Statistical Learning

Instructor: **Gérard Biau**

Main references

1. Bach, F. (2024). Learning from first principles, The MIT press, Cambridge, Massachusetts.

2. Biau, G. and Devroye, L. (2015). Lectures on the nearest neighbor method, Springer, Cham.

3. Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: A survey of some recent advances, *ESAIM: Probability and Statistics*, **Vol. 9**, pp. 323–375.

4. Devroye, L., Györfi, L., and Lugosi, G. (1996). A probabilistic theory of pattern recognition, Springer, New York.

5. Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). A distribution-free theory of nonparametric regression, Springer, New York.

6. Linder, T. (2002). Learning-theoretic methods in vector quantization, in *Principles of Nonparametric Learning*, ed. Györfi, L., pp. 163–210, Springer, Vienna.

Basics of Supervised Learning

监督学习基础

Context and notation

- **Random pair**: $(X, Y) \in \mathscr{X} \times \mathscr{Y}$. Often, but not always, $\mathscr{X} \subseteq \mathbb{R}^d$.
- X is the **input** and Y is the **output** (response, label, class, etc.).
- Examples:
  二元分类 – **Binary classification**: $\mathscr{Y} = \{0, 1\}$ or $\mathscr{Y} = \{-1, 1\}$.
  多元分类 – **Multi-category classification**: $\mathscr{Y} = \{1, \dots, k\}$.
  回归 – **Regression**: $\mathscr{Y} = \mathbb{R}$.
- Notation:
  - $p(dx, dy)$ the distribution of (X, Y).
  - $\mu(dx)$ the distribution of X. $\mu(A) = \mathbb{P}(X \in A)$
  - $r(x) = \mathbb{E}(Y|X = x)$ the regression function. 回归函数
  - In binary classification, $\eta(x) = \mathbb{P}(Y = 1|X = x)$ $(= r(x)$ when $\mathscr{Y} = \{0, 1\})$.

⚠ Are X and Y independent? **Not necessarily**.

⚠ Do we have $Y = \varphi(X)$? **Not necessarily**.

- **Objective**: find a predictor $f : \mathscr{X} \to \mathscr{Y}$ such that $f(X) \approx Y$.
  目的 找到预测器
- In the classification setting, f is called a **classifier**.
  分类器

Loss function and risk 损失函数和风险

损失函数 • **Loss function**: $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}_+$. **Interpretation**: $\ell(y, z)$ is the loss incurred when predicting z while the true output is y.

- Examples:
  - 二元分类 **Binary classification**: $\mathscr{Y} = \{0,1\}$ and $\ell(y,z) = \mathbf{1}_{[y\neq z]}$ (0-1 loss).
  - 多元分类 **Multi-category classification**: $\mathscr{Y} = \{1,\dots,k\}$ and $\ell(y,z) = \mathbf{1}_{[z\neq y]}$.
  - 回归 **Regression**: $\mathscr{Y} = \mathbb{R}$ and $\ell(y,z) = (y-z)^2$ (squared loss). Absolute loss: $\ell(y,z) = |y-z|$.
- The **risk** (generalization performance, testing error) of a predictor $f : \mathscr{X} \to \mathscr{Y}$ is

风险 $$\mathscr{R}(f) = \mathbb{E}\ell(\underbrace{Y}_{\text{真实值}}, \underbrace{f(X)}_{\text{预测值}}) = \int_{\mathscr{X}\times\mathscr{Y}} \ell(y, f(x))p(dx,dy).$$

- Examples:
  - **Binary classification**: $\mathscr{Y} = \{0,1\}$, $\ell(y,z) = \mathbf{1}_{[y\neq z]}$, and $\mathscr{R}(f) = \mathbb{E}\mathbf{1}_{[Y\neq f(X)]} = \mathbb{P}(f(X)\neq Y)$.
  - **Multi-category classification**: $\mathscr{Y} = \{1,\dots,k\}$, $\ell(y,z) = \mathbf{1}_{[y\neq z]}$, and $\mathscr{R}(f) = \mathbb{P}(f(X)\neq Y)$.
  - **Regression**: $\mathscr{Y} = \mathbb{R}$, $\ell(y,z) = (y-z)^2$, and $\mathscr{R}(f) = \mathbb{E}(Y - f(X))^2$.

Bayes risk and Bayes predictor Bayes风险 和 Bayes预测器

- Bayes风险 **Bayes risk**: $\mathscr{R}^* = \inf_{f:\mathscr{X}\to\mathscr{Y}} \mathscr{R}(f)$. 不一定能为 0
- Bayes预测器 **Bayes predictor**: any $f^* : \mathscr{X} \to \mathscr{Y}$ such that 满足 $\mathscr{R}(f^*) = \mathscr{R}^*$ (non necessarily 不一定唯一 unique).
- The **excess risk** 超额风险 of $f : \mathscr{X} \to \mathscr{Y}$ is $\mathscr{R}(f) - \mathscr{R}^*$ ($\geqslant 0$).
- Examples:
  - **Binary classification**: $\mathscr{Y} = \{0,1\}$, $\ell(y,z) = \mathbf{1}_{[y\neq z]}$. The **Bayes classifier** is

$$f^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x) \\ 0 & \text{otherwise,} \end{cases}$$

i.e., $f^*(x) = \mathbf{1}_{[\eta(x)>1/2]}$. Moreover,

$$\mathscr{R}^* = \mathbb{E}\min(\eta(X), 1-\eta(X)) = \frac{1}{2} - \frac{1}{2}\mathbb{E}|2\eta(X)-1|.$$

Proof. Let $f : \mathscr{X} \to \{0,1\}$ be an arbitrary Borel measurable function. Then
任意的

$$\mathbb{P}(f(X) \neq Y) = 1 - \mathbb{P}(f(X) = Y).$$

Thus,

$$\begin{aligned}\mathbb{P}(f(X) \neq Y) - \mathbb{P}(f^*(X) \neq Y) &= \mathbb{P}(f^*(X) = Y) - \mathbb{P}(f(X) = Y)\\ &= \mathbb{E}(\mathbb{P}(f^*(X) = Y|X) - \mathbb{P}(f(X) = Y|X))\\ &\geqslant 0.\end{aligned}$$

To prove this inequality, just note that

$$\begin{aligned}\mathbb{P}(f(X) = Y|X) &= \mathbb{P}(f(X) = 1, Y = 1|X) + \mathbb{P}(f(X) = 0, Y = 0|X)\\ &= \mathbf{1}_{[f(X)=1]}\mathbb{P}(Y = 1|X) + \mathbf{1}_{[f(X)=0]}\mathbb{P}(Y = 0|X).\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{P}(f^*(X) = Y|X) &= \mathbf{1}_{[f^*(X)=1]}\mathbb{P}(Y = 1|X) + \mathbf{1}_{[f^*(X)=0]}\mathbb{P}(Y = 0|X)\\ &= \max(\mathbb{P}(Y = 0|X), \mathbb{P}(Y = 1|X)),\end{aligned}$$

by definition of f^*. ■

with probability

- **Remark**: $\mathscr{R}^* = 0 \Leftrightarrow Y = \varphi(X)$ wp 1.
  Y可以由X表达
- **Regression**: $\mathscr{Y} = \mathbb{R}$, $\ell(y,z) = (y-z)^2$, $\mathbb{E}Y^2 < \infty$. The Bayes predictor is $f^*(x) = r(x)$, it is μ-almost surely unique, and $\mathscr{R}^* = \mathbb{E}(Y - r(X))^2$.
  回归函数
  难以计算,不知道分布

Learning from data 从数据中学习

- In practice, the distribution of (X,Y) is **unknown**.
- **Sample**: $\mathscr{D}_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, i.i.d. copies of (X,Y).
- The pair (X,Y) and $\mathscr{D}_n$ are **independent**.
- A **predictor**: $f_n(x) = f_n(x\,;\mathscr{D}_n) : \mathscr{X} \to \mathscr{Y}$. ⚠ It is random.
  预测器
- The **risk** of f_n is
  风险

和(X,Y)独立,此时不求期望
一开始视作随机变量

$$\mathscr{R}(f_n) = \mathbb{E}(\ell(Y, f_n(X))|\mathscr{D}_n) = \int_{\mathscr{X}\times\mathscr{Y}} \ell(y, f_n(x))p(dx,dy).$$

组成 f_n

⚠ One has $\mathbb{E}\mathscr{R}(f_n) = \mathbb{E}\ell(Y, f_n(X))$.

对 D_n 也求期望

- **Objective** 目的: construct f_n such that $\mathscr{R}(f_n) \approx \mathscr{R}^*$.
- **Consistency** 一致性: for a **certain** distribution 某些分布 of (X, Y), $\mathbb{E}\mathscr{R}(f_n) \to \mathscr{R}^*$ as $n \to \infty$.
- **Universal consistency** 全局一致性: for **any** distribution 任意分布 of (X, Y), $\mathbb{E}\mathscr{R}(f_n) \to \mathscr{R}^*$ as $n \to \infty$.

  ↳ 是 L_1-收敛
  ↳ 而且 $\mathcal{R}(f_n)$ 有界，所以也是 $\mathbb{P}$-收敛
- **PAC bounds**: for a given $\delta \in (0, 1)$ and $\varepsilon > 0$,

$$\mathbb{P}(\mathscr{R}(f_n) - \mathscr{R}^* \leqslant \varepsilon) \geqslant 1 - \delta.$$

- Two main approaches: empirical risk minimization 经验风险最小化 and local averaging 局部均值.

Concentration inequalities 集中不等式

Theorem 2.1 — Hoeffding's inequality *Hoeffding不等式*

Let $X_1, \dots, X_n$ be **independent** real-valued random variables. Assume that each X_i takes its values in $[a_i, b_i]$ $(a_i < b_i)$ wp 1. Then, for all $t > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i \geqslant t\Big) \leqslant e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$$

and 取 $-X_i$

$$\mathbb{P}\Big(\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i \leqslant -t\Big) \leqslant e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

In particular,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i\Big| \geqslant t\Big) \leqslant 2e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

The proof is a consequence of the following lemma.

Lemma 2.1. *Let X be a real-valued random variable with $\mathbb{E}X = 0$ and $X \in [a, b]$ $(a < b)$ wp 1. Then, for all $s \geqslant 0$,*

$$\mathbb{E}e^{sX} \leqslant e^{s^2(b-a)^2/8}.$$

Proof. Note that, by the convexity of the exponential function,

$$e^{sx} \leqslant \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}, \quad a \leqslant x \leqslant b.$$

Exploiting $\mathbb{E}X = 0$, and introducing the notation $p = -\frac{a}{b-a}$, we obtain

$$\begin{aligned}
\mathbb{E}e^{sX} &\leqslant \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\
&= \left(1 - p + pe^{s(b-a)}\right)e^{-ps(b-a)} \\
&\stackrel{\text{def}}{=} e^{\phi(u)},
\end{aligned}$$

where $u = s(b-a)$ and $\phi(t) = -pt + \log(1-p+pe^t)$. The derivative of ϕ is

$$\phi'(t) = -p + \frac{p}{p+(1-p)e^{-t}},$$

and therefore $\phi(0) = \phi'(0) = 0$. Moreover,

$$\phi''(t) = \frac{p(1-p)e^{-t}}{(p+(1-p)e^{-t})^2} \leqslant 1/4.$$

Thus, by Taylor's theorem, for some $\theta \in [0, u]$,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leqslant \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

■

Proof of Hoeffding's inequality:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i \geqslant t\right) &\leqslant e^{-st}\mathbb{E}e^{s\sum_{i=1}^n (X_i - \mathbb{E}X_i)} \\
&\quad \text{(by Markov's inequality)} \\
&\leqslant e^{-st}\prod_{i=1}^n e^{s^2(b_i-a_i)^2/8} \\
&\quad \text{(by independence and Lemma 2.1)} \\
&= e^{-st}e^{s^2\sum_{i=1}^n (b_i-a_i)^2/8} = e^{-2t^2/\sum_{i=1}^n (b_i-a_i)^2},
\end{aligned}$$

by choosing $s = 4t/\sum_{i=1}^n (b_i-a_i)^2$. The other two inequalities are immediate consequences.
■

Theorem 2.2 — Bounded difference inequality 有界差分不等式

Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $\mathscr{X}$ wp 1. Assume that $g : \mathscr{X}^n \to \mathbb{R}$ is Borel measurable and satisfies

$$\sup_{\substack{(x_1,\ldots,x_n)\in\mathscr{X}^n \\ x_i' \in \mathscr{X}}} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leqslant c_i, \; 1 \leqslant i \leqslant n,$$

for some positive constants $c_1, \ldots, c_n$ (**bounded difference assumption**). Then, for all $t > 0$,
有界差分假设

$$\mathbb{P}\big(g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \geqslant t\big) \leqslant e^{-2t^2/\sum_{i=1}^n c_i^2}$$

and

$$\mathbb{P}\big(g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \leqslant -t\big) \leqslant e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

In particular,

$$\mathbb{P}\big(|g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n)| \geqslant t\big) \leqslant 2e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

Lemma 2.2. *Let $\alpha > 0$, and let $X_1, \ldots, X_n$ be real-valued random variables such that, for all $s > 0$ and all $1 \leqslant i \leqslant n$, $\mathbb{E}e^{sX_i} \leqslant e^{s^2\alpha^2/2}$. Then, if $n \geqslant 2$,*

$$\mathbb{E} \max_{1 \leqslant i \leqslant n} X_i \leqslant \alpha\sqrt{2 \log n}.$$

If, in addition, $\mathbb{E}e^{-sX_i} \leqslant e^{s^2\alpha^2/2}$ for all $s > 0$ and $1 \leqslant i \leqslant n$, then, for any $n \geqslant 1$,

$$\mathbb{E} \max_{1 \leqslant i \leqslant n} |X_i| \leqslant \alpha\sqrt{2 \log(2n)}.$$

Proof. By Jensen's inequality, for all $s > 0$,

$$\begin{aligned} e^{s\mathbb{E}\max_{1\leqslant i \leqslant n} X_i} &\leqslant \mathbb{E}e^{s\max_{1\leqslant i\leqslant n} X_i} = \mathbb{E} \max_{1 \leqslant i \leqslant n} e^{sX_i} \\ &\leqslant \sum_{i=1}^n \mathbb{E}e^{sX_i} \leqslant ne^{s^2\alpha^2/2}. \end{aligned}$$

Thus,

$$\mathbb{E} \max_{1 \leqslant i \leqslant n} X_i \leqslant \frac{\log n}{s} + \frac{s\alpha^2}{2},$$

and taking $s = \sqrt{2 \log n}/\alpha$ yields the first inequality. Finally, note that $\max_{1\leqslant i\leqslant n} |X_i| = \max(X_1, -X_1, \ldots, X_n, -X_n)$ and apply the first inequality to prove the second one. ∎

LINEAR LEAST-SQUARES REGRESSION

线性最小二乘回归

Context and notation

- Regression setting:
  - $\mathscr{Y} = \mathbb{R}$ and $\ell(y, z) = (y - z)^2$.
  - $\mathbb{E}Y^2 < \infty$, $\mathbb{E}f(X)^2 < \infty$, $\mathscr{R}(f) = \mathbb{E}(Y - f(X))^2$, and $f^*(x) = \mathbb{E}(Y|X = x)$.
- **Least-squares regression**: 最小二乘回归
  - Choose a parametric family of predictors $\{f_\theta : \mathscr{X} \to \mathbb{R}, \theta \in \Theta\}$, with $\mathbb{E}f_\theta(X)^2 < \infty$.
  - Minimize the **empirical risk** 最小化经验风险

$$\mathscr{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\theta(X_i))^2.$$

  - **Estimator** 估计量: $\theta_n \in \arg\min_{\theta\in\Theta} \mathscr{R}_n(\theta)$.

⚠ In most cases $f^* \notin \{f_\theta, \theta \in \Theta\}$. f_{θ_n}不一定是f^*，因为f_θ的范围只是参数预测器

- **Linear** least-squares regression: 线性最小二乘回归
  - $\Theta = \mathbb{R}^d$ and a known **feature vector** 特征向量 $\varphi(x) \in \mathbb{R}^d$ such that $f_\theta(x) = \varphi(x)^\top\theta$. x是输入的数据，经过$\varphi(\cdot)$提取特征变成向量$\varphi(x)$

  ⚠ $\mathbb{E}\|\varphi(X)\|_2^2 < \infty$ and linearity is in θ ($\|\alpha\|_2^2 = \sum_{j=1}^{d}\alpha_j^2$ is the squared ℓ^2-norm of α).
  - Empirical risk:

经验风险
$$\mathscr{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \varphi(X_i)^\top\theta)^2.$$

  - When $\mathscr{X} \subseteq \mathbb{R}^d$, extensions are possible. Examples: $\varphi(x) = (x^\top, 1)^\top \in \mathbb{R}^{d+1}$ and $\varphi(x)$ = collection of monomials.

$$\Phi = \begin{bmatrix} \varphi^{(1)}(X_1) & \varphi^{(2)}(X_1) & \cdots & \varphi^{(d)}(X_1) \\ \varphi^{(1)}(X_2) & \varphi^{(2)}(X_2) & \cdots & \varphi^{(d)}(X_2) \\ \vdots & \vdots & & \vdots \\ \varphi^{(1)}(X_n) & \varphi^{(2)}(X_n) & \cdots & \varphi^{(d)}(X_n) \end{bmatrix}$$

- **Matrix notation**: 矩阵记号
  - $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ the **response vector**. 响应向量
  - $\Phi = (\varphi(X_1) \mid \cdots \mid \varphi(X_n))^\top \in \mathbb{R}^{n \times d}$ the **design matrix**. 设计矩阵
  - Empirical risk:

$$\mathscr{R}_n(\theta) = \frac{1}{n}\|\mathbf{Y} - \Phi\theta\|_2^2.$$

  - Least-squares estimator: $\theta_n \in \arg\min_{\theta \in \Theta} \mathscr{R}_n(\theta)$.

Ordinary least-squares estimator 普通最小二乘估计量

- **Assumption**: the matrix $\Phi \in \mathbb{R}^{n \times d}$ has full rank d (and thus $d \leqslant n$). 假设
- **Remark**: this is equivalent to $\Phi^\top\Phi \in \mathbb{R}^{d \times d}$ invertible. The matrix $\Sigma_n = \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d \times d}$ is the non-centered empirical covariance matrix. 非中心经验协方差矩阵
- **Definition**: θ_n is called the **ordinary least-squares** (OLS) estimator. 定义

Proposition 3.1. *The OLS estimator exists and is unique. It is given by*
OLS估计量 的 存在唯一性

$$\theta_n = (\Phi^\top\Phi)^{-1}\Phi^\top\mathbf{Y} = \frac{1}{n}\Sigma_n^{-1}\Phi^\top\mathbf{Y}.$$

Proof. Since the function $\mathscr{R}_n(\cdot)$ is coercive (i.e., going to infinity at infinity) and continuous, it admits at least a minimizer. Moreover, it is differentiable, so a minimizer θ_n must satisfy $\nabla\mathscr{R}_n(\theta_n) = 0$. For any $\theta \in \mathbb{R}^d$, we have

$$\mathscr{R}_n(\theta) = \frac{1}{n}\Big(\|\mathbf{Y}\|_2^2 - 2\theta^\top\Phi^\top\mathbf{Y} + \theta^\top\Phi^\top\Phi\theta\Big) \quad \text{and} \quad \nabla\mathscr{R}_n(\theta) = \frac{2}{n}\Big(\Phi^\top\Phi\theta - \Phi^\top\mathbf{Y}\Big).$$

The condition $\nabla\mathscr{R}_n(\theta_n) = 0$ leads to $\Phi^\top\Phi\theta_n = \Phi^\top\mathbf{Y}$, and therefore $\theta_n = (\Phi^\top\Phi)^{-1}\Phi^\top\mathbf{Y}$. ∎

Proposition 3.2. *The vector of predictions* $\Phi\theta_n = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top\mathbf{Y}$ *is the orthogonal projection of* $\mathbf{Y} \in \mathbb{R}^n$ *onto* $\mathrm{im}(\Phi) \subseteq \mathbb{R}^n$*, the column space of* Φ*.* 预测向量 正交投影

Proof. The operator $\Pi = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top \in \mathbb{R}^{n \times n}$ is the orthogonal projection on $\mathrm{im}(\Phi)$. To see this, observe that for any $a \in \mathbb{R}^d$, $\Pi\Phi a = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top\Phi a = \Phi a$. Therefore, $\Pi u = u$ for all $u \in \mathrm{im}(\Phi)$. Moreover, since $\mathrm{im}(\Phi)^\perp = \mathrm{null}(\Phi^\top)$, $\Phi^\top(u') = 0$ for all $u' \in \mathrm{im}(\Phi)^\perp$. Thus, $\Pi u' = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top u' = 0$. These properties characterize the orthogonal projection onto $\mathrm{im}(\Phi)$. ∎

- Inverting $\Phi^\top\Phi$ may be unstable + important computational cost for large d → **numerical resolution** by QR factorization or gradient descent is preferred. QR分解 梯度下降

Statistical analysis: Fixed design 统计分析：固定设计

- **Context and assumptions**:
  - The input data $x_1, \ldots, x_n$ are **deterministic** 确定性的 (and so is the matrix $\Phi \in \mathbb{R}^{n \times d}$).
  - The matrix $\Sigma_n = \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d \times d}$ is **invertible** 可逆的.
  - There exists $\theta^* \in \mathbb{R}^d$ such that

    $$Y_i = \varphi(x_i)^\top \theta^* + \varepsilon_i, \quad 1 \leqslant i \leqslant n,$$

    where $\varepsilon_1, \ldots, \varepsilon_n$ are independent real-valued random variables, with $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}\varepsilon_i^2 = \sigma^2$.
  - **Notation**: $\mathbf{Y} = \Phi\theta^* + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$.

- **Risk minimization**: 风险最小化
  - **Objective**: minimize 最小化

    $$\mathscr{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \varphi(x_i)^\top\theta)^2 = \frac{1}{n}\|\mathbf{Y} - \Phi\theta\|_2^2.$$

  - The risk of $\theta \in \mathbb{R}^d$ is

    $$\mathscr{R}(\theta) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \varphi(x_i)^\top\theta)^2\Big) = \mathbb{E}\Big(\frac{1}{n}\|\mathbf{Y} - \Phi\theta\|_2^2\Big),$$

    and the risk of the OLS estimator θ_n is

    组成θn (有随机性)

    $$\mathscr{R}(\theta_n) = \mathbb{E}\Big(\frac{1}{n}\|\mathbf{Y}' - \Phi\theta_n\|_2^2 \mid Y_1, \ldots, Y_n\Big),$$

    where $Y_1', \ldots, Y_n'$ are i.i.d., independent of, and distributed as, $Y_1, \ldots, Y_n$.

  ⚠ $\mathscr{R}(\theta_n)$ is random, function of $Y_1, \ldots, Y_n$.
  - Bayes risk: $\mathscr{R}^* = \inf_{\theta \in \mathbb{R}^d} \mathscr{R}(\theta)$.

Theorem 3.1 — Fixed design setting 固定设计

One has $\mathscr{R}^* = \mathscr{R}(\theta^*) = \sigma^2$ and, for all $\theta \in \mathbb{R}^d$, $\mathscr{R}(\theta) - \mathscr{R}^* = \|\theta - \theta^*\|_{\Sigma_n}^2$, where $\|\theta\|_{\Sigma_n}^2 = \theta^\top \Sigma_n \theta$. Moreover, the **OLS estimator θ_n** satisfies the following properties:

1. $\mathbb{E}\theta_n = \theta^*$ and $\operatorname{var}(\theta_n) = \mathbb{E}(\theta_n - \theta^*)(\theta_n - \theta^*)^\top = \frac{\sigma^2}{n}\Sigma_n^{-1}$.
2. $\mathbb{E}\mathscr{R}(\theta_n) - \mathscr{R}^* = \frac{\sigma^2 d}{n}$.

Proof. Recall that $\mathbf{Y} = \Phi\theta^* + \boldsymbol{\varepsilon}$, with $\mathbb{E}\boldsymbol{\varepsilon} = 0$ and $\mathbb{E}\|\boldsymbol{\varepsilon}\|_2^2 = n\sigma^2$. Thus, for all $\theta \in \mathbb{R}^d$,

$$
\begin{aligned}
\mathscr{R}(\theta) &= \mathbb{E}\Big(\frac{1}{n}\|\mathbf{Y} - \Phi\theta\|_2^2\Big) = \mathbb{E}\Big(\frac{1}{n}\|\Phi\theta^* + \boldsymbol{\varepsilon} - \Phi\theta\|_2^2\Big) \\
&= \frac{1}{n}\mathbb{E}\Big(\|\Phi(\theta^* - \theta)\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 + 2(\Phi(\theta^* - \theta))^\top \boldsymbol{\varepsilon}\Big) \\
&= \sigma^2 + \frac{1}{n}(\theta - \theta^*)^\top \Phi^\top \Phi(\theta - \theta^*) \\
&\quad \text{(since } \mathbb{E}\boldsymbol{\varepsilon} = 0) \\
&= \sigma^2 + (\theta - \theta^*)^\top \Sigma_n (\theta - \theta^*).
\end{aligned}
$$

This shows that $\mathscr{R}^* = \mathscr{R}(\theta^*) = \sigma^2$. Moreover, $\mathscr{R}(\theta) - \mathscr{R}^* = \|\theta - \theta^*\|_{\Sigma_n}^2$.

