Tutorial Session 3: Linear Regression

Joanina, Ken, Augustin

The Tree of CA

- ML
  - Supervised
    - Classification
      - Binary Classification
        - NCC
        - Perceptron
        - LDA
    - Regression
      - Linear Regression
  - Unsupervised
    - Clustering
    - Dimensionality Reduction

Classification vs. Regression

Binary Classification

$f : \mathbb{R}^d \to \{0, 1\}$

$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} - \beta \geq 0$

Linear Regression

$f : \mathbb{R}^d \to \mathbb{R}$

$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} - \beta$

From the data to the linear function

- Given some data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ with respective class labels $y_1, \ldots, y_n \in \mathbb{R}$, we assume a linear relationship between $\mathbf{x}_i$ and corresponding label y_i:

$$\begin{aligned} x_{i,1}w_1 + x_{i,2}w_2 + \ldots + x_{i,d}w_d - \beta &= y_i \\ \mathbf{w}^\top \mathbf{x}_i - \beta &= y_i \end{aligned}$$

From the data to the linear function

- Given some data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ with respective class labels $y_1, \ldots, y_n \in \mathbb{R}$, we assume a linear relationship between $\mathbf{x}_i$ and corresponding label y_i:

$$\begin{aligned} x_{i,1}w_1 + x_{i,2}w_2 + \ldots + x_{i,d}w_d - \beta &= y_i \\ \mathbf{w}^\top \mathbf{x}_i - \beta &= y_i \end{aligned}$$

- this gives us an *unsolvable* set of n linear equations:

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} & 1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} & 1 \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ \beta \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

From the data to the linear function

- Given some data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ with respective class labels $y_1, \ldots, y_n \in \mathbb{R}$, we assume a linear relationship between $\mathbf{x}_i$ and corresponding label y_i:

$$\begin{aligned} x_{i,1}w_1 + x_{i,2}w_2 + \ldots + x_{i,d}w_d - \beta &= y_i \\ \mathbf{w}^\top \mathbf{x}_i - \beta &= y_i \end{aligned}$$

- this gives us an *unsolvable* set of n linear equations:

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} & 1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} & 1 \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ \beta \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- for which the *next best* solution is given by: $\mathbf{w} = (XX^\top)^{-1}X\mathbf{y}^\top$

OLS Error Function

$$\mathcal{E}_{lsq}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = ||\mathbf{y} - \mathbf{w}^\top X||^2$$

Task 1.1

Consider a data set with three data points,

$$x_1 = 0, x_2 = 1, x_3 = 2$$

with respective labels

$$y_1 = 0, y_2 = 1, y_3 = 0.$$

We want to fit a simple linear model $f(x) = w \cdot x$ to the data using ordinary least squares (OLS). Compute $\mathbf{w}$.

Polynomial Regression

$$
\begin{aligned}
h(x) &= w_1 + w_2 x + w_3 x^2 + \ldots + w_d x^{d-1} \\
&= w_1 \phi_1(x) + w_2 \phi_2(x) + w_3 \phi_2(x) + \ldots + w_d \phi_d(x) \\
&= \mathbf{w}^\top \phi(x)
\end{aligned}
$$

$$
\phi(x) : \mathbb{R} \ni x \mapsto \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^{d-1} \end{bmatrix} \in \mathbb{R}^d
$$

Polynomial Regression - The Solution

$$\Phi = [\phi(x_1), \phi(x_2), \ldots, \phi(x_n)]$$

$$\Leftrightarrow$$

$$\mathbf{w} = (\Phi\Phi^\top)^{-1}\Phi\mathbf{y}^\top$$

Task 1.2

Now we want to fit a polynomial model $g(x) = w_1 \cdot x + w_2 \cdot x^2 = \mathbf{w}^\top \cdot \boldsymbol{\phi}(x)$ where we have defined a mapping $\boldsymbol{\phi} : \mathbb{R} \mapsto \mathbb{R}^2$ with

