Theory of Machine Learning

Prof. Damien Garreau

Julius-Maximilians-Universität Würzburg

Winter term 2024–2025

1. Course organization

Organization of the course

- **Wuestudy Course ID:** 08134700
- **Name on Wuecampus:** Theory of Machine Learning
- **Who?**
  - **Lectures:** myself
  - **Exercises:** M. Taimeskhanov
- **Format** = slides (available on Moodle after each lecture)
- **Exercises** = mostly pen and paper, regular coding (in Python)
- **Schedule:**
  1. lectures on Fridays, 4-5:30pm
  2. exercise sessions on Fridays, 2-3:30pm *(starting next week)*
- **Room:** SE 2, CAIDAS building

Evaluation

- do not forget to register to the exam
- **Evaluation:**
  - written exam at the end of the semester
  - content = definitions, similar derivations to the exercises, more ambitious problem
  - exercises sessions $\rightarrow$ bonus points
- **How does the bonus work?**
  - attend the sessions
  - send your work to Magamed at the end of the session
  - global grade $\rightarrow$ up to 10% bonus
- **Examples:** (based on 10 sessions)
  - exam = 76%, I attended all exercise sessions and made a good effort for each: I get full bonus and my final grade is $76 + 10 = 86\%$
  - exam = 96%, I attended all exercise sessions and made a good effort for each: I get full bonus and my final grade is $96 + 10 = 100\%$
  - exam = 76%, I skipped two sessions and during one session I was not paying attention and handed out something subpar: bonus = 7.5%, final grade = 83.5%

Goals and pre-requisites

- **Pre-requisites:**
  - linear algebra (matrix, eigenvectors, diagonalization)
  - analysis (derivative, gradient, global maximum)
  - probability theory (random variable, density, expectation)
  - I am glad to interrupt the lecture if some maths notion is not clear
- **Goals of the lecture:**
  - know about the **basic vocabulary**
  - look into the **details of the fundamental machine learning algorithms** (linear regression, gradient descent, etc.)
  - prove **key easy theoretical results** (*e.d.*, convergence rate for least squares)
  - **check experimentally** that these results hold

Outline I

1. Course organization
2. Introduction
   First concepts
   Empirical risk minimization
3. Linear least-square regression
   Framework
   Ordinary least-squares
   Fixed design analysis
   Ridge regression
   Random design analysis
4. Generalization bounds
   Uniform bounds via concentration
   Rademacher complexity
   Approximation error
5. Kernel methods
   Positive semi-definite kernels

Outline II

Reproducing kernel Hilbert spaces
More examples
The kernel trick and applications
The representer theorem
Kernel ridge regression
Kernel logistic regression
Generalization guarantees

Useful resources

- **Main references:**
  - *for general learning theory:* Francis Bach, *Learning Theory from First Principles*, 2023
  - *for methodology:* Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, 2001 (second edition: 2009)
  - *for kernel methods specifically:* Bernhard Schölkopf, Alexander Smola, *Learning with kernels*, MIT Press, 2002
- **Wikipedia:** as good as ever.
- **Wolfram alpha:** if you have computations to make and you do not know want to use a proper language: `https://www.wolframalpha.com/`
- **Remedials:**
  - *linear algebra:* Gilbert Strang, *Introduction to Linear Algebra*, Cambridge Press, 2009
  - *probability theory:* William Feller, *An introduction to probability theory and its applications*, Wiley, 1950

2. Introduction

2.1. First concepts

Fundamental example

- **Fundamental example:** image classification
- input = image x
- **Goal:** given any input, we want to predict which object / animal is in the image
- output = label y

$\longmapsto$ "lion"

- **Successful philosophy:** instead of defining the function f ourselves, we are going to *learn it* from data

Supervised learning

Definition: we call *predictor* (or *model*) any mapping between inputs and outputs.

- supervised learning $\to$ we will find a good predictor using *annotated* examples
- **Remark (i):** why is it difficult?
  - output may not be a deterministic function of input
  - link between the two may be incredibly complex
  - only a few observations available, potentially not where we want them
  - high dimensionality
  - ...
- **Remark (ii):** large part of machine learning: *unsupervised learning* (no annotations)
- **Examples:** clustering, dimension reduction, etc.
- out of the scope of this lecture

Input space

Definition: we call *input space* (or *domain*, or *domain set*) the set of all possible inputs of our machine learning model. We will denote it by $\mathcal{X}$.

- **Example (i):** tabular data = spreadsheet data; x has well-defined *features* such as age, `income`, `has_a_car`
- **Example (ii):** text data = ordered sequence of tokens; generally have to be pre-processed to be understood by our computer
- **Example (iii):** images = $H \times W \times C$ arrays of numbers

$$\in [\![0, 255]\!]^{299 \times 299 \times 3}$$

Input space as vector space

- **Remark:** elements $x \in \mathcal{X}$ are usually described as *vectors*
- **Reminder:** vectors are 1D arrays of number, here are two vectors with three *coordinates*:

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \qquad v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

- they can be
  - *added:* $(u + v)_i = u_i + v_i$
  - *multiplied by a number:* $(\lambda u)_i = \lambda u_i$
- vectors belong to a **vector space**, its dimension is the number of coordinates
- $\dim = d \Rightarrow$ canonical identification with $\mathbb{R}^d$
- **Intuition:** d copies of $\mathbb{R}$ with a special structure
- **Remark:** d typically high in modern machine learning
- **Example:** ImageNet images $\rightarrow 299 \times 299 \times 3 = 268,203$

Classification and regression

- we will consider two fundamental tasks: **classification** and **regression**
  - in classification, we want to associate to each $x \in \mathcal{X}$ a given *class*
  - in regression, we want to associate to each $x \in \mathcal{X}$ a given *value*
- **Example (i):** for each image on my hard drive, I want to predict what appears in it

```
{0: 'tench, Tinca tinca',
 1: 'goldfish, Carassius auratus',
 2: 'great white shark, white shark, man-eater, man-eating shark, Carcharodon carcharias',
 3: 'tiger shark, Galeocerdo cuvieri',
 4: 'hammerhead, hammerhead shark',
 5: 'electric ray, crampfish, numbfish, torpedo',
 6: 'stingray',
 7: 'cock',
 8: 'hen',
 9: 'ostrich, Struthio camelus',
```

- **Example (ii):** for each customer in my database, I want to predict how many euros he will spend next year

Labels / responses

Definition: we call *target space* (or *output space*) the set of all possible outputs of our machine learning model. We will denote it by $\mathcal{Y}$.

- **Example (i):** in image classification, $\mathcal{Y}$ is the set of all names of object and animals of the dataset
- we identify it with $\{1, 2, \ldots, 1000\} = [1000]$
- **Remark (i):** no notion of order (3 is not better than 2)
- **Remark (ii):** we will often restrict ourselves to $\mathcal{Y} = \{0, 1\}$ or $\{-1, +1\}$ for simplicity
- **Example (ii):** in regression, $\mathcal{Y} = \mathbb{R}$ (or $\mathbb{R}^k$ if we want to predict several targets simultaneously)

Training data

Definition: we call *training data* (or *training set*) a *finite* sequence of elements of $\mathcal{X} \times \mathcal{Y}$, denoted as

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \ .$$

Here, n is the size of the training set.

- **Example (i):** S is a collection of 10^6 images, each associated to the correct label
- **Example (ii):** S is a spreadsheet with the customer data from the last 25 years
- **Remark:** in real-life, there are many complications:
  - labels may be *corrupt*
  - some data ($=$ feature value for some observations) may be *missing*
- we do not consider these complications in this lecture

Machine learning algorithm

- we can now be a bit more precise:

Definition: we call *machine learning algorithm* a mapping A transforming a training set $S \in (\mathcal{X} \times \mathcal{Y})^n$ into a predictor $f : \mathcal{X} \to \mathcal{Y}$. Thus $f = A(S)$.

- of course, we want to devise a "good" algorithm
- **Question:** what does good even mean?
- **Definition that machine learning uses:** performance on new, unseen data
- there are two difficulties here: we need to define
  1. performance
  2. new, unseen data

Loss functions

Definition: we call loss function any mapping $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

- **Intuitively:** $\ell(y, y')$ measures the cost of predicting y' whereas the true target is y
- generally, we require that:
  - ℓ is symmetric;
  - ℓ has non-negative (≥ 0) values
  - $\ell(y, y) = 0$.
- **Example (i):** classification $\to$ $0 - 1$ loss

$$\ell(y, y') = \mathbb{1}_{y \neq y'} \,.$$

- here, $\mathbb{1}_E = 1$ if E is true, 0 otherwise
- **Remark:** does not matter how many classes

Loss functions

- **Example (ii):** regression $\to \mathcal{Y} \subseteq \mathbb{R} \to$ square loss

$$\ell(y, y') = (y - y')^2 \,.$$

- other possibility: absolute loss

$$\ell(y, y') = |y - y'| \;.$$

- **Other examples:** structured prediction,[1] functional regression, etc.
- **Remark (i):** in addition to the properties already lister, regression loss tend to tend to ∞ when the prediction errs far away from the ground truth
- **Remark (ii):** loss function also tend to be convex, but there are exceptions

[1] Osokin, Bach, Lacoste-Julien, *On structured prediction theory with calibrated convex surrogate losses*, NeurIPS, 2017

Expected risk: informal definition

- we model new, unseen data by a random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with distribution p
- **Intuition:** new annotated data coming from the same distribution as the training data
- **Informal definition:** expected risk is the expected loss on new data
- **Reminder:** expectation $=$ average value of a random variable
- in the discrete case, $X \in \{x_1, \ldots, x_p\}$,

$$\mathbb{E}[X] = \sum_{i=1}^{p} x_i \cdot \mathbb{P}(X = x_i) \, .$$

- **Intuition:** sum of outcome values weighted by how often they occur

Expected risk

- let us give a formal definition:

Definition: for a given data distribution p and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we define the *expected risk* (or *test error*) of a predictor $f : \mathcal{X} \to \mathcal{Y}$ as

$$\mathcal{R}(f) := \mathbb{E}\left[\ell(Y, f(X))\right] .$$

- **Remark (i):** depends on both the loss function and the data distribution p
- **Remark (ii):** hidden assumption: data distribution is equal to p...
- unfortunately, we do not know the data distribution...
- expected risk is the key quantity: ideally, we want to find f such that it is minimal

Special cases

- general definition, often specified in two key examples:
- **Binary classification:** $\mathcal{Y} = \{0, 1\}$ and $\ell(y, y') = \mathbb{1}_{y \neq y'}$, risk can be rewritten as

$$\begin{aligned}\mathcal{R}(f) = \mathbb{E}\left[\mathbb{1}_{Y \neq f(X)}\right] &= 0 \cdot \mathbb{P}\left(Y = f(X)\right) + 1 \cdot \mathbb{P}\left(f(X) \neq Y\right) \\ &= \mathbb{P}\left(f(X) \neq Y\right) .\end{aligned}$$

- **Remark:** probability of disagreement $= 1-$ accuracy
- **Regression:** $\mathcal{Y} = \mathbb{R}$ and $\ell(y, y') = (y - y')^2$

$$\mathcal{R}(f) = \mathbb{E}\left[(Y - f(X))^2\right]$$

- also known as **mean squared error** ($=$ MSE)
- in any case, *lower is better*

Expected risk

- **Example (i):** in the classification setting, consider the following predictor:

$$\forall x \in \mathcal{X}, \qquad f(x) = 1\,.$$

- let us assume balanced data, that is, $\mathbb{P}\left(Y = 0\right) = \mathbb{P}\left(Y = 1\right) = 1/2$
- then the expected risk of f is

$$\mathcal{R}(f) = \mathbb{P}\left(f(X) \neq Y\right) = \mathbb{P}\left(Y \neq 1\right) = \mathbb{P}\left(Y = 0\right) = 1/2\,.$$

- **Example (ii):** regression setting, assume that $Y = X + \varepsilon$, with $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$
- consider $f(x) = x$ (perfect predictor!)

