5/18/23

Lecture 16

Multivariate Gaussian
Linear regression
D. D.

Posted
LVIII Linear Regression
- Double descent (soon)
- for regression
Poll Results !!

Multivariate Normal

$$S = \mathbb{R}^d \ni \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

Bivariate $d=2$

$$S = \mathbb{R}^2 \quad \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\Big(\underbrace{\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}}_{\vec{\mu} \in \mathbb{R}^2}, \overbrace{\Sigma}^{\text{covariance}}\Big) \iff x, y \text{ jointly Normal}$$

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

σ_x^2: Var X, σ_{xy}: Cov(X,Y), σ_y^2: Var Y

symmetric

$(\Sigma \geq 0)$ $\quad \Sigma > 0$

positive definite

$\iff |\Sigma| > 0$

$$f_{xy}(x,y) = \frac{1}{\underbrace{2\pi|\Sigma|^{1/2}}_{Z}} e^{-\frac{1}{2}[x-\mu_x,\ y-\mu_y]\Sigma^{-1}\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}}$$

like $\sim \frac{1}{\sigma^2}$

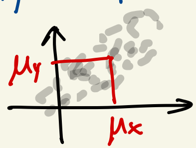
$$\boxed{v^T}\ \boxed{\Sigma^{-1}}\ \boxed{v} = \square \in \mathbb{R}$$

ok in the limit

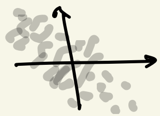
$$v^T \Sigma^{-1} v \geq 0 \quad \text{when } \Sigma > 0$$

$$\text{Cov}(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = 0 \quad \text{if } X, Y \text{ independent}$$

$$> 0$$



$$< 0$$



Correlation coefficient $\quad \rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

Multivariate $N(\mu, \Sigma)$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \longrightarrow \mu_i = E[X_i]$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ \sigma_{ij} & \sigma_2^2 & \\ & & \ddots \; \sigma_d^2 \end{bmatrix}$$

with $\Sigma > 0 \iff v^T \Sigma v > 0$ for any $v \neq 0$

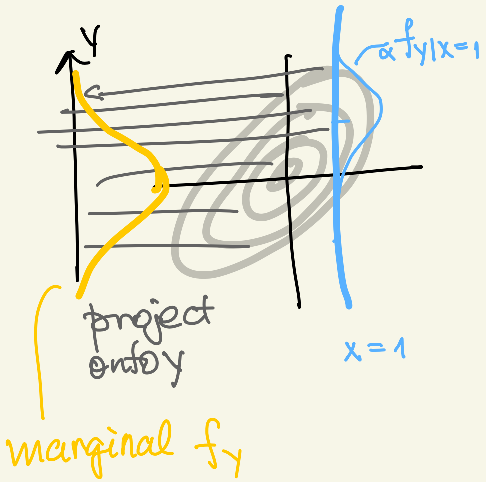
$\sigma_{ij} = \sigma_{ji} = \mathrm{Cov}(X_i, X_j)$

$\sigma_i^2 = \mathrm{Var}\, X_i$

$$f_X = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

joint of $(X_1, \dots X_d)$

$$\boxed{x-\mu} \; \boxed{\Sigma^{-1}} \; \boxed{x-\mu} = \boxed{\;} \geq 0$$



$\sigma_{ij} = 0$ for all $i \neq j$:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix} \iff X_{1,2,\dots,d} \text{ mutually independent}$$

Lecture Notes I – Regression

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

May, 2023

Prediction problems

Linear regression

Linear regression for non-linear f

Reading: Ch.

