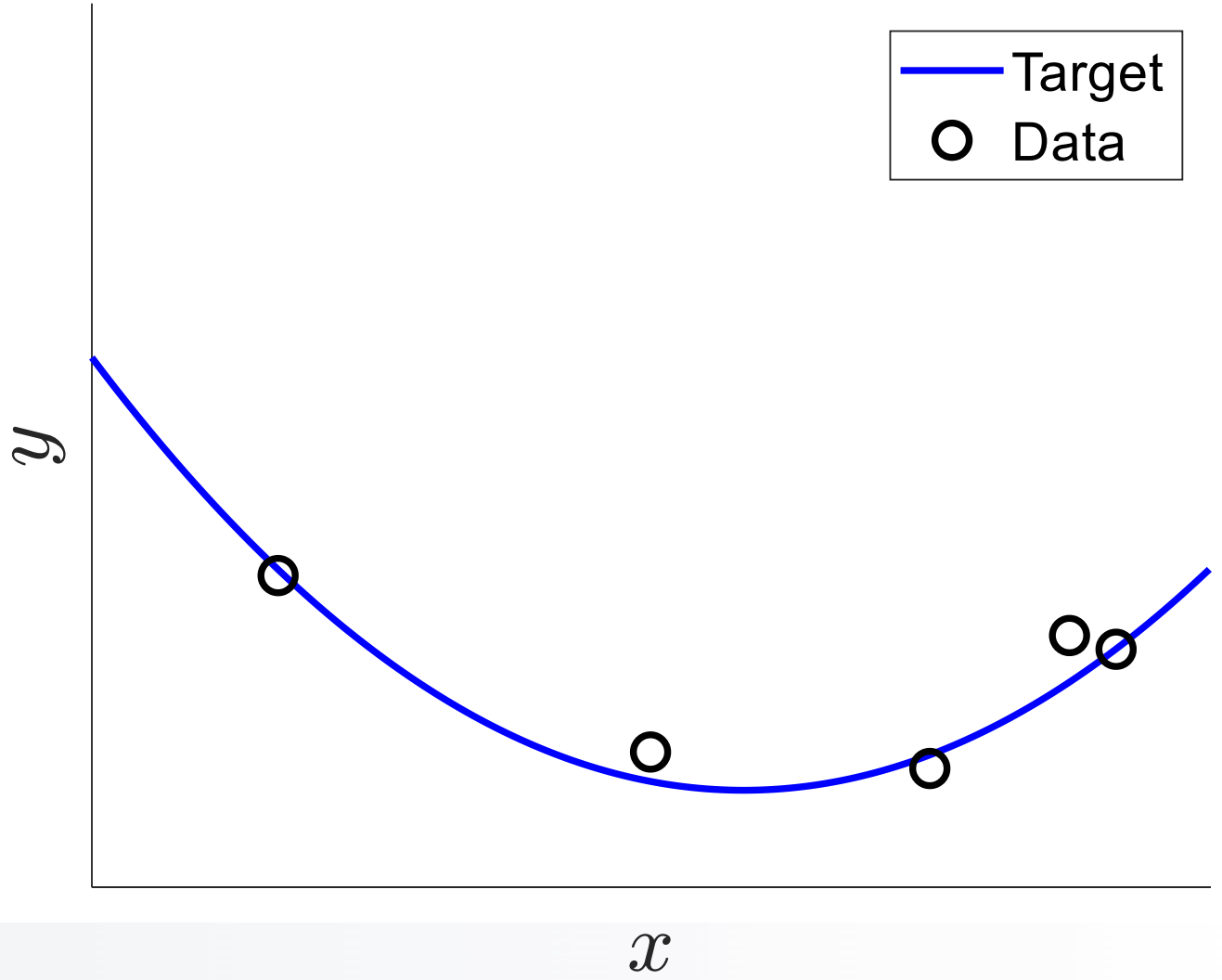
- Simple function to learn
- $N = 5$ points
- Model: 4-th order polynomial