$$\mathcal{R}(f) = \mathbb{E}\left[(Y - f(X))^2\right] = \mathbb{E}\left[(X + \varepsilon - X)^2\right] = \mathbb{E}\left[\varepsilon^2\right] = \sigma^2 > 0\,.$$

- **Reminder:** $\mathrm{Var}\left(\varepsilon\right) = \mathbb{E}\left[(\varepsilon - \mathbb{E}\left[\varepsilon\right])^2\right]$

Bayes risk

- **Question:** what is the *best* prediction function for our criterion (expected risk)?
- **Intuitively:** we want to find f that **minimizes** expected risk

Definition: we define the *Bayes risk* as the minimal possible risk over all possible predictors, for a given loss function and data distribution. Formally,

$$\mathcal{R}^\star := \inf_f \mathcal{R}(f) = \inf_f \mathbb{E}\left[\ell(Y, f(X))\right] .$$

- **Reminder:** $\inf_{x \in E} r(x)$ is the minimal value of $r(x)$ on the set E
- **Remark (i):** this is not necessarily $= 0$
- **Remark (ii):** $\mathcal{R}^\star$ is our true yardstick

Bayes predictors

- in some cases, one can actually give predictors achieving $\mathcal{R}^\star$

Definition: we call *Bayes predictor* any predictor with minimal risk and denote it by $f^\star$. Formally,

$$\mathcal{R}(f^\star) = \mathcal{R}^\star \left(= \inf_f \mathcal{R}(f) = \inf_f \mathbb{E}\left[\ell(Y, f(X))\right] \right) .$$

- **Question:** how do we do that?
- first step $=$ using the **tower property**: let g be a predictor,

$$\mathcal{R}(g) = \mathbb{E}_{x\sim p}[\mathbb{E}\left[\ell(Y, g(x)) \mid X = x\right]]$$

Reminder: conditional probability

Proposition: given two events A and B such that $\mathbb{P}(B) \neq 0$, we define the *conditional probability* of A "given" B by

$$\mathbb{P}(A \,|\, B) := \frac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)} .$$

- **Example:** let us consider two Bernoulli with parameter $1/2$, A_1 and A_2
- we can compute

$$\mathbb{P}(A_1 + A_2 = 1 \,|\, A_1 = 0) = \frac{\mathbb{P}(A_1 + A_2 = 1 \text{ and } A_1 = 0)}{\mathbb{P}(A_1 = 0)} = \frac{\mathbb{P}(A_1 = 0 \text{ and } A_2 = 1)}{\mathbb{P}(A_1 = 0)}$$
$$= \frac{1/4}{1/2} = \frac{1}{2} .$$

Reminder: conditional expectation

Proposition: let X and Y be discrete rndom variables. The *conditional expectation* of X given Y is given by

$$\mathbb{E}\left[X \mid Y = y\right] = \sum_x x \cdot \mathbb{P}\left(X = x \,|\, Y = y\right) .$$

- **Remark:** undefined if $\mathbb{P}\left(Y = y\right) = 0$ (but still possible for continuous random variables)
- **Example:**

$$\begin{aligned}\mathbb{E}\left[A_1 + A_2 \mid A_1 = 0\right] &= 0 \cdot \mathbb{P}\left(A_1 + A_2 = 0 \,|\, A_1 = 0\right) + 1 \cdot \mathbb{P}\left(A_1 + A_2 = 1 \,|\, A_1 = 0\right) \\ &\quad + 2 \cdot \mathbb{P}\left(A_1 + A_2 = 2 \,|\, A_1 = 0\right) \\ &= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} + 2 \cdot 0 = \frac{1}{2} .\end{aligned}$$

Reminder: tower property

Proposition: Let X and Y be two random variables. Then $\mathbb{E}_Y[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$.

- **Proof (in the discrete case):** using the previous slide:

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}[X \mid Y]] &= \sum_y \left(\sum_x x \cdot \mathbb{P}(X = x \mid Y = y) \right) \mathbb{P}(Y = y) \\
&= \sum_x x \cdot \sum_y \mathbb{P}(X = x \mid Y = y) \mathbb{P}(Y = y) \\
&= \sum_x x \cdot \sum_y \mathbb{P}(X = x, Y = y) \\
&= \sum_x x \cdot \mathbb{P}(X = x) \\
\mathbb{E}_Y[\mathbb{E}[X \mid Y]] &= \mathbb{E}[X] \qquad \square
\end{aligned}$$

Back to Bayes predictors

- according to the tower property:

$$\mathcal{R}(g) = \mathbb{E}_{x\sim p}[\mathbb{E}\left[\ell(Y, g(x)) \mid X = x\right]]$$

- **Remark:** $\mathbb{E}\left[\ell(Y, g(x)) \mid X = x\right]$ is also sometimes called the *conditional risk*
- we can *define* $f^\star$ such that, for all $x \in \mathcal{X}$, it minimizes

$$C(g, x) := \mathbb{E}\left[\ell(Y, g(x)) \mid X = x\right] .$$

- by positivity of the integral, this gives us the best possible risk

Bayes predictors

- summarizing everything:

Proposition: The expected risk is minimized at a *Bayes predictor* $f^\star : \mathcal{X} \to \mathcal{Y}$ satisfying for all $x \in \mathcal{X}$

$$f^\star(x) \in \arg\min_{z \in \mathcal{Y}} \mathbb{E}\left[\ell(Y, z) \mid X = x\right] .$$

All Bayes predictor have the same risk, equal to the Bayes risk. It can be computed as

$$\mathcal{R}^\star = \mathbb{E}_{x \sim p}\left[\inf_{z \in \mathcal{Y}} \mathbb{E}\left[\ell(Y, z) \mid X = x\right]\right] .$$

- **Remark:** $f^\star$ seems complicated to compute... and it is
- we can still get some interesting statements

Examples

- **Binary classification:** for the $0-1$ loss, Bayes predictor can be written

$$f^\star(x) \in \arg\min_{z\in\{0,1\}} \mathbb{P}\left(Y \neq z \,|\, X = x\right) = \arg\max_{z\in\{0,1\}} \mathbb{P}\left(Y = z \,|\, X = x\right) \, .$$

- set $\eta(x) = \mathbb{P}\left(Y = 1 \,|\, X = x\right)$, then $f^\star(x) = \mathbb{1}_{\eta(x) > 1/2}$
- Bayes risk is equal to

$$\mathcal{R}^\star = \mathbb{E}\left[\min(\eta(x), 1 - \eta(x))\right] \, .$$

- **Regression:** for the square loss, Bayes predictor is such that

$$f^\star(x) \in \arg\min_{z\in\mathbb{R}} \mathbb{E}\left[(Y - z)^2 \mid X = x\right] = \mathbb{E}\left[Y \mid X = x\right]$$

2.2. Empirical risk minimization

Empirical risk

- **Reminder:** we do not have access to data distribution

Definition: for fixed training data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$, we define the *empirical risk* of a predictor $f : \mathcal{X} \to \mathcal{Y}$ as

$$\hat{\mathcal{R}}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) \, .$$

- **Intuition:** good proxy for $\mathcal{R}$ if n is large enough:

$$\hat{\mathcal{R}}(f) \approx \mathcal{R}(f) \, .$$

Empirical risk minimization

- let $\mathcal{H}$ be a class of models
- ideally, we would like to find
$$f^\star \in \operatorname*{arg\,min}_{h\in\mathcal{H}} \mathcal{R}(h)\,.$$
- **Problem:** we do not know p... and even if we did it would still be a very difficult problem
- **Idea:** replace $\mathcal{R}$ by the empirical risk
- this leads to empirical risk minimization (ERM):[2]

$$\boxed{\hat{f} \in \operatorname*{arg\,min}_{f\in\mathcal{H}} \hat{\mathcal{R}}(f) = \operatorname*{arg\,min}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i))\,.}$$

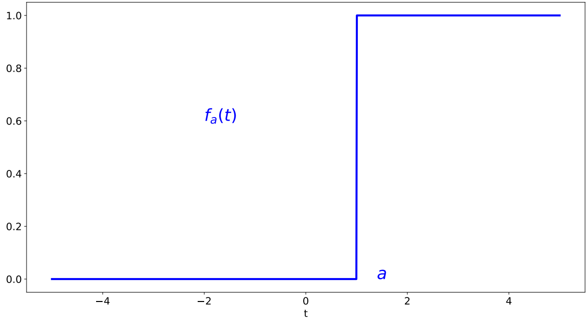
[2]Vapnik, *Principles of risk minimization for learning theory*, NIPS, 1991

Empirical risk minimization: example

- let us give a simple example
- take $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0,1\}$, 0-1 loss, and "bump functions:"

$$\mathcal{H} = \left\{ f_a : \mathbb{R} \to \mathbb{R}, \forall t \in \mathbb{R}, f_a(t) = \mathbb{1}_{t \geq a} \right\} .$$

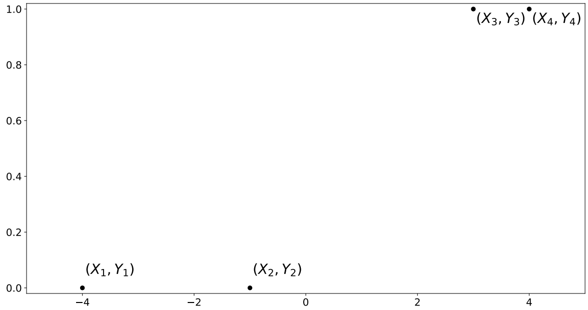
- **Visually,** elements of $\mathcal{H}$ look like:



Empirical risk minimization: example

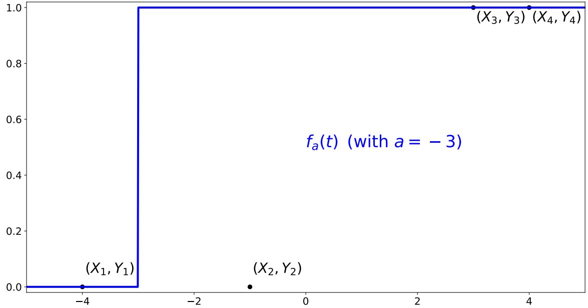
- take the following datapoints:

$$(X_1, Y_1) = (-4, 0),\ (X_2, Y_2) = (-1, 0),\ (X_3, Y_3) = (3, 1),\ (X_4, Y_4) = (4, 1)\,.$$



Empirical risk minimization: example

- for each candidate f_a, we can compute the associated empirical risk:

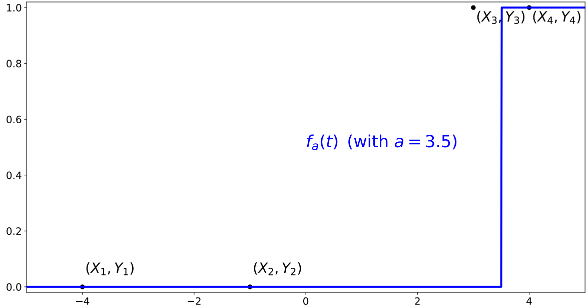


- here we have

$$\hat{\mathcal{R}}(f_a) = \frac{1}{4}(0 + 1 + 0 + 0) = \frac{1}{4}\,.$$

Empirical risk minimization: example

- for each candidate f_a, we can compute the associated empirical risk:

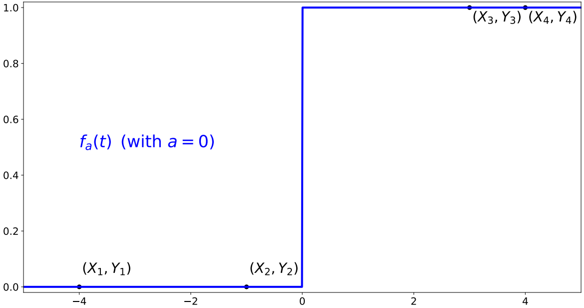


- here we have

$$\hat{\mathcal{R}}(f_a) = \frac{1}{4}(0 + 0 + 1 + 0) = \frac{1}{4}\,.$$

Empirical risk minimization: example

- we notice that several candidates achieve empirical risk $= 0$:

