Linear combination of features $\quad \hat{y} = \underset{\downarrow}{\Theta_0} + \Theta_1 x_1 + \cdots + \Theta_d x_d = X\Theta \rightsquigarrow (d+1 \times 1)$

bias term

X: $n \times (d+1)$

Training data

Now, we need define a goal or objective function: Minimize the difference between y_{actual} & $\hat{y}_{predicted}$

$$L(\Theta) = E(\Theta) = \frac{1}{N} \sum_{i=1}^{N} (y_a - \hat{y}_p)^2 = E\left[(y_a - \hat{y}_p)^2 \right]$$

$$y = X\Theta \Rightarrow \underset{(d+1 \times 1)}{\Theta} = X^{-1} y \qquad X^{-1} = (X^T X)^{-1} X^T$$

$$\underset{(d+1 \times 1)}{\Theta} = (X^T X)^{-1} X^T y$$

$$\theta^{\{t+1\}} \longleftarrow \theta^{\{t\}} - \alpha \frac{\partial L(\theta)}{\partial \theta}$$

⇒ GD : In each iteration, we need to go over All data points

⇒ SGD : " " " , " " " " " One "

⇒ BGD : " " " , " " " " " a subset or batch "

$$X = \begin{array}{c} sqf \\ \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix} \end{array} \quad Y = \begin{array}{c} \text{Rent price} \\ \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix} \end{array} \quad Z = \begin{array}{c} sqf \quad sqf^2 \quad sqf^3 \quad \ldots \quad sqf^d \\ \begin{bmatrix} \quad & \quad & \quad & \quad & \quad \\ \quad \\ \quad \end{bmatrix} \end{array}$$

overfitting

$$E(\theta) = L(\theta) = E\left[(y_a - \hat{y}_p)^2\right] = \underbrace{bias^2 + Variance}_{\text{bias-variance trade off}}$$

$$X = \begin{array}{cc} h & w \\ \multicolumn{2}{c}{\begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix}} \end{array} \quad Z = \begin{array}{cccccccc} h & w & h^2 & w^2 & hw & h^2w & hw^2 & \ldots \\ \multicolumn{8}{c}{\begin{bmatrix} \quad \\ \quad \end{bmatrix}} \end{array}$$

Regularized Linear Regression

Mahdi Roozbahani
Georgia Tech

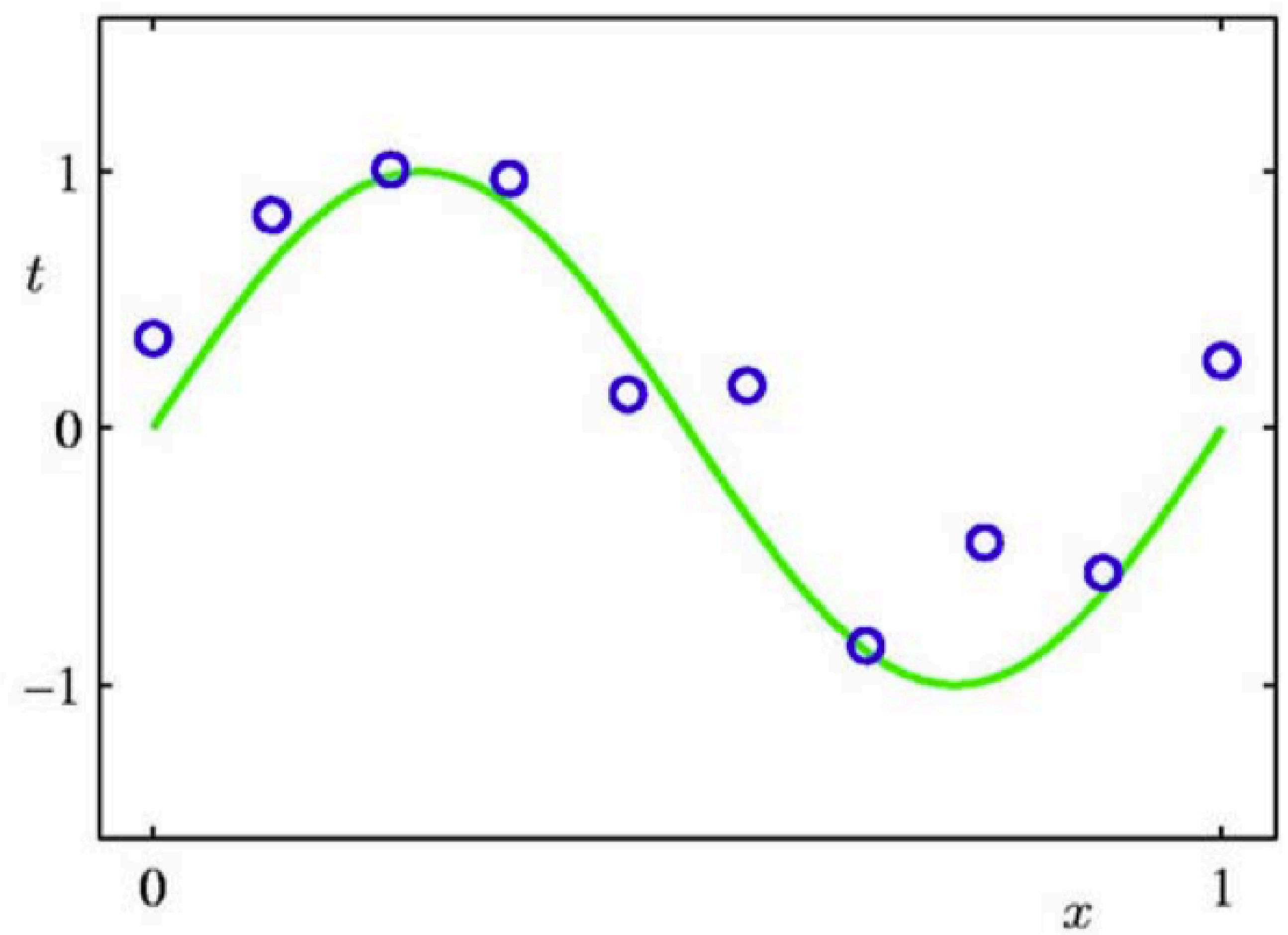
These slides are inspired based on slides from Andrew Zisserman, Jonathan Taylor, Chao Zhang and Yaser Abu-Mostafa.


EVERY GROUP PROJECT
DOES 99% OF THE WORK
HAS NO IDEA WHAT'S GOING ON THE WHOLE TIME
SAYS HE'S GOING TO HELP BUT HE'S NOT
DISAPPEAR AT THE VERY BEGINNING AND DOESN'T SHOW UP AGAIN TIL THE VERY END

Outline

- Overfitting and regularized learning
- Ridge regression
- Lasso regression
- Determining regularization strength

Regression: Recap

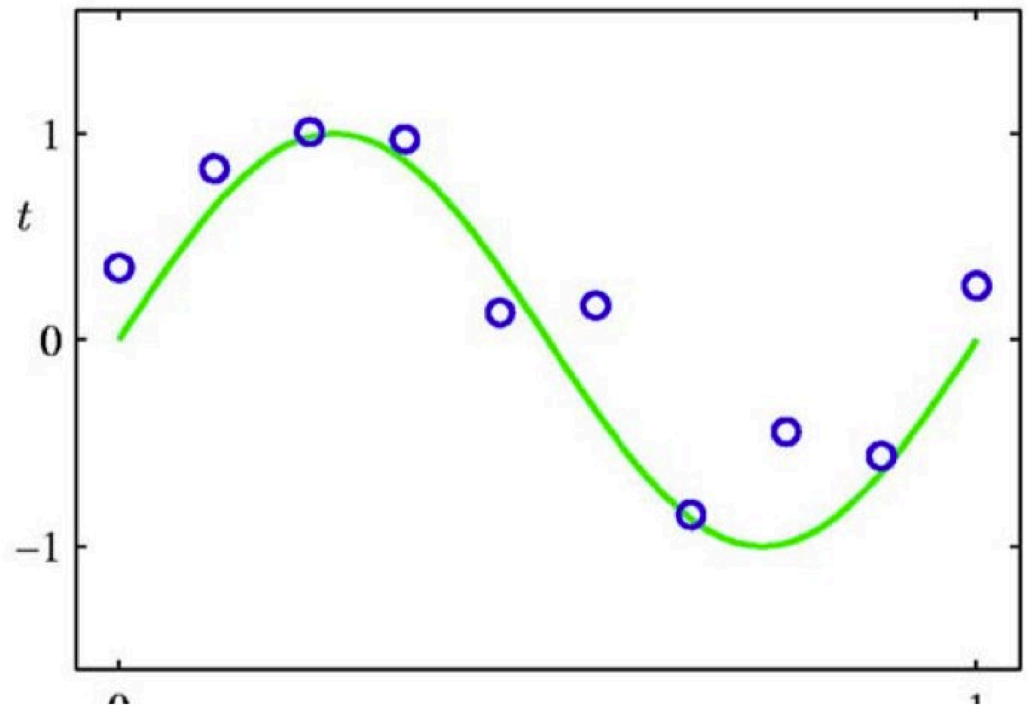


- Suppose we are given a training set of N observations

$$(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N)$$

- Regression problem is to estimate y(x) from this data

Regression: Recap



- Want to fit a polynomial regression model

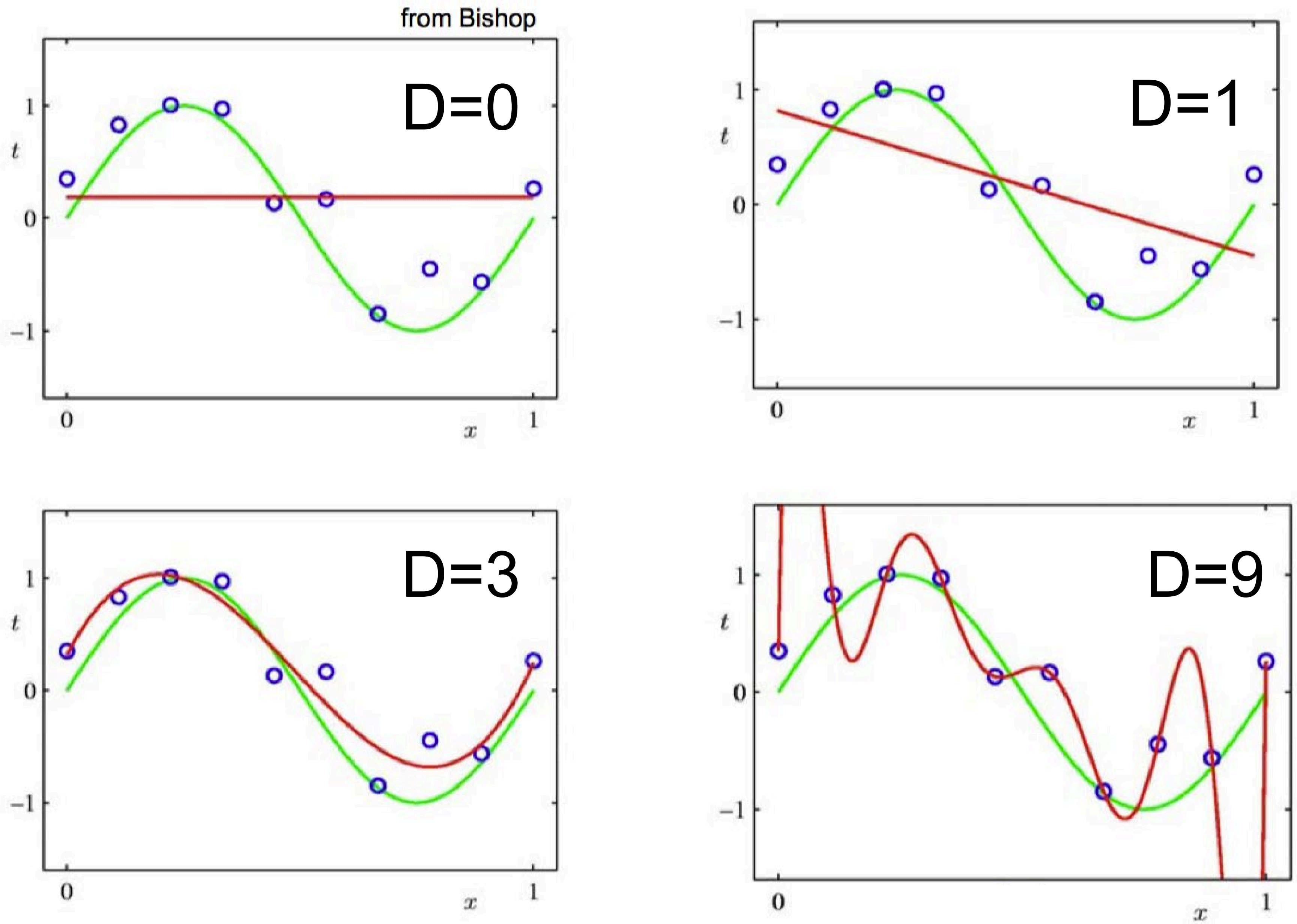
$$\theta_0 + \theta_1 z_1 + \theta_2 z_2 + \theta_d z_d$$

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_d x^d + \epsilon$$

- $z = \{1, x, x^2, \ldots, x^d\} \in R^d$ and $\theta = (\theta_0, \theta_1, \theta_2, \ldots, \theta_d)^T$

$$\boxed{y = z\theta}$$

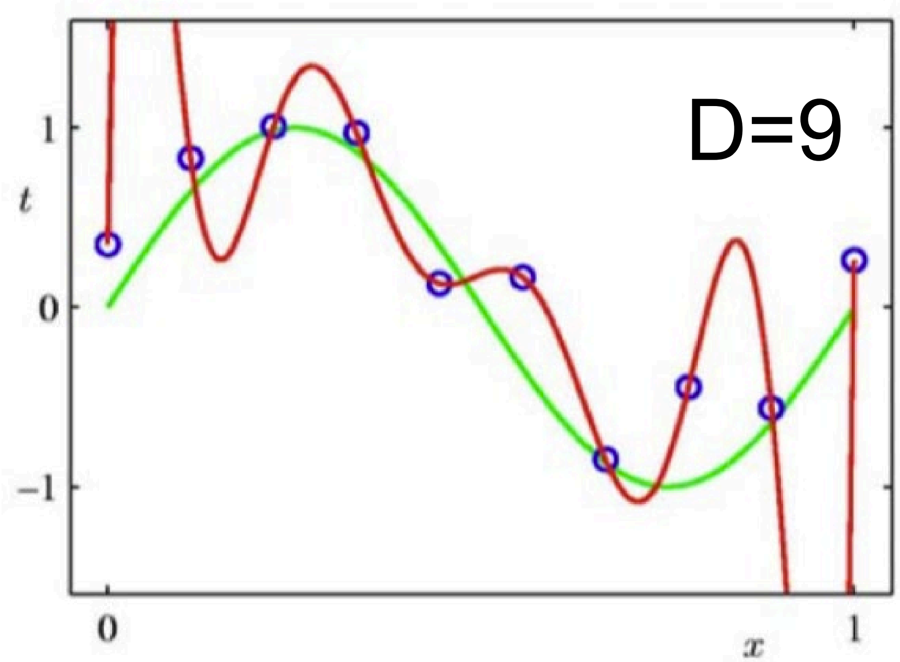
Which One is Better?

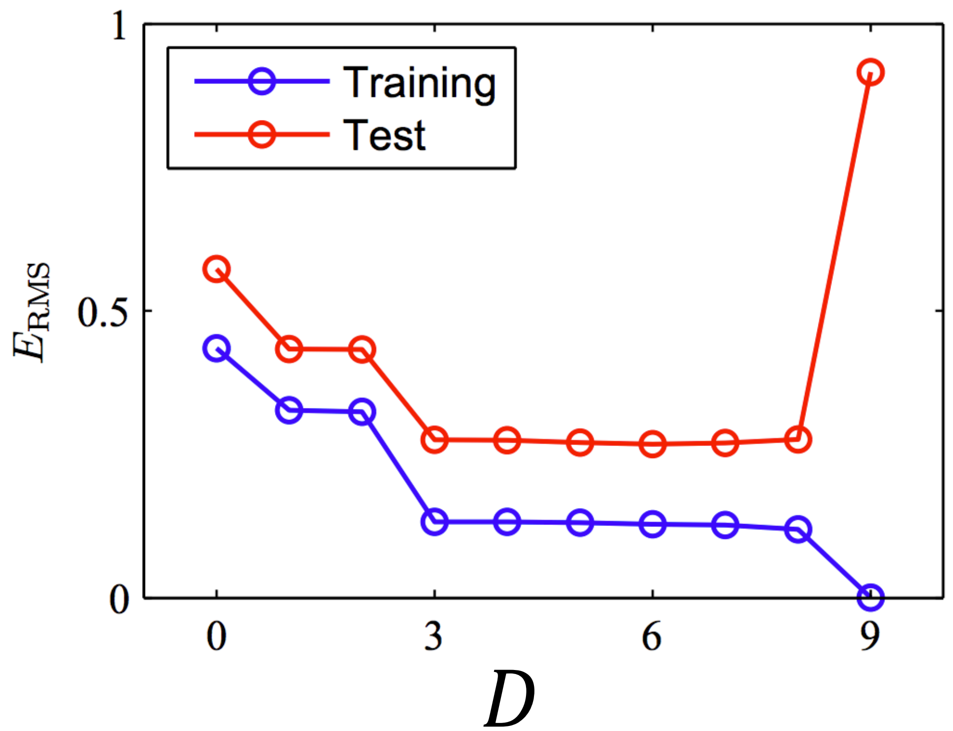


- Can we increase the maximal polynomial degree to very large, such that the curve passes through all training points?