Prediction

- **Data** $\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$
- **Inputs** $x^{1:n} \in \mathbb{R}$, or $\mathbb{R}^d$
- **Outputs** $y^{1:n}$

- **Goal** Learn/estimate $f(x)$ **predictor** for y

- By type of output
  - **Classification** if $y \in S_Y$ discrete
  - $y \in \{0, 1\}$ (or $\{\pm 1\}$) binary classification
  - $y \in \{1, 2, \ldots m\}$ multiclass classification

  - **Regression** if $y \in \mathbb{R}$ continuous

Prediction

- **Data** $\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$
- **Inputs** $x^{1:n} \in \mathbb{R}$, or $\mathbb{R}^d$
- **Outputs** $y^{1:n}$

- **Goal** Learn/estimate $f(x)$ **predictor** for y

- By type of output
  - **Classification** if $y \in S_Y$ discrete
  - $y \in \{0, 1\}$ (or $\{\pm 1\}$) binary classification
  - $y \in \{1, 2, \ldots m\}$ multiclass classification

  - **Regression** if $y \in \mathbb{R}$ continuous

- **Model family** $\mathcal{F} = \{f_\theta : \mathbb{R}^d \to S_Y, \theta \in \Theta\}$
  - $\mathcal{F} = \{$ linear functions $\}$ linear regression/classification
  - $\mathcal{F} = \{$ polynomes of degree 2, 3, $\ldots\}$ polynomial regression/classification
  - $\mathcal{F} = \{\frac{1}{1+e^{-\beta^T x + \beta_0}}\}$ logistic regression
  - neural network regression/classification
  - kernel regression/classification
  - $\mathcal{F} = \{$ monotonic functions $\}$ isotonic regression
  - support vector regression/classification
  - regression/classification *trees* (and random forests)

Linear regression → predict y from x

$\begin{cases} x \in \mathbb{R} \\ x \in \mathbb{R}^d \end{cases}$

Model is linear

- $y^i = \beta_0 + \beta_1 x^i + \epsilon^i$ for $x^{1:n} \in \mathbb{R}$ (univariate regression)
- $y^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \ldots \beta_d x_d^i + \epsilon^i$ for $x^{1:n} \in \mathbb{R}^d$ (multivariate regression)

  $y \in \mathbb{R}$ is output/response

  $x_j^{1:n}$ for $j = 1 : d$ are input(s)/covariates/features/attributes/...

  $\beta_{1:d}$ are **regression coefficients**, β_0 is intercept

  $\epsilon^{1:n} \in \mathbb{R}$ is noise, $\epsilon^{1:n} \sim N(0, \sigma^2)$ i.i.d.

$$y^i = \beta_0 + \beta_1 x_1^i + \cdots + \beta_d x_d^i + \underbrace{\varepsilon^i}_{\text{noise}} \leftarrow \text{Model}$$

$\varepsilon^i \sim N(0, \sigma^2)$ iid

Data $\mathcal{D} = \{(x^i, y^i)\}_{i=1:n}$

Wanted: parameters β_0, $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$, σ^2

Linear regression model in matrix form

For a single data point (x^i, y^i)

$$\underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_d \end{bmatrix} \in \mathbb{R}^d \qquad \underline{x^i} = \begin{bmatrix} x^i_1 \\ x^i_2 \\ \dots \\ x^i_d \end{bmatrix} \in \mathbb{R}^d \tag{1}$$

Then,

$$y^i = \beta_0 + (x^i)^T \beta + \epsilon^i \tag{2}$$

$y = \beta_0 + \beta^T x + \varepsilon$

For all $\mathcal{D}$

$$y = \begin{bmatrix} y^1 \\ y^2 \\ \dots \\ y^n \end{bmatrix} \in \mathbb{R}^n \qquad \mathrm{X} = \begin{bmatrix} (x^1)^T \\ (x^2)^T \\ \dots \\ (x^n)^T \end{bmatrix} \in \mathbb{R}^{n \times d} \qquad \epsilon = \begin{bmatrix} \epsilon^1 \\ \epsilon^2 \\ \dots \\ \epsilon^n \end{bmatrix} \tag{3}$$

$$y = \beta_0 1 + \mathrm{X}\beta + \epsilon \qquad \text{with } Cov(\epsilon) = \sigma^2 I_d \tag{4}$$