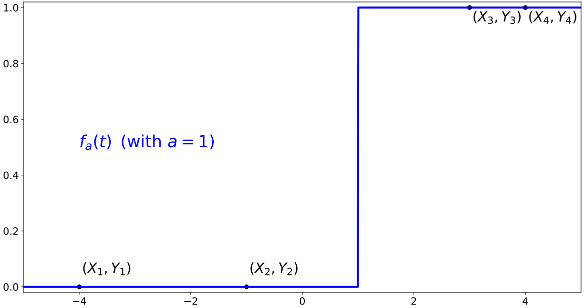


- here we have

$$\hat{\mathcal{R}}(f_a) = \frac{1}{4}(0 + 0 + 0 + 0) = 0\,.$$

Empirical risk minimization: example

- we notice that several candidates achieve empirical risk $= 0$:

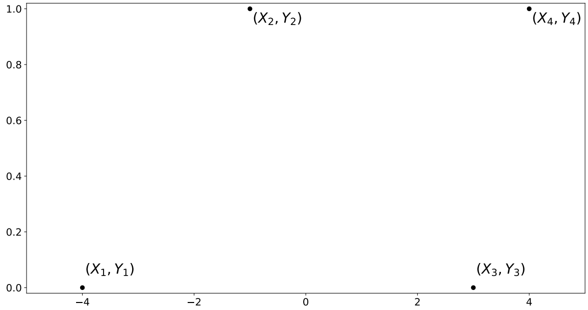


- here we have

$$\hat{\mathcal{R}}(f_a) = \frac{1}{4}(0 + 0 + 0 + 0) = 0\,.$$

Empirical risk minimization: example

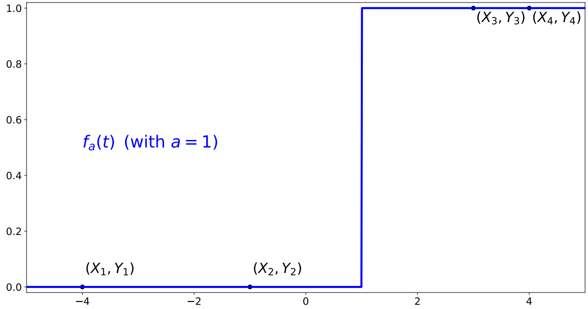
- f_a with $a \in (-1, 3)$ are all **empirical risk minimizers**
- we can pick any of them
- not always the case:



- **Question:** can you find a candidate with empirical risk $= 0$?

Generalization

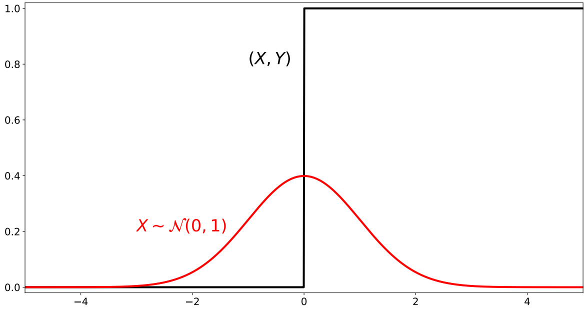
- back to the “separable” case:



- **Question:** does $\hat{\mathcal{R}}(f) = 0$ say something about $\mathcal{R}(f)$?

Generalization

- **Answer:** it depends (on the true data distribution)
- **Example:** assume $X \sim \mathcal{N}(0, 1)$, and $Y = \mathbb{1}_{X \geq 0}$



- we can compute the (true) risk for different candidates

Generalization

- **Example:**

$$
\begin{aligned}
\mathcal{R}(f_1) &= \mathbb{P}\left(f_1(X) \neq Y\right) && \text{(definition of the risk)} \\
&= \mathbb{P}\left(\mathbb{1}_{X \geq 1} \neq \mathbb{1}_{X \geq 0}\right) && \text{(definition of } f_a \text{ and data distribution)} \\
&= \mathbb{P}\left(X \in [0, 1]\right) \\
&= \frac{1}{\sqrt{2\pi}} \int_0^1 \mathrm{e}^{\frac{-x^2}{2}} \mathrm{d}x && \text{(density of a } \mathcal{N}\left(0, 1\right)\text{)} \\
\mathcal{R}(f_1) &\approx 0.34
\end{aligned}
$$

- this is not zero!
- one predictor, though, has zero risk in that case: f_0
- it is the **Bayes predictor**

Overfitting

- **Problem:** in extreme cases, this can be a severe issue
- this is in particular true when the hypotheses class $\mathcal{H}$ is too large
- **Example:** assume $\mathcal{H}$ is the set of all measurable functions
- consider a fixed training set (x_i, y_i) and let

$$h(x) = \begin{cases} y_i & \text{if } \exists i \in \{1, \ldots, n\} \text{ s.t. } x = x_i \\ 0 & \text{otherwise.} \end{cases}$$

- in particular, $h \in \mathcal{H}$ (since $\mathcal{H}$ contains all functions), and

$$\forall i \in [n], \quad h(x_i) = y_i \, .$$

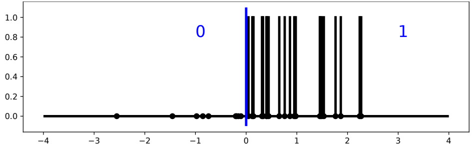
- in that case,

$$\hat{\mathcal{R}}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{h(x_i) \neq y_i} = 0 \, .$$

- empirical risk $= 0$ (interpolating)

Overfitting, ctd.

- **As in the previous example:** assume $Y = \mathbb{1}_{X \geq 0}$ and $X \sim \mathcal{N}(0, 1)$
- h looks like:



- since X has a density, $\mathbb{P}(X = x_i) = 0$
- thus we will always predict 0 on new datapoints
- let us compute the true risk:

$$\mathcal{R}(h) = \mathbb{P}(h(X) \neq Y) = \mathbb{P}(0 \neq \mathbb{1}_{X \geq 0}) = 1/2\,.$$

- this is essentially the **worst we can get**, despite having 0 training error

How to prevent overfitting?

- **Solution I:** reduce size of $\mathcal{H}$
- typical situation: parameterized space $f_\theta : \mathcal{X} \to \mathcal{Y}$, with $\theta \in \Theta$
- in this situation, ERM becomes

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta \in \Theta} \hat{\mathcal{R}}(f_\theta) = \operatorname*{arg\,min}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))$$

- we can control the number of parameters
- **Solution II:** regularize (not exclusive), that is, minimize

$$\hat{\mathcal{R}}(f_\theta) + \lambda \Omega(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)) + \lambda \Omega(\theta) \,.$$

- **Example:** $\Omega(\theta) = \lambda \left\| \theta \right\|^2$ with $\lambda > 0$ some hyperparameter

Empirical risk minimization: summary

- **Pros:**
  - general framework
  - can be solved approximately when $\mathcal{H}$ is parameterized
- **Cons:**
  - non-separable data
  - non-convexity $\to$ optimization problem can be hard
  - overfitting
- **Other approaches:** local averaging
- **Idea:** we know $\mathbb{E}[Y \mid X = x]$ or $\mathbb{P}(Y = 1 \mid X = x)$ are "the best we can do"
- $\to$ let us approximate them directly
- typical example = k-nearest neighbors[3]

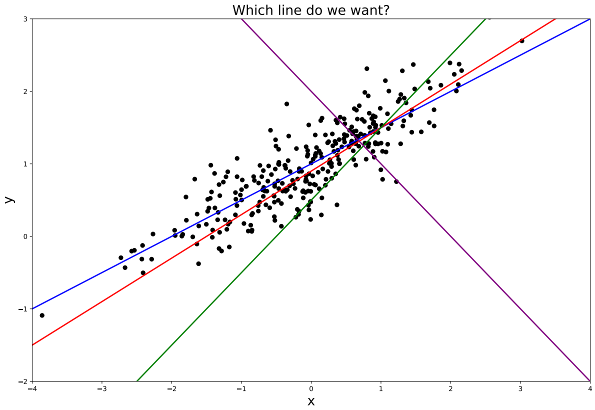
[3] Fix, Hodges, *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*, USAF report, 1951

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 $f^\star(x) = \mathbb{E}[Y \mid X = x]$
- **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**
- $\rightarrow$ 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 $Y := (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$ 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

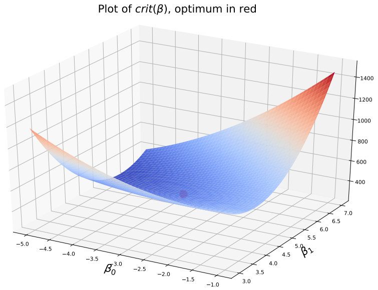
3.2. Ordinary least-squares

Ordinary Least Squares

- **Reminder:** we want to minimize

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

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



Calculus aparte

- **Reminder:** let $f : \mathbb{R}^N \to \mathbb{R}^M$, 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 $1 \in \{1, \dots, N\}$, 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

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

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 $(X^\top X)^{-1} = (X^\top X)^\dagger$

Conclusion on least squares

- now we can look at the solutions:

Theorem (James, 1978): Let $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$. If $AA^\dagger b = b$, the complete set of solutions of $Ax = b$ is given by

$$z = A^\dagger b + (\mathrm{I}_d - A^\dagger A) w \,,$$

for $w \in \mathbb{R}^d$.

- $A^\dagger A$ is an orthogonal projection, $I_d - A^\dagger A$ is the orthogonal projection on $\mathrm{Im}(A^\dagger A)^\perp$ and

$$\begin{aligned} \|A^\dagger b + (\mathrm{I}_d - A^\dagger A) w\|^2 &= \|(A^\dagger A) A^\dagger b + (\mathrm{I}_d - A^\dagger A) w\|^2 \\ &= \|A^\dagger b\|^2 + \|(\mathrm{I}_d - A^\dagger A) w\|^2 \,. \end{aligned}$$

- taking the Moore-Penrose pseudo-inverse guarantees that **we take the solution with smallest Euclidean norm**.

Gradient descent

- yet another possibility: gradient descent
- **Idea:** minimize $\hat{\mathcal{R}}$ following the steepest descent line
- formally, build the sequence of iterates

$$\begin{cases} \theta^{(0)} & = \theta_0 \\ \theta^{(t+1)} & = \theta^{(t)} - \gamma \nabla \hat{\mathcal{R}}(\theta^{(t)}) \end{cases}$$

  with $\gamma > 0$ the *stepsize*
- if convergence, then $\nabla \hat{\mathcal{R}} = 0$: minimizer
- computational complexity for each step is reduced to $\mathcal{O}(d)$
- it T steps, with $T \ll d^2$, much faster

3.3. Fixed design analysis

Setting

- **Fixed design:** in this section, we assume that Φ is *deterministic*
- namely, fixed, deterministic $x_1, \ldots, x_n \in \mathcal{X}$
- **Assumption I:** there exists $\theta^\star \in \mathbb{R}^d$ such that

$$\forall i \in [n], \qquad Y_i = \varphi(x_i)^\top \theta^\star + \varepsilon_i \, ,$$

  with ε_i noise variables
- in matrix notation, we still have:

$$Y = \Phi\theta^\star + \varepsilon \, .$$

- **Assumption II:** the ε_is are independent, have zero mean, and variance $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$
- **Remark (i):** we do not assume identically distributed
- **Remark (ii):** variance assumption is sometimes called *homoscedasticity*

Mahalanobis distance

- for any positive-definite matrix A, we set

$$\forall u \in \mathbb{R}^d, \qquad \|u\|_A^2 := u^\top A u\,.$$

- **Remark (i):** taking $A = \mathrm{I}$, we recover the Euclidean norm
- **Remark (ii):** intuition when A is diagonal: weighting the features
- the function

$$d_A(x, y) := \|x - y\|_A$$

  is often called *Mahalanobis distance*

Excess risk

- under our assumptions, we now turn to the computation of the Bayes risk and excess risk of ordinary least squares
- **Definition:** excess risk = true risk − Bayes risk
- **Notation:** we set $\hat{\Sigma} := \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d\times d}$ the (empirical) covariance matrix

Proposition (excess risk of OLS): under assumptions I and II, for any $\theta \in \mathbb{R}^d$, we have $\mathcal{R}^\star = \sigma^2$ and

$$\mathcal{R}(\theta) - \mathcal{R}^\star = \|\theta - \theta^\star\|_{\hat{\Sigma}}^2 \, .$$

- **Remark (i):** in the presence of noise ($\sigma^2 > 0$), the Bayes risk is positive
- **Remark (ii):** excess risk is the squared distance between our parameter and the true parameter in the geometry defined by $\hat{\Sigma}$

Excess risk, ctd.