No, this can lead to **overfitting**!

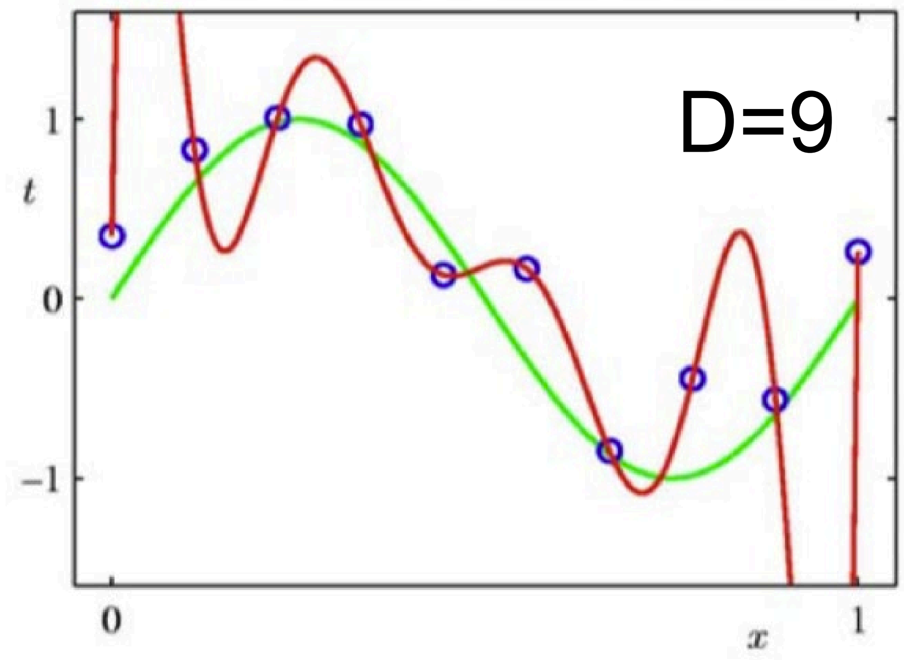
The Overfitting Problem





- The training error is very low, but the error on test set is large.
- The model captures not only patterns but also noisy nuisances in the training data.

The Overfitting Problem

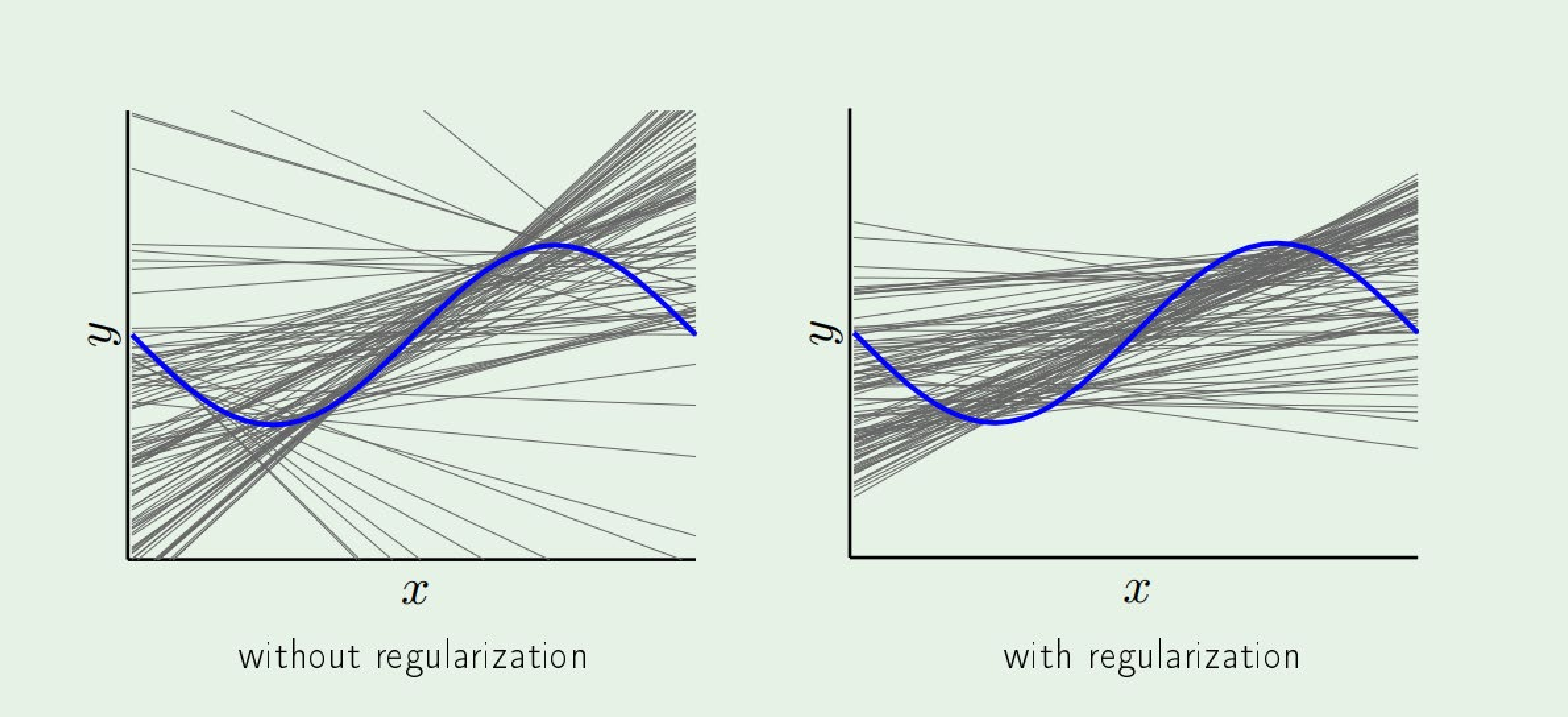


θ

- In regression, overfitting is often associated with large Weights (severe oscillation)
  Parameters
- How can we address overfitting?

Regularization

(smart way to cure overfitting disease)



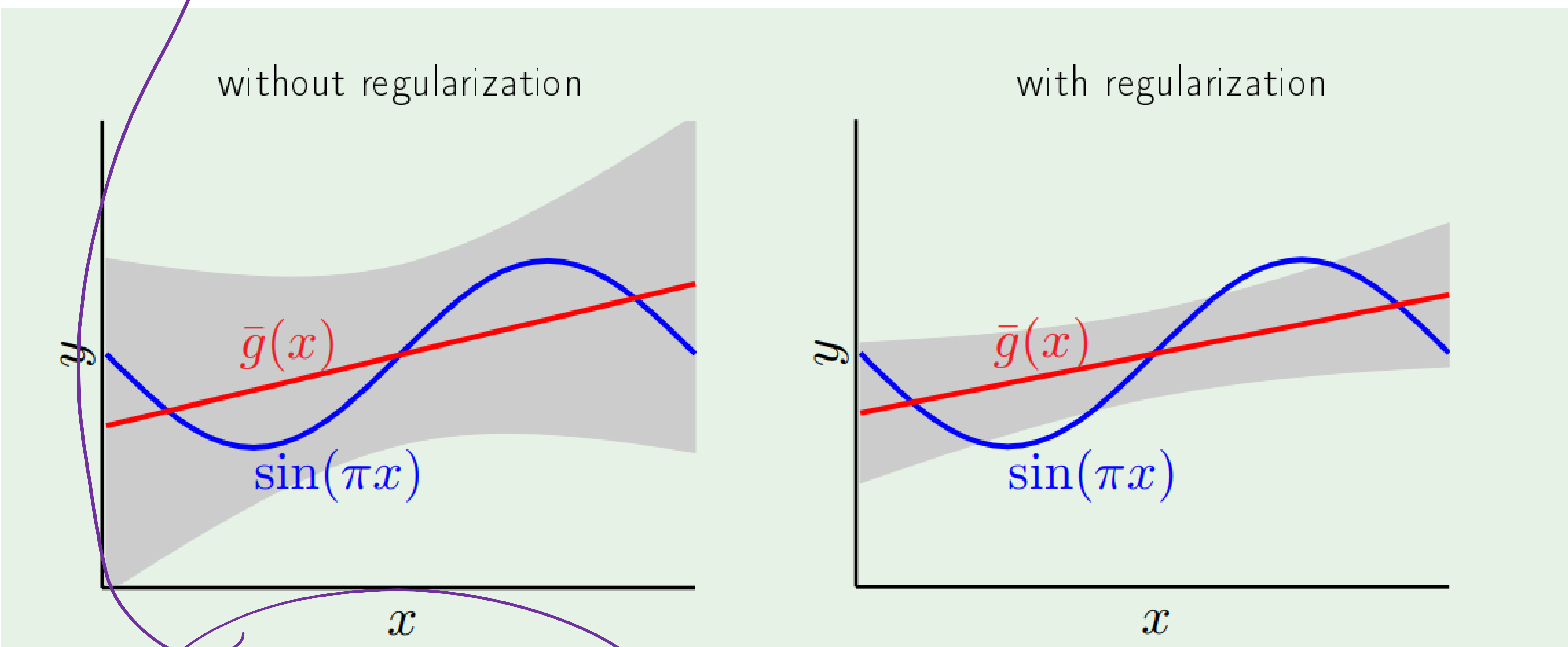
Put a brake on fitting

Fit a linear line on sinusoidal with just two data points

Who is the winner?

$E(\Theta) = L(\Theta) = 0.21^2 + 1.69$

$\bar{g}(x)$: average over all lines



bias=0.21; var=1.69

bias=0.23; var=0.33

Polynomial Model

Want to fit a polynomial regression model

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_d x^d + \epsilon$$

Let's rewrite it as:

$$y = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \cdots + \theta_d z_\mathrm{d} + \epsilon = \boldsymbol{z\theta}$$

Regularizing is just constraining the weights ($\boldsymbol{\theta}$)

For example: let's do a **hard** constraining

$$y = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \cdots + \theta_d z_\mathrm{d}$$

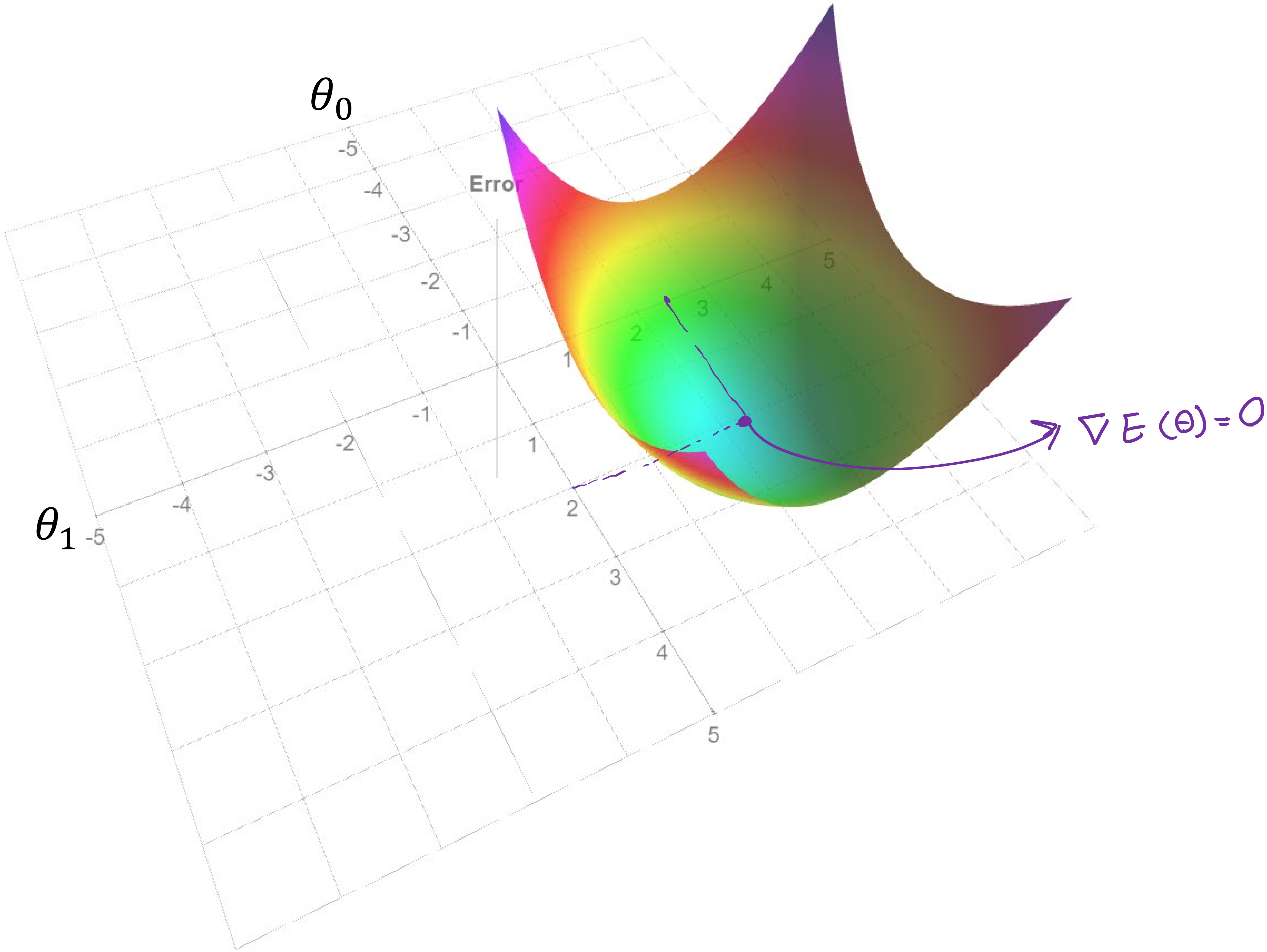
subject to

$$\theta_d = 0 \; for \; d > 2$$

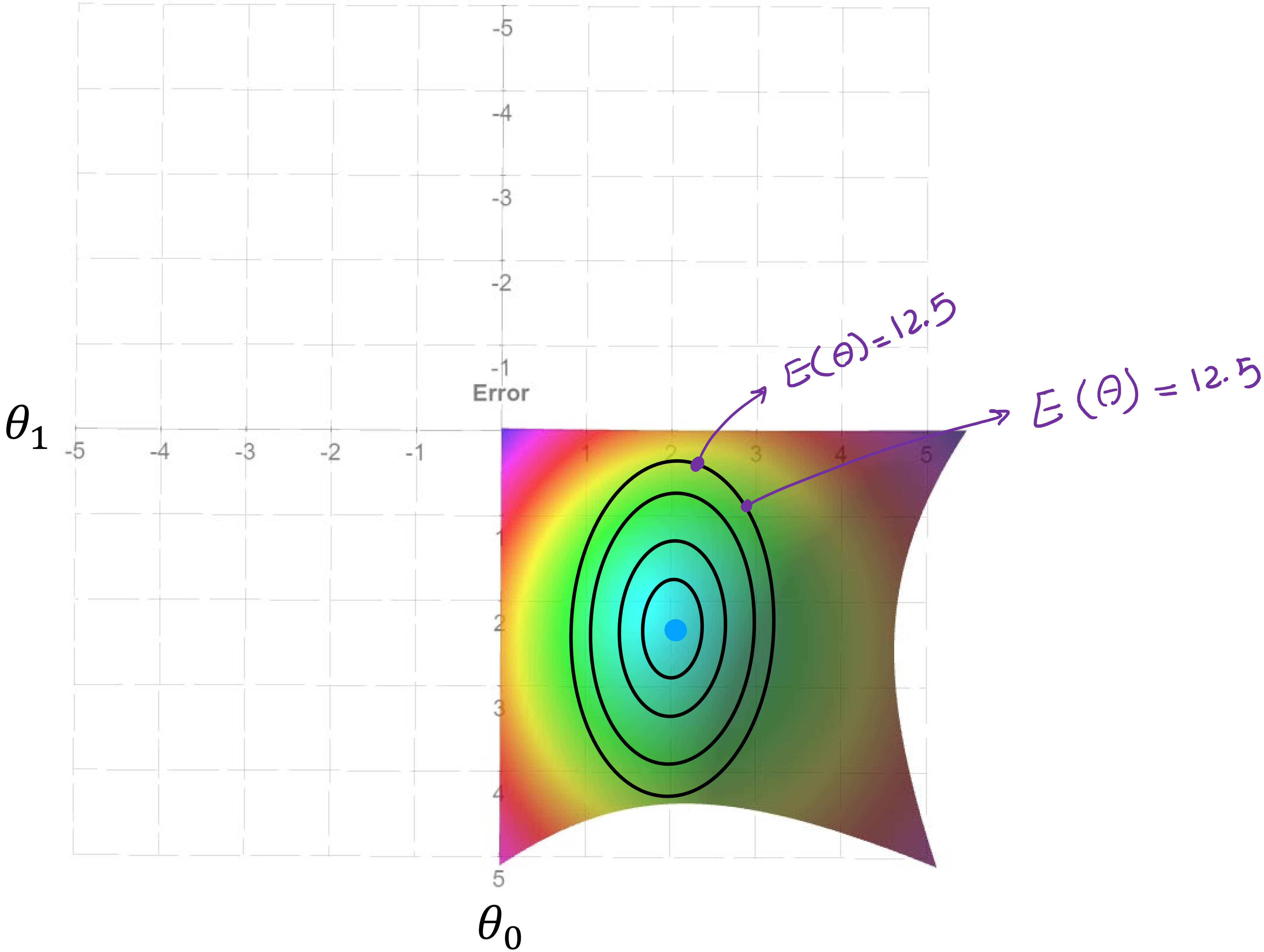
$$y = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + 0 + \cdots + 0$$