$$\boldsymbol{\phi}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

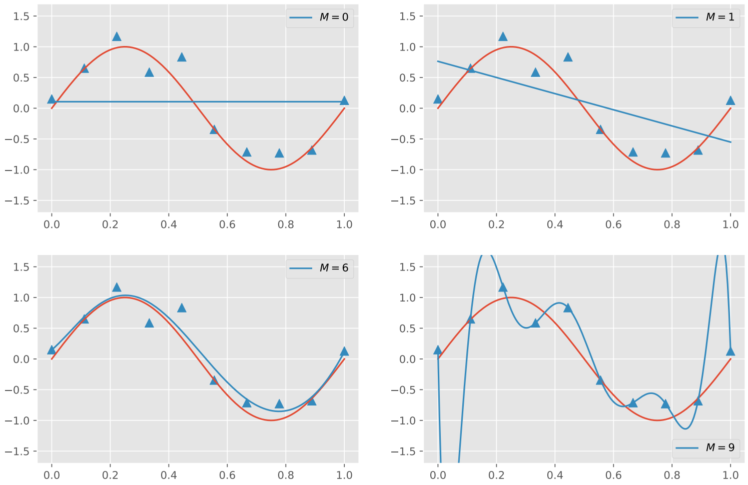
and a weight vector

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

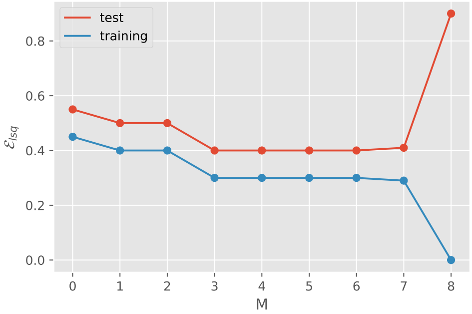
.

Compute $\mathbf{w}$.

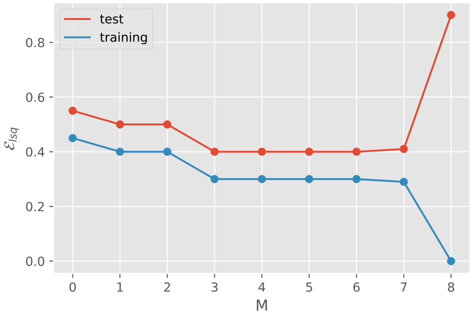
Polynomial Regression - Fitting Sine Curve



Overfitting for high degrees



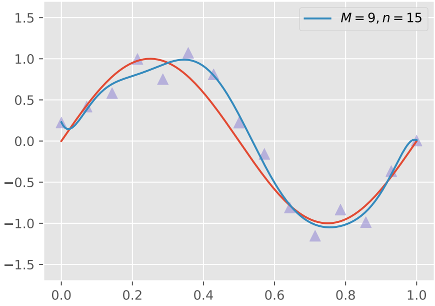
Overfitting for high degrees

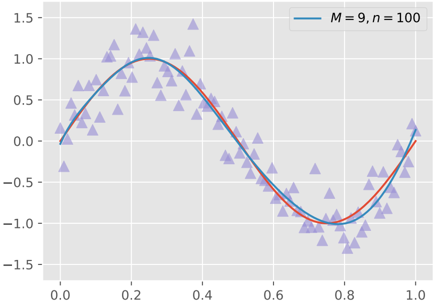


	$M=0$	$M=1$	$M=6$	$M=9$
w_0	0.11	-1.31	33.24	-102297.34
w_1		0.76	-135.49	470852.74
w_2			207.87	-911589.91
w_3			-129.08	963843.87
w_4			19.42	-604186.44
w_5			4.01	227748.07
w_6			0.15	-49782.98
w_7				5656.36
w_8				-244.40
w_9				0.15

Increasing number of samples

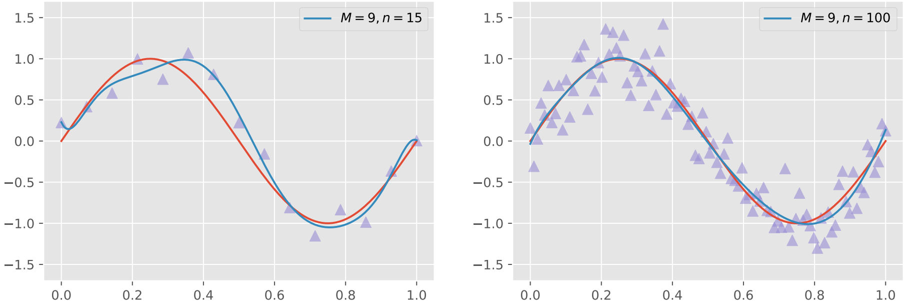
- Increasing the size of the data set reduces the overfitting problem





Increasing number of samples

- Increasing the size of the data set reduces the overfitting problem



$\Rightarrow$ But usually the number of samples is limited!

