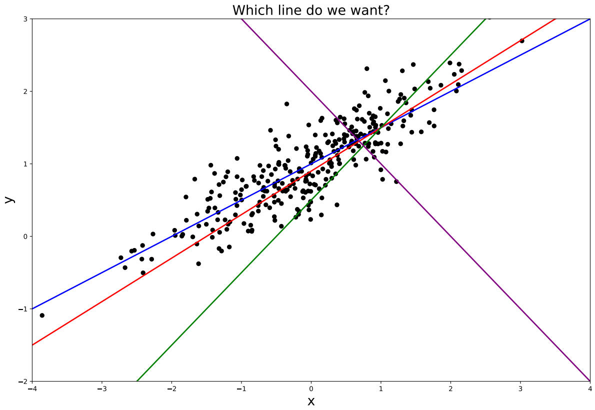
3. Linear least-square regression

3.1. Framework

Intuition

- **Goal:** find the “best” hyperplane going through our training data



Least-square framework

- **reminders:** regression $\Rightarrow \mathcal{Y} = \mathbb{R}$
- square loss $\ell(y, y') = (y - y')^2$
- we know that the optimal predictor is $f^\star(x) = \mathbb{E}\left[Y \mid X = x\right]$
- **Notation:** $\varphi : \mathcal{X} \to \mathbb{R}^d$ some feature function
- ERM on the class of functions

$$f_\theta(x) = \varphi(x)^\top \theta = \sum_{j=1}^{d} \varphi(x)_j \theta_j \, ,$$

  with $\theta \in \mathbb{R}^d$
- **Remark:** linear in θ, not necessarily in x!
- **Overall:** minimize

$$\hat{\mathcal{R}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - \varphi(X_i)^\top \theta)^2 \, .$$

Random design

- mathematically, more interesting to see (x_i, y_i) as **random variables**
- $\to$ we write (X_i, Y_i) instead of (x_i, y_i)

Key assumption: (X_i, Y_i) are independent, identically-distributed (i.i.d.) copies of (X, Y).

- from now on, we will work in this framework
- **Remark:** *distribution shift* is a current research topic[4]
- **Key difference:**

$$\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i))$$

is a *random variable*

[4]Sugiyama, Kawanabe, *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*, MIT Pres, 2012

Example 1: linear regression

- **Question:** what is φ? and why is it useful?
- univariate inputs: $\mathcal{X} = \mathbb{R}$
- take $d = 2$
- **Why?** allowing an *intercept*: $\varphi(x) = (1, x)^\top$ and

$$\Phi = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

Example 2: polynomial regression

- consider again univariate inputs: $\mathcal{X} = \mathbb{R}$
- take $d = p + 1$, with p maximal degree
- set $\varphi(x) = (1, x, x^2, \dots, x^p)^\top$, and

$$\Phi = \begin{pmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^p \\ \vdots & \vdots & & \vdots & \\ 1 & X_n & X_n^2 & \cdots & X_n^p \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}$$

- true strength of the linear model: non-linear transformations of the entries

Matrix notation

- let $Y := (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ the response vector
- let $\Phi \in \mathbb{R}^{n \times d}$ the matrix of inputs
- row i of $\Phi = \varphi(X_i)^\top$
- with these notation,
$$\hat{\mathcal{R}}(\theta) = \frac{1}{n} \left\| Y - \Phi\theta \right\|^2 .$$
- **Reminder:**
$$\left\| u \right\|^2 = \langle u, u \rangle = u^\top u = \sum_{j=1}^{d} u_j^2$$
denotes the Euclidean norm

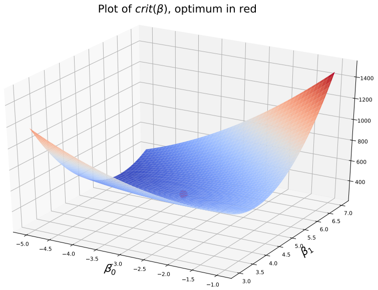
3.2. Ordinary least-squares

Ordinary Least Squares

- **Reminder:** we want to minimize

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n} \| Y - \Phi\theta \|^2 .$$

- now we have to work a bit because crit is a function of d variables:



Calculus aparte

- **Reminder:** let $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$, then the *gradient* of f is defined as

$$\nabla f = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_M}{\partial x_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_N} & \frac{\partial f_2}{\partial x_N} & \cdots & \frac{\partial f_M}{\partial x_N} \end{pmatrix} \in \mathbb{R}^{N \times M}$$

- **Example:** when f is real-valued $(M = 1)$, ∇f is a vector, thus a column

Calculus aparte, ctd.

- let us consider first the function $f : x \mapsto Ax$, with $x \in \mathbb{R}^N$ and $A \in \mathbb{R}^{M\times N}$ a fixed matrix
- let $j \in \{1, \ldots, M\}$, then we know that

$$(Ax)_j = A_{j,1}x_1 + A_{j,2}x_2 + \cdots + A_{j,N}x_N \,.$$

- let $i \in \{1, \ldots, N\}$, then

$$\frac{\partial}{\partial x_i}(Ax)_j = A_{j,i} \,.$$

- we deduce from this computation that

$$\boxed{\forall A \in \mathbb{R}^{M\times N}, \qquad \nabla(Ax) = A^\top}$$

Calculus aparte, ctd.