$$E(\theta) = \frac{1}{N}\sum_{i=1}^{n}\left(y^{i} - z_{i}\theta\right)^{2}$$

$$\theta = (z^T z)^{-1} z^T y$$



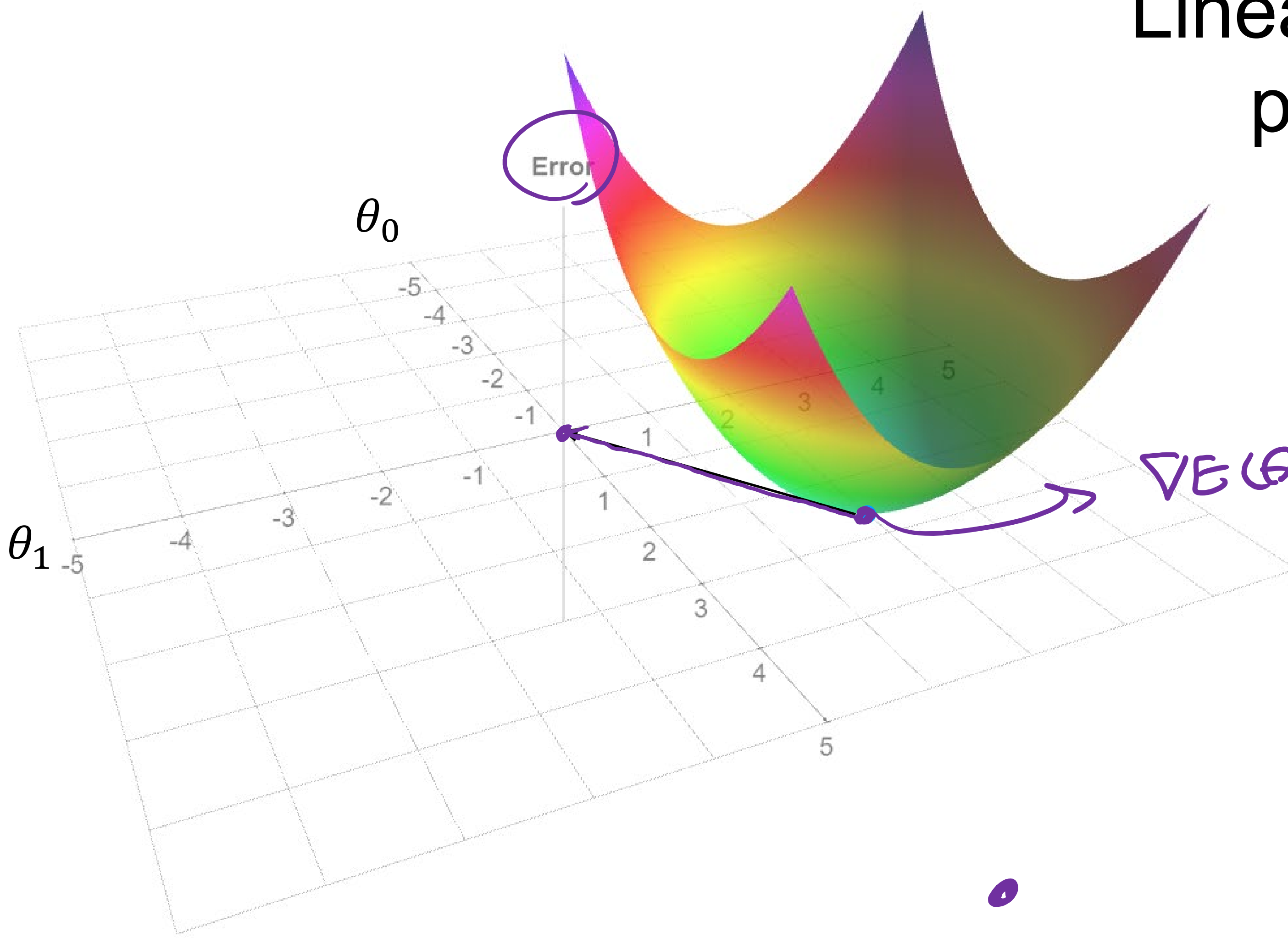
Project the same graph on x-y using contour plot



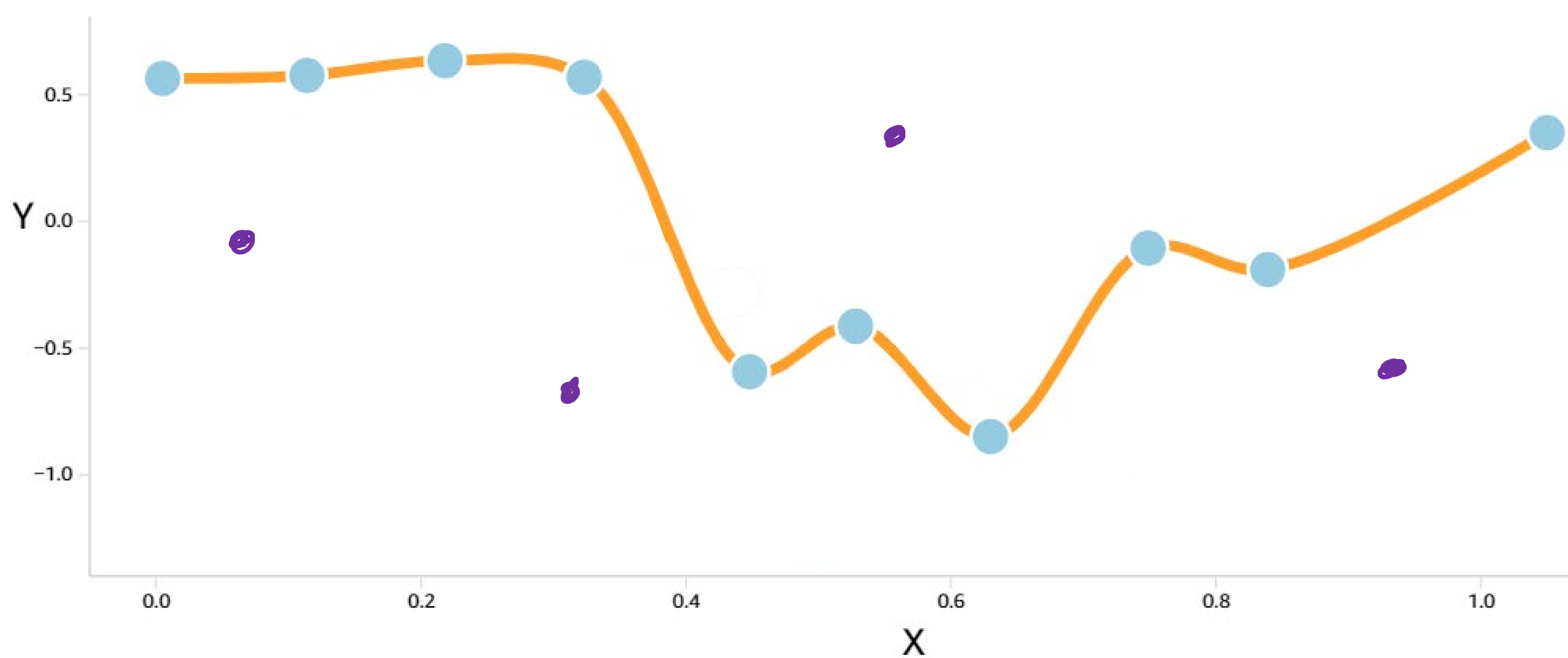
Linear regression with a very high polynomial degree solution

