Georgia Tech

Regularized Linear Regression

Mahdi Roozbahani
Georgia Tech

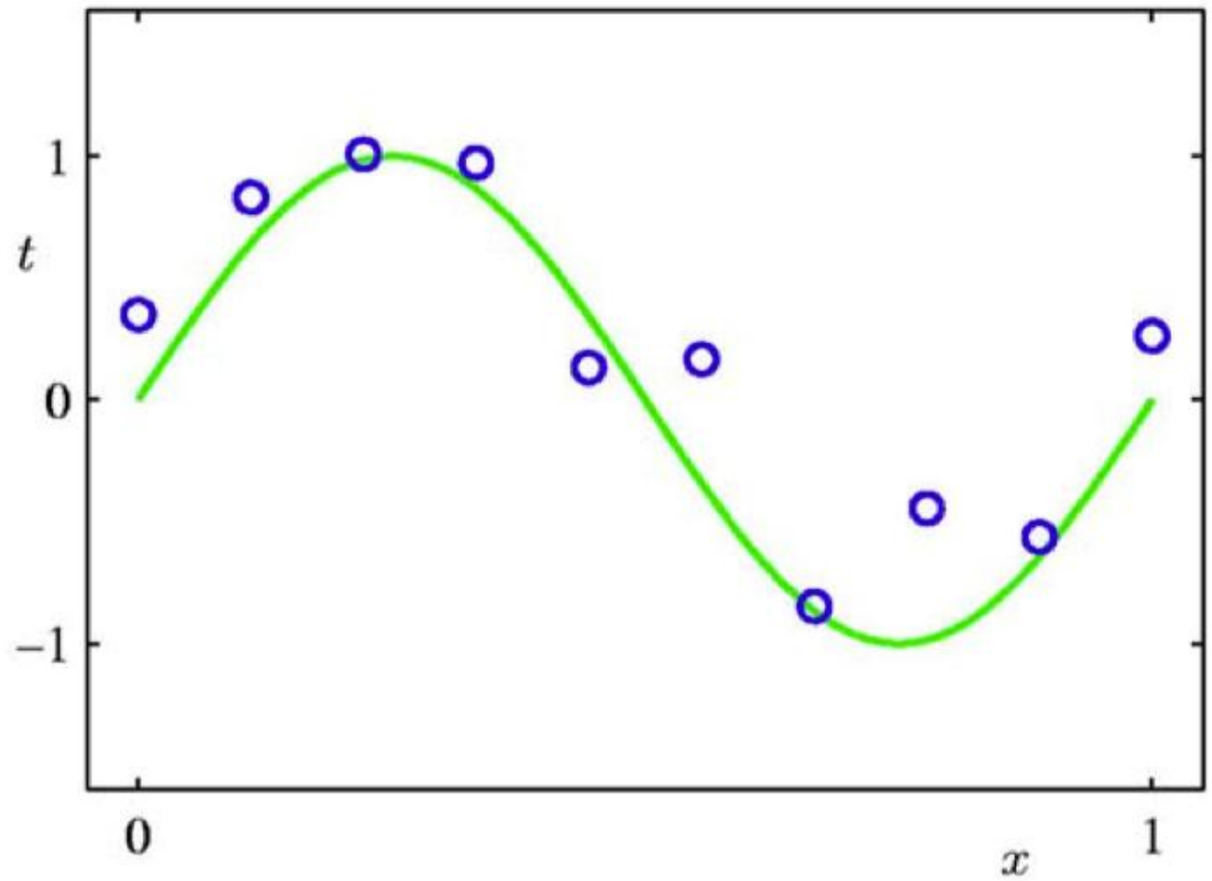
These slides are inspired based on slides from Andrew Zisserman, Jonathan Taylor, Chao Zhang and Yaser Abu-Mostafa.


THE BEST WAY TO
EXPLAIN OVERFITTING

Outline

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

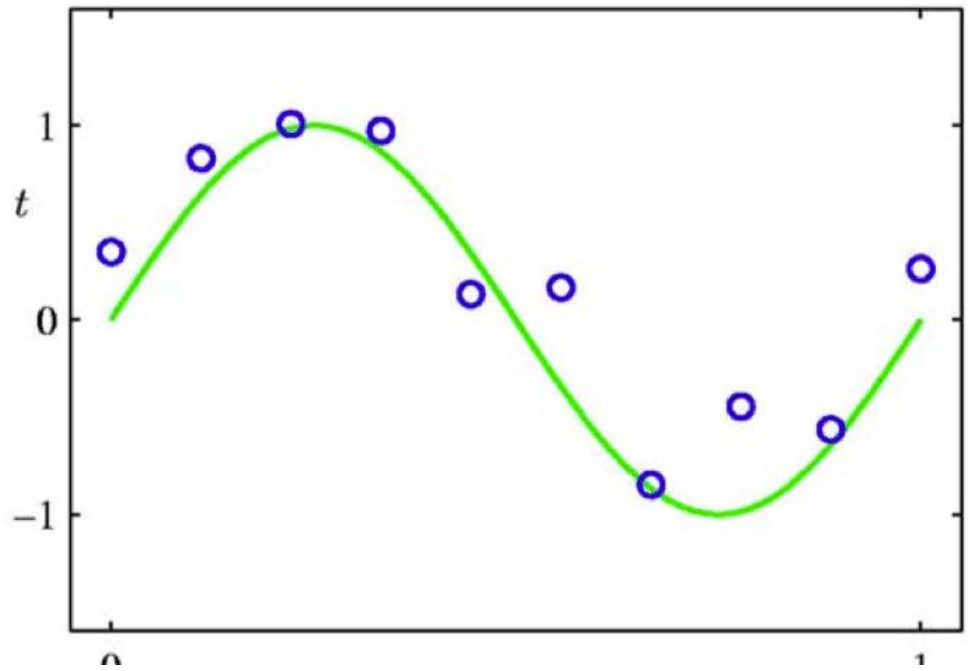
Regression: Recap



- Suppose we are given a training set of N observations

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

Regression: Recap



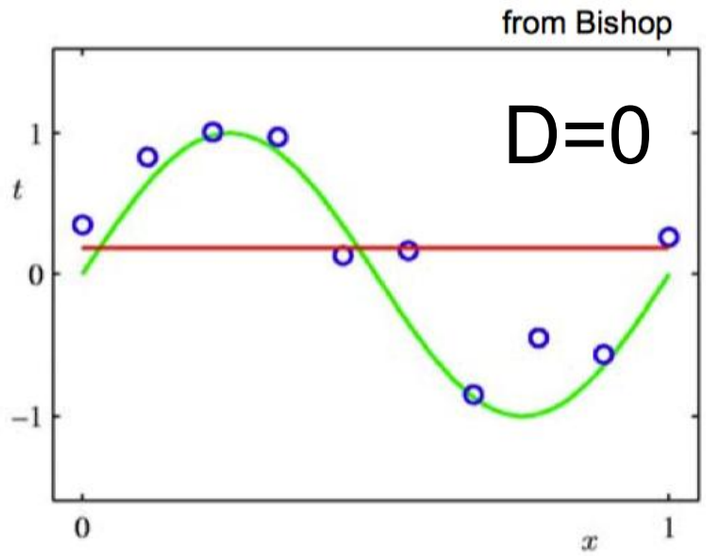
- Want to fit a polynomial regression model

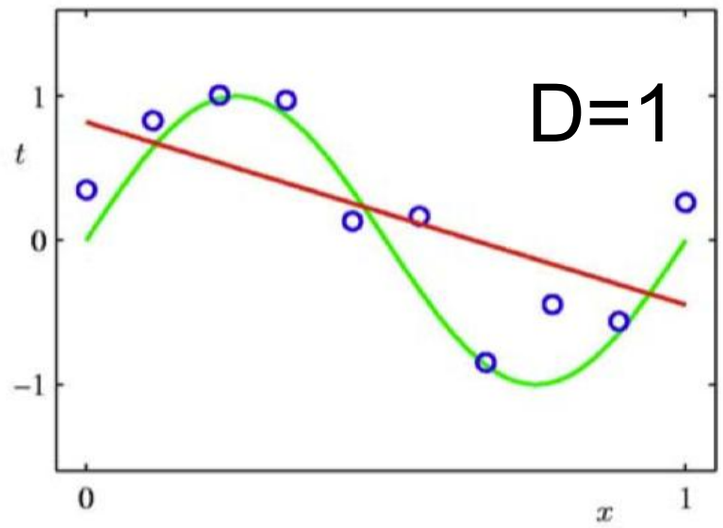
$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_d x^d + \epsilon$$

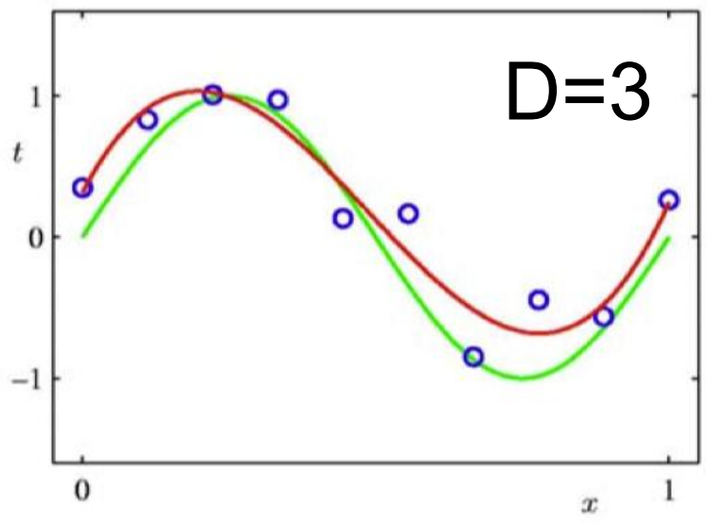
- $z = \{1, x, x^2, \ldots, x^d\} \in R^d$ and $\theta = (\theta_0, \theta_1, \theta_2, \ldots, \theta_d)^\top$

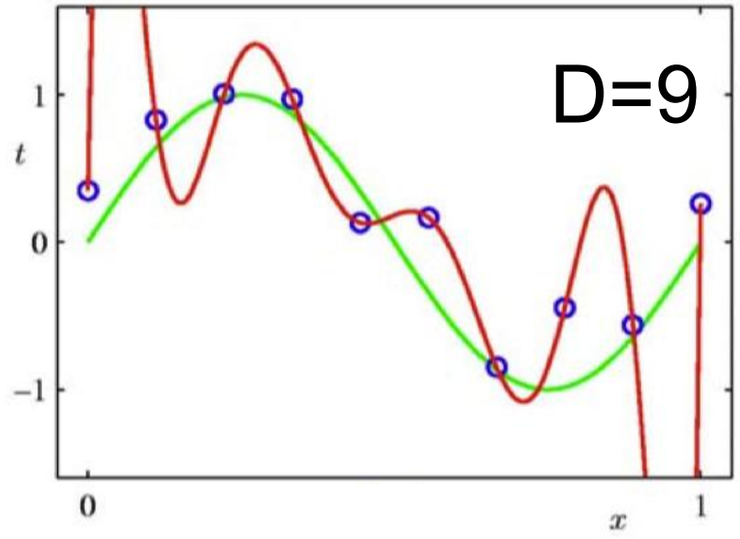
$$y = z\theta$$

Which One is Better?





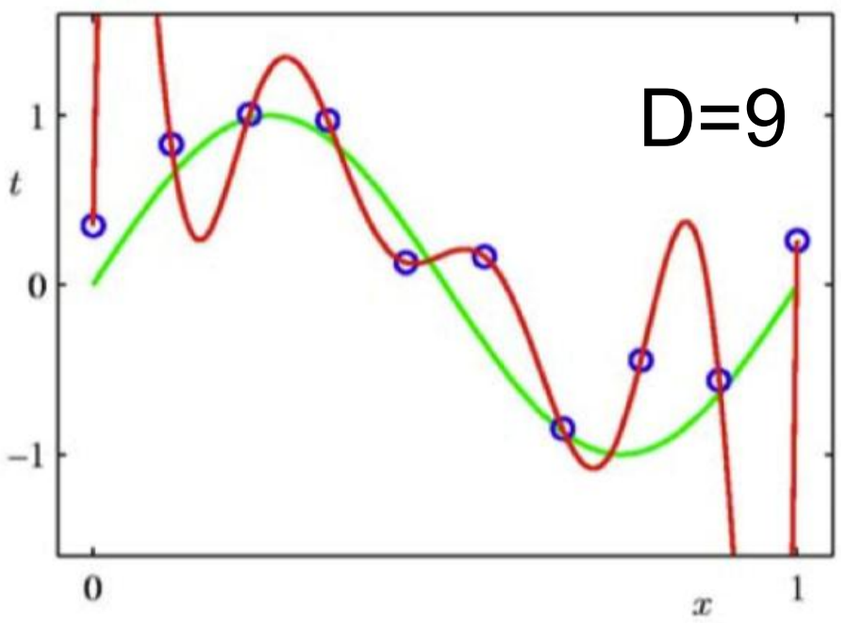


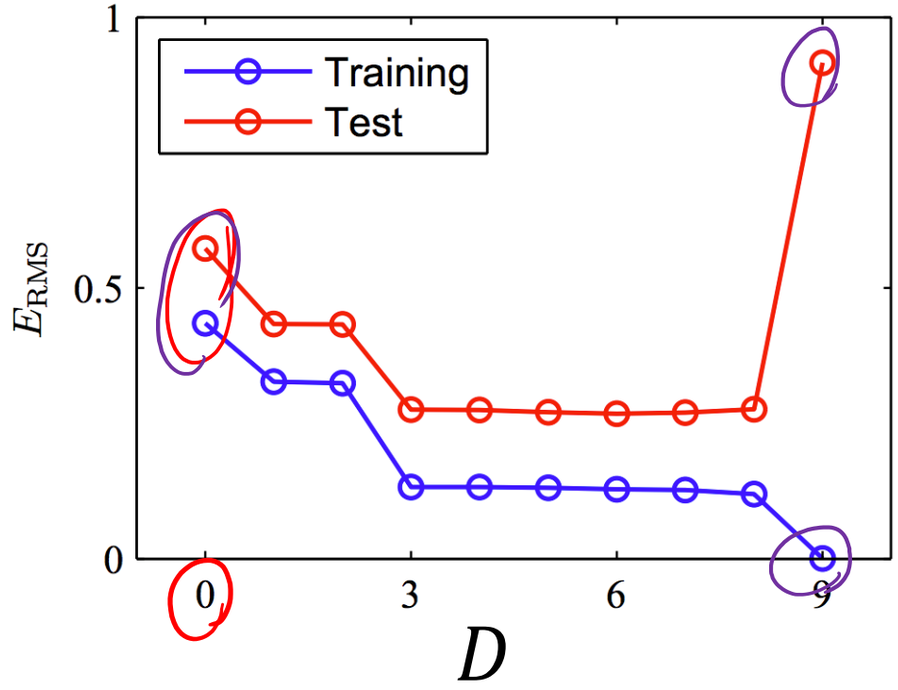


- Can we increase the maximal polynomial degree to very large, such that the curve passes through all training points?