- more complicated: let $B \in \mathbb{R}^{N \times N}$ and define $f : x \mapsto x^\top Bx$
- set $1 \in \{1, \ldots, N\}$, then

$$(Bx)_j = B_{j,1}x_1 + B_{j,2}x_2 + \cdots + B_{j,N}x_N \, .$$

- we deduce that

$$x^\top Bx = \sum_{j,k=1}^{n} B_{j,k}x_j x_k \, .$$

- therefore,

$$\frac{\partial}{\partial x_i}(x^\top Bx) = \sum_{j=1}^{n}(B_{i,j} + B_{j,i})x_j \, .$$

- in a concise form:

$$\boxed{\forall B \in \mathbb{R}^{N \times N}, \qquad \nabla(x^\top Bx) = (B + B^\top)x}$$

Closed-form solution (i)

- $\hat{\mathcal{R}}$ is a convex smooth function $\Rightarrow$ look at critical point
- back to the definition:

$$\begin{aligned}\hat{\mathcal{R}}(\theta) &= \frac{1}{n}\left\|Y - \Phi\theta\right\|^2 \\ &= \frac{1}{n}\left(\left\|Y\right\|^2 - 2\theta^\top\Phi^\top Y + \theta^\top\Phi^\top\Phi\theta\right)\end{aligned}$$

- from the previous slides, we deduce

$$\nabla\hat{\mathcal{R}}(\theta) = \frac{2}{n}\left(\Phi^\top\Phi\theta - \Phi^\top Y\right)$$

- setting to zero yields the **normal equations**:

$$\Phi^\top\Phi\hat{\theta} = \Phi^\top Y\,.$$

Closed-form solution (ii)

Proposition: Assume that Φ has full column rank. Then the unique minimizer of $\hat{\mathcal{R}}$ is given by

$$\hat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top Y \, .$$

- when it exists, we will refer to $\hat{\theta}$ as the *ordinary least squares* (OLS) solution
- **Remark (i):** Φ full column rank $\Leftrightarrow$ $\Phi^\top \Phi$ positive-definite (in particular, invertible)
- **Remark (ii):** if $\varphi = \mathrm{id}$, recover the well-know formula:

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y \, .$$

- **Remark (iii):** $\Phi\hat{\theta}$ (vector of predictions) $=$ orthogonal projection of Y onto $\mathrm{Im}\,(\Phi)$

Numerical resolution, invertible case

- inverting matrices is *hard* (costly + unstable)
- **What is done in practice:** QR factorization: write

$$\Phi = QR$$

with $Q \in \mathbb{R}^{n\times d}$ such that $Q^\top Q = \mathrm{I}$ and $R \in \mathbb{R}^{d\times d}$ upper triangular
- fast, and more stable
- then

$$\Phi^\top \Phi = R^\top Q^\top QR = R^\top R$$

which means

$$\left(\Phi^\top \Phi\right)\hat{\theta} = \Phi^\top Y$$

if, and only if,

$$R^\top R\hat{\theta} = R^\top Q^\top Y \quad \Leftrightarrow \quad R\hat{\theta} = Q^\top Y$$

- last step = triangular linear system (easy)

Numerical resolution, non-invertible case

Definition-Theorem (singular value decomposition): Let $A \in \mathbb{R}^{M \times N}$. Then there exist (i) $U \in \mathbb{R}^{M \times M}$ orthogonal, (ii) $V \in \mathbb{R}^{N \times N}$ orthogonal, and (iii) $\Sigma \in \mathbb{R}^{M \times N}$ diagonal with positive entries such that

$$A = U \Sigma V^\top .$$

The matrix Σ is unique up to ordering of its diagonal elements.

- we call $\sigma_i := \Sigma_{ii}$ the **singular values** of A
- they are the square roots of the eigenvalues of $A^\top A$
- only $\operatorname{rank}(A)$ of them are non-zero
- the columns of U (resp. V) are the eigenvectors of $AA^\top$ (resp. $A^\top A$)

Generalized inverse

- pseudo-inverse of a diagonal matrix:

$$\begin{pmatrix} d_1 & 0 & \cdots & 0 & 0 & \\ 0 & \ddots & \ddots & \vdots & \vdots & \cdots \\ \vdots & \ddots & \ddots & 0 & 0 & \cdots \\ 0 & \cdots & 0 & d_p & 0 & \end{pmatrix} \mapsto \begin{pmatrix} d_1^\dagger & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_p^\dagger \\ 0 & \cdots & 0 & 0 \\ & \vdots & \vdots & \end{pmatrix}$$

  where $x^\dagger = x^{-1}$ is $x \neq 0$ and 0 otherwise
- the **Moore-Penrose pseudo-inverse** of M is then defined as

$$M^\dagger = V\Sigma^\dagger U^\top .$$

  We always have $M^\dagger M M^\dagger = M^\dagger$ and $MM^\dagger M = M$.
- **Example:** if M is invertible, then $M^{-1} = M^\dagger$.
- from now on, we set $(X^\top X)^{-1} = (X^\top X)^\dagger$