$\mathbb{R}^n \ni y^i = \beta_0 \; + \; \varepsilon^i \; + \; \overset{d}{(x^i)^T} \; \beta \; d \iff \boxed{y = \beta_0 1 + \varepsilon + X\beta}$

Solution by Maximum Likelihood

- **Likelihood** Probability $y \,|\, \mathrm{X}$, parameters
- Parameters $\beta_0, \beta_{1:d}, \sigma^2$
- $\beta_0^{ML}, \beta_{1:d}^{ML}, (\sigma^2)^{ML} = \underset{\beta_0, \beta_{1:d}, \sigma^2}{\mathrm{argmax}}\, l(y|\mathrm{X},\, \beta_0, \beta_{1:d}, \sigma^2)$

The (log)-likelihood

- What is random? **the noise** $\epsilon^{1:n}$
- Express noise as function of $(x^{1:n}, y^{1:n})$

for all i :
$$\epsilon^i = y^i - \beta_0 - \beta^T x^i \quad \sim N(0, \sigma^2) \tag{5}$$

- Likelihood
- Let $p_{0,\sigma^2}(\epsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\epsilon^2}{2\sigma^2}} = N(\epsilon; 0, \sigma^2)$
- Then

$$L(\beta_0, \beta_{1:d}, \sigma^2) = \prod_{i=1}^{n} p_{0,\sigma^2}(\epsilon^i) \tag{6}$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\epsilon^i)^2}{2\sigma^2}} \tag{7}$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y^i - \beta_0 - \beta^T x^i)^2}{2\sigma^2}} \tag{8}$$

ϵ^i

- **log-likelihood**

$$l(\beta_0, \beta_{1:d}, \sigma^2) = \tag{9}$$

$$= \sum_{i=1}^{n} \left\{ -\frac{1}{2} \ln \sigma^2 - \frac{1}{2} \ln(2\pi) - \frac{1}{2} \left(y^i - \beta_0 - \beta^T x^i\right)^2 \frac{1}{2\sigma^2} \right\} \tag{10}$$

$$= -\frac{n}{2} \ln \sigma^2 - \frac{1}{2} \ln(2\pi) + \text{constant} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^i - \beta_0 - \beta^T x^i\right)^2 \tag{11}$$

Maximizing the log-likelihood w.r.t β

Calculus

X^+ d n

$X^+X = I_d$

$XX^+ \neq I_n$

- For simplicity, let $\beta_0 = 0$; hence $y^i = \beta^T x^i + \epsilon^i$
- log-likelihood

$$l(\beta_{1:d}, \sigma^2) \;=\; -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^i - \beta_0 - \beta^T x^i\right)^2 + \text{constant} \tag{12}$$

- For any σ^2,

$$\operatorname*{argmax}_{\beta} l(\sigma^2, \beta) \;=\; \operatorname*{argmin}_{\beta} \underbrace{\sum_{i=1}^{n}\left(y^i - \beta_0 - \beta^T x^i\right)^2}_{=(y-X\beta)^T(y-X\beta)} \tag{13}$$

a **Least Squares Problem**

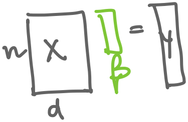
- In matrix form $\min_\beta \|y - X\beta\|^2$ $= (y - X\beta)^T(y - X\beta)$
- Solution

$$\boxed{\beta^{ML} = (X^TX)^{-1}X^Ty} \tag{14}$$

X^+ pseudo inverse

with $(X^TX)^{-1}X^T \equiv X^\dagger$ the pseudoinverse of X
- β^{ML} is linear in y!