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

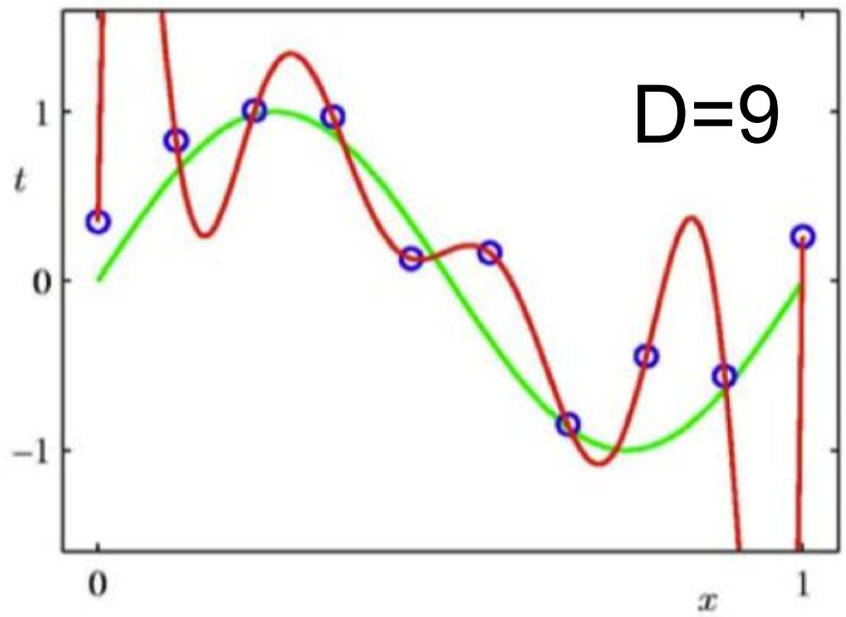
The Overfitting Problem





- 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.

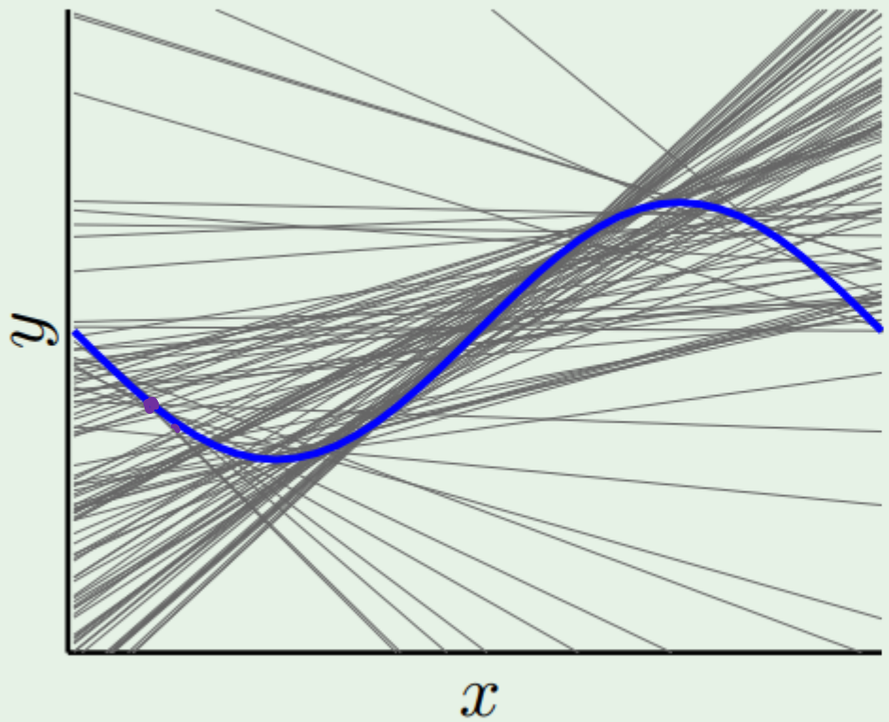
The Overfitting Problem



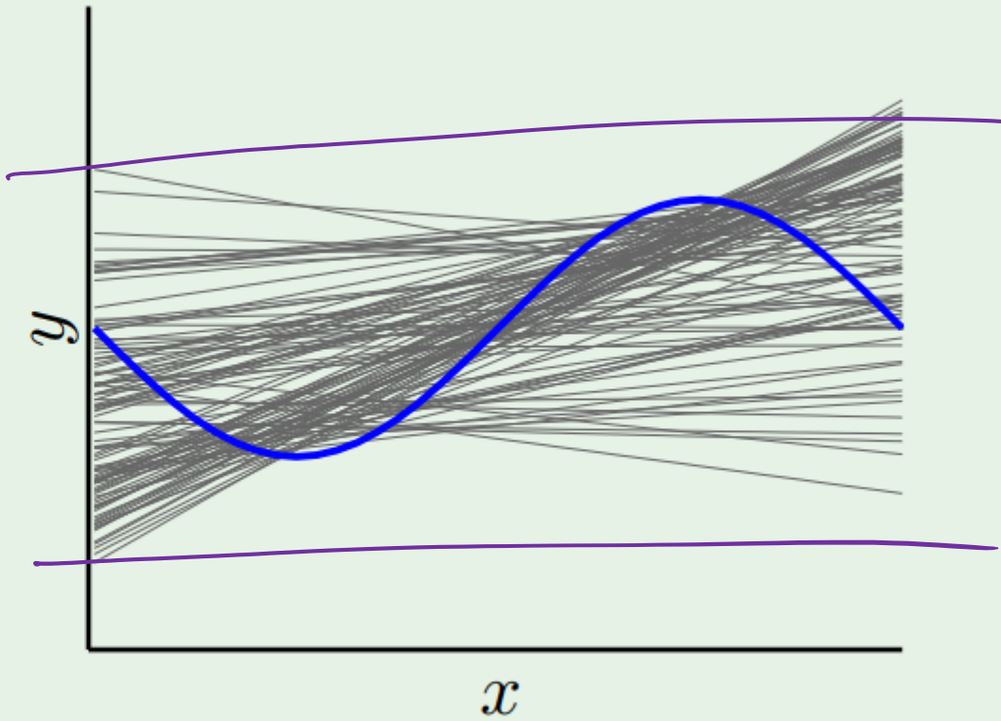
- In regression, overfitting is often associated with large Weights (severe oscillation)
- How can we address overfitting?

Regularization

(smart way to cure overfitting disease)



without regularization



with regularization

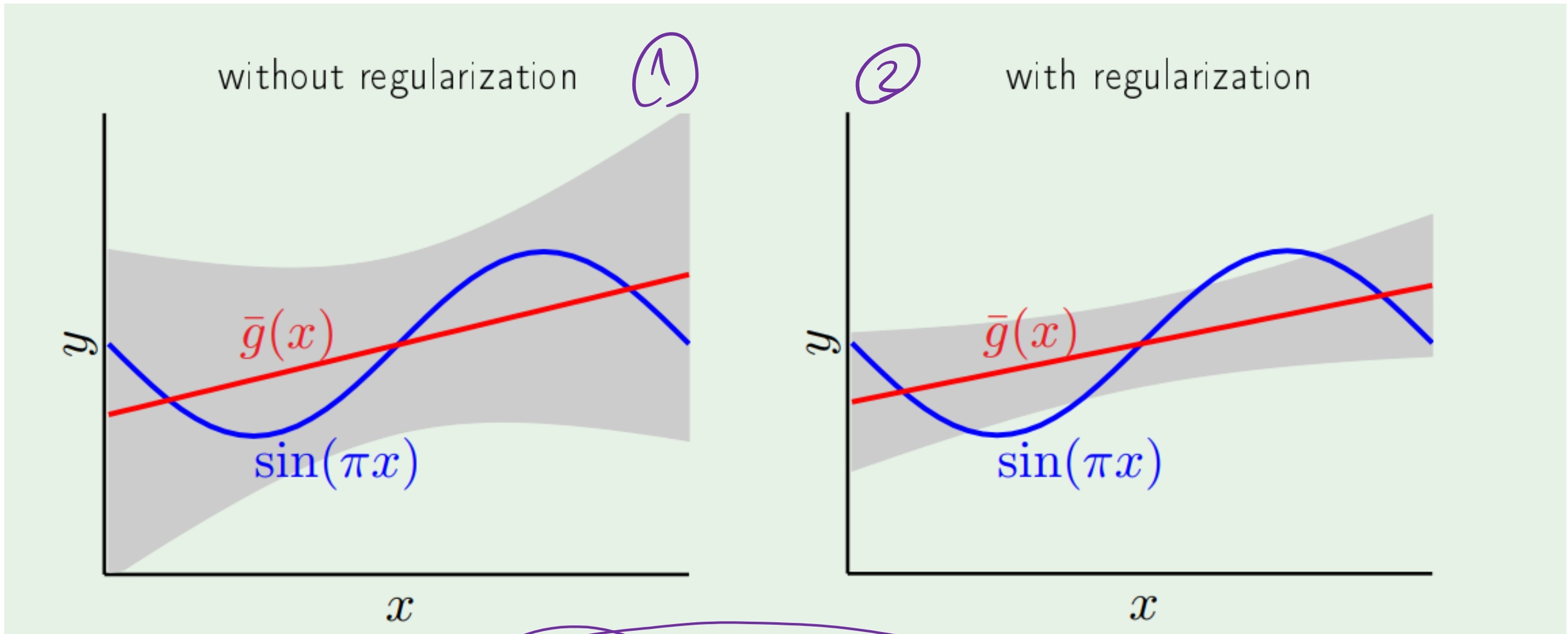
Put a brake on fitting

Fit a linear line on sinusoidal with just two data points

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

Who is the winner?

$\bar{g}\,(x)$: 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:

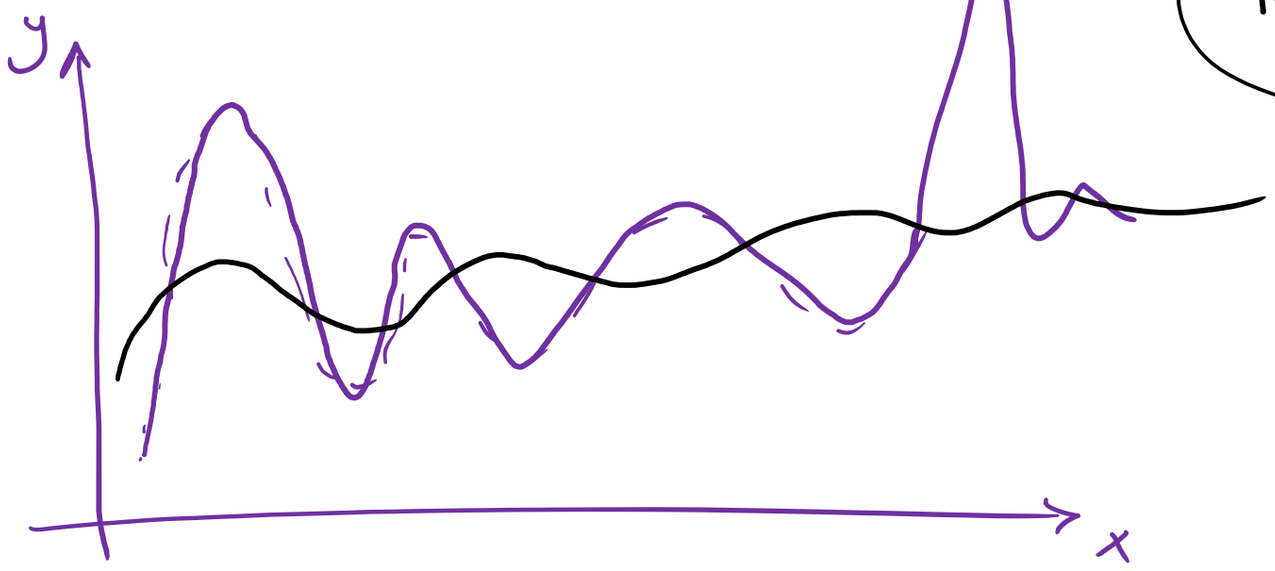
$$y = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \cdots + \theta_d z_{\mathrm{d}} + \epsilon = \boldsymbol{z\theta}$$