Next, observing that $\mathbb{E}\mathbf{Y} = \Phi\theta^*$, we have $\mathbb{E}\theta_n = (\Phi^\top\Phi)^{-1}\Phi^\top\Phi\theta^* = \theta^*$. In addition, $\theta_n - \theta^* = (\Phi^\top\Phi)^{-1}\Phi^\top(\Phi\theta^* + \boldsymbol{\varepsilon}) - \theta^* = (\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{\varepsilon}$. Thus, using $\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top = \sigma^2 I_n$, we obtain

$$
\operatorname{var}(\theta_n) = \mathbb{E}\Big((\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\Phi(\Phi^\top\Phi)^{-1}\Big) = \sigma^2(\Phi^\top\Phi)^{-1}(\Phi^\top\Phi)(\Phi^\top\Phi)^{-1} = \sigma^2(\Phi^\top\Phi)^{-1},
$$

i.e., $\operatorname{var}(\theta_n) = \frac{\sigma^2}{n}\Sigma_n^{-1}$.

To prove the last assertion, just note that

$$
\begin{aligned}
\mathbb{E}\mathscr{R}(\theta_n) - \mathscr{R}^* &= \mathbb{E}\|\theta_n - \theta^*\|_{\Sigma_n}^2 = \mathbb{E}(\theta_n - \theta^*)^\top \Sigma_n (\theta_n - \theta^*) \\
&= \mathbb{E}\operatorname{tr}((\theta_n - \theta^*)^\top \Sigma_n (\theta_n - \theta^*)) = \mathbb{E}\operatorname{tr}((\theta_n - \theta^*)(\theta_n - \theta^*)^\top \Sigma_n) \\
&\quad \text{(since } \operatorname{tr}(AB) = \operatorname{tr}(BA)) \\
&= \operatorname{tr}(\operatorname{var}(\theta_n)\Sigma_n) = \operatorname{tr}\Big(\frac{\sigma^2}{n}\Sigma_n^{-1}\Sigma_n\Big) = \frac{\sigma^2}{n}\operatorname{tr}(I_d) = \frac{\sigma^2 d}{n}.
\end{aligned}
$$

■

- **Conclusion**: in the fixed design setting, the OLS has excess risk $\sigma^2 d/n$.

⚠ d/n needs to be small → regularization (ridge and Lasso regression).

Statistical analysis: Random design 统计分析：随机设计

- **Context and assumptions**:
  - **Sample**: $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, i.i.d. copies of (X, Y).
  - The matrix $\Sigma_n = \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d\times d}$ is **random**. The noncentered covariance matrix $\Sigma = \mathbb{E}\varphi(X)\varphi(X)^\top \in \mathbb{R}^{d\times d}$ is **deterministic**.
  - There exists $\theta^* \in \mathbb{R}^d$ such that

    $$Y = \varphi(X)^\top\theta^* + \varepsilon,$$

    where $\varepsilon \perp\!\!\!\perp X$, $\mathbb{E}\varepsilon = 0$, and $\mathbb{E}\varepsilon^2 = \sigma^2$.
  - **Notation**: $\mathbf{Y} = \Phi\theta^* + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$.

- **Risk minimization**: 风险最小化
  - **Objective**: minimize

    $$\mathscr{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \varphi(X_i)^\top\theta)^2 = \frac{1}{n}\|\mathbf{Y} - \Phi\theta\|_2^2.$$

  - The risk of $\theta \in \mathbb{R}^d$ is

    $$\mathscr{R}(\theta) = \mathbb{E}(Y - \varphi(X)^\top\theta)^2,$$

    and the risk of the OLS estimator θ_n is

    $$\mathscr{R}(\theta_n) = \mathbb{E}((Y - \varphi(X)^\top\theta_n)^2 \mid \mathscr{D}_n).$$

  ⚠ The Bayes predictor $f^*(x) = \mathbb{E}(Y|X = x) = \varphi(x)^\top\theta^*$ belongs to the family $\{f_\theta(x) = \varphi(x)^\top\theta, \theta \in \mathbb{R}^d\}$.
  - Bayes risk: $\mathscr{R}^* = \inf_{\theta\in\mathbb{R}^d}\mathscr{R}(\theta)$.

Theorem 3.2 — Random design setting 随机设计

One has $\mathscr{R}^* = \mathscr{R}(\theta^*) = \sigma^2$ and, for all $\theta \in \mathbb{R}^d$, $\mathscr{R}(\theta) - \mathscr{R}^* = \|\theta - \theta^*\|_\Sigma^2$, where $\|\theta\|_\Sigma^2 = \theta^\top\Sigma\theta$. Moreover, assuming that Σ_n is invertible, the **OLS estimator θ_n** satisfies $\mathbb{E}\mathscr{R}(\theta_n) - \mathscr{R}^* = \frac{\sigma^2}{n}\mathbb{E}\,\mathrm{tr}(\Sigma\Sigma_n^{-1})$.

Proof. For all $\theta \in \mathbb{R}^d$, one has

$$\begin{aligned}
\mathscr{R}(\theta) &= \mathbb{E}(Y - \varphi(X)^\top \theta)^2 = \mathbb{E}(\varphi(X)^\top \theta^* + \varepsilon - \varphi(X)^\top \theta)^2 \\
&= \mathbb{E}\Big((\varphi(X)^\top(\theta^* - \theta))^2 + \varepsilon^2 + 2\varphi(X)^\top(\theta^* - \theta)\varepsilon\Big) \\
&= \sigma^2 + \mathbb{E}(\theta^* - \theta)^\top \varphi(X)\varphi(X)^\top(\theta^* - \theta) \\
&\quad (\text{since } \varepsilon \perp\!\!\!\perp X \text{ and } \mathbb{E}\varepsilon = 0) \\
&= \sigma^2 + (\theta^* - \theta)^\top \Sigma(\theta^* - \theta).
\end{aligned}$$

This shows that $\mathscr{R}^* = \mathscr{R}(\theta^*) = \sigma^2$. Moreover, $\mathscr{R}(\theta) - \mathscr{R}^* = \|\theta - \theta^*\|_\Sigma^2$.

To prove the last assertion, notice that $\theta_n = \frac{1}{n}\Sigma_n^{-1}\Phi^\top \mathbf{Y} = \frac{1}{n}\Sigma_n^{-1}\Phi^\top(\Phi\theta^* + \boldsymbol{\varepsilon}) = \theta^* + \frac{1}{n}\Sigma_n^{-1}\Phi^\top \boldsymbol{\varepsilon}$. Therefore,

$$\begin{aligned}
\mathbb{E}\mathscr{R}(\theta_n) - \mathscr{R}^* &= \mathbb{E}\bigg(\Big(\frac{1}{n}\Sigma_n^{-1}\Phi^\top \boldsymbol{\varepsilon}\Big)^\top \Sigma\Big(\frac{1}{n}\Sigma_n^{-1}\Phi^\top \boldsymbol{\varepsilon}\Big)\bigg) \\
&= \mathbb{E}\,\mathrm{tr}\bigg(\Sigma\Big(\frac{1}{n}\Sigma_n^{-1}\Phi^\top \boldsymbol{\varepsilon}\Big)\Big(\frac{1}{n}\Sigma_n^{-1}\Phi^\top \boldsymbol{\varepsilon}\Big)^\top\bigg) = \frac{1}{n^2}\mathbb{E}\,\mathrm{tr}(\Sigma\Sigma_n^{-1}\Phi^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \Phi\Sigma_n^{-1}) \\
&\quad (\text{since } \mathrm{tr}(AB) = \mathrm{tr}(BA)) \\
&= \frac{1}{n^2}\mathbb{E}\,\mathrm{tr}(\Sigma\Sigma_n^{-1}\Phi^\top \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\Phi\Sigma_n^{-1}) = \frac{\sigma^2}{n^2}\mathbb{E}\,\mathrm{tr}(\Sigma\Sigma_n^{-1}\Phi^\top\Phi\Sigma_n^{-1}) \\
&= \frac{\sigma^2}{n}\mathbb{E}\,\mathrm{tr}(\Sigma\Sigma_n^{-1}).
\end{aligned}$$

■

⚠ The matrix Σ_n need not be invertible.
不必须 可逆的

- If $\varphi(X) \sim \mathscr{N}(0, \Sigma)$, Σ invertible, and $n > d + 1$, then
  Σ 可逆

$$\mathbb{E}\mathscr{R}(\theta_n) - \mathscr{R}^* = \frac{\sigma^2 d}{n} \times \frac{1}{1 - (d+1)/n} \approx \frac{\sigma^2 d}{n}.$$

Empirical Risk Minimization

经验风险最小化

Context and notation

- **Sample**: $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, i.i.d. copies of (X, Y).
- A loss function $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}_+$.
- A family $\mathscr{F} = \{f : \mathscr{X} \to \mathscr{Y}\}$ of predictors. Often $\mathscr{F} = \{f_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$.
- **Empirical risk minimization**: choose $f_n \in \mathscr{F}$ such that
  经验风险最小化
  $$f_n \in \arg\min_{f \in \mathscr{F}} \mathscr{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$
- **Objective**: bound the excess risk
  目标 约束超额风险
  $$\mathscr{R}(f_n) - \mathscr{R}^* = \mathscr{R}(f_n) - \inf_{f:\mathscr{X} \to \mathscr{Y}} \mathscr{R}(f),$$
  where $\mathscr{R}(f_n) = \mathbb{E}(\ell(Y, f_n(X)) \mid \mathscr{D}_n)$.

Convexification of the risk 风险函数的凸化

- **Binary classification**: $\mathscr{Y} = \{-1, 1\}$, $\ell(y, z) = \mathbf{1}_{[z \neq y]}$ (0-1 loss).
  二元分类
- **Empirical risk minimization**: choose $f_n \in \mathscr{F}$ such that
  经验风险最小化
  $$f_n \in \arg\min_{f \in \mathscr{F}} \mathscr{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[f(X_i) \neq Y_i]}.$$
- **Problem**: computationally hard. **Idea**: use convex surrogates.
  计算困难 凸代理损失函数
- We consider ± 1-classifiers of the form
  score
  $$g(x) = \begin{cases} 1 & \text{if } f(x) > 0 \\ -1 & \text{otherwise,} \end{cases}$$
  where $f : \mathscr{X} \to \mathbb{R}$. The risk of g is $\mathscr{R}(g) = \mathbb{E}\mathbf{1}_{[g(X) \neq Y]}$.

$Y \in \{-1,1\}$

$\mathbb{P}(g(X) \neq Y)$

- **Key**: $\mathbb{P}(Yf(X) < 0) \leqslant \mathscr{R}(g) \leqslant \mathbb{P}(Yf(X) \leqslant 0)$.
- **Notation 1**: $\mathscr{R}(f)$ instead of $\mathscr{R}(g)$.
- **Notation 2**: $\Phi_{0\text{-}1}(u) = \mathbf{1}_{[u \leqslant 0]}$ (0-1 loss function).
  损失函数
- One has $\mathscr{R}(f) \approx \mathbb{E}[\Phi_{0\text{-}1}(Yf(X))]$ and $\mathscr{R}_n(f) \approx \frac{1}{n}\sum_{i=1}^{n} \Phi_{0\text{-}1}(Y_i f(X_i))$.

  对于分类问题

  希望损失函数在负数时很大，在正数时很小
- **Idea**: smooth $\Phi_{0\text{-}1}$ by a convex **loss function** $\Phi : \mathbb{R} \to \mathbb{R}_+$.
  光滑化 凸损失函数
- **Φ-risks**: $\mathscr{R}_\Phi(f) = \mathbb{E}\Phi(Yf(X))$ and $\mathscr{R}_{n,\Phi}(f) = \frac{1}{n}\sum_{i=1}^{n} \Phi(Y_i f(X_i))$.
- The product $Yf(X)$ is the **margin**. Large margin = good confidence.
  间隔

⚠ Note the shift of notation $g \rightsquigarrow f$.

- Examples (see Figure 4.1):
  - 平方损失 **Squared loss**: $\Phi(u) = (1-u)^2$. One has $\Phi(Yf(X)) = (1 - Yf(X))^2 = (Y - f(X))^2 \to$ least-squares regression.
  - 指数损失 **Exponential loss**: $\Phi(u) = e^{-u}$. One has $\Phi(Yf(X)) = e^{-Yf(X)}$.
  - 逻辑损失 **Logistic loss**: $\Phi(u) = \log_2(1 + e^{-u})$. One has

$$\Phi(Yf(X)) = \log_2(1 + e^{-Yf(X)}) = -\log_2(\sigma(Yf(X))),$$

    where $\sigma(v) = \frac{1}{1+e^{-v}}$ is the sigmoid function.
  - 合页损失 **Hinge loss**: $\Phi(u) = \max(1-u, 0) \to$ support vector machines.

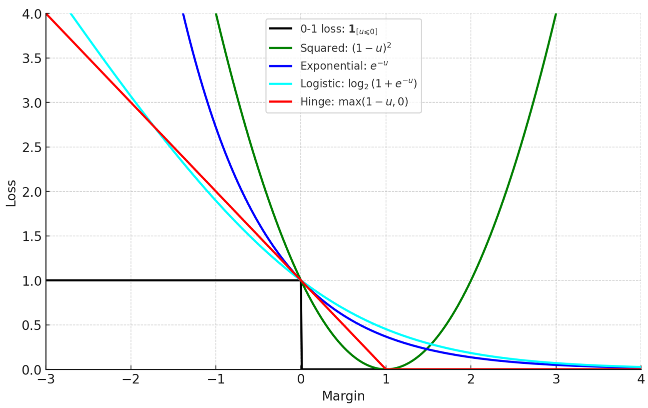


Figure 4.1: 0-1 loss and classical convex losses.

回归函数

$\eta(x) = \mathbb{E}[Y|X] = \mathbb{P}(Y=1|X)$

对于 0-1分类

- Bayes classifier:

  Beyes分类器

$$g^*(x) = \begin{cases} 1 & \text{if } 2\eta(x) - 1 > 0 \\ -1 & \text{otherwise,} \end{cases}$$

  i.e.,

$$g^*(x) = \begin{cases} 1 & \text{if } f^*(x) > 0 \\ -1 & \text{otherwise,} \end{cases}$$

  where $f^*(x) = 2\eta(x) - 1$.

- **Question**: what is $f^* \in \arg\min_{f:\mathscr{X}\to\mathbb{R}} \mathscr{R}_\Phi(f)$?

- **Definition**: The **conditional Φ-risk** of $f : \mathscr{X} \to \mathbb{R}$ is

  定义 条件Φ-风险

$$\mathbb{E}(\Phi(Yf(X)) \mid X = x) = \eta(x)\Phi(f(x)) + (1 - \eta(x))\Phi(-f(x)) \stackrel{\text{def}}{=} C_{\eta(x)}(f(x)),$$

  where $C_\eta(\alpha) = \eta\Phi(\alpha) + (1-\eta)\Phi(-\alpha)$, $\eta \in [0, 1]$.

- Φ is **(classification)-calibrated** if, for any $\eta \in [0, 1]$,

  (分类) - 校准

$$\text{(positive optimal prediction)} \quad \eta > 1/2 \Leftrightarrow \arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subseteq \mathbb{R}_+^* \tag{4.1}$$

  正最优预测

$$\text{(negative optimal prediction)} \quad \eta < 1/2 \Leftrightarrow \arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subseteq \mathbb{R}_-^*. \tag{4.2}$$

  负最优预测

Proposition 4.1. *Let $\Phi : \mathbb{R} \to \mathbb{R}_+$ be convex. Then Φ is classification-calibrated if and only if Φ is differentiable at 0 and $\Phi'(0) < 0$.*

分类校准的 在0点可微，并且Φ'(0)<0

Proof. Since Φ is convex, so is C_η for any $\eta \in [0, 1]$. Thus, we simply consider left and right derivatives at zero to obtain conditions about the location of minimizers, with only two possibilities:

$$\arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subseteq \mathbb{R}_+^* \Leftrightarrow C'_\eta(0_+) = \eta\Phi'(0_+) - (1-\eta)\Phi'(0_-) < 0 \tag{4.3}$$

$$\arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subseteq \mathbb{R}_-^* \Leftrightarrow C'_\eta(0_-) = \eta\Phi'(0_-) - (1-\eta)\Phi'(0_+) > 0. \tag{4.4}$$

1. Assume that Φ is calibrated. By letting η tend to 1/2 in (4.3), we see that $C'_{1/2}(0_+) = \frac{1}{2}[\Phi'(0_+) - \Phi'(0_-)] \leqslant 0$. But, since ϕ is convex, $\Phi'(0_+) - \Phi'(0_-) \geqslant 0$. Therefore, the left and right derivatives are equal, which implies that Φ is differentiable at 0. Then $C'_\eta(0) = (2\eta - 1)\Phi'(0)$ and, according to (4.1) and (4.3), one must have $\Phi'(0) < 0$.

2. Assume that Φ is differentiable at 0 and $\Phi'(0) < 0$. Then $C'_\eta(0) = (2\eta - 1)\Phi'(0)$, and identities (4.1) and (4.2) are immediate consequences of (4.3) and (4.4).

对于指数损失 $C_\eta(\alpha) = \eta \cdot e^{-\alpha} + (1-\eta)\cdot e^{\alpha}$

$C'_\eta(\alpha) = -\eta \cdot e^{-\alpha} + (1-\eta)\cdot e^{\alpha}$

$C'_\eta(\alpha) = 0 \iff \eta \cdot e^{-\alpha} = (1-\eta)\cdot e^{\alpha}$

$e^{2\alpha} = \frac{1-\eta}{\eta}$

$\alpha = \frac{1}{2}\cdot \log(\frac{1-\eta}{\eta})$

对于逻辑损失 $C_\eta(\alpha) = \eta \cdot \log_2(1+e^{-\alpha}) + (1-\eta)\cdot \log_2(1+e^{\alpha})$

$C'_\eta(\alpha) = \frac{\eta}{\ln 2}\cdot\frac{1}{1+e^{-\alpha}}\cdot(-e^{-\alpha}) + \frac{1-\eta}{\ln 2}\cdot\frac{1}{1+e^{\alpha}}\cdot e^{\alpha}$

$C'_\eta(\alpha) = 0 \iff \frac{\eta}{1+e^{-\alpha}}\cdot e^{-\alpha} = \frac{1-\eta}{1+e^{\alpha}}\cdot e^{\alpha}$

$\eta\cdot(e^{-\alpha}+1) = (1-\eta)\cdot(e^{\alpha}+1)$

$\alpha = \log(\frac{1-\eta}{\eta})$

对于平方损失 $C_\eta(\alpha) = \eta(1-\alpha)^2 + (1-\eta)\cdot(1+\alpha)^2$

$C'_\eta(\alpha) = -\eta\cdot 2(1-\alpha) + 2(1-\eta)\cdot(1+\alpha)$

$C'_\eta(\alpha) = 0 \iff (1-\eta)(1+\alpha) = \eta(1-\alpha)$

$(1-\eta) + \alpha(1-\eta) = \eta - \eta\cdot\alpha$

$\alpha = 2\eta - 1$

∎

- Convex and classification-calibrated losses:
  上凸的 分类校准的 损失函数
  - 平方损失 **Squared loss**: $f^*(x) = 2\eta(x) - 1$.
  - 指数损失 **Exponential loss**: $f^*(x) = \frac{1}{2}\log(\frac{\eta(x)}{1-\eta(x)})$.
  - 逻辑损失 **Logistic loss**: $f^*(x) = \log(\frac{\eta(x)}{1-\eta(x)})$.
  - 合页损失 **Hinge loss**: $f^*(x) = 2\mathbf{1}_{[\eta(x)>1/2]} - 1$ (Bayes classifier itself!).
- **Last step**: connect $\mathscr{R}(f) - \mathscr{R}^*$ with $\mathscr{R}_\Phi(f) - \mathscr{R}^*_\Phi$.
- **Tool**: $H(\eta) = \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha)$.

Theorem 4.1 — Excess risks 超额风险

Let ϕ be convex and classification-calibrated. Assume that there exist constants $c \geqslant 0$ and $s \geqslant 1$ satisfying

$$\left|\frac{1}{2} - \eta\right|^s \leqslant c^s(1 - H(\eta)), \quad \eta \in [0,1].$$

Then, for any function $f : \mathscr{X} \to \mathbb{R}$,

$$\mathscr{R}(f) - \mathscr{R}^* \leqslant 2c(\mathscr{R}_\Phi(f) - \mathscr{R}^*_\Phi)^{1/s}.$$

- Examples:
  - **Squared loss**: $H(\eta) = 4\eta(1-\eta)$, $c = 1/2$, and $s = 2$.
  - **Exponential loss**: $H(\eta) = 2\sqrt{\eta(1-\eta)}$, $c = 1/\sqrt{2}$, and $s = 2$.
  - **Logistic loss**: $H(\eta) = -\eta\log_2\eta - (1-\eta)\log_2(1-\eta)$, $c = 1/\sqrt{2}$, and $s = 2$.
  - **Hinge loss**: $H(\eta) = 2\min(\eta, 1-\eta)$, $c = 1/2$, and $s = 1$.

对于平方损失

$H(\eta) = C_\eta(2\eta-1)$

$= \eta\cdot(1-2\eta+1)^2 + (1-\eta)(1+2\eta-1)^2$

$= 4\cdot\eta(1-\eta)^2 + 4\cdot\eta^2(1-\eta)$

$= 4\eta(1-\eta)$

对于指数损失

$H(\eta) = C_\eta(\frac{1}{2}\cdot\log(\frac{\eta}{1-\eta}))$

$= \eta\cdot e^{-\frac{1}{2}\log(\frac{\eta}{1-\eta})} + (1-\eta)\cdot e^{\frac{1}{2}\log(\frac{\eta}{1-\eta})}$

$= \eta\cdot\sqrt{\frac{1-\eta}{\eta}} + (1-\eta)\cdot\sqrt{\frac{\eta}{1-\eta}}$

$= 2\cdot\sqrt{\eta\cdot(1-\eta)}$

$|\frac{1}{2}-\eta|^2 \leq \frac{1}{2}\cdot(1-2\cdot\sqrt{\eta(1-\eta)})$

$\frac{1}{4}+\eta^2-\eta \leq \frac{1}{2} - \sqrt{\eta(1-\eta)}$

$\sqrt{\eta(1-\eta)} \leq \frac{1}{4} + \eta(1-\eta)$ 由 $(\sqrt{\eta(1-\eta)} - \frac{1}{2})^2 \geq 0$ 可得

对于逻辑损失

$C_\eta(\log(\frac{\eta}{1-\eta})) = \eta\cdot\log_2(1+\frac{1-\eta}{\eta}) + (1-\eta)\cdot\log_2(1+\frac{\eta}{1-\eta})$

$= -\eta\cdot\log_2(\eta) - (1-\eta)\cdot\log_2(1-\eta)$

Notational convention
符号惯例

- 回归 **Regression setting**: a predictor is a function $f : \mathscr{X} \to \mathbb{R}$. Loss: $\ell(y, f(x)) = (y - f(x))^2$.
- 分类 **Classification setting**: a classifier is a function $g : \mathscr{X} \to \{-1, 1\}$ (or $\{0, 1\}$), of the form

$$g(x) = \begin{cases} 1 & \text{if } f(x) > 0 \\ -1 & \text{otherwise,} \end{cases}$$

  where $f : \mathscr{X} \to \mathbb{R}$. The classifier g is the **classifier associated** with f. Loss: $\ell(y, g(x)) = \Phi(yf(x))$.
- **Unified notation**: the loss is $\ell(y, f(x))$ and the risk is $\mathscr{R}(f)$.
- **Advantage**: we only consider real-valued functions.

Risk minimization decomposition 风险最小化分解

- A family $\mathscr{F} = \{f : \mathscr{X} \to \mathbb{R}\}$ of predictors.

⚠ $\mathscr{Y} = \mathbb{R} \to$ regression **and** classification.

- Often $\mathscr{F} = \{f_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$. Example: linear models and neural networks.
- **Empirical risk minimization**: choose $f_n \in \mathscr{F}$ such that
经验风险最小化

$$f_n \in \arg\min_{f \in \mathscr{F}} \mathscr{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

- Risk decomposition: 风险分解

$$\begin{aligned} \mathscr{R}(f_n) - \mathscr{R}^* &= [\mathscr{R}(f_n) - \inf_{f \in \mathscr{F}} \mathscr{R}(f)] + [\inf_{f \in \mathscr{F}} \mathscr{R}(f) - \mathscr{R}^*] \\ &= \textbf{estimation error} + \textbf{approximation error}. \end{aligned}$$

估计误差 近似误差

⚠ The **estimation error** is random, the **approximation error** is deterministic.

- **Small** $\mathscr{F}$: restrictive. **Large** $\mathscr{F}$: overfitting.
- Bounding $\inf_{f \in \mathscr{F}} \mathscr{R}(f) - \mathscr{R}^*$ requires assumptions on (X, Y).

