KERNEL METHODS - PRACTICAL SESSION № 1

AMMI 2020 18/05/2020

Linear Algebra Recap

Exercise 1

Let $A \in \mathbb{R}^{m \times n}$ be a real matrix. For each function f, **specify its codomain and compute** ∇f

(a) $f : \begin{cases} \mathbb{R} \to ? \; \mathbb{R}^{m \times m} \\ x \to xA \end{cases}$

(b) f: $\begin{cases} \mathbb{R}^n \to ? \; \mathbb{R}^m \\ \mathrm{x} \to A\mathrm{x} \end{cases}$

(c) f: $\begin{cases} \mathbb{R}^{m \times n} \to ? \; \mathbb{R} \\ X \to Tr(A^T X) \end{cases}$

Exercise 2

Let $X \in \mathbb{R}^{n \times p}$, $\lambda > 0$.
Prove that $M = X^T X + \lambda I_p$ **is invertible.**
Hint: Prove that the eigenvalues of M *are larger than* λ

Linear Regression

Using notations from the slides:

- $Y = (y_1, ..., y_n)^T \in \mathbb{R}^n$ the vector of outcomes
- $X = (\mathrm{x}_1, ..., \mathrm{x}_n)^T \in \mathbb{R}^{n \times p}$ the matrix (n rows=samples, p columns=features)
- $\beta \in \mathbb{R}^p$ the linear model's parameters

Exercise 3

(a) State the loss function for ordinary least squares (OLS)

(b) Formulate OLS as a minimization problem

(c) Compute the solution $\hat{\beta}^{OLS}$

(d) What happens if $X^T X$ is singular (not invertible)?

Exercise 4 - Bias and Variance of OLS

slide 16

Exercise 5 - Optimality of OLS

slide 17

Ridge Regression

Exercise 6 - Bias and Variance

slide 33

Exercise 7 - Performance

slide 34

Ridge Logistic Regression

Exercise 8

slide 46

EXERCISE 01

① $A \in \mathbb{R}^{m \times m}$

a] $f: \begin{cases} \mathbb{R} \longrightarrow \mathbb{R}^{m \times m} \\ x \longrightarrow xA \end{cases}$ $\quad f(x) = xA$; $\nabla f \in \mathbb{R}^{m \times m}$

$\nabla f_{ij} = f'_{ij}(x)$; $f_{ij}(x) = f(x)_{ij} = A_{ij} x \Rightarrow \nabla f_{ij} = A_{ij} \Rightarrow \boxed{\nabla f = A}$

b] $f: \begin{cases} \mathbb{R}^m \longrightarrow \mathbb{R}^{m \times 1} \\ x \rightsquigarrow Ax \end{cases}$ $\quad$ m dimensions, m variables

$\nabla f \in \mathbb{R}^{m \times m}$

$\underset{m \times m}{A} \; \underset{m \times 1}{x}$

$\nabla f_{ij} = \frac{\partial f_i}{\partial x_j}$ $\quad f_i(x) = A_i^T x = \sum_k A_{ik} x_k$

$\frac{\partial f_i}{\partial x_j} = A_{ij}$

$\Rightarrow \boxed{\nabla f = A}$

c] $f: \begin{cases} \mathbb{R}^{m \times m} \longrightarrow \mathbb{R} \\ X \rightsquigarrow Tr(A^T X) \in \mathbb{R} \end{cases}$

$Tr(A^T X) \in \mathbb{R}^{m \times m} \times \mathbb{R}^{m \times m} \to \mathbb{R}^{m \times m}$

$Tr(M) = \sum$ of diagonal elements of M.

$\nabla f_{ij} = \frac{\partial f}{\partial x_{ij}}$ with $f(x) = \sum A_{ij} X_{ij}$

$\frac{\partial f}{\partial x_j} = A_{ij} \Rightarrow \boxed{\nabla f = A}$

If we have $f: \mathbb{R}^m \longrightarrow \mathbb{R}$

$f(x) = f(x_0) + (x - x_0)^T \nabla f(x_0) + O(\|x - x_0\|)$

$\approx f(x_0) + f'(x_0)(x - x_0) + O(x - x_0)$

$\nabla' f_{ij} = \frac{\partial^2 f}{\partial x_{ij}}$

EXERCISE 2

Let $X \in \mathbb{R}^{n \times p}$, $\lambda > 0$. Proove that $M = X'X + \lambda I_p$ is invertible.