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

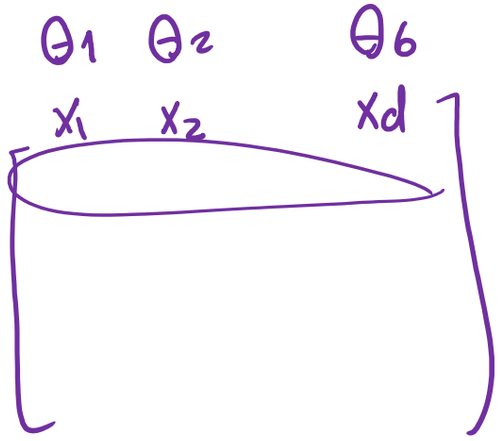
→ LCF

hyper Parameter

$$|\theta_0| + |\theta_1| + \cdots + |\theta_d| \leq 100$$



$X =$



$y =$

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d \approx y^{\{13\}}$$

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

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

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

z_1 z_2 z_3 ... z_d

h h^2 h^3 ... h^d

x_1 x_2 ... x_d

$x = [\ \ ]$ $\quad x_d = h^d$

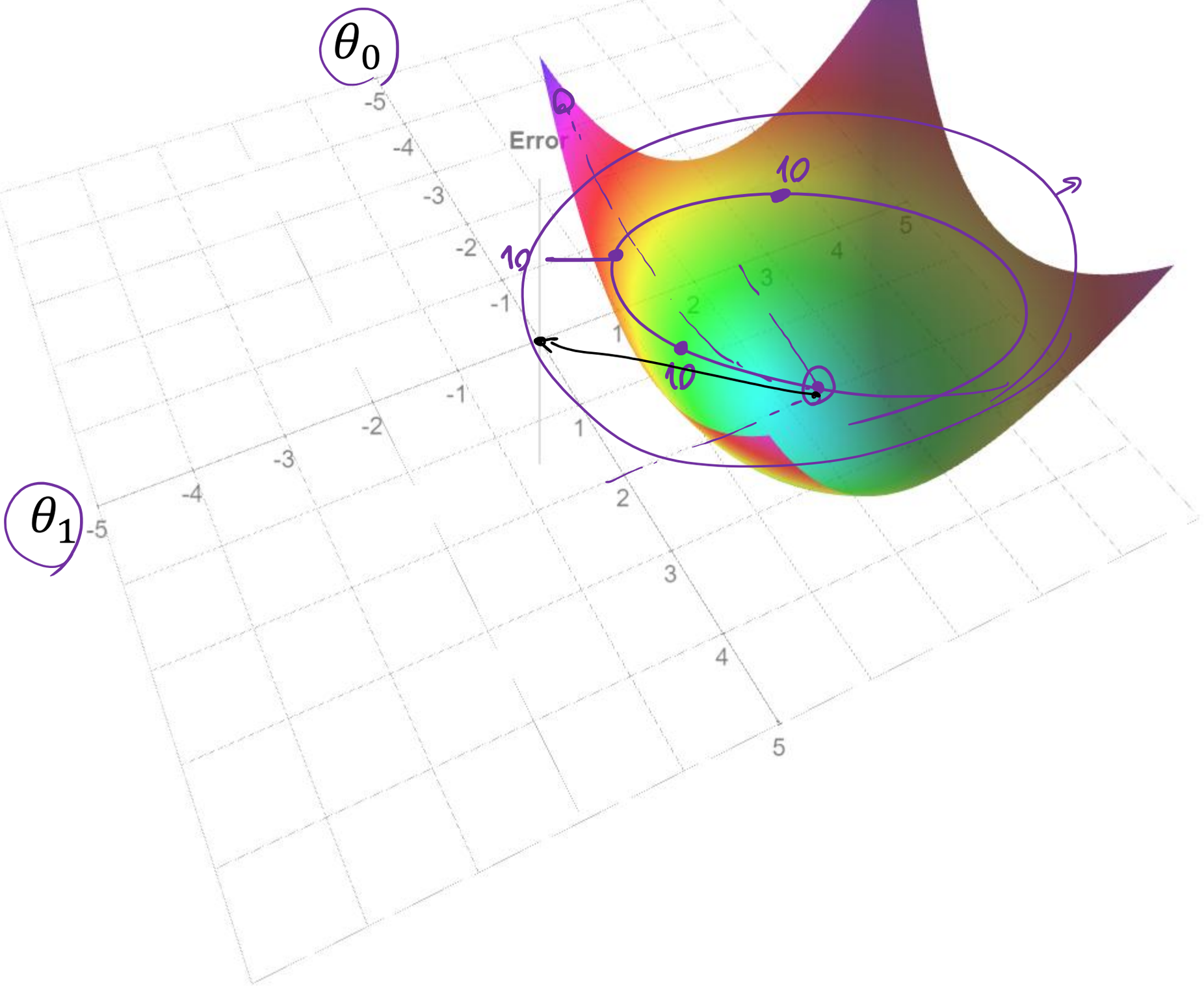
subject to

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

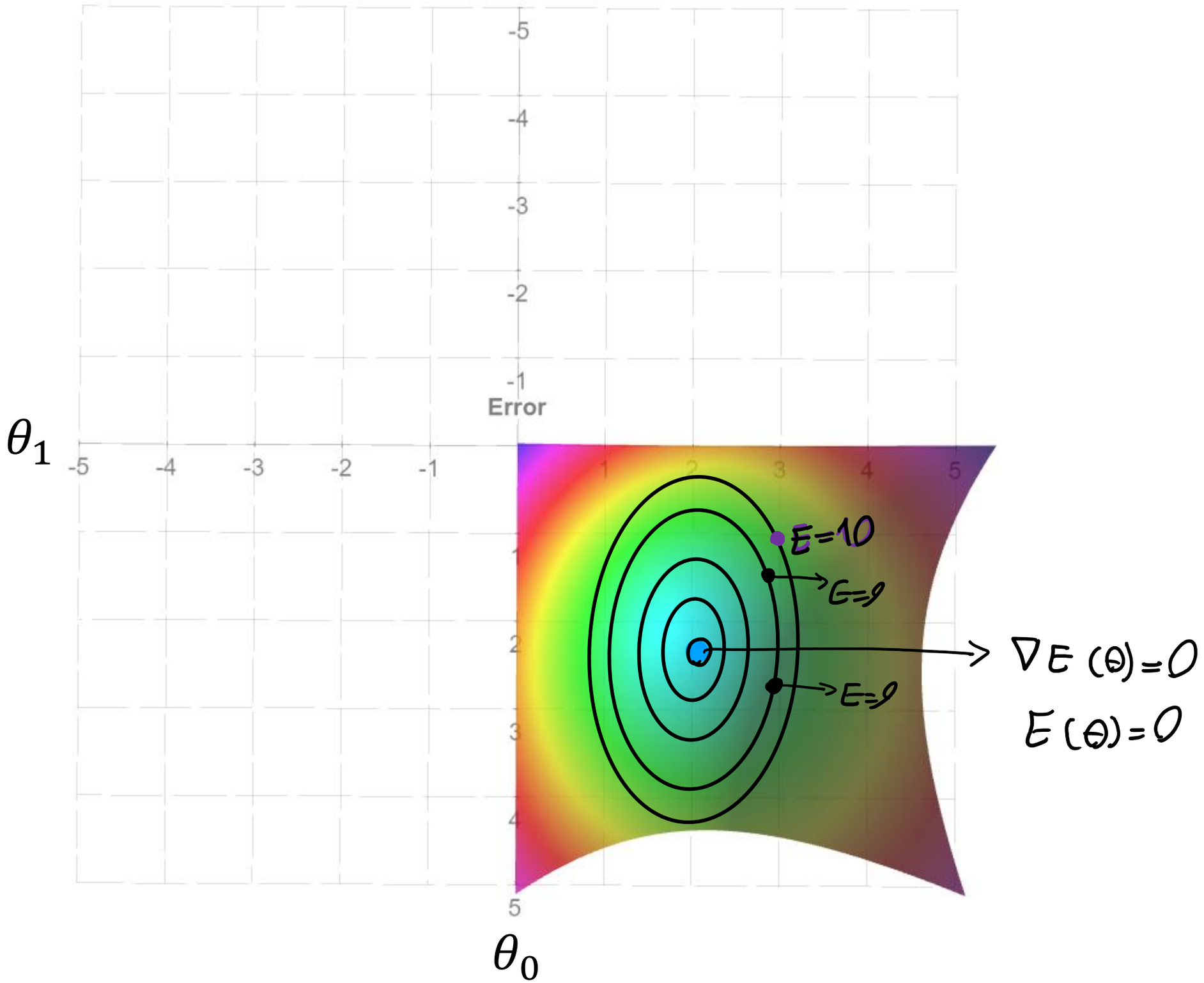
$$y = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + 0 + \cdots + 0$$

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

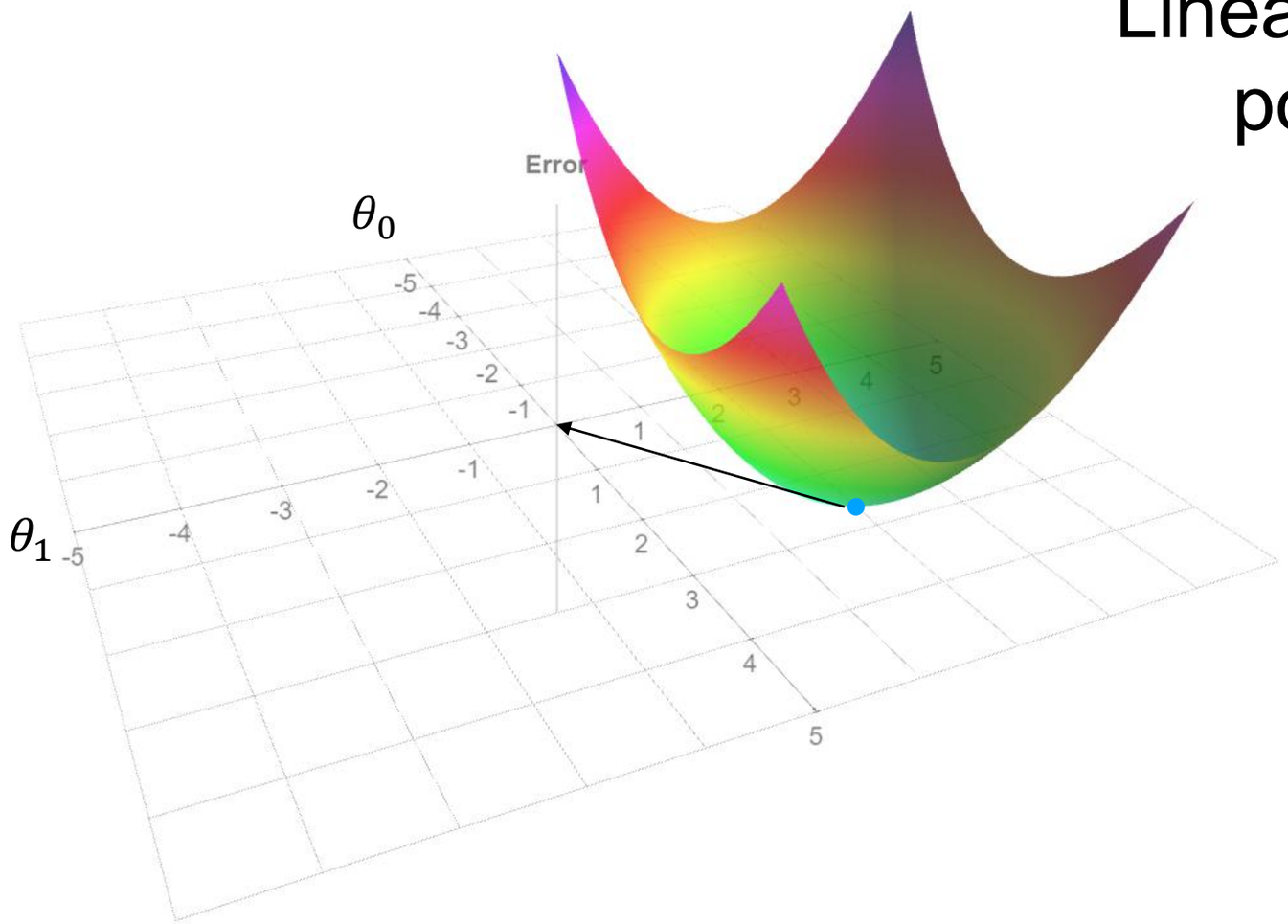
Min

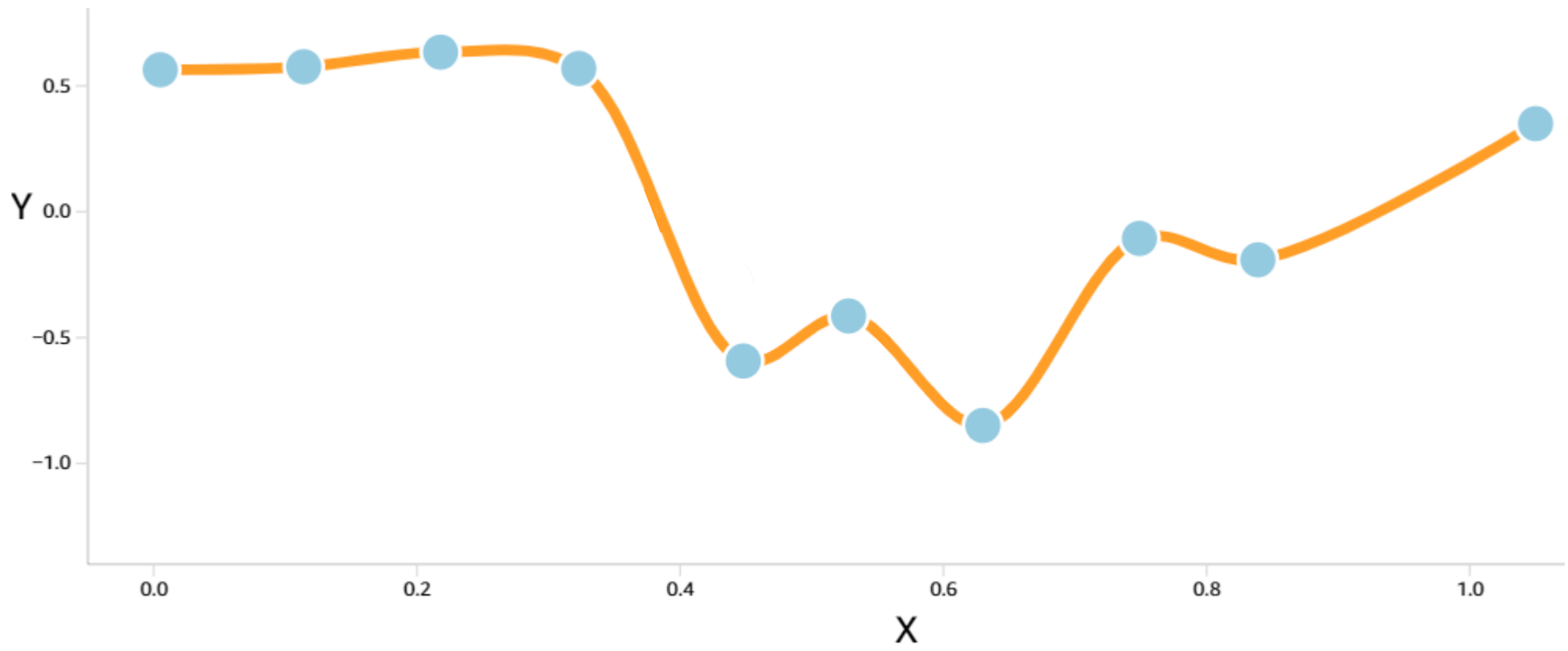


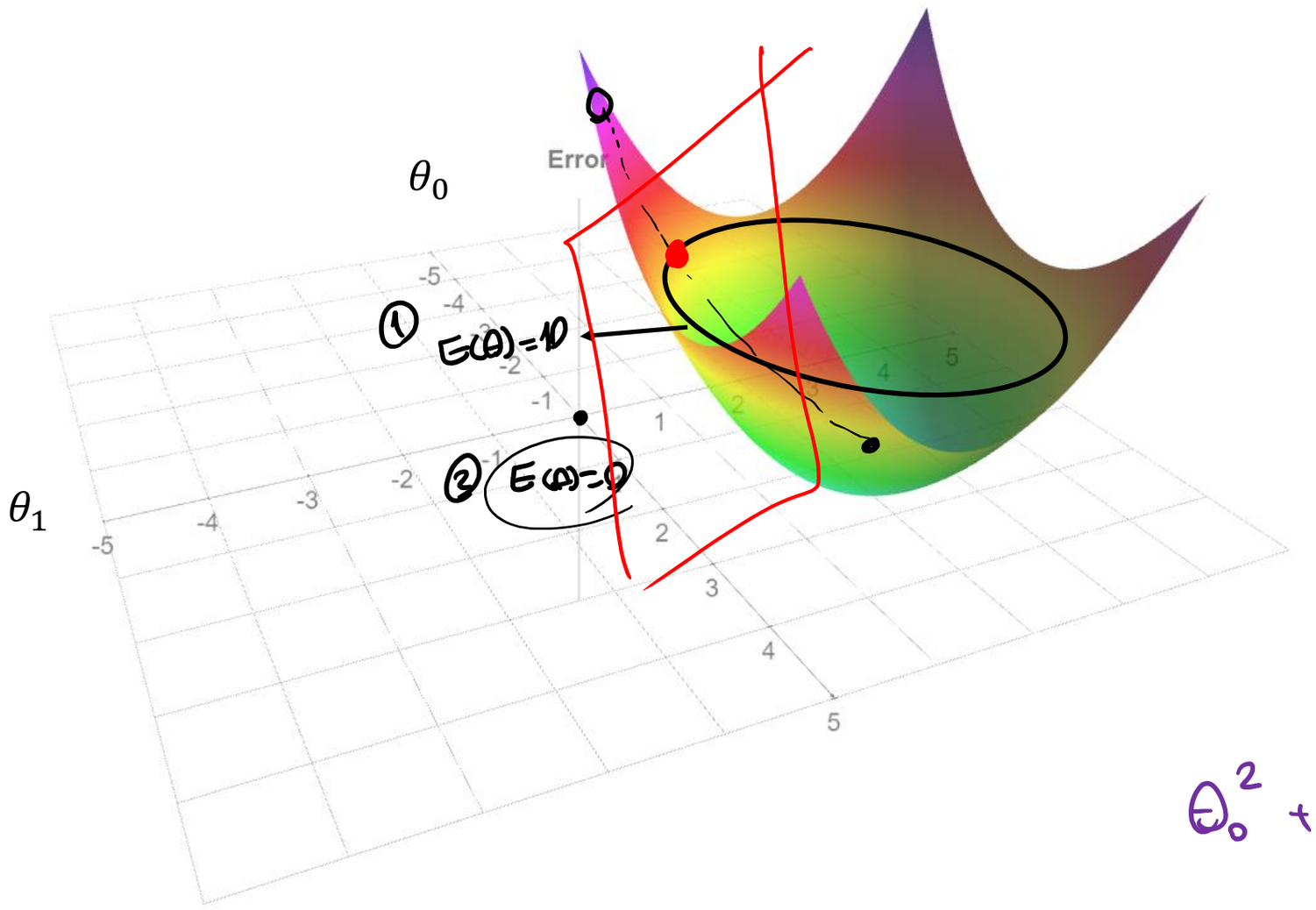
Project the same graph on x-y using contour plot