≈ solve $X\beta = y$ with error



⇒ solution $\beta = X^+ y$

Statistical properties of β^{ML}

- Assume the true model is $y^i = \beta^T x^i + \epsilon^i$ with $\epsilon \sim N(0, \sigma^2)$.
- Here β, σ^2 are true parameters and we assume we know them.
-
- β^{ML} is a random variable. Let us calculate its mean and standard deviation.
- **Expectation** of β^{ML}

$$
\begin{aligned}
E[\beta^{ML}] &= E[X^\dagger y] = E[\mathrm{X}^\dagger(\mathrm{X}\beta + \epsilon)] && (15)\\
&= E[\mathrm{X}^\dagger \mathrm{X}]\beta + E[X^\dagger \epsilon)] && (16)\\
&= \underbrace{\mathrm{X}^\dagger \mathrm{X}}_{I_d}\beta + X^\dagger \underbrace{E[\epsilon)]}_{0} = \beta && (17)
\end{aligned}
$$

Hence, $E[\beta^{ML}] = \beta$ and we say that β^{ML} is **unbiased**

- **Covariance** of β^{ML}
- By a similar (but longer) calculation, we obtain that

$$Cov(\beta^{ML}) = \sigma^2(\mathrm{X}\mathrm{X}^T)^{-1} \in \mathbb{R}^{d \times d} \tag{18}$$

- In fact, β^{ML} has a Normal distribution. Remember from (15)

$$\beta^{ML} = \beta + X^\dagger \epsilon. \tag{19}$$

This is a linear transformation of ϵ, a Gaussian variable. Therefore,

$$\beta^{ML} \sim N(\beta, \sigma^2(\mathrm{X}\mathrm{X}^T). \tag{20}$$

- This is a **multivariate Normal** distribution over $\mathbb{R}^d$,

Estimation of σ^2

- Let $\boxed{\hat{\epsilon}^i = y^i - (x^i)^T \beta^{ML}}$, for $i = 1 : n$
- $\hat{\epsilon}^i$ called the residuals when β^{ML} used

- If we plug in β^{ML} in the log-likelihood equation (12) we obtain

like $(x^i - \hat{\mu})^2$ for $N(\mu, \sigma^2)$

$$l(\sigma^2, \beta^{ML}) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\hat{\epsilon}^i)^2 + \text{constant} \quad (21)$$

known known

- Maximizing this expression w.r.t. σ^2 gives

$$\boxed{(\sigma^2)^{ML} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\epsilon}^i)^2} \quad (22)$$

- The predictor is $f(x) = x^T \beta^{ML}$
- Hence, for a new $x \in \mathbb{R}^d$, our guess of y is $\hat{y} = f(x)$

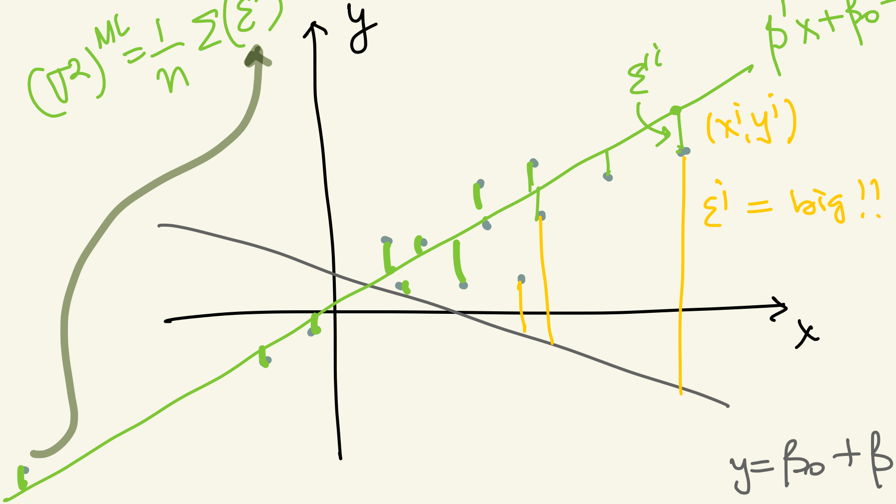
new x : $f(x) = (\beta^{ML})^T x$ $\leftarrow$ guess for new y !