Proof: we know that $Y = \Phi\theta^\star + \varepsilon$, thus

$$
\begin{aligned}
\mathcal{R}(\theta) &= \mathbb{E}\left[\frac{1}{n}\|Y - \Phi\theta\|^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\|\Phi\theta^\star + \varepsilon - \Phi\theta\|^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\|\Phi(\theta^\star - \theta)\|^2 + 2\varepsilon^\top\Phi(\theta^\star - \theta) + \|\varepsilon\|^2\right] \\
&= \sigma^2 + \frac{1}{n}(\theta - \theta^\star)^\top\Phi^\top\Phi(\theta - \theta^\star)\,. \qquad (\mathbb{E}[\varepsilon_i] = 0,\ \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2)
\end{aligned}
$$

Since $\hat{\Sigma}$ is invertible, $\theta^\star$ is the unique global minimizer and the minimum value is σ^2. $\square$

Bias / variance decomposition

Proposition (bias-variance): Let $\hat\theta \in \mathbb{R}^d$. Then, under assumption I and II,

$$\underbrace{\mathbb{E}\left[\mathcal{R}(\hat\theta)\right] - \mathcal{R}^\star}_{\text{expected excess risk}} = \underbrace{\left\|\mathbb{E}[\hat\theta] - \theta^\star\right\|_{\hat\Sigma}^2}_{\text{bias}} + \underbrace{\mathbb{E}\left[\left\|\hat\theta - \mathbb{E}[\hat\theta]\right\|_{\hat\Sigma}^2\right]}_{\text{variance}}$$

Proof: using the previous proposition:

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\hat\theta)\right] - \mathcal{R}^\star &= \mathbb{E}\left[\left\|\hat\theta - \theta^\star\right\|_{\hat\Sigma}^2\right] \\ &= \mathbb{E}\left[\left\|\hat\theta - \mathbb{E}[\hat\theta] + \mathbb{E}[\hat\theta] - \theta^\star\right\|_{\hat\Sigma}^2\right],\end{aligned}$$

then develop. $\square$

Expectation and variance

- **Reminder:** the OLS solution is given by

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

Proposition (mean and variance of OLS): Let $\hat{\theta}$ be the OLS solution. Assume I and II. Then $\hat{\theta}$ satisfies

$$\mathbb{E}[\hat{\theta}] = \theta^\star \qquad \text{and} \qquad \mathrm{Var}(\hat{\theta}) = \frac{\sigma^2}{n}\hat{\Sigma}^{-1} \, .$$

- **Remark (i):** in the language of statistics, we say that $\hat{\theta}$ is an *unbiased estimator* of $\theta^\star$
- **Remark (ii):** the matrix $\hat{\Sigma}^{-1}$ is sometimes called the *precision* matrix

Expectation and variance, proof

Proof: We know that $\mathbb{E}\left[Y\right] = \Phi\theta^\star$, thus

$$\mathbb{E}[\hat{\theta}] = (\Phi^\top\Phi)^{-1}\Phi^\top\Phi\theta^\star = \theta^\star\,.$$

We deduce that

$$\begin{aligned}\hat{\theta} - \theta^\star &= (\Phi^\top\Phi)^{-1}\Phi^\top(\Phi\theta^\star + \varepsilon) - \theta^\star \\ &= (\Phi^\top\Phi)^{-1}\Phi^\top\varepsilon\,,\end{aligned}$$

from which we compute the variance

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) &= \mathbb{E}\left[(\Phi^\top\Phi)^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi(\Phi^\top\Phi)^{-1}\right] \\ &= \sigma^2(\Phi^\top\Phi)^{-1}(\Phi^\top\Phi)(\Phi^\top\Phi)^{-1} && (\mathbb{E}\left[\varepsilon_i\varepsilon_j\right] = \sigma^2\mathbb{1}_{i=j}) \\ &= \sigma^2(\Phi^\top\Phi)^{-1}\,.\end{aligned}$$

$\square$

Excess risk of OLS

Proposition (expected excess risk of OLS): Assume I and II. Then the (expected) excess risk of the ERM is equal to

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star = \frac{\sigma^2 d}{n} .$$

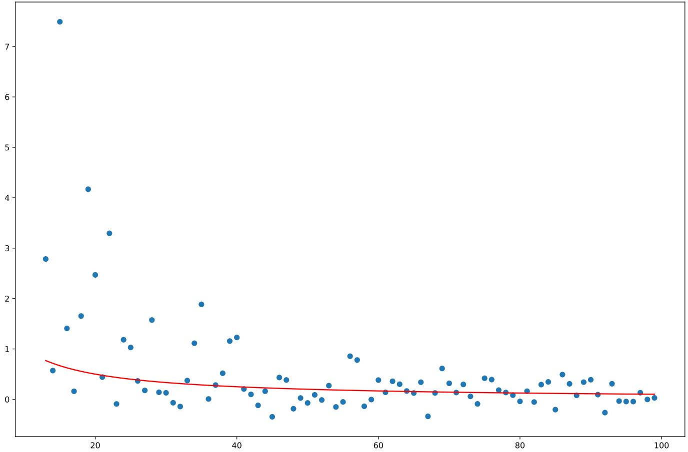
- **Remark (i):** decreasing when $n \to +\infty$ (consistency)
- **Remark (ii):** but, for fixed n, quite bad when $d \approx n$...
- **Remark (iii):** one can show that

$$\mathbb{E}\left[\hat{\mathcal{R}}(\hat{\theta})\right] = \frac{n-d}{n}\sigma^2 = \sigma^2 - \frac{d}{n}\sigma^2 ,$$

  thus training error *underestimates* test error, which is

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] = \sigma^2 + \frac{d}{n}\sigma^2 .$$

Excess risk of OLS, illustration



- **Figure:** excess risk as a function of n (one simulation per n). Gaussian noise, dimension 10, $\theta^\star = \mathbb{1}$. In red, the expected value $\sigma^2 d/n$.

Excess risk of OLS, proof

Proof: Using our previous computations:

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star &= \mathbb{E}\left[\left\|\hat{\theta} - \theta^\star\right\|_{\hat{\Sigma}}^2\right] \\
&= \mathbb{E}\left[\mathrm{trace}\left((\hat{\theta} - \theta^\star)^\top \hat{\Sigma}(\hat{\theta} - \theta^\star)\right)\right] && \text{(definition of } \|\cdot\|_{\hat{\Sigma}}\text{)} \\
&= \mathbb{E}\left[\mathrm{trace}\left((\hat{\theta} - \theta^\star)(\hat{\theta} - \theta^\star)^\top \hat{\Sigma}\right)\right] && \text{(cyclic property of the trace)} \\
&= \mathrm{trace}\left(\mathrm{Var}(\hat{\theta})\hat{\Sigma}\right) && \text{(linearity)} \\
&= \mathrm{trace}\left(\frac{\sigma^2}{n}\hat{\Sigma}^{-1}\hat{\Sigma}\right) && \text{(variance computation)} \\
&= \frac{\sigma^2}{n}\mathrm{trace}\left(\mathrm{I}_d\right)
\end{aligned}
$$

$\square$

3.4. Ridge regression

Introduction

- **Reminder:** when $n \approx d$, OLS does not fare too good
- even more complicated when $d > n$
- yet, this is a common occurrence
- **Possible solution:** L^2 regularization

Definition: let $\lambda > 0$. With our notation, the ridge least-squares estimator $\hat{\theta}_\lambda$ is defined as the minimizer of

$$\frac{1}{n}\left\| Y - \Phi\theta \right\|^2 + \lambda \left\| \theta \right\|^2 .$$

- one can easily show the following:

Proposition: we have $\hat{\theta}_\lambda = \frac{1}{n}(\hat{\Sigma} + \lambda \, \mathrm{I}_d)^{-1} \Phi^\top Y$.

A note on invertibility

- in the previous proposition we inverted the matrix $M := \hat{\Sigma} + \lambda \, \mathsf{I}_d$
- **Why can we do that?**
- $\hat{\Sigma}$ is positive semi-definite, $\lambda \, \mathsf{I}_d$ "pushes" the spectrum in $\mathbb{R}_+^\star$
- more rigorously, if M was not invertible, one would have

$$\det \left(\frac{1}{n} \Phi^\top \Phi + \lambda \, \mathsf{I}_d \right) = 0 \, .$$

- meaning that $-\lambda$ would be an eigenvalue of $\Phi^\top \Phi$: this is not possible
- **Note:** this was the main motivation when first introduced[5]

[5] Hoerl, Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, 1970

Fixed design analysis

- as with OLS, we can compute the expected excess risk
- only a bit more complicated because of the regularization...
- bias-variance decomposition still holds:

Proposition (ridge bias-variance decomposition): Let $\hat{\theta}_\lambda$ as before. Under assumption I and II,

$$\mathbb{E}[\mathcal{R}(\hat{\theta}_\lambda)] - \mathcal{R}^\star = \left\| \mathbb{E}[\hat{\theta}_\lambda] - \theta^\star \right\|_{\hat{\Sigma}}^2 + \mathbb{E}\left[\left\| \hat{\theta}_\lambda - \mathbb{E}[\hat{\theta}_\lambda] \right\|_{\hat{\Sigma}}^2 \right]$$

- *Proof:* did not depend on $\hat{\theta}$'s exact expression □

Rewriting $\mathbb{E}[\hat{\theta}_\lambda]$

- we will then use the following:

Lemma: Let $\hat{\theta}_\lambda$ be the ridge regressor. Assume that I and II hold. Then

$$\mathbb{E}[\hat{\theta}_\lambda] = \theta^\star - \lambda(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\theta^\star .$$

- *Proof:*

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_\lambda] &= \mathbb{E}\left[\frac{1}{n}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\Phi^\top Y\right] && \text{(def. of } \hat{\theta}_\lambda\text{)} \\
&= \mathbb{E}\left[\frac{1}{n}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\Phi^\top(\Phi\theta^\star + \varepsilon)\right] && \text{(assumption I)} \\
&= \frac{1}{n}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\Phi^\top\Phi\theta^\star && \text{(linearity + } \varepsilon \text{ centered)}
\end{aligned}$$

Rewriting $\mathbb{E}[\hat{\theta}_\lambda]$

- now, by definition of $\hat{\Sigma}$,

$$\mathbb{E}[\hat{\theta}_\lambda] = (\hat{\Sigma} + \lambda \,\mathsf{I}_d)^{-1}\hat{\Sigma}\theta^\star \,.$$

- finally, since for any matrix A

$$(A + \lambda \,\mathsf{I})^{-1}A = \mathsf{I} - \lambda(A + \lambda \,\mathsf{I})^{-1} \,,$$

  we deduce the result. $\square$

Excess risk

Proposition (ridge excess risk): assume I and II, let $\hat{\theta}_\lambda$ as before. Then

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta}_\lambda)\right] - \mathcal{R}^\star = \lambda^2 \left(\theta^\star\right)^\top (\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2} \hat{\Sigma} \theta^\star + \frac{\sigma^2}{n} \operatorname{trace}\left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2}\right) .$$

