PIOTR ZWIERNIK

ADVANCED THEORY OF STATISTICS

Introduction

These lecture notes are intended to offer complement to the first part of STA3000. In designing this course we aimed at giving a modern treatment of mathematical statistics. Thus, we discuss important topics in theoretical statistics, like exponential families, statistical decision theory, or empirical processes and develop basic intuition behind what is multivariate statistics and what are its basic problems and techniques.

These goals heavily affected the exposition. Our approach is to use the multivariate notation whenever possible, to emphasize connections to convex analysis, and to present some results in the high-dimensional statistics. We will try to show that important of convex analysis for statistical theory goes much beyond convex optimization used for the maximum likelihood estimation or its regularized versions.

The whole material is divided into twelve 3-hour lectures. The notes contain more detailed material than presented in the lecture. Preparing these lecture notes we benefited from several excellent textbooks or lecture notes:

1. Robert W. Keener, *Theoretical statistics*.
2. Lehmann, Romano, *Testing Statistical Hypotheses*.
3. Sundberg, *Statistical Modelling by Exponential Families*.
4. Wainwright, *High-dimensional statistics*.
5. Martin, *Lecture Notes on Advanced Statistical Theory*.
6. Rigolett, High-dimensional Statistics.
7. van der Vaart, Asymptotic Statistics

Special thanks go to Morris Greenberg, Ichiro Hashimoto, Vishakh Patel, Emily Somerset, Qiang Sun, Leonard Wang, and Zhenghang Xu for helping me to improve the notes.

Contents

I Topics in Statistical Inference 9

1 *Exponential families (2 weeks)* 11

1.1 *Definition and examples* 11

1.2 *Basic results* 14

1.3 *Convexity and the MLE* 17

1.4 *Marginal and conditional distributions** 22

1.5 *Conditional inference for canonical parameter** 25

1.6 *Kullback-Leibler divergence* 26

1.7 *Generalized Linear Models* 29

1.8 *Diaconis-Ylvisaker conjugate priors* 31

1.9 *Exercises* 32

2 *Statistical Decision Theory (1 week)* 35

2.1 *Decision theoretic framework** 35

2.2 *Randomized decision rules** 38

2.3 *Bayesian and minimax rules** 40

2.4 *Admissibility and Rao-Blackwell* 47

2.5 *Stein's paradox* 49

2.6 *Minimizing risk under constraints* 54

2.7 *Exercises* 62

3 *Hypothesis testing and multiple testing (2 weeks)* 67

3.1 *Neyman-Pearson lemma** 68

3.2 *Some constructions of UMP tests** 73

3.3 *Sequential testing** 76

3.4 *Motivating multiple testing* 81

3.5 *Family-wise error rate* 81

3.6 *False discovery rate* 86

3.7 *Exercises* 90

II *Statistical Learning Theory: An Empirical Process Perspective* 93

4 *Motivation and examples (1 week)* 95

4.1 *Uniform law of large numbers* 95

4.2 *The Uniform Central Limit Theory* 100

4.3 *Exercises* 103

5 *Concentration of measure (3 weeks)* 105

5.1 *Basic inequalities* 105

5.2 *Sub-gaussian and sub-exponential random variables* 109

5.3 *Martingale-based methods* 114

5.4 *Lipschitz functions of Gaussian variables* 118

5.5 *Exercises* 121

6 *More advanced techniques (1-2 weeks)* 123

6.1 *Maximal inequalities* 123

6.2 *Rademacher complexity and bounds on suprema* 126

6.3 *Polynomial discrimination and VC dimension* 128

6.4 *Chaining* 132

6.5 *Exercises* 135

7 *Applications in statistics (1-2 weeks)* 137
7.1 *Sub-Gaussian sequence model with sparsity* 137
7.2 *Fixed design linear regression* 140
7.3 *Constrained least squares estimator* 142
7.4 *LASSO regression* 143
7.5 *Random matrices* 147
7.6 *Exercises* 152

III Some topics from the first semester 155

8 *Classical large sample theory* 157
8.1 *Preliminaries* 157
8.2 *Delta Method* 160
8.3 *M-estimators* 162
8.4 *Generalized likelihood ratio test* 165
8.5 *Limits of Bayesian procedures* 168
8.6 *Exercises* 170

IV Mathematical appendix 171

A *Real Analysis* 173
A.1 *Vector spaces* 173
A.2 *Continuity and semicontinuity* 175
A.3 *Differentiation* 177

B *Convex Analysis* 181
B.1 *Convexity and hyperplane separation* 181
B.2 *Convex functions and optimization* 184

C *Probability* 189
C.1 *Continuity of probability* 189
C.2 *Martingales* 190

Part I

Topics in Statistical Inference

1
Exponential families (2 weeks)

Exponential families were discussed briefly in the first semester. The goal of this section is to provide a more detailed treatment of multivariate exponential families in connection with convexity, sufficiency, and hypothesis testing. This chapter is mostly based on two books on exponential families by Lawrence D. Brown[1] and by Rolf Sundberg [2], and on the lecture notes „Topics in Information Geometry" of our colleague Ting-Kam Leonard Wong.

[1] Lawrence D. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1986

[2] Rolf Sundberg. *Statistical modelling by exponential families*, volume 12 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, Cambridge, 2019

1.1 *Definition and examples*

Basic definition and univariate examples appeared in the first part of the lecture. We focus on developing uniform notation in the multivariate case. Consider a random vector $\mathbf{X} = (X_1, \dots, X_m)$ with values in the state space $\mathcal{X} \subseteq \mathbb{R}^m$ equipped with a σ-finite measure μ.[3]

[3] Typically the measure μ is either the counting measure or the Lebesgue measure.