Ridge Regression

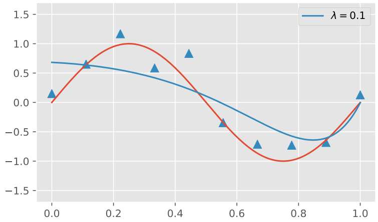
- Regression with penalization: restrict large values for $\mathbf{w}$
- Often it is important to control the complexity of the solution $\mathbf{w}$

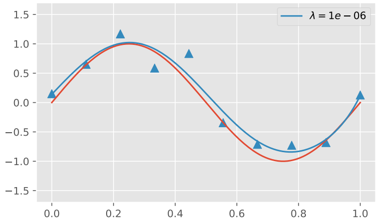
$$\mathcal{E}_{RR}(\mathbf{w}) = ||y - \mathbf{w}^\top X||^2 + \lambda ||w||^2$$

- Solution is given by

$$\mathbf{w} = (XX^\top + \lambda I)^{-1} X \mathbf{y}^\top$$

Ridge Regression - Fitting Sine Curve





- $M = 9$

	$\lambda = 0.1$	$\lambda = 10^{-6}$	$\lambda = 0$
w_0	0.83	28.58	-102297.34
w_1	0.69	-34.35	470852.74
w_2	0.51	-26.34	-911589.91
w_3	0.26	13.48	963843.87
w_4	-0.07	36.93	-604186.44
w_5	-0.48	12.88	227748.07
w_6	-0.94	-36.35	-49782.98
w_7	-1.21	-0.41	5656.36
w_8	-0.27	5.55	-244.40
w_9	0.68	0.14	0.15

Task 4

1. Show that $\hat{\mathbf{y}}_{\mathsf{OLS}}$ is invariant under arbitrary transformations A, but $\hat{\mathbf{w}}_{\mathsf{OLS}}$ is not.

Task 4

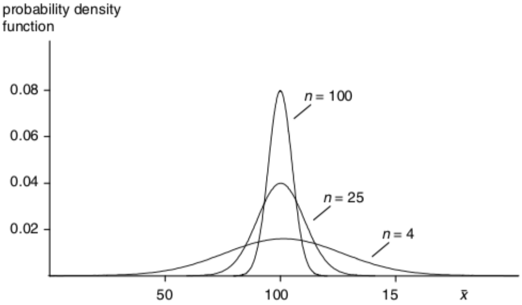
1. Show that $\hat{\mathbf{y}}_{\mathsf{OLS}}$ is invariant under arbitrary transformations A, but $\hat{\mathbf{w}}_{\mathsf{OLS}}$ is not.

2. Show that $\hat{\mathbf{y}}_{\mathsf{RR}}$ is invariant under orthogonal transformations A.

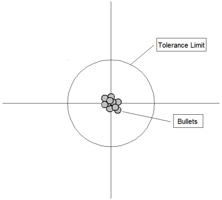
Bias-Variance Tradeoff - Simple Estimators

- Estimators are themselves random variables

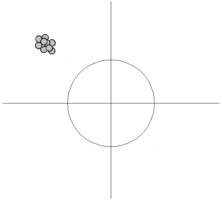
$\Rightarrow$ They have their own distribution!