Linear regression with a very high polynomial degree solution



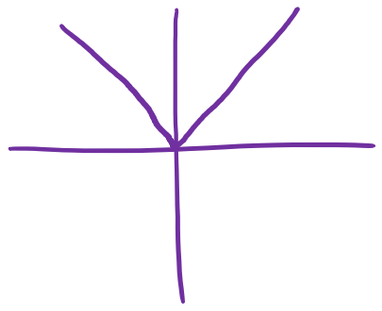




$$|\theta| \leq 100 = C$$

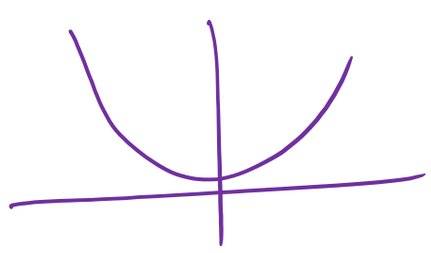
$$|\theta_0| + |\theta_1| + \cdots + |\theta_d| \leq C$$

$$g(\theta) = |\theta|$$



$$\theta_0^2 + \theta_1^2 + \cdots + \theta_d^2 \leq C$$

$$y = \theta^2$$

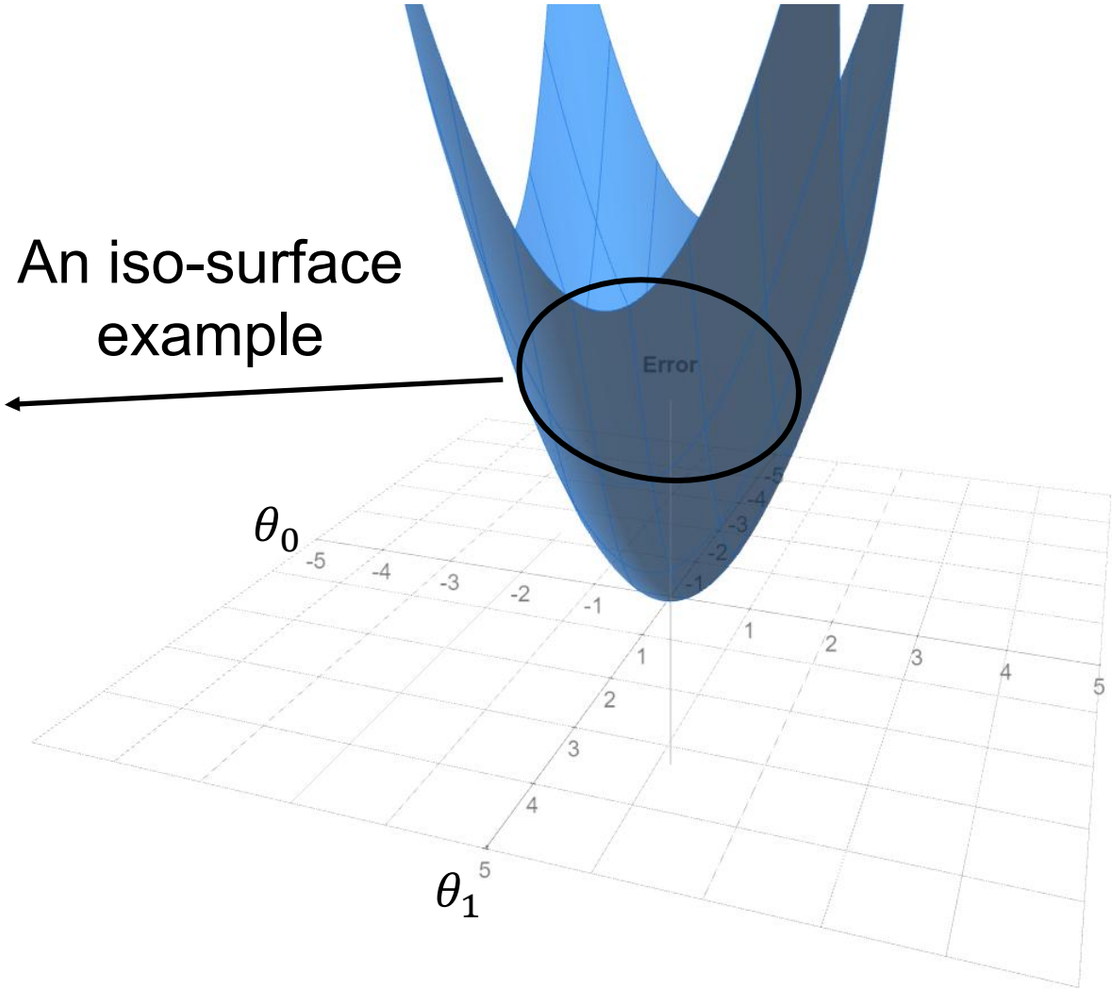


How can we get an optimal solution with a positive error for a model that overfits?

We need to introduce a constraint

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

An iso-surface example



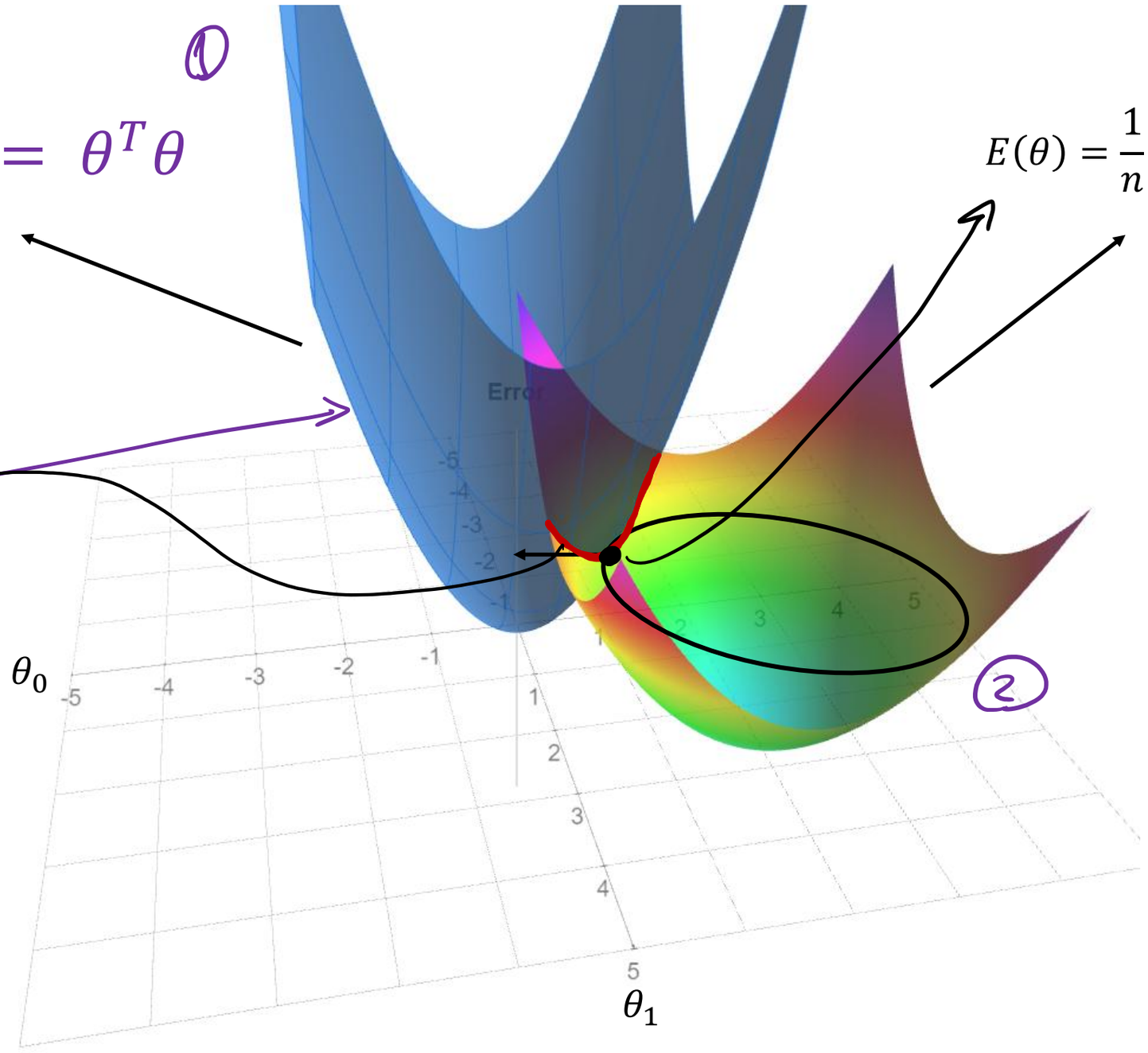
Error function together with a new introduced constraint

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

$$\theta^T\theta \leq c$$

$$\theta^T\theta = c$$

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



Let's define the Lagrange function

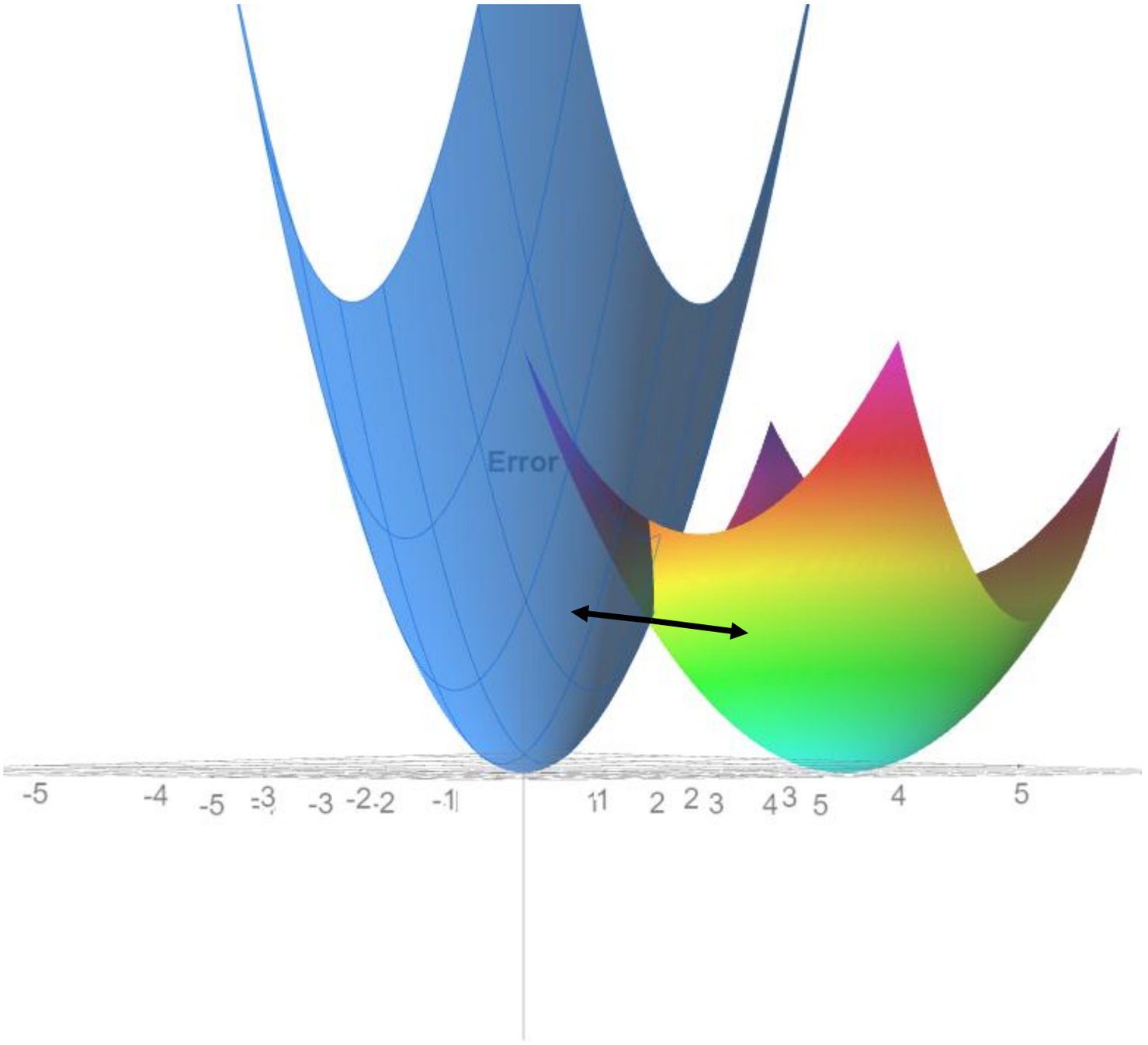
$$L(\theta, \lambda) = E(\theta) + \lambda g(\theta)$$

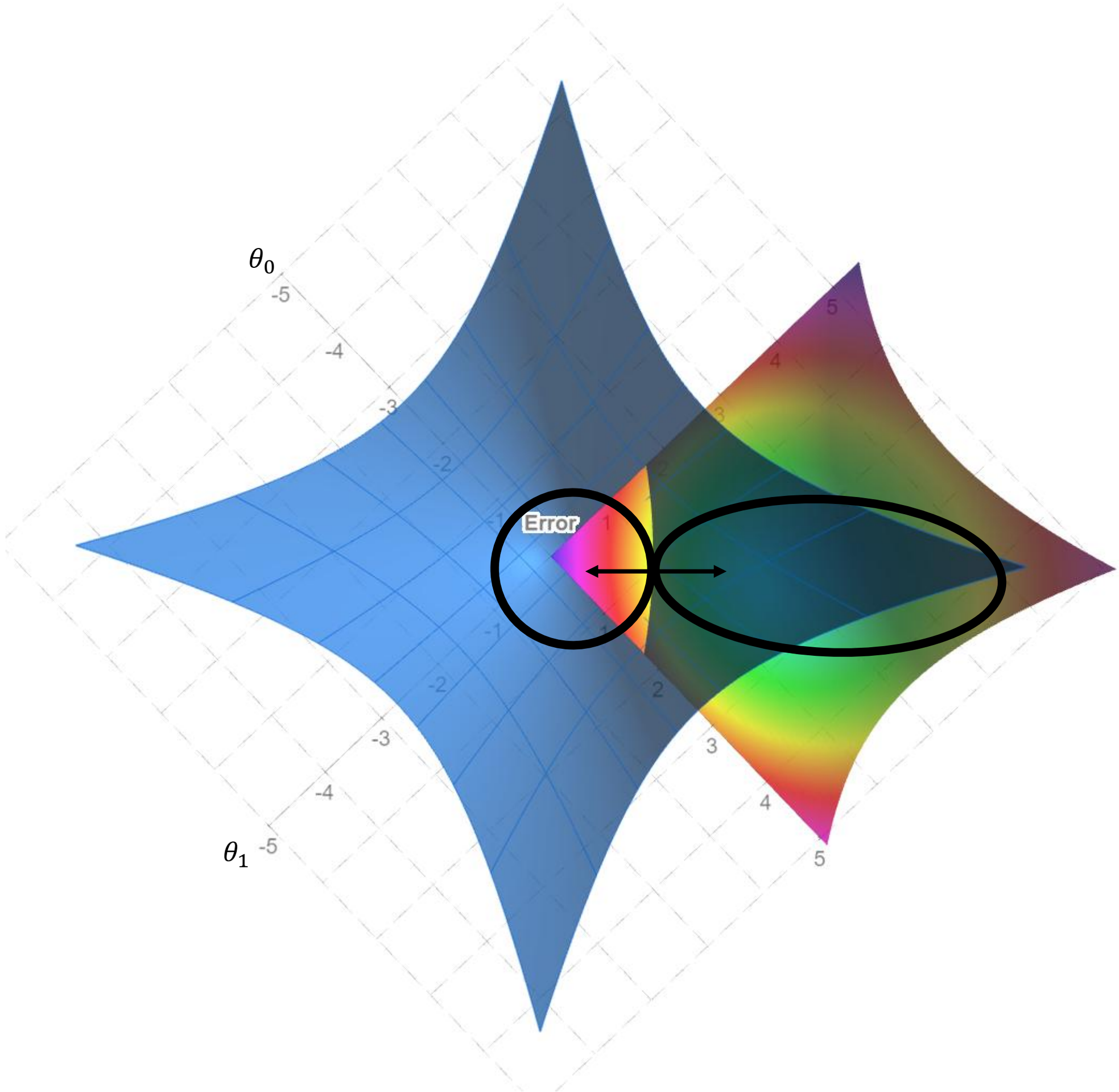
$$L(\theta, \lambda) = E(\theta) + \lambda \theta^T \theta$$