$(\sigma^2)^{ML} = \frac{1}{n}\sum(\varepsilon^i)^2$

$\beta' x + \beta_0' = y \longrightarrow$ high likelihood maximum!

$x \in \mathbb{R}$

$\beta, \beta_0 \in \mathbb{R}$



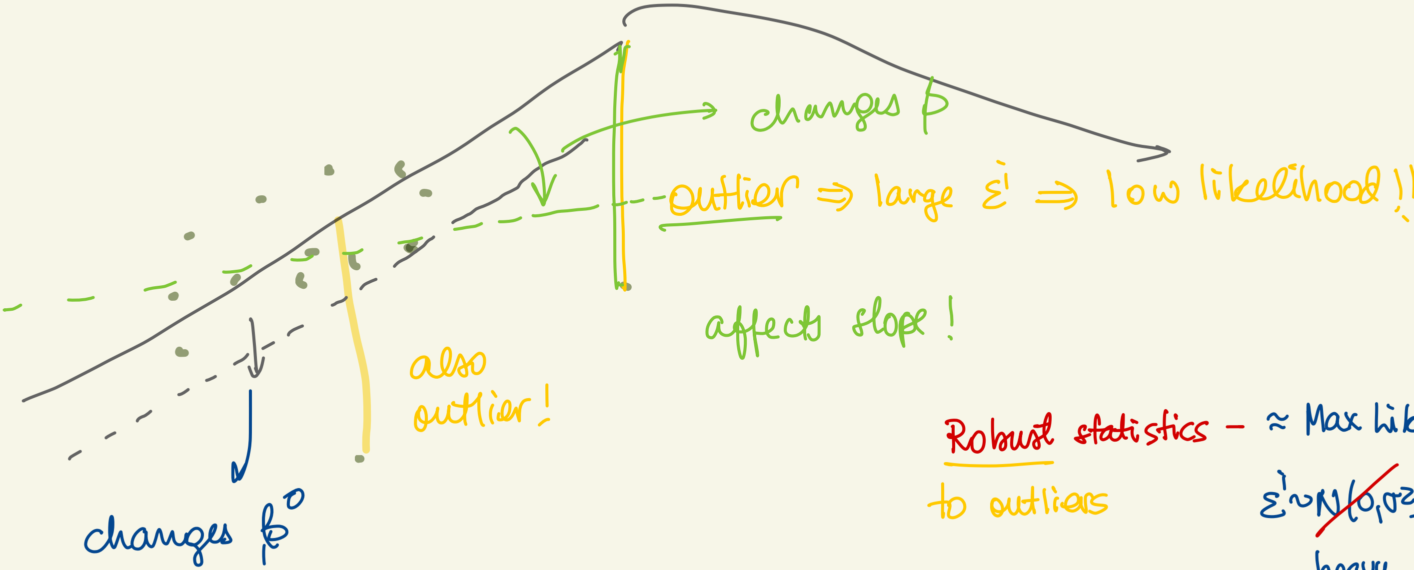
$y = \beta_0 + \beta x \longrightarrow$ low likelihood!

$$\begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_0 \end{bmatrix} + \underbrace{\begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix}}_{X} [\beta] + \begin{bmatrix} \varepsilon^1 \\ - \\ \vdots \\ \varepsilon^n \end{bmatrix}$$

$$\boxed{\beta^{ML} = X^+ y = \frac{1}{\sum (x^i)^2} [x^1 \dots x^n] \begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix} = \frac{1}{(\sum x^i)^2} \sum_{i=1}^{n} x^i y^i}$$

if $\beta_0 = 0$:

$$X^+ = (X^T X)^{-1} X = \left(\sum_{i=1}^{n} (x^i)^2\right)^{-1} \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix} = \frac{1}{\sum_{i=1}^{n} (x^i)^2} \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix}$$