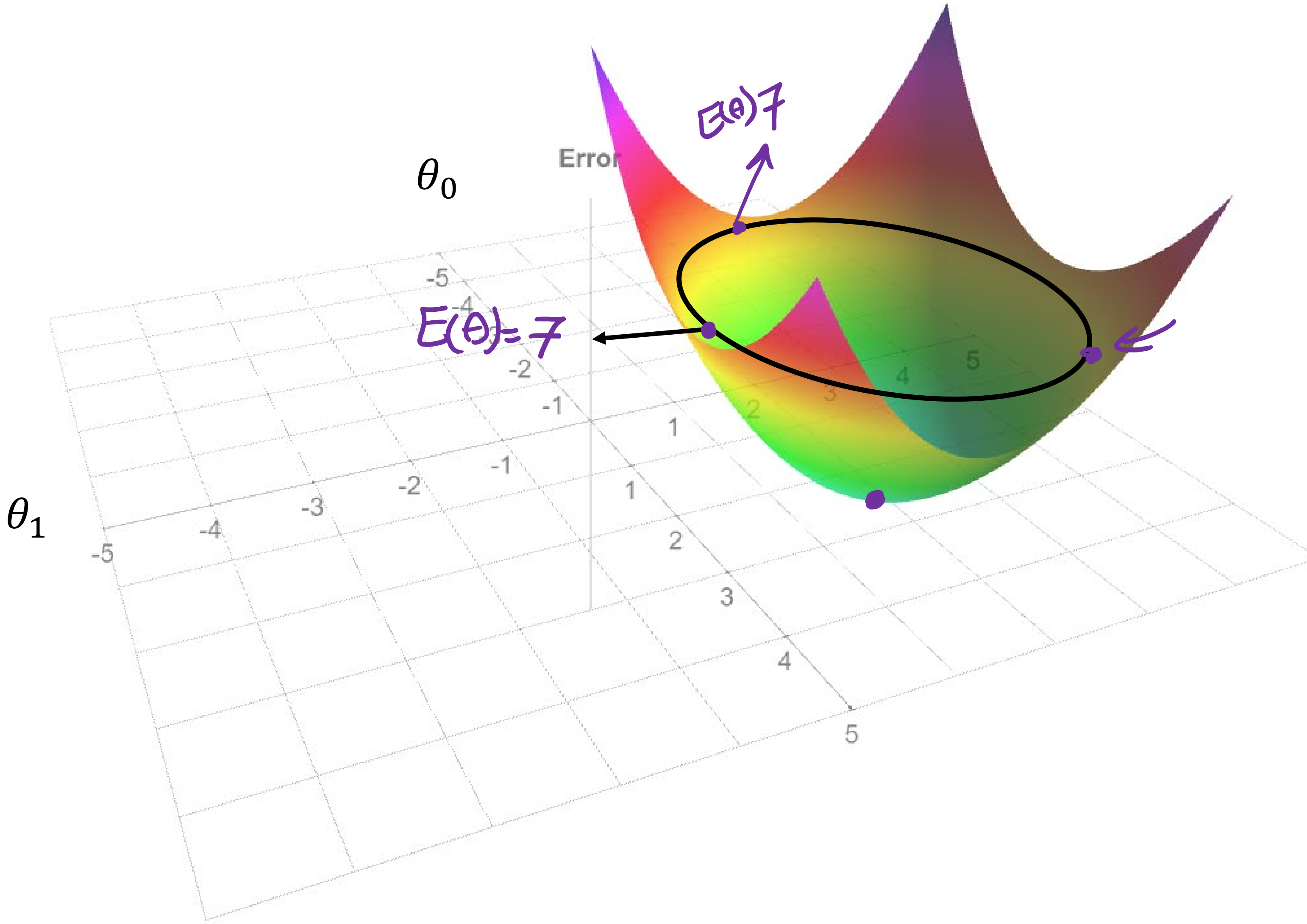
$E(\theta) = 0$

$E(\theta) = \text{bias}^2 + \text{Variance}$

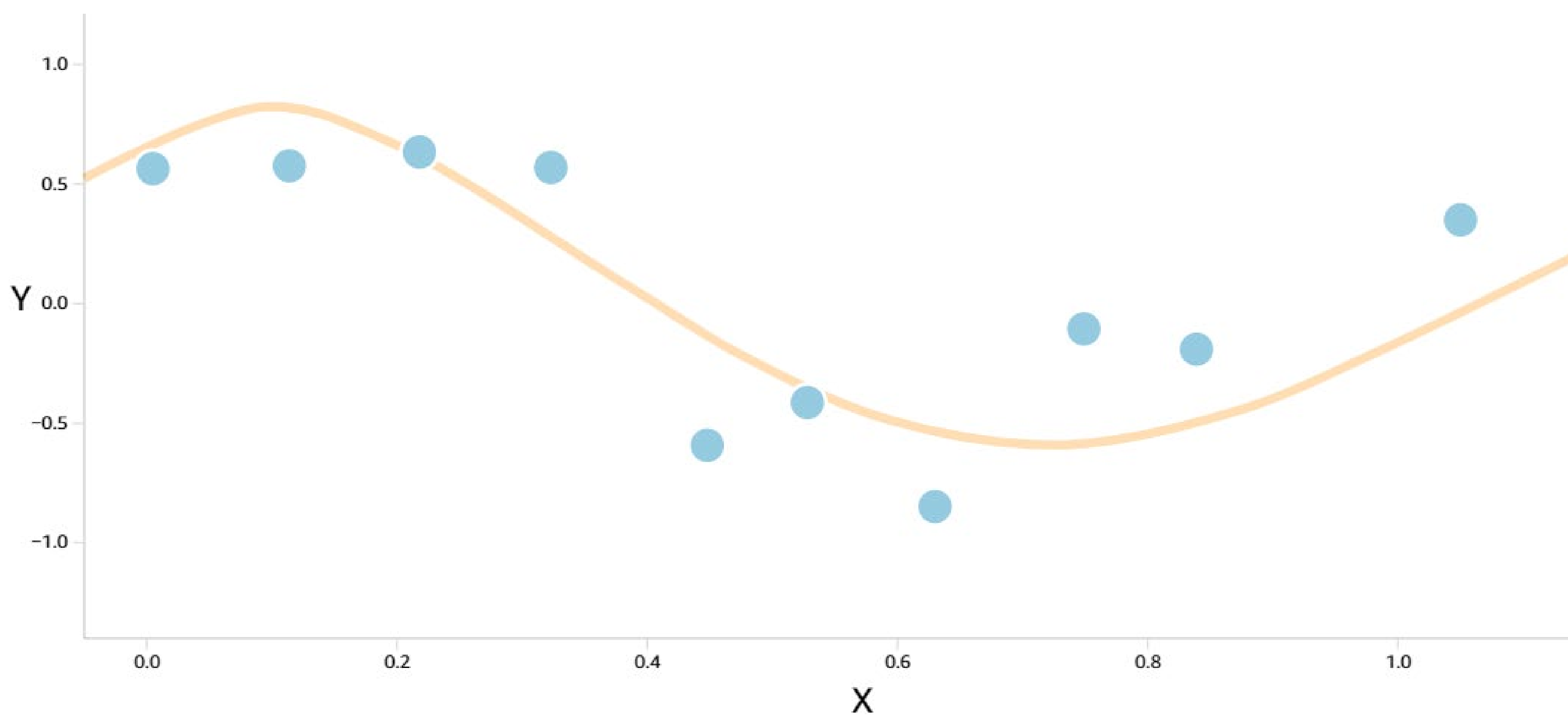


$\nabla E(\theta) = 0$





$$E(\theta) = bias^2 + variance$$



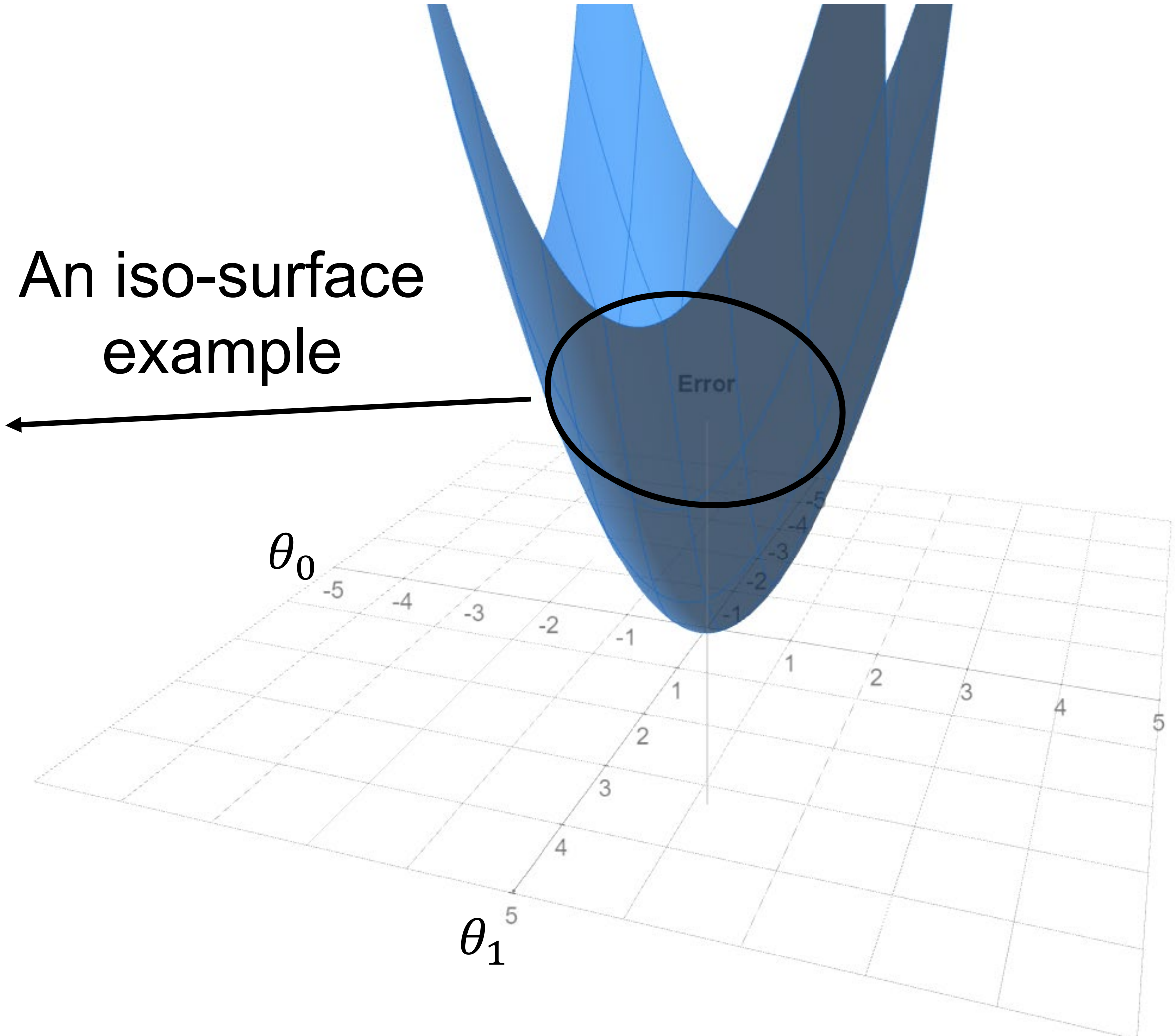
How can we get an optimal solution with a positive error for a model that overfits?

We need to introduce a constraint

$\Theta = \begin{bmatrix} \Theta_0 \\ \Theta_1 \end{bmatrix} \qquad \Theta^T\Theta = \Theta_0^2 + \Theta_1^2$

$$g(\theta) = \theta_0^2 + \theta_1^2 = \theta^T\theta = C$$

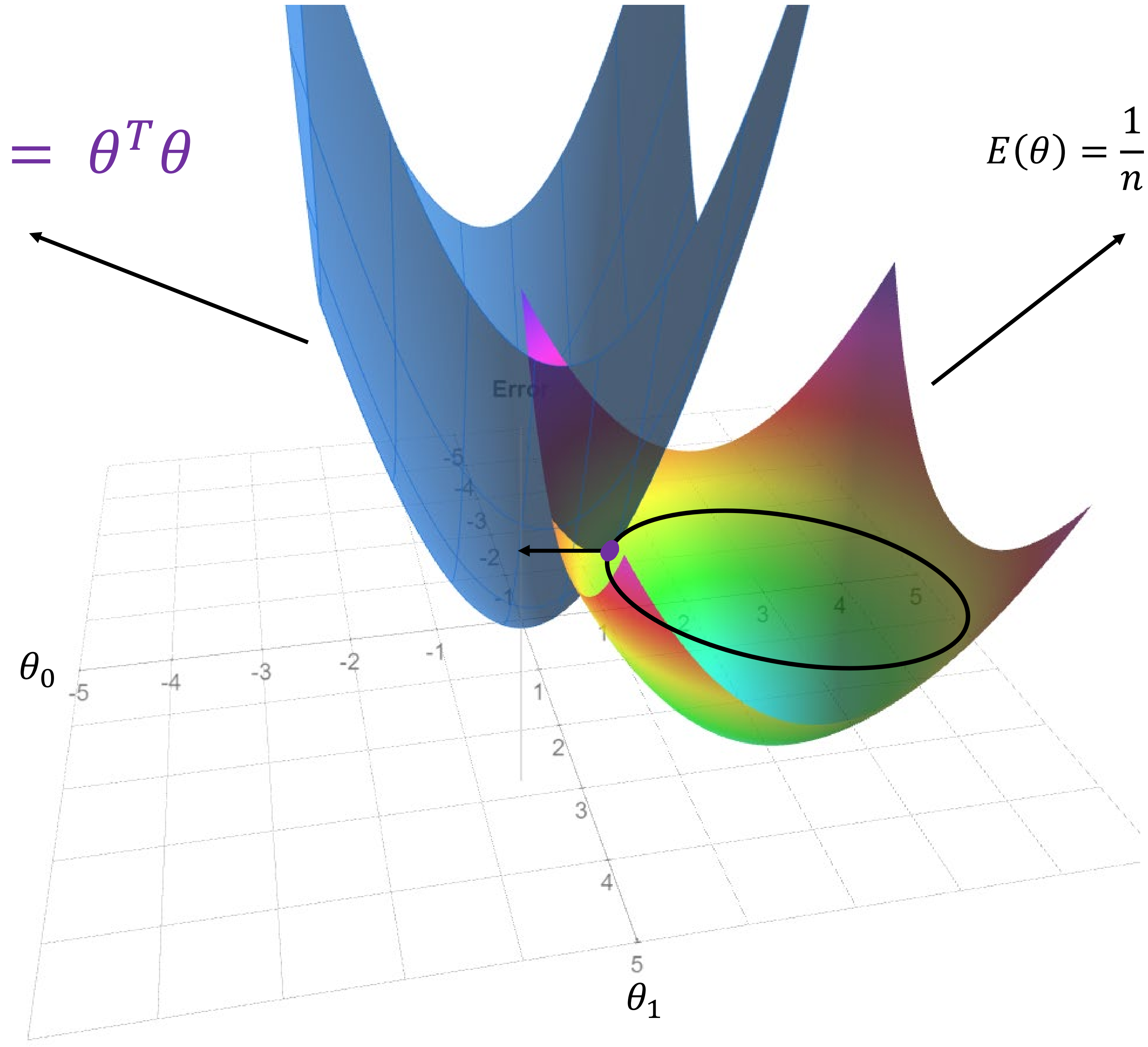
An iso-surface example



Error function together with a new introduced constraint

$$g(\theta) = \theta_0^2 + \theta_1^2 = \theta^T \theta$$

$$E(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y^i - z_i\theta\right)^2$$



Let's define the Lagrange function

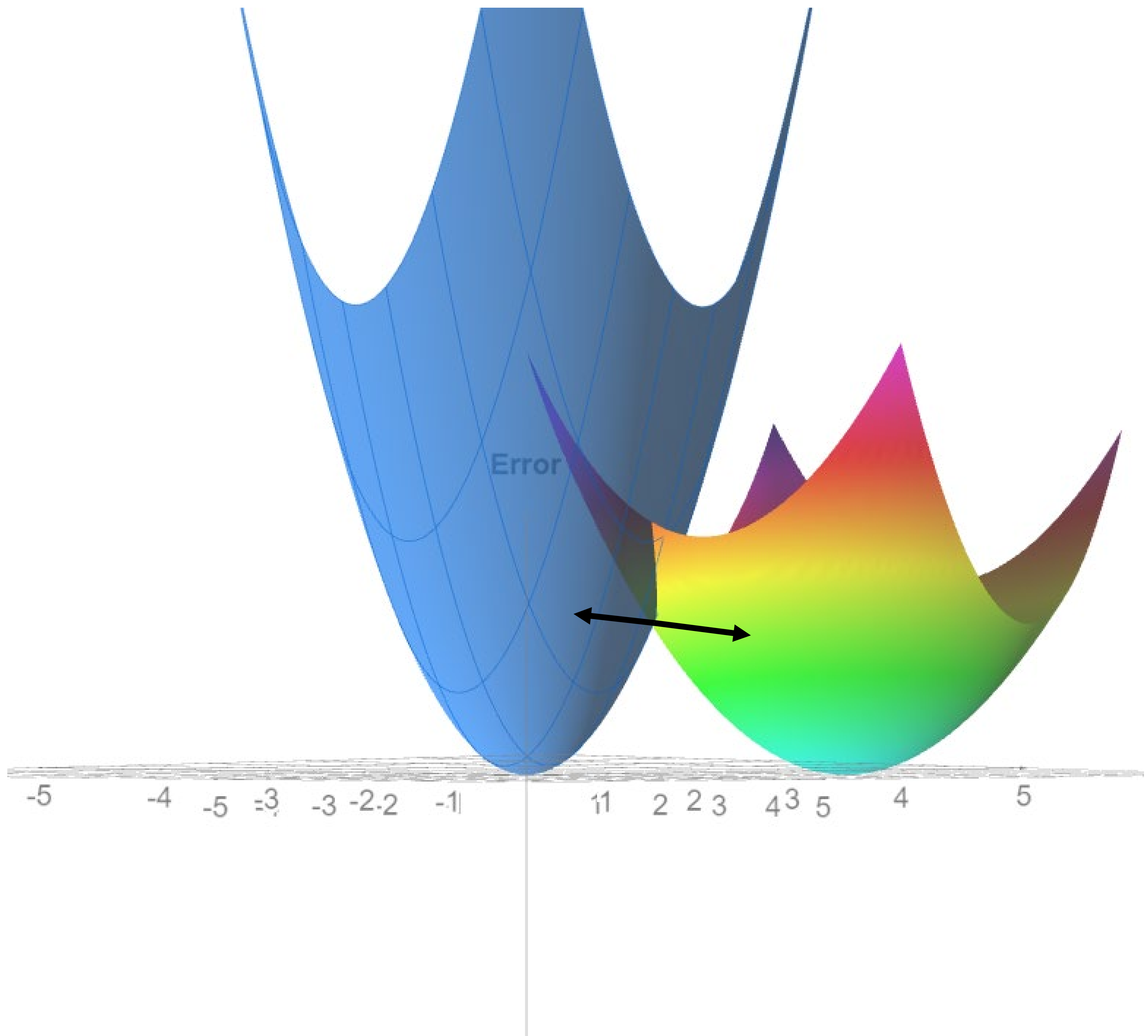
$$L(\theta, \lambda) = E(\theta) + \lambda g(\theta)$$

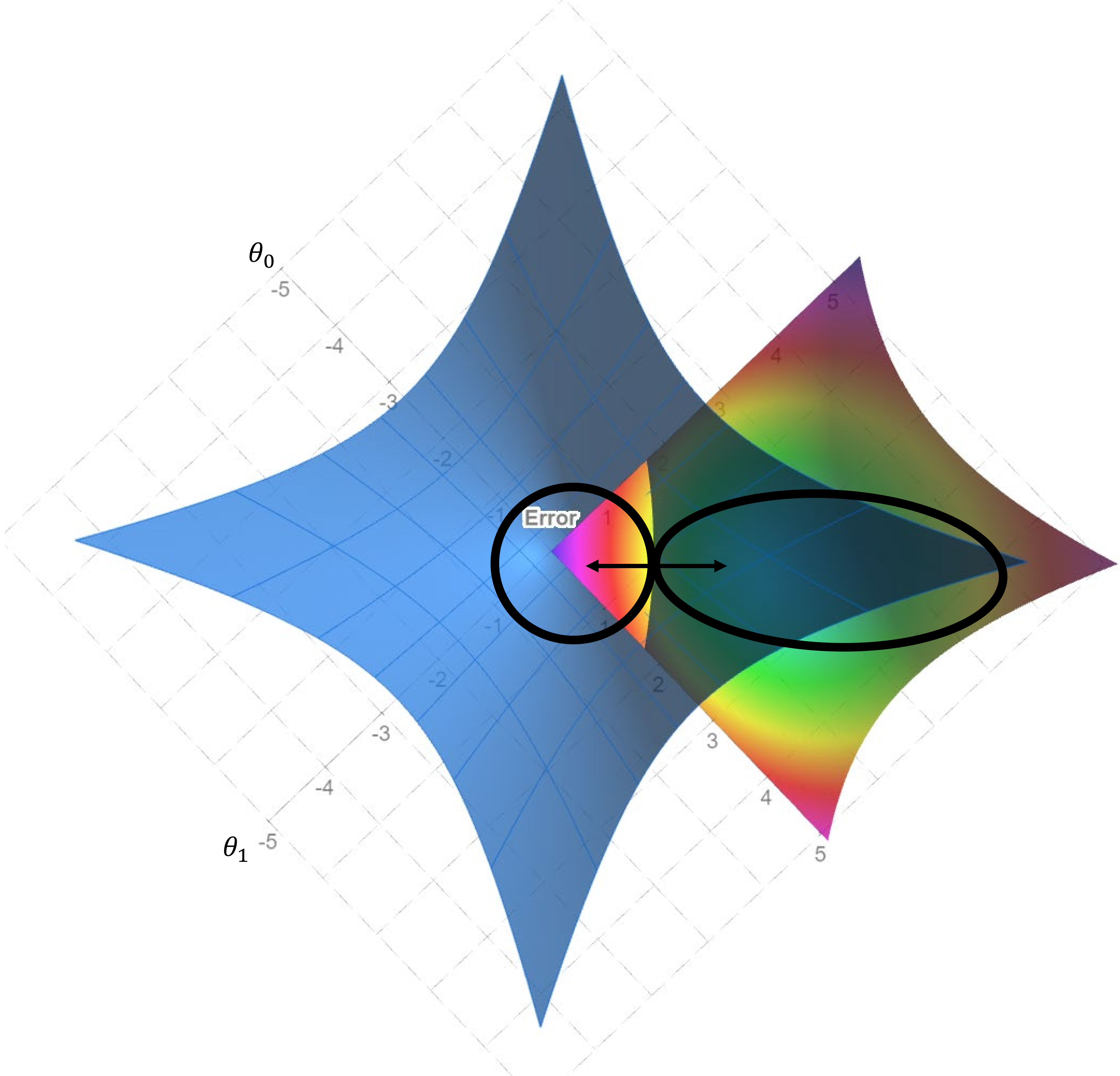
$$L(\theta, \lambda) = E(\theta) + \lambda \theta^T \theta$$

$$\nabla L(\theta, \lambda) = 0 \qquad \nabla[E(\theta) + \lambda \theta^T \theta] = 0$$

$$\nabla[E(\theta)] + \lambda \nabla[\theta^T \theta] = 0$$

How to enforce the gradient of Lagrange function to be zero





θ_0
θ_1
Error
-5
-4
-3
-2
-1
1
2
3
4
5

Let's calculate the gradients

Gradient of constraint $g(\theta)$

$$\nabla[\theta^T\theta] = 2\theta$$

$$\nabla E(\theta) \sim -\nabla[\theta^T\theta]$$

$$\nabla[E(\theta)] + \lambda\nabla[\theta^T\theta] = 0$$

$$\nabla[E(\theta)] = -\lambda\nabla[\theta^T\theta]$$

$$\nabla E(\theta) = -2\lambda\theta$$

$E(\theta)$ s.t. $\theta^T\theta = c$ $\longleftrightarrow$ $E(\theta) + \delta\,\theta^T\theta$