Definition 1.1.1. *A parametric statistical model for $\mathbf{X}$ is an exponential family with canonical parameter vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$ and canonical statistics $\boldsymbol{t}(\mathbf{x}) = (t_1(\mathbf{x}), \dots, t_d(\mathbf{x}))$, if it admits a density f with respect to μ and f has the form*

$$f(\mathbf{x}; \boldsymbol{\theta}) \;=\; h(\mathbf{x}) \exp\Big\{ \big\langle \boldsymbol{\theta}, \boldsymbol{t}(\mathbf{x}) \big\rangle - A(\boldsymbol{\theta}) \Big\}. \tag{1.1}$$

Formally, we define $\mathcal{X}$ as the smallest closed set satisfying $\mathbb{P}_\theta(X \in \mathcal{X}) = 1$. This definition does not depend on the choice of $\boldsymbol{\theta}$ because all $\mathbb{P}_{\boldsymbol{\theta}}$ have the same support.

Remark 1.1.2. *For notational purposes it is often easier to subsume $h(\mathbf{x})$ into the underlying measure and use the formulation*

$$f(\mathbf{x}; \boldsymbol{\theta}) \;=\; \exp\Big\{ \big\langle \boldsymbol{\theta}, \boldsymbol{t}(\mathbf{x}) \big\rangle - A(\boldsymbol{\theta}) \Big\}. \tag{1.2}$$

Another useful reformulation is when $h(\mathbf{x})$ in (1.1) is itself a density. In this case $A(\mathbf{0}) = 0$.

Example 1.1.3 (Bernoulli variable and logistic regression). *If* $X \sim \text{Bern}(p)$ *with* $p \in (0,1)$ *we can write its distribution in the exponential form. For* $x \in \{0,1\}$ *we have*

$$f(x;p) = p^x(1-p)^{1-x} = \exp\left\{x \log\left(\frac{p}{1-p}\right) + \log(1-p)\right\}.$$

We have $t(x) = x$ *and the canonical parameter is the logit of* p*:*

$$\text{logit}(p) \;:=\; \log\frac{p}{1-p} \;=\; \theta.$$

We also have $h(x) = 1$ *and* $A(\theta) = \log(1+e^\theta)$*. In logistic regression, we model* $\text{logit}(p)$ *as a linear function of regressors.*

Example 1.1.4 (Univariate Gaussian). *If* $X \sim N(\mu, \sigma^2)$ *then*

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} + \frac{1}{2}\left(\log(\frac{1}{\sigma^2}) - \frac{\mu^2}{\sigma^2}\right)},$$

which can be written as a two dimensional exponential family with

$$\mathbf{t}(x) = \left(x, -\frac{x^2}{2}\right), \qquad \boldsymbol{\theta} = \left(\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2}\right).$$

Then

$$h(\mathbf{x}) = \frac{1}{\sqrt{2\pi}}, \qquad A(\boldsymbol{\theta}) = -\frac{1}{2}\left(\log(\theta_2) - \frac{\theta_1^2}{\theta_2}\right).$$

If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ *are i.i.d. then the joint distribution of this sample* $\mathbf{X}_{1:n} = (X_1, \ldots, X_n)$ *is also Gaussian with the same canonical parameters and the density*

$$f(\mathbf{x}_{1:n}; \boldsymbol{\theta}) \;=\; \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_i x_i^2 + \frac{\mu}{\sigma^2}\sum_i x_i + \frac{n}{2}\left(\log(\frac{1}{\sigma^2}) - \frac{\mu^2}{\sigma^2}\right)}$$

and so it forms an exponential family with the same canonical parameter and with the sufficient statistics $\left(\sum_i x_i, -\frac{1}{2}\sum_i x_i^2\right)$*. We know, of course, that this distribution is the n-variate Gaussian with parameters* $\mu\mathbf{1}$ *and* $\sigma^2 I_n$*.*

Consider the function

$$Z(\boldsymbol{\theta}) \;:=\; \int_{\mathcal{X}} h(\mathbf{x}) \exp\left\{\langle \boldsymbol{\theta}, \boldsymbol{t}(\mathbf{x})\rangle\right\} \mu(\mathrm{d}\mathbf{x}), \tag{1.3}$$

where we put $Z(\boldsymbol{\theta}) = +\infty$ if this integral is infinite. Because f is a density function, it follows that

$$A(\boldsymbol{\theta}) \;=\; \log Z(\boldsymbol{\theta}).$$

The function $A(\boldsymbol{\theta})$ plays a special role in this theory and it has many names: the log-partition function, Laplace transform, or the cumulant function. We define the space of canonical parameters as

$$\Theta \;=\; \{\theta \in \mathbb{R}^d :\; Z(\boldsymbol{\theta}) < +\infty\}. \tag{1.4}$$

It is implicit in (1.1) that $\boldsymbol{\theta} \in \Theta$. An exponential family model is then specified by possibly constraining to $\Theta_0 \subseteq \Theta$. The dimension d of $\boldsymbol{\theta}$ is called the order of this exponential family.