Robust statistics – $\approx$ Max Likelihood
to outliers

$\varepsilon^i \sim \cancel{N(0,\sigma^2)}$

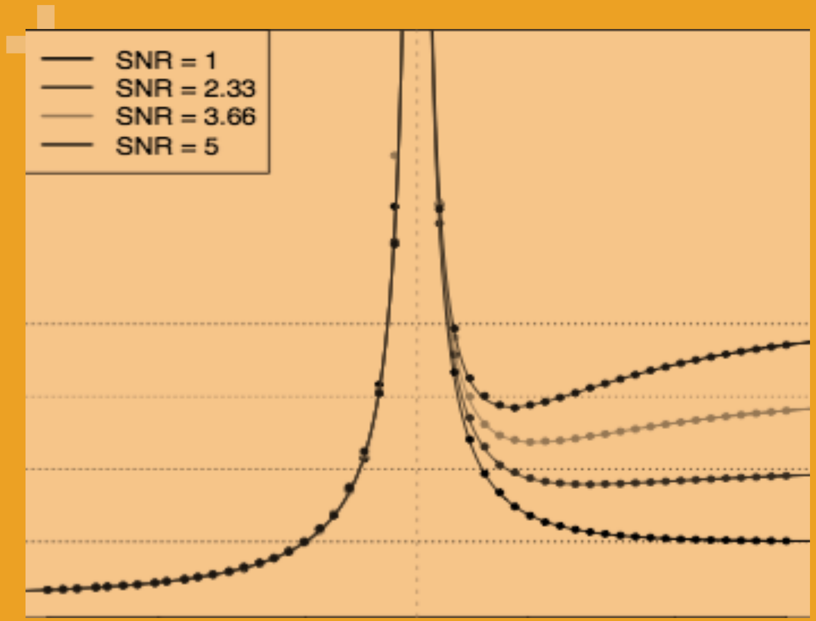
heavy tailed model

Linear regression → Linear regr for non-linear models

Logistic regression

Overparametrized (linear) regression

Bias - Variance Tradeoff for Neural Network



Double Descent

STAT 535+LPL2019

Beyond the Bias-Variance trade-off

Marina Meila
University of Washington

Prediction with "overparametrized" model

$S = \mathbb{R}^d \ni x$