Let $M' = X^T X \in \mathbb{R}^{p \times p}$, we will show that M' has non-negative eigenvalues.
Let μ be an eigenvalue of M', and y an eigenvector.

$M'y = \mu y \Rightarrow y^T M' y = y^T \mu y = \mu y^T y = \mu \|y\|^2$ ($\mu \geq 0$)

$M' = X^T X \Rightarrow y^T M' y = y^T X^T X y = (Xy)^T (Xy) = \|Xy\|^2$

$$\|Xy\|^2 = \mu \|y\|^2 \Rightarrow \mu = \frac{\|Xy\|^2}{\|y\|^2} \geq 0$$

$\Rightarrow$ The eigenvalues of M are equal to λ + the eigenvalues of M'

$$\mu \in Eig(M') \Leftrightarrow \mu + \lambda \in Eig(M) \quad (1)$$
(≥ 0, > 0)

Therefore the eigenvalues of M are $\geq \lambda$

So M is invertible $\Rightarrow$ (M is invertible $\Leftrightarrow$ its eigenvalues are $\neq 0$)

Proof of (1): M' is symmetric ($M'^T = M'$) so there exist $U \in \mathbb{R}^{p \times p}$ orthogonal ($U^T U = I$) such that $M' = U^T \Delta U$; $\Delta = Diag(\lambda_1 \dots \lambda_n)$ the λ_i's are eigenvalues of M'.

$$M = M' + \lambda I = U^T \Delta U + \lambda U^T U = U^T (\Delta + \lambda) U$$

$$= U^T \begin{pmatrix} \lambda_1 + \lambda & 0 & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & & & \lambda_n + \lambda \end{pmatrix} U \Rightarrow$$ The eigenvalues of M are $\lambda_i + \lambda$

EXERCISE 3

a) State the loss function for ordinary least squares (OLS)

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

$= \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$ $\qquad \varepsilon = y - \beta^T x$

The criterion: $MSE(\beta) = \frac{1}{n} \sum (y_i - \beta^T x_i)^2$

$$= \frac{1}{n} (y - X\beta)^T (y - X\beta) = \frac{1}{n} \|y - X\beta\|^2$$

b) Formulate OLS as a minimization Problem.

ordinary Least Square

$$\min_{\beta} MSE(\beta), \quad MSE: \mathbb{R}^p \longrightarrow \mathbb{R}$$

c) Compute the solution $\hat{\beta}^{OLS}$

$$\nabla MSE(\tilde{\beta}) = 0$$

$$\nabla MSE = \frac{1}{n}\left[(y-X\beta)^T(y-X\beta)\right] = \frac{2}{n}(y-X\beta)(-X) = \frac{2}{n}X^T(X\beta - y)$$

$$\nabla MSE = \frac{\partial MSE}{\partial \beta}$$

OLS convex quadratic reaches optimum when $\nabla MSE(\beta^*)$

$$X^T(X\beta^* - y) = 0$$

$$X^T X \beta^* = X^T y$$

* If X^TX non singular $\beta^* = (X^TX)^{-1}X^Ty$

$A^TB = B^TA$.

$$\nabla MSE = \frac{2}{n}X^T(X\beta - y)$$

d) What happens if X^TX is singular (not invertible)

when X^TX is singular

⇒ pseudo inverse → Find solution when there is no solution.