Example 1.1.5. *Consider the univariate Gaussian case in Example 1.1.4. Since $\mu \in \mathbb{R}$, $\sigma^2 > 0$, the space of canonical parameters is $\Theta = \{(\theta_1, \theta_2) : \theta_1 \in \mathbb{R}, \theta_2 > 0\}$. Taking $\mu = 0$ corresponds to fixing a linear subspace $\Theta_0 = \{\boldsymbol{\theta} \in \Theta : \theta_1 = 0\}$.*

In our basic set-up both the parameters and the sufficient statistics lie in $\mathbb{R}^d$ but exactly the same definition can be provided for a general d-dimensional vector space with a given inner product. The most relevant example is when the underlying vector space is the space $\mathbb{S}^m$ of all symmetric $m \times m$ matrices. Here the standard inner product is given by $\langle A, B\rangle = \operatorname{tr}(AB)$. Denote by $\mathbb{S}^m_+$ the set of positive definite matrices in $\mathbb{S}^m_+$.

Show that $\operatorname{tr}(AB) = \sum_{ij} A_{ij}B_{ij}$

Example 1.1.6 (Centered multivariate Gaussian distribution)**.** *Consider the m-variate Gaussian distribution with the zero mean vector and covariance matrix $\Sigma \in \mathbb{S}^m_+$. Let $K = \Sigma^{-1}$. The density with respect to the Lebesgue measure is*

$$f(\mathbf{x}; K) = \frac{1}{(2\pi)^{m/2}}\sqrt{\det(K)}\exp\{-\tfrac{1}{2}\mathbf{x}^\intercal K\mathbf{x}\}.$$

This is an exponential family with $h(\mathbf{x}) = \frac{1}{(2\pi)^{m/2}}$, $A(K) = -\frac{1}{2}\log\det(K)$. Denoting $\boldsymbol{t}(\mathbf{x}) = -\frac{1}{2}\mathbf{x}\mathbf{x}^\intercal \in \mathbb{S}^m$ we get

$$-\tfrac{1}{2}\mathbf{x}^\intercal K\mathbf{x} \;=\; \operatorname{tr}(K\boldsymbol{t}(\mathbf{x})) = \langle K, \boldsymbol{t}(\mathbf{x})\rangle.$$

← Exercise 1.9.1

Given a sample $\boldsymbol{x}_{1:n} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ from a distribution with density $f(\boldsymbol{x}; \boldsymbol{\theta})$, the log-likelihood function is

← Exercise 1.9.2

$$\ell_n(\boldsymbol{\theta}) \;:=\; \frac{1}{n}\sum_{i=1}^n \log f(\boldsymbol{x}_i; \boldsymbol{\theta}).$$

Note that we normalize the log-likelihood by n to get the interpretation as the expectation of $\log f(\mathbf{x}; \boldsymbol{\theta})$ under the sample distribution.

Proposition 1.1.7 (The log-likelihood function)**.** *If $\boldsymbol{x}_{1:n} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is a sample from the exponential family (1.1) then, denoting*

← Exercise 1.9.3

$$\overline{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{t}(\boldsymbol{x}_i),$$

the log-likelihood takes the form

$$\ell_n(\boldsymbol{\theta}) \;=\; \langle \boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n \rangle - A(\boldsymbol{\theta}) + (\text{constant}). \tag{1.5}$$

In the statistical/ML practice we typically go the other way around. We first define a suitable sufficient statistics that should contain all the relevant information of the data. This choice defines then an exponential family.

Example 1.1.8 (Exponential random graph model)**.** *Each graph can be associated to its adjacency matrix $A \in \{0,1\}^{n\times n}$. This is a symmetric matrix with zeros on the diagonal and so each graph is an element $\mathbf{x} \in \mathcal{X} = \{0,1\}^{\binom{n}{2}}$ with entries x_{ij} for $1 \le i < j \le n$. For the simplest example, consider $\mathbf{t}(\mathbf{x}) = \sum_{i<j} x_{ij} \in \mathbb{R}$, which is simply the number of edges of the underlying graph. The corresponding exponential family has one parameter θ and is of the form*

$$f(\mathbf{x};\boldsymbol{\theta}) \;=\; \exp\left\{\theta \sum_{i<j} x_{ij} - A(\theta)\right\} \;\propto\; \prod_{i<j} e^{\theta x_{ij}}.$$

In other words, each edge x_{ij} is an independent Bernoulli variable with the success probability $p = e^\theta/(1+e^\theta)$ and computing the normalizing constant is easy. This is the famous Erdős-Renyi model. Other statistics of the graph will give different exponential models.

Another model for binary variables that uses the graph structure is the Ising model.

Example 1.1.9 (The Ising model)**.** *Let G be a graph over m nodes representing m binary random variables $X_i \in \{-1,1\}$ for $i = 1,\dots,m$. Consider the model*[4]

$$f(\mathbf{x};\boldsymbol{\theta}) \;\propto\; \exp\{\sum_{ij\in G} \theta_{ij} x_i x_j\}.$$

Here computing the normalizing constant is generally hard. This and similar examples motivated developing methods that do not rely on computing this normalizing constant (e.g. variational inference).

[4] The statistical interpretation of this modelling construction in terms of conditional independence comes from the Hammersley-Clifford theorem.

← Exercise 1.9.4

← Exercise 1.9.5

There are many models with a given sufficient statistics. What is then special about the exponential families? To answer this question we need to develop more theory and we will see that this is connected to maximizing the entropy; c.f. Theorem 1.6.5.

1.2 *Basic results*

We assume throughout that there is *no* hyperplane $H = \{\boldsymbol{x} : \langle \boldsymbol{\alpha}, \boldsymbol{t}(\boldsymbol{x})\rangle = c\}$ such that $\mathbb{P}_\theta(H) = 1$. In other words, no entry of the vector $\boldsymbol{t}$ can be written as an affine combination of the remaining entries. Similarly, we assume that there is no hyperplane

For example $\langle (1,2), (x_1,x_2)\rangle = x_1 + 2x_2 = 1$ is a simple hyperplane in $\mathbb{R}^2$.