- **Remark (i):** when $\lambda \to 0$, we recover the OLS result
- **Remark (ii):** we have an exact description of the bias / variance evolution w.r.t. λ (!)
- **Remark (iii):** bias increases with λ, variance decreases, $\lambda = 0$ not optimal (in general)
- **Remark (iv):** the quantity $\operatorname{trace}\left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2}\right)$ is called "degrees of freedom" $\approx$ implicit number of parameters

Excess risk, proof

- *Proof:* we plug the alternative expression of $\mathbb{E}[\hat{\theta}_\lambda]$ into the bias / variance decomposition
- the bias term is clear, variance yields

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{\theta}_\lambda - \mathbb{E}[\hat{\theta}_\lambda]\right\|_{\hat{\Sigma}}^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\Phi^\top \varepsilon\right\|_{\hat{\Sigma}}^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\mathrm{trace}\left(\varepsilon^\top \Phi(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\Phi^\top \varepsilon\right)\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\mathrm{trace}\left(\Phi^\top \varepsilon\varepsilon^\top \Phi(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\right)\right] \\
&\qquad \text{(trace cyclic property)} \\
&= \frac{\sigma^2}{n}\mathrm{trace}\left(\hat{\Sigma}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\right) . \qquad \left(\mathbb{E}\left[\varepsilon\varepsilon^\top\right] = \sigma^2 \mathrm{I}_d\right)
\end{aligned}
$$

Excess risk, proof

- finally, since
$$(\hat{\Sigma} + \lambda \mathsf{I}_d)(\hat{\Sigma} + \lambda \mathsf{I}_d)^{-1} = (\hat{\Sigma} + \lambda \mathsf{I}_d)^{-1}(\hat{\Sigma} + \lambda \mathsf{I}_d) = \mathsf{I}_d \,,$$
we deduce that
$$\hat{\Sigma}(\hat{\Sigma} + \lambda \mathsf{I}_d)^{-1} = (\hat{\Sigma} + \lambda \mathsf{I}_d)^{-1}\hat{\Sigma} \left(= \mathsf{I}_d - \lambda(\hat{\Sigma} + \lambda \mathsf{I}_d)^{-1} \right) \,.$$
- together with the trace cyclic property, this allows us to write
$$\mathrm{trace}\left(\hat{\Sigma}(\hat{\Sigma} + \lambda \mathsf{I}_d)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda \mathsf{I}_d)^{-1}\right) = \mathrm{trace}\left(\hat{\Sigma}^2(\hat{\Sigma} + \lambda \mathsf{I}_d)^{-2}\right)$$
and to conclude. $\square$

Choice of regularization

Proposition (choice of regularization parameter): Assume that I and II hold. Set

$$\lambda^\star := \frac{\sigma \operatorname{trace}(\hat{\Sigma})^{1/2}}{\|\theta^\star\| \sqrt{n}}$$

as regularization parameter. Then

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta}_{\lambda^\star})\right] - \mathcal{R}^\star \leq \frac{\sigma \operatorname{trace}(\hat{\Sigma})^{1/2} \|\theta^\star\|}{\sqrt{n}} \, .$$

- **Remark (i):** of course, in practice, we know neither σ, nor $\theta^\star$...
- **Remark (ii):** $\lambda^\star$ maybe not optimal for the true risk
- **Remark (iii):** slower rate of convergence, but σ instead of σ^2

Choice of regularization, proof

- we take for granted that all eigenvalues of $\lambda(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2}\hat{\Sigma}$ are smaller than $1/2$
- as a consequence:

$$
\begin{aligned}
B &= \lambda^2 (\theta^\star)^\top (\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2} \hat{\Sigma} \theta^\star \\
&= \lambda (\theta^\star)^\top \left[\lambda (\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2} \hat{\Sigma} \right] \theta^\star \\
&\leq \frac{\lambda}{2} \left\| \theta^\star \right\|^2 .
\end{aligned}
$$

- in the same fashion:

$$
\begin{aligned}
V &= \frac{\sigma^2}{n} \mathrm{trace} \left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2} \right) \\
&= \frac{\sigma^2}{\lambda n} \mathrm{trace} \left(\hat{\Sigma} \left(\lambda (\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2} \hat{\Sigma} \right) \right) \leq \frac{\sigma^2}{2 \lambda n} \mathrm{trace} \left(\hat{\Sigma} \right) .
\end{aligned}
$$

Proof, ctd.

- putting both bounds together, we get

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta}_\lambda)\right] - \mathcal{R}^\star \leq \frac{\lambda}{2}\left\|\theta^\star\right\|^2 + \frac{\sigma^2}{2\lambda n}\mathrm{trace}\left(\hat{\Sigma}\right) .$$

- minimizing in λ yields

$$\lambda^\star = \frac{\sigma \mathrm{trace}\left(\hat{\Sigma}\right)^{1/2}}{\left\|\theta^\star\right\|\sqrt{n}} .$$

- one readily checks the last inequality.

Dimension free bound?

- recall that our upper bound reads

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta}_{\lambda^\star})\right] - \mathcal{R}^\star \leq \frac{\sigma \operatorname{trace}(\hat{\Sigma})^{1/2} \left\|\theta^\star\right\|}{\sqrt{n}} .$$

- no explicit dependency in d
- under some assumptions (e.g., sparsity), $\|\theta^\star\| \ll d$
- moreover, if $\|\varphi(x)\| \leq R$,

$$\begin{aligned} \operatorname{trace}\left(\hat{\Sigma}\right) &= \sum_{j=1}^{d} \hat{\Sigma}_{j,j} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \varphi(x_i)_j^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\|\varphi(x_i)\right\|^2 \leq R^2 . \end{aligned}$$

3.5. Random design analysis

Random design analysis

- **Random design**: (X_i, Y_i) i.i.d. from some distribution p on $\mathcal{X} \times \mathcal{Y}$
- **Goal:** prove the same excess risk bound (i.e., $\approx \frac{\sigma^2 d}{n}$)
- **Important:** we make the same assumptions, transposed to the random design setting:
  - **Assumption I:** $\exists \theta^\star \in \mathbb{R}^d$ such that

$$\forall i \in [n], \qquad Y_i = \varphi(X_i)^\top \theta^\star + \varepsilon_i \, ,$$

  - **Assumption II:** the noise distribution of ε_i is independent from that of X_i, $\mathbb{E}\left[\varepsilon_i\right] = 0$, and $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$.
- notable consequence of our assumptions:

$$\mathbb{E}\left[Y_i \mid X_i\right] = \varphi(X_i)^\top \theta^\star \, .$$

Excess risk

- the excess risk has a similar decomposition:

Proposition (excess risk for random design least-squares regression): Assume that I and II hold. Then $\mathcal{R}^\star = \sigma^2$, and

$$\forall \theta \in \mathbb{R}^d, \qquad \mathcal{R}(\theta) - \mathcal{R}^\star = \|\theta - \theta^\star\|_\Sigma^2 \, ,$$

where $\Sigma := \mathbb{E}\left[\varphi(X)\varphi(X)^\top\right]$.

- **Intuition:** $\hat{\Sigma}$ is replace by its expectation, which is Σ
- (recall that $\hat{\Sigma} = \frac{1}{n}\Phi^\top\Phi$)

Excess risk, proof

- **Proof:** let (X_0, Y_0) be a "new" observation, with noise ε_0

$$
\begin{aligned}
\mathcal{R}(\theta) &= \mathbb{E}\left[(Y_0 - \theta^\top \varphi(X_0))^2\right] \\
&= \mathbb{E}\left[(\varphi(X_0)^\top \theta^\star + \varepsilon_0 - \theta^\top \varphi(X_0))^2\right] && \text{(AI)} \\
&= \mathbb{E}\left[(\varphi(X_0)^\top \theta^\star - \theta^\top \varphi(X_0))^2\right] + 2\mathbb{E}\left[\varepsilon_0 (\theta^\star - \theta)^\top \varphi(X_0)\right] + \mathbb{E}\left[\varepsilon_0^2\right]
\end{aligned}
$$

- by independence, and since the noise is centered,

$$
\mathbb{E}\left[\varepsilon_0 (\theta^\star - \theta)^\top \varphi(X_0)\right] = \mathbb{E}\left[\varepsilon_0\right] \mathbb{E}\left[(\theta^\star - \theta)^\top \varphi(X_0)\right] = 0 \,.
$$

- now we can conclude:

$$
\begin{aligned}
\mathcal{R}(\theta) &= \mathbb{E}\left[((\theta^\star - \theta)^\top \varphi(X_0))^2\right] + \mathbb{E}\left[\varepsilon_0^2\right] && \text{(AII)} \\
&= (\theta - \theta^\star)^\top \mathbb{E}\left[\varphi(X_0)\varphi(X_0)^\top\right] (\theta - \theta^\star) + \sigma^2 && \text{(linearity)} \\
&= (\theta - \theta^\star)^\top \Sigma (\theta - \theta^\star) + \sigma^2 \,. \quad \square && \text{(definition of } \Sigma\text{)}
\end{aligned}
$$

Excess risk of OLS

- we now use the previous result to investigate $\hat{\theta}$:

Proposition: Assume that I and II hold. Assume further that $\hat{\Sigma}$ is almost surely invertible. Then the expected excess risk of the OLS estimator is equal to

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star = \frac{\sigma^2}{n}\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\right)\right] .$$

- **Remark (i):** $\hat{\Sigma}$ has the same definition, but is now a *random* quantity
- **Remark (ii):** under reasonable assumptions (e.g., density), $\hat{\Sigma}$ is almost surely invertible
- **Intuition:** $\det(\hat{\Sigma}) = 0$ is a "zero-measure" condition

Excess risk of OLS, proof

- from the definition of $\hat{\theta}$,

$$\hat{\theta} = \frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top Y = \frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top(\Phi\theta^\star + \varepsilon) = \theta^\star + \frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\,.$$

- using the previous result:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star &= \mathbb{E}\left[\left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\right)^\top \Sigma\left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\right)\right] \\
&= \mathbb{E}\left[\mathrm{trace}\left(\Sigma\left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\right)\left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\right)^\top\right)\right] \qquad \text{(cyclic property)} \\
&= \frac{1}{n^2}\mathbb{E}\left[\mathrm{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi\hat{\Sigma}^{-1}\right)\right]
\end{aligned}$$

Excess risk of OLS, proof ctd.

- now we use properties of the conditional expectation:

$$\begin{aligned}
\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi\hat{\Sigma}^{-1}\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi\hat{\Sigma}^{-1}\right) \mid X_1,\ldots,X_n\right]\right] \\
&\qquad \text{(tower property)} \\
&= \mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\mathbb{E}\left[\varepsilon\varepsilon^\top \mid X_1,\ldots,X_n\right]\Phi\hat{\Sigma}^{-1}\right)\right] \\
&\qquad (\Phi, \hat{\Sigma} \text{ are } X_1,\ldots,X_n\text{-measurable}) \\
&= \mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\mathbb{E}\left[\varepsilon\varepsilon^\top\right]\Phi\hat{\Sigma}^{-1}\right)\right] \quad \text{(independence)} \\
&= \sigma^2\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\Phi\hat{\Sigma}^{-1}\right)\right] \qquad (\mathbb{E}\left[\varepsilon\varepsilon^\top\right] = \sigma^2\, \mathrm{I}_d) \\
&= \sigma^2\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\right)\right].
\end{aligned}$$

□

Gaussian design

- to be more precise, we need to specify a distribution for the $\varphi(X_i)$s

Proposition: Assume that I and II hold. Assume further that $\varphi(X) \sim \mathcal{N}(0, \Sigma)$. Then the expected risk of OLS is given by

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star = \frac{\sigma^2 d}{n - d - 1} \, .$$

- **Remark:** we (nearly) recover the $\sigma^2 d/n$ bound from fixed design!