Approximation error 近似误差

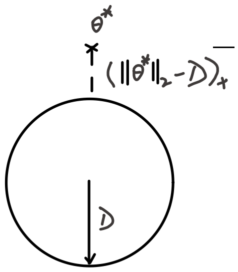
- Focus on $\mathscr{F} = \{f_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$:
  参数化函数空间
  - 假设 **Assumption**: there is $\theta^* \in \cancel{\Theta}\ \mathbb{R}^d$ such that $\mathscr{R}(f_{\theta^*}) = \inf_{\theta \in \mathbb{R}^d} \mathscr{R}(f_\theta)$.
  - One has

$$\begin{aligned}\inf_{\theta\in\Theta} \mathscr{R}(f_\theta) - \mathscr{R}^* &= [\inf_{\theta\in\Theta} \mathscr{R}(f_\theta) - \inf_{\theta\in\mathbb{R}^d} \mathscr{R}(f_\theta)] + [\inf_{\theta\in\mathbb{R}^d} \mathscr{R}(f_\theta) - \mathscr{R}^*] \\ &= [\inf_{\theta\in\Theta} \mathscr{R}(f_\theta) - \mathscr{R}(f_{\theta^*})] + [\mathscr{R}(f_{\theta^*}) - \mathscr{R}^*].\end{aligned}$$

  - The second term is **incompressible**. 第二项是不可压缩的.
  - The first term can be seen as a "distance" between θ^* and Θ.
  - 假设 **Assumption**: there exists $G \geqslant 0$ such that $\ell(Y, f_\theta(X))$ is **G-Lipschitz continuous** wp 1 with respect to the second variable.
  - Example: $Y = \pm 1$, choosing $\ell(y, f_\theta(x)) = \log_2(1 + e^{-y f_\theta(x)})$, one has $G = 1/\log 2$.
  - For each $\theta \in \Theta$,

$$\begin{aligned}\mathscr{R}(f_\theta) - \mathscr{R}(f_{\theta^*}) &\leqslant \mathbb{E}|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))| \\ &\leqslant G\,\mathbb{E}|f_\theta(X) - f_{\theta^*}(X)|.\end{aligned}$$

  - Example: $f_\theta(x) = \varphi(x)^\top \theta$ and $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_2 \leqslant D\}$.
  - In this case,



$$\begin{aligned}\inf_{\theta\in\Theta} \mathscr{R}(f_\theta) - \mathscr{R}(f_{\theta^*}) &\leqslant G \inf_{\|\theta\|_2 \leqslant D} \mathbb{E}|\varphi(X)^\top(\theta - \theta^*)| \\ \text{C-S} &\leqslant G\,\mathbb{E}\|\varphi(X)\|_2 \inf_{\|\theta\|_2 \leqslant D} \|\theta - \theta^*\|_2 \\ &= G\,\mathbb{E}\|\varphi(X)\|_2 (\|\theta^*\|_2 - D)_+.\end{aligned}$$

  - The bound is zero if $\|\theta^*\|_2 = D$ (well-specified model).

Estimation error 估计误差

Lemma 4.1. *One has*

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

$$\mathscr{R}(f_n) - \inf_{f\in\mathscr{F}} \mathscr{R}(f) \leqslant \sup_{f\in\mathscr{F}} (\mathscr{R}(f) - \mathscr{R}_n(f)) + \sup_{f\in\mathscr{F}} (\mathscr{R}_n(f) - \mathscr{R}(f)).$$

In particular, $\mathscr{R}(f_n) - \inf_{f\in\mathscr{F}} \mathscr{R}(f) \leqslant 2 \sup_{f\in\mathscr{F}} |\mathscr{R}_n(f) - \mathscr{R}(f)|$.

Proof. We have

$$\mathscr{R}(f_n) - \inf_{f\in\mathscr{F}} \mathscr{R}(f) = \mathscr{R}(f_n) - \mathscr{R}_n(f_n) + \mathscr{R}_n(f_n) - \inf_{f\in\mathscr{F}} \mathscr{R}(f).$$

Clearly,

$$\mathscr{R}(f_n) - \mathscr{R}_n(f_n) \leqslant \sup_{f\in\mathscr{F}}(\mathscr{R}(f) - \mathscr{R}_n(f)),$$

$f_n = \operatorname{argmin}_{f\in\mathscr{F}} R_n(f)$

and

$$\mathscr{R}_n(f_n) - \inf_{f\in\mathscr{F}} \mathscr{R}(f) = \inf_{f\in\mathscr{F}} \mathscr{R}_n(f) - \inf_{f\in\mathscr{F}} \mathscr{R}(f) \leqslant \sup_{f\in\mathscr{F}}(\mathscr{R}_n(f) - \mathscr{R}(f)).$$

This shows the first statement of the lemma. The second one is an immediate consequence. ■

- We need **uniform deviations** 一致偏差 of random variables from their means.
- The **easy** case $|\mathscr{F}| < \infty$:
  - 假设 **Assumption**: $\sup_{f\in\mathscr{F}} \ell(Y, f(X)) \leqslant \ell_\infty$ wp 1.
  - Example: $Y = \pm 1$, $g(x) = \mathbf{1}_{[f(x)>0]}$, where $\|f\|_\infty \leqslant B$. Choosing $\ell(y, f(x)) = \log_2(1 + e^{-yf(x)})$, one has $\ell_\infty = \log_2(1 + e^B)$.
  - By Hoeffding's inequality, for **each** $f \in \mathscr{F}$, for all $t > 0$,
  $$\mathbb{P}(|\mathscr{R}_n(f) - \mathscr{R}(f)| \geqslant t) \leqslant 2e^{-2nt^2/\ell_\infty^2}.$$
  - Thus,
  $$\mathbb{P}(\sup_{f\in\mathscr{F}} |\mathscr{R}_n(f) - \mathscr{R}(f)| \geqslant t) \leqslant 2|\mathscr{F}|e^{-2nt^2/\ell_\infty^2}.$$
  - In other words, for any $\delta \in (0, 1)$, wp at least $1 - \delta$,
  $$\begin{aligned}\sup_{f\in\mathscr{F}} |\mathscr{R}_n(f) - \mathscr{R}(f)| &\leqslant \frac{\ell_\infty}{\sqrt{2n}}\sqrt{\log\left(\frac{2|\mathscr{F}|}{\delta}\right)} \\ &\leqslant \ell_\infty\sqrt{\frac{\log(2|\mathscr{F}|)}{2n}} + \frac{\ell_\infty}{\sqrt{2n}}\sqrt{\log\left(\frac{1}{\delta}\right)}.\end{aligned}$$
  - Also, using Lemma 2.1 and Lemma 2.2,
  $$\mathbb{E} \sup_{f\in\mathscr{F}} |\mathscr{R}_n(f) - \mathscr{R}(f)| \leqslant \ell_\infty\sqrt{\frac{\log(2|\mathscr{F}|)}{2n}}.$$
- The **easy** case of quadratic functions 二次方程:

- Context: $\mathscr{F} = \{f_\theta(x) = \varphi(x)^\top\theta, \theta \in \Theta\}$, $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_2 \leqslant D\}$, $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.

假设 – Assumptions: $\mathbb{E}Y^2 < \infty$ and $\mathbb{E}\|\varphi(X)\|_2^2 < \infty$.

– For each $\theta \in \Theta$,

$$\begin{aligned}\mathscr{R}_n(f_\theta) - \mathscr{R}(f_\theta) = \theta^\top\Big(\frac{1}{n}\sum_{i=1}^n \varphi(X_i)\varphi(X_i)^\top - \mathbb{E}\varphi(X)\varphi(X)^\top\Big)\theta \\ - 2\theta^\top\Big(\frac{1}{n}\sum_{i=1}^n Y_i\varphi(X_i) - \mathbb{E}Y\varphi(X)\Big) \\ + \Big(\frac{1}{n}\sum_{i=1}^n Y_i^2 - \mathbb{E}Y^2\Big).\end{aligned}$$

Thus,

对于每一项都可由TCL收敛到0

$$\begin{aligned}&\sup_{\|\theta\|_2 \leqslant D} |\mathscr{R}_n(f_\theta) - \mathscr{R}(f_\theta)| \\ &\leqslant D^2\Big\|\frac{1}{n}\sum_{i=1}^n \varphi(X_i)\varphi(X_i)^\top - \mathbb{E}\varphi(X)\varphi(X)^\top\Big\|_{\mathrm{op}} \\ &\quad + 2D\Big\|\frac{1}{n}\sum_{i=1}^n Y_i\varphi(X_i) - \mathbb{E}Y\varphi(X)\Big\|_2 + \Big|\frac{1}{n}\sum_{i=1}^n Y_i^2 - \mathbb{E}Y^2\Big|,\end{aligned}$$

where $\|M\|_{\mathrm{op}} = \sup_{\|u\|_2=1}\|Mu\|_2$ is the operator norm of the matrix M. In particular, $|u^\top Mu| \leqslant \|M\|_{\mathrm{op}}\|u\|_2^2$ for any vector u, and $\|M\|_{\mathrm{op}} \leqslant \|M\|_2$.

- **Conclusion**: $\sup_{\|\theta\|_2 \leqslant D} |\mathscr{R}_n(f_\theta) - \mathscr{R}(f_\theta)| = \mathrm{O}_{\mathbb{P}}(1/\sqrt{n})$.

Rademacher complexity Rademacher 复杂度

- Assumption: $\sup_{f\in\mathscr{F}} \ell(Y, f(X)) \leqslant \ell_\infty$ wp 1. 依概率收敛
- **Notation**: $Z_i = (X_i, Y_i)$, $1 \leqslant i \leqslant n$,

$$H(Z_1, \dots, Z_n) = \sup_{f\in\mathscr{F}}(\mathscr{R}(f) - \mathscr{R}_n(f)),$$

and

$$\bar{H}(Z_1, \dots, Z_n) = \sup_{f\in\mathscr{F}}(\mathscr{R}_n(f) - \mathscr{R}(f)).$$

- By the bounded difference inequality, wp at least $1-\delta$,

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

and

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

- **Conclusion**: wp at least $1-\delta$,

$$H(Z_1,\ldots,Z_n) + \bar{H}(Z_1,\ldots,Z_n) \leqslant \mathbb{E}H(Z_1,\ldots,Z_n) + \mathbb{E}\bar{H}(Z_1,\ldots,Z_n) + \frac{\ell_\infty}{\sqrt{n}}\sqrt{2\log\left(\frac{2}{\delta}\right)}.$$

- **Focus** on $\mathbb{E}\sup_{f\in\mathscr{F}}(\mathscr{R}(f)-\mathscr{R}_n(f))$ and $\mathbb{E}\sup_{f\in\mathscr{F}}(\mathscr{R}_n(f)-\mathscr{R}(f))$.

- **General context**:
  - A random variable $Z\in\mathscr{Z}$.
  - A sample $\{Z_1,\ldots,Z_n\}$, i.i.d. copies of Z.
  - A class of functions $\mathscr{H}=\{h:\mathscr{Z}\to\mathbb{R}\}$.
  - **Target**: $Z=(X,Y)$ and $\mathscr{H}=\{h:(x,y)\mapsto \ell(y,f(x)), f\in\mathscr{F}\}$.
    目标 损失函数
  - **Rationale**:

$$\sup_{f\in\mathscr{F}}(\mathscr{R}(f)-\mathscr{R}_n(f)) = \sup_{h\in\mathscr{H}}\left(\mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^n h(Z_i)\right)$$

and

$$\sup_{f\in\mathscr{F}}(\mathscr{R}_n(f)-\mathscr{R}(f)) = \sup_{h\in\mathscr{H}}\left(\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}h(Z)\right).$$

- Rademacher complexity of $\mathscr{H}$: H 的 Rademacher复杂度

$$\mathbf{R}_n(\mathscr{H}) = \mathbb{E}\sup_{h\in\mathscr{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i h(Z_i),$$

where $\sigma_1,\ldots,\sigma_n$ are i.i.d. Rademacher random variables ($\mathbb{P}(\sigma_i=\pm1)=1/2$), independent of $Z_1,\ldots,Z_n$. (Note that $\mathbf{R}_n(\mathscr{H})\geqslant 0$. Why?)

对称性

Proposition 4.2 (Symmetrization). *One has*

$$\mathbb{E}\sup_{h\in\mathscr{H}}\left(\mathbb{E}h(Z)-\frac{1}{n}\sum_{i=1}^{n}h(Z_i)\right)\leqslant 2\mathbf{R}_n(\mathscr{H})$$

and

$$\mathbb{E}\sup_{h\in\mathscr{H}}\left(\frac{1}{n}\sum_{i=1}^{n}h(Z_i)-\mathbb{E}h(Z)\right)\leqslant 2\mathbf{R}_n(\mathscr{H}).$$

Proof. Introduce a "ghost sample" $Z_1',\dots,Z_n'$, independent of the Z_i and distributed identically. We have

$$\begin{aligned}
\mathbb{E}\sup_{h\in\mathscr{H}}\left(\mathbb{E}h(Z)-\frac{1}{n}\sum_{i=1}^{n}h(Z_i)\right) &= \mathbb{E}\sup_{h\in\mathscr{H}}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}h(Z_i')-\frac{1}{n}\sum_{i=1}^{n}h(Z_i)\mid Z_1,\dots,Z_n\right)\\
&\leqslant \mathbb{E}\sup_{h\in\mathscr{H}}\frac{1}{n}\left(\sum_{i=1}^{n}h(Z_i')-\sum_{i=1}^{n}h(Z_i)\right)\\
&\quad(\text{since }\sup\mathbb{E}(\cdot)\leqslant\mathbb{E}\sup(\cdot))\\
&= \mathbb{E}\sup_{h\in\mathscr{H}}\frac{1}{n}\left(\sum_{i=1}^{n}\Big(h(Z_i')-h(Z_i)\Big)\right).
\end{aligned}$$

Now, let $\sigma_1,\dots,\sigma_n$ be independent Rademacher random variables, independent of the Z_i and Z_i'. Then

由于 Z_i' 和 Z_i 同分布，$E[h(Z_i')-h(Z_i)]=E[h(Z_i)-h(Z_i')]$
以及 σ_i 的分布，可得两式相等

$$\begin{aligned}
\mathbb{E}\sup_{h\in\mathscr{H}}\frac{1}{n}\left(\sum_{i=1}^{n}\Big(h(Z_i')-h(Z_i)\Big)\right) &= \mathbb{E}\sup_{h\in\mathscr{H}}\frac{1}{n}\left(\sum_{i=1}^{n}\sigma_i\Big(h(Z_i')-h(Z_i)\Big)\right)\\
&\leqslant 2\mathbb{E}\sup_{h\in\mathscr{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)=2\mathbf{R}_n(\mathscr{H}).
\end{aligned}$$

The proof of the second statement is similar. ■

集中原则

Proposition 4.3 (Contraction principle). *Let $b,a_i:\Theta\to\mathbb{R}$ be functions, and let $\varphi_i:\mathbb{R}\to\mathbb{R}$ be 1-Lipschitz-continuous functions, $1\leqslant i\leqslant n$. Then*

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left(b(\theta)+\sum_{i=1}^{n}\sigma_i\varphi_i(a_i(\theta))\right)\right]\leqslant\mathbb{E}\left[\sup_{\theta\in\Theta}\left(b(\theta)+\sum_{i=1}^{n}\sigma_i a_i(\theta)\right)\right].$$

Back to learning

- With $Z=(X,Y)$ and $\mathscr{H}=\{h:(x,y)\mapsto\ell(y,f(x)),f\in\mathscr{F}\}$, one has

$$\mathbb{E}\sup_{f\in\mathscr{F}}(\mathscr{R}(f)-\mathscr{R}_n(f))\leqslant 2\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(Y_i,f(X_i))$$

  and

$$\mathbb{E}\sup_{f\in\mathscr{F}}(\mathscr{R}_n(f)-\mathscr{R}(f))\leqslant 2\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(Y_i,f(X_i)).$$

- **Assumption**: there exists $G \geqslant 0$ such that $\ell(Y, f_\theta(X))$ is ***G*-Lipschitz continuous** wp 1 with respect to the second variable.

- **Contraction principle** applied conditionally on $\mathscr{D}_n$ with $b = 0$, $\Theta = \{(f(X_1), \dots, f(X_n)), f \in \mathscr{F}\} \subseteq \mathbb{R}^n$, $a_i(\theta) = \theta_i$, and $\varphi_i(u_i) = \ell(Y_i, u_i)$:

$$\mathbb{E}\Big(\sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(Y_i, f(X_i)) \mid \mathscr{D}_n\Big) \leqslant G\,\mathbb{E}\Big(\sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \mid \mathscr{D}_n\Big),$$

  and thus

$$\begin{aligned}\mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(Y_i, f(X_i)) &\leqslant G\,\mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \\ &= G\mathbf{R}_n(\mathscr{F}).\end{aligned}$$

- **Conclusion**:

$$\begin{aligned}\mathbb{E}\mathscr{R}(f_n) - \inf_{f \in \mathscr{F}} \mathscr{R}(f) &\leqslant 4G\mathbf{R}_n(\mathscr{F}) \\ &= 4G\,\mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i),\end{aligned}$$

  and, for any $\delta \in (0, 1)$, wp at least $1 - \delta$,

  估计误差

$$\begin{aligned}\mathscr{R}(f_n) - \inf_{f \in \mathscr{F}} \mathscr{R}(f) &\leqslant 4G\mathbf{R}_n(\mathscr{F}) + \frac{\ell_\infty}{\sqrt{n}} \sqrt{2 \log\Big(\frac{2}{\delta}\Big)} \\ &= 4G\,\mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) + \frac{\ell_\infty}{\sqrt{n}} \sqrt{2 \log\Big(\frac{2}{\delta}\Big)}.\end{aligned}$$

⚠ **Binary classification**, loss $\ell(y, z) = \mathbf{1}_{[z \neq y]}$:
二元分类

$$\mathbb{E}\mathscr{R}(g_n) - \inf_{g \in \mathscr{G}} \mathscr{R}(g) \leqslant 4\mathbb{E} \sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbf{1}_{[g(X_i) \neq Y_i]}$$

$\to$ **combinatorics and Vapnik-Chervonenkis (VC) theory**.

Ball-constrained linear predictions 球约束线性预测

- **Context**: $\mathscr{F} = \{f_\theta(x) = \varphi(x)^\top \theta, \theta \in \Theta\}$, where $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_2 \leqslant D\}$.
  线性预测函数族

- With $\Phi \in \mathbb{R}^{n\times d}$ the design matrix and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)^\top$, one has

设计矩阵

$\phi^\top\phi = \cdots$

$$
\begin{aligned}
\mathbf{R}_n(\mathscr{F}) &= \mathbb{E}\Big(\sup_{\|\theta\|_2 \leqslant D} \frac{1}{n}\sum_{i=1}^n \sigma_i \varphi(X_i)^\top \theta \Big) = \mathbb{E} \sup_{\|\theta\|_2 \leqslant D} \Big(\frac{1}{n} \boldsymbol{\sigma}^\top \Phi \theta \Big) \\
&= \frac{D}{n} \mathbb{E}\|\Phi^\top \boldsymbol{\sigma}\|_2 \leqslant \frac{D}{n} \sqrt{\mathbb{E}\|\Phi^\top \boldsymbol{\sigma}\|_2^2} \\
&\quad \text{(since } \sup_{\|\theta\|_2 \leqslant 1} u^\top \theta = \|u\|_2 \text{ and by Jensen's inequality)} \\
&= \frac{D}{n} \sqrt{\mathbb{E}\,\mathrm{tr}(\Phi^\top \boldsymbol{\sigma}\boldsymbol{\sigma}^\top \Phi)} = \frac{D}{n}\sqrt{\mathbb{E}\,\mathrm{tr}(\Phi^\top \Phi)} \\
&\quad \text{(since } \mathbb{E}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top = I_n\text{)} \\
&= \frac{D}{n} \sqrt{\sum_{i=1}^n \mathbb{E}\|\varphi(X_i)\|_2^2} = \frac{D}{\sqrt{n}} \sqrt{\mathbb{E}\|\varphi(X)\|_2^2}.
\end{aligned}
$$

$\mathrm{tr}(\sigma^\top \Phi \Phi^\top \sigma)$

$\mathrm{tr}(\Phi^\top \sigma \sigma^\top \Phi)$

- This bound is **dimension-free**.

- **Estimation error**:

  - **Assumptions**: there exists $G \geqslant 0$ such that $\ell(Y, f_\theta(X))$ is ***G*-Lipschitz continuous** wp 1 with respect to the second variable, and $\mathbb{E}\|\varphi(X)\|_2^2 \leqslant R^2$.
  - With f_{θ_n} the minimizer of the empirical risk, one has

$$
\mathbb{E}\mathscr{R}(f_{\theta_n}) - \inf_{\|\theta\|_2 \leqslant D} \mathscr{R}(f_\theta) \leqslant \frac{4GRD}{\sqrt{n}}.
$$

- **Approximation error**:

  - **Assumption**: there is $\theta^* \in \cancel{\Theta}$ $\mathbb{R}^d$ such that $\mathscr{R}(f_{\theta^*}) = \inf_{\theta \in \mathbb{R}^d} \mathscr{R}(f_\theta)$.
  - One has

$$
\begin{aligned}
\inf_{\|\theta\|_2 \leqslant D} \mathscr{R}(f_\theta) - \mathscr{R}(f_{\theta^*}) &\leqslant G \inf_{\|\theta\|_2 \leqslant D} \mathbb{E}|f_\theta(X) - f_{\theta^*}(X)| \\
&= G \inf_{\|\theta\|_2 \leqslant D} \mathbb{E}|\varphi(X)^\top(\theta - \theta^*)| \\
&\leqslant G\,\mathbb{E}\|\varphi(X)\|_2 \inf_{\|\theta\|_2 \leqslant D} \|\theta - \theta^*\|_2 \\
&\leqslant GR \inf_{\|\theta\|_2 \leqslant D} \|\theta - \theta^*\|_2 \\
&= GR(\|\theta^*\|_2 - D)_+.
\end{aligned}
$$

approximation

estimation

- **Conclusion**:

$$\mathbb{E}\mathscr{R}(f_{\theta_n}) - \mathscr{R}(f_{\theta^*}) \leqslant \frac{4GRD}{\sqrt{n}} + GR(\|\theta^*\|_2 - D)_+.$$

- If D is too large: **overfitting**. If D is too small: **underfitting**.

Kernel Methods

核方法

正定核函数 定义

- **Definition**: A symmetric function 对称函数 $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is a **positive-definite kernel** 正定核 if

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geqslant 0$$

for all $n \geqslant 1$, all $(x_1, \dots, x_n) \in \mathscr{X}^n$, and all $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$.

Theorem 5.1 — Moore-Aronszajn Moore-Aronszajn（正定核的判定定理）

The function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is a positive-definite kernel if and only if there exists a Hilbert space $\mathscr{H}$ and a function $\varphi : \mathscr{X} \to \cancel{\mathscr{X}}$ $\mathcal{H}$ 函数域 such that, for all $x, x' \in \cancel{\mathscr{X}}$ $\mathcal{X}$, $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathscr{H}}$.

- The Hilbert space $\mathscr{H}$ is the completion of the **space of functions** 以下形式函数 的 完备函数空间 $f : \mathscr{X} \to \mathbb{R}$ of the form $f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$.

再生性质

- **Reproducing properties**: for all $x \in \mathscr{X}$, $k(\cdot, x) \in \mathscr{H}$ 函数 and, for any $f \in \mathscr{H}$ 函数, $f(x) = \langle f, k(\cdot, x) \rangle_{\mathscr{H}}$.

$\varphi : \mathcal{X} \longrightarrow \mathcal{H}$

$x \longmapsto \varphi(x) = k(\cdot, x)$

- **Feature map** 特征映射: $\varphi(x) = k(\cdot, x)$ 函数 $\in \mathscr{H}$. In particular,

$$\langle \varphi(x), \varphi(x') \rangle_{\mathscr{H}} = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathscr{H}} = k(x, x').$$

- $\mathscr{H}$ is called the **reproducing kernel Hilbert space** 再生核 希尔伯特空间 (feature space) (特征空间) associated with k.

⚠ Hilbert space $\cancel{\Rightarrow}$ $\Leftarrow$ RKHS but the converse is not true. Example: $L^2(\mathbb{R}^d)$ is **not** a RKHS.

⚠ No assumption on the input space $\mathscr{X}$.

任何 H 中的函数 f 都可以写成以下形式

⚠ Each $f \in \mathscr{H}$ is of the form $f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathscr{H}}$, where $\theta \in \mathscr{H}$. In addition, $\|f\|_{\mathscr{H}} = \|\theta\|_{\mathscr{H}}$.

- Examples:
  - 线性核 – **Linear kernel**: $\mathscr{X} = \mathbb{R}^d$, $k(x, x') = x^\top x'$. It corresponds to linear functions $f_\theta(x) = \theta^\top x$, with $\|f_\theta\|_{\mathscr{H}} = \|\theta\|_2$.
  - 多项式核 – **Polynomial kernel**: $\mathscr{X} = \mathbb{R}^d$ and for r a positive integer,

$$\begin{aligned} k(x, x') &= (x^\top x')^r \\ &= \sum_{\alpha_1+\cdots+\alpha_d=r} \binom{r}{\alpha_1, \ldots, \alpha_d} (x_1^{\alpha_1} \cdots x_d^{\alpha_d})((x_1')^{\alpha_1} \cdots (x_d')^{\alpha_d}). \end{aligned}$$

    Explicit feature map: $\varphi(x) = \left(\binom{r}{\alpha_1,\ldots,\alpha_d}^{1/2} x_1^{\alpha_1} \cdots x_d^{\alpha_d}\right)_{\alpha_1+\cdots+\alpha_d=r}$.
    显式特征映射
    The set of functions is the set of degree-r homogeneous polynomials on $\mathbb{R}^d$, with dimension $\binom{d+r-1}{r}$.
  - 指数核 – **Exponential kernel**: $k(x, x') = \exp(-\|x - x'\|_2 / r)$, where $r > 0$ is the bandwidth.
  - 高斯核 – **Gaussian kernel**: $k(x, x') = \exp(-\|x - x'\|_2^2 / r^2)$.
  - Kernels on point clouds, texts, sequences, images, graphs, etc.

Generalization guarantees 泛化保证

- **Context**: a kernel k on $\mathscr{X} \times \mathscr{X}$ and a loss function $\ell(Y, f(X))$ that is ***G*-Lipschitz continuous** wp 1 with respect to the second variable.