$c\uparrow$ $\delta\downarrow$

$$\nabla E(\theta) + 2\lambda\theta = 0$$

Let's do integration

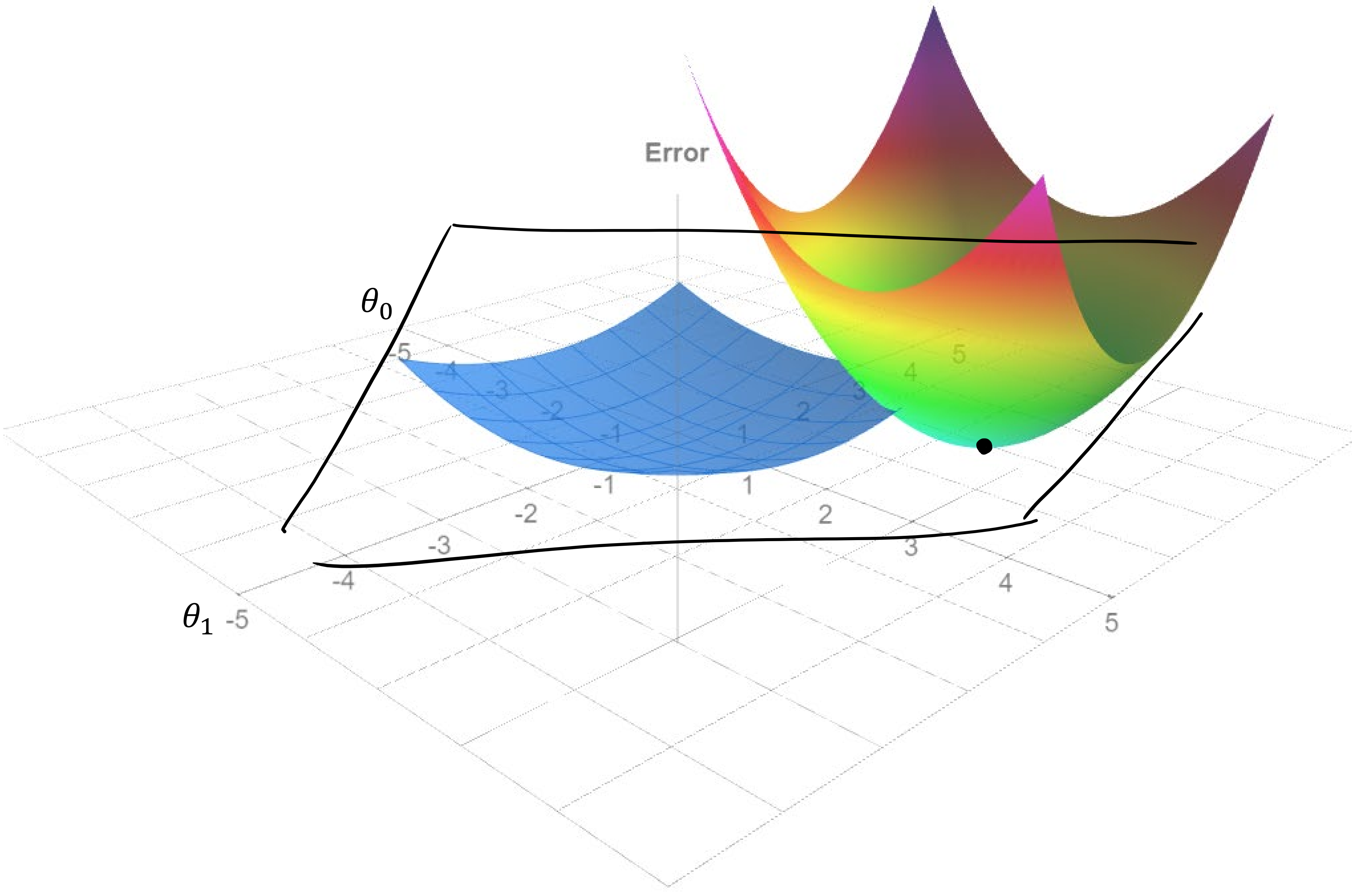
minimize

Regularization term

$$\boxed{E(\theta) + \lambda\theta^T\theta}$$

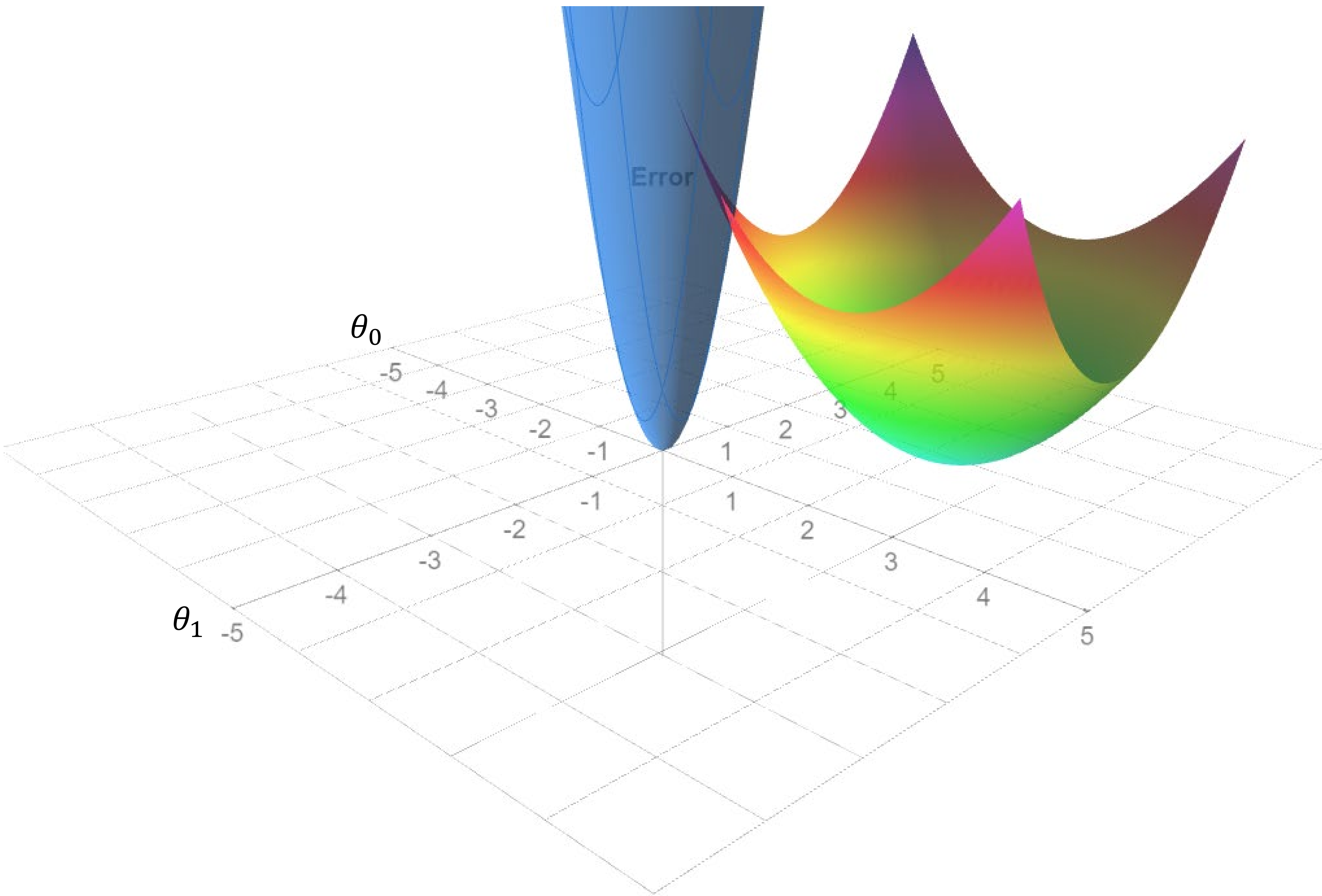
The effect of low Lambda

$$E(\theta) + \frac{\lambda}{N}\theta^T\theta$$



The effect of high Lambda

$$E(\theta) + \frac{\lambda}{N}\theta^T\theta$$



Regularized Learning

Now we know Why this term leads to the regularization of parameters

Minimize $$E(\theta) + \lambda\theta^T\theta$$

Regularized Error

$$\tilde{E}(\theta) = \frac{1}{N}\sum_{i=1}^{n}\left(y^i - z_i\theta\right)^2 + \frac{\lambda}{2N}\|\theta\|_2^2$$

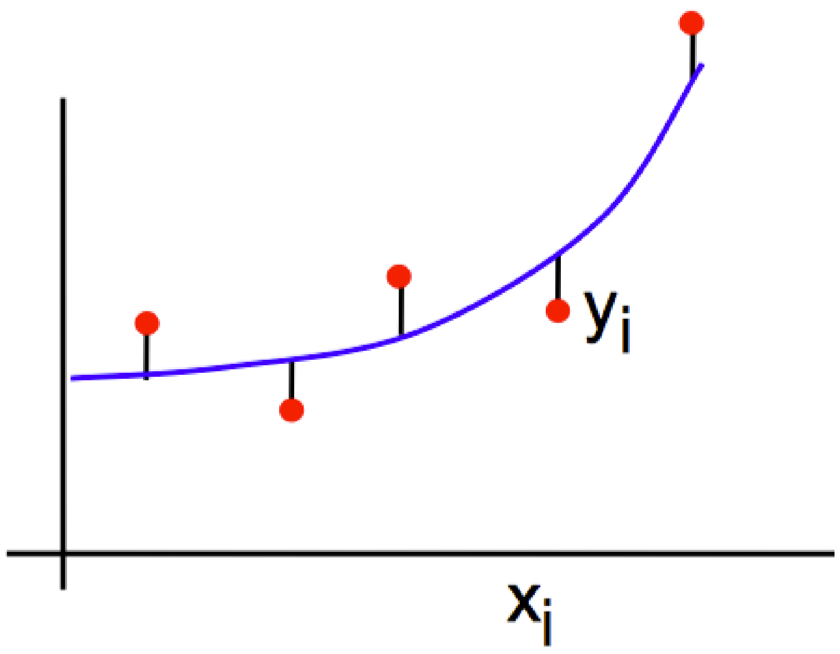
L2 Regularization term

Outline

- Overfitting and regularized learning
- Ridge regression
- Lasso regression
- Determining regularization strength

Ridge Regression

$$\tilde{E}(\theta) = \frac{1}{N}\sum_{i=1}^{n}\left(y^{i} - z_i\theta\right)^2 + \lambda\|\theta\|_2^2$$



$$\theta_0 + \theta_1 z_1 + \theta_2 z_2 + \cdots + \theta_d z_d + \epsilon = \boldsymbol{z\theta}$$

General form

$$\tilde{E}(\theta) = \frac{1}{N}\sum_{i=1}^{n}\left(y^{i} - z_i\theta\right)^2 + \lambda\|\theta\|_2^2$$

Matrix form

$$\tilde{E}(\theta) = \frac{1}{N}(y - z\theta)^{\mathrm{T}}(\mathrm{y} - z\theta) + \frac{\lambda}{2}\|\theta\|_2^2$$

$$\frac{\partial \tilde{E}(\theta)}{\partial \theta} = -z^T(y - z\theta) + \lambda\theta$$

$$(z^T z + \lambda I)\theta = z^T y$$

$$\boxed{\theta_{reg} = (z^T z + \lambda I)^{-1} z^T y}$$

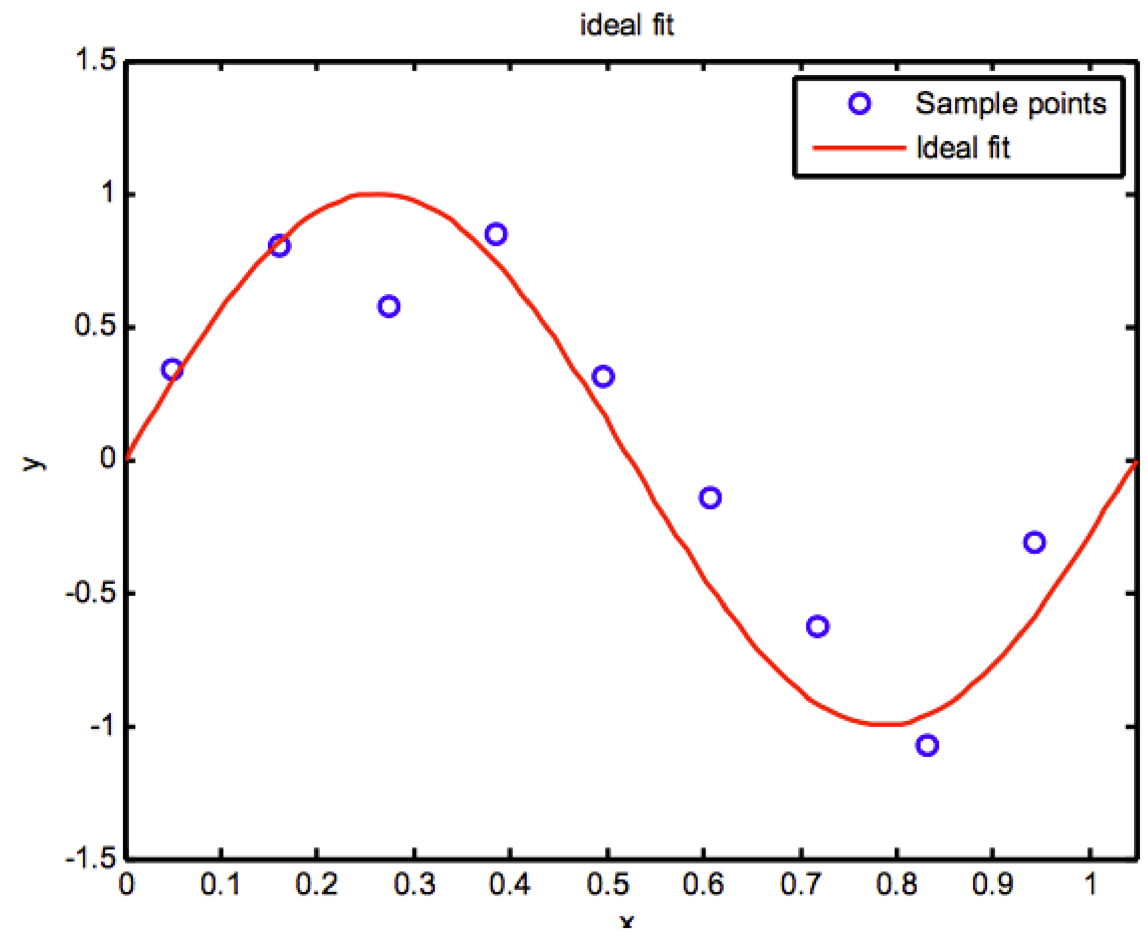
$$\theta_{overfitted} = (z^T z)^{-1} z^T y$$

Ridge Regression Example

- The red curve is the true function (which is not a polynomial)

- The data points are samples from the curve with added noise in y.