$$E_{in} = 0 \qquad E_{out} = 0$$



Overfitting example

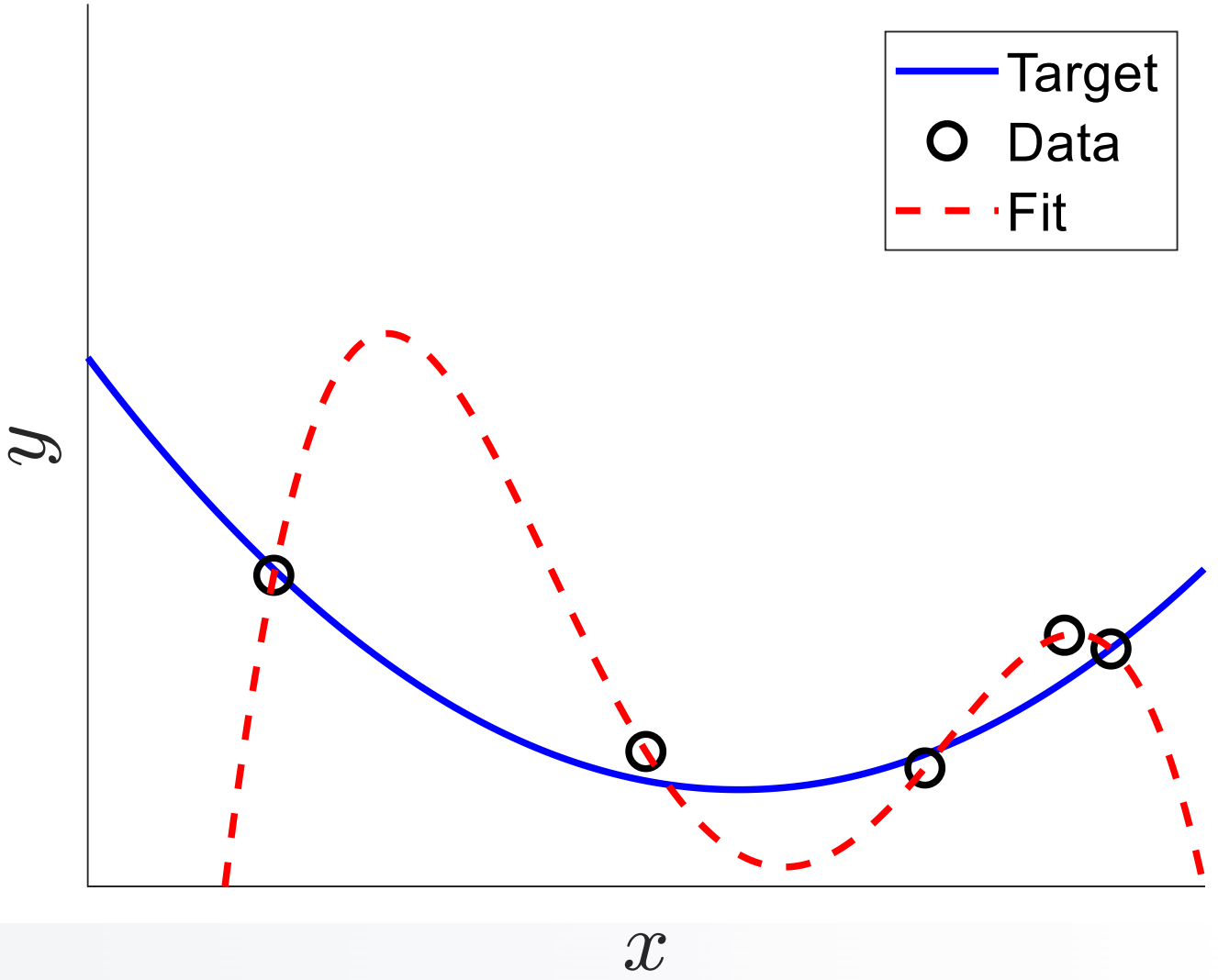
- Simple function to learn
- $N = 5$ **noisy** points
- Model: 4-th order polynomial