⇒ Regularize with small λ; $(X^TX)^{-1}$ becomes $(X^TX + \lambda I_m)^{-1}$ → which is inv. (Ex. 2)

ridge regression no closed forme / unstable

EXERCISE 4 - BIAS AND VARIANCE OF OLS (SLIDE 16)

Show that $\hat{\beta}^{OLS} = (X^TX)^{-1}X^Ty$

satisfies $\begin{cases} E(\hat{\beta}^{OLS}) = \beta^* \\ Var(\hat{\beta}^{OLS}) = E(\hat{\beta}^{OLS} - \beta^*)(\hat{\beta}^{OLS} - \beta^*)^T = \sigma^2(X^TX)^{-1} \end{cases}$

For $A \in \mathbb{R}^p$, $Var(A) \in \mathbb{R}^{p \times p}$ — is the covariance matrix.

$$Var(A) = E\left[(A - E(A))(A - E(A))^T\right]$$

1-dimension $Var(A) = E(A - E(A))$

$Var(A)_{ij} = Cov(A_i, A_j) = E(A_i, A_j) - E(A_i)E(A_j)$

In this exercise: $E(\varepsilon) = 0$; $Var(\varepsilon) = \sigma^2 I$

$\Rightarrow Cov(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$ and $Var(\varepsilon_i) = \sigma^2$

Let define $P = (X^T X)^{-1} X^T$; and assume: $y = X\beta^* + \varepsilon$

$$\hat{\beta}^{OLS} = Py = (X^T X)^{-1} X^T (X\beta^* + \varepsilon) = (X^T X)^{-1} X^T X \beta^* + P\varepsilon$$
$$= \beta^* + P\varepsilon$$

E is linear

$$E[\hat{\beta}^{OLS}] = E(\beta^*) + E(P\varepsilon) = \beta^* + P\,E(\varepsilon) = \beta^* \Rightarrow \hat{\beta}^{OLS} \text{ is unbiased}$$

($E(\varepsilon) = 0$)

$$\Rightarrow Var(\hat{\beta}^{OLS}) = Var(\hat{\beta}^{OLS} - \beta^*) = Var(\beta^* + P\varepsilon - \beta^*) = Var(P\varepsilon)$$
$$= P\,Var(\varepsilon)\,P^T$$

$Var(\varepsilon) = \sigma^2 I$

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

$$Var(\hat{\beta}^{OLS}) = \sigma^2 P P^T = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}$$

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

$\left((M^{-1})^T = (M^T)^{-1}\right)$

$$= \sigma^2 (X^T X)^{-1}$$

EXERCISE 5

Gauss Markov theorem:

Least squares estimator is Best Linear unbiased Estimator.

Assumption

$y = X\beta^* + \varepsilon$

$E(\varepsilon) = 0$ $\quad Var(\varepsilon) = \sigma^2 I$

$\hat{\beta}^{OLS}$ has the lowest variance among linear unbiased estimators

if $\tilde{\beta} = Cy$ (C is a matrix) unbiased ($E\tilde{\beta} = \beta^*$) then $Var(\hat{\beta}^{OLS}) \leq Var(\tilde{\beta})$

(⚠ This condition does not mean $Var(\hat{\beta}^{OLS})_{ij} \leq Var(\tilde{\beta})_{ij}$ $\forall i, j$

i.e.:

$M = Var(\tilde{\beta}) - Var(\hat{\beta}^{OLS})$ is a positive semidefinite matrix

i.e. $\forall X_0,\ X_0^T M X_0 \geq 0$

PROOF

Let $\tilde{\beta} = Cy$ be an unbiased estimator, $C \in \mathbb{R}^{p \times n}$

Define: $D := C - (X^TX)^{-1}X^T$ (*)

We are going to show that $Var(\hat{\beta}^{OLS}) \leq Var(\tilde{\beta})$

→ STEP 1: show that $DX = 0$

$\tilde{\beta}$ is unbiased: $E(\tilde{\beta}) = \beta^*$ (1)

$E(\tilde{\beta}) = E(Cy) = CE(y) = CE(X\beta^* + \varepsilon) = CX\beta^* + C\underbrace{E(\varepsilon)}_{=0} = CX\beta^*$ (2)

Combining (1) and (2) $\beta^* = CX\beta^*$

$\Rightarrow (CX - I)\beta^* = 0$ Recall $D = C - (X^TX)^{-1}X^T$

$= DX\beta^* = 0$ $DX = CX - (X^TX)^{-1}X^TX = CX - I$

This is true for all $\beta^* \Rightarrow \underline{DX = 0}$ □

→ step 2: show that $Var(\tilde{\beta}) = Var(\hat{\beta}^{OLS}) + \sigma^2 DD^T$