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

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

How to enforce the gradient of Lagrange function to be zero




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

Let's calculate the gradients

Gradient of constraint $g(\theta)$ $\nabla[\theta^T\theta] = 2\theta$

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

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

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

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

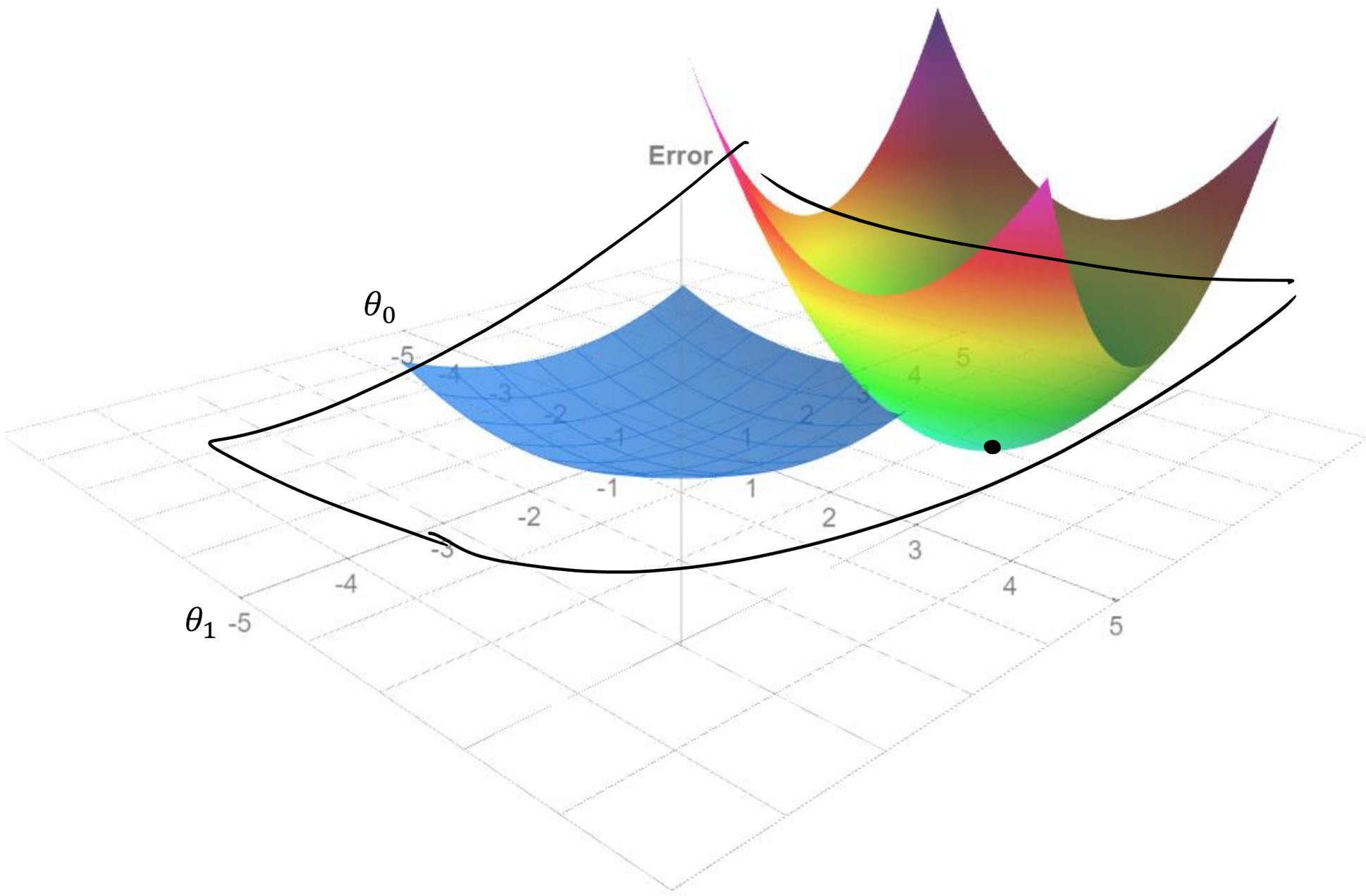
Let's do integration $E(\theta) + \lambda\theta^T\theta$

The effect of low Lambda

$E(\theta) = \frac{1}{N} \sum (y_a^{(i)} - y_b^{(i)})^2$

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

λ = large



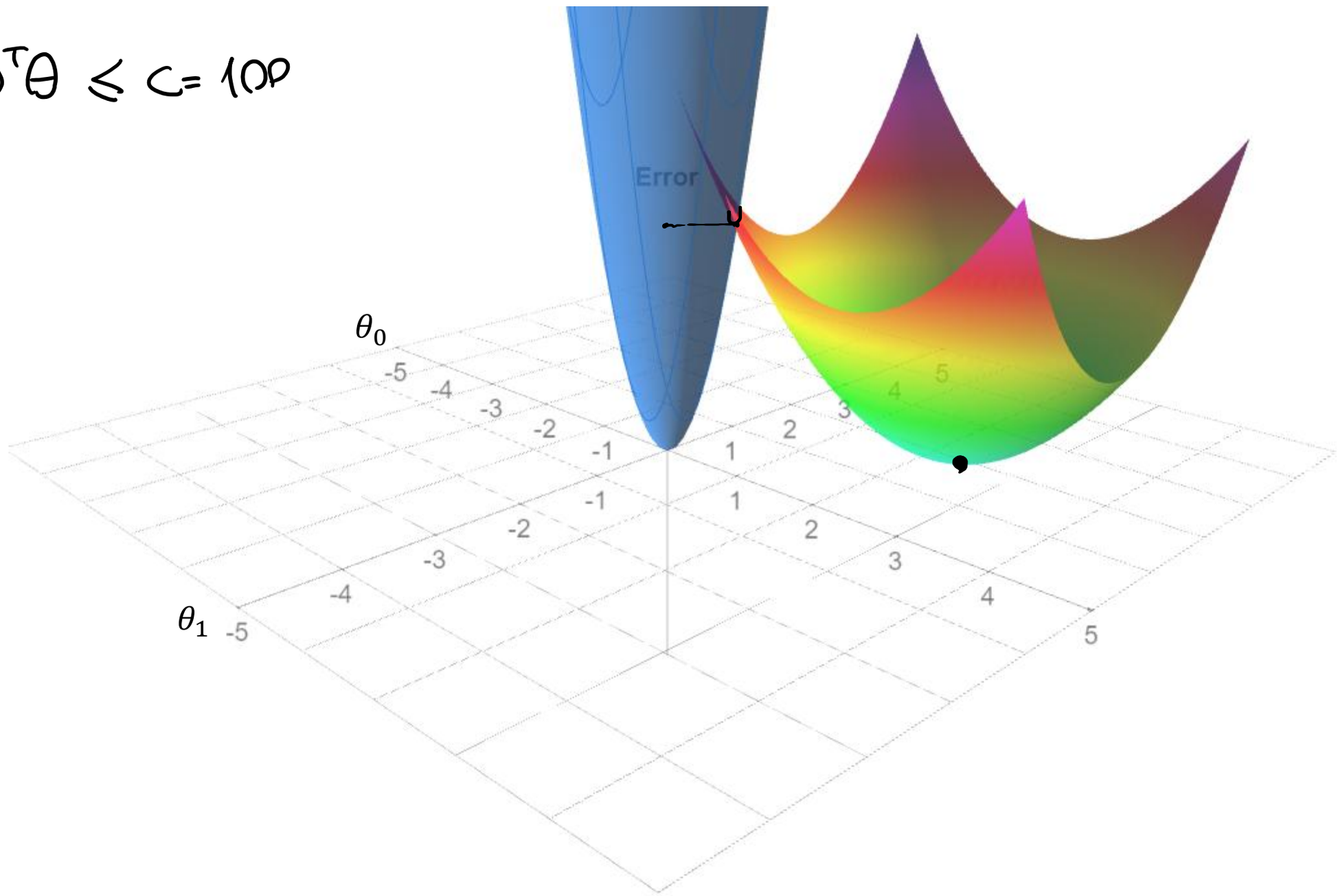
The effect of high Lambda

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

C = small

$\theta^T\theta \leq 99$

$\theta^T\theta \leq C = 100$



$$\text{Min} \quad E(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y_a^{\{ij\}} - X^{\{ij\}}\theta\right)^2$$

$\|\theta\|^2 = \theta^T\theta \qquad \|\theta\|^2 \leq C \rightarrow^{s.t} \theta^T\theta = C \qquad g(\theta) = \theta^T\theta - C$

$$L(\theta, \delta) = E(\theta) + \delta(\theta^T\theta - C) \Rightarrow \frac{\partial L}{\partial \theta} = 0 \; ; \; \frac{\partial L}{\partial \delta} = 0$$

Let's say someone gave us δ already $\rightsquigarrow$ $\delta \rightarrow$ constant

$\theta^T\theta = C$

$\theta_1^2 + \theta_2^2 + \cdots + \theta_1^2 = C$

Min

$L(\theta) = E(\theta) + \delta\theta^T\theta - \cancel{\delta C} = E(\theta) + \delta(\theta^T\theta)$

$\delta\uparrow$ large $\quad \downarrow C = \theta^T\theta$

$\downarrow\delta$ small $\quad \theta^T\theta$

Regularized Learning

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

Overfit

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

Regularized Error

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

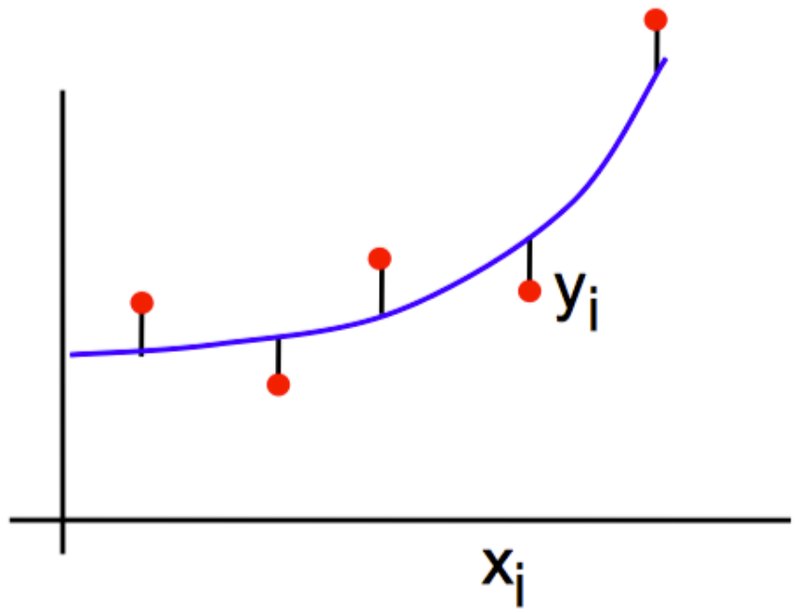
L2 Regularization term

Outline

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

Ridge Regression

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



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

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

Convex

Convex

$\theta^T D \theta$

$D = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & \\ \vdots & \vdots & \vdots & \end{bmatrix}$

$\lambda\theta^T I\theta$

$\lambda|\theta|$

$\lambda(\theta_0^2 + \theta_1^2 + \cdots + \theta_d^2)$

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

$\theta_1 X + \theta_2 X^2 + \theta_3 X^3$

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

$\frac{\partial(\frac{\lambda}{2}\theta^2)}{\partial\theta} = \lambda\theta$

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

$$\rightarrow \theta = (z^T z + \lambda I)^{-1} z^T y$$

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

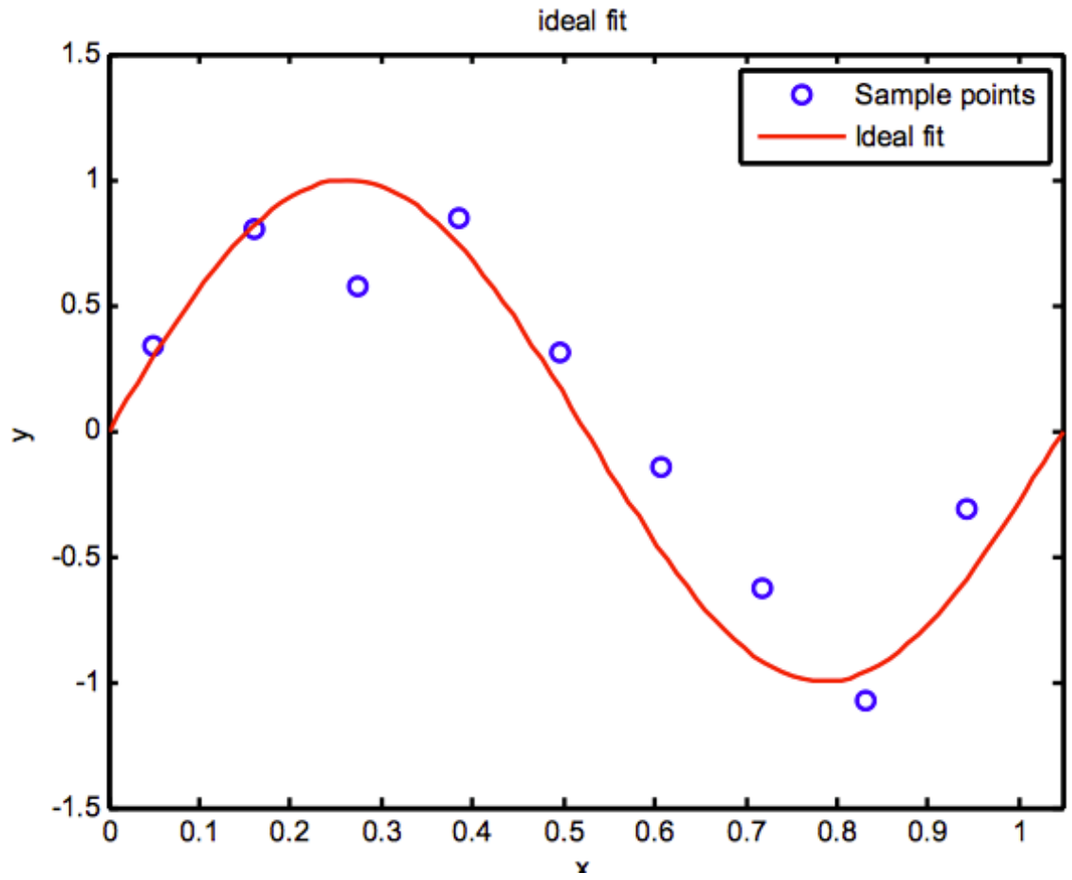
$\approx \frac{1}{z^T z + \lambda I}$

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.