Overfitting example

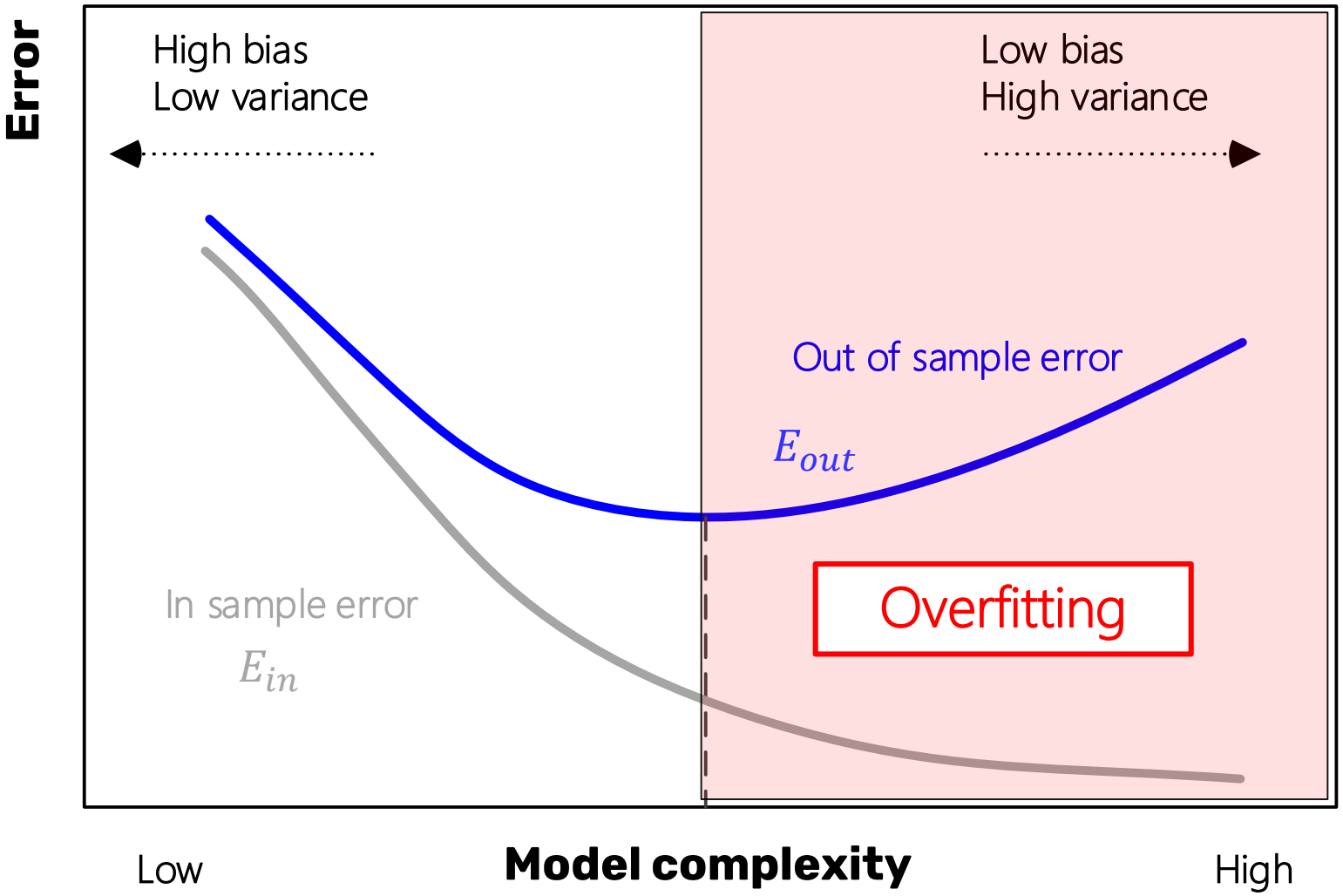
- Simple function to learn
- $N = 5$ **noisy** points
- Model: 4-th order polynomial