We reuse the notation $P = (X^TX)^{-1}X^T$

Since $DX = 0$, $DP^T = PD^T = 0$

From (*) $C = D + P$

$\Rightarrow \tilde{\beta} = CY = (D+P)(X\beta^* + \varepsilon) = \underbrace{DX\beta^*}_{=0} + D\varepsilon + \underbrace{PX\beta^*}_{I} + P\varepsilon$

$= \beta^* + (D+P)\varepsilon$

$Var(\tilde{\beta}) = Var(\beta^* + (D+P)\varepsilon) = (D+P)\underbrace{Var(\varepsilon)}_{\sigma^2 I}(D+P)^T$ $(Var(\beta^*) = 0)$

$= \sigma^2(DD^T + \underbrace{DP^T}_{=0} + \underbrace{PD^T}_{=0} + \underbrace{PP^T}_{(X^TX)^{-1}}) = \sigma^2 DD^T + \underbrace{\sigma^2(X^TX)^{-1}}_{= Var(\hat{\beta}^{OLS})}$

$\Rightarrow Var(\tilde{\beta}) = Var(\hat{\beta}^{OLS}) + \sigma^2 DD^T$ (EX.4) □

Note: $y \in \mathbb{R}^p$ random vector, $C \in \mathbb{R}^{p \times p}$ fixed. then $E(Cy) = CE(y)$ and $Var(Cy) = C\,Var(y)\,C^T$

→ step 3 (final) $M = Var(\tilde{\beta}) - Var(\hat{\beta}^{OLS})$ is positive semi-definite (Pd)

- $M = Var(\tilde{\beta}) - Var(\hat{\beta}^{OLS}) = \sigma^2 D D^T$ (from step 2)
- Let $x_0 \in \mathbb{R}^n$: $x_0^T M x_0 = \sigma^2 x_0^T D D^T x_0 = \sigma^2 (D^T x_0)^T (D^T x_0) = \sigma^2 \|D^T x_0\|^2 \geq 0$

This is fine $\forall x_0$: M is positive semi-definite (P.d.)

Finally: $Var(\tilde{\beta}) \geq Var(\hat{\beta}^{OLS})$

$\tilde{\beta}$ has a "large" variance than $\hat{\beta}^{OLS}$

Why does this matter?

Suppose we observe measurements $y = \mu + \varepsilon$ (what we want to estimate; ε: noise)

n measurements $y_i = \mu + \varepsilon_i$; $\mathbb{E}(\varepsilon_i) = 0$; $Var(\varepsilon) = \sigma^2 I$

we want an estimator $\hat{\mu}$ of μ.

which is the best estimator?

① $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i$; ② $\hat{\mu}_2 = y_i$; ③ $\hat{\mu}_3 = \frac{1}{n}\sum y_i + \frac{1}{n}$

→ Bias: $\mathbb{E}(\hat{\mu}_1) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu = \mathbb{E}(\hat{\mu}_2)$

$\mathbb{E}(\mu_3) = \mu + \frac{1}{n} \neq \mu$

⇒ $\hat{\mu}_1$ and $\hat{\mu}_2$ are unbiased but $\hat{\mu}_3$ is biased

→ Variance $Var(\hat{\mu}_1) = Var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(\mu) + Var(\varepsilon)$

$= \frac{1}{n^2}\left(n\,Var(\mu) + n\sigma^2 I\right) = \frac{\sigma^2}{n} = Var(\hat{\mu}_3)$ ($Var(\mu) = 0$)

$Var(\hat{\mu}_2) = \sigma^2$

→ Error: $Er(\hat{\mu}) = bias^2(\hat{\mu}) + Var(\hat{\mu})$

$Er(\hat{\mu}_1) = bias^2(\hat{\mu}_1) + Var(\hat{\mu}_1) = \mathbb{E}\left(\mathbb{E}(\hat{\mu}_1)\right) + \frac{\sigma^2}{n} = \frac{\sigma^2}{n}$ ($\mathbb{E}(\hat{\mu}_1) = \hat{\mu}$, $\hat{\mu} = 0$)