$\{\boldsymbol{\theta} : \langle \boldsymbol{\theta}, \boldsymbol{\beta} \rangle = c\}$ containing Θ. An exponential family (1.1) with these two properties is called **minimal**. A canonical example of a family that is not minimal appears for discrete data.

Example 1.2.1 (Bernoulli variable)**.** *The Bernoulli distribution in Example 1.1.3 could be alternatively written in form of a two-dimensional exponential family*

$$p(x) \;=\; \exp\{\log(1-p)\mathbb{1}(x=0) + \log(p)\mathbb{1}(x=1)\} \qquad \text{for } x \in \{0,1\}$$

with $\boldsymbol{t}(x) = (\mathbb{1}(x=0), \mathbb{1}(x=1))$, $\boldsymbol{\theta} = (\log(1-p), \log(p))$, $A(\boldsymbol{\theta}) = 0$. *Clearly, this representation is not minimal as* $\mathbb{1}(x=0) + \mathbb{1}(x=1) = 1$. *An easy fix is to define the new (minimal) sufficient statistics* $x = \mathbb{1}(x=1)$ *and rewrite the above using the fact that* $\mathbb{1}(x=0) = 1 - x$. *This gives the representation in Example 1.1.3.*

A slightly more complicated version of this example appears in the vector case. But the idea is similar and the difficulty is purely notational.

Example 1.2.2 (Binary vectors)**.** *Consider a binary vector* $X = (X_1, \ldots, X_m)$ *with the probability distribution* $p(\mathbf{x})$ *for* $\mathbf{x} \in \mathcal{X} = \{0,1\}^m$. *As in the Bernoulli case, we can write*

$$p(\mathbf{x}) = \exp\{\langle \boldsymbol{\theta}, \boldsymbol{t}(\mathbf{x}) \rangle\} \qquad \text{for } \mathbf{x} \in \mathcal{X},$$

where $\boldsymbol{\theta}, \boldsymbol{t}(\mathbf{x}) \in \mathbb{R}^{\mathcal{X}}$. *By this, we mean that* $\boldsymbol{\theta}$ *and* $\boldsymbol{t}(\mathbf{x})$ *are themselves functions on* $\mathcal{X}$: $\boldsymbol{\theta}(\mathbf{y})$ *and* $\boldsymbol{t}(\mathbf{x}, \mathbf{y})$ *for* $\mathbf{y} \in \mathcal{X}$, *such that* $\boldsymbol{\theta}(\mathbf{y}) = \log p(\mathbf{y})$, *and*

$$\boldsymbol{t}(\mathbf{x}, \mathbf{y}) \;=\; \mathbb{1}(\mathbf{x} = \mathbf{y}) \;=\; \begin{cases} 1 & \text{if } \mathbf{x} = \mathbf{y}, \\ 0 & \text{otherwise.} \end{cases}$$

The inner product in $\mathbb{R}^{\mathcal{X}}$ *simply means*

$$\langle \boldsymbol{\theta}, \boldsymbol{t}(\mathbf{x}) \rangle \;=\; \sum_{\mathbf{y} \in \mathcal{X}} \boldsymbol{\theta}(\mathbf{y}) \boldsymbol{t}(\mathbf{x}, \mathbf{y})$$

However $\sum_{\mathbf{x} \in \mathcal{X}} \boldsymbol{t}(\mathbf{x}) = \mathbf{1} \in \mathbb{R}^{\mathcal{X}}$ *so this representation is not minimal. We reduce the dimension by rewriting*

$$\boldsymbol{t}(\mathbf{0}) \;=\; \mathbf{1} - \sum_{\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}} \boldsymbol{t}(\mathbf{x}).$$

Show that this defines a minimal exponential family with $\Theta = \mathbb{R}^{\mathcal{X} \setminus \{\mathbf{0}\}}$ *and canonical parameters* $\theta(\mathbf{y}) = \log p(\mathbf{y}) - \log p(\mathbf{0})$ *and*

$$A(\boldsymbol{\theta}) \;=\; \log\left(1 + \sum_{\mathbf{x} \neq \mathbf{0}} e^{\langle \boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x}) \rangle_0}\right),$$

where the inner product $\langle \cdot, \cdot \rangle_0$ *is defined in the space* $\mathbb{R}^{\mathcal{X} \setminus \{\mathbf{0}\}}$, *that is,*

$$\langle \boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x}) \rangle_0 \;=\; \sum_{\mathbf{y} \neq \mathbf{0}} \boldsymbol{\theta}(\mathbf{y}) \boldsymbol{t}(\mathbf{x}, \mathbf{y}).$$

Making sure that there is no hyperplane containing $t(\mathcal{X})$ can be easily done by changing the sufficient statistics as in the examples above[5]. By redefining Θ if necessary, we can also always without loss assume that the canonical parameters are not contained in an affine subspace.

[5] If one coordinate of $t(\mathcal{X})$ can be written as an affine combination of the others, we simply replace this coordinate with this combination obtaining a sufficient statistics with one dimension lower. This process can be repeated.