$$E_{in} = 0 \qquad E_{out} = \textbf{huge}$$

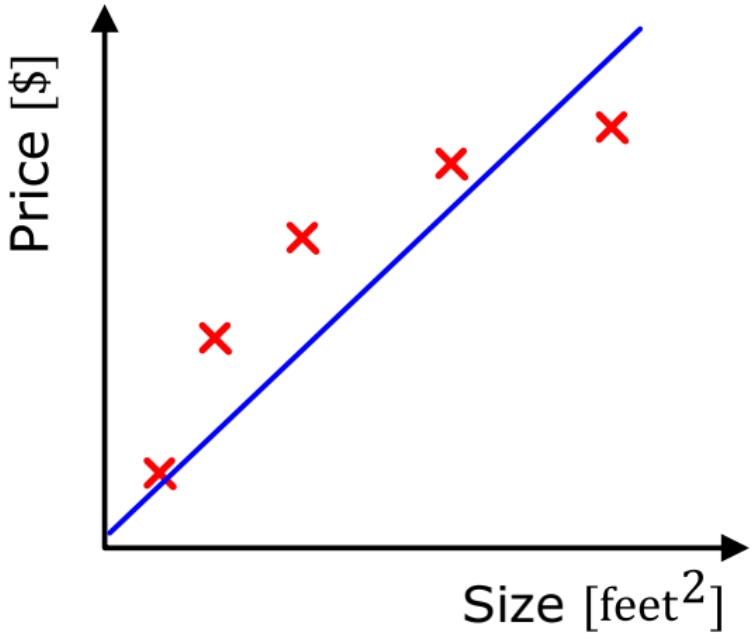


Overfitting vs. model complexity

- We talk of **overfitting** when decreasing E_{in} leads to increasing E_{out}
- Major source of failure for machine learning systems
- Overfitting leads to bad generalization
- A model can exhibit bad generalization even if it does not overfit

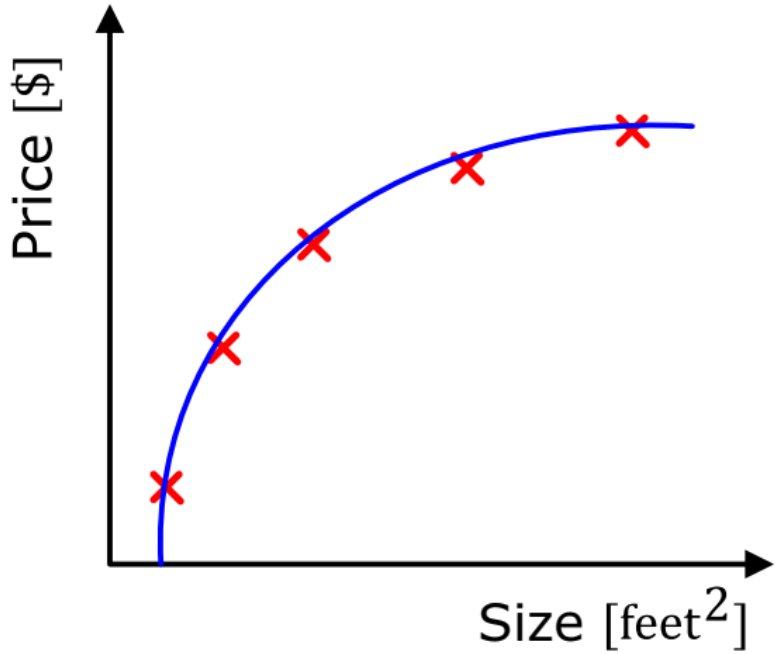


Overfitting vs. model complexity

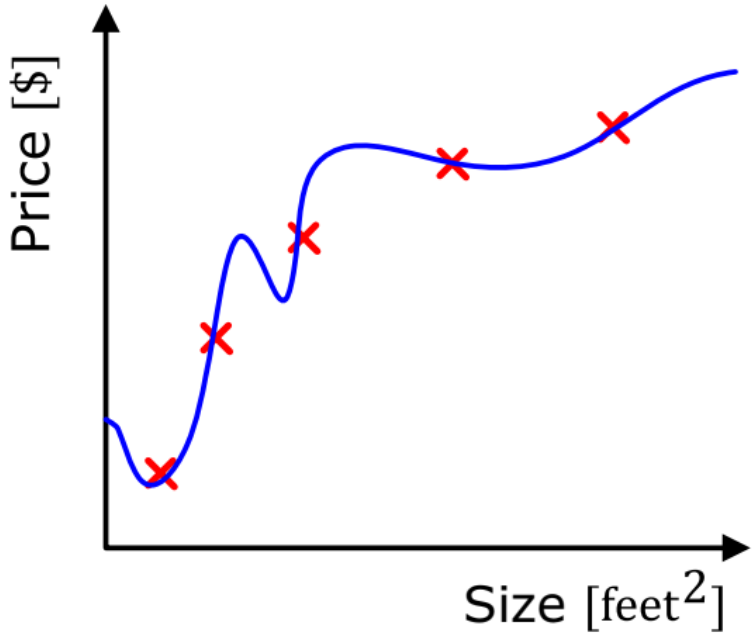


Underfit

High bias



Just right



Overfit

High variance

Bias-variance tradeoff revisited

Let the stochastic noise η be a random variable with zero mean and variance σ^2

$$\mathbb{E}_{\mathcal{D},\mathbf{x},\eta}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - (f(\mathbf{x}) + \eta(\mathbf{x}))\right)^2\right] =$$