$Er(\hat{\mu}_2) = 0 + \sigma^2 = \sigma^2$

$Er(\hat{\mu}_3) = \mathbb{E}(\mu)\mathbb{E}(\mu) + \mathbb{E}(\mu)\mathbb{E}\left(\frac{1}{n}\right) + \mathbb{E}\left(\frac{1}{n}\right)\mathbb{E}(\mu) + \mathbb{E}\left(\frac{1}{n}\right)\mathbb{E}\left(\frac{1}{n}\right) + \frac{\sigma^2}{n} = \frac{1}{n^2} + \frac{\sigma^2}{n}$ ($\mathbb{E}(\mu) = 0$, $\mathbb{E}\left(\frac{1}{n}\right) = \frac{1}{n}$)

⇒ $\hat{\mu}_1$ and $\hat{\mu}_2$ are both unbiased, but $\hat{\mu}_1$ is much better than $\hat{\mu}_2$

⇒ $\hat{\mu}_3$ is biased, and $\hat{\mu}_2$ is not, but $\hat{\mu}_3$ is much better than $\hat{\mu}_2$

↳ Variance matters just as much as bias.

EXERCISE 6 (Slides 33) Ridge Regression

$$\begin{cases} \text{bias}\left(\hat{\beta}^{\text{Ridge}}\right) = ? = \frac{-\lambda}{1+\lambda}\beta^* \\ \text{Var}\left(\hat{\beta}_2^{\text{Ridge}}\right) = ? = \frac{1}{(1+\lambda)^2}\text{Var}\left(\hat{\beta}^{OLS}\right) \end{cases}$$

loss ($\lambda \geq 0$)

$$L(\beta) = MSE(\beta) + \lambda\|\beta\|^2 = \frac{1}{n}(y - X\beta)^T(y - X\beta) + \lambda\|\beta\|^2$$

gradiant

$$\nabla J(\beta) = \frac{2}{n}X^T(X\beta - y) + 2\lambda\beta$$

$$\nabla J\left(\hat{\beta}_\lambda^{\text{Ridge}}\right) = 0 \iff \frac{2}{n}X^TX\beta + 2\lambda\beta = \frac{2}{n}X^Ty$$

$$\Rightarrow \frac{2}{n}\beta(X^TX + \lambda n I) = \frac{2}{n}X^Ty \Rightarrow \boxed{\hat{\beta}_\lambda^{\text{Ridge}} = (X^TX + \lambda n I)^{-1}X^Ty}$$

N.B.: when $\lambda > 0$, $(X^TX + \lambda n I)^{-1}$ always exist (Exercise 2)

Assumption : $X^TX = nI$ (**) (orthogonal vectors X_i)

$$\Rightarrow \hat{\beta}_\lambda^{\text{Ridge}} = \frac{1}{n(1+\lambda)}X^Ty$$

$$y = X\beta^* + \varepsilon$$

$$= \frac{1}{n(1+\lambda)}\left(\underbrace{X^TX}_{nI\ (**)}\beta^* + X^T\varepsilon\right) = \frac{1}{(1+\lambda)}\beta^* + \frac{1}{n(1+\lambda)}X^T\varepsilon$$

$$\Rightarrow \hat{\beta}_\lambda^{\text{Ridge}} = \frac{1}{(1+\lambda)}\beta^* + \frac{1}{n(1+\lambda)}X^T\varepsilon$$

Taking expectations :

$$\mathbb{E}\left(\hat{\beta}_\lambda^{\text{Ridge}}\right) = \frac{1}{1+\lambda}\mathbb{E}(\beta^*) + \frac{1}{n(1+\lambda)}X^T\underbrace{\mathbb{E}(\varepsilon)}_{0} = \frac{1}{1+\lambda}\beta^*$$

⇒ Hence

$$\text{bias}\left(\hat{\beta}_\lambda^{\text{Ridge}}\right) = \mathbb{E}\left(\hat{\beta}_\lambda^{\text{Ridge}}\right) - \beta^* = -\frac{\lambda}{1+\lambda}\beta^*$$