- **Constrained problem**:
  (球) 约束问题

$$f_n \in \arg\min_{f \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) \quad \text{such that } \|f\|_{\mathscr{H}} \leqslant D.$$

- 假设
  Assumptions: one has $\mathbb{E}|f(X)|^2 < \infty$ for all $f \in \mathscr{H}$, there exists $f^* \in \arg\min_{f:\mathscr{X}\to\mathbb{R}} \mathscr{R}(f)$, and $\mathbb{E}|f^*(X)|^2 < \infty$.
  不一定在H中
  选择一个核函数近似f^*

- **Excess risk**:
  超额风险

$$\begin{aligned} \mathscr{R}(f) - \mathscr{R}(f^*) &\leqslant \mathbb{E}|\ell(Y, f(X)) - \ell(Y, f^*(X))| \leqslant G\,\mathbb{E}|f(X) - f^*(X)| \\ &\leqslant G\sqrt{\mathbb{E}|f(X) - f^*(X)|^2} = G\|f - f^*\|_{L^2(\mu)}. \end{aligned}$$

- If $\sup_{x\in\mathscr{X}} k(x,x) \leqslant R^2$, then

$$\mathbb{E}\mathscr{R}(f_n) - \mathscr{R}(f^*) \leqslant \frac{4GDR}{\sqrt{n}} + G \inf_{\|f\|_{\mathscr{H}}\leqslant D} \|f - f^*\|_{L^2(\mu)}.$$

- The proof is similar to that of the linear model seen in the previous chapter and uses the following lemma.

Lemma 5.1. *Let* $\mathscr{F} = \{f \in \mathscr{H}, \|f\|_{\mathscr{H}} \leqslant D\}$. *Then*

$$\mathbf{R}_n(\mathscr{F}) \leqslant \frac{D}{n}\mathbb{E}\sqrt{\sum_{i=1}^n k(X_i, X_i)}.$$

↳ 核矩阵的迹

Proof. Observe that

Ramacher 复杂度

$$\begin{aligned}
\mathbf{R}_n(\mathscr{F}) &= \mathbb{E} \sup_{\|f\|_{\mathscr{H}}\leqslant D} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) \\
&= \frac{1}{n}\mathbb{E} \sup_{\|f\|_{\mathscr{H}}\leqslant D} \sum_{i=1}^n \sigma_i \langle f, k(\cdot, X_i)\rangle_{\mathscr{H}} \\
&= \frac{1}{n}\mathbb{E} \sup_{\|f\|_{\mathscr{H}}\leqslant D} \Big\langle f, \sum_{i=1}^n \sigma_i k(\cdot, X_i)\Big\rangle_{\mathscr{H}} \\
&= \frac{D}{n}\mathbb{E}\Big\|\sum_{i=1}^n \sigma_i k(\cdot, X_i)\Big\|_{\mathscr{H}},
\end{aligned}$$

(def) $f(x) = \langle f, k(\cdot,x)\rangle_H$

$\leqslant \|f\|_H \cdot \|\sum_{i=1}^n \sigma_i k(\cdot, X_i)\|_H$

by the Cauchy-Schwarz inequality. Next, by Jensen's inequality, for any vectors $a_1, \dots, a_n$ in $\mathscr{H}$,

$$\left(\mathbb{E}\Big\|\sum_{i=1}^n \sigma_i a_i\Big\|_{\mathscr{H}}\right)^2 \leqslant \mathbb{E}\Big\|\sum_{i=1}^n \sigma_i a_i\Big\|_{\mathscr{H}}^2.$$

$a_i = k(\cdot, X_i)$

The conclusion follows from

$$\mathbb{E}\Big\|\sum_{i=1}^n \sigma_i a_i\Big\|_{\mathscr{H}}^2 = \mathbb{E}\sum_{i,j=1}^n \sigma_i\sigma_j\langle a_i, a_j\rangle_{\mathscr{H}} = \sum_{i=1}^n \|a_i\|_{\mathscr{H}}^2.$$

↳ σ_i 和 σ_j 独立

$i\neq j$ 时, $E[\sigma_i\sigma_j] = \underbrace{E[\sigma_i]}_{0}\cdot\underbrace{E[\sigma_j]}_{0} = 0$

■

Representer theorem 表征定理

- In practice, one solves the **penalized problem** 惩罚问题

$$\inf_{\theta\in\mathscr{H}} \sum_{i=1}^n \ell(Y_i, \underbrace{\langle\theta, \varphi(X_i)\rangle_{\mathscr{H}}}_{= f(X_i)}) + \frac{\lambda}{2}\underbrace{\|\theta\|_{\mathscr{H}}^2}_{= \|f\|_H^2}, \tag{5.1}$$

where $\lambda > 0$ is a regularization parameter.

Theorem 5.2 — Representer theorem 表征定理

Consider a feature map $\varphi : \mathscr{X} \to \mathscr{H}$. Let $(x_1, \ldots, x_n) \in \overset{\mathscr{X}^n}{\cancel{\mathscr{H}}^n}$, and assume that the functional $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ is strictly increasing with respect to the last variable. Then the infimum of

$$\Psi\Big(\langle \theta, \varphi(x_1)\rangle_{\mathscr{H}}, \ldots, \langle \theta, \varphi(x_n)\rangle_{\mathscr{H}}, \|\theta\|_{\mathscr{H}}^2\Big)$$

can be obtained by restricting to vectors θ of the form

$$\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i),$$

dimension finie

where $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$.

Proof. Since the context is clear, we drop the underscore notation $\mathscr{H}$ in the dot products and norms throughout the proof. Let $\theta \in \mathscr{H}$ and $\mathscr{H}_{\mathscr{D}} = \{\sum_{i=1}^n \alpha_i \varphi(x_i), (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n\} \subseteq \mathscr{H}$ be the linear span of the observed feature vectors. Let $\theta_{\mathscr{D}} \in \mathscr{H}_{\mathscr{D}}$ and $\theta_\perp \in \mathscr{H}_{\mathscr{D}}^\perp$ be such that $\theta = \theta_{\mathscr{D}} + \theta_\perp$. Then, for all $i \in \{1, \ldots, n\}$,

$$\langle \theta, \varphi(x_i)\rangle = \langle \theta_{\mathscr{D}}, \varphi(x_i)\rangle + \overset{=0}{\langle \theta_\perp, \varphi(x_i)\rangle} = \langle \theta_{\mathscr{D}}, \varphi(x_i)\rangle,$$

since $\theta_\perp \in \mathscr{H}_{\mathscr{D}}^\perp$. Moreover, according to the Pythagorean theorem, $\|\theta\|^2 = \|\theta_{\mathscr{D}}\|^2 + \|\theta_\perp\|^2$. Therefore,

$$\begin{aligned}
\Psi\Big(\langle \theta, \varphi(x_1)\rangle, \ldots, \langle \theta, \varphi(x_n)\rangle, \|\theta\|^2\Big) &= \Psi\Big(\langle \theta_{\mathscr{D}}, \varphi(x_1)\rangle, \ldots, \langle \theta_{\mathscr{D}}, \varphi(x_n)\rangle, \|\theta_{\mathscr{D}}\|^2 + \|\theta_\perp\|^2\Big) \\
&\geqslant \Psi\Big(\langle \theta_{\mathscr{D}}, \varphi(x_1)\rangle, \ldots, \langle \theta_{\mathscr{D}}, \varphi(x_n)\rangle, \|\theta_{\mathscr{D}}\|^2\Big),
\end{aligned}$$

with equality if and only if $\theta_\perp = 0$ since Ψ is strictly increasing with respect to the last variable. Thus,

$$\inf_{\theta \in \mathscr{H}} \Psi\Big(\langle \theta, \varphi(x_1)\rangle, \ldots, \langle \theta, \varphi(x_n)\rangle, \|\theta\|^2\Big) = \inf_{\theta \in \mathscr{H}_{\mathscr{D}}} \Psi\Big(\langle \theta_{\mathscr{D}}, \varphi(x_1)\rangle, \ldots, \langle \theta_{\mathscr{D}}, \varphi(x_n)\rangle, \|\theta_{\mathscr{D}}\|^2\Big),$$

which is the desired result. ■

- **Conclusion**:
  - For $\lambda > 0$, the infimum of (5.1) can be obtained by **restricting** to vectors θ of the form $\theta = \sum_{i=1}^n \alpha_i \varphi(x_i)$.
  - This is a **finite-dimensional** space.
    有限维空间
- **Kernel matrix** $K \in \mathbb{R}^{n \times n}$:
  核矩阵

$$K_{ij} = \langle \varphi(X_i), \varphi(X_j)\rangle_{\mathscr{H}} = k(X_i, X_j).$$

- For $\theta = \sum_{i=1}^n \alpha_i \varphi(X_i)$ and $\alpha = (\alpha_1, \ldots, \alpha_n)^\top \in \mathbb{R}^n$,

$$\langle \theta, \varphi(X_j) \rangle_{\mathscr{H}} = \sum_{i=1}^n \alpha_i k(X_i, X_j) = (K\alpha)_j$$

$$\langle \theta, \varphi(X_i) \rangle_{H} = \langle \sum_{j=1}^n \alpha_j \cdot \varphi(X_j), \varphi(X_i) \rangle_{H} = \sum_{j=1}^n \alpha_j \cdot \langle \varphi(X_j), \varphi(X_i) \rangle_{H} = \sum_{j=1}^n \alpha_j \cdot k(X_j, X_i)$$

and

$$\|\theta\|^2_{\mathscr{H}} = \sum_{i,j=1}^n \alpha_i \alpha_j \langle \varphi(X_i), \varphi(X_j) \rangle_{\mathscr{H}} = \sum_{i,j=1}^n \alpha_i \alpha_j k(X_i, X_j) = \alpha^\top K \alpha.$$

- **Conclusion**:
总结

实际计算的是这个

$$\inf_{\theta \in \mathscr{H}} \sum_{i=1}^n \ell(Y_i, \underbrace{\langle \theta, \varphi(X_i) \rangle_{\mathscr{H}}}_{\text{不需要计算}}) + \frac{\lambda}{2} \underbrace{\|\theta\|^2_{\mathscr{H}}}_{\text{不需要计算}} = \inf_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \ell(Y_i, \underbrace{(K\alpha)_i}_{\text{转换为 计算核矩阵}}) + \frac{\lambda}{2} \alpha^\top K \alpha.$$

- **Prediction function**: for $x \in \mathscr{X}$,
预测函数

$$f(x) = \langle \theta, \varphi(x) \rangle_{\mathscr{H}} = \sum_{i=1}^n \alpha_i \langle \varphi(x), \varphi(X_i) \rangle_{\mathscr{H}} = \sum_{i=1}^n \alpha_i k(x, X_i).$$

- **Take-home messages**:
  - The input observations are summarized in the kernel matrix and the kernel function. 输入观察结果总结为 核矩阵 和 核函数
  - This is independent of the dimension of $\mathscr{H}$.
  - Explicit computing of the feature vector $\varphi(X)$ is **never needed**, as we solely need **dot products**. 特征向量 φ(X)的显式计算是不需要的，因为我们只需要点积
  - This is the **kernel trick**.
核技巧

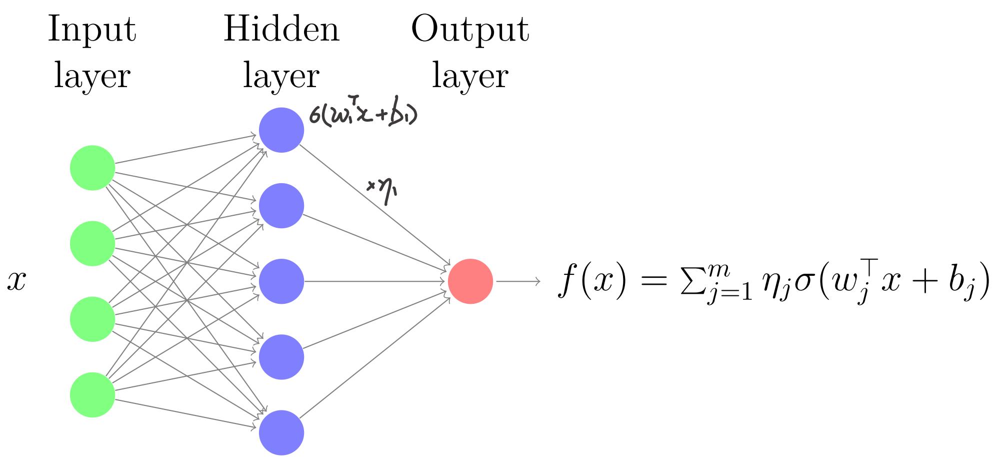
Neural Networks

神经网络

- We consider $\mathscr{X} = \mathbb{R}^d$ and prediction functions of the form
  预测函数

$$f(x) = \sum_{j=1}^{m} \eta_j \sigma(w_j^\top x + b_j),$$

where $w_j \in \mathbb{R}^d$, $b_j \in \mathbb{R}$ are the **input weights**, $\eta_j \in \mathbb{R}$ are the **output weights**, $1 \leqslant j \leqslant m$, and σ is an **activation function**.



- Typical activations (see Figure 6.1):
  激活函数
  - **Step function**: $\sigma(u) = \mathbf{1}_{[u \geqslant 0]}$.
  - **Sigmoid**: $\sigma(u) = \frac{1}{1+e^{-u}}$.
  - **ReLU**: $\sigma(u) = \max(u, 0)$.
  - **Hyperbolic tangent**: $\sigma(u) = \tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$.
- Each function $x \mapsto \sigma(w_j^\top x + b_j)$ is called a **neuron**.
- This is a neural network with **one hidden layer** $\rightarrow$ easy extension to multiple layers.

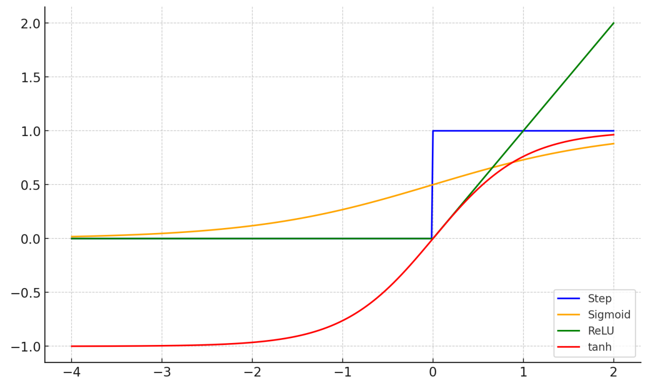


Figure 6.1: Typical activation functions.

Estimation error 估计误差

- **Notation and assumptions**:
  - $\|X\|_2 \leqslant R$ wp 1.
  - 参数 $\theta = ((\eta_j), (\overset{d维}{w_j}), (b_j), 1 \leqslant j \leqslant m) \in \mathbb{R}^{m(d+2)}$, $\eta = (\eta_1, \ldots, \eta_m) \in \mathbb{R}^m$.
  - 参数空间 $\Theta = \{\theta \in \mathbb{R}^{m(d+2)} : \|\eta\|_1 \leqslant D, \|w_j\|_2^2 + b_j^2/R^2 = 1, 1 \leqslant j \leqslant m\}$.
  - 预测函数 $f_\theta(x) = \sum_{j=1}^n \eta_j \sigma(w_j^\top x + b_j)$.
  - 预测函数空间 $\mathscr{F} = \{f_\theta, \theta \in \Theta\}$.
  - 激活函数 The activation function σ is **G_σ-Lipschitz continuous**.
  - 损失函数 The loss function $\ell(Y, f_\theta(X))$ is **G-Lipschitz continuous** wp 1 with respect to the second variable.

- We have

$$\mathbf{R}_n(\mathscr{F}) = \mathbb{E} \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \sigma_i f_\theta(X_i) = \mathbb{E} \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \eta_j \sigma_i \sigma(w_j^\top X_i + b_j).$$

Using the ℓ_1-constraint on η and $\sup_{\|\eta\|_1 \leqslant D} u^\top \eta = D\|u\|_\infty$, we are led to

$$\begin{aligned}\mathbf{R}_n(\mathscr{F}) &\leqslant D\,\mathbb{E} \sup_{j \in \{1,\ldots,m\}} \sup_{\|w_j\|_2^2 + b_j^2/R^2 = 1} \cancel{\sup_{s \in \{-1,1\}}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \sigma(w_j^\top X_i + b_j) \right| \\ &= D\,\mathbb{E} \sup_{\|w\|_2^2 + b^2/R^2 = 1} \sup_{s \in \{-1,1\}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \sigma(w^\top X_i + b) \right|.\end{aligned}$$

Thus, since σ is G_σ-Lipschitz continuous, by Proposition 4.3,

$$\mathbf{R}_n(\mathscr{F}) \leqslant 2DG_\sigma\,\mathbb{E} \sup_{\|w\|_2^2+b^2/R^2=1} \cancel{\sup_{s\in\{-1,1\}}} \cancel{s} \left| w^\top \underbrace{\Big(\frac{1}{n}\sum_{i=1}^n \sigma_i X_i\Big)}_{z} + b \underbrace{\Big(\frac{1}{n}\sum_{i=1}^n \sigma_i\Big)}_{t} \right|.$$

Observe that, by the Cauchy-Schwarz inequality,

$$\sup_{\|w\|_2^2+b^2/R^2=1} z^\top w + t^\top b = \sup_{\|w\|_2^2+c^2=1} |z^\top w + (Rt)^\top c| = (\|z\|_2^2 + R^2 t^2)^{1/2}.$$

?

We obtain

$$\mathbf{R}_n(\mathscr{F}) \leqslant 2DG_\sigma\,\mathbb{E}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i X_i\Big\|_2^2 + R^2\Big(\frac{1}{n}\sum_{i=1}^n \sigma_i\Big)^2\Big)^{1/2}.$$

We conclude, by Jensen's inequality, that

$\mathbb{E}[\sigma_i^2]=1$

$$\begin{aligned}\mathbf{R}_n(\mathscr{F}) &\leqslant 2DG_\sigma\Big[\mathbb{E}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i X_i\Big\|_2^2 + R^2\Big(\frac{1}{n}\sum_{i=1}^n \sigma_i\Big)^2\Big)\Big]^{1/2}\\ &= 2DG_\sigma\Big(\frac{1}{n}\mathbb{E}\|X\|_2^2 + \frac{R^2}{n}\Big)^{1/2} \leqslant \frac{2\sqrt{2}DG_\sigma R}{\sqrt{n}} \leqslant \frac{4DG_\sigma R}{\sqrt{n}}.\end{aligned}$$

- **Conclusion**: if $\|\theta\|_1 \leqslant D$ and $\|w_j\|_2^2 + b_j^2/R^2 = 1$ for all $j \in \{1, \ldots, m\}$,
  总结

$$\mathbb{E}\mathscr{R}(f_n) - \inf_{f\in\mathscr{F}} \mathscr{R}(f) \leqslant \frac{16GDG_\sigma R}{\sqrt{n}}.$$

⚠ The number of parameters is **irrelevant**. What matters is the overall norm of the weights.

Approximation properties 近似性质

- $\mathscr{F}_m$ = the class of all neural networks with one hidden layer of m nodes.
  神经网络函数族 n个结点的一个隐藏层
- **±1 binary classification**: for $f \in \mathscr{F}_m$, the associated classifier is $g(x) = 2\mathbf{1}_{[f(x)>0]} - 1$.
- **Loss**: $\ell(y, f(x)) = \mathbf{1}_{[f(x)\neq y]}$. **Risk**: $\mathscr{R}(f) = \mathbb{P}(g(X) \neq Y)$.
  损失函数 风险函数
- **Notation**: $\eta(x) = \mathbb{P}(Y = 1|X = x)$, $f^*(x) = 2\eta(x) - 1$, $g^*(x) = 2\mathbf{1}_{[f^*(x)>0]} - 1$.

Lemma 6.1. *One has*

$$\begin{aligned}\mathscr{R}(f)-\mathscr{R}^* &= \mathbb{E}|2\eta(X)-1|\mathbf{1}_{[g(X)\neq g^*(X)]}\\ &\leqslant \mathbb{E}|2\eta(X)-1-f(X)|.\end{aligned}$$

Proof. Observe that

$$\begin{aligned}&\mathbb{P}(g(X)\neq Y|X)\\ &\quad= 1-\mathbb{P}(g(X)=Y|X) = 1-\Big(\mathbb{P}(g(X)=1,Y=1|X)+\mathbb{P}(g(X)=-1,Y=-1|X)\Big)\\ &\quad= 1-\Big(\mathbf{1}_{[g(X)=1]}\mathbb{P}(Y=1|X)+\mathbf{1}_{[g(X)=-1]}\mathbb{P}(Y=-1|X)\Big)\\ &\quad= 1-\Big(\mathbf{1}_{[g(X)=1]}\eta(X)+\mathbf{1}_{[g(X)=-1]}(1-\eta(X))\Big).\end{aligned}$$

Similarly,

$$\mathbb{P}(g^*(X)\neq Y|X) = 1-\Big(\mathbf{1}_{[g^*(X)=1]}\eta(X)+\mathbf{1}_{[g^*(X)=-1]}(1-\eta(X))\Big).$$

Therefore,

$$\begin{aligned}&\mathbb{P}(g(X)\neq Y|X)-\mathbb{P}(g^*(X)\neq Y|X)\\ &\quad= \eta(X)(\mathbf{1}_{[g^*(X)=1]}-\mathbf{1}_{[g(X)=1]})+(1-\eta(X))(\underset{1-\mathbb{1}\{g^*(x)=1\}}{\mathbf{1}_{[g^*(X)=-1]}}-\underset{1-\mathbb{1}\{g(x)=1\}}{\mathbf{1}_{[g(X)=-1]}})\\ &\quad= (2\eta(X)-1)(\mathbf{1}_{[g^*(X)=1]}-\mathbf{1}_{[g(X)=1]})\\ &\quad= |2\eta(X)-1|\mathbf{1}_{[g(X)\neq g^*(X)]}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathscr{R}(f)-\mathscr{R}^* &= \mathbb{P}(g(X)\neq Y)-\mathbb{P}(g^*(X)\neq Y)\\ &= \mathbb{E}|2\eta(X)-1|\mathbf{1}_{[g(X)\neq g^*(X)]}\\ &\leqslant \mathbb{E}|2\eta(X)-1-f(X)|,\end{aligned}$$

since $g(x)\neq g^*(x)$ implies $|2\eta(x)-1-f(x)|\geqslant|2\eta(x)-1|$. ■

如果 $g(x)=1$, $g^*(x)=-1$, 则 $2\eta(x)-1>0$, $f(x)\leq 0$

Theorem 6.1 — Approximation error 近似误差

For the activation function (激活函数) $\sigma(u)=\mathbf{1}_{[u\geqslant 0]}$, one has

$$\lim_{m\to\infty}\inf_{f\in\mathscr{F}_m}\mathscr{R}(f)=\mathscr{R}^*$$

for all distributions of (X,Y).