← Exercise 1.9.6

Example 1.2.3 (Gaussian graphical models)**.** *Consider the centered m-variate Gaussian model in Example 1.1.6 with the canonical parameter $K = \Sigma^{-1} \in \Theta = \mathbb{S}^m_+$. Fix a graph G over m nodes and the subset*

$$\Theta_G \;=\; \{K \in \Theta : K_{ij} = 0 \text{ if } ij \notin G\}.$$

This linear constrain defines an exponential family with sufficient statistic $(-\frac{1}{2}x_i x_j)_{ij\in G}$. The corresponding model, is typically called the Gaussian graphical model, which forms a popular dimension reduction technique. Its relevance in practice comes from the fact that in the Gaussian distribution $K_{ij} = 0$ if and only X_i is independent of X_j given all the remaining variables in the system.[6]

← Exercise 1.9.5

[6] Caroline Uhler. Gaussian graphical models. In *Handbook of Graphical Models*, pages 217–238. CRC Press, 2018

Remark 1.2.4. *The rest of this section is not esssential if you skip Section 1.4 and Section 1.5.*

Definition 1.2.5. *A minimal exponential family is called **full** if its parameter space is maximal, that is, Θ_0 equals the canonical space Θ.*

Some relevant examples of a non-full exponential family are when the parameter space Θ_0 is a convex subset of Θ (convex exponential families) or when it forms a lower dimensional manifold (curved exponential families). Note however that if this manifold is a linear subspace we again get a full exponential family after a reparametrization.

← Exercise 1.9.6

Recall from the first semester that a statistic $\mathbf{t}(\mathbf{x})$ is **sufficient** for $\boldsymbol{\theta}$ if the conditional distribution of $\mathbf{x}$ given $\mathbf{t}(\mathbf{x})$ does not depend on $\boldsymbol{\theta}$. Sufficiency of $\mathbf{t}(\mathbf{x})$ in exponential families can by argued by Proposition 1.2.7 below or directly by Fisher-Neyman Factorization Theorem, which states that $\mathbf{t}(\mathbf{x})$ is sufficient for $\boldsymbol{\theta}$ if and only if $p(\mathbf{x};\boldsymbol{\theta}) = h(\mathbf{x})g_{\boldsymbol{\theta}}(\mathbf{t}(\mathbf{x}))$ for some $h, g_{\boldsymbol{\theta}}$. Typically the canonical statistics is also minimal sufficient, that is, for any other sufficient statistic $\mathbf{t}'(\mathbf{x})$,

$$\mathbf{t}'(\mathbf{x}) = \mathbf{t}'(\mathbf{y}) \qquad \Longrightarrow \qquad \mathbf{t}(\mathbf{x}) = \mathbf{t}(\mathbf{y}).$$

Proposition 1.2.6 (Minimal sufficiency of $\boldsymbol{t}$)**.** *In a full exponential family the statistic $\boldsymbol{t}(\mathbf{x})$ is minimally sufficient for $\boldsymbol{\theta}$.*

Proof. Consider any other sufficient statitsics $\mathbf{t}'$. If $\mathbf{t}'(\mathbf{x}) = \mathbf{t}'(\mathbf{y})$ then the factorization theorem shows that $\frac{f(\mathbf{x},\boldsymbol{\theta})}{f(\mathbf{y};\boldsymbol{\theta})}$ is independent of $\boldsymbol{\theta}$. On the other hand

$$\frac{f(\mathbf{x},\boldsymbol{\theta})}{f(\mathbf{y};\boldsymbol{\theta})} \;=\; \frac{h(\mathbf{x})}{h(\mathbf{y})} \exp\{\langle \boldsymbol{\theta}, \mathbf{t}(\mathbf{x}) - \mathbf{t}(\mathbf{y})\rangle\},$$

which can be constant in $\boldsymbol{\theta}$ if and only if $\langle \boldsymbol{\theta}, \mathbf{t}(\mathbf{x}) - \mathbf{t}(\mathbf{y}) \rangle$ is constant. Because the family is full (in particular $\boldsymbol{\theta}$ is not contained in an affine subspace), this happens if and only if $\mathbf{t}(\mathbf{x}) = \mathbf{t}(\mathbf{y})$. □

Note that in the proof we only used that $\boldsymbol{\theta}$ is not contained in an affine space, so this result generalizes to curved exponential families.

Proposition 1.2.7 (Distribution of the sufficient statistic)**.** *Suppose* $\mathbf{X}$ *has distribution in the exponential family* (1.1)*. Then, under certain regularity conditions, the distribution of* $\boldsymbol{t} = \boldsymbol{t}(\mathbf{X})$ *is*

$$f(\boldsymbol{t}; \boldsymbol{\theta}) = g(\boldsymbol{t}) \exp\{\langle \boldsymbol{\theta}, \boldsymbol{t} \rangle - A(\boldsymbol{\theta})\}, \tag{1.6}$$

where the structure function $g(\boldsymbol{t})$ *in the discrete case is*

In the discrete case the proof is elementary.

$$g(\mathbf{t}) = \sum_{\mathbf{t}(\mathbf{x})=\mathbf{t}} h(\mathbf{x}),$$

and in the continuous case

The last expression is written rather informally. The integral is computed with respect to the measure on the set $\{\mathbf{x} : \mathbf{t}(\mathbf{x}) = \mathbf{t}\}$ induced from the Lebesgue measure. Here the proof is non-trivial and we skip it.

$$g(\mathbf{t}) = \int_{\mathbf{t}(\mathbf{x})=\mathbf{t}} h(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

Example 1.2.8. *Consider a centered Gaussian distribution with* $\Sigma \in \mathbb{S}_+^m$*. Given a random sample* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *from this distribution, the statistics* $n\overline{\boldsymbol{\mu}}_n := \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\intercal$ *has the Wishart distribution.*

1.3 Convexity and the MLE

Recall a basic version of the Hölder's inequality.