Large λ = large bias

Small λ = small bias

⇒ when $\lambda = 0$ (no regularization), bias $= 0$, $\hat{\beta}_\lambda^{\text{Ridge}} = \hat{\beta}^{OLS}$

when $\lambda = +\infty$, bias $= -\beta^*$ because $\hat{\beta}_\lambda^{\text{Ridge}} = 0$

Assumption

$$X^T X = nI$$

$$X_i^T X_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

$$\hat{\beta}_\lambda^{Ridge} = \frac{1}{n(1+\lambda)} X^T y$$

$$Var(\hat{\beta}_\lambda^{Ridge}) = \frac{1}{n^2(1-\lambda)^2} Var(X^T y) = \frac{1}{n^2(1-\lambda)^2} X^T \underbrace{Var(\varepsilon)}_{\sigma^2 I} X$$

$$= \frac{\sigma^2}{n^2(1-\lambda)^2} \underbrace{X^T X}_{nI} = \frac{\sigma^2}{n(1+\lambda)^2}$$

Large λ = small variance

Small λ = large variance

Unolet out assumptions

$$Var(\hat{\beta}_\lambda^{RIDGE}) = \frac{\sigma^2}{n(1+\lambda)^2} \quad ; \quad Var(\hat{\beta}^{OLS}) = (X^T X)^{-1}\sigma^2 = \frac{\sigma^2}{n}$$

$$Var(\hat{\beta}_\lambda^{Ridge}) = \frac{1}{(1+\lambda)^2} Var(\hat{\beta}^{OLS})$$

$\lambda \to 0$: $\hat{\beta}_\lambda^{Ridge} \to \hat{\beta}^{OLS}$: no bias, large variance

$\lambda \to +\infty$: $\hat{\beta}_\lambda^{Ridge} \to 0$: Large bias, no variance

EXERCISE 7. Expectation of total error

$$f(\lambda) = \mathbb{E}_{S, X_0}\left[bias^2(X_0^T \hat{\beta}_\lambda^{Ridge}) + Var(X_0^T \hat{\beta}_\lambda^{Ridge}) \right]$$

$\mathbb{E} X_0 = 0$

$\mathbb{E} X_0 X_0^T = I$

→ Let us compute the two terms:

- $bias^2(X_0^T \hat{\beta}_\lambda) = (X_0^T bias(\hat{\beta}_\lambda))^2$

$$= \frac{\lambda^2}{(1+\lambda)^2} (X_0^T \beta^*)^2$$

$$= \frac{\lambda^2}{(1+\lambda)^2} \beta^T X_0 X_0^T \beta^*$$

$bias(\hat{\beta}) = \beta^*$

$(X_0^T \beta^*)^2 = \sum X_{0,i} X_{0,j} \beta_i \beta_j^*$

$= \beta^T X_0 X_0^T \beta^*$

- $Var(X_0^T \hat{\beta}_\lambda^{Ridge}) = X_0^T Var(\hat{\beta}_\lambda^{Ridge}) X_0 = \frac{\sigma^2}{n(1+\lambda)} X_0^T X_0$

→ Taking expectations, with $\mathbb{E}(X_0 X_0^T) = I$, we got

$$f(\lambda) = \frac{\lambda^2}{(1+\lambda)^2} \|\beta\|^2 + \frac{\sigma^2 p}{n(1+\lambda)^2}$$

$$\mathbb{E}(X_0^T X_0) = \sum_{i=1}^{p} \mathbb{E}(X_{0,i}^2) = p$$

→ Finding the optimum λ^*

Define $a = \|\beta^*\|^2 = f(+\infty)$; $b = \frac{\sigma^2 p}{n} = f(0)$

$$f(\lambda) = \frac{1}{(1+\lambda)^2}(a\lambda^2 + b) \quad ; \quad f'(\lambda) = \frac{2}{(1+\lambda)^3}(a\lambda - b)$$

$$f'(\lambda) = 0 \iff \lambda^* = \frac{b}{a} = \frac{\sigma^2 p}{n\|\beta^*\|^2}$$

Keeping our notations : $f(\lambda^*) = \frac{1}{(1+\frac{b}{a})^2}\left(a\cdot\left(\frac{b}{a}\right)^2 + b\right) = \frac{ab}{a+b}$