$\mathcal{F}_p = \{$ neural network, RForest, $\dots \}$

f_θ

$|\theta| = p$ # parameters

$\mathcal{F}_{p_1}, \mathcal{F}_{p_2}, \dots \mathcal{F}_K$

train

$\downarrow \quad \downarrow \quad \downarrow$

$f_1 \quad f_2 \quad f_K$

$\mathcal{D} = \{(x^i, y^i)\}_{i=1:n}$

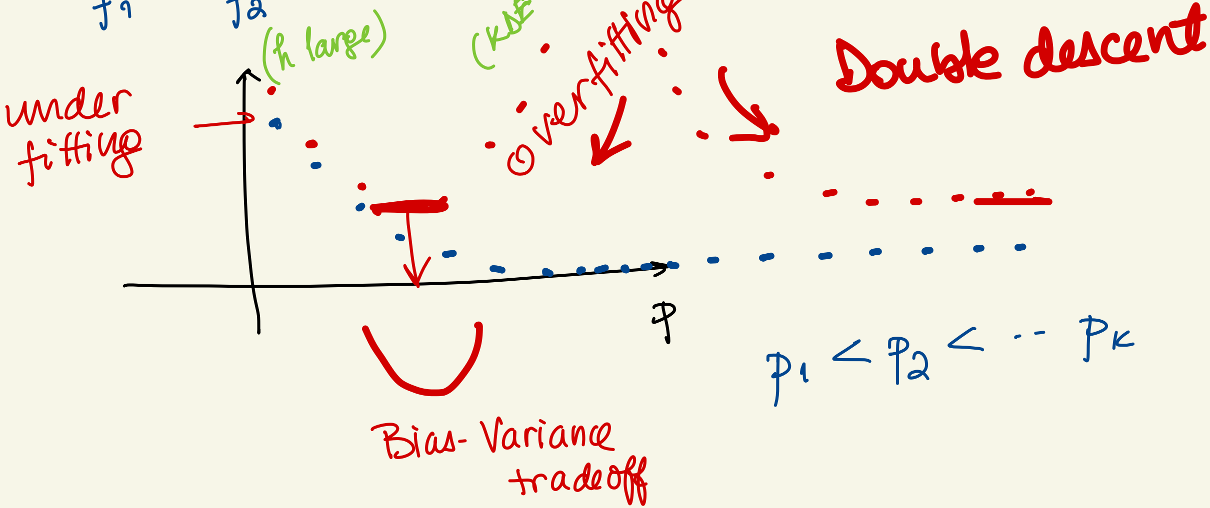
$\mathcal{D}^{test}$ large

trained on $\mathcal{D}$

Least Squares

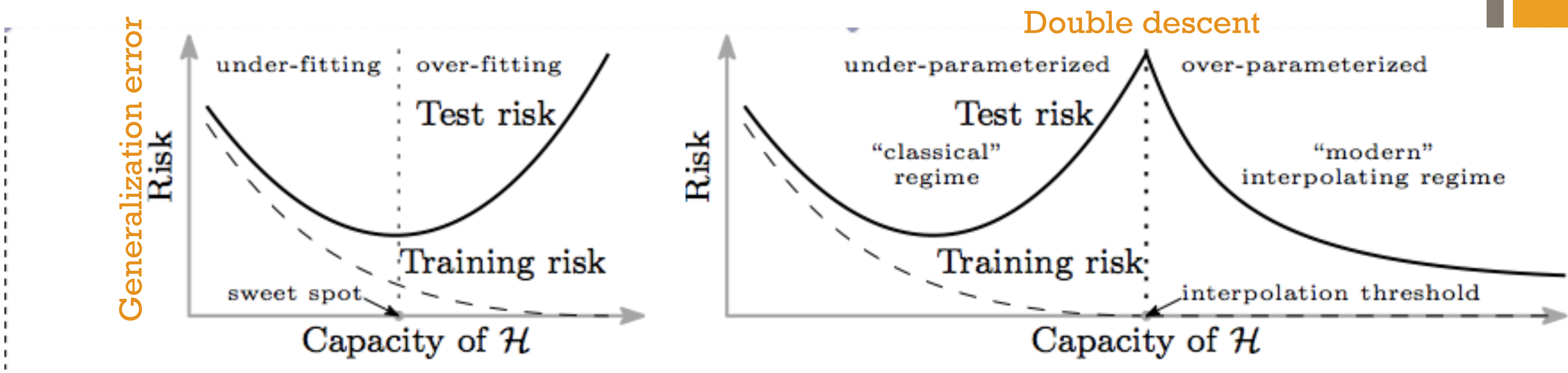
Training error $= \sum_{i=1}^{n} (y^i - f(x^i))^2$ from $\mathcal{D}$