- There is a choice in both the degree, D, of the basis functions used, and in the strength of the regularization

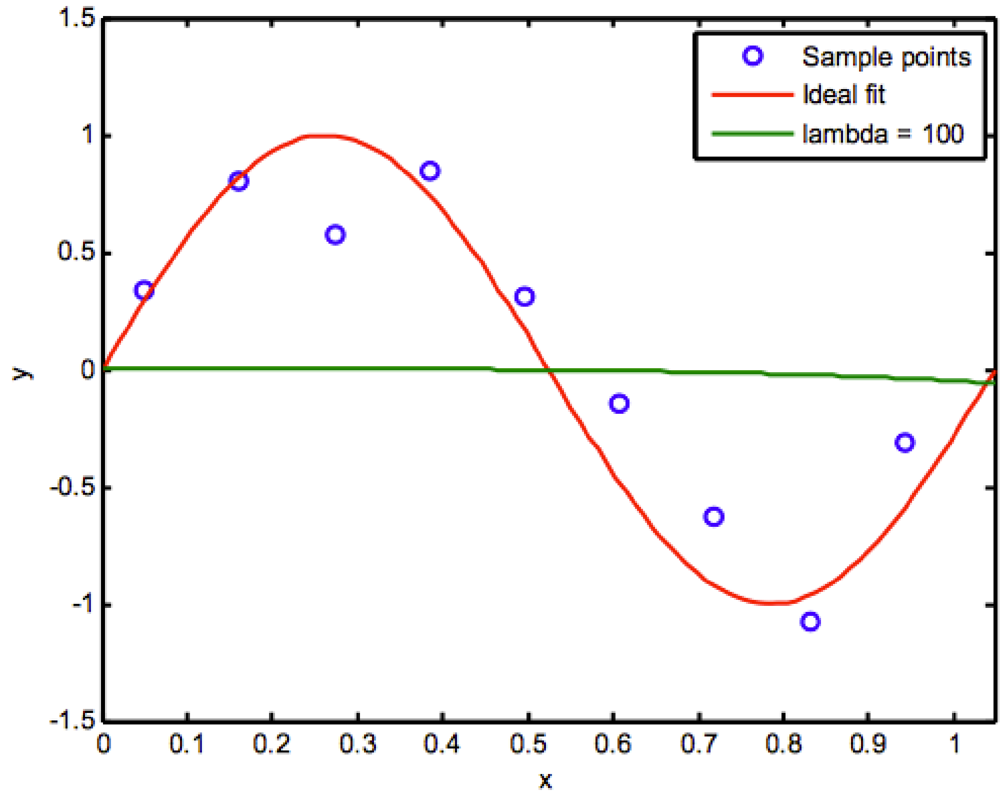


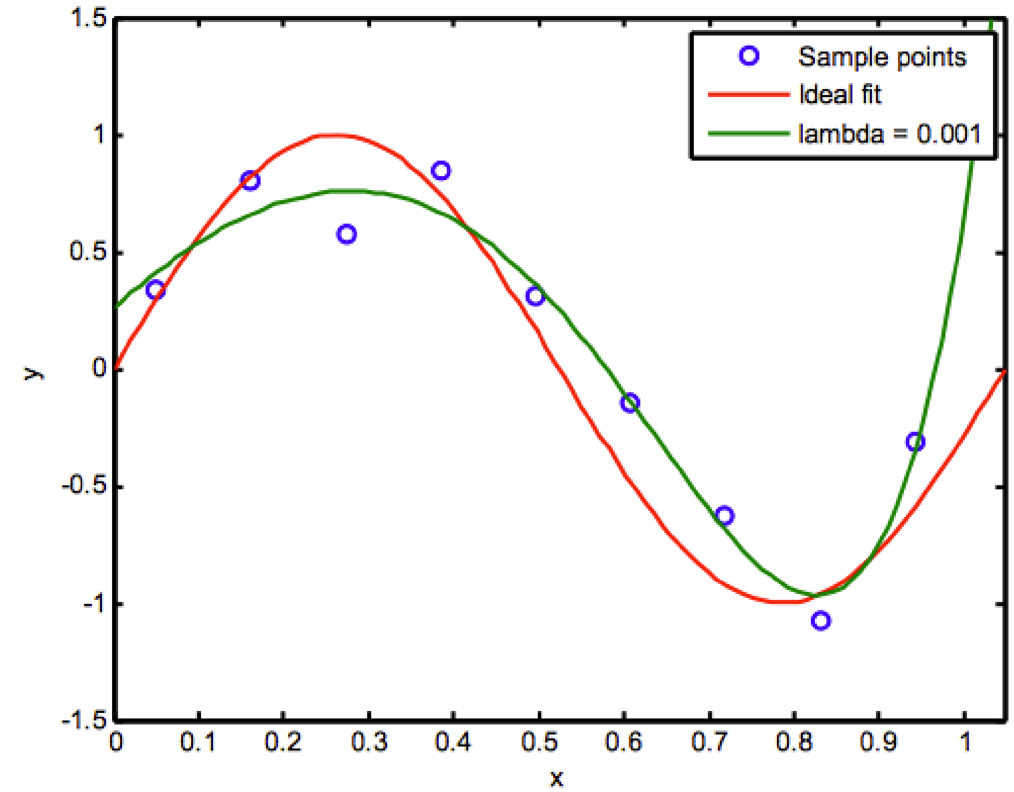
$$f(x,\theta) = z\theta \qquad z: x \rightarrow z$$

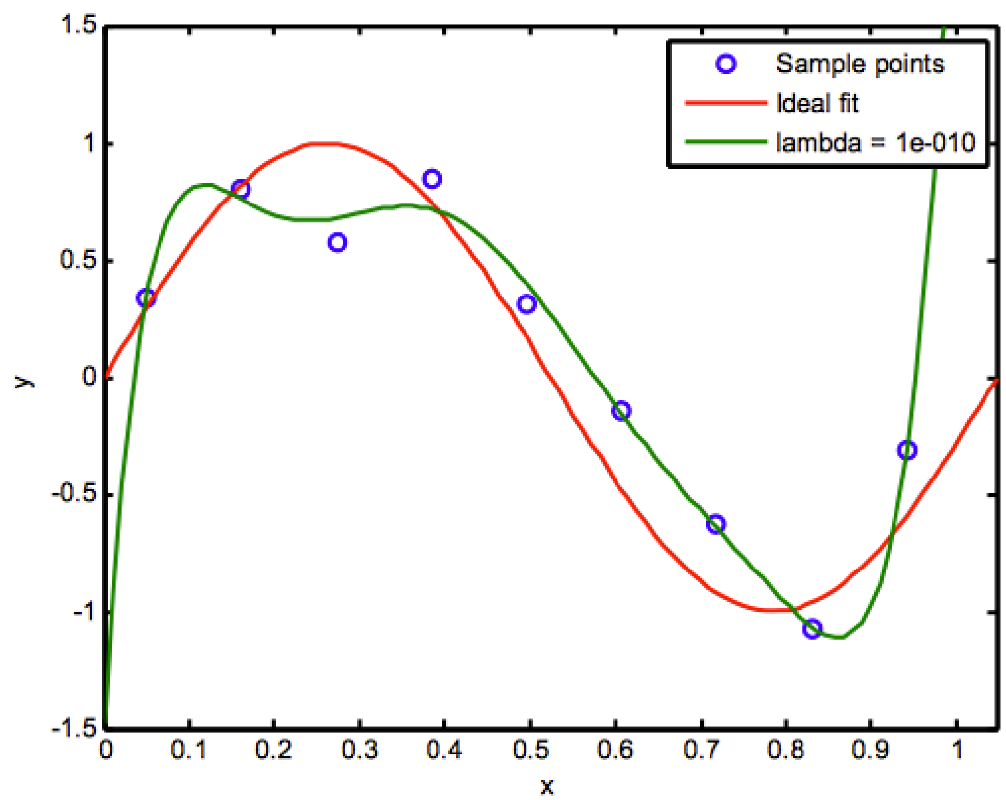
$$\tilde{E}(\theta) = \frac{1}{N}\sum_{i=1}^{n}\left(y^i - z_i\theta\right)^2 + \lambda\|\theta\|_2^2 \qquad \theta \in \mathbb{R}^{D+1}$$

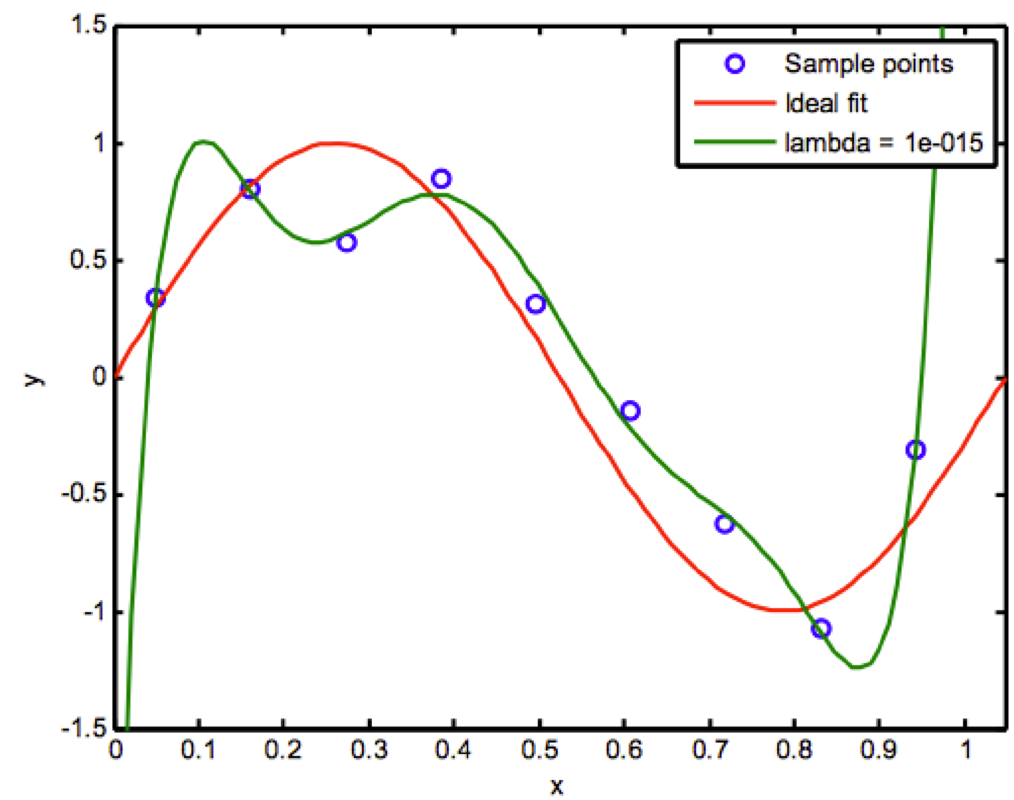
$\frac{\delta}{2N}\|\theta\|_2^2$

N = 9 samples, D = 7

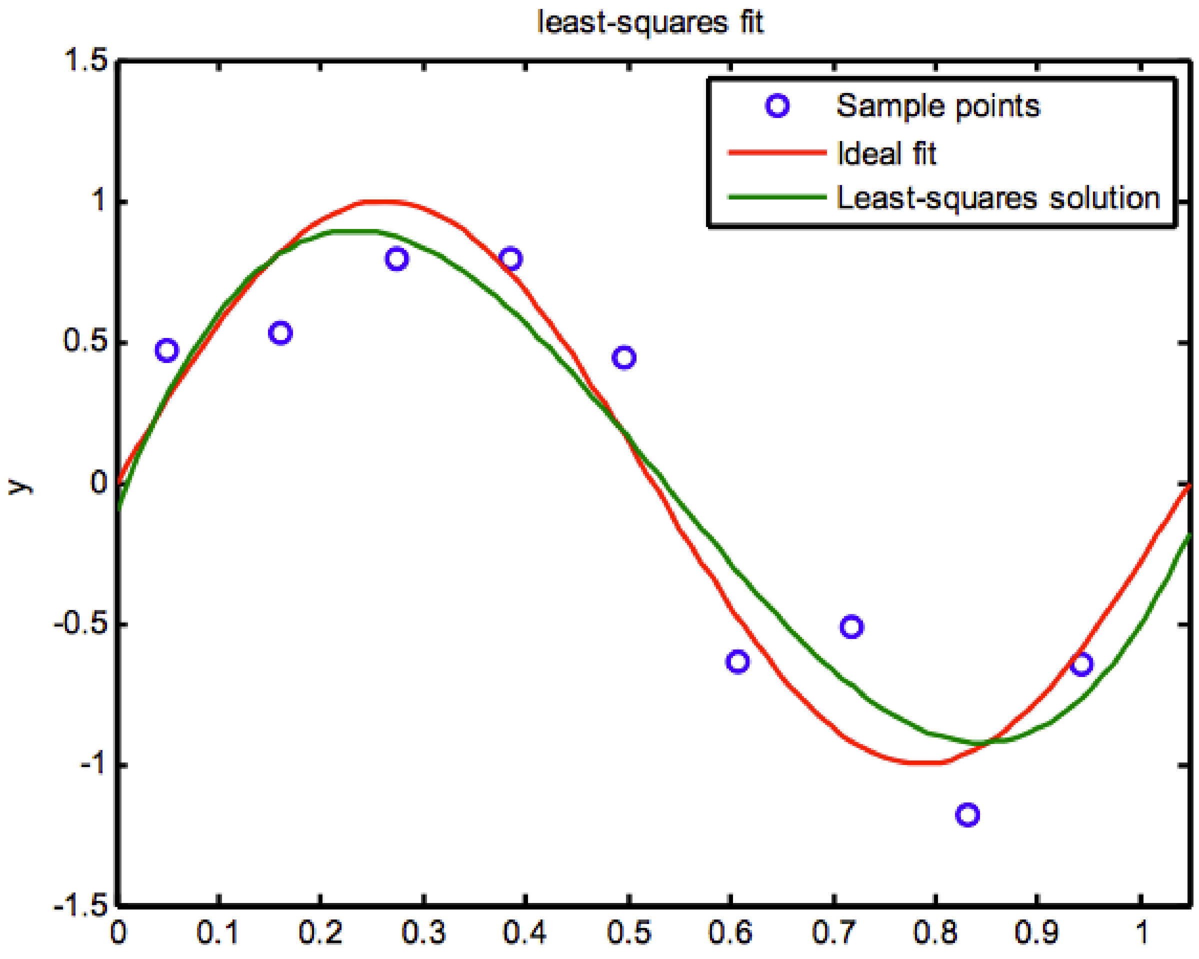




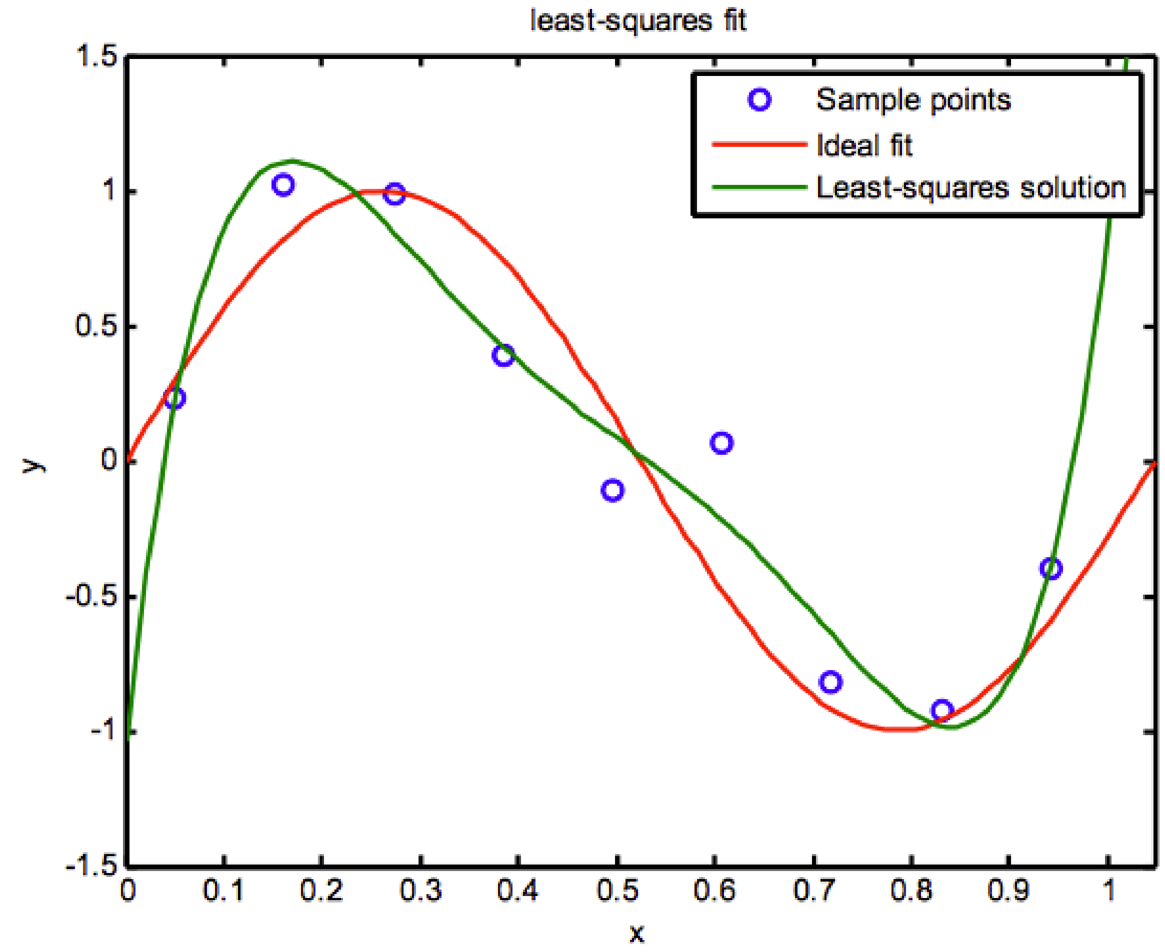




D = 3


least-squares fit
Sample points
Ideal fit
Least-squares solution
y
1.5
1
0.5
0
-0.5
-1
-1.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

D = 5


least-squares fit
Sample points
Ideal fit
Least-squares solution
y
x
1.5
1
0.5
0
-0.5
-1
-1.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Outline

- Overfitting and regularized learning
- Ridge regression
- Lasso regression
- Determining regularization strength

Regularized Regression

$$\tilde{E}(\theta) = \frac{1}{N}\sum_{i=1}^{n}\left(y^i - z_i\theta\right)^2 + \lambda\|\theta\|_2^2$$

Squared loss\Error $\frac{1}{N}\sum_{i=1}^{n}\left(y^i - z_i\theta\right)^2$

L2 Regularizer $\lambda\|\theta\|_2^2$

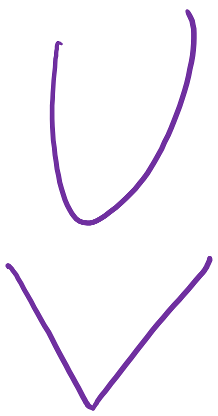
Now let's look at another regularization choice.