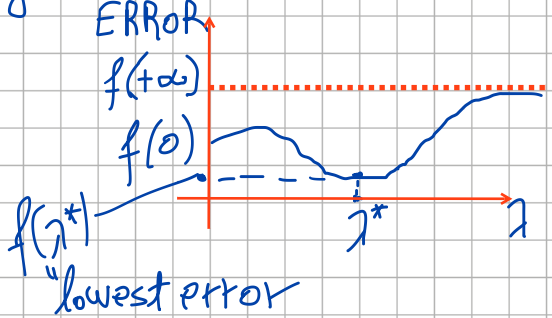
$$f(\lambda^*) = \frac{ab}{a+b} = \frac{f(0)f(+\infty)}{f(0)+f(+\infty)}$$

$$\left.\begin{array}{l} a \geq 0 \Rightarrow \frac{ab}{a+b} \leq \frac{a(a+b)}{a+b} = a \\ b \geq 0 \Rightarrow \frac{ab}{a+b} \leq \frac{(a+b)b}{a+b} = b \end{array}\right\} \frac{ab}{a+b} \leq \min(a,b)$$

$f(X)$ = error function $\exists$ optimal λ^* that gives the best error
Cannot compute it directly. Find with cross-validation.

$$\Rightarrow f(\lambda^*) \leq \min(f(0), f(+\infty))$$

Using λ^* is better than not using regularization.



conclusion (courbe)
Some regularization is better than
and better than the estimator 0.

EXERCISE 8 — RIDGE LOGISTIC REGRESSION (SLIDE 46)

$$\beta^{new} = \beta^{old} - \left[\nabla^2_\beta J(\beta^{old})\right]^{-1} \nabla_\beta J(\beta^{old})$$

Lemma: β^{new} is also the solution to $\min_\beta J_q(\beta) = \frac{1}{n}\sum_{i=1}^{n} \ln(1+e^{-y_i \beta^T x_i}) + \lambda \|\beta\|^2$ were

$$J_q(\beta) := J(\beta^{old}) + \nabla J(\beta^{old})(\beta - \beta^{old}) + \frac{1}{2}(\beta - \beta^{old})^T \nabla^2_\beta J(\beta^{old})(\beta - \beta^{old})$$

J_q is the quadratic approximation to J in β^{old}

(analogy in 1 dimension: $J_q(x) = J'(x_0)(x-x_0) + \frac{1}{2}J''(x_0)(x-x_0)^2$)

Proof:

$$\nabla J_q(\beta) = \nabla J(\beta^{old}) + \nabla^2_\beta J(\beta^{old})(\beta - \beta^{old})$$

$$= 0 \iff \beta = \beta^{old} - \left(\nabla^2_\beta J(\beta^{old})\right]^{-1} \nabla_\beta J(\beta^{old})$$

$$= \beta^{new}$$

writting all the terms of J_q that depend on β:

$$\Rightarrow \beta^T \nabla J(\beta^{old}) = \frac{1}{n}\beta^T X^T P y + 2\lambda \beta^T \beta^{old}$$

$$\Rightarrow \frac{1}{2}\beta^T \nabla^2_\beta J(\beta^{old})\beta = \frac{1}{2n}\beta^T X^T W X \beta + \lambda \beta^T \beta$$

$$\Rightarrow -\beta^T \nabla^2_\beta J(\beta^{old})\beta^{old} = -\frac{1}{n}\beta^T X^T W X \beta^{old} - 2\lambda \beta^T \beta^{old}$$

Putting it all together

$$J_q(\beta) = -\frac{1}{n}\beta^T X^T W X \beta^{old} + \frac{1}{n}\beta^T X^T P y + \frac{1}{2n}\beta^T X^T W X \beta + \lambda \beta^T \beta + C$$

$$= -\frac{1}{n}\beta^T X^T W \underbrace{(X\beta^{old} - W^{-1} P y)}_{z} + \frac{1}{2n}\beta^T X^T W X \beta + \lambda \beta^T \beta + C$$

$$= \frac{1}{2n}(X\beta - z)^T W + \lambda \|\beta\|^2 + C$$