The proof is a consequence of the next two propositions.

$$ER(f_n)-R^* = ER(f_n)-\inf_{f\in\mathcal{F}_m}R(f)+\underbrace{\inf_{f\in\mathcal{F}_m}R(f)-R^*}_{\xrightarrow{m\to\infty}0}$$

隐藏层神经元数

Proposition 6.1. *Let $(\mathscr{F}_m)_m$ be a sequence of classes of functions $f : \mathbb{R}^d \to \mathbb{R}$. Assume that for every $a, b \in \mathbb{R}^d$ and every continuous function h on $[a, b]^d$,*

函数族序列

$$\lim_{m\to\infty} \inf_{f\in\mathscr{F}_m} \sup_{x\in[a,b]^d} |h(x) - f(x)| = 0.$$

*Then, for **any** distribution of (X, Y),*

$$\lim_{m\to\infty} \inf_{f\in\mathscr{F}_m} \mathscr{R}(f) = \mathscr{R}^*.$$

Proof. For fixed $\varepsilon > 0$, find a, b such that $\mu([a, b]^d) \geqslant 1 - \varepsilon/3$, where μ is the distribution of X. Choose a continuous function η_ε vanishing off $[a, b]^d$ such that

$$\mathbb{E}|2\eta(X) - 1 - \eta_\varepsilon(X)| \leqslant \frac{\varepsilon}{6}.$$

For all m large enough, find $f \in \mathscr{F}_m$ such that

$$\sup_{x\in[a,b]^d} |\eta_\varepsilon(x) - f(x)| \leqslant \frac{\varepsilon}{6}.$$

For $g(x) = 2\mathbf{1}_{[f(x)>0]} - 1$, we have, by Lemma 6.1,

$$\begin{aligned}
\mathscr{R}(f) - \mathscr{R}^* &\leqslant \mathbb{E}|2\eta(X) - 1 - f(X)|\mathbf{1}_{[X\in[a,b]^d]} + \frac{\varepsilon}{3}\\
&\leqslant \mathbb{E}|\eta_\varepsilon(X) - f(X)|\mathbf{1}_{[X\in[a,b]^d]} + \mathbb{E}|2\eta(X) - 1 - \eta_\varepsilon(X)| + \frac{\varepsilon}{3}\\
&\leqslant \sup_{x\in[a,b]^d} |\eta_\varepsilon(x) - f(x)| + \mathbb{E}|2\eta(X) - 1 - \eta_\varepsilon(X)| + \frac{\varepsilon}{3}\\
&\leqslant \varepsilon.
\end{aligned}$$

We conclude that, for all m large enough,

$$\inf_{f\in\mathscr{F}_m} \mathscr{R}(f) - \mathscr{R}^* \leqslant \varepsilon.$$

■

对任意连续函数 h(x)

Proposition 6.2. *For every continuous function $h : [a, b]^d \to \mathbb{R}$ and for every $\varepsilon > 0$, there exists a neural network with one hidden layer and activation function $\sigma(u) = \mathbf{1}_{[u\geqslant 0]}$, of the form*

存在 只有一个隐藏层、激活函数为 σ(u) 的神经网络 ψ(x)

$$\psi(x) = \sum_{j=1}^{m} \eta_j \sigma(w_j^\top x + b_j),$$

such that

可以使得

$$\sup_{x\in[a,b]^d} |h(x) - \psi(x)| \leqslant \varepsilon. \implies \inf_{f\in\mathscr{F}_m} |h(x) - f(x)| \leqslant \varepsilon$$

Proof. Fix $\varepsilon > 0$ and take the Fourier series approximation of $h(x)$. By the Stone-Weierstrass theorem, there exists a positive integer M, nonzero real coefficients $\alpha_1, \dots, \alpha_M, \beta_1, \dots, \beta_M$, and real numbers $m_{i,j}$ for $1 \leqslant i \leqslant M$, $1 \leqslant j \leqslant d$, such that

$$\sup_{x \in [a,b]^d} \left| h(x) - \sum_{i=1}^{M} (\alpha_i \cos(m_i^\top x) + \beta_i \sin(m_i^\top x)) \right| \leqslant \frac{\varepsilon}{2},$$

where $m_i = (m_{i,1}, \dots, m_{i,d})^\top$, $1 \leqslant i \leqslant M$. It is clear that every continuous function on the real line can be arbitrarily closely approximated uniformly on compact intervals by one-dimensional neural networks, i.e., by functions of the form $\sum_{i=1}^k c_i \sigma(a_i x + b_i)$. Just observe that the indicator function of an interval $[b, c]$ may be written as $\sigma(x - b) + \sigma(-x + c) - 1$. This implies that the sin and cos functions can be approximated arbitrarily closely by neural networks on compact intervals. In particular, there exist neural networks $u_i(x)$, $v_i(x)$ with $1 \leqslant i \leqslant M$ (i.e., mappings from $\mathbb{R}^d$ to $\mathbb{R}$), such that

$$\sup_{x \in [a,b]^d} |\cos(m_i^\top x) - u_i(x)| \leqslant \frac{\varepsilon}{4M|\alpha_i|}$$

and

$$\sup_{x \in [a,b]^d} |\sin(m_i^\top x) - v_i(x)| \leqslant \frac{\varepsilon}{4M|\beta_i|}.$$

Therefore, applying the triangle inequality, we get

$$\sup_{x \in [a,b]^d} \left| \sum_{i=1}^{M} (\alpha_i \cos(m_i^\top x) + \beta_i \sin(m_i^\top x)) - \sum_{i=1}^{M} (\alpha_i u_i(x) + \beta_i v_i(x)) \right| \leqslant \frac{\varepsilon}{2}.$$

Since the u_i and v_i are neural networks, their linear combination

$$\psi(x) = \sum_{i=1}^{M} (\alpha_i u_i(x) + \beta_i v_i(x))$$

is a neural network too and, in fact,

$$\begin{aligned}
\sup_{x \in [a,b]^d} |h(x) - \psi(x)| &\leqslant \sup_{x \in [a,b]^d} \left| h(x) - \sum_{i=1}^{M} (\alpha_i \cos(m_i^\top x) + \beta_i \sin(m_i^\top x)) \right| \\
&\quad + \sup_{x \in [a,b]^d} \left| \sum_{i=1}^{M} (\alpha_i \cos(m_i^\top x) + \beta_i \sin(m_i^\top x)) - \psi(x) \right| \\
&\leqslant \frac{2\varepsilon}{2} = \varepsilon.
\end{aligned}$$

■

STONE'S THEOREM

Stone 定理

Plug-in principle Plug in 原则

- We consider $\mathscr{X} = \mathbb{R}^d$.
- Starting point:

$$g^*(x) = \begin{cases} 1 & \text{if } r(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

回归函数

最佳分类器

- **Idea**: estimate $r(x)$ from the training data $\mathscr{D}_n \rightsquigarrow r_n(x)$.
从训练集中估计 r(x)
- **Plug-in classifier**: Plug-in 分类器

$$g_n(x) = \begin{cases} 1 & \text{if } r_n(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

- **Question 1**: connection $r_n \leftrightarrow \mathscr{R}(g_n)$?
- **Question 2**: which choice for r_n?
- Plug-in $\rightsquigarrow$ **regression estimation** problem.
回归估计问题

Theorem 7.1 — Classification and regression 分类和回归

Let r_n be a **regression function estimator** of r, and let g_n be the **corresponding plug-in classifier**. Then
回归函数估计
相关 plug in 分类器

$$0 \leqslant \mathscr{R}(g_n) - \mathscr{R}^* \leqslant 2 \int_{\mathbb{R}^d} |r_n(x) - r(x)| \mu(dx).$$

In particular, for **all $p \geqslant 1$**,

$$0 \leqslant \mathscr{R}(g_n) - \mathscr{R}^* \leqslant 2 \Big(\int_{\mathbb{R}^d} |r_n(x) - r(x)|^p \mu(dx) \Big)^{1/p},$$

and

$$0 \leqslant \mathbb{E}\mathscr{R}(g_n) - \mathscr{R}^* \leqslant 2\big(\mathbb{E}|r_n(X) - r(X)|^p\big)^{1/p}.$$

- **Take-home message**:

$$\mathbb{E}\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \to 0$$

implies that the corresponding plug-in classifier g_n is **consistent**.

Proof. We have

$$\begin{aligned}
&\mathbb{P}(g_n(X) \neq Y | X, \mathscr{D}_n) \\
&\quad = 1 - \mathbb{P}(g_n(X) = Y | X, \mathscr{D}_n) \\
&\quad = 1 - \Big(\mathbb{P}(g_n(X) = 1, Y = 1 | X, \mathscr{D}_n) + \mathbb{P}(g_n(X) = 0, Y = 0 | X, \mathscr{D}_n)\Big) \\
&\quad = 1 - \Big(\mathbf{1}_{[g_n(X)=1]}\mathbb{P}(Y = 1 | X, \mathscr{D}_n) + \mathbf{1}_{[g_n(X)=0]}\mathbb{P}(Y = 0 | X, \mathscr{D}_n)\Big) \\
&\quad = 1 - \Big(\mathbf{1}_{[g_n(X)=1]} r(X) + \mathbf{1}_{[g_n(X)=0]}(1 - r(X))\Big),
\end{aligned}$$

where, in the last equality, we used the independence of (X, Y) and $\mathscr{D}_n$. Similarly,

$$\mathbb{P}(g^*(X) \neq Y | X) = 1 - \Big(\mathbf{1}_{[g^*(X)=1]} r(X) + \mathbf{1}_{[g^*(X)=0]}(1 - r(X))\Big).$$

Therefore,

$$\begin{aligned}
&\mathbb{P}(g_n(X) \neq Y | X, \mathscr{D}_n) - \mathbb{P}(g^*(X) \neq Y | X) \\
&\quad = r(X)(\mathbf{1}_{[g^*(X)=1]} - \mathbf{1}_{[g_n(X)=1]}) + (1 - r(X))(\mathbf{1}_{[g^*(X)=0]} - \mathbf{1}_{[g_n(X)=0]}) \\
&\quad = (2r(X) - 1)(\mathbf{1}_{[g^*(X)=1]} - \mathbf{1}_{[g_n(X)=1]}) \\
&\quad = |2r(X) - 1| \mathbf{1}_{[g_n(X) \neq g^*(X)]}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathscr{R}(g_n) - \mathscr{R}^* &= \mathbb{P}(g_n(X) \neq Y | \mathscr{D}_n) - \mathbb{P}(g^*(X) \neq Y) \\
&= 2\int_{\mathbb{R}^d} |r(x) - 1/2| \mathbf{1}_{[g_n(x) \neq g^*(x)]} \mu(dx) \\
&\leqslant 2\int_{\mathbb{R}^d} |r_n(x) - r(x)| \mu(dx),
\end{aligned}$$

在此条件下，如果 $r_n(x) > \frac{1}{2}$，则 $r(x) \leq \frac{1}{2}$
故

since $g_n(x) \neq g^*(x)$ implies $|r_n(x) - r(x)| \geqslant |r(x) - 1/2|$. The other assertions follow from Hölder's and Jensen's inequality, respectively. ■

Local average estimators 局部均值估计

- Definition: $r_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i$.

$Y_i = 0$或1

- **Important**: each $W_{ni}(x)$ is a function of x and $X_1, \ldots, X_n$ (and **not** of $Y_1, \ldots, Y_n$).
- **Weight vector**: $(W_{n1}(x), \ldots, W_{nn}(x))$.
  权重向量
- Interpretation: X_i "close" to x should provide more information.
- Often (but not always) $(W_{n1}(x), \ldots, W_{nn}(x))$ is a **probability vector**.
- Equivalently: $r_n(x) = \sum_{i=1}^n W_{ni}(x)\mathbf{1}_{[Y_i=1]}$.
- Companion **plug-in classifier**:

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n W_{ni}(x)Y_i > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

- Whenever $\sum_{i=1}^n W_{ni}(x) = 1$: 即权重向量是概率向量时

$(1-\mathbb{1}_{\{Y_i=1\}})$

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n W_{ni}(x)\mathbf{1}_{[Y_i=1]} > \sum_{i=1}^n W_{ni}(x)\mathbf{1}_{[Y_i=0]} \\ 0 & \text{otherwise.} \end{cases}$$

- **Example 1: kernel estimator** 例 1：核估计
  - **Definition**:

$$r_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)Y_i}{\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)}.$$

  核函数 K – **Kernel** K: a nonnegative real-valued function on $\mathbb{R}^d$.
  带宽 h – **Bandwidth** h: a positive real number (may depend on n).
  权重 – **Weights**:

$$W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)}.$$

  - If both denominator and numerator are **zero**: $r_n(x) = \frac{1}{n}\sum_{i=1}^n Y_i$.
    如果分子和分母都是 0
  - Kernels:

    简单核 ▷ **Naive**: $K(z) = \mathbf{1}_{[\|z\|_2 \leqslant 1]}$,

$$r_n(x) = \frac{\sum_{i=1}^n \mathbf{1}_{[\|x-X_i\|_2 \leqslant h]}Y_i}{\sum_{j=1}^n \mathbf{1}_{[\|x-X_j\|_2 \leqslant h]}}.$$

Epanechnikov 核 ▷ **Epanechnikov**: $K(z) = (1 - \|z\|_2^2)\mathbf{1}_{[\|z\|_2 \leqslant 1]}$.

高斯核 ▷ **Gaussian**: $K(z) = e^{-\|z\|_2^2}$.

- **Example 2: nearest neighbor (NN) estimator** 例 2：最小邻近估计
  - **Definition**:
    ▷ $(X_{(1)}(x), Y_{(1)}(x)), \dots, (X_{(n)}(x), Y_{(n)}(x))$ **reordering** of $\mathscr{D}_n$ according to
    $$\|X_{(1)}(x) - x\|_2 \leqslant \cdots \leqslant \|X_{(n)}(x) - x\|_2.$$
    ▷ Whenever $\|X_i - x\|_2 \leqslant \|X_j - x\|_2$ and $i < j$, we declare X_i **closer** to x.
    ▷ **NN estimator**: $r_n(x) = \sum_{i=1}^n v_{ni} \overset{\text{排序后的 Y}}{Y_{(i)}(x)}$, where $\sum_{i=1}^n v_{ni} = 1$.
  - $(\Sigma_1, \dots, \Sigma_n)$: **permutation** of $(1, \dots, n)$ such that X_i is the Σ_i-th nearest neighbor of x for all i.

局部均值 – **Local averaging**: $r_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i$, where $W_{ni}(x) = v_{n\Sigma_i}$.
  - ***k*-NN estimator**:
    $$v_{ni} = \begin{cases} \frac{1}{k} & \text{for } 1 \leqslant i \leqslant k \\ 0 & \text{for } k < i \leqslant n. \end{cases}$$
    只有 k 个邻近点有权重
  - To keep in mind: $r_n(x) = \frac{1}{k}\sum_{i=1}^k Y_{(i)}(x)$.

Theorem 7.2 — Stone Stone 定理

Let $r_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i$, with $(W_{n1}(x), \dots, W_{nn}(x))$ a **probability vector**. Assume that for any distribution of X, the weights satisfy the following conditions:

1. There is a constant C such that, for every Borel measurable function $f : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}|f(X)| < \infty$,
$$\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)|f(X_i)|\Big) \leqslant C\mathbb{E}|f(X)| \quad \text{for all } n \geqslant 1.$$
2. For all $a > 0$,
$$\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\mathbf{1}_{[\|X_i - X\|_2 > a]}\Big) \to 0.$$
局部性

3. One has

$$\mathbb{E} \max_{1 \leqslant i \leqslant n} W_{ni}(X) \to 0.$$

Then the corresponding plug-in classifier g_n is universally consistent, i.e., $\mathbb{E}\mathscr{R}(g_n) \to \mathscr{R}^*$ for all distributions of (X, Y). 一致收敛的

- Comments:
  - Condition 1 is merely technical.
  - Condition 2 ensures that $r_n(X)$ is asymptotically mostly influenced by the data points close to X.
  - Condition 3 states that asymptotically all weights become small.
  - No single observation has a too large contribution to the estimator.
  - The number of points in the averaging must tend to infinity.

Proof. According to Theorem 7.1, it suffices to prove that for every distribution of (X, Y),

$$\mathbb{E}|r_n(X) - r(X)|^2 = \mathbb{E} \int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \to 0.$$

Introduce the notation

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{ni}(x) r(X_i).$$

Then, by the simple inequality $(a+b)^2 \leqslant 2(a^2 + b^2)$, we have

$$\begin{aligned}
\mathbb{E}|r_n(X) - r(X)|^2 &= \mathbb{E}|r_n(X) - \hat{r}_n(X) + \hat{r}_n(X) - r(X)|^2 \\
&\leqslant 2\Big(\mathbb{E}|r_n(X) - \hat{r}_n(X)|^2 + \mathbb{E}|\hat{r}_n(X) - r(X)|^2\Big).
\end{aligned} \tag{7.1}$$

Therefore, it is enough to show that both terms on the right-hand side tend to zero as n tends to infinity. Since the W_{ni} are nonnegative and sum to one, by Jensen's inequality, one has

$$\begin{aligned}
\mathbb{E}|\hat{r}_n(X) - r(X)|^2 &= \mathbb{E}\Big|\sum_{i=1}^{n} W_{ni}(X)(r(X_i) - r(X))\Big|^2 \\
&\leqslant \mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)|r(X_i) - r(X)|^2\Big).
\end{aligned}$$

If the function r, which satisfies $0 \leqslant r \leqslant 1$, is continuous with compact support, then it is uniformly continuous as well: for every $\varepsilon > 0$, there is an $a > 0$ such that for $\|x - x'\|_2 \leqslant a$, $|r(x) - r(x')|^2 \leqslant \varepsilon$. Thus, since $|r(x) - r(x')| \leqslant 1$,

$$\begin{aligned}
\mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)|r(X_i) - r(X)|^2\Big) &\leqslant \mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)\mathbf{1}_{[\|X_i - X\|_2 > a]}\Big) + \mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)\varepsilon\Big) \\
&= \mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)\mathbf{1}_{[\|X_i - X\|_2 > a]}\Big) + \varepsilon.
\end{aligned}$$

Therefore, by condition 2, since ε is arbitrary,

$$\mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)|r(X_i) - r(X)|^2\Big) \to 0.$$

In the general case, since the set of continuous functions with compact support is dense in $L^2(\mu)$, for every $\varepsilon > 0$ we can choose r_ε taking values in $[0,1]$ and such that

$$\mathbb{E}|r(X) - r_\varepsilon(X)|^2 \leqslant \varepsilon.$$

By this choice, using the inequality $(a+b+c)^2 \leqslant 3(a^2+b^2+c^2)$ (which follows from the Cauchy-Schwarz inequality),

$$\begin{aligned}
&\mathbb{E}|\hat{r}_n(X) - r(X)|^2 \\
&\quad\leqslant \mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)|r(X_i) - r(X)|^2\Big) \\
&\quad\leqslant 3\mathbb{E}\bigg(\sum_{i=1}^{n} W_{ni}(X)\Big(|r(X_i) - r_\varepsilon(X_i)|^2 + |r_\varepsilon(X_i) - r_\varepsilon(X)|^2 + |r_\varepsilon(X) - r(X)|^2\Big)\bigg).
\end{aligned}$$

Thus, using condition 1,

$$\begin{aligned}
&\mathbb{E}|\hat{r}_n(X) - r(X)|^2 \\
&\quad\leqslant 3C\mathbb{E}|r(X) - r_\varepsilon(X)|^2 + 3\mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)|r_\varepsilon(X_i) - r_\varepsilon(X)|^2\Big) + 3\mathbb{E}|r_\varepsilon(X) - r(X)|^2 \\
&\quad\leqslant 3C\varepsilon + 3\mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)|r_\varepsilon(X_i) - r_\varepsilon(X)|^2\Big) + 3\varepsilon.
\end{aligned}$$

Therefore, $\mathbb{E}|\hat{r}_n(X) - r(X)|^2 \to 0$.

To handle the first term of the right-hand side of (7.1), observe that, for all $i \neq j$,

$$\begin{aligned}
&\mathbb{E}\Big(W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)(Y_j - r(X_j))\Big) \\
&\quad= \mathbb{E}\Big[\mathbb{E}\Big(W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)(Y_j - r(X_j)) \mid X, X_1, \ldots, X_n, Y_i\Big)\Big] \\
&\quad= \mathbb{E}\Big[W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)\mathbb{E}\Big(Y_j - r(X_j) \mid X, X_1, \ldots, X_n, Y_i\Big)\Big] \\
&\quad= \mathbb{E}\Big[W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)\mathbb{E}(Y_j - r(X_j) \mid X_j)\Big] \\
&\qquad\text{(by independence of } (X_j, Y_j) \text{ and } X, X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n, Y_i) \\
&\quad= \mathbb{E}\Big[W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)(r(X_j) - r(X_j))\Big] \\
&\quad= 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}|r_n(X) - \hat{r}_n(X)|^2 &= \mathbb{E}\Big|\sum_{i=1}^{n} W_{ni}(X)(Y_i - r(X_i))\Big|^2 \\
&= \sum_{i,j=1}^{n} \mathbb{E}\Big(W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)(Y_j - r(X_j))\Big) \\
&= \sum_{i=1}^{n} \mathbb{E}\Big(W_{ni}^2(X)(Y_i - r(X_i))^2\Big).
\end{aligned}$$

We conclude that

$$\mathbb{E}|r_n(X) - \hat{r}_n(X)|^2 \leqslant \mathbb{E}\sum_{i=1}^n W_{ni}^2(X) \leqslant \mathbb{E}\Big(\max_{1\leqslant i\leqslant n} W_{ni}(X) \sum_{j=1}^n W_{nj}(X)\Big)$$
$$= \mathbb{E}\max_{1\leqslant i\leqslant n} W_{ni}(X) \to 0$$

by condition 3, and the theorem is proved. ■

The k-NN estimator 最小 k 邻近估计

- Reordering $(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$ according to

  按距离排序 $$\|X_{(1)}(x) - x\|_2 \leqslant \cdots \leqslant \|X_{(n)}(x) - x\|_2.$$

- Whenever $\|X_i - x\|_2 \leqslant \|X_j - x\|_2$ and $i < j$, we declare X_i closer to x.
- **k-NN regression function estimator**: $r_n(x) = \frac{1}{k}\sum_{i=1}^k Y_{(i)}(x)$.
  最小 k 邻近回归函数估计
- **k-NN classifier**:

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^k \mathbf{1}_{[Y_{(i)}(x)=1]} > \sum_{i=1}^k \mathbf{1}_{[Y_{(i)}(x)=0]} \\ 0 & \text{otherwise.} \end{cases}$$

⚠ If X has a density, then there is **no distance tie**.

Theorem 7.3 — Universal consistency 一致收敛性

Assume that $k \to \infty$ and $k/n \to 0$. Then the k-NN classifier is **universally consistent**, i.e., $\mathbb{E}\mathscr{R}(g_n) \to \mathscr{R}^*$ for **all** distributions of (X, Y).

- k is large but small with respect to n: **bias/variance compromise**.
- Proof's agenda: verify Stone's conditions 1-3.
- **Simplification**: distance ties $\|X_i - X\|_2 = \|X_j - X\|_2$ occur with zero probability.
- **Definition**: The **support** of μ is defined by
  测度的支集

  $$\text{supp}(\mu) = \big\{x \in \mathbb{R}^d : \mu(B(x,\rho)) > 0 \text{ for all } \rho > 0\big\},$$

  where $B(x, \rho)$ is the closed ball in $\mathbb{R}^d$ with center at x and radius ρ.

- **Properties**:
  - $\text{supp}(\mu)$ is a **closed** set.
  - $\text{supp}(\mu)$ is the **smallest** closed subset of $\mathbb{R}^d$ of μ-measure one.
  - One has $\mathbb{P}(X \in \text{supp}(\mu)) = 1$.

Lemma 7.1. *If* $x \in \text{supp}(\mu)$ *and* $k/n \to 0$, *then*

$$\|X_{(k)}(x) - x\|_2 \to 0 \quad \textit{wp 1.}$$

Proof. Take $\varepsilon > 0$ and note, since x belongs to the support of μ, that $\mu(B(x, \varepsilon)) > 0$. Observe that

$$\Big[\|X_{(k)}(x) - x\|_2 > \varepsilon\Big] = \Big[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[X_i \in B(x,\varepsilon)]} < \frac{k}{n}\Big].$$

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[X_i \in B(x,\varepsilon)]} \to \mu(B(x, \varepsilon)) \quad \text{wp 1.}$$

Since $k/n \to 0$, we conclude that $\|X_{(k)}(x) - x\|_2 \to 0$ wp 1. ■

Lemma 7.2. *Let* ν *be a probability measure on* $\mathbb{R}^d$. *Fix* $x' \in \mathbb{R}^d$ *and let, for* $a \geqslant 0$,

$$B_a(x') = \Big\{x \in \mathbb{R}^d : \nu\big(B(x, \|x' - x\|_2)\big) \leqslant a\Big\}.$$

Then

$$\nu(B_a(x')) \leqslant \gamma_d a,$$

where γ_d *is a positive constant depending only upon* d.

Proof. Fix $x' \in \mathbb{R}^d$ and let $\mathscr{C}_1, \ldots, \mathscr{C}_{\gamma_d}$ be a collection of cones of angle $0 < \theta \leqslant \pi/6$ covering $\mathbb{R}^d$, all centered at x' but with different central directions (such a covering is always possible). In other words,

$$\bigcup_{j=1}^{\gamma_d} \mathscr{C}_j = \mathbb{R}^d.$$

We leave it as an easy exercise to show that if $u \in \mathscr{C}_j$, $u' \in \mathscr{C}_j$, and $\|u - x'\|_2 \leqslant \|u' - x'\|_2$, then $\|u - u'\|_2 \leqslant \|u' - x'\|_2$ (see Figure 7.1). In addition,

$$\nu(B_a(x')) \leqslant \sum_{j=1}^{\gamma_d} \nu(\mathscr{C}_j \cap B_a(x')).$$

Let $x^* \in \mathscr{C}_j \cap B_a(x')$. Then, by the geometrical property of cones mentioned above, we have

$$\nu\Big(\mathscr{C}_j \cap B(x', \|x^* - x'\|_2) \cap B_a(x')\Big) \leqslant \nu\Big(B(x^*, \|x' - x^*\|_2)\Big) \leqslant a.$$

Since x^* was arbitrary, we conclude that

$$\nu(\mathscr{C}_j \cap B_a(x')) \leqslant a.$$

■

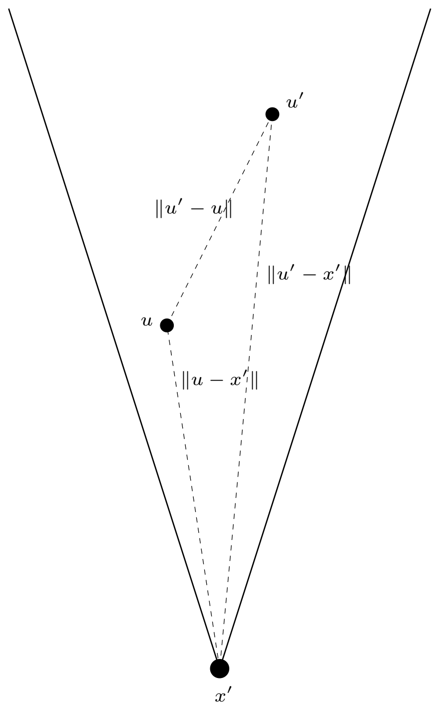


Figure 7.1: The geometrical property of a cone of angle $0 < \theta \leqslant \pi/6$ (in dimension 2).

Corollary 7.1. *If distance ties occur with zero probability, then*

$$\sum_{i=1}^{n} \mathbf{1}_{[\boldsymbol{X} \text{ is among the } k\text{-NN of } \boldsymbol{X_i} \text{ in } \{X_1,\ldots,X_{i-1},\boldsymbol{X},X_{i+1},\ldots,X_n\}]} \leqslant k\gamma_d,$$

wp 1.

Proof. We apply Lemma 7.2 with $a = k/n$ and ν the empirical measure μ_n associated with $X_1, \ldots, X_n$. With these choices,

$$B_{k/n}(X) = \Big\{x \in \mathbb{R}^d : \mu_n\Big(B(x, \|X - x\|_2)\Big) \leq k/n\Big\}$$

and, wp 1,

$$\begin{aligned} X_i \in B_{k/n}(X) &\Leftrightarrow \mu_n\Big(B(X_i, \|X - X_i\|_2)\Big) \leq k/n \\ &\Leftrightarrow X \text{ is among the } k\text{-NN of } X_i \text{ in } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}. \end{aligned}$$

(Note that the second equivalence uses the fact that distance ties occur with zero probability.) Thus, by Lemma 7.2, we conclude that, wp 1,

$$\begin{aligned} &\sum_{i=1}^{n} \mathbf{1}_{[X \text{ is among the } k\text{-NN of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}]} \\ &\quad = \sum_{i=1}^{n} \mathbf{1}_{[X_i \in B_{k/n}(X)]} = n \times \mu_n(B_{k/n}(X)) \leq k\gamma_d. \end{aligned}$$

■

Stone 引理

Lemma 7.3 (Stone's lemma)**.** *Assume that distance ties occur with zero probability. Then, for every Borel measurable function $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}|f(X)| < \infty$,*

$$\sum_{i=1}^{k} \mathbb{E}\big|f(X_{(i)}(X))\big| \leqslant k\gamma_d \mathbb{E}|f(X)|,$$

where γ_d is a positive constant depending only upon d.