Proposition 1.3.1 (Hölder's inequality)**.** *If* f, g *are two functions on a measurable space* $(\mathcal{X}, \mu)$ *then for every* $p, q \in [1, \infty]$ *such that* $1/p + 1/q = 1$ *we have*

$$\int_{\mathcal{X}} |f(\mathbf{x}) g(\mathbf{x})| \mu(\mathrm{d}\mathbf{x}) \;\leq\; \left(\int_{\mathcal{X}} |f(\mathbf{x})|^p \mu(\mathrm{d}\mathbf{x}) \right)^{1/p} \left(\int_{\mathcal{X}} |g(\mathbf{x})|^q \mu(\mathrm{d}\mathbf{x}) \right)^{1/q}. \tag{1.7}$$

Moreover, if $p, q > 1$ *then* (1.7) *holds as equality if and only if* $|f|^p$ *and* $|g|^q$ *are linearly dependent in* $L^1(\mathcal{X})$ *meaning that there exist real numbers* $\alpha, \beta \geq 0$ *such that* $\alpha |f|^p = \beta |g|^q$ *µ-almost everywhere.*

$L^1(\mathcal{X})$ is the vector space of all functions f on $\mathcal{X}$ with the property that $\|f\|_1 := \int_{\mathcal{X}} |f(x)| \mu(\mathrm{d}\mathbf{x})$ is finite. Formally, we identify two functions that are equal almost surely.

Here is an important fundamental fact about exponential families.

Theorem 1.3.2. *For every exponential family* (1.1) *with* Θ *defined in* (1.4) *we have:*

(i) Θ *is a convex set and the function* $A(\boldsymbol{\theta})$ *is convex on* Θ*.*

(ii) $\mathbb{P}_{\boldsymbol{\theta}_1} = \mathbb{P}_{\boldsymbol{\theta}_2}$ *if and only if*

$$A((1-\lambda)\boldsymbol{\theta}_1 + \lambda \boldsymbol{\theta}_2) = (1-\lambda) A(\boldsymbol{\theta}_1) + \lambda A(\boldsymbol{\theta}_2) \qquad \text{for all } \lambda \in (0, 1).$$

(iii) If the exponential family is minimal then A is strictly convex on Θ and $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$ if $\theta_1 \neq \theta_2 \in \Theta$.

(iv) A is lower semi-continuous on $\mathbb{R}^d$ and is continuous in the interior of Θ.

For the definitions and basic results see Section A.2.2.

Proof. (i) Let $\theta_1, \theta_2 \in \Theta$, $\lambda \in (0,1)$ and denote $\theta_\lambda = (1-\lambda)\theta_1 + \lambda\theta_2$. By the Hölder's inequality with $p = 1/(1-\lambda)$ and $q = 1/\lambda$:

$$\begin{aligned} Z(\theta_\lambda) &= \int h(\mathbf{x}) e^{\langle \theta_\lambda, \mathbf{t}(\mathbf{x})\rangle} \mu(\mathrm{d}\mathbf{x}) \\ &= \int \left(h(\mathbf{x}) e^{\langle \theta_1, \mathbf{t}(\mathbf{x})\rangle}\right)^{1-\lambda} \left(h(\mathbf{x}) e^{\langle \theta_2, \mathbf{t}(\mathbf{x})\rangle}\right)^{\lambda} \mu(\mathrm{d}\mathbf{x}) \\ &\leq \left(\int h(\mathbf{x}) e^{\langle \theta_1, \mathbf{t}(\mathbf{x})\rangle} \mu(\mathrm{d}\mathbf{x})\right)^{1-\lambda} \left(\int h(\mathbf{x}) e^{\langle \theta_2, \mathbf{t}(\mathbf{x})\rangle} \mu(\mathrm{d}\mathbf{x})\right)^{\lambda} \\ &= Z(\theta_1)^{1-\lambda} Z(\theta_2)^{\lambda}. \end{aligned}$$

Taking the logs we get convexity of A. Now convexity of Θ follows easily.

(ii) The Hölder's inequality above is strict unless

$$\langle \theta_1 - \theta_2, \mathbf{t}(\mathbf{x})\rangle \equiv \text{const} \qquad (\mu \text{ a.s.}). \tag{1.8}$$

This last assertion is equivalent to $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$.

(iii) If (1.8) holds for some $\theta_1 \neq \theta_2$ then the exponential family is not minimal.

(iv) By Fatou's lemma $Z(\theta)$ is lower semicontinuous and so $A(\theta)$ is lower semicontinuous. Any convex function defined and finite on a convex set $\Theta \subset \mathbb{R}^d$ must be continuous on the interior of Θ. □

← Exercise B.2.10

The following definition will be important in the rest of this chapter.

Definition 1.3.3. *A minimal exponential family is **regular** if its canonical parameter space Θ is open.*

In the finite discrete case we are always regular.

← Exercise 1.9.7

Most of the exponential families you will ever encounter are regular. But Θ does not need to be always open. An instance of a non-regular exponential family is given by the inverse Gaussian[7].

[7] See Section 3.2.1 in Rolf Sundberg's book

Proposition 1.3.4. *In a regular exponential family, $A(\theta)$ is smooth and*

$$\nabla A(\theta) = \mathbb{E}_\theta(\mathbf{t}(\mathbf{X})) =: \mu(\theta) \tag{1.9}$$

$$\nabla^2 A(\theta) = \mathrm{var}_\theta(\mathbf{t}(\mathbf{X})) =: V(\theta). \tag{1.10}$$

Taking higher derivatives, we obtain higher cumulants of $\mathbf{t}(\mathbf{X})$.