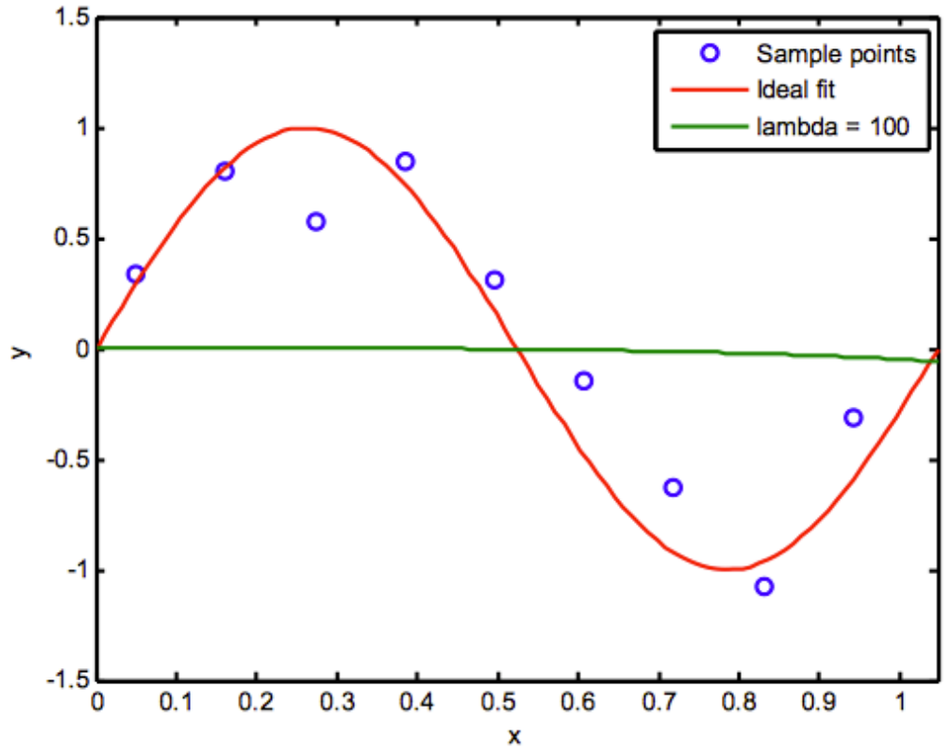
- There is a choice in both the degree, D, of the basis functions used, and in the strength of the regularization

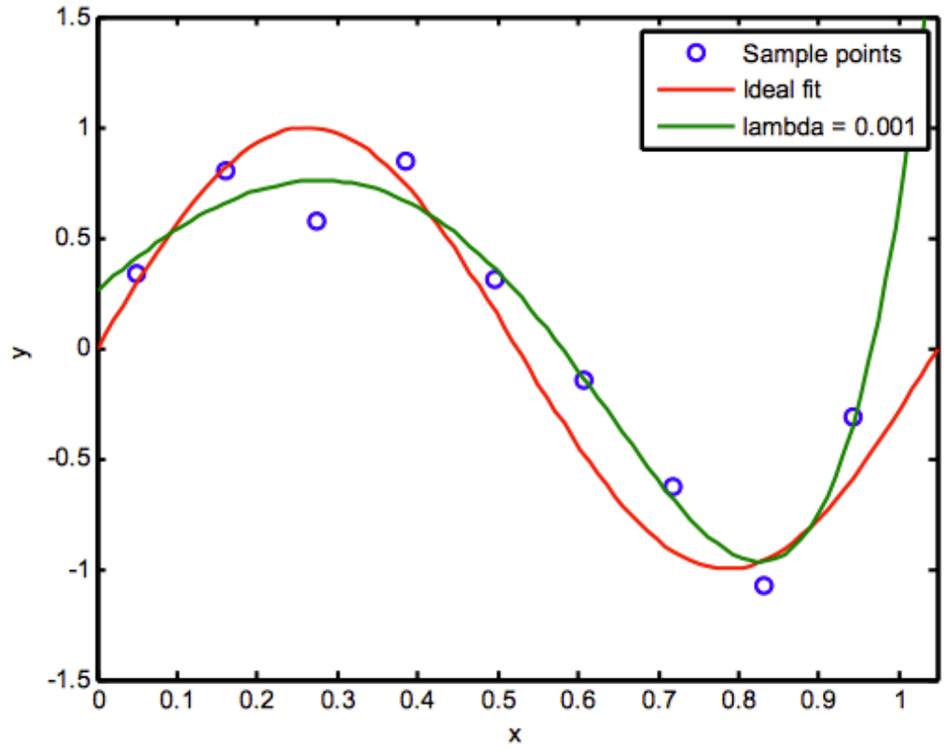


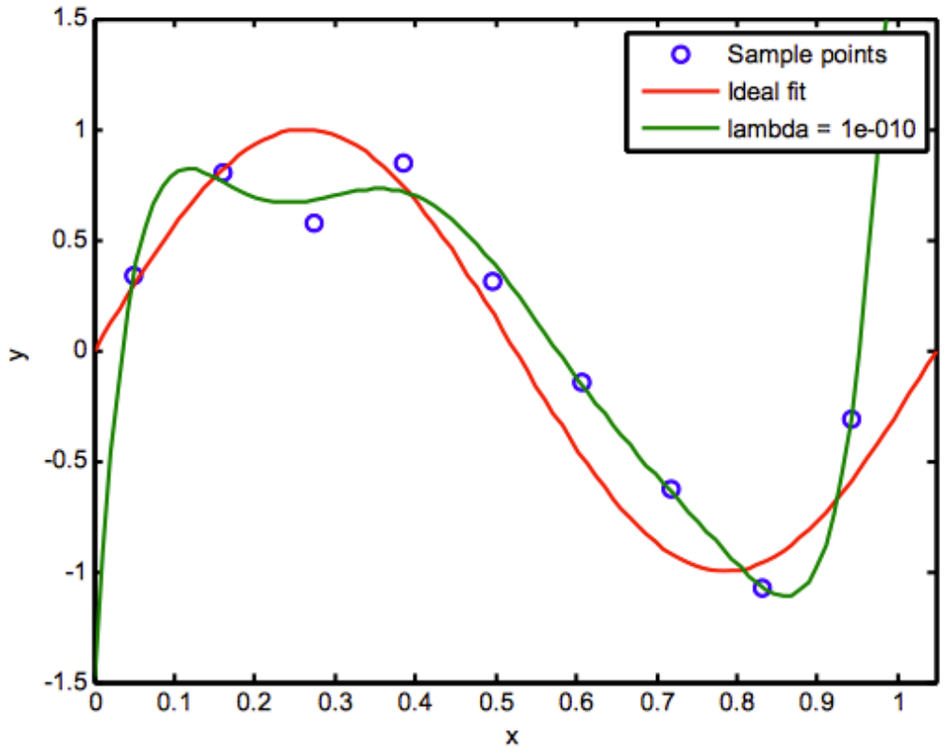
$$f(x,\theta) = z\theta \qquad z: x \to z$$

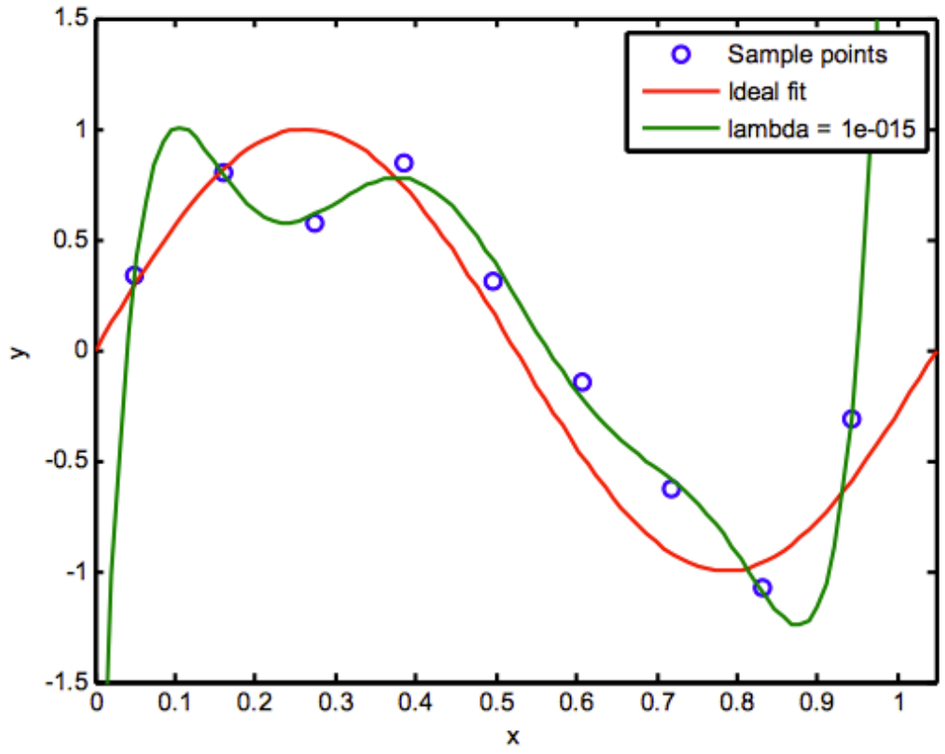
$$\tilde{E}(\theta) = \frac{1}{N}\sum_{i=1}^{n}\left(y^{\{i\}} - z^{\{i\}}\theta\right)^2 + \lambda\|\theta\|_2^2 \quad \theta \in \mathbb{R}^{D+1}$$

N = 9 samples, D = 7

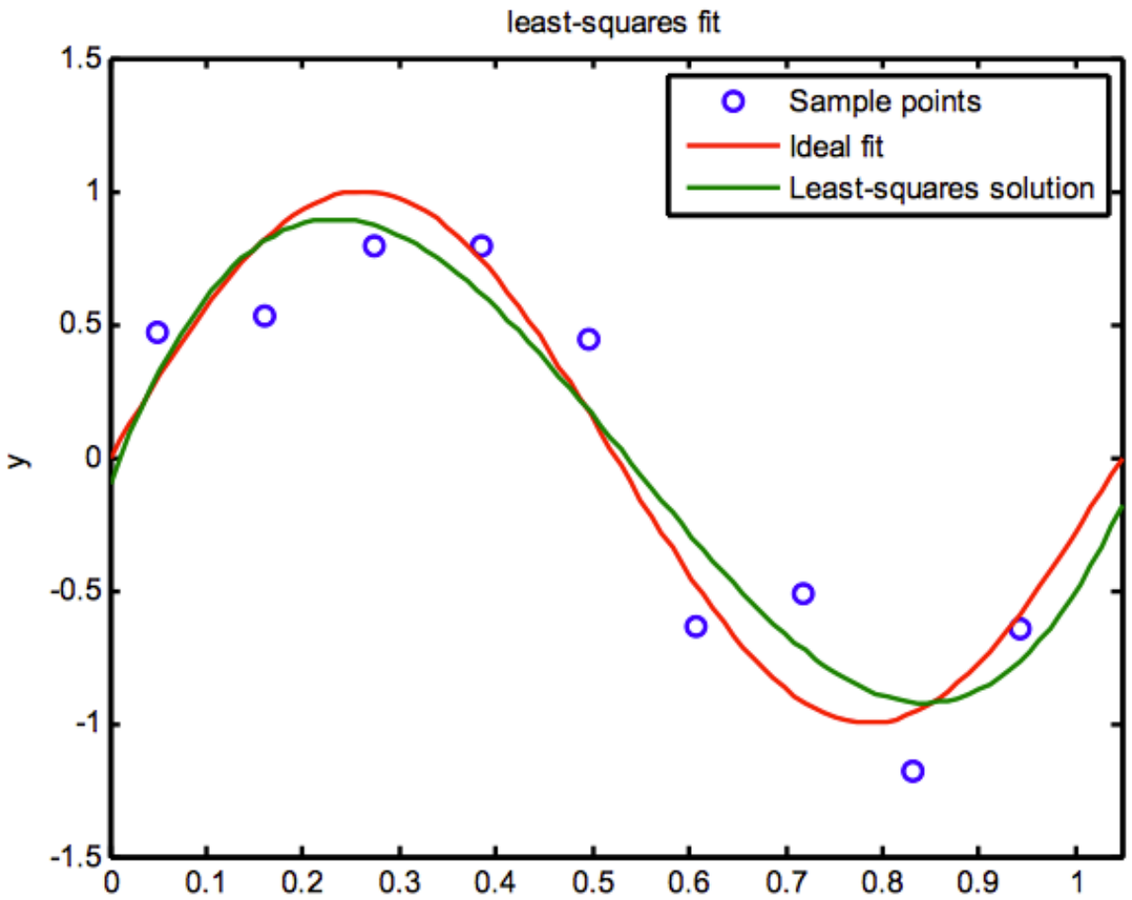




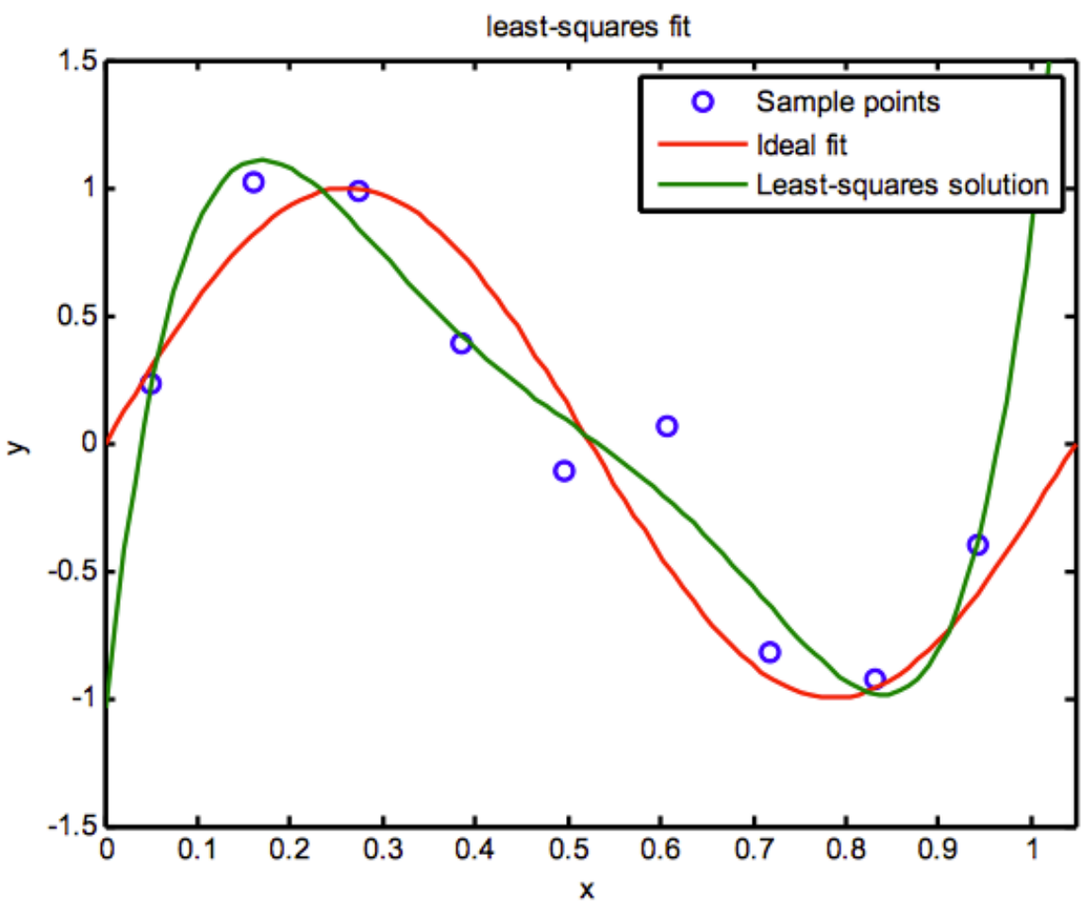




D = 3


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

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

Now let's look at another regularization choice.