Proof. Take f as in the lemma. Then

$$\begin{aligned}\sum_{i=1}^{k} \mathbb{E}\big|f(X_{(i)}(X))\big| &= \mathbb{E}\Big(\sum_{i=1}^{n} |f(X_i)| \mathbf{1}_{[X_i \text{ is among the } k\text{-NN of } X \text{ in } \{X_1,\ldots,X_n\}]}\Big) \\ &= \mathbb{E}\Big(|f(X)| \sum_{i=1}^{n} \mathbf{1}_{[X \text{ is among the } k\text{-NN of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}]}\Big) \\ &\quad \text{(by exchanging } X \text{ and } X_i) \\ &\leq \mathbb{E}(|f(X)| k\gamma_d),\end{aligned}$$

by Corollary 7.1. ■

- **To do**: verify the conditions of Stone's theorem with $W_{ni}(x) = 1/k$ if X_i is among the k nearest neighbors of x and $W_{ni}(x) = 0$ otherwise.

- Condition 3 is clear since $k \to \infty$.

- Condition 2: note that

$$\mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)\mathbf{1}_{[\|X_i - X\|_2 > a]}\Big) = \mathbb{E}\Big(\frac{1}{k}\sum_{i=1}^{k} \mathbf{1}_{[\|X_{(i)}(X) - X\|_2 > a]}\Big).$$

So,

lemme 7.1

$$\mathbb{P}(\|X_{(k)}(X) - X\|_2 > a) \to 0 \Rightarrow \mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)\mathbf{1}_{[\|X_i - X\|_2 > a]}\Big) \to 0.$$

But, for all $a > 0$,

$$\mathbb{P}(\|X_{(k)}(X) - X\|_2 > a) = \int_{\mathbb{R}^d} \mathbb{P}(\|X_{(k)}(x) - x\|_2 > a)\mu(dx).$$

Assuming that $k/n \to 0$, the conclusion follows by the Lebesgue dominated convergence theorem.

- **Condition 1**: take f such that $\mathbb{E}|f(X)| < \infty$. We have to show that for some constant C

$$\mathbb{E}\Big(\frac{1}{k}\sum_{i=1}^{n}|f(X_i)|\mathbf{1}_{[X_i \text{ is among the } k\text{-NN of } X]}\Big) \leqslant C\mathbb{E}|f(X)|.$$

Since

$$\mathbb{E}\Big(\frac{1}{k}\sum_{i=1}^{n}|f(X_i)|\mathbf{1}_{[X_i \text{ is among the } k\text{-NN of } X]}\Big) = \mathbb{E}\Big(\frac{1}{k}\sum_{i=1}^{k}\big|f(X_{(i)}(X))\big|\Big),$$

this is precisely the statement of Stone's lemma 7.3.

$$\leqslant \frac{1}{k}\cdot k\cdot \gamma_d\cdot \mathbb{E}|f(x)|$$
$$= \gamma_d\cdot \mathbb{E}|f(x)|$$

Choice of k k 的选择

- Choosing k by minimizing the empirical error is **not** a good idea. Why?

如果直接对所有样本最小化风险，会得到 $k=1$ 时，risque $=0$，并不合适（过拟合）

- **Data splitting**:
  - A **training** set $\mathscr{D}_m = \{(X_1,Y_1),\dots,(X_m,Y_m)\}$.
  - A **testing** set $\mathscr{D}_\ell = \{(X_{m+1},Y_{m+1}),\dots,(X_n,Y_n)\}$, with $m+\ell=n$.
- Candidates: $\mathscr{G}_m = \{g_k, 1\leqslant k\leqslant m\} \to k$-NN classifiers using $\mathscr{D}_m$.

使用 Dm 的 k-NN 分类器

- **Strategy**: choose $g_n^* \in \mathscr{G}_m$ such that

$$g_n^* \in \arg\min_{g_k\in\mathscr{G}_m} \frac{1}{\ell}\sum_{i=m+1}^{n}\mathbf{1}_{[g_k(X_i)\neq Y_i]}.$$

在 test 上最小化损失

Theorem 7.4 — Choice of k by data-splitting 通过划分数据集选择 k

One has

$$\mathbb{E}\big(\mathscr{R}(g_n^*) - \inf_{g_k\in\mathscr{G}_m}\mathscr{R}(g_k)\big) \leqslant 2\sqrt{\frac{\log(2\cancel{m}m)}{2\ell}}.$$

lemme 4.1

由 Stone 可得

- The classifier g_n^* is **universally consistent** provided

$$\lim_{n\to\infty} m = \infty \quad\text{and}\quad \lim_{n\to\infty}\frac{\ell}{\log m} = \infty.$$

$\mathbb{E}[$

PARTITIONING CLASSIFIERS AND TREES

划分分类器 和 决策树

Partitioning classifiers 划分分类器

- **Principle**: partition $\mathbb{R}^d$ into disjoint cells $A_1, A_2, \ldots$
  划分 不相交区域
- Classification by a **majority vote** in each cell.
  通过每个区域的 多数投票 来进行分类
- **Classifier**:

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{1}_{[X_i \in A(x)]} \mathbf{1}_{[Y_i=1]} > \sum_{i=1}^n \mathbf{1}_{[X_i \in A(x)]} \mathbf{1}_{[Y_i=0]} \\ 0 & \text{otherwise,} \end{cases}$$

  where $A(x)$ = cell containing x.
- ***X*-property**: the partitions depend **only** on $X_1, \ldots, X_n$ (and **not** on $Y_1, \ldots, Y_n$).
  划分只取决于 X
- **Notation**: $\operatorname{diam}(A) = \sup_{(x,y)\in A^2} \|x - y\|_2$ and $N(x) = \sum_{i=1}^n \mathbf{1}_{[X_i \in A(x)]}$.
  区域 A 中的点的最大距离 区域 A 中的点的数量

Theorem 8.1 — Partitioning classifiers 划分分类器

(划分分类器) 用于判断收敛性的一般定理

Let g_n be a partitioning classifier with the ***X*-property**. If

1. $\operatorname{diam}(A(X)) \to 0$ in probability,

   and

2. $N(X) \to \infty$ in probability,

then $\mathbb{E}\mathscr{R}(g_n) \to \mathscr{R}^*$.

回归函数

Proof. Let $r(x) = \mathbb{E}(Y|X = x)$. From Theorem 7.1, we recall that we need only show that $\mathbb{E}|r_n(X) - r(X)| \to 0$, where

$$r_n(x) = \frac{1}{N(x)} \sum_{i=1}^n \mathbf{1}_{[X_i \in A(x)]} Y_i.$$

已知X在区域A(x)中，r(x)的条件期望

Introduce $\bar{r}(x) = \mathbb{E}(r(X) \mid X \in A(x))$. By the triangle inequality,

$$\mathbb{E}|r_n(X) - r(X)| \leqslant \mathbb{E}|r_n(X) - \bar{r}(X)| + \mathbb{E}|\bar{r}(X) - r(X)|.$$

不太清晰

By conditioning on the random variable $N(x)$, and upon noticing that $\mathbb{P}(Y = 1 | X \in A(x)) = \bar{r}(x)$, it is easy to see that $N(x)r_n(x)$ is distributed as $\text{Bin}(N(x), \bar{r}(x))$, a binomial random variable with parameters $N(x)$ and $\bar{r}(x)$. Thus,

$$\mathbb{E}\Big(|r_n(X) - \bar{r}(X)| \,\Big|\, X, \mathbf{1}_{[X_1 \in A(X)]}, \ldots, \mathbf{1}_{[X_n \in A(X)]}\Big)$$

C-S

由 $|r_n(X) - \bar{r}(X)| \leq 1$ 直接放缩

$$\leqslant \mathbb{E}\left(\left|\frac{\text{Bin}(N(X), \bar{r}(X))}{N(X)} - \bar{r}(X)\right| \mathbf{1}_{[N(X)>0]} \,\Big|\, X, \mathbf{1}_{[X_1 \in A(X)]}, \ldots, \mathbf{1}_{[X_n \in A(X)]}\right) + \mathbf{1}_{[N(X)=0]}$$

$\leq \sqrt{E\left|\frac{Bin(N(x), \bar{r}(X))}{N(x)} - \bar{r}(x)\right|^2}$

二项分布的方差

$$\leqslant \sqrt{\frac{\bar{r}(X)(1 - \bar{r}(X))}{N(X)}} \mathbf{1}_{[N(X)>0]} + \mathbf{1}_{[N(X)=0]},$$

by the Cauchy-Schwarz inequality. Taking expectations, we see that

即 对上述式子再求一次期望

$$\mathbb{E}|r_n(X) - \bar{r}(X)| \leqslant \mathbb{E}\left(\frac{1}{2\sqrt{N(X)}} \mathbf{1}_{[N(X)>0]}\right) + \mathbb{P}(N(X) = 0).$$

Both terms on the right-hand side tend to zero as n tends to infinity by condition 2.

Next, for $\varepsilon > 0$, find a uniformly continuous $[0, 1]$-valued function r_ε with compact support 有紧支集 so that $\mathbb{E}|r(X) - r_\varepsilon(X)| \leqslant \varepsilon$. By the triangle inequality,

并且使得

$$\begin{aligned}\mathbb{E}|\bar{r}(X) - r(X)| &\leqslant \mathbb{E}|\bar{r}(X) - \bar{r}_\varepsilon(X)| + \mathbb{E}|\bar{r}_\varepsilon(X) - r_\varepsilon(X)| + \mathbb{E}|r_\varepsilon(X) - r(X)| \\ &\stackrel{\text{def}}{=} \mathbf{I} + \mathbf{II} + \mathbf{III},\end{aligned}$$

where $\bar{r}_\varepsilon(x) = \mathbb{E}(r_\varepsilon(X) \mid X \in A(x))$. Clearly, $\mathbf{III} \leqslant \varepsilon$ by choice of r_ε. Observe that, for all x,

对于z是常数

$$\mathbf{II} \longrightarrow \left|\frac{1}{\mu(A(x))} \int_{A(x)} r_\varepsilon(z)\mu(dz) - r_\varepsilon(x)\right| \leqslant \frac{1}{\mu(A(x))} \int_{A(x)} |r_\varepsilon(z) - r_\varepsilon(x)|\mu(dz).$$

由 r_ε 的一致连续性 $\leq \varepsilon$

Thus, since r_ε is uniformly continuous, we can find a $\theta = \theta(\varepsilon) > 0$ such that

$$\mathbf{II} \leqslant \varepsilon + \mathbb{P}(\text{diam}(A(X)) > \theta).$$

$\int_{A(\varepsilon)} |r_\varepsilon(z) - r_\varepsilon(x)| \cdot \mathbb{1}\{diam(A(x)) \leq \theta\} \cdot \mu(dz)$

Therefore, $\mathbf{II} \leqslant 2\varepsilon$ for all n large enough, by condition 1. Finally, $+ \int_{A(\varepsilon)} |r_\varepsilon(z) - r_\varepsilon(x)| \cdot \mathbb{1}\{diam(A(x)) > \theta\} \cdot \mu(dz)$

≤ 1

$$\mathbf{I} \leqslant \int_{\mathbb{R}^d} \mathbb{E}\Big(|r(X) - r_\varepsilon(X)| \,\Big|\, X \in A(x)\Big)\mu(dx) = \mathbf{III} \leqslant \varepsilon.$$

Taking together these steps prove the theorem. $\sum_i \int_{A_i} E\big(|r(x) - r_\varepsilon(x)| \,\big|\, X \in A_i\big) \mu(dx)$ ■

$= \sum_i \int_{A_i} \frac{\quad}{\mu(A_i)}$

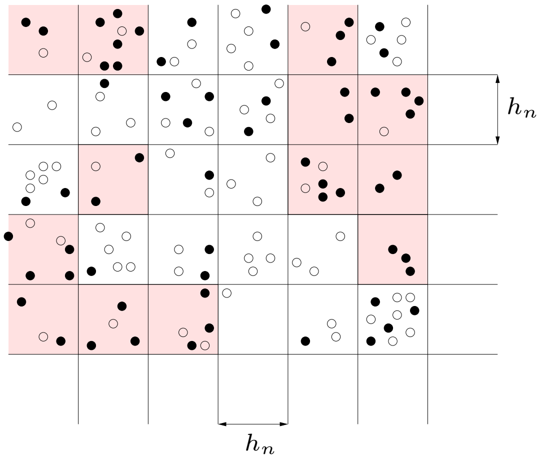
Example 1: cubic histogram classifier 立方直方图分类器

- **Definition**: $A_{n1}, A_{n2}, \ldots$ a partition of $\mathbb{R}^d$ into cubes of size h.
- So, each cell $= \prod_{j=1}^d [k_j h, (k_j + 1)h)$, where the k_j are integers.

Theorem 8.2 — Cubic histogram classifier 立方直方图分类器

Assume that $h \to 0$ and $nh^d \to \infty$. Then the cubic histogram classifier is **universally consistent**, i.e., $\mathbb{E}\mathscr{R}(g_n) \to \mathscr{R}^*$ for **all** distributions of (X, Y).

一致收敛的



Proof. We check the two simple conditions of Theorem 8.1. Clearly, the diameter of each cell is $\sqrt{d}h$. Therefore condition 1 follows trivially. To show condition 2, we need to prove that for any $M < \infty$, $\mathbb{P}(N(X) \leqslant M) \to 0$. Let S be an arbitrary ball centered at the origin. Then the number of cells intersecting S is not more than $c_1 + c_2/h^d$ for some positive constants c_1, c_2. Let μ_n be the empirical measure associated with $X_1, \ldots, X_n$. Then

$$
\begin{aligned}
&\mathbb{P}(N(X) \leqslant M) \\
&\quad \leqslant \sum_{j: A_{nj} \cap S \neq \emptyset} \mathbb{P}(X \in A_{nj}, N(X) \leqslant M) + \mathbb{P}(X \in S^c) \\
&\quad \leqslant \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) \leqslant 2M/n}} \mu(A_{nj}) + \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) > 2M/n}} \mu(A_{nj}) \mathbb{P}(n\mu_n(A_{nj}) \leqslant M) + \mu(S^c) \\
&\quad \leqslant \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) > 2M/n}} \mu(A_{nj}) \mathbb{P}\Big(\mu_n(A_{nj}) - \mu(A_{nj}) \leqslant M/n - \mu(A_{nj})\Big) + \mu(S^c) \\
&\quad \leqslant \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) > 2M/n}} \mu(A_{nj}) \mathbb{P}\Big(\mu_n(A_{nj}) - \mu(A_{nj}) \leqslant -\mu(A_{nj})/2\Big) + \mu(S^c).
\end{aligned}
$$

$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in A_{nj}}\right) = \frac{\mathbb{P}(X \in A_{nj}) \cdot (1 - \mathbb{P}(X \in A_{nj}))}{n}$

Thus, by Chebyshev's inequality,

$$
\mathbb{P}(N(X) \leqslant M) \leqslant \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) > 2M/n}} 4\mu(A_{nj}) \frac{\operatorname{var}(\mu_n(A_{nj}))}{(\mu(A_{nj}))^2} + \mu(S^c).
$$

Therefore,

$$\begin{aligned}\mathbb{P}(N(X) \leqslant M) &\leqslant \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j:A_{nj}\cap S\neq\emptyset\\ \mu(A_{nj})>2M/n}} 4\mu(A_{nj})\frac{1}{n\mu(A_{nj})} + \mu(S^c)\\ &\leqslant \frac{2M+4}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \mu(S^c)\\ &\to \mu(S^c),\end{aligned}$$

because $nh^d \to \infty$. Since S is arbitrary, the proof of the theorem is complete. ■

Example 2: tree classifiers 树分类器

二叉树 • **Binary trees**:

- **Definition**: Recursive binary partitioning of $\mathbb{R}^d$, represented by a tree.
  递归二元分类
- A node has exactly either zero or two **children**.
- A node with zero children is called a **leaf**.
  左分支 和 右分支
- If $u \leftrightarrow A$ and $u_L, u_R \leftrightarrow A_L, A_R$, then $A = A_L \cup A_R$ and $A_L \cap A_R = \emptyset$.
- The **root** $\leftrightarrow \mathbb{R}^d$ and the **leaves** $\leftrightarrow$ a **partition of** $\mathbb{R}^d$.
- We pass from A to A_L and A_R by **answering a question** on x:

  "Is $x^{(j)} \geqslant \alpha$?", for some coordinate j and some α.

- $\mathbb{R}^d$ is partitioned into **hyperrectangles**.
- **Principle**: x is passed into the root and then **iteratively transmitted** to the child nodes. This is repeated until a leaf is reached.

树分类器 • **Tree classifier**: for $x \in A$,

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{1}_{[X_i\in A]}\mathbf{1}_{[Y_i=1]} > \sum_{i=1}^n \mathbf{1}_{[X_i\in A]}\mathbf{1}_{[Y_i=0]}\\ 0 & \text{otherwise.}\end{cases}$$

• **Two questions**:

1. Do we cut?
2. In the affirmative, where do we cut?

- Many tree species (median, centered, CART, etc.).
- **Median tree classifier**: 中位数树分类器
  - **At each node**: find the **median** according to one coordinate.
  - n points $\to$ two children with sizes $\lfloor (n-1)/2 \rfloor$ and $\lceil (n-1)/2 \rceil$.
  - The median itself stays behind and is **not** sent down to the subtrees.
  - Repeat this for k levels of nodes, in a **rotational** manner.
  - 2^k leaf regions, each having **at least** $n/2^k - 2$ and **at most** $n/2^k$ points.

Theorem 8.3 — Median tree classifier 中位数 树分类器

Assume that X has a density. If $k \to \infty$ and $\frac{n}{k 2^k} \to \infty$, then the median tree classifier is **consistent**, i.e., $\mathbb{E}\mathscr{R}(g_n) \to \mathscr{R}^*$. (Note: the conditions on k are fulfilled if $k \leqslant \log_2 n - 2\log_2 \log_2 n$, $k \to \infty$.)

- **Extensions**: label-dependent cuts, CART algorithm, random forests, boosting, etc.

Quantization and Clustering

量化 和 聚类

Basic definitions

- **Quantization** 量化: probabilistic principle to compress information.
- **Context**: a random variable X taking values in $(\mathbb{R}^d, \|\cdot\|_2)$.
- **Assumption** 假设: $\mathbb{E}\|X\|_2^2 < \infty \Leftrightarrow \int_{\mathbb{R}^d} \|x\|_2^2 \mu(dx) < \infty$.
- **Definition**: Let $k \geqslant 1$ be an integer. A **quantizer q of order k** (k 阶量化器 q) is a Borel measurable function $q : \mathbb{R}^d \to \mathscr{C} \subseteq \mathbb{R}^d$, with $|\mathscr{C}| \leqslant k$.
- A quantizer q of order k is characterized by:
  1. A **codebook** $\mathscr{C} = \{c_1, \ldots, c_k\}$.
  2. A **partition** $\mathscr{P} = \{A_1, \ldots, A_k\}$ of $\mathbb{R}^d$, with $q(x) = c_j \Leftrightarrow x \in A_j$.
- **Notation**: $q = (\mathscr{C}, \mathscr{P})$.
- **Definition**: The **distortion** 失真 (for X or μ) of a quantizer $q = (\mathscr{C}, \mathscr{P})$ of order k is

  $$D(\mu, q) = \mathbb{E}\|X - q(X)\|_2^2 = \int_{\mathbb{R}^d} \|x - q(x)\|_2^2 \mu(dx).$$

  The **minimal distortion** at the order k is $D_k^*(\mu) = \inf_q D(\mu, q)$, where the infimum is taken over all quantizers of order k.
- The smaller the distortion, the better the compression.
- The compression quality improves with k.

Lemma 9.1. *One has $D_k^*(\mu) \downarrow 0$ as $k \to \infty$.*

Proof. Clearly, the minimal distortion is a nonincreasing function of the order k. Since $\mathbb{R}^d$ is a Polish space, the bounded measure ν defined for every Borel subset A of $\mathbb{R}^d$ by

$$\nu(A) = \int_A \|x\|_2^2 \mu(dx)$$

is tight, i.e., for all $\varepsilon \in (0, 1]$ there exists a compact K with $\nu(K) \geq 1 - \varepsilon$. Let $\{c_1, c_2, \ldots\}$ be a countable and dense subset of $\mathbb{R}^d$. Since K is compact, one has, for all k large enough,

$$K \subseteq B \stackrel{\text{def}}{=} \bigcup_{j=1}^{k} B(c_j, \sqrt{\varepsilon}).$$

Thus, $\nu(B) \geq 1 - \varepsilon$. Define now q_{k+1} as the quantizer of order $k + 1$ with codebook $\{c_1, \ldots, c_k, 0\}$ (assuming, without loss of generality, that $0 \notin \{c_1, c_2, \ldots\}$) and partition $\{A_1, \ldots, A_k, B^c\}$, with $A_1 = B(c_1, \sqrt{\varepsilon})$ and, for $j \in \{2, \ldots, k\}$, $A_j = B(c_j, \sqrt{\varepsilon}) \setminus A_{j-1}$. Since $\|x - c_j\|_2 \leqslant \sqrt{\varepsilon}$ when $x \in A_j$, we have

$$\begin{aligned}
D^*_{k+1}(\mu) \leqslant D_{k+1}(\mu, q_{k+1}) &= \int_{\mathbb{R}^d} \|x - q_{k+1}(x)\|_2^2 \mu(dx) \\
&= \sum_{j=1}^{k} \int_{A_j} \|x - c_j\|_2^2 \mu(dx) + \int_{B^c} \|x\|_2^2 \mu(dx) \\
&\leqslant \varepsilon \mu\Big(\bigcup_{j=1}^{k} A_j\Big) + \nu(B^c) \leqslant 2\varepsilon,
\end{aligned}$$

which concludes the proof. ∎

Nearest neighbor (NN) quantizers 最小邻近（NN）量化器

- **Context**: quantizers of order k.
- **Voronoi partition**: for $\mathscr{C} = \{c_1, \ldots, c_k\}$, the Voronoi partition $\mathscr{P}_{\mathrm{V}}(\mathscr{C})$ is

$$\begin{aligned}
A_1 &= \Big\{x \in \mathbb{R}^d : \|x - c_1\|_2 \leqslant \|x - c_\ell\|_2, \forall \ell = 1, \ldots, k\Big\}, \text{ and} \\
A_j &= \Big\{x \in \mathbb{R}^d : \|x - c_j\|_2 \leqslant \|x - c_\ell\|_2, \forall \ell = 1, \ldots, k\Big\} \setminus \bigcup_{t=1}^{j-1} A_t,
\end{aligned}$$

  for $2 \leqslant j \leqslant k$ (see Figure 9.1).
- **Definition**: A quantizer of order k is a **NN quantizer** if its partition is the **Voronoi partition** associated with its codebook. Thus, a NN quantizer takes the form $q = (\mathscr{C}, \mathscr{P}_{\mathrm{V}}(\mathscr{C}))$, where $|\mathscr{C}| \leqslant k$.

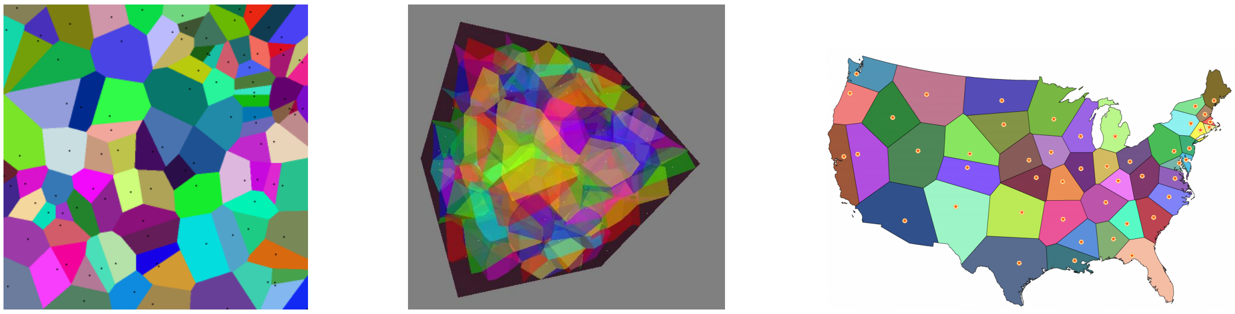


Figure 9.1: A Voronoi partition in dimension $d = 2$ (left), $d = 3$ (middle), and a bonus (right).

- A NN quantizer is entirely characterized by its codebook, via the rule

$$\|x - q(x)\|_2 = \min_{c_j \in \mathscr{C}} \|x - c_j\|_2.$$

- **Vocabulary**: the c_j are the **centers** or the **centroids**.