Gaussian design, proof

- define $Z := \Sigma^{-1/2}\varphi(X)$
- properties of Gaussian vectors: $Z \sim \mathcal{N}(0, \mathrm{I}_d)$
- we see that

$$\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\right)\right] = \operatorname{trace}\left(\mathbb{E}\left[\Sigma(\Sigma^{1/2}Z\Sigma^{1/2}Z^\top)^{-1}\right]\right)$$
$$= \operatorname{trace}\left(\mathbb{E}\left[(ZZ^\top)^{-1}\right]\right) .$$

- $(ZZ^\top)^{-1}$ has the *inverse Wishart distribution*
- we read in the tables:

$$\mathbb{E}\left[(ZZ^\top)^{-1}\right] = \frac{1}{n-d-1}\mathrm{I}_d$$

and conclude. $\square$

4. Generalization bounds

Reminder: risk decomposition

- **Reminder:**

$$\mathcal{R}(f) - \mathcal{R}^\star = \left[\mathcal{R}(f) - \inf_{h \in \mathcal{H}} \mathcal{R}(h)\right] + \left[\inf_{h \in \mathcal{H}} \mathcal{R}(h) - \mathcal{R}^\star\right]$$

$$\text{excess risk} = \text{estimation error} + \text{approximation error}$$

- **Estimation error:**
  - always non-negative
  - random if there is randomness in the creation of f
  - characterizes how much we loose by picking the wrong predictor in a given class
- **Approximation error:**
  - deterministic, does not depend on f, **only on the class of functions** $\mathcal{H}$
  - characterizes how much we loose by restricting ourselves to a given class

Decomposition of the estimation error

- **Notation (i):** $f_{\mathcal{H}} \in \arg\min_{f \in \mathcal{H}} \mathcal{R}(f)$, best predictor in our function class
- **Notation (ii):** $\hat{f}$ empirical risk minimizer
- **Useful decomposition:**

$$\begin{aligned}
\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) &= \mathcal{R}(\hat{f}) - \mathcal{R}(f_{\mathcal{H}}) && \text{(def. of } f_{\mathcal{H}}\text{)} \\
&= \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) + \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f_{\mathcal{H}}) + \hat{\mathcal{R}}(f_{\mathcal{H}}) - \mathcal{R}(f_{\mathcal{H}}) \\
&\leq \sup_{f \in \mathcal{H}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} + \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f_{\mathcal{H}}) + \sup_{f \in \mathcal{H}} \left\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right\}
\end{aligned}$$

- middle term is ≤ 0 by definition, and we get

$$\boxed{\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) \leq 2 \sup_{f \in \mathcal{H}} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right| .}$$

Decomposition of the estimation error, ctd.

- **Remark (i):** no more dependency in $\hat{f}$, we only need to control functions (but we do need uniform control)
- **Remark (ii):** if $\hat{f}$ not global minimizer, say

$$\hat{\mathcal{R}}(\hat{f}) \leq \inf_{f \in \mathcal{H}} \hat{\mathcal{R}}(f) + \varepsilon \,,$$

  we need to add ε to our bound
- **Remark (iii):** bound usually grows with size of $\mathcal{H}$ and decreases with n

4.1. Uniform bounds via concentration

Concentration inequalities

- informally speaking: random variable is "close" to its expectation with high probability
- **Example:** Markov, Chebyshev
- more involved:

Proposition (Hoeffding's inequality): let $Z_1, \ldots, Z_n$ be independent random variables such that $Z_i \in [0,1]$ almost surely, then, for any $t \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}\left[Z_i\right])\right| \geq t\right) \leq 2\exp\left(-2nt^2\right) .$$

Single function

- assume that $\mathcal{H} = \{f_0\}$ and ℓ a bounded loss function
- then we can control

$$\sup_{f\in\mathcal{H}} \left|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right| = \hat{\mathcal{R}}(f_0) - \mathcal{R}(f_0) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_0(X_i)) - \mathbb{E}\left[\ell(Y, f_0(X))\right] .$$

- indeed, since the observations are i.i.d., we can use Hoeffding on the $Z_i := \ell(Y_i, f_0(X_i))$
- common expectation $= \mathcal{R}(f_0)$
- for any $\delta \in (0, 1/2)$,

$$\mathbb{P}\left(\left|\hat{\mathcal{R}}(f_0) - \mathcal{R}(f_0)\right| \geq \frac{1}{\sqrt{2n}}\sqrt{\log\frac{1}{\delta}}\right) \leq 2\exp\left(-2n\frac{1}{2n}\log 1/\delta\right) = 2\delta .$$

Single function

- scaling by ℓ_∞, we obtain:

Proposition: Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. observations of p and f_0 be a fixed predictor. Then, for any $\delta \in (0, 1/2)$, with probability greater than $1 - 2\delta$,

$$\mathcal{R}(f_0) - \hat{\mathcal{R}}(f_0) < \frac{\ell_\infty}{\sqrt{2n}} \sqrt{\log \frac{1}{\delta}}\,,$$

where ℓ_∞ is an upper bound on $\ell(Y_i, f(X_i))$.

From sup to expectation

- **Problem:** there is often more than one function in $\mathcal{H}$...
- still possible, using for instance:

Proposition (McDiarmid's inequality):[6] Let $Z_1, \ldots, Z_n$ be independent random variables and F a function such that

$$|F(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n) - F(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c\,.$$

Then

$$\mathbb{P}\left(|F(Z_1, \ldots, Z_n) - \mathbb{E}\left[F(Z_1, \ldots, Z_n)\right]| \geq t\right) \leq 2\exp\left(-2t^2/(nc^2)\right)\,.$$

[6] McDiarmid, *On the method of bounded differences*, Survey in Combinatorics, 1989

Application of McDiarmid

- set $Z_i := (X_i, Y_i)$, and

$$H(Z_1, \ldots, Z_n) := \sup_{f \in \mathcal{H}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} .$$

- Mc Diarmid tells us that, with probability higher than $1 - \delta$,

$$H(Z_1, \ldots, Z_n) - \mathbb{E}\left[H(Z_1, \ldots, Z_n)\right] \leq \frac{\ell_\infty \sqrt{2}}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}} .$$

- getting bound on $\mathbb{E}\left[H(Z_1, \ldots, Z_n)\right]$ automatically yields bound on $\sup_{f \in \mathcal{H}} \left\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right\}$
- by symmetry, upper bound on $\sup_{f \in \mathcal{H}} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right|$

4.2. Rademacher complexity

Rademacher complexity

- set $Z := (X, Y)$ and $\mathcal{G} := \{(x, y) \mapsto \ell(y, f(x))\}$, with f in some function class $\mathcal{H}$
- **Recall:** we want to bound

$$\sup_{f \in \mathcal{H}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} = \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}\left[g(Z)\right] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right\} .$$

- set $\mathcal{D} := \{Z_1, \ldots, Z_n\}$ the data

Definition: We call *Rademacher complexity* of the function class $\mathcal{G}$ the quantity

$$R_n(\mathcal{G}) := \mathbb{E}_{\varepsilon, \mathcal{D}} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(Z_i) \right] ,$$

where the ε_is are independent Rademacher random variables (that is, $\mathbb{P}\left(\varepsilon_i = \pm 1\right) = 1/2$).

Rademacher complexity, first properties

- **Intuition:** expectation of maximal dot-product with random labels
- measures the *capacity* of the set $\mathcal{G}$

Properties: Rademacher complexity satisfies the following properties:

- if $\mathcal{G} \subset \mathcal{G}'$, then $R_n(\mathcal{G}) \leq R_n(\mathcal{G}')$;
- $R_n(\mathcal{G} + \mathcal{G}') = R_n(\mathcal{G}) + R_n(\mathcal{G}')$;
- $R_n(\alpha \mathcal{G}) = |\alpha| \, R_n(\mathcal{G})$;
- if g_0 is a function, $R_n(\mathcal{G} + \{g_0\}) = R_n(\mathcal{G})$;
- $R_n(\mathcal{G}) = R_n(\mathrm{conv}(\mathcal{G}))$.

Symmetrization

- **Question:** why is it useful?
- Rademacher complexity directly controls expected uniform deviation

Proposition (symmetrization): With the previous notation,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\left[g(Z)\right]\right\}\right] \leq 2R_n(\mathcal{G})\,,$$

and

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right\}\right] \leq 2R_n(\mathcal{G})\,.$$

Symmetrization, proof

- let $\mathcal{D}' := \{Z'_1, \ldots, Z'_n\}$ be an independent copy of $\mathcal{D}'$
- in particular, one has $\mathbb{E}\left[g(Z'_i) \mid \mathcal{D}\right] = \mathbb{E}\left[g(Z)\right]$
- we write

$$\mathbb{E}\left[\sup_{g \in \mathcal{G}} \left\{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right\}\right] = \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left\{\mathbb{E}\left[g(Z'_i) \mid \mathcal{D}\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right\}\right]$$
$$= \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[g(Z'_i) - g(Z_i) \mid \mathcal{D}\right]\right\}\right] .$$

Symmetrization, proof ctd.

- since the sup of expectation is $\leq$ than expectation of the sup,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(Z)\right]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right\}\right]\leq\mathbb{E}\left[\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_i')-g(Z_i))\right\}\mid\mathcal{D}\right]\right]$$
$$=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_i')-g(Z_i))\right\}\right]$$

  by the tower property.
- we notice that

$$g(Z_i')-g(Z_i)\quad\text{and}\quad\varepsilon_i(g(Z_i')-g(Z_i))\text{ have the same distribution}$$

  (this is what we call symmetrization)

Symmetrization proof, ctd.

- thus

$$
\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_i') - g(Z_i))\right\}\right] &= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(g(Z_i') - g(Z_i))\right\}\right] \\
&\leq \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right\}\right] + \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}-\varepsilon_i g(Z_i)\right\}\right] \\
&= 2R_n(\mathcal{G})
\end{aligned}
$$

since ε and $-\varepsilon$ have the same distribution. $\square$

Example: linear predictors

- let Ω be a norm on $\mathbb{R}^d$
- assume $\mathcal{H} = \{\theta^\top \varphi(x), \Omega(\theta) \le D\}$
- then

$$R_n(\mathcal{H}) = \mathbb{E}\left[\sup_{\Omega(\theta)\le D} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \theta^\top \varphi(X_i)\right]$$
$$= \mathbb{E}\left[\sup_{\Omega(\theta)\le D} \frac{1}{n}\varepsilon^\top \Phi\theta\right]$$
$$= \frac{D}{n}\mathbb{E}\left[\Omega^\star(\Phi^\top \varepsilon)\right] ,$$

where $\Omega^\star$ is the *dual norm* of Ω:

$$\Omega^\star(u) := \sup_{\Omega(\theta)\le 1} u^\top \theta \, .$$

Example: linear predictors, ctd.

- **Claim:** when $p \in [1, +\infty)$ and Ω is the p-norm (*see exercise*), $\Omega^\star$ is the q-norm with $1/p + 1/q = 1$
- for the 2-norm:

$$
\begin{aligned}
R_n(\mathcal{H}) &= \frac{D}{n}\mathbb{E}\left[\left\|\Phi^\top \varepsilon\right\|\right] \\
&\leq \frac{D}{n}\sqrt{\mathbb{E}\left[\left\|\Phi^\top \varepsilon\right\|^2\right]} && \text{(Jensen's inequality)} \\
&= \frac{D}{n}\sqrt{\mathbb{E}\left[\mathrm{trace}\left(\Phi^\top \varepsilon\varepsilon^\top \Phi\right)\right]} \\
&= \frac{D}{n}\sqrt{\mathbb{E}\left[\mathrm{trace}\left(\Phi^\top \Phi\right)\right]} = \frac{D}{n}\sqrt{\sum_{i=1}^n \mathbb{E}\left[(\Phi^\top\Phi)_{i,i}\right]} = \frac{D}{n}\sqrt{\sum_{i=1}^n \mathbb{E}\left[\left\|\varphi(X_i)\right\|^2\right]} \\
&= \frac{D}{\sqrt{n}}\sqrt{\mathbb{E}\left[\left\|\varphi(x)\right\|^2\right]} \qquad \Rightarrow \text{dimension-free bound with the same rate!}
\end{aligned}
$$

Example: linear predictors, ctd.