Proof. Note that

$$K(\boldsymbol{s}) \;:=\; A(\boldsymbol{\theta}_0+\boldsymbol{s}) - A(\boldsymbol{\theta}_0) \;=\; \log \mathbb{E}_{\boldsymbol{\theta}_0} e^{\langle \boldsymbol{s}, \boldsymbol{t}(\mathbf{x})\rangle},$$

which shows that $K(\boldsymbol{s})$ is the cumulant generating function of $\mathbf{t}(X)$ with respect to the distribution $\mathbb{P}_{\boldsymbol{\theta}_0}$. Since $\boldsymbol{\theta}_0$ is an interior point of Θ, $K(\boldsymbol{s})$ is well-defined in a neighbourhood of zero and so all cumulants exist (see, for example, p. 267 in [8]). Using the chain rule, we get that, for any $r \geq 1$, the r-th order derivatives satisfy $\nabla^r_{\boldsymbol{s}} K(\boldsymbol{s}) = \nabla^r_{\boldsymbol{\theta}} A(\boldsymbol{\theta}_0+\boldsymbol{s})$, where the notation ∇^r is hopefully intuitive and it denotes a $d \times \cdots \times d$ array with the $(i_1, \ldots, i_r)$-th entry:

[8] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013

$$(\nabla^r_{\boldsymbol{s}} K(\mathbf{0}))_{i_1 \cdots i_r} = \frac{\partial^r}{\partial s_{i_1} \cdots \partial s_{i_r}} K(\mathbf{0}), \qquad (\nabla^r_{\boldsymbol{\theta}} A(\boldsymbol{\theta}_0))_{i_1 \cdots i_r} = \frac{\partial^r}{\partial \theta_{i_1} \cdots \partial \theta_{i_r}} A(\boldsymbol{\theta}_0).$$

By taking $\boldsymbol{s} = \mathbf{0}$, we obtain

$$\nabla^r_{\boldsymbol{s}} K(\mathbf{0}) \;=\; \nabla^r_{\boldsymbol{\theta}} A(\boldsymbol{\theta}_0), \tag{1.11}$$

The expression on the left in (1.11) gives precisely the r-th order cumulants of $\mathbf{t}(X)$. If $r = 1$ this is $\mathbb{E}_{\boldsymbol{\theta}_0}(\mathbf{t}(X))$ and if $r = 2$ then this is $\mathrm{var}_{\boldsymbol{\theta}_0}(\mathbf{t}(X))$. □

If $A(\boldsymbol{\theta})$ is smooth then $\log f(\mathbf{x};\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{t}(\mathbf{x})\rangle - A(\boldsymbol{\theta})$ is smooth too. In this case the **Fisher information matrix** satisfies

$$I(\boldsymbol{\theta}) \;:=\; -\mathbb{E}_{\boldsymbol{\theta}}(\nabla^2_{\boldsymbol{\theta}} \log f(\mathbf{x};\boldsymbol{\theta})) \;=\; \nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta}).$$

Note that the observed information

$$J(\boldsymbol{\theta}) \;:=\; -\nabla^2_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}) \;=\; I(\boldsymbol{\theta})$$

and so, in particular, it does not depend on the data. We obtain the following result.

Proposition 1.3.5. *In a regular exponential family, the log-likelihood function, given in (1.5), is a smooth and strictly concave function of the canonical parameter $\boldsymbol{\theta}$. The score function $U(\boldsymbol{\theta}) = \nabla \ell_n(\boldsymbol{\theta})$ satisfies*

$$U(\boldsymbol{\theta}) \;=\; \overline{\boldsymbol{\mu}}_n - \mu(\boldsymbol{\theta})$$

and the observed information $J(\boldsymbol{\theta})$ equals the expected (Fisher) information $I(\boldsymbol{\theta})$, and they are both given by the variance of $\mathbf{t}$,

$$I(\boldsymbol{\theta}) \;=\; J(\boldsymbol{\theta}) \;=\; V(\boldsymbol{\theta}).$$

Recall that $\mathcal{X}$ is formally defined as the support of $\mathbb{P}_{\boldsymbol{\theta}}$. Define $\mathcal{K}$ to be the convex hull of the image $\mathbf{t}(\mathcal{X})$:

$$\mathcal{K} \;:=\; \mathrm{conv}(\mathbf{t}(\mathcal{X})). \tag{1.12}$$

For example, for the multivariate Gaussian distribution in Example 1.1.6, $\mathbf{x} \in \mathbb{R}^m$ and so $\mathbf{t}(\mathbf{x}) = -\frac{1}{2}\mathbf{x}\mathbf{x}^\top$ is a rank-one negative semi-definite matrix. In this case, the convex hull of all such matrices, namely $\mathcal{K}$, is the cone of all negative semidefinite matrices.

The next results offers an alternative parametrization for the exponential model.

Proposition 1.3.6. *In a minimal regular exponential family:*

(i) The mapping $\mu : \mathbb{R}^d \to \mathbb{R}^d$ given by $\boldsymbol{\theta} \mapsto \nabla A(\boldsymbol{\theta})$ is one-to-one on Θ.

(ii) The log-likelihood function with data $\overline{\boldsymbol{\mu}} \in \mathcal{K}$ has a maximum in Θ if and only if $\overline{\boldsymbol{\mu}} \in \mu(\Theta)$ and then the maximum $\overline{\boldsymbol{\theta}}$ is given uniquely as $\overline{\boldsymbol{\theta}} = \mu^{-1}(\overline{\boldsymbol{\mu}})$.

(iii) $\mu(\Theta) = \operatorname{int}(\mathcal{K})$ and so in particular $\mu(\Theta)$ is open.