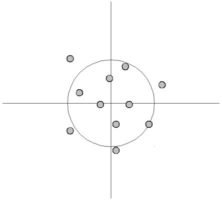
Bias-Variance Tradeoff Visualized



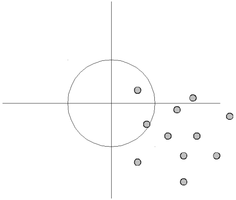
low bias, low variance



high bias, low variance

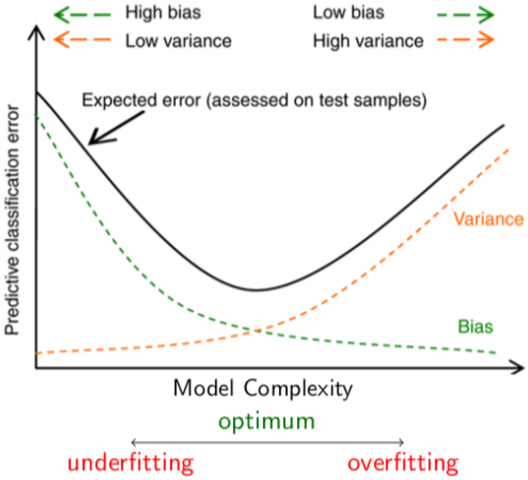


low bias, high variance

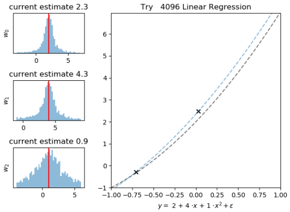


high bias, high variance

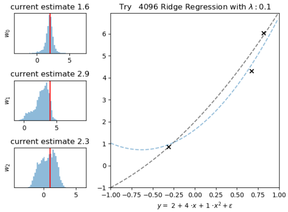
Bias-Variance Tradeoff



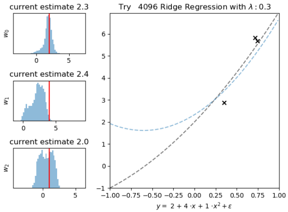
Bias and Variance in (Regularized) Linear Regression



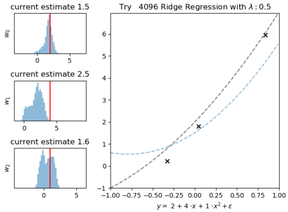
$\lambda = 0.0$



$\lambda = 0.1$



$\lambda = 0.3$



$\lambda = 0.5$

Task 2

1. If the number of data points n increases, the variance of $\hat{\mathbf{w}}$ will

   (a) decrease (b) increase (c) remain the same.

Task 2

1. If the number of data points n increases, the variance of $\hat{\mathbf{w}}$ will

   (a) decrease (b) increase (c) remain the same.

2. If the noise variance σ_ϵ^2 increases, the variance of $\hat{\mathbf{w}}$ will

   (a) decrease (b) increase (c) remain the same.

Task 2

1. If the number of data points n increases, the variance of $\hat{\mathbf{w}}$ will
   (a) decrease (b) increase (c) remain the same.
2. If the noise variance σ_ϵ^2 increases, the variance of $\hat{\mathbf{w}}$ will
   (a) decrease (b) increase (c) remain the same.
3. If the data variance σ_x^2 increases, the variance of $\hat{\mathbf{w}}$ will
   (a) decrease (b) increase (c) remain the same.

Task 2

1. If the number of data points n increases, the variance of $\hat{\mathbf{w}}$ will

   (a) decrease (b) increase (c) remain the same.

2. If the noise variance σ_ϵ^2 increases, the variance of $\hat{\mathbf{w}}$ will

   (a) decrease (b) increase (c) remain the same.

3. If the data variance σ_x^2 increases, the variance of $\hat{\mathbf{w}}$ will

   (a) decrease (b) increase (c) remain the same.

4. If the true slope w increases, the variance of $\hat{\mathbf{w}}$ will

   (a) decrease (b) increase (c) remain the same.

Task 3

1. Draw a sketch showing two curves: training error vs. the number of features m and test error vs. the number of features m.

Task 3

1. Draw a sketch showing two curves: training error vs. the number of features m and test error vs. the number of features m.
2. Annotate the plot with the two terms "Overfitting" and "Underfitting"

Task 3

1. Draw a sketch showing two curves: training error vs. the number of features m and test error vs. the number of features m.
2. Annotate the plot with the two terms "Overfitting" and "Underfitting"
3. Draw two more curves in a second sketch: The bias of $\hat{f}$ and the variance of $\hat{f}$ against the number of features m.

Task 3

4. Suppose we choose m such that we are in the "overfitting" region, but we use ridge regression with a (good) regularisation parameter $\lambda > 0$. Compared to OLS,
   (a) will the training error decrease,increase or is it ambigious?
   (b) will the test error decrease, increase or is it ambigious?
   (c) will the bias of $\hat{f}$ decrease, increase or is it ambigious?
   (d) will the variance of $\hat{f}$ decrease, increase or is it ambigious?