- we can get a bound on the estimation error:

Proposition: assume that ℓ is L-Lipschitz and continuous. Consider linear predictors with bounded coefficients, that is, $f_\theta(x) = \theta^\top \varphi(x)$ with $\|\theta\| \leq D$. Assume further that $\mathbb{E}\left[\|\varphi(X)\|^2\right] \leq R^2$. Let $\hat{f}$ be the empirical risk minimizer. Then

$$\mathbb{E}\left[\mathcal{R}(\hat{f})\right] \leq \inf_{\|\theta\| \leq D} \mathcal{R}(f_\theta) + \frac{4LRD}{\sqrt{n}} \, .$$

- **Remark (i):** does not depend on exact expression of the loss
- **Remark (ii):** does not depend on the dimension

Proof of the proposition

- recall the decomposition of the estimation error:

$$\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) \leq 2 \sup_{f \in \mathcal{H}} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right| .$$

- by symmetrization:

$$\mathbb{E}\left[\mathcal{R}(\hat{f})\right] - \inf_{f \in \mathcal{H}} \mathcal{R}(f) \leq 4 R_n(\mathcal{H}) .$$

- set $\mathcal{F} := \{f_\theta, \|\theta\| \leq D\}$. Since the loss is L-Lipschitz, by contraction (*see exercise*),

$$R_n(\mathcal{H}) \leq L R_n(\mathcal{F}) .$$

- by previous computation,

$$R_n(\mathcal{F}) \leq \frac{DR}{\sqrt{n}} .$$

□

4.3. Approximation error

Further decomposition

- **Reminder:** approximation error is defined as

$$\inf_{f \in \mathcal{H}} \mathcal{R}(f) - \mathcal{R}^\star \,.$$

- deterministic, small if function class is large
- let us focus on parametric models, in particular $\mathcal{H} = \{f_\theta, \theta \in \Theta\}$
- $\theta^\star$ parameter corresponding to $f^\star$
- typically does not belong to Θ!
- further decomposition of the approximation error:

$$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \mathcal{R}^\star = \left(\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^p} \mathcal{R}(f_\theta) \right) + \left(\inf_{\theta \in \mathbb{R}^p} \mathcal{R}(f_\theta) - \mathcal{R}^\star \right) \,.$$

- **Remark:** both positive terms

Incompressible approximation error

- **Recall:**

$$\inf_{\theta\in\Theta}\mathcal{R}(f_\theta)-\mathcal{R}^\star=\left(\inf_{\theta\in\Theta}\mathcal{R}(f_\theta)-\inf_{\theta\in\mathbb{R}^p}\mathcal{R}(f_\theta)\right)+\left(\inf_{\theta\in\mathbb{R}^p}\mathcal{R}(f_\theta)-\mathcal{R}^\star\right)\,.$$

- let us start with the second term
- for rich model class, this **goes to zero**

Upper bounds

- now focus on $\inf_{\theta\in\Theta}\mathcal{R}(f_\theta) - \inf_{\theta\in\mathbb{R}^p}\mathcal{R}(f_\theta)$
- this term is typically upper bounded by a **distance** between the best candidate in Θ and the best candidate in $\mathbb{R}^d$
- **Example:** $f_\theta(x) = \theta^\top\varphi(x)$, $\Theta = \{\theta\in\mathbb{R}^d, \|\theta\|\leq D\}$
- for a L-Lipschitz loss, we write

$$\begin{aligned}\inf_{\theta\in\Theta}\mathcal{R}(f_\theta) - \inf_{\theta\in\mathbb{R}^p}\mathcal{R}(f_\theta) &= \mathbb{E}\left[\ell(\theta_1^\top\varphi(X), Y) - \ell((\theta^\star)^\top\varphi(X), Y)\right] \\ &\leq L\mathbb{E}\left[\|\varphi(X)\|\cdot\|\theta_1 - \theta^\star\|\right] \\ &\leq L\mathbb{E}\left[\|\varphi(X)\|\right]\cdot\left(\|\theta^\star\| - D\right)_+ .\end{aligned}$$

- **Remark:** equal to zero if $\|\theta^\star\|\leq D$ (well-specified model)

5. Kernel methods

5.1. Positive semi-definite kernels

Representation of the data

- **What we have seen so far:** linear classification / linear regression
- works well if the data is linearly separable
- **Problem:** that is not always the case!
- what if we could transport the data to another space where it is well-behaved?
- for instance a very high-dimensional space
- first we define a (positive-definite) *kernel*
- **many** definitions in maths, introduced in machine learning by Aizerman, Braverman, and Rozonoer in the 60s[7]

[7] Aizerman, Braverman, Rozonoer, *Theoretical foundations of the potential function method in pattern recognition learning*, Automation and Remote Control, 1964

Positive semi-definite kernels

Definition: a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *positive semi-definite kernel* if $k(x, x') = k(x', x)$ for any $x, x' \in \mathcal{X}$, and

$$\forall x_1, \ldots, x_n \in \mathcal{X}, \forall c_1, \ldots, c_n \in \mathbb{R}, \quad \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \,.$$

- in other words, the Gram matrix $K = (k(x_i, x_j))_{i,j=1}^{n}$ is positive definite for any input data $x_1, \ldots, x_n$
- *kernel methods* take this K as input
- **Remark:** this is *costly*, $\mathcal{O}\left(n^2\right)$ whatever we do, with possible dependency in the dimensionality of the data
- Beware: unlike the name suggests, k has no reason to be *positive*

Fundamental example

- suppose that $\mathcal{X} = \mathbb{R}$
- then $k(x, y) := xy$ is a positive definite kernel
- **Why?** first, we check that $k(x, y) = k(y, x)$
- second, let $n \geq 1$, $x_1, \ldots, x_n \in \mathbb{R}^d$, and $c_1, \ldots, c_n \in \mathbb{R}$, then

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j x_i x_j \\
&= \left(\sum_{i=1}^{n} c_i x_i \right)^2 \\
&\geq 0 \,.
\end{aligned}$$

Fundamental example, ctd.

- we can extend this example: set $k(x,y) := x^\top y$ on $\mathcal{X} = \mathbb{R}^d$
- let $n \geq 1$, $x_1, \ldots, x_n \in \mathbb{R}^d$, and $c_1, \ldots, c_n \in \mathbb{R}$, then

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=1}^n c_i c_j k(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n c_i c_j x_i^\top x_j \\
&= \left\| \sum_{i=1}^n c_i x_i \right\|^2 \\
&\geq 0 \,.
\end{aligned}$$

- $k(x,y) := x^\top y$ is usually called the **linear kernel**
- **Intuition:** kernels are a generalization of inner product

Other examples

- **Polynomial kernel:**
$$\mathcal{X} = \mathbb{R}^d, \qquad k(x, y) = (x^\top y + c)^k \,.$$
- **min kernel:**
$$\mathcal{X} = \mathbb{R}, \qquad k(x, y) = \min(x, y) \,.$$
- **Gaussian kernel:**
$$\mathcal{X} = \mathbb{R}^d, \qquad k(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\nu^2}\right) \,.$$
- **Exponential kernel:**
$$\mathcal{X} = \mathbb{R}^d, \qquad k(x, y) = \exp\left(\frac{-\|x - y\|}{2\nu}\right) \,.$$
- ...

Choosing the bandwidth

- Gaussian and Laplace kernel: one has to choose the bandwidth parameter ν
- indeed, if ν is *too large* with respect to the typical value of $\|x_i - x_j\|$, then $K \approx I_n$
- in the other direction, if ν is *too small*, then $K \approx \mathbf{1}\mathbf{1}^\top$
- both cases are degenerate: whatever we do with K is not going to work very well
- one possible solution: **median heuristic**[8]

$$\nu = \mathrm{Med}\{\|x_i - x_j\| , \quad 1 \leq i,j \leq n\} .$$

- preferable to the mean (too sensitive to extreme values)
- we can pick other quantiles

[8] Garreau, Jitkrittum, Kanagawa, *Large sample analysis of the median heuristic*, 2017

Hilbert spaces

> **Definition:** A *Hilbert space* is a real or complex vector space which is also a complete metric space with respect to the distance function induced by the inner product.

- **Remark:** recall the linear kernel, all we used were properties of inner product
- let $\Phi : \mathcal{X} \to \mathcal{H}$ be some mapping, $\mathcal{H}$ a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$
- then $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ is positive definite:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle \Phi(x), \Phi(y) \rangle = \left\| \sum_{i=1}^{n} c_i \Phi(x_i) \right\|^2 \geq 0 \,,$$

by linearity of the inner product.

Kernel as inner products

- **Remarkable fact:** the converse statement is true!

Theorem:[9] For any kernel k on $\mathcal{X}$, there exists a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that

$$\forall x, y \in \mathcal{X}, \qquad k(x, y) = \langle \Phi(x), \Phi(y) \rangle \,.$$

- **Reminder:** Hilbert space = inner product + *complete* for the associated norm (Cauchy sequences converge in $\mathcal{H}$)
- **Consequence:** we can think of any kernel as a dot product in the *feature space*
- **Main idea:** forget about Φ and work only with kernel evaluations (more on that later)

[9] Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society, 1950

Proof in the finite case

- assume that $\mathcal{X} = \{x_1, \ldots, x_N\}$ is finite of size N
- any kernel k is entirely defined by the $N \times N$ positive semi-definite matrix $K := (k(x_i, x_j))_{i,j=1}^N$
- we can diagonalize K in an orthonormal basis $(u_1, \ldots, u_N)$ with associated (non-negative) eigenvalues $\lambda_1, \ldots, \lambda_N$: $K = U \Lambda U^\top$, with $U_{:,i} = u_i$, $\Lambda = \operatorname{diag}(\lambda)$, $UU^\top = U^\top U = \mathrm{I}$
- then we write

$$k(x_i, x_j) = \left(\sum_{\ell=1}^N \lambda_\ell u_\ell u_\ell^\top \right)_{i,j} = \sum_{\ell=1}^N \lambda_\ell (u_\ell)_i (u_\ell)_j = \langle \Phi(x_i), \Phi(x_j) \rangle \, ,$$

with

$$\Phi(x_i) := \left(\sqrt{\lambda_1}(u_1)_i, \cdots, \sqrt{\lambda_n}(u_N)_i \right)^\top \, .$$

$\square$

5.2. Reproducing kernel Hilbert spaces

Function spaces

- among all spaces in the previous statement, one of them has interesting properties
- in particular, it is a **space of functions**
- *i.e.*, we can map each point $x \in \mathcal{X}$ to a *function* $\Phi(x) = k_x \in \mathcal{H}$
- **Example:** $\mathcal{X} = \mathbb{R}$, we map each x to the function $t \mapsto xt$
- $\rightarrow$ space of linear functions
- more complicated in general...

Reproducing Kernel Hilbert Space (RKHS)

Definition: let $\mathcal{X}$ be a set and $\mathcal{H}$ be a function space forming a Hilbert space with inner product $\langle \cdot, \cdot \rangle$. The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if

- $\mathcal{H}$ contains all functions of the form $k_x : t \mapsto k(x, t)$
- for every $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the *reproducing property* holds:

$$f(x) = \langle f, k_x \rangle \,.$$

- if a reproducing kernel exists, then $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS)

Equivalent definition

Theorem: the Hilbert space $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is a RKHS if, and only if, for any $x \in \mathcal{X}$, the mapping $f \mapsto f(x)$ is continuous.