Proof. (i) By Theorem 1.3.2, minimality implies that A is strictly convex. Thus, for every $\boldsymbol{m} \in \mathbb{R}^d$, the function $\langle \boldsymbol{\theta}, \boldsymbol{m} \rangle - A(\boldsymbol{\theta})$ is strictly concave. In particular, it has at most one stationary point in Θ, that is, at most one $\overline{\boldsymbol{\theta}} \in \Theta$ such $\boldsymbol{m} = \nabla A(\overline{\boldsymbol{\theta}})$.

(ii) This is just (i) rephrased.

(iii) To get the first inclusion $\mu(\Theta) \subseteq \operatorname{int}(\mathcal{K})$, we first show that $\mu(\Theta) \subseteq \overline{\mathcal{K}}$, where the latter denotes the closure of $\mathcal{K}$. For every $\boldsymbol{\theta} \in \Theta$ and for every $c \in \mathbb{R}$, if there exists $\boldsymbol{u} \in \mathbb{R}^d$ such that $\langle \boldsymbol{u}, \mathbf{t}(\mathbf{X}) \rangle \leq c$ almost surely ($\mathbf{t}(\mathcal{X})$ is contained in the given half-space) then $\langle \boldsymbol{u}, \mathbb{E}_{\boldsymbol{\theta}}(\mathbf{t}(\mathbf{X})) \rangle \leq c$ for every $\boldsymbol{\theta} \in \Theta$. This implies that if a half-space H contains $\mathcal{K}$ then it also contains $\mu(\Theta)$. The intersection of all such halfplanes is equal to $\overline{\mathcal{K}}$; this is a standard application of the Hyperplane Separation Theorem B.1.3. This shows that $\mu(\Theta) \subseteq \overline{\mathcal{K}}$. To prove that $\mu(\Theta) \subseteq \operatorname{int}(\mathcal{K})$, we argue by contradiction. Suppose that $\boldsymbol{\mu}(\boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \Theta$ lies in the boundary of $\mathcal{K}$, $\partial\mathcal{K} = \overline{\mathcal{K}} \setminus \operatorname{int}(\mathcal{K})$. Then, again by the Hyperplane Separation Theorem, there exists a closed halfplane $H = \{\boldsymbol{\mu} : \langle \boldsymbol{u}, \boldsymbol{\mu} \rangle \leq c\}$ such that $\operatorname{int}(\mathcal{K}) \subseteq H$ but $\boldsymbol{\mu}(\boldsymbol{\theta}_0) \in \partial H$. In particular,

← Exercise B.1.4

$$Z := \left\langle \boldsymbol{u}, \mathbf{t}(\mathbf{X}) - \boldsymbol{\mu}(\boldsymbol{\theta}_0) \right\rangle \leq 0 \quad \text{almost surely.}$$

Note however that, since $\mathbb{E}_{\boldsymbol{\theta}} \boldsymbol{t}(X) = \mu(\boldsymbol{\theta})$, $\mathbb{E}_{\boldsymbol{\theta}_0} Z = 0$, which implies that $Z = 0$ $\mathbb{P}_{\boldsymbol{\theta}_0}$-almost surely. This however contradicts minimality of the exponential family. We conclude that $\mu(\Theta) \subseteq \operatorname{int}(\mathcal{K})$.

It remains to show that the opposite inclusion $\operatorname{int}(\mathcal{K}) \subseteq \mu(\Theta)$ also holds. We again argue by contradiction. Let $\mathbf{t}_0 \in \operatorname{int}(\mathcal{K}) \setminus \mu(\Theta)$ then the equation $\mathbf{t}_0 = \nabla A(\boldsymbol{\theta})$ has no solution, or equivalently, the log-likelihood function $\ell(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{t}_0 \rangle - A(\boldsymbol{\theta})$ is not bounded above on Θ. To get a contradiction, fix any $\boldsymbol{\theta}_0 \in \Theta$ and consider half-lines

$$L_{\boldsymbol{u}} := \{\boldsymbol{\theta}_\lambda = \boldsymbol{\theta}_0 + \lambda \boldsymbol{u} : \lambda \geq 0\}.$$

Now $L_{\boldsymbol{u}}$ can be either not entirely contained in Θ (case 1) or it can be contained in Θ (case 2). We will show that over each $L_{\boldsymbol{u}}$ the value of $-\ell(\boldsymbol{\theta}_\lambda)$ goes to infinity as $\lambda \to \infty$ (in case 1) or as we approach the boundary of Θ (in case 2). In consequence, maximizing ℓ over Θ can be reduced to a compact subset $\{\boldsymbol{\theta} : \ell(\boldsymbol{\theta}) \geq \ell(\boldsymbol{\theta}_0)\}$ and thus the optimum must exist. But then it must be of the form $\mathbf{t}_0 = \nabla A(\hat{\boldsymbol{\theta}})$ leading to a contradiction.

Case 1: Since the half-line $L_{\boldsymbol{u}}$ is not contained in Θ, for some λ_0 the point $\boldsymbol{\theta}_{\lambda_0}$ lies on the boundary of Θ. Since Θ is open, $A(\boldsymbol{\theta}_{\lambda_0}) = \infty$ and hence $\ell(\boldsymbol{\theta}_{\lambda_0}) = -\infty$. Thus $-\ell(\boldsymbol{\theta}_\lambda) \to \infty$ as $\lambda \to \lambda_0$.

Case 2: In this case we can take $\lambda \to \infty$ without leaving Θ. We will still show that $-\ell(\boldsymbol{\theta}_\lambda) \to \infty$. Note that

$$e^{