Properties of NN quantizers 最小邻近量化器的性质

Proposition 9.1. *Let q_{NN} be a NN quantizer with codebook $\mathscr{C} = \{c_1, \dots, c_k\}$. Then*

$$D(\mu, q_{\mathrm{NN}}) = \mathbb{E} \min_{1 \leqslant j \leqslant k} \|X - c_j\|_2^2 = \int_{\mathbb{R}^d} \min_{1 \leqslant j \leqslant k} \|x - c_j\|_2^2 \mu(dx).$$

*In addition, for **any** quantizer $q = (\mathscr{C}, \mathscr{P})$, $D(\mu, q_{\mathrm{NN}}) \leqslant D(\mu, q)$.*

Proof. Let $\mathscr{P}_{\mathrm{V}}(\mathscr{C}) = \{A_{\mathrm{V},1}, \dots, A_{\mathrm{V},k}\}$ be the Voronoi partition associated with $\mathscr{C}$. Then

$$\begin{aligned} D(\mu, q_{\mathrm{NN}}) &= \int_{\mathbb{R}^d} \|x - q_{\mathrm{NN}}(x)\|_2^2 \mu(dx) = \sum_{j=1}^{k} \int_{A_{\mathrm{V},j}} \|x - c_j\|_2^2 \mu(dx) \\ &= \int_{\mathbb{R}^d} \min_{1 \leqslant j \leqslant k} \|x - c_j\|_2^2 \mu(dx). \end{aligned}$$

This shows the first statement. Next, for $\mathscr{P} = \{A_1, \dots, A_k\}$,

$$\begin{aligned} D(\mu, q_{\mathrm{NN}}) &= \sum_{j=1}^{k} \int_{A_j} \min_{1 \leqslant j \leqslant k} \|x - c_j\|_2^2 \mu(dx) \\ &\leqslant \sum_{j=1}^{k} \int_{A_j} \|x - c_j\|_2^2 \mu(dx) \\ &= \int_{\mathbb{R}^d} \|x - q(x)\|_2^2 \mu(dx) = D(\mu, q), \end{aligned}$$

by definition of the distortion. ∎

- **Conclusion**: if quantizers with **minimal distortion** exist, they are **NN quantizers**.
- **Notation**: $q_{\mathrm{NN}} = (\mathbf{c}, \mathscr{P}_{\mathrm{V}}(\mathbf{c}))$, with $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$ and distortion
$$W(\mu, \mathbf{c}) \stackrel{\text{def}}{=} D(\mu, q_{\mathrm{NN}}).$$

Theorem 9.1 — Optimal quantizer 最优量化器

There exists a quantizer with minimal distortion.

Sketch of proof. According to Proposition 9.1, we have to prove that there exists $\mathbf{c}^* \in \mathbb{R}^{dk}$ such that
$$W(\mu, \mathbf{c}^*) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, \mathbf{c}).$$
One first shows (omitted) that there exists an $R > 0$ such that
$$\inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, \mathbf{c}) = \inf_{\|\mathbf{c}\|_2 \leqslant R} W(\mu, \mathbf{c}).$$
Then we prove that the function $\mathbb{R}^{dk} \ni \mathbf{c} \mapsto W(\mu, \mathbf{c})$ is continuous. To this aim, observe that the function $x \mapsto \min_{1 \leqslant j \leqslant k} \|x - c_j\|_2$ is continuous. Therefore, for $\mathbf{c}_0 = (c_{0,1}, \ldots, c_{0,k}) \in \mathbb{R}^{dk}$, one has
$$\begin{aligned}
\lim_{\mathbf{c} \to \mathbf{c}_0} W(\mu, \mathbf{c}) &= \int_{\mathbb{R}^d} \lim_{\mathbf{c} \to \mathbf{c}_0} \min_{1 \leqslant j \leqslant k} \|x - c_j\|_2^2 \mu(dx) \\
&\quad \text{(by the Lebesgue dominated convergence theorem)} \\
&= \int_{\mathbb{R}^d} \min_{1 \leqslant j \leqslant k} \|x - c_{0,j}\|_2^2 \mu(dx) \\
&\quad \text{(by continuity)} \\
&= W(\mu, \mathbf{c}_0),
\end{aligned}$$
which shows that $W(\mu, \cdot)$ is continuous.

It follows from the continuity of $W(\mu, \cdot)$ and the compactness of the ball $B(0, R)$ of $\mathbb{R}^{dk}$ that the infimum of $W(\mu, \cdot)$ is achieved at some $\mathbf{c}^* \in \mathbb{R}^{dk}$. But then the quantizer $q^* = (\mathbf{c}^*, \mathscr{P}_{\mathrm{V}}(\mathbf{c}^*))$ has minimal distortion since
$$W(\mu, \mathbf{c}^*) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, c) = \inf_{q} D(\mu, q) = D_k^*(\mu).$$
∎

Empirical quantization 经验量化器

- In practice, the distribution of X is **unknown**.
- **Sample**: $X_1, \ldots, X_n$ i.i.d., distributed as (and independent of) X.
- **Objective**: construct a "good" $q_n(\cdot) = q_n(\cdot\,; X_1, \ldots, X_n)$.
- The **distortion** of q_n is naturally defined by

$$D(\mu, q_n) = \mathbb{E}\big(\|X - q_n(X)\|_2^2 \mid X_1, \ldots, X_n\big) = \int_{\mathbb{R}^d} \|x - q_n(x)\|_2^2 \mu(dx).$$

⚠ It is a random quantity.

- **Empirical distortion**:
经验失真

$$D(\mu_n, q) = \int_{\mathbb{R}^d} \|x - q(x)\|_2^2 \mu_n(dx) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - q(X_i)\|_2^2,$$

where $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ is the empirical measure.

- For $q_{\text{NN}} = (\mathbf{c}, \mathscr{P}_{\text{V}}(\mathbf{c}))$, with $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$,

$$D(\mu_n, q_{\text{NN}}) = W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leqslant j \leqslant k} \|X_i - c_j\|_2^2.$$

- A quantizer is **consistent** if

$$\mathbb{E} D(\mu, q_n) \to D_k^*(\mu) \quad \text{as } n \to \infty.$$

- Natural choice: q_n^* that **minimizes** the empirical distortion over all NN quantizers.
- **Definition**: $\mathbf{c}_n^* = (c_{n,1}^*, \ldots, c_{n,k}^*)$ such that

$$W(\mu_n, \mathbf{c}_n^*) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu_n, \mathbf{c}).$$

So,

$$q_n^* = (\mathbf{c}_n^*, \mathscr{P}_{\text{V}}(\mathbf{c}_n^*)).$$

Quantization and clustering 量化和聚类

- q_n^* allows a **clustering** of $X_1, \ldots, X_n$ into k groups.
- **Principle**: X_i is affected to group j if $q_n^*(X_i) = j$.
- **Cluster** $\sharp j$ = the X_i such that $\|X_i - c_{n,j}^*\|_2 \leqslant \|X_i - c_{n,\ell}^*\|_2$, $\forall \ell = 1, \ldots, k$.
- Computation of q_n^* is often a NP hard problem $\to$ ***k*-means algorithm**.
- **Basic idea**: for $\mathscr{C} = \{c_1, \ldots, c_k\}$ and $\mathscr{P} = \{A_1, \ldots, A_k\}$, let $q = (\mathscr{C}, \mathscr{P})$ and $q_n = (\mathscr{C}_n, \mathscr{P})$, with $\mathscr{C}_n = \{c_{n,1}, \ldots, c_{n,k}\}$ such that

$$c_{n,j} = \arg\min_{y \in \mathbb{R}^d} \sum_{i=1}^n \|X_i - y\|_2^2 \mathbf{1}_{[X_i \in A_j]} = \frac{\sum_{i=1}^n X_i \mathbf{1}_{[X_i \in A_j]}}{\sum_{i=1}^n \mathbf{1}_{[X_i \in A_j]}}, \quad 1 \leqslant j \leqslant k.$$

Then

$$\begin{aligned} D(\mu_n, q) &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \|X_i - c_j\|_2^2 \mathbf{1}_{[X_i \in A_j]} \\ &\geq \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^n \|X_i - c_{n,j}\|_2^2 \mathbf{1}_{[X_i \in A_j]} \\ &= D(\mu_n, q_n). \end{aligned}$$

k-means algorithm

1. `Initialization of the algorithm`: $\mathscr{C}^{(1)} = \{c_1^{(1)}, \ldots, c_k^{(1)}\}$ and $\mathscr{P}_V^{(1)} = \{A_1^{(1)}, \ldots, A_k^{(1)}\}$.
2. `Lloyd's iteration`: compute $\mathscr{C}^{(\ell+1)} = \{c_1^{(\ell+1)}, \ldots, c_k^{(\ell+1)}\}$ from $\mathscr{C}^{(\ell)} = \{c_1^{(\ell)}, \ldots, c_k^{(\ell)}\}$ via the iteration

$$c_j^{(\ell+1)} = \frac{\sum_{i=1}^n X_i \mathbf{1}_{[X_i \in A_j^{(\ell)}]}}{\sum_{i=1}^n \mathbf{1}_{[X_i \in A_j^{(\ell)}]}}, \qquad 1 \leqslant j \leqslant k,$$

   where $\{A_1^{(\ell)}, \ldots, A_k^{(\ell)}\}$ is the Voronoi partition associated with $\mathscr{C}^{(\ell)}$.
3. The algorithm **stops** after a finite number of iterations.

⚠ The output codebook is **not** $\mathbf{c}_n^*$.

Consistency of q_n^*

- **Reminder**: $\mathbf{c}_n^* = (c_{n,1}^*, \ldots, c_{n,k}^*)$ such that
$$W(\mu_n, \mathbf{c}_n^*) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu_n, \mathbf{c}).$$
So,
$$q_n^* = (\mathbf{c}_n^*, \mathscr{P}_{\mathrm{V}}(\mathbf{c}_n^*)).$$

- **Definition**: Let ν_1 and ν_2 be probability measures on $\mathbb{R}^d$ with finite second moment. The **Wasserstein distance ρ_W** between ν_1 and ν_2 is
$$\rho_W(\nu_1, \nu_2) = \inf_{X \overset{\mathscr{D}}{=} \nu_1, Y \overset{\mathscr{D}}{=} \nu_2} \sqrt{\mathbb{E}\|X - Y\|_2^2}.$$

- **Property 1**: There exists (X_0, Y_0) such that $X_0 \overset{\mathscr{D}}{=} \nu_1$, $Y_0 \overset{\mathscr{D}}{=} \nu_2$, and
$$\rho_W(\nu_1, \nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|_2^2}.$$

- **Property 2**: One has $\rho_W(\nu_n, \nu) \to 0$ if and only if
$$\nu_n \Rightarrow \nu \quad \text{and} \quad \int_{\mathbb{R}^d} \|x\|_2^2 \nu_n(dx) \to \int_{\mathbb{R}^d} \|x\|_2^2 \nu(dx).$$

Proposition 9.2. *Let ν_1 and ν_2 be probability measures on $\mathbb{R}^d$ with finite second moment. If q is a NN quantizer, then*
$$\left| D(\nu_1, q)^{1/2} - D(\nu_2, q)^{1/2} \right| \leqslant \rho_W(\nu_1, \nu_2).$$

Proof. Let (X_0, Y_0) be such that $X_0 \overset{\mathscr{D}}{=} \nu_1$, $Y_0 \overset{\mathscr{D}}{=} \nu_2$, and
$$\rho_W(\nu_1, \nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|_2^2}.$$
For $q = (\mathbf{c}, \mathscr{P}_{\mathrm{V}}(\mathbf{c}))$, one has
$$\begin{aligned}
D(\nu_1, q)^{1/2} = W(\nu_1, \mathbf{c})^{1/2} &= \sqrt{\mathbb{E} \min_{1 \leqslant j \leqslant k} \|X_0 - c_j\|_2^2} \\
&= \sqrt{\mathbb{E}(\min_{1 \leqslant j \leqslant k} \|X_0 - c_j\|_2)^2} \\
&\leqslant \sqrt{\mathbb{E}\Big(\min_{1 \leqslant j \leqslant k} (\|X_0 - Y_0\|_2 + \|Y_0 - c_j\|_2)\Big)^2} \\
&= \sqrt{\mathbb{E}\Big(\|X_0 - Y_0\|_2 + \min_{1 \leqslant j \leqslant k} \|Y_0 - c_j\|_2\Big)^2} \\
&\leqslant \sqrt{\mathbb{E}\|X_0 - Y_0\|_2^2} + \sqrt{\mathbb{E} \min_{1 \leqslant j \leqslant k} \|Y_0 - c_j\|_2^2} \\
&\quad \text{(by the Cauchy-Schwarz inequality)} \\
&= \rho_W(\nu_1, \nu_2) + D(\nu_2, q)^{1/2}.
\end{aligned}$$

One shows with similar arguments that $D(\nu_2, q)^{1/2} \leqslant \rho_W(\nu_1, \nu_2) + D(\nu_1, q)^{1/2}$, and the result follows. ■

Theorem 9.2 — Consistency of q_n^*

One has $D(\mu, q_n^*) \to D_k^*(\mu)$ wp 1, and $\mathbb{E}D(\mu, q_n^*) \to D_k^*(\mu)$.

Proof. Since the context is clear, we write $\|\cdot\|$ instead of $\|\cdot\|_2$ throughout the proof. If q^* is a NN quantizer optimal for μ, then, by Proposition 9.2,

$$\begin{aligned}
0 &\leqslant D(\mu, q_n^*)^{1/2} - D_k^*(\mu)^{1/2} \\
&= \Big[D(\mu, q_n^*)^{1/2} - D(\mu_n, q_n^*)^{1/2}\Big] + \Big[D(\mu_n, q_n^*)^{1/2} - D(\mu, q^*)^{1/2}\Big] \\
&\leqslant \Big[D(\mu, q_n^*)^{1/2} - D(\mu_n, q_n^*)^{1/2}\Big] + \Big[D(\mu_n, q^*)^{1/2} - D(\mu, q^*)^{1/2}\Big] \\
&\leqslant 2\,\rho_W(\mu, \mu_n).
\end{aligned} \tag{9.1}$$

But $\rho_W(\mu_n, \mu) \to 0$ wp 1, since $\mathbb{P}(\mu_n \Rightarrow \mu) = 1$ (by Varadarajan's theorem) and, wp 1,

$$\int_{\mathbb{R}^d} \|x\|^2 \mu_n(dx) \to \int_{\mathbb{R}^d} \|x\|^2 \mu(dx)$$

(by the strong law of large numbers). We conclude that $D(\mu, q_n^*) \to D_k^*(\mu)$ wp 1.

To prove the second assertion, we introduce $\mathscr{M}(\mu, \mu_n)$, the (random) set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and μ_n, respectively. By definition,

$$\rho_W^2(\mu, \mu_n) = \inf_{\nu \in \mathscr{M}(\mu, \mu_n)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \nu(dx, dy).$$

Let $C > 0$ be an arbitrary constant, and let $\mathscr{A}$ be the subset of $\mathbb{R}^d \times \mathbb{R}^d$ defined by

$$\mathscr{A} = \Big\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \max(\|x\|, \|y\|) \leqslant C\Big\}.$$

One has, for all $\nu \in \mathscr{M}(\mu, \mu_n)$,

$$\begin{aligned}
&\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \nu(dx, dy) \\
&\quad = \int_{\mathscr{A}} \|x - y\|^2 \nu(dx, dy) + \int_{\mathscr{A}^c} \|x - y\|^2 \nu(dx, dy) \\
&\quad \leqslant \int_{\mathscr{A}} \|x - y\|^2 \nu(dx, dy) + 2\int_{\mathscr{A}^c} \|x\|^2 \nu(dx, dy) + 2\int_{\mathscr{A}^c} \|y\|^2 \nu(dx, dy) \\
&\qquad \text{(since } \|x - y\|^2 \leqslant 2\|x\|^2 + 2\|y\|^2\text{)} \\
&\quad \leqslant \int_{\mathscr{A}} \|x - y\|^2 \nu(dx, dy) + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbf{1}_{[\|x\| > C]} \mu(dx) + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbf{1}_{[\|x\| \leqslant C, \|y\| > C]} \nu(dx, dy) \\
&\qquad + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbf{1}_{[\|y\| > C]} \mu_n(dy) + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbf{1}_{[\|x\| > C, \|y\| \leqslant C]} \nu(dx, dy) \\
&\quad \leqslant \int_{\mathscr{A}} \|x - y\|^2 \nu(dx, dy) + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbf{1}_{[\|x\| > C]} \mu(dx) + 2C^2 \mu_n(\|y\| > C) \\
&\qquad + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbf{1}_{[\|y\| > C]} \mu_n(dy) + 2C^2 \mu(\|x\| > C).
\end{aligned}$$

Therefore, by Markov's inequality,

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^2\nu(dx,dy) \leqslant \int_{\mathscr{A}}\|x-y\|^2\nu(dx,dy)&\\ +2\int_{\mathbb{R}^d}\|x\|^2\mathbf{1}_{[\|x\|>C]}\mu(dx)+2\int_{\mathbb{R}^d}\|y\|^2\mathbf{1}_{[\|y\|>C]}\mu_n(dy)&\\ +2\int_{\mathbb{R}^d}\|y\|^2\mathbf{1}_{[\|y\|>C]}\mu_n(dy)+2\int_{\mathbb{R}^d}\|x\|^2\mathbf{1}_{[\|x\|>C]}\mu(dx).&\end{aligned}$$

Taking the infimum over $\mathscr{M}(\mu,\mu_n)$ on the right-hand side, and then expectation on both sides, we conclude that

$$\mathbb{E}\rho_W^2(\mu,\mu_n)\leqslant \mathbb{E}\inf_{\nu\in\mathscr{M}(\mu,\mu_n)}\int_{\mathscr{A}}\|x-y\|^2\nu(dx,dy)+8\int_{\mathbb{R}^d}\|x\|^2\mathbf{1}_{[\|x\|>C]}\mu(dx).$$

For fixed $C>0$, the first term on the right-hand side tends to zero as n tends to infinity by the first statement and the Lebesgue dominated convergence theorem. Since $\int_{\mathbb{R}^d}\|x\|^2\mu(dx)<\infty$, the second term can be made arbitrarily small by taking C sufficiently large. Putting all the pieces together, we see that $\mathbb{E}\rho_W^2(\mu,\mu_n)$ tends to zero, and the result easily follows from inequality (9.1). ■

Theorem 9.3 — Rate of convergence

If $\|X\|_2\leqslant R$ wp 1, then

$$\mathbb{E}D(\mu,q_n^*)-D_k^*(\mu)\leq\frac{12kR^2}{\sqrt{n}}.$$

- $\|X\|_2\leqslant R$ is called the **peak power constraint**.
- **Take-home message**: the rate of convergence is independent of d.

Proof. Let us start with some preliminary remarks.

1. Let $\sigma_1,\dots,\sigma_n$ be i.i.d. Rademacher random variables, independent of $X_1,\dots,X_n$, and let $\mathscr{F}$ be a collection of real-valued functions on $\mathbb{R}^d$. Then, by the contraction principle,

$$\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i|f(X_i)|\leqslant\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(X_i).$$

2. If $\|X\|_2\leqslant R$ wp 1, then the optimal codevectors are in $B_R\stackrel{\text{def}}{=}B(0,R)$. To see this, just note that if $\|c\|_2>R$ and p is the projection onto B_R, then, for all $x\in B_R$,

$$\begin{aligned}\|x-c\|_2^2&=\|x-p(c)\|_2^2+\|p(c)-c\|_2^2-2\langle x-p(c),c-p(c)\rangle\\&\geq\|x-p(c)\|_2^2.\end{aligned}$$

Thus, the distortion is smaller for codevectors in B_R.

3. If $X \stackrel{\mathscr{D}}{=} \mu$, then

$$W(\mu, \mathbf{c}) = \mathbb{E} \min_{1 \leqslant j \leqslant k} \|X - c_j\|_2^2 = \mathbb{E}\|X\|_2^2 + \mathbb{E} \min_{1 \leqslant j \leqslant k} \Big(-2\langle X, c_j\rangle + \|c_j\|_2^2\Big).$$

The last two remarks show that minimizing $W(\mu, \cdot)$ over $\mathbb{R}^{dk}$ is identical to minimizing $\bar{W}(\mu, \cdot)$ over B_R^k, where

$$\bar{W}(\mu, \mathbf{c}) = \mathbb{E} \min_{1 \leqslant j \leqslant k} f_{c_j}(X), \quad f_c(x) = -2\langle x, c\rangle + \|c\|_2^2.$$

The same principle holds with μ_n in place of μ.

We are now ready to prove the theorem. Observe that

$$\begin{aligned} D(\mu, q_n^*) - D_k^*(\mu) &= W(\mu, \mathbf{c}_n^*) - \inf_{\mathbf{c} \in B_R^k} W(\mu, \mathbf{c}) \\ &= \bar{W}(\mu, \mathbf{c}_n^*) - \inf_{\mathbf{c} \in B_R^k} \bar{W}(\mu, \mathbf{c}) \\ &= \Big[\bar{W}(\mu, \mathbf{c}_n^*) - \bar{W}(\mu_n, \mathbf{c}_n^*)\Big] + \Big[\inf_{\mathbf{c} \in B_R^k} \bar{W}(\mu_n, \mathbf{c}) - \inf_{\mathbf{c} \in B_R^k} \bar{W}(\mu, \mathbf{c})\Big] \\ &\leqslant \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu, \mathbf{c}) - \bar{W}(\mu_n, \mathbf{c})) + \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})). \end{aligned}$$

We are thus interested in upper bounds for the maximal deviation

$$\mathbb{E} \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})),$$

and note that the other term can be similarly bounded. Let $X_1', \dots, X_n'$ be a ghost sample, independent of $X_1, \dots, X_n$ and $\sigma_1, \dots, \sigma_n$. Then

$$\begin{aligned} &\mathbb{E} \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})) \\ &\quad = \mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \sum_{i=1}^n \Big(\min_{1 \leqslant j \leqslant k} f_{c_j}(X_i) - \mathbb{E} \min_{1 \leqslant j \leqslant k} f_{c_j}(X)\Big) \\ &\quad = \mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \mathbb{E}\Big(\sum_{i=1}^n \Big(\min_{1 \leqslant j \leqslant k} f_{c_j}(X_i) - \min_{1 \leqslant j \leqslant k} f_{c_j}(X_i')\Big) \mid X_1, \dots, X_n\Big). \end{aligned}$$

Thus, upon noting that $\sup \mathbb{E}(\cdot) \leqslant \mathbb{E} \sup(\cdot)$,

$$\begin{aligned} \mathbb{E} \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})) &\leqslant \mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \sum_{i=1}^n \Big(\min_{1 \leqslant j \leqslant k} f_{c_j}(X_i) - \min_{1 \leqslant j \leqslant k} f_{c_j}(X_i')\Big) \\ &\leqslant 2\mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \sum_{i=1}^n \sigma_i \min_{1 \leqslant j \leqslant k} f_{c_j}(X_i). \end{aligned}$$

The proof proceeds now by induction on k, using the contraction principle. Let

$$S_k = \mathbb{E} \sup_{(c_1, \dots, c_k) \in B_R^k} \frac{1}{n} \sum_{i=1}^n \sigma_i \min_{1 \leqslant j \leqslant k} f_{c_j}(X_i).$$

Case $k = 1$. Since $\|X\|_2 \leqslant R$,

$$\begin{aligned}
S_1 &= \mathbb{E} \sup_{c \in B_R} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Big(-2\langle X_i, c\rangle + \|c\|_2^2 \Big) \\
&\leqslant 2\mathbb{E} \sup_{c \in B_R} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle X_i, c\rangle + \mathbb{E} \sup_{c \in B_R} \frac{\|c\|_2^2}{n} \sum_{i=1}^{n} \sigma_i \\
&\leqslant 2\mathbb{E} \sup_{c \in B_R} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle X_i, c\rangle + \frac{R^2}{n} \mathbb{E} \Big| \sum_{i=1}^{n} \sigma_i \Big|.
\end{aligned}$$

Thus, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
S_1 &\leqslant 2\mathbb{E} \sup_{c \in B_R} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle X_i, c\rangle + \frac{R^2}{\sqrt{n}} \\
&= 2\mathbb{E} \sup_{c \in B_R} \frac{1}{n} \Big\langle \sum_{i=1}^{n} \sigma_i X_i, c \Big\rangle + \frac{R^2}{\sqrt{n}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
S_1 &\leqslant \frac{2R}{n} \mathbb{E} \Big\| \sum_{i=1}^{n} \sigma_i X_i \Big\|_2 + \frac{R^2}{\sqrt{n}} \\
&\leqslant 2R \sqrt{\frac{\mathbb{E}\|X\|_2^2}{n}} + \frac{R^2}{\sqrt{n}} \\
&\quad \text{(by the Cauchy-Schwarz inequality)} \\
&\leqslant \frac{3R^2}{\sqrt{n}}.
\end{aligned}$$

Case $k = 2$. Using the equality $\min(a, b) = \frac{a+b}{2} - \frac{|a-b|}{2}$, we may write

$$\begin{aligned}
S_2 &= \mathbb{E} \sup_{(c_1, c_2) \in B_R^2} \frac{1}{2n} \sum_{i=1}^{n} \sigma_i \Big(f_{c_1}(X_i) + f_{c_2}(X_i) - |f_{c_1}(X_i) - f_{c_2}(X_i)| \Big) \\
&\leqslant S_1 + \mathbb{E} \sup_{(c_1, c_2) \in B_R^2} \frac{1}{2n} \sum_{i=1}^{n} \sigma_i |f_{c_1}(X_i) - f_{c_2}(X_i)|.
\end{aligned}$$

Applying the contraction principle, we obtain

$$S_2 \leqslant S_1 + \mathbb{E} \sup_{(c_1, c_2) \in B_R^2} \frac{1}{2n} \sum_{i=1}^{n} \sigma_i (f_{c_1}(X_i) - f_{c_2}(X_i)) \leqslant 2S_1.$$

Case $k = 3$. Since $S_2 \leqslant 2S_1$,

$$S_3 \leqslant \frac{S_1 + S_2}{2} + \frac{S_1 + S_2}{2} \leqslant 3S_1.$$

Repeating this process, we find

$$S_k \leqslant kS_1 \leqslant \frac{3kR^2}{\sqrt{n}}.$$

Finally,

$$\mathbb{E}D(\mu, q_n^*) - D_k^*(\mu) \leqslant 4S_k \leqslant \frac{12kR^2}{\sqrt{n}},$$

and the proof is complete. ■

Problem 1

Exercise 1

Let (X, Y) be a random pair taking values in $\mathbb{R} \times \{0, 1\}$, where X is uniformly distributed on $[-2, 2]$. We assume that

$$Y = \begin{cases} \mathbf{1}_{[U \leqslant 2]} & \text{if } X \leqslant 0 \\ \mathbf{1}_{[U > 1]} & \text{if } X > 0, \end{cases}$$

where U is a random variable uniformly distributed on $[0, 10]$, independent of X. Compute the Bayes rule and the Bayes risk associated with (X, Y).