The Lasso Regularization (L1 norm) and sparsity

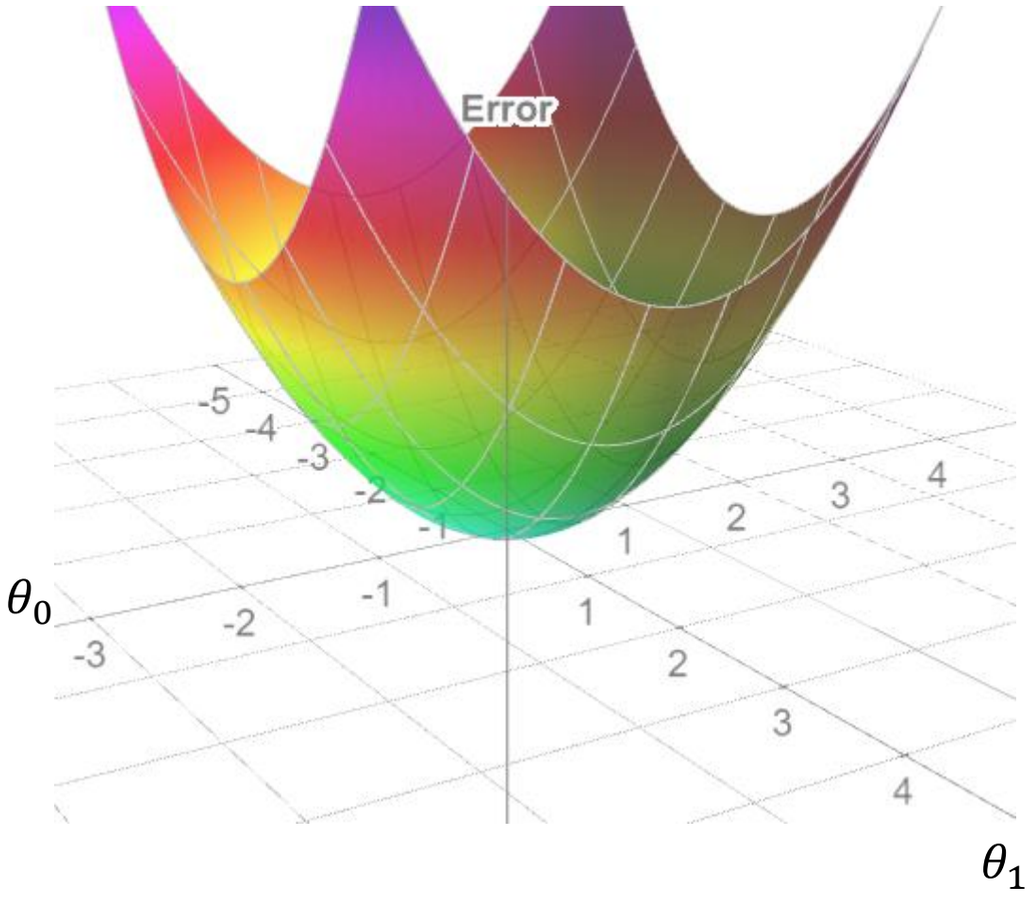
Lasso = **L**east Absolute **S**hrinkage and **S**election **O**perator

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

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

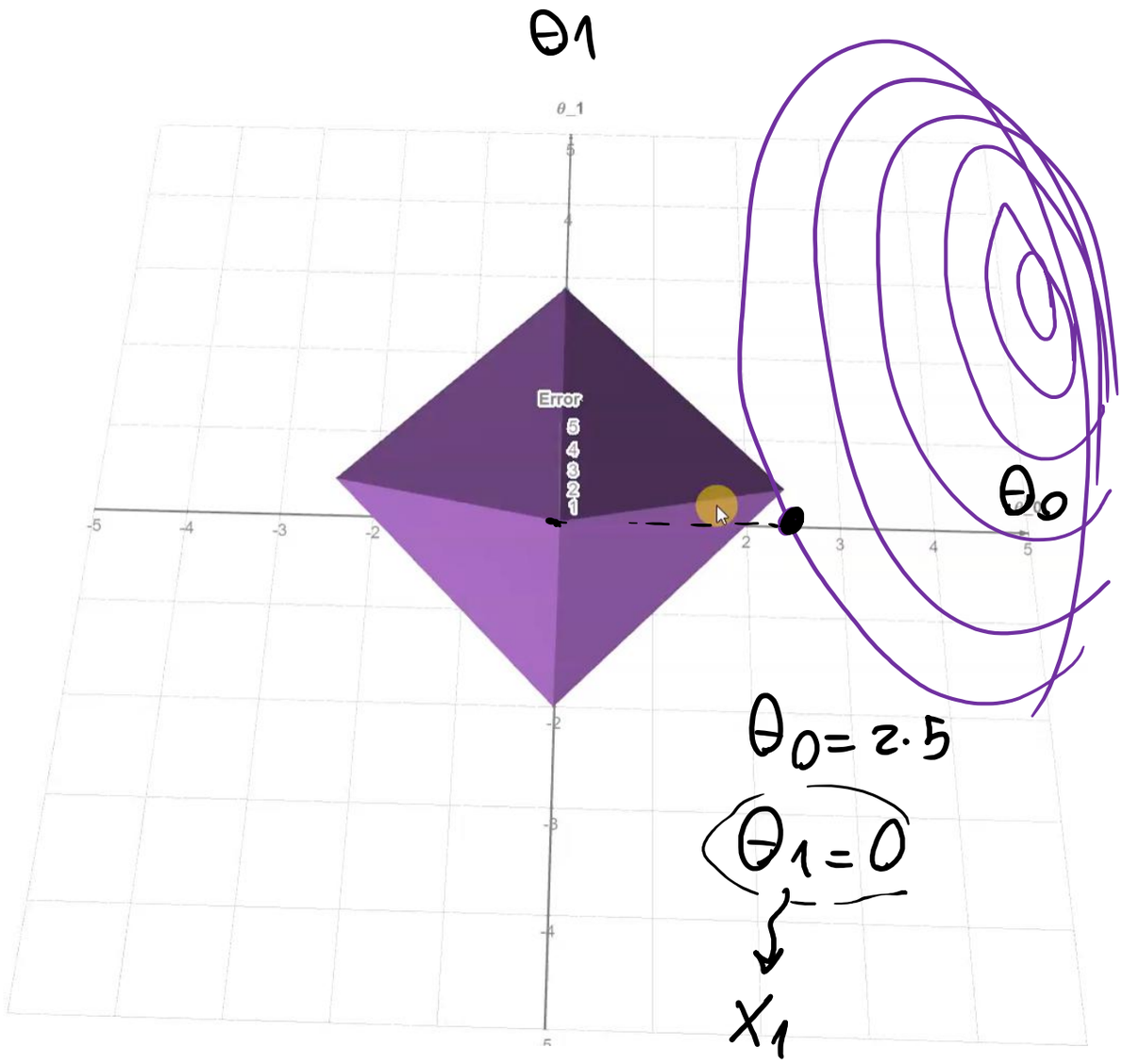
Ridge Regularizer

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



Lasso Regularizer

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



[Animation]

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

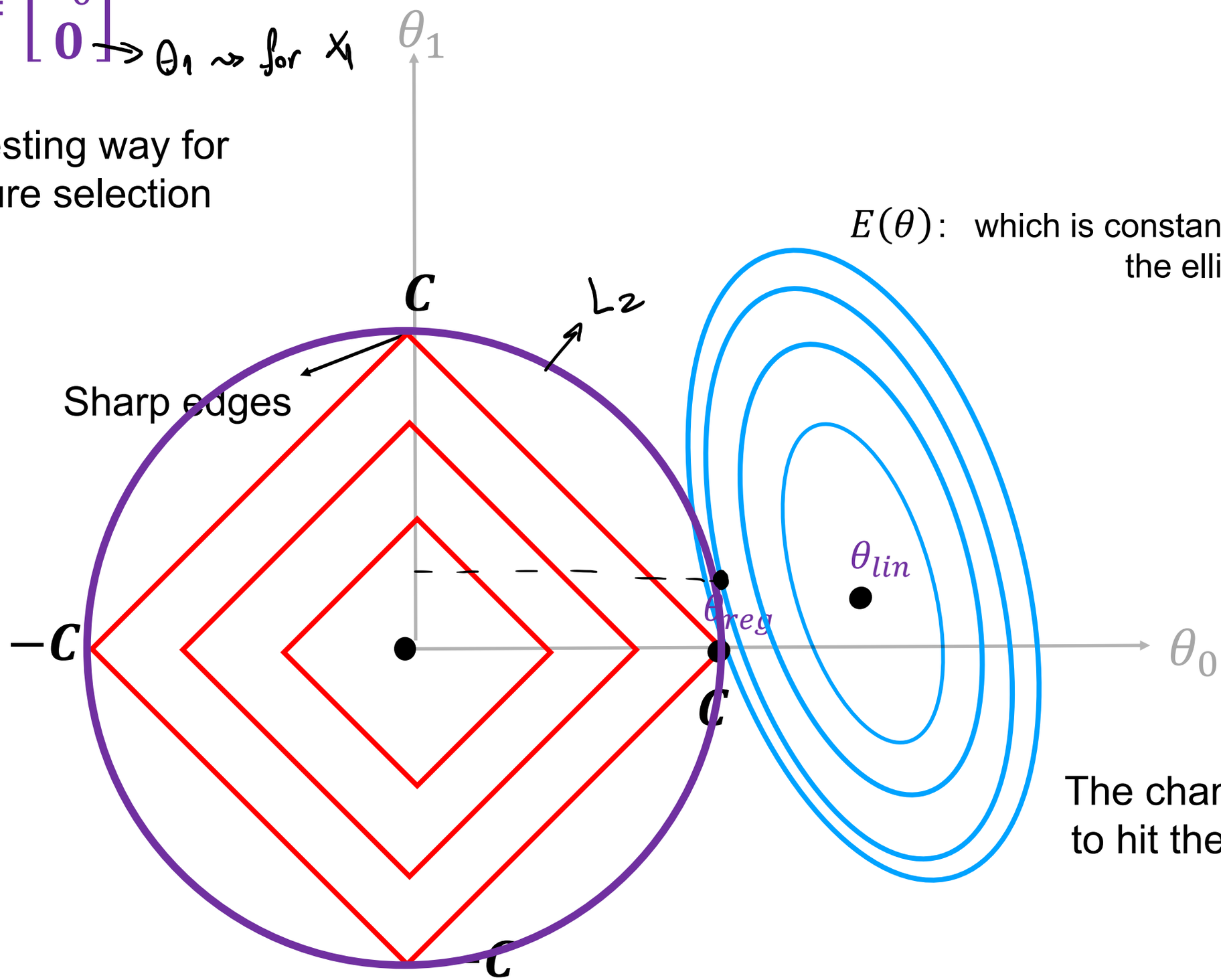
$$\theta = \begin{bmatrix} \theta_0 \\ \mathbf{0} \end{bmatrix}$$

bias term

θ_1 → for x_1

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

Interesting way for feature selection



$E(\theta)$: which is constant on the surface of the ellipsoid

The chance is very high to hit the sharp corners first

Graph

Ridge versus Lasso

Ridge

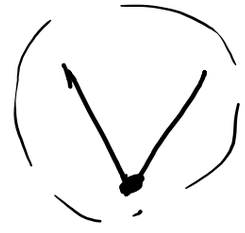
$$\tilde{E}(\theta) = \frac{1}{N}(y - z\theta)^{\mathrm{T}}(\mathrm{y} - z\theta) + \lambda\|\theta\|_2^2$$

It is a convex model

Both mean squared error and L2 regularizer are differentiable.

We can get a closed form solution

Lasso



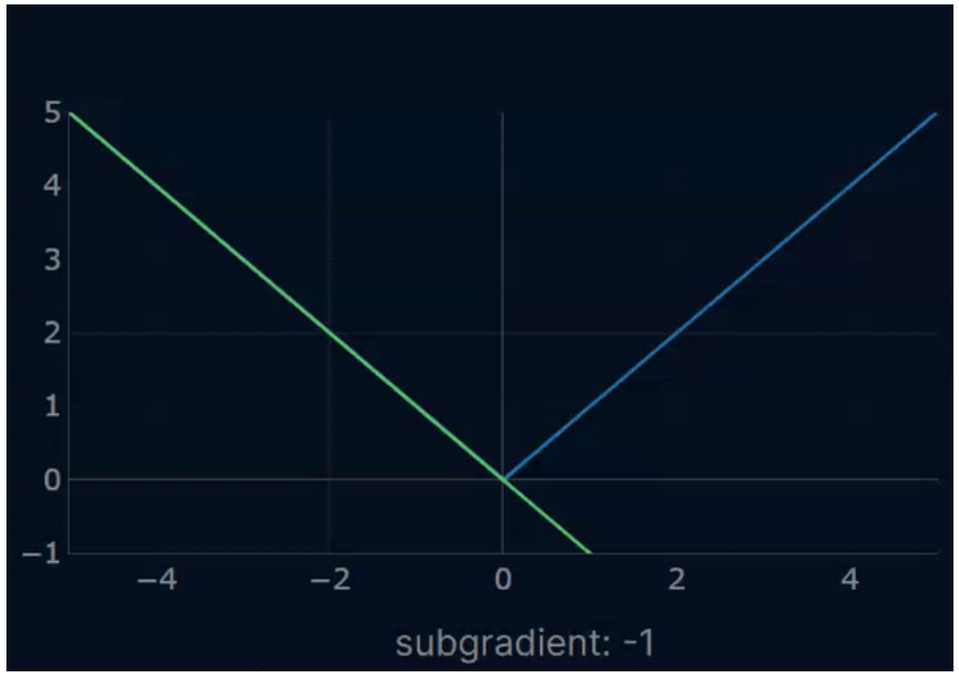
$$\tilde{E}(\theta) = \frac{1}{N}(y - z\theta)^{\mathrm{T}}(\mathrm{y} - z\theta) + \lambda\|\theta\|_1$$

It is a convex model

L1 regularizer is NOT differentiable.