The Lasso Regularization (L1 norm) and sparsity

Lasso = **L**east Absolute **S**hrinkage and **S**election **O**perator

$\|\theta\|_2^2$

$\|\theta\|_1$

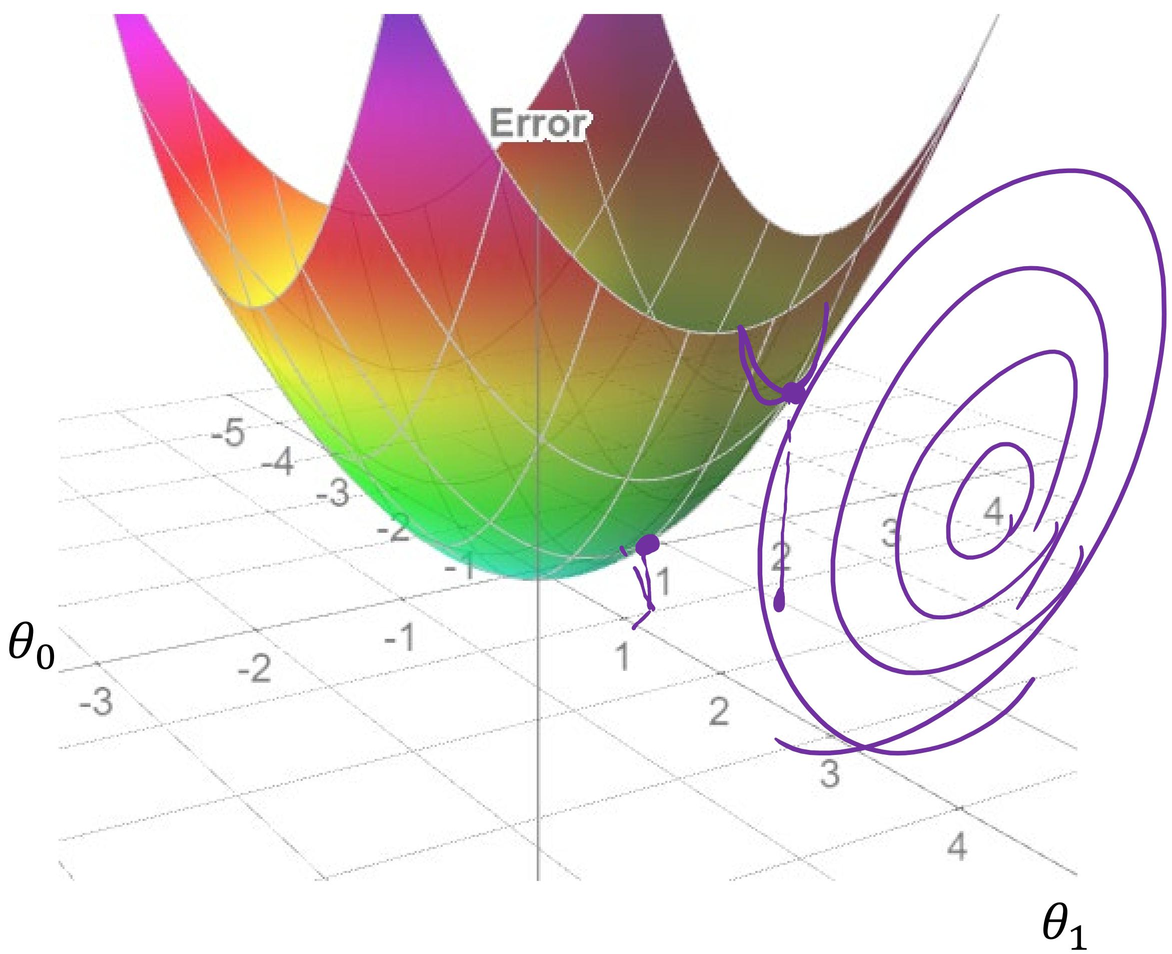


$$\tilde{E}(\theta) = \frac{1}{N}\sum_{i=1}^{n}\left(y^i - z_i\theta\right)^2 + \lambda\|\theta\|_1$$

L1 norm induces sparsity. This means that some of the weights become zero, and the feature contribution will be completely removed. L1 Regularizer could be used for feature selection

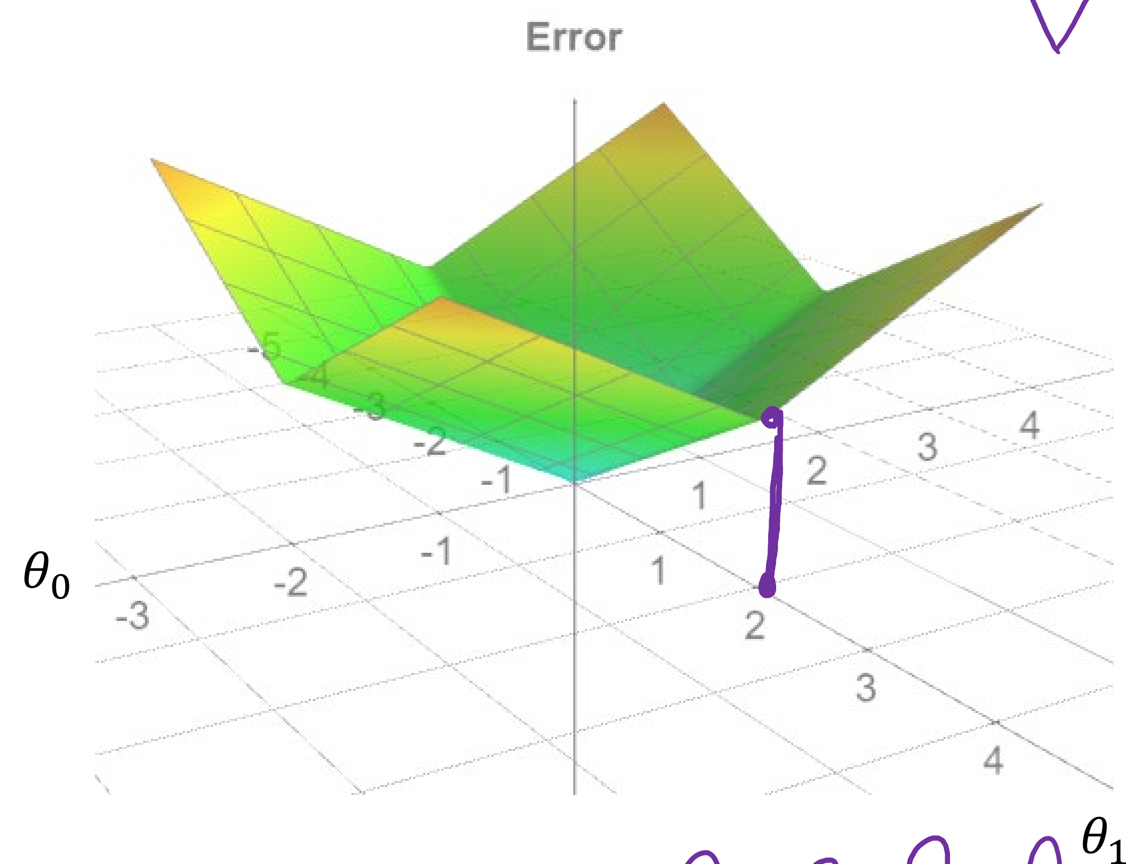
Ridge Regularizer

$$g(\theta) = \theta_0^2 + \theta_1^2 = \theta^T\theta$$



Lasso Regularizer

$$g(\theta) = \theta_0 + \theta_1 = \theta$$



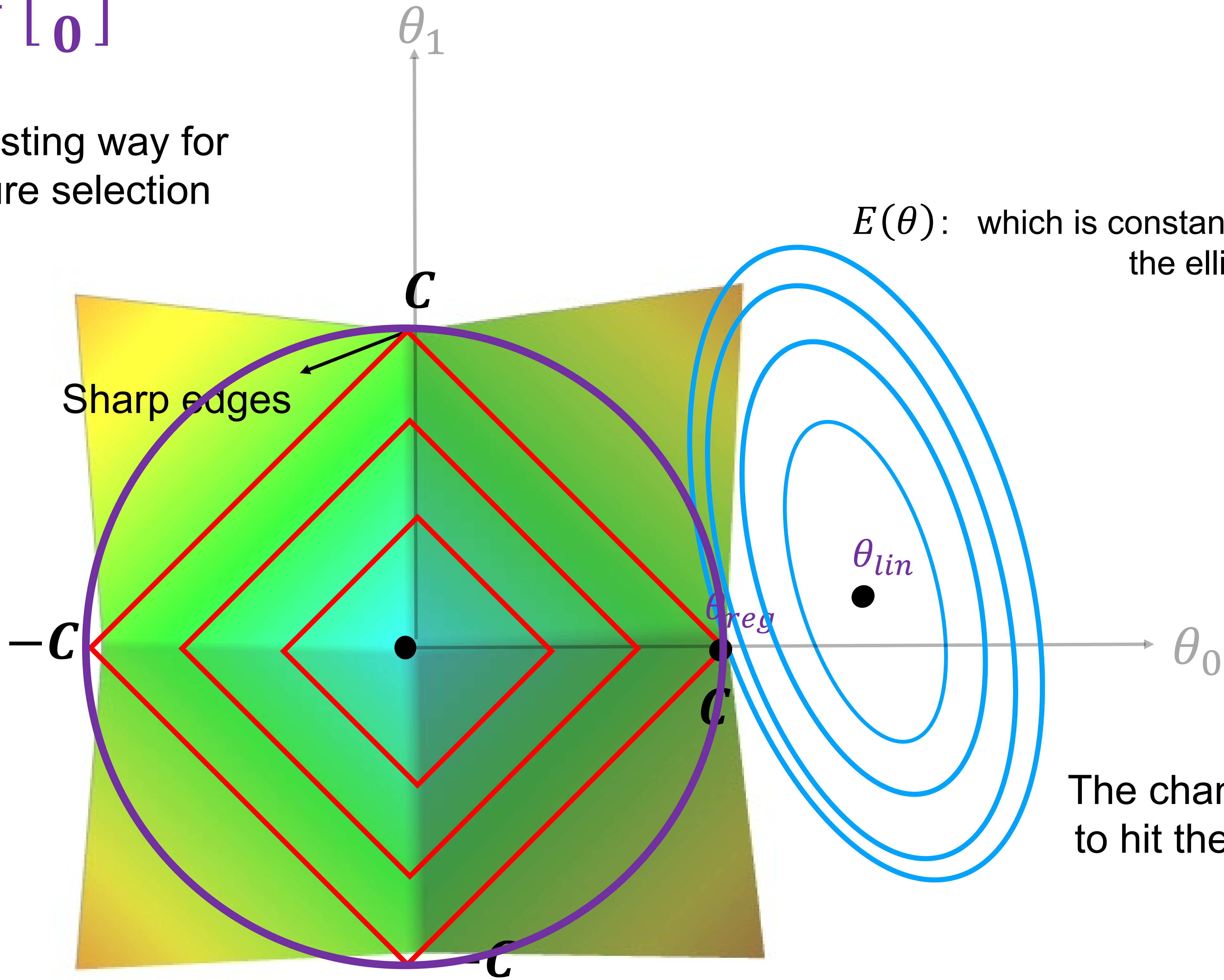
$\theta_1 = 2 \quad \theta_0 = 0$

Let's say we have two parameters (θ_0 and θ_1)

$$Min\ E(\theta) = \frac{1}{N}(zw - y)^T(z\theta - y) + \lambda\|\theta\|_1$$

$$\theta = \begin{bmatrix} \theta_0 \\ \mathbf{0} \end{bmatrix}$$

Interesting way for feature selection



The chance is very high to hit the sharp corners first

Ridge versus Lasso

Ridge

$$\tilde{E}(\theta) = \frac{1}{N}(y - z\theta)^{\mathrm{T}}(\mathrm{y} - z\theta) + \lambda\|\theta\|_2^2$$

It is a convex model

Both mean squared error and L2 regularizer are differentiable.

We can get a closed form solution

Lasso

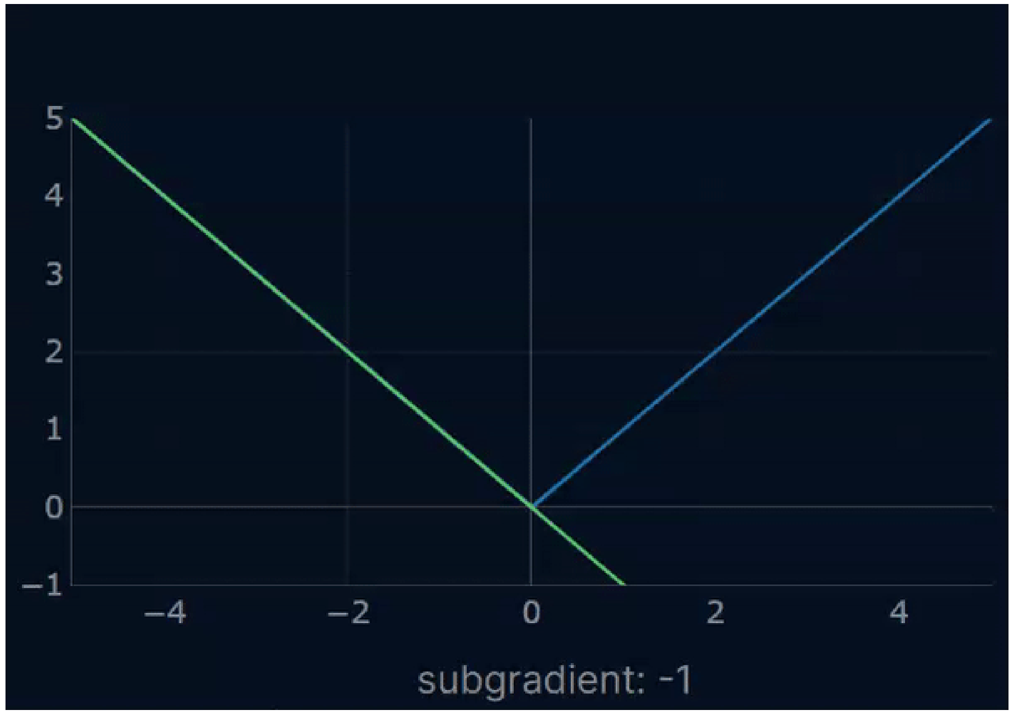
$$\tilde{E}(\theta) = \frac{1}{N}(y - z\theta)^{\mathrm{T}}(\mathrm{y} - z\theta) + \lambda\|\theta\|_1$$

It is a convex model

L1 regularizer is NOT differentiable.

We can **NOT** get a closed form solution

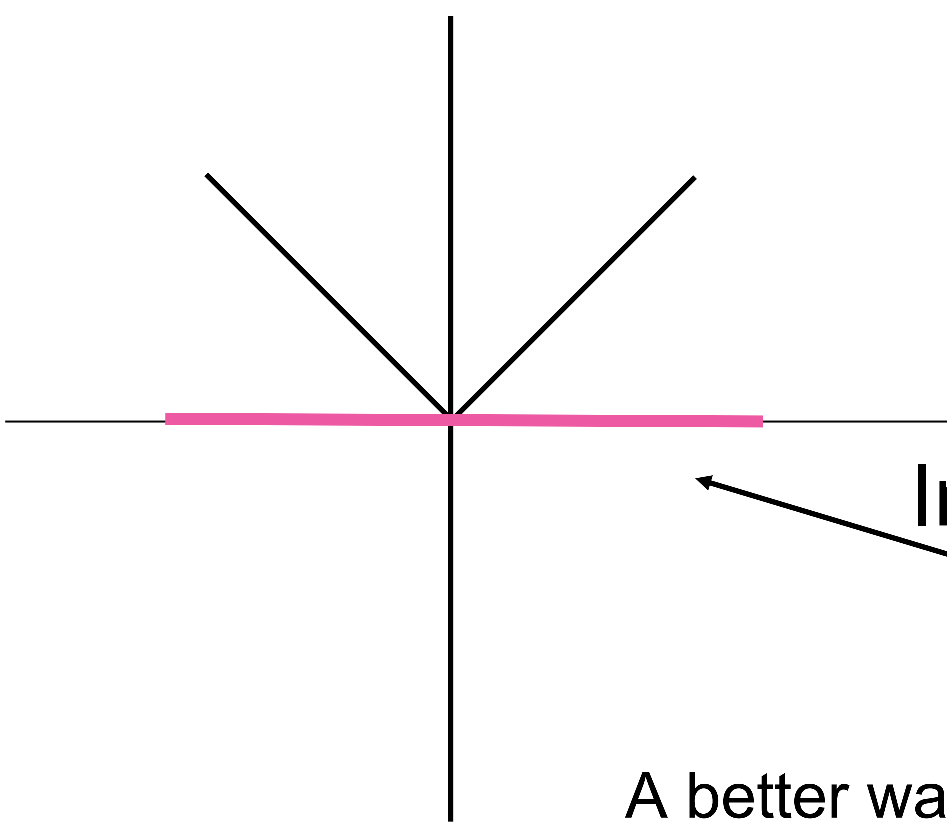
Sub-gradient Descend in Lasso



$$\tilde{E}(\theta) = \frac{1}{N}(y - z\theta)^{\mathrm{T}}(\mathrm{y} - z\theta) + \lambda\|\theta\|_1$$

$$\frac{\partial \tilde{E}(\theta)}{\partial \theta} = -z^T(y - z\theta) + \frac{\partial(\lambda\|\theta\|_1)}{\partial \theta}$$

Using Sub-gradient



$$\frac{\partial \tilde{E}(\theta)}{\partial \theta} = -z^T(y - z\theta) + \lambda sign(\theta)$$

In $sign$ function, we use this sub-gradient line as our under-estimator (below our function)

A better way: Proximal gradient descent with soft-thresholding

Outline

- Overfitting and regularized learning
- Ridge regression
- Lasso regression
- Determining regularization strength

Leave-One-Out Cross Validation

LOOCV

$\delta = 0.02$ (2)

$\delta = 0.01$ (1)

For every $i = 1, \ldots, n$:

- train the model on every point except i,
- compute the test error on the held out point.