Instead of $y = f(\boldsymbol{x})$, we observe $\tilde{y} = f(\boldsymbol{x}) + \eta(\boldsymbol{x})$

$$\underbrace{\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\text{var}} + \underbrace{\mathbb{E}_{\mathbf{x}}\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\text{bias}} + \underbrace{\mathbb{E}_{\eta,\mathbf{x}}\left[\left(\eta(\mathbf{x})\right)^2\right]}_{\sigma^2}$$

- The error σ^2 can not be driven to zero
- The stochastic noise contributes to the variance of the chosen hypotesis, causing overfitting

Outline

1. Overfitting

2. **Regularization: the technique**

3. Regularization: types

4. Regularization: bias and variance

A cure for overfitting

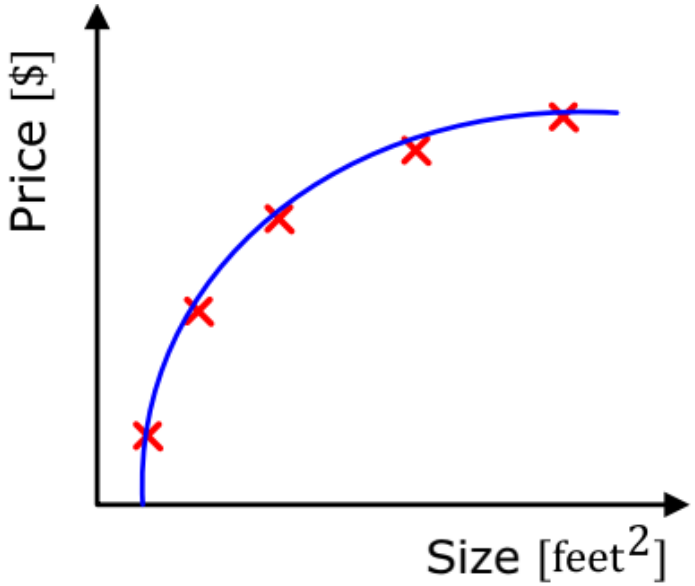
Regularization is the first line of defense against overfitting

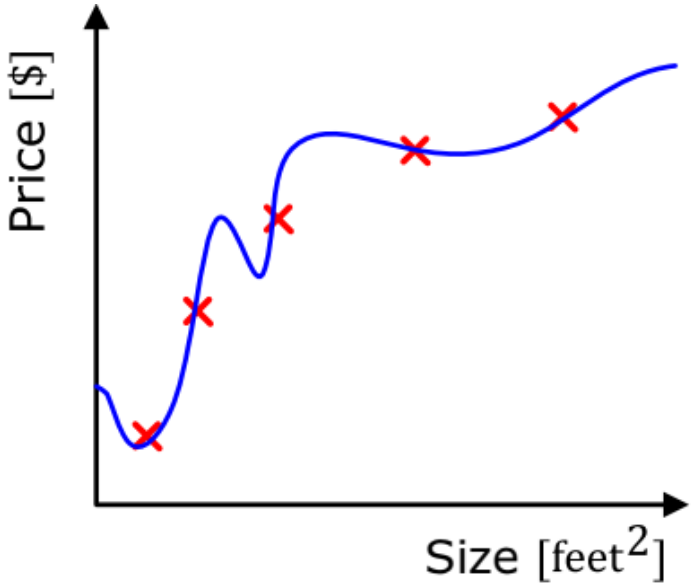
- We have seen that **complex models** are more prone to overfitting
  - ✓ This is because they are more powerful, and thus they can fit the noise

- **Simple models** exhibit less variance because of their limited expressivity. This gain in variance often is greater than their greater bias
  - ✓ However, if we stick only to simple models, we may not end up with a satisfying approximation of the target function f

*How can we retain the benefits of **both** worlds?*

A cure for overfitting





$\mathcal{H}_2$: $y = \theta_0 + \theta_1 x + \theta_2 x^2$

$\mathcal{H}_4$: $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

We can "recover" the model $\mathcal{H}_2$ from the model $\mathcal{H}_4$ by **imposing** $\theta_3 = \theta_4 = 0$