- *Proof of* $\Rightarrow$*:* let k be a reproducing kernel, $x \in \mathcal{X}$ and $f_n \to f$ in $\mathcal{H}$
- we write

$$\begin{aligned} |f_n(x) - f(x)| &= |\langle f_n - f, k_x \rangle| \\ &\leq \|f_n - f\| \cdot \|k_x\| \end{aligned}$$

  by Cauchy-Schwarz inequality.
- $\|f_n - f\| \to 0$ and we can conclude
- **Remark:** $\|k_x\|^2 = \langle k_x, k_x \rangle = k(x, x)$, thus $|f(x)| \leq \|f\| \cdot k(x, x)^{1/2}$

Continuity ctd.

- *Proof of* $\Leftarrow$*:* let $x \in \mathcal{X}$
- by the reproducing property, $L : x \mapsto f(x)$ is a *linear functional*
- Riesz theorem: there exists ℓ_x such that $L(x) = \langle f, \ell_x \rangle$
- define $k(x, y) := \ell_y(x)$
- one can check readily the RKHS properties. $\square$

Uniqueness

Theorem: if $\mathcal{H}$ is a RKHS, then it has a unique reproducing kernel. Conversely, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be the reproducing kernel of at most one RKHS.

- we talk about *the* RKHS associated to k
- *Proof:* let k and k' be two reproducing kernels
- then for all $x \in \mathcal{X}$,

$$\begin{aligned}\|k_x - k'_x\|^2 &= \langle k_x - k'_x, k_x - k'_x \rangle \\ &= k_x(x) - k'_x(x) - k_x(x) + k'_x(x) \\ &= 0\end{aligned}$$

$\square$

Equivalence psd / RKHS

Theorem: a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if, and only if, it is a reproducing kernel.

- **Idea:** build $\mathcal{H}$ as the completion of

$$\mathcal{H}_0 := \left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}$$

- **Remark:** showing that a kernel is positive definite is enough to get Φ and $\mathcal{H}$ with the reproducing property "for free"

Example

- **Example:** polynomial kernel of degree 2:

$$k(x,y) = (x^\top y)^2 \,.$$

- **Claim:**

$$k(x,y) = \langle xx^\top, yy^\top \rangle_F \,,$$

  thus k is positive definite
- **Question:** what is the RKHS?
- we know that $\mathcal{H}$ contains all the functions

$$f(x) = \sum_i a_i k(x_i, x) = \sum_i a_i \langle x_i x_i^\top, xx^\top \rangle = \langle \sum_i a_i x_i x_i^\top, xx^\top \rangle$$

Example, ctd.

- spectral theorem: any symmetric matrix can be decomposed as $\sum_i a_i x_i x_i^\top$
- candidate RKHS: set a quadratic functions

$$f_S(x) = \langle S, xx^\top \rangle = x^\top S x \,,$$

  with S symmetric matrix of size $d \times d$
- inner product on $\mathcal{H}$:

$$\langle f_S, f_{S'} \rangle = \langle S, S' \rangle_F \,.$$

- we can check that $\mathcal{H}$ is a Hilbert space (isomorphic to $\mathcal{S}^{d \times d}$)
- finally, we check the reproducing property

5.3. More examples

Elementary properties

Proposition: Let $k_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a (potentially infinite) family of p.d. kernels. Then

- for any $\lambda_1, \ldots, \lambda_p \geq 0$, the sum $\sum_{i=1}^{p} \lambda_i k_i$ is positive definite
- for any $a_1, \ldots, a_p \in \mathbb{N}$, the product $k_1^{a_1} \cdots k_p^{a_p}$ is positive definite
- if it exists, the limit $k = \lim_{p \to +\infty} k_p$ is positive definite

Moreover, let $\mathcal{X}_i$ be a sequence of sets and k_i positive kernels on each $\mathcal{X}_i$. Then

$$k((x_1, \ldots, x_p), (y_1, \ldots, y_p)) := \prod_{i=1}^{p} k_i(x_i, y_i)$$

and

$$k((x_1, \ldots, x_p), (y_1, \ldots, y_p)) := \sum_{i=1}^{p} k_i(x_i, y_i)$$

are positive definite kernels.

Taking the exponential

Theorem: if k is a positive definite kernel, then e^k as well.

- *Proof:* we write
$$\mathrm{e}^{k(x,y)} = \lim_{n \to +\infty} \sum_{p=0}^{n} \frac{k(x,y)^p}{p!} \, ,$$
then reason step by step.
- by the product property, $k(x,y)^p$ is a kernel for any $p \geq 0$
- as a positive linear combination of kernels, $\sum_{p=0}^{n} \frac{k(x,y)^p}{p!}$ is a kernel for all $n \geq 1$
- finally, e^k is a kernel as a limit of kernels. $\square$

5.4. The kernel trick and applications

The kernel trick

- input data $x_1, \ldots, x_n \in \mathcal{X}$
- $k : \mathcal{X} \times \mathcal{X}$ kernel with associated RKHS $\mathcal{H}$
- we call $\Phi : \mathcal{X} \to \mathcal{H}$ the feature map
- **Idea:** imagine that our algorithm only depends on scalar products $x_i^\top x_j$
- then we can map the x_i to $\mathcal{H}$ and replace the inner products by kernel evaluations, since

$$\langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j)\,.$$

- simple "trick" with many, many applications

Example

- **Example:** computing distances
- suppose that our algo relies on distance computation
- that is, $\|x - y\|^2$
- we can write

$$\begin{aligned}\|\Phi(x) - \Phi(y)\|^2 &= \langle \Phi(x) - \Phi(y), \Phi(x) - \Phi(y) \rangle \\ &= \langle \Phi(x), \Phi(x) \rangle - 2\langle \Phi(x), \Phi(y) \rangle + \langle \Phi(y), \Phi(y) \rangle \\ \|\Phi(x) - \Phi(y)\|^2 &= k(x,x) - 2k(x,y) + k(y,y)\,.\end{aligned}$$

- in other words,

$$d_{\mathcal{H}}(x,y) = \sqrt{k(x,x) - 2k(x,y) + k(y,y)}\,.$$

- as promised, **we do not need to know Φ!**

5.5. The representer theorem

Motivation

- let us imagine that we take $\mathcal{H}$ as hypothesis class
- starting from regularized ERM, our optimization problem will look like

$$\arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|f\|^2 \right\} . \tag{$\star$}$$

- we penalize by the norm because it is an indicator of the *smoothness* of f
- **Why?** Cauchy-Schwarz + exercise:

$$|f(x) - f(y)| = |\langle f, k_x - k_y \rangle| \leq \|f\| \cdot \|k_x - k_y\| = \|f\| \cdot d_{\mathcal{H}}(x, y) .$$

- Eq. ($\star$) is a complicate problem, potentially *infinite-dimensional*
- **Question:** how to solve it in practice?

The representer theorem

Theorem: let $\mathcal{H}$ be the RKHS associated to k defined on $\mathcal{X}$. Let $S = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ be a finite set of points. Let $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ be a function, increasing in the last variable. Then any solution to the minimization problem

$$\arg\min_{f \in \mathcal{H}} \Psi(f(x_1), \ldots, f(x_n), \|f\|)$$

admits a representation of the form

$$\forall x \in \mathcal{X}, \qquad f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)\,.$$

- **Main consequence:** Eq. $(\star)$ is actually a finite-dimensional problem (!)

Practical use

- recall that we defined $K := (k(x_i, x_j))_{i,j=1}^n$
- before turning to concrete examples, we notice that we can simply express the key quantities
- for instance, for any $1 \leq j \leq n$,

$$f(x_j) = \sum_{i=1}^n \alpha_i k(x_i, x_j) = (K\alpha)_j \,.$$

- in the same way,

$$\|f\|^2 = \left\| \sum_{i=1}^n \alpha_i k(x_i, \cdot) \right\|^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) = \alpha^\top K \alpha \,.$$

5.6. Kernel ridge regression

Kernel Ridge Regression[10] (KRR)

- regression setting: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$
- $\mathcal{Y} \subseteq \mathbb{R}$, but $\mathcal{X}$ could be anything
- we have a kernel k on $\mathcal{X}$
- same idea than with ridge regression:

$$\hat{f} \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \left\| f \right\|^2 \right\} .$$

- here effect of the regularization is to make $\hat{f}$ smoother

[10] Cristianini and Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, 2000

Solving KRR

- representer theorem $\Rightarrow$

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)\,,$$

  for some $\alpha \in \mathbb{R}^n$
- as per the previous remark, we know that

$$(\hat{f}(x_1), \ldots, \hat{f}(x_n))^\top = K\alpha\,,$$

  and

$$\|\hat{f}\|^2 = \alpha^\top K\alpha\,.$$

- thus KRR can be re-written as

$$\hat{\alpha} \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n}(K\alpha - y)^\top (K\alpha - y) + \lambda \alpha^\top K\alpha \right\}\,.$$

Solving KRR, ctd.

- convex, smooth objective $\Rightarrow$ set the gradient to zero
- $\hat{\alpha}$ has to be solution of

$$0 = \frac{-2}{n} K(y - K\alpha) + 2\lambda K\alpha = \frac{2}{n} K\left[(K + n\lambda I_n)\alpha - y\right]$$

- since $\lambda > 0$, $K + n\lambda I_n$ is invertible
- a solution is given by

$$\hat{\alpha} = (K + n\lambda I_n)^{-1} y\,.$$

- **Remark:** if $k =$ linear kernel, $K = XX^\top$
- solution we found solving "regular" ridge regression is

$$\hat{\beta} = (X^\top X + n\lambda\, \mathrm{I}_d)^{-1} X^\top y\,.$$

Solving KRR, ctd.

- actually leads to the same solution
- can compare the predictions:
- on one side,
$$K\hat{\alpha} = K(K + n\lambda\,\mathrm{I}_n)^{-1} = XX^\top(XX^\top + n\lambda\,\mathrm{I}_n)^{-1}y\,.$$
- on the other side,
$$X\hat{\beta} = X(X^\top X + n\lambda\,\mathrm{I}_d)^{-1}X^\top y$$
- *Proof:* Woodbury identity:
$$(\mathrm{I} + AA^\top)^{-1} = \mathrm{I} - A(\mathrm{I} + A^\top A)^{-1}A^\top\,.$$
- (Woodbury actually has a more general statement)

Uniqueness

- **Reminder:**
$$\hat{\alpha} = (K + n\lambda\,\mathrm{I}_n)^{-1} y\,.$$
- **Remark:** not the only solution if K is singular
- **Why?** $K + \lambda n\,\mathrm{I}$ and $(K + \lambda n\,\mathrm{I})^{-1}$ both leave $\ker K$ stable, can add ε such that $K\varepsilon = 0$
- but correspond to same element in the RKHS!
- **Why:** compute (squared) norm of the difference:
$$\left\| \sum_i \alpha_i k(\cdot, x_i) - \sum_i (\alpha_i + \varepsilon_i) k(\cdot, x_i) \right\|^2 = (\alpha - \varepsilon)^\top K (\alpha - \varepsilon) = 0\,.$$

5.7. Kernel logistic regression

Kernel Logistic Regression[11] (KLR)

- classification setting: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$
- $\mathcal{Y} = \{0, 1\}$, but $\mathcal{X}$ could be anything
- we have a kernel k on $\mathcal{X}$
- kernelized version of logistic regression:

$$\hat{f} \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i f(x_i)}\right) + \lambda \left\| f \right\|^2 \right\} .$$

- same regularization effect

[11] Green, Yandell, *Semi-parametric generalized linear models*, Generalized linear models, 1985

Solving KLR

- no explicit solution, but convex and smooth
- again, we can use the representer theorem:

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

  for some $\alpha \in \mathbb{R}^n$
- again, $(\hat{f}(x_1), \ldots, \hat{f}(x_n))^\top = K\alpha$ and $\|\hat{f}\|^2 = \alpha^\top K \alpha$
- we can rewrite KLR as

$$\hat{\alpha} \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \frac{1}{n} \left\{ \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i (K\alpha)_i}\right) + \lambda \alpha^\top K \alpha \right\}.$$

- this can be solved (approximately) by gradient descent

Illustration