Testing err $= \sum_{i \in \mathcal{D}^{test}} (y^i - f(x^i))^2$



$p_1 < p_2 < \dots p_K$

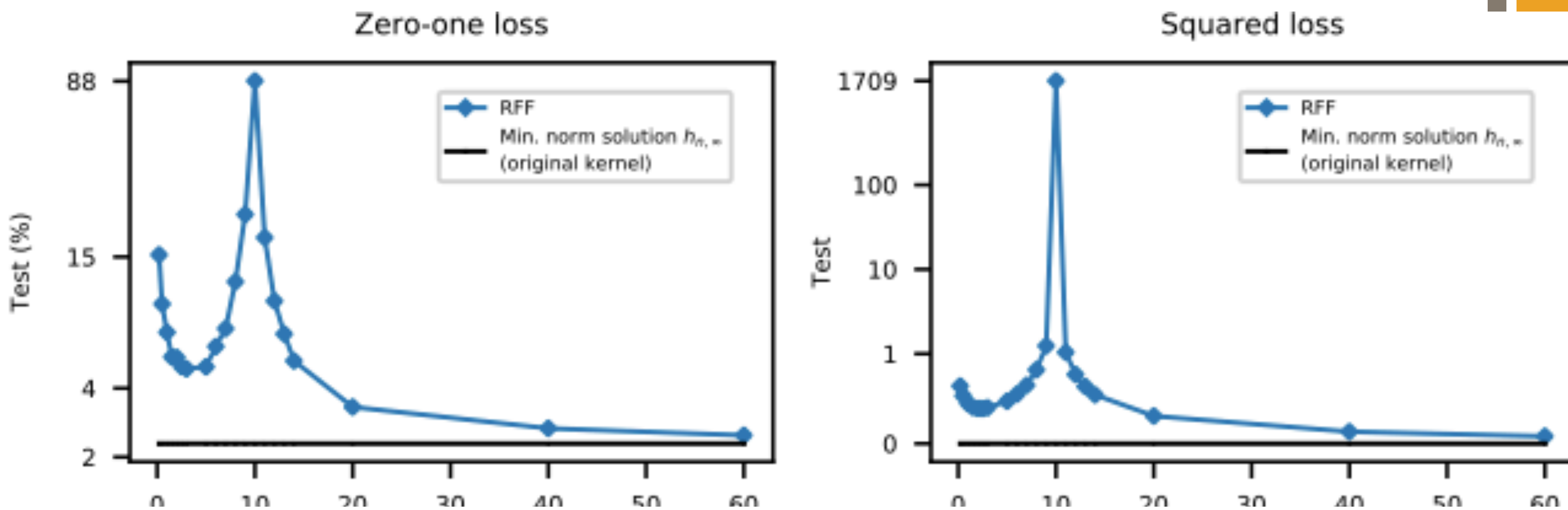
What is observed



Belkin, Hsu, Ma, Mandal 2018

- Classical regime p < N
- Modern/Deep Learning/High dimensional regime N > n
  - Think N fixed, p increases, gamma=p/N
  - Training error = 0 (interpolation)
  - Test error decreases with p (or gamma)

+
What is observed



Belkin, Hsu, Ma, Mandal 2018

- Double descent curves for the generalization error
  - Random Fourier Features (RFF)
  - ReLU 2 layer networks (with random first layer weights)
  - Random Forests, l2-Adaboost
  - Linear regression
- With and without noise

$$\begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [\beta_0] + \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix}}_{} [\beta] + \begin{bmatrix} \varepsilon^1 \\ \vdots \\ \varepsilon^n \end{bmatrix} \Longrightarrow \begin{bmatrix} \varepsilon^1 \\ \vdots \\ \varepsilon^n \end{bmatrix} = \begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix} - X\beta$$

dummy variable

$$\underbrace{\begin{bmatrix} x^1 & 1 \\ x^2 & 1 \\ \cdots & \cdots \\ x^n & 1 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta \\ \beta_0 \end{bmatrix}}_{\beta}$$

$$\boxed{\begin{bmatrix} x^1 & \cdots & x^n \\ 1 & \cdots & 1 \end{bmatrix}}$$

$$X^+ = (X^T X)^{-1} X^T$$

$$X^T X = \begin{bmatrix} \sum_{i=1}^{n} (x^i)^2 & \sum_{i=1}^{n} x^i \\ \sum_{i=1}^{n} x^i & n \end{bmatrix} \in \mathbb{R}^{2\times 2} \quad \succeq 0$$

invertible if X is full rank