This can be done by minimizing, along with $J(\boldsymbol{\theta})$, also the value of the parameters θ_3, θ_4

A cure for overfitting

More generally, instead of minimizing the in-sample error E_{in} (i.e. the cost function $J(\boldsymbol{\theta})$), minimize the **augmented error:**

For simplicity, suppose $J(\boldsymbol{\theta})$ as squared error function

$$E_{aug}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left(h(\boldsymbol{x}(i),\boldsymbol{\theta}) - f\left(\boldsymbol{x}(i)\right)\right)^2 + \lambda \cdot \sum_{j=1}^{d-1}\left(\theta_j\right)^2$$

- Usually we do not want to penalize the intercept θ_0, so j starts from 1
- The term $\Omega(h) = \sum_{j=1}^{d-1}\left(\theta_j\right)^2$ is called **regularizer**
- The regularizer is a penalty term which depends on the hypothesis h
- The term λ (regularization hyper-parameter) weights the importance of minimizing $J(\boldsymbol{\theta})$, with respect to minimizing $\Omega(h)$.

A cure for overfitting

The minimization of E_{aug} can be viewed as a **constrained** minimization problem

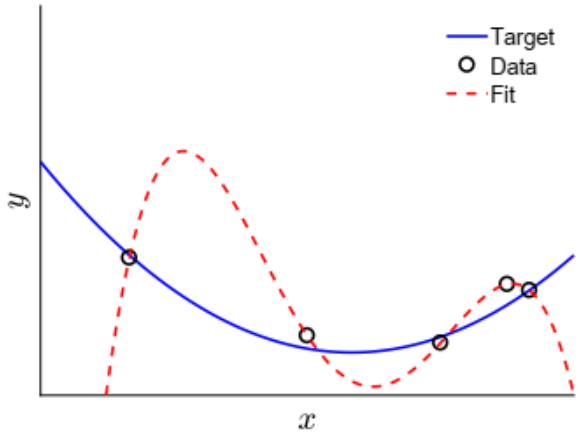
$$\text{minimize } E_{in}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left(h(\boldsymbol{x}(i),\boldsymbol{\theta}) - f\left(\boldsymbol{x}(i)\right)\right)^2$$

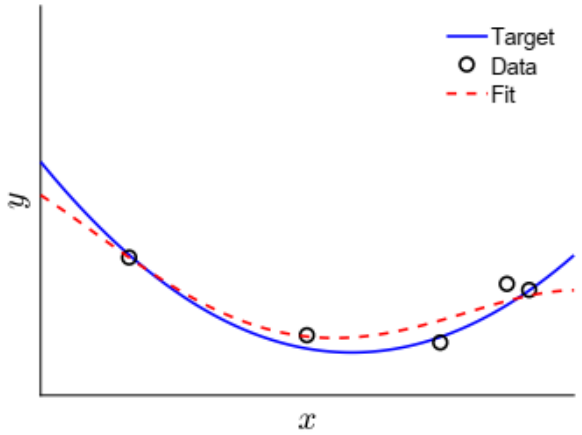
$$\text{subject to } \boldsymbol{\theta}^T\boldsymbol{\theta} \leq C$$

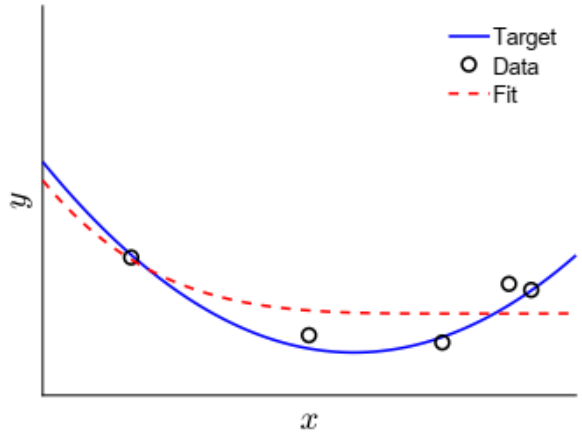
- With this view, we are explicitly **constraining the weights to not have large values**
- There is a relation between C and λ in such a way that if $C \uparrow$ then $\lambda \downarrow$
- In fact, bigger C means that the weights can be greater. This is equal to set for a lower λ, because the regularization term will be less important, and therefore the weights will not be shrunken as much

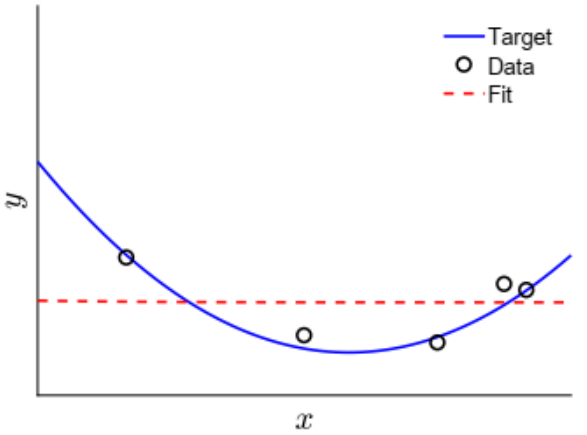
Effect of λ

λ_1 | $\lambda_2 > \lambda_1$ | $\lambda_3 > \lambda_2$ | $\lambda_4 > \lambda_3$









Overfit → → **Underfit**

Augmented error

General form of the augmented error

$$E_{aug}(\boldsymbol{\theta}) = E_{in}(\theta) + \lambda\Omega(h)$$

Recalling the VC generalization bound

$$E_{out}(\boldsymbol{\theta}) \leq E_{in}(\theta) + \Omega(\mathcal{H})$$

- $\Omega(h)$ is a measure of complexity of a specific hypothesis $h \in \mathcal{H}$
- $\Omega(\mathcal{H})$ measures the complexity of the hypothesis space $\mathcal{H}$
- The two quantities are obviously related, in the sense that a more complex hypothesis space $\mathcal{H}$ is described by more complex function h

The augmented error E_{aug} is **better** than E_{in} as a proxy for E_{out}

Augmented error

The holy Grail of machine learning would be to have a formula for E_{out} to minimize

- In this way, it would be possible to **directly minimize** the out of sample error instead of the in sample one
- **Regularization** helps by estimating the quantity $\Omega(h)$, which, added to E_{in}, gives E_{aug}, an estimate of E_{out}

Outline

1. Overfitting

2. Regularization: the technique

3. **Regularization: types**

4. Regularization: bias and variance

Choice of the regularizer

There are many choices of possible regularizers. The most used ones are:

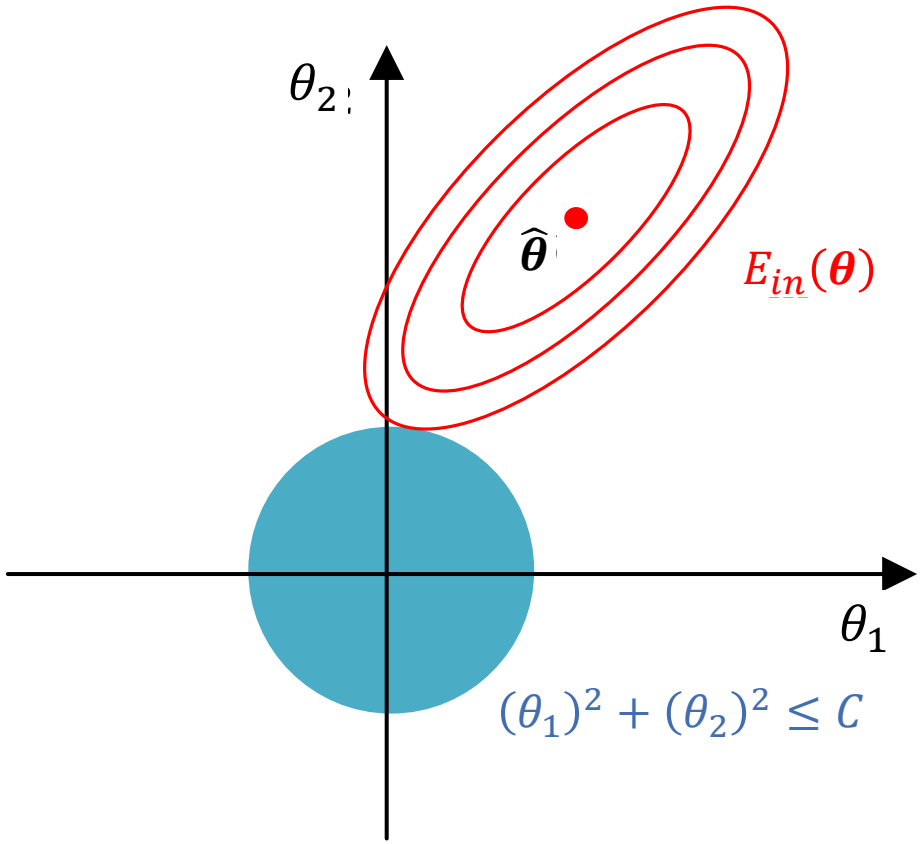
- L_2 **regularizer:** also called Ridge regression $\Omega(h) = \sum_{j=1}^{d-1}(\theta_j)^2$
- L_1 **regularizer:** also called Lasso regression $\Omega(h) = \sum_{j=1}^{d-1}|\theta_j|$
- **Elastic-net regularizer:** $\Omega(h) = \sum_{j=1}^{d-1}\beta(\theta_j)^2 + (1-\beta)\sum_{j=1}^{d-1}|\theta_j|$

The different regularizers behaves differently:

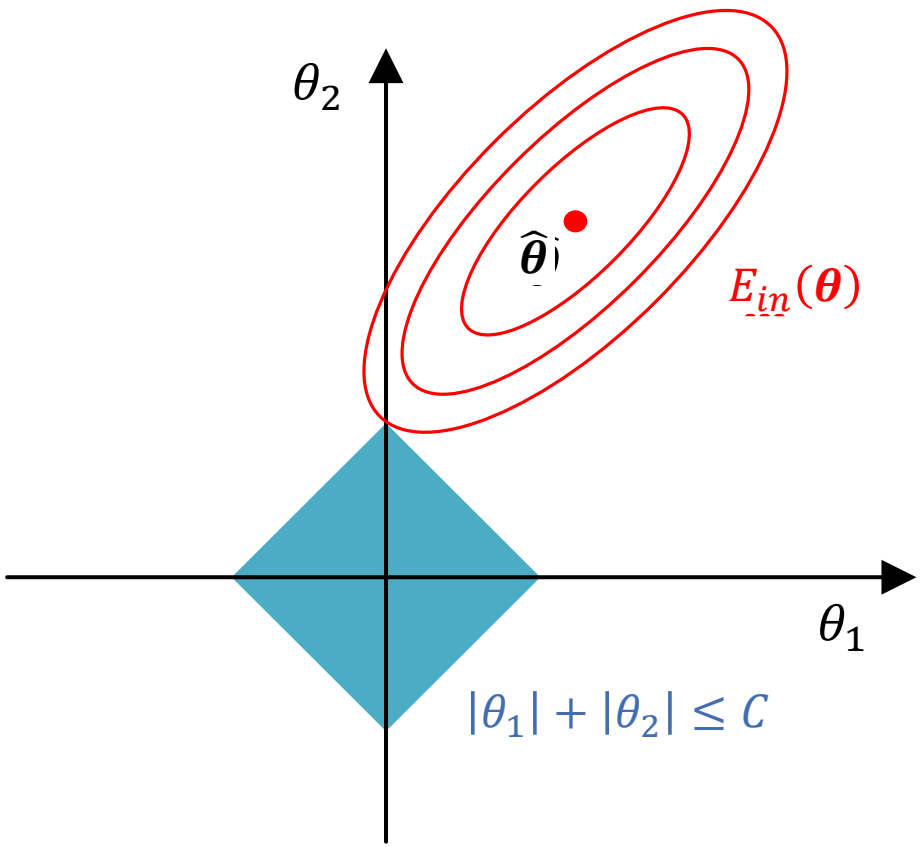
- The ridge penalty tends to shrink all coefficients to a **lower value**
- The lasso penalty tends to set more coefficients **exactly to zero**
- The elastic-net penalty is a compromise between ridge and lasso, with the β value controlling the two contributions

Geometrical interpretation

Ridge



Lasso



Outline

1. Overfitting
2. Regularization: the technique
3. Regularization: types
4. **Regularization: bias and variance**

Regularization and bias-variance

The effects of the regularization procedure can be observed in the **bias and variance** terms

- Regularization **trades bias** in order to considerably **decrease the variance** of the model
- Regularization strives for **smoother** hypothesis, thus reducing the opportunities to overfit
- The amount of regularization λ **has to be chosen specifically** for each type of regularizer
- Usually λ is chosen by **cross-validation**