Exercise 2

Let (X, Y) be a random pair taking values in $\mathbb{R}_+ \times \{-1, 1\}$. We let $\eta(x) = \mathbb{P}(Y = 1|X = x)$ and assume that $\eta(x) = x/(c + x)$, where c is a positive constant.

1. Show that the Bayes risk $\mathscr{R}^*$ associated with (X, Y) is

$$\mathscr{R}^* = \mathbb{E}\Big(\frac{\min(c, X)}{c + X}\Big).$$

2. Provide an expression of $\mathscr{R}^*$ when X is uniformly distributed on $[0, \alpha c]$, where $\alpha \geqslant 1$.

Exercise 3

Let (X, Y) be a random pair taking values in $\mathbb{R}^3 \times \{0, 1\}$. The three components of X are denoted by T, B, and E, respectively. The variable T represents the average number of hours per week that a student spends watching TV, and the variable B the average number of hours per week he/she spends in bars. The component E is an abstract quantity measuring extra negative factors such as laziness and learning difficulties. Unfortunately, E is intangible, and not available to the observer.

Finally, the random variable Y simply models the student's results: $Y = 1$ or $Y = 0$ according to whether he/she fails or passes a course. It is assumed that

$$Y = \begin{cases} 1 & \text{if } T + B + E < 7 \\ 0 & \text{otherwise.} \end{cases}$$

It is also assumed that T, B, and E are independent with an exponential distribution (with parameter 1). The Bayes rule associated with $((T, B), Y)$ is denoted by $g^*(T, B)$.

1. What is $\mathscr{R}^*$, the Bayes risk associated with $((T, B, E), Y)$?
2. Give the expression of $\mathbb{P}(Y = 1|T, B)$.
3. Deduce from the above $g^*(T, B)$.
4. What is the probability density of the random variable $T + B$?
5. Provide the numerical expression of $\mathbb{P}(g^*(T, B) \neq Y)$.
6. What is the error incurred by a student who decides that $Y = 1$, independently of T and B?

PROBLEM 2

Rademacher均值

A. Rademacher averages. Given a set $A \subseteq \mathbb{R}^n$ of vectors $a = (a_1, \ldots, a_n)$, the Rademacher complexity of A is defined by

$$\mathbf{R}_n(A) = \mathbb{E} \sup_{a \in A} \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i,$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables.

1. Prove that if $A = \{a^{(1)}, \ldots, a^{(N)}\} \subseteq \mathbb{R}^n$ is a **finite** set, then

$$\mathbf{R}_n(A) \leqslant \max_{1 \leqslant j \leqslant N} \|a^{(j)}\|_2 \frac{\sqrt{2 \log N}}{n}.$$

Solution. The result is clear if $\max_{1 \leqslant j \leqslant N} \|a^{(j)}\|_2 = 0$ or $N = 1$. Thus, in the sequel, we assume that $\max_{1 \leqslant j \leqslant N} \|a^{(j)}\|_2 > 0$ and $N > 1$. Observe that, for all $s > 0$, by independence, for $a = (a_1, \ldots, a_n) \in A$,

$$\begin{aligned}
\mathbb{E} \exp\left(\frac{s}{n} \sum_{i=1}^{n} \sigma_i a_i\right) = \prod_{i=1}^{n} \mathbb{E} \exp\left(\frac{s}{n} \sigma_i a_i\right) &\leqslant \prod_{i=1}^{n} \exp\left(\frac{s^2 a_i^2}{2n^2}\right) \\
&\quad \text{(by Lemma 2.1)} \\
&= \exp\left(\frac{s^2 \|a\|_2^2}{2n^2}\right) \\
&\leqslant \exp\left(\frac{s^2 \max_{1 \leqslant j \leqslant N} \|a^{(j)}\|_2^2}{2n^2}\right).
\end{aligned}$$

Therefore, using Lemma 2.2 with $\alpha = \max_{1 \leqslant j \leqslant N} \|a^{(j)}\|_2 / n$, we conclude that

$$\mathbf{R}_n(A) = \mathbb{E} \max_{1 \leqslant j \leqslant N} \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i^{(j)} \leqslant \max_{1 \leqslant j \leqslant N} \|a^{(j)}\|_2 \frac{\sqrt{2 \log N}}{n}.$$

■

碎裂系数 和 VC维度

B. Shatter coefficients and VC dimension. For $X_1, \ldots, X_n$ i.i.d. random variables taking values in a set $\mathscr{X}$ and a class of indicators $\mathscr{F} = \{f = \mathbf{1}_A, A \in \mathscr{A}\}$, with $|\mathscr{A}| \geqslant 2$, we let

$$\mathbf{R}_n(\mathscr{F}) = \mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) = \mathbb{E} \sup_{A \in \mathscr{A}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbf{1}_{[X_i \in A]},$$

where $\sigma_1, \ldots, \sigma_n$ are independent of the X_i. The n-th **shatter coefficient** of $\mathscr{A}$ is defined by

$$\mathbf{S}_{\mathscr{A}}(n) = \max_{x_1^n} |\mathscr{F}(x_1^n)|,$$

where $x_1^n = (x_1, \ldots, x_n)$ and

$$\mathscr{F}(x_1^n) = \big\{(f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n, f \in \mathscr{F}\big\}$$

(note that it is a finite subset of $\mathbb{R}^n$, why?).

1. Show that $\mathbf{S}_{\mathscr{A}}(1) = 2$, $2 \leqslant \mathbf{S}_{\mathscr{A}}(n) \leqslant 2^n$, and

$$\mathbf{S}_{\mathscr{A}}(k) < 2^k \text{ for some } k > 1 \Leftrightarrow \mathbf{S}_{\mathscr{A}}(n) < 2^n \text{ for all } n \geqslant k.$$

2. Prove that

$$\mathbf{R}_n(\mathscr{F}) \leqslant \sqrt{\frac{2\log(\mathbf{S}_{\mathscr{A}}(n))}{n}}.$$

Definition: The **VC dimension** $V_{\mathscr{A}}$ of $\mathscr{A}$ is the **largest integer** $n_0 \geqslant 1$ for which $\mathbf{S}_{\mathscr{A}}(n_0) = 2^{n_0}$. If $\mathbf{S}_{\mathscr{A}}(n) = 2^n$ for all $n \geqslant 1$, then $V_{\mathscr{A}} = \infty$.

3. Show that if $|\mathscr{A}| < \infty$, then $\mathbf{S}_{\mathscr{A}}(n) \leqslant |\mathscr{A}|$ and $V_{\mathscr{A}} \leqslant \log_2 |\mathscr{A}|$.

4. Prove that if $\mathscr{A} = \{(-\infty, a], a \in \mathbb{R}\}$, then $V_{\mathscr{A}} = 1$. Similarly, if $\mathscr{A} = \{[a, b], (a, b) \in \mathbb{R}^2\}$, then $V_{\mathscr{A}} = 2$.

5. What is $V_{\mathscr{A}}$ for $\mathscr{A} = \{$all convex polygons of $\mathbb{R}^2\}$?

Two important results:

Theorem 11.1 — VC dimension of affine spaces

Let $\mathscr{G}$ be a finite-dimensional vector space of functions $\mathbb{R}^p \to \mathbb{R}$, and let

$$\mathscr{A} = \big\{\{x \in \mathbb{R}^p, g(x) \geqslant 0\}, g \in \mathscr{G}\big\}.$$

Then $V_{\mathscr{A}} \leqslant \dim \mathscr{G}$. Consequence: if $\mathscr{A}$ = subsets of $\mathbb{R}^p$ of the form $\{x \in \mathbb{R}^p : a^\top x + b \geqslant 0, a \in \mathbb{R}^p, b \in \mathbb{R}\}$, then $V_{\mathscr{A}} \leqslant p + 1$.

Theorem 11.2 — Sauer

If $V_{\mathscr{A}} < \infty$, then, for all $n \geqslant 1$, $\mathbf{S}_{\mathscr{A}}(n) \leqslant \sum_{i=1}^{V_{\mathscr{A}}} \binom{n}{i}$.

6. Exploit Sauer's inequality to prove that $\mathbf{S}_{\mathscr{A}}(n) \leqslant (n+1)^{V_{\mathscr{A}}}$. Conclude that:

   – Either $V_{\mathscr{A}} = \infty \to \mathbf{S}_{\mathscr{A}}(n) = 2^n$ for all $n \geqslant 1$.

   – Either $V_{\mathscr{A}} < \infty \to \mathbf{S}_{\mathscr{A}}(n) \leqslant (n+1)^{V_{\mathscr{A}}}$ for all $n \geqslant 1$.

   In particular, it is impossible to have $\mathbf{S}_{\mathscr{A}}(n) \sim 2^{\sqrt{n}}$, for example.

7. Establish that

$$\mathbf{R}_n(\mathscr{F}) \leqslant \sqrt{\frac{2V_{\mathscr{A}} \log(n+1)}{n}}.$$

8. **Bonus**: show that, for all distributions,

$$\begin{aligned}\mathbb{E} \sup_{A \in \mathscr{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[X_i \in A]} - \mathbb{P}(X_1 \in A) \right| &\leqslant 2\mathbb{E} \sup_{A \in \mathscr{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbf{1}_{[X_i \in A]} \right| \\ &\leqslant 4\sqrt{\frac{V_{\mathscr{A}} \log(n+1)}{n}}\end{aligned}$$

   (Vapnik-Chervonenkis inequality).

C. Back to learning.

→ **Context**:

– An i.i.d. sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \in \mathscr{X} \times \{-1, 1\}$.

– A collection of classifiers $\mathscr{G} = \{g : \mathscr{X} \to \{-1, 1\}\}$.

– A minimizer g_n of the empirical risk $\mathscr{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[g(X_i) \neq Y_i]}$.

– The estimation error

$$\mathbb{E}\mathscr{R}(g_n) - \inf_{g \in \mathscr{G}} \mathscr{R}(g) \leqslant 4\mathbb{E} \sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbf{1}_{[g(X_i) \neq Y_i]}.$$

1. Why is it adapted to consider the class $\mathscr{F}$ of all **indicator functions** of the form $f = \mathbf{1}_{[(x,y):g(x) \neq y]}$, $g \in \mathscr{G}$?

2. Let $\mathscr{A} = \{A_g, g \in \mathscr{G}\}$, where $A_g = \{(x, y) \in \mathscr{X} \times \{-1, 1\}, g(x) \neq y\}$. Show that, for all $n \geqslant 1$, $\mathbf{S}_{\bar{\mathscr{A}}}(n) = \mathbf{S}_{\mathscr{A}}(n)$, where

$$\bar{\mathscr{A}} = \Big\{\{x \in \mathscr{X}, g(x) = 1\}, g \in \mathscr{G}\Big\}.$$

In particular, $V_{\bar{\mathscr{A}}} = V_{\mathscr{A}}$.

Solution. Observe that

$$\mathscr{A} = \Big\{\bar{A} \times \{-1\} \cup \bar{A}^c \times \{1\}, \bar{A} \in \bar{\mathscr{A}}\Big\},$$

where the sets $\bar{A}$ are of the form $\{x \in \mathscr{X}, g(x) = 1\}$, and the sets in $\mathscr{A}$ are sets of pairs (x, y) for which $g(x) \neq y$.

Let N be a positive integer. We show that for any n pairs from $\mathscr{X} \times \{-1, 1\}$, if N sets from $\mathscr{A}$ pick N different subsets of the n pairs, then there are N corresponding sets in $\bar{\mathscr{A}}$ that pick N different subsets of n points in $\mathscr{X}$, and vice versa. Fix n pairs

$$(x_1, -1), \ldots, (x_m, -1), (x_{m+1}, 1), \ldots, (x_n, 1).$$

Note that since ordering does not matter, we may arrange any n pairs in this manner. Assume that for a certain set $\bar{A} \in \bar{\mathscr{A}}$, the corresponding set $A = \bar{A} \times \{-1\} \bigcup \bar{A}^c \times \{1\} \in \mathscr{A}$ picks out the pairs

$$(x_1, -1), \ldots, (x_k, -1), (x_{m+1}, 1), \ldots, (x_{m+\ell}, 1),$$

that is, the set of these pairs is the intersection of A and the n pairs. Again, we can assume without loss of generality that the pairs are ordered in this way. This means that $\bar{A}$ picks from the set $\{x_1, \ldots, x_n\}$ the subset $\{x_1, \ldots, x_k, x_{m+\ell+1}, \ldots, x_n\}$, and the two subsets uniquely determine each other. This shows $\mathbf{S}_{\mathscr{A}}(n) \leqslant \mathbf{S}_{\bar{\mathscr{A}}}(n)$. The other direction is proved in exactly the same way. ■

3. Conclude that

$$\mathbb{E}\mathscr{R}(g_n) - \inf_{g \in \mathscr{G}} \mathscr{R}(g) = \mathrm{O}\left(\sqrt{\frac{V_{\mathscr{G}} \log n}{n}}\right),$$

where we denote $V_{\mathscr{G}}$ instead of $V_{\bar{\mathscr{A}}}$.

4. **Example 1**: let

$$g(x) = \begin{cases} 1 & \text{if } a^\top x + a_0 > 0 \\ -1 & \text{otherwise,} \end{cases}$$

where $a \in \mathbb{R}^d$ and $a_0 \in \mathbb{R}$. Prove that $V_{\mathscr{G}} \leqslant d + 1$.

5. **Example 2**: let

$$\bar{\mathscr{A}} = \Big\{\Big\{x \in \mathbb{R}^d, \sum_{j=1}^{d} (x^{(j)} - a_j)^2 \leqslant a_0\Big\}, (a_0, a_1, \ldots, a_d) \in \mathbb{R}^{d+1}\Big\}.$$

Prove that $V_{\mathscr{G}} \leqslant d + 2$.

Problem 3

Throughout the problem, we let $\mathscr{B}$ be the Borel subsets of $\mathbb{R}^d$.

A. Preliminaries. Let f and g be two probability densities on $\mathbb{R}^d$, that is, nonnegative functions such that

$$\int f = \int g = 1.$$

(All integrals are evaluated with respect to the Lebesgue measure.)

1. Show that
$$\int |f-g| = 2\int_{A_{fg}} (f-g),$$
where A_{fg} is the set $\{f > g\}$, i.e.,
$$A_{fg} = \{x \in \mathbb{R}^d, f(x) > g(x)\}.$$

2. Deduce that
$$\int |f-g| = 2 \sup_{B \in \mathscr{B}} \left| \int_B f - \int_B g \right|.$$

This result is known as **Scheffé's theorem**.

B. A selection problem. Assume we are given a sample of independent random variables $X_1, \ldots, X_n$ with common **unknown** density f. We denote by $\mathscr{F}$ a collection of densities parameterized by θ:

$$\mathscr{F} = \{f_\theta, \theta \in \Theta\}.$$

Our goal is to select in $\mathscr{F}$ the "best" possible density, using only $X_1, \ldots, X_n$.

1. Let μ_n be the empirical measure associated with $X_1, \ldots, X_n$. Explain why the strategy that chooses θ in Θ by minimizing the quantity
$$\sup_{B \in \mathscr{B}} \left| \int_B f_\theta - \mu_n(B) \right|$$
is not a good idea.

2. Introduce the collection of sets

$$\mathscr{A} = \big\{\{f_\theta > f_{\theta'}\}, (\theta, \theta') \in \Theta^2\big\}.$$

In order to choose the "best" density in $\mathscr{F}$, a possible route is to minimize in θ the following criterion:

$$\Delta(\theta) = \sup_{A \in \mathscr{A}} \Big| \int_A f_\theta - \mu_n(A) \Big|.$$

We denote by θ_n an element of Θ such that $\Delta(\theta_n) = \inf_{\theta \in \Theta} \Delta(\theta)$.

2.a Let θ^* be an element of Θ such that

$$\int |f_{\theta^*} - f| = \inf_{\theta \in \Theta} \int |f_\theta - f|.$$

Prove that

$$\int |f_{\theta_n} - f_{\theta^*}| \leqslant 4 \sup_{A \in \mathscr{A}} \Big| \int_A f_{\theta^*} - \mu_n(A) \Big|.$$

2.b Next, show that

$$\int |f_{\theta_n} - f| \leqslant 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + 4\Delta_n,$$

where Δ_n is some explicit random quantity.

2.c Recall the definition of $\mathbf{S}_{\mathscr{A}}(n)$, the shatter coefficient of n points by the class $\mathscr{A}$.

2.d Show that

$$\mathbb{E}\Big(\int |f_{\theta_n} - f| \Big) \leqslant 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + \mathrm{O}\bigg(\sqrt{\frac{\log(\mathbf{S}_{\mathscr{A}}(n))}{n}} \bigg).$$

2.e Provide a statistical interpretation of this inequality.

C. Application. On the real line $\mathbb{R}$, we let $\mathscr{F}$ be the set of Gaussian densities, parameterized by their mean and variance, i.e.,

$$\mathscr{F} = \Big\{ f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/(2\sigma^2)}, \theta = (m, \sigma^2) \in \mathbb{R} \times (0, \infty) \Big\}.$$

1. Prove that $\mathscr{A}$ is contained in a class of sets $\mathscr{B}_2$ that can be easily described.

2. Determine the VC dimension V of $\mathscr{B}_2$.

3. Conclude that

$$\mathbb{E}\Big(\int |f_{\theta_n} - f|\Big) \leqslant 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + \mathrm{O}\left(\sqrt{\frac{V \log n}{n}}\right).$$

Problem 4

A. Preliminaries. We start with some independent questions, which will be useful later in the problem.

1. Let Z be a real random variable with second order moment. Prove that, for all $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}Z \geqslant t) \leqslant \frac{\operatorname{var}(Z)}{\operatorname{var}(Z) + t^2}.$$

   (Hint: if Z is centered, then $t \leqslant \mathbb{E}((t - Z)\mathbf{1}_{[Z<t]})$.)

2. Let Z be a binomial random variable with parameters $n \in \mathbb{N}^*$ and $p \in (0, 1)$. Prove that

$$\mathbb{E}\Big(\frac{1}{Z}\mathbf{1}_{[Z>0]}\Big) \leqslant \frac{2}{(n+1)p}.$$

   (Hint: start by providing a upper bound on $\mathbb{E}(\frac{1}{1+Z})$.)

3. Let $(p_1, \ldots, p_k)$ be a probability vector (i.e., $p_i \geqslant 0$ and $\sum_{i=1}^k p_i = 1$). Show that

$$\sum_{i=1}^k p_i(1 - p_i)^n \leqslant \frac{k}{en}.$$

B. The problem. Let k be a positive integer and let (X, Y) be a pair of random variables taking values in $\{1, \ldots, k\} \times \{0, 1\}$. The distribution of the **discrete** random variable X is thus fully described by the probability vector $(p_1, \ldots, p_k)$, where $p_i = \mathbb{P}(X = i)$. We let $\eta(x) = \mathbb{P}(Y = 1|X = x)$ and denote by $\mathscr{R}^*$ the Bayes risk associated with (X, Y).

Assume we are given a sample $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of independent random variables, all distributed as (and independent of) (X, Y). We consider the natural classifier g_n defined for all $x \in \{1, \ldots, k\}$ by

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{1}_{[X_i=x]}\mathbf{1}_{[Y_i=1]} > \sum_{i=1}^n \mathbf{1}_{[X_i=x]}\mathbf{1}_{[Y_i=0]} \\ 0 & \text{otherwise.} \end{cases}$$

(By convention, an empty sum is zero.) We let

$$\mathscr{R}(g_n) = \mathbb{P}(g_n(X) \neq Y | \mathscr{D}_n).$$

The main objective of the problem is to establish that

$$\mathbb{E}\mathscr{R}(g_n) - \mathscr{R}^* \leqslant \sqrt{\frac{k}{2(n+1)}} + \frac{k}{en}. \tag{13.1}$$

Warm-up.

1. Prove in one line that $\mathbb{E}\mathscr{R}(g_n) \to \mathscr{R}^*$ as n tends to infinity.

2. Show that
$$\mathscr{R}(g_n) \geqslant \sum_{x:\sum_{i=1}^n \mathbf{1}_{[X_i=x]}=0} \eta(x) p_x.$$

3. Deduce that
$$\mathbb{E}\mathscr{R}(g_n) \geqslant \sum_{x=1}^k \eta(x) p_x (1-p_x)^n.$$

4. We assume in this question (and only in this question) that $\eta(x) = 1$ for all x.

   4.a What is the value of $\mathscr{R}^*$?

   4.b Find a vector $(p_1, \ldots, p_k)$ such that $\mathbb{E}\mathscr{R}(g_n) \geqslant 1/2$ for all $k \geqslant 2n$.

   4.c Conclusion?

Proof of inequality (13.1).

1. In the sequel, we let $N(x) = \sum_{i=1}^n \mathbf{1}_{[X_i=x]}$. Rewrite $g_n(x)$ using $N(x)$ (with the convention $0/0 = 0$).

2. What is, conditionally on $\mathbf{1}_{[X_1=x]}, \ldots, \mathbf{1}_{[X_n=x]}$, the distribution of the random variable $Z(x) = \sum_{i=1}^n \mathbf{1}_{[X_i=x]} Y_i$?

3. Prove that
$$\mathbb{E}\mathscr{R}(g_n) = \sum_{x=1}^k p_x \big(\eta(x) + (1-2\eta(x))\mathbb{P}(\mathrm{Bin}(N(x), \eta(x)) > N(x)/2)\big),$$
where the notation $\mathrm{Bin}(N(x), \eta(x))$ means a binomial random variable with parameters $N(x)$ and $\eta(x)$ (null by convention if $N(x) = 0$).

4. Deduce that

$$\mathbb{E}\mathscr{R}(g_n) \leqslant \sum_{x=1}^{k} p_x\big(\xi(x) + \big(1 - 2\xi(x)\big)\mathbb{P}(\mathrm{Bin}(N(x), \xi(x)) \geqslant N(x)/2)\big),$$

where $\xi(x) = \min(\eta(x), 1 - \eta(x))$. (Hint: observe that $\mathbb{P}(\mathrm{Bin}(m, p) \leqslant m/2) = \mathbb{P}(\mathrm{Bin}(m, 1 - p) \geqslant m/2)$.)

5. Next, show that

$$\mathbb{E}\mathscr{R}(g_n) - \mathscr{R}^* \leqslant \sum_{x=1}^{k} p_x(1 - 2\xi(x))\mathbb{E}\Big(\frac{1}{1 + (1 - 2\xi(x))^2 N(x)}\Big).$$

6. Prove that

$$\mathbb{E}\mathscr{R}(g_n) - \mathscr{R}^* \leqslant \sum_{x=1}^{k} p_x \mathbb{E}\Big(\frac{1}{2\sqrt{N(x)}}\mathbf{1}_{[N(x)>0]} + (1 - 2\xi(x))\mathbf{1}_{[N(x)=0]}\Big).$$

7. Conclude that

$$\mathbb{E}\mathscr{R}(g_n) - \mathscr{R}^* \leqslant \sum_{x=1}^{k} p_x(1 - p_x)^n + \frac{1}{2}\sum_{x=1}^{k} p_x\sqrt{\mathbb{E}\Big(\frac{1}{N(x)}\mathbf{1}_{[N(x)>0]}\Big)}.$$

8. Establish inequality (13.1).

C. The multivariate case with independent components. We assume in this last section that X is a multivariate random variable taking values on $\{0, 1\}^d$. We let $X = (X^{(1)}, \ldots, X^{(d)})$ (each $X^{(j)}$ is thus taking values in $\{0, 1\}$) and assume that $X^{(1)}, \ldots, X^{(d)}$ are **independent conditionally** to $Y = 1$, and also **independent conditionally** to $Y = 0$. We let

$$p(j) = \mathbb{P}(X^{(j)} = 1|Y = 1), \quad q(j) = \mathbb{P}(X^{(j)} = 1|Y = 0),$$

and $p = \mathbb{P}(Y = 1)$, and assume that all these quantities are strictly comprised between 0 and 1.

1. For $x = (x^{(1)}, \ldots, x^{(d)})$, what are $\mathbb{P}(X = x|Y = 1)$ and $\mathbb{P}(X = x|Y = 0)$?

2. Give the expression of the Bayes rule g^* associated with the pair (X, Y).

3. Letting

$$\alpha_0 = \ln\left(\frac{p}{1-p}\right) + \sum_{j=1}^{d} \ln\left(\frac{1-p(j)}{1-q(j)}\right)$$

and

$$\alpha_j = \ln\left(\frac{p(j)}{q(j)} \cdot \frac{1-q(j)}{1-p(j)}\right), \quad 1 \leqslant j \leqslant d,$$

write g^* as a function of α_0 and the α_j.

4. Why is this result interesting?