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

Sub-gradient Descend in Lasso



$$\theta^{t+1} \leftarrow \theta^{t} - \alpha \frac{\partial \tilde{E}(\theta)}{\partial \theta}$$

$$\theta = \begin{bmatrix} 0.01 \\ \vdots \end{bmatrix}$$

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

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

$f(\theta) = |\theta| \xrightarrow{+} f(\theta) = \theta \Rightarrow f'(\theta) = 1$

$\searrow f(\theta) = -\theta \Rightarrow f'(\theta) = -1$

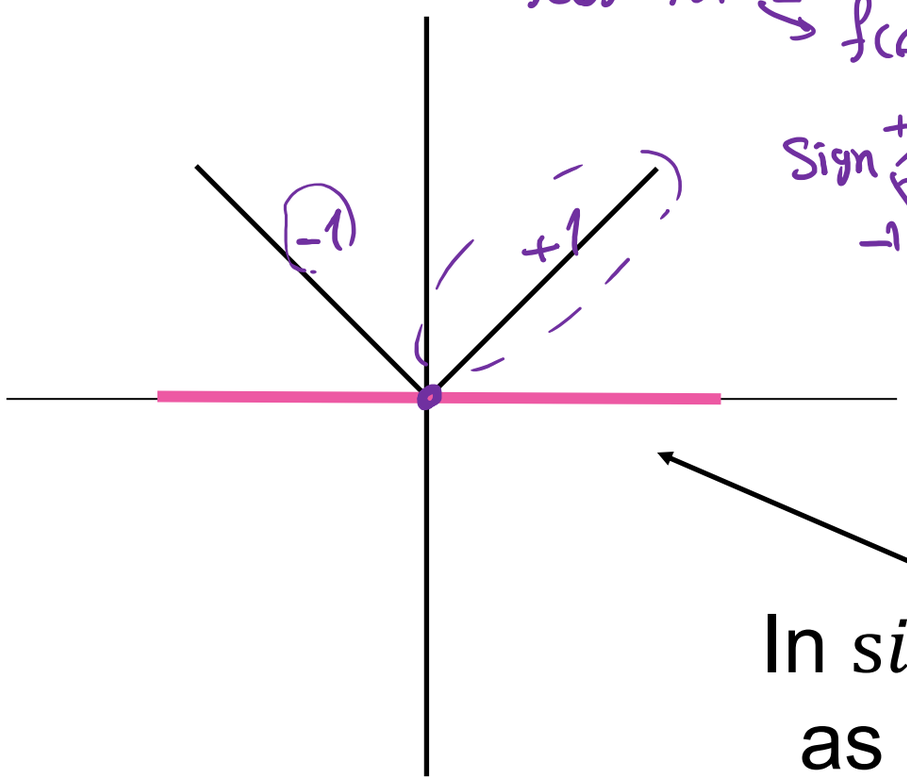
$\text{Sign}: + \rightarrow +1$, $0 \rightarrow 0$, $- \rightarrow -1$

$\frac{\partial |\theta|}{\partial \theta} = \text{Sign}(\theta)$

Using Sub-gradient

$$\frac{\partial \tilde{E}(\theta)}{\partial \theta} = -z^{T}(y - z\theta) + \lambda sign(\theta) = 0$$

$\theta = ?$



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

Outline

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

We need optimize hyper parameters

Training %70 Validation %10 Testing %20

Leave-One-Out Cross Validation

$\lambda = 0.01$ $\lambda = 0.03$ $Z_{n \times d+1}$ $Y_{n \times 1}$

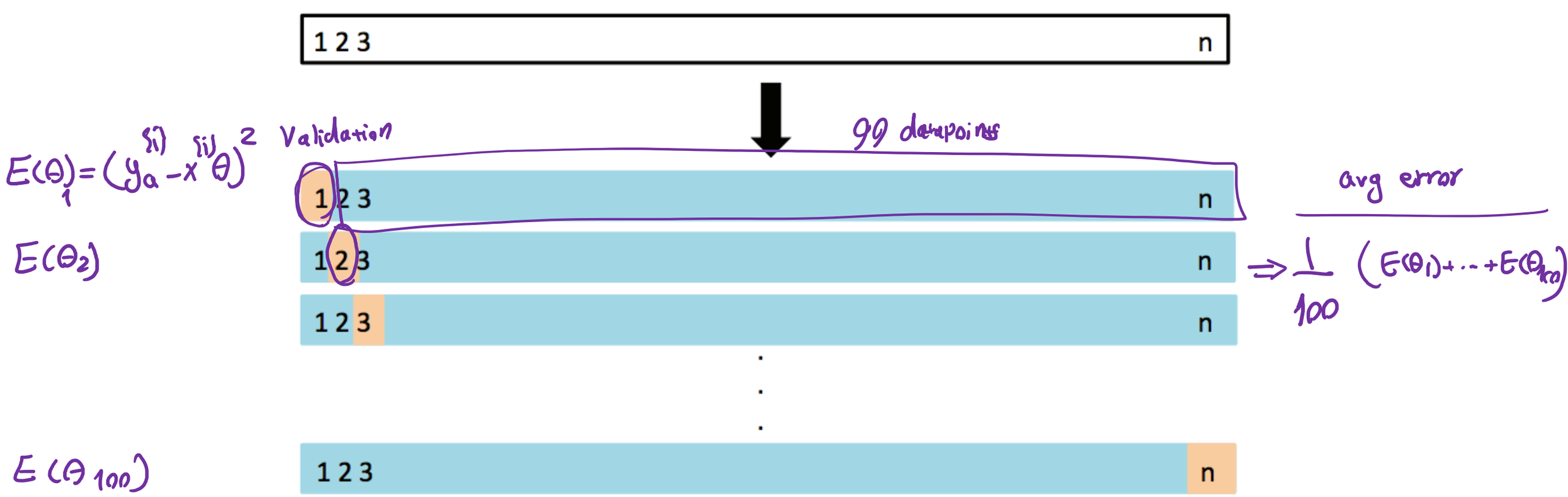
$\lambda = 0.02$

For every $i = 1, \ldots, n$: $\theta_{d+1 \times 1} = (\underbrace{Z^T Z + \lambda I}_{d+1 \times d+1})^{-1} Z^T_{d+1 \times n} Y_{n \times 1}$

$E(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_a^{[i]} - x^{[i]}\theta)^2$

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

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



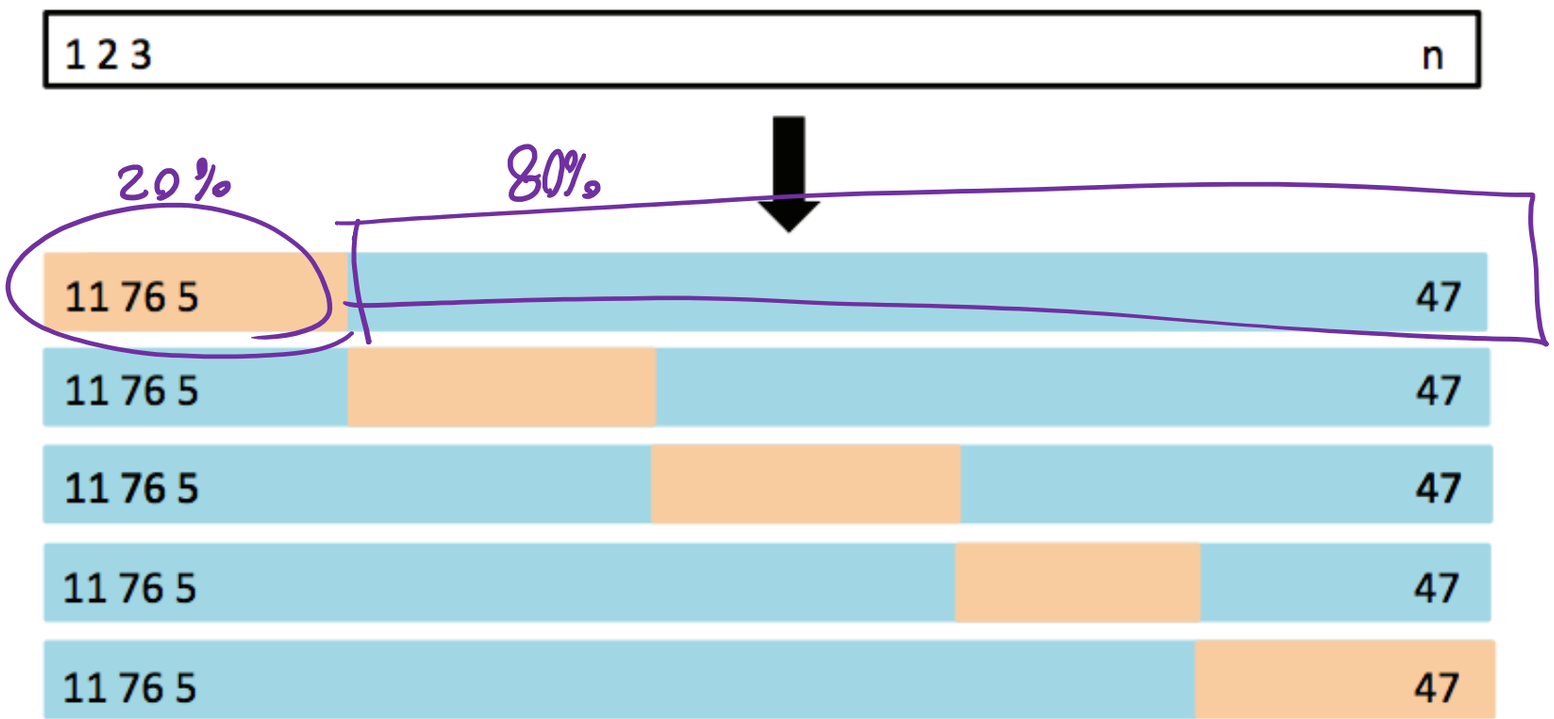
K-Fold Cross Validation

Split the data into k subsets or *folds*.

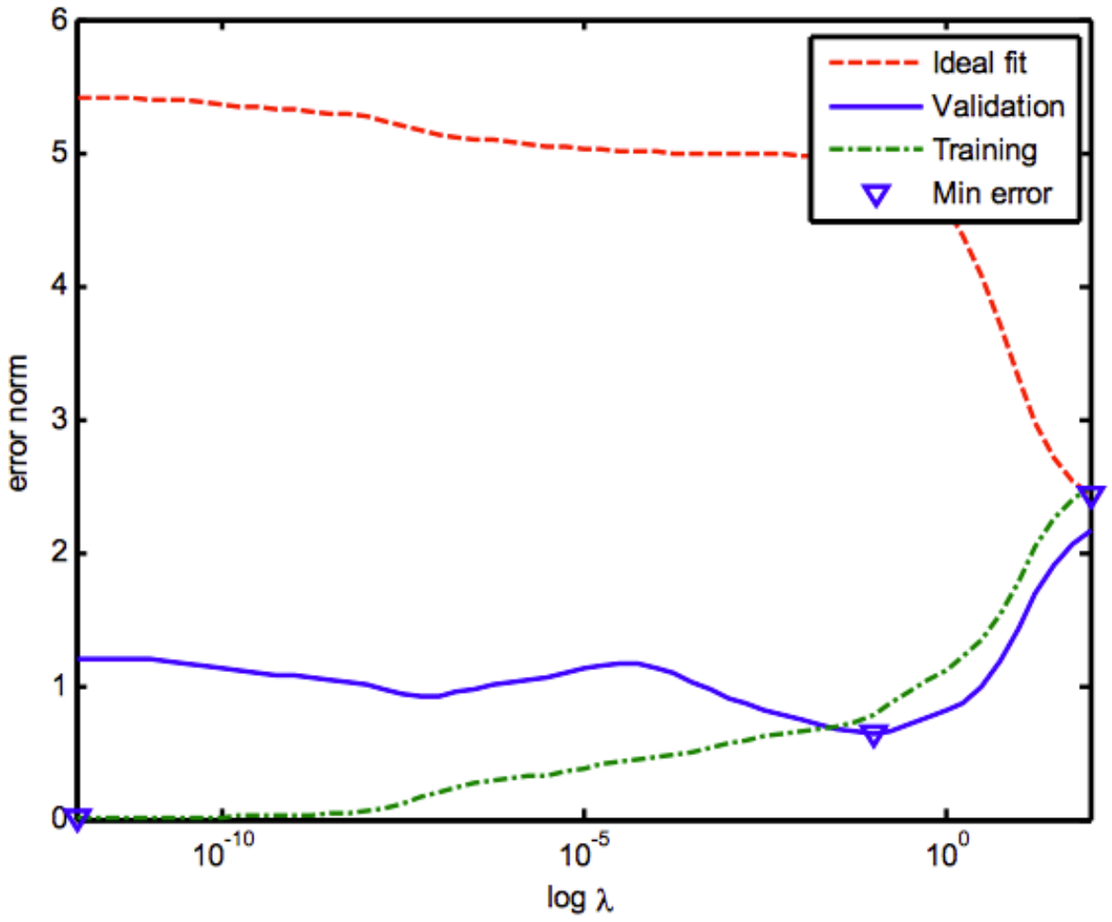
For every $i = 1, \ldots, k$:

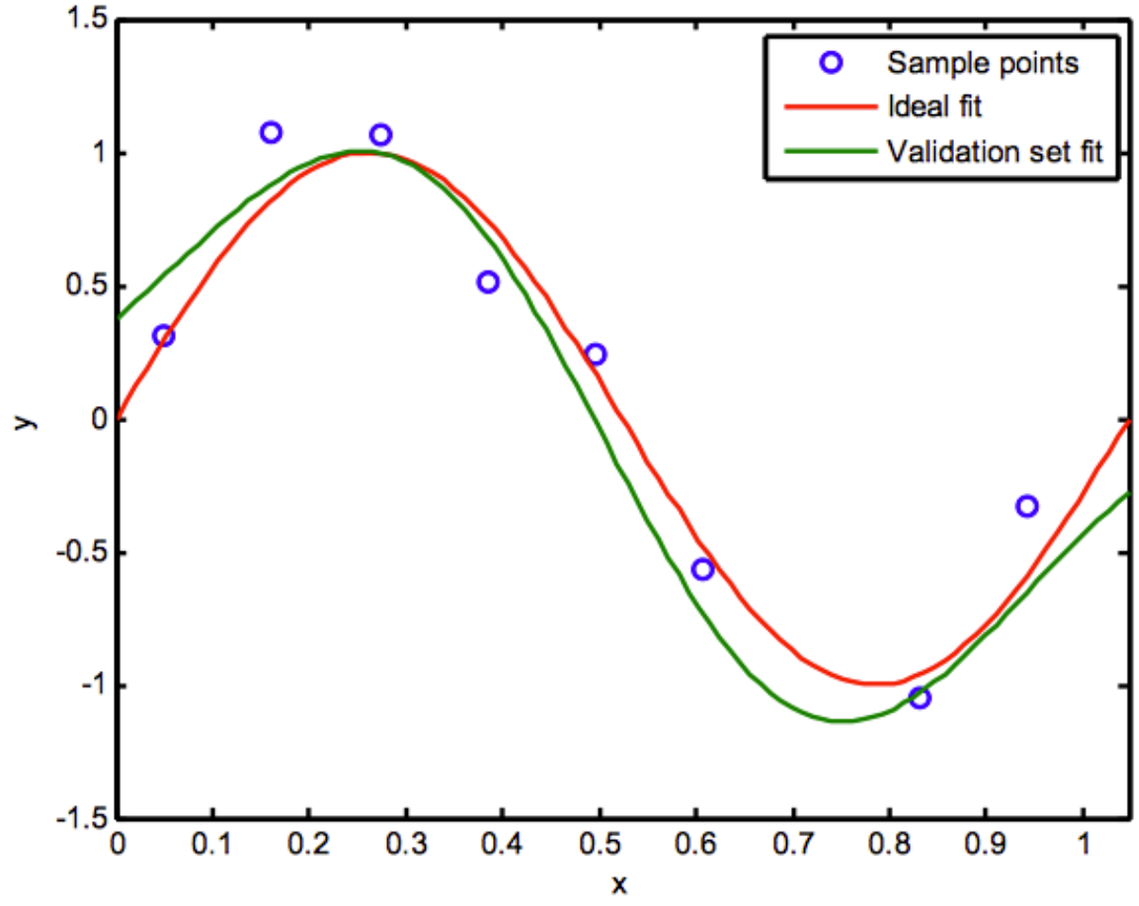
- train the model on every fold except the ith fold,
- compute the test error on the ith fold.

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