$\Theta = (Z^T Z + \lambda I)^{-1} Z^T y$

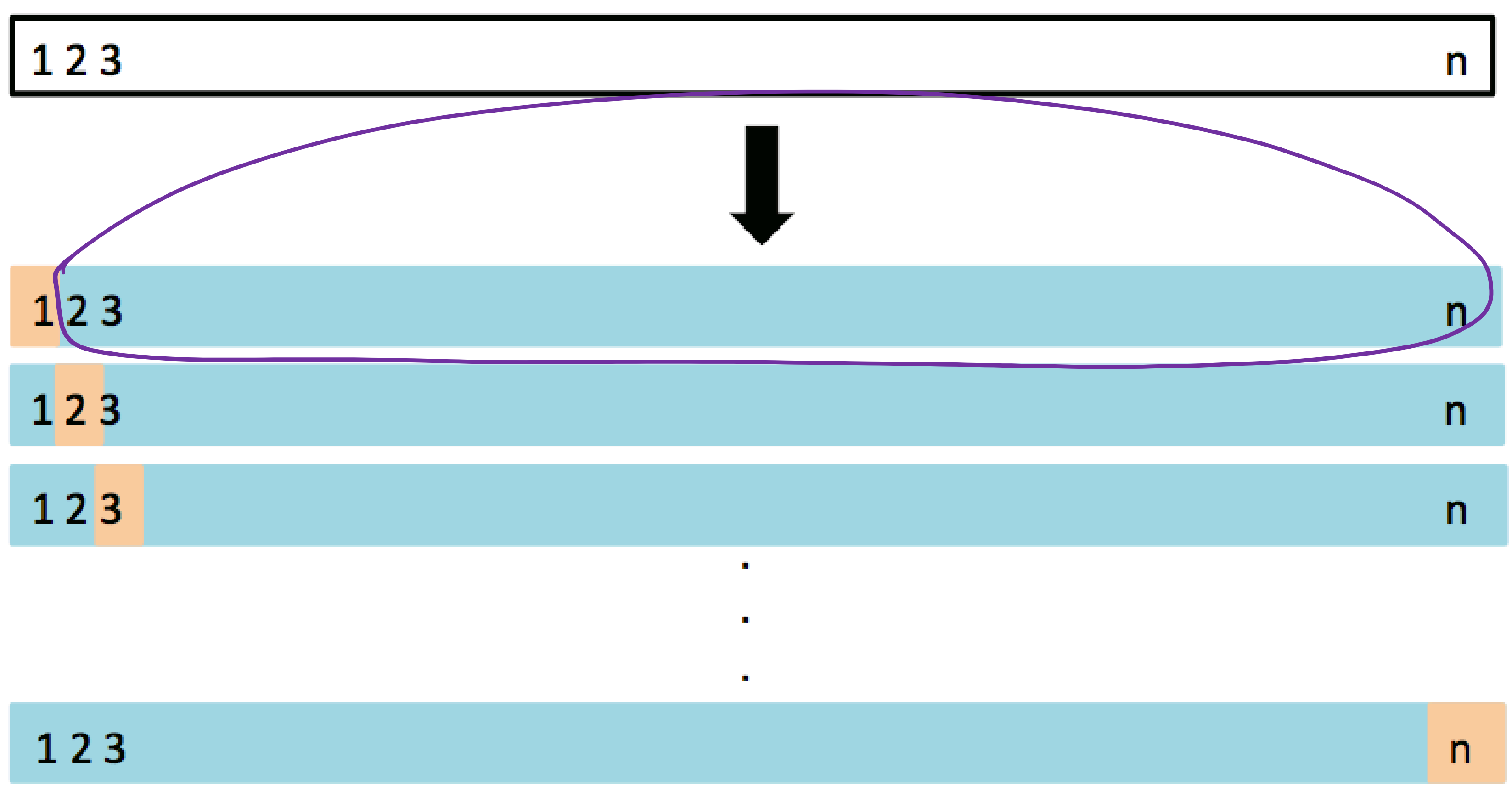
Average the test errors. $\mathrm{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i^{(-i)})^2$

$E(\Theta) = \frac{1}{N} \sum_{i=1}^{N} (y_a - \hat{y}_p)^2$

$E(\Theta)_1$

$E(\Theta)_2$

$E(\Theta)_n$



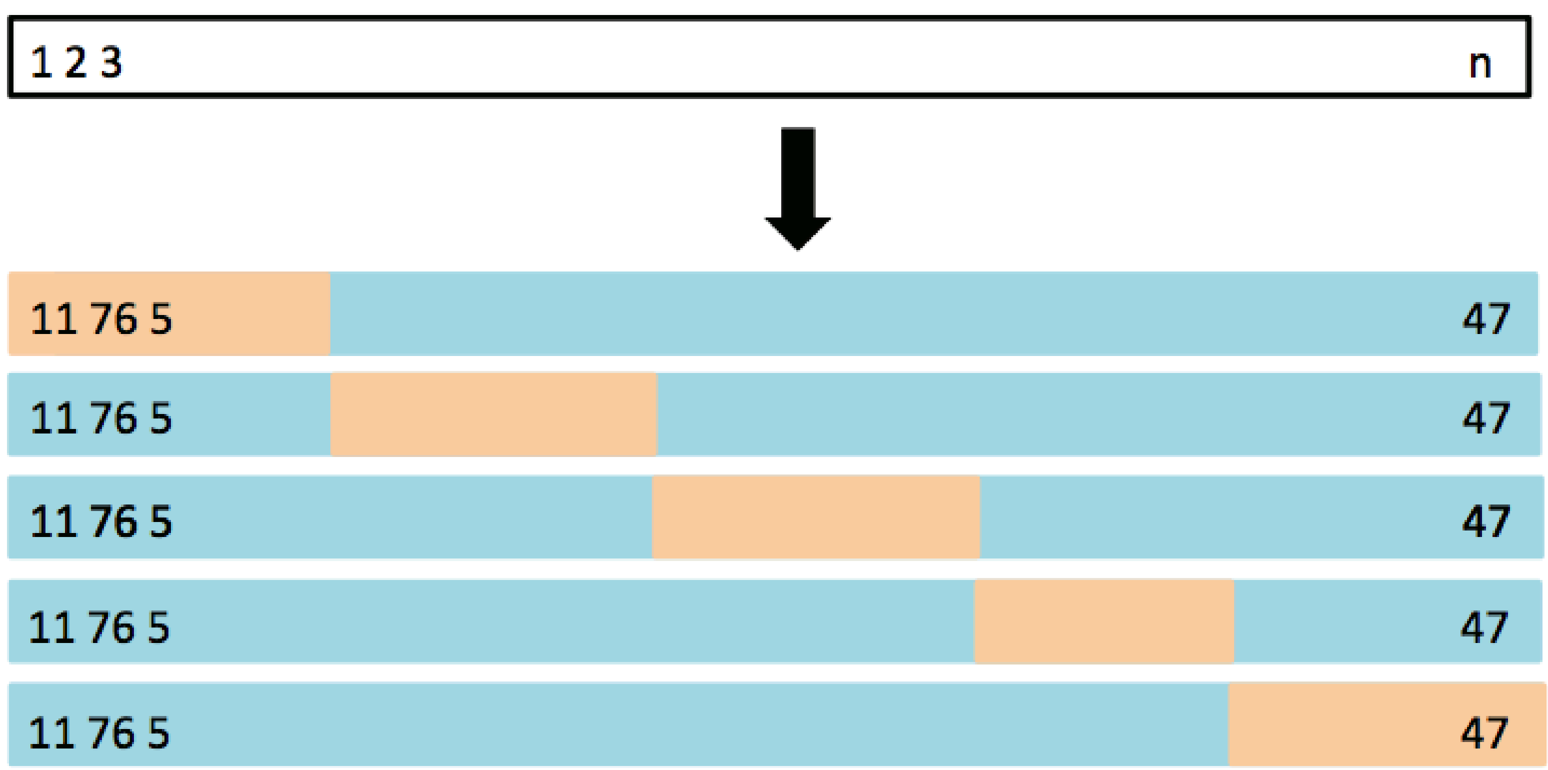
K-Fold Cross Validation

Split the data into k subsets or *folds*.

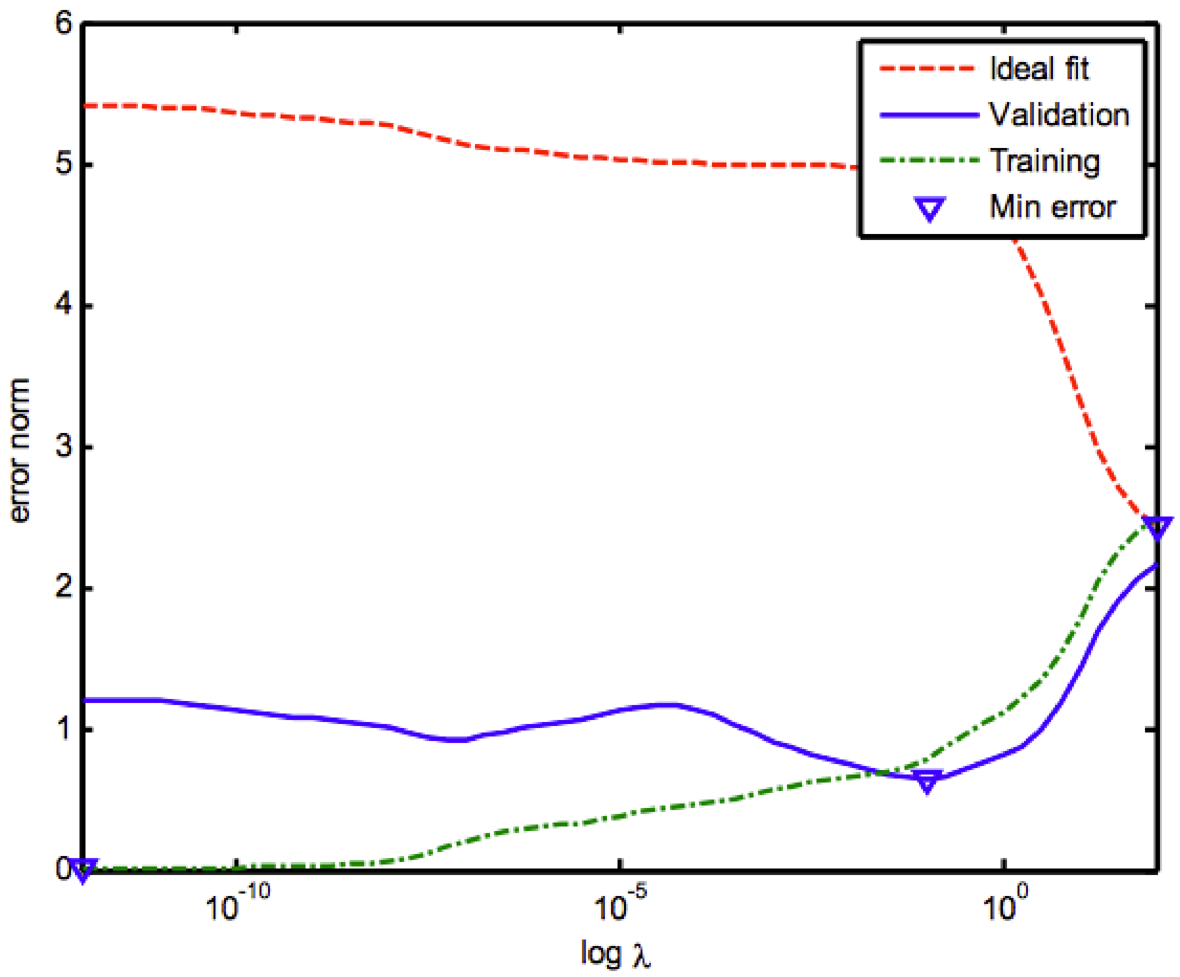
For every $i = 1, \ldots, k$:

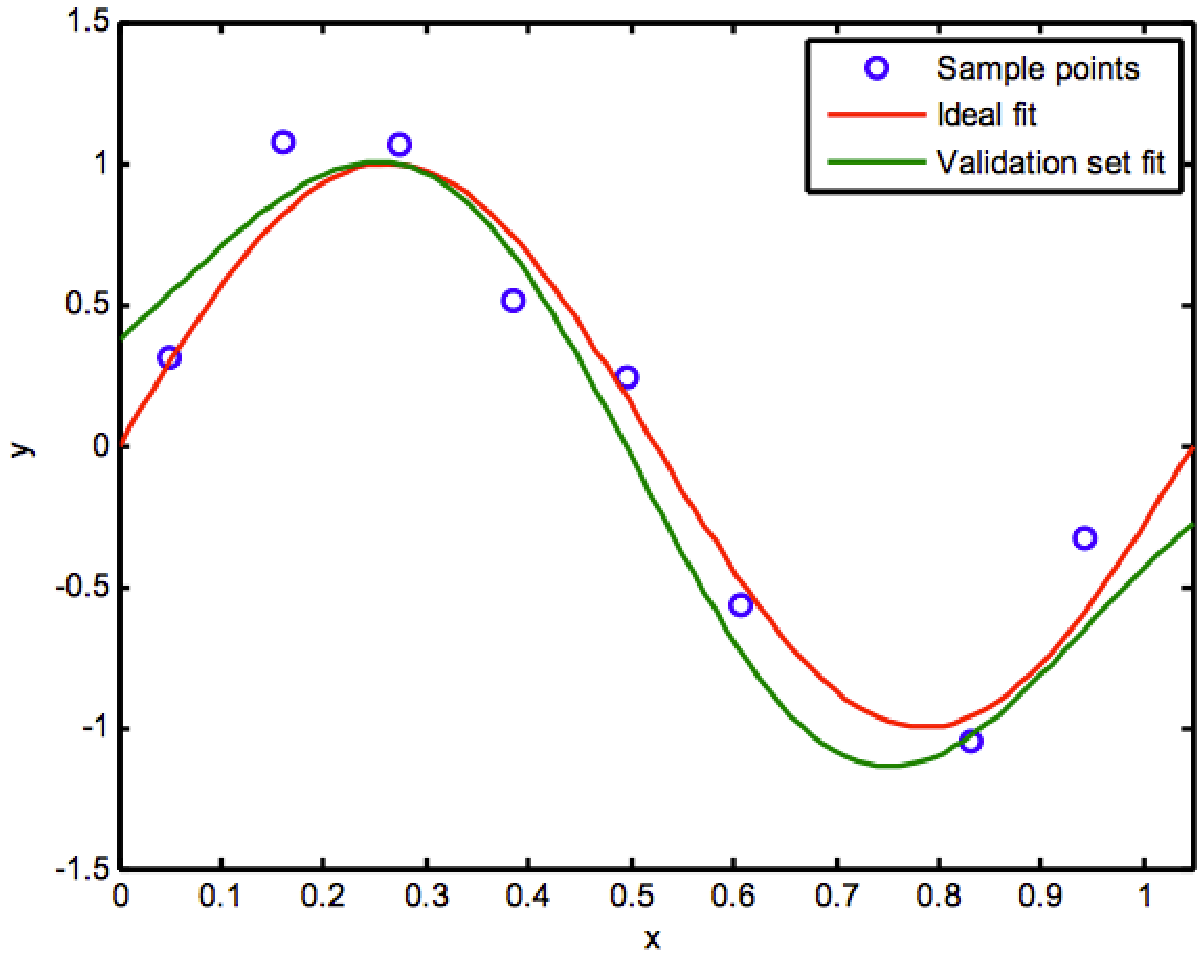
- train the model on every fold except the ith fold,
- compute the test error on the ith fold.

Average the test errors.



Choosing λ Using Validation Dataset





Pick up the lambda with the lowest mean value of rmse calculated by Cross Validation approach

Take-Home Messages

- What is overfitting
- What is regularization
- How does Ridge regression work
- Sparsity properties of Lasso regression
- How to choose the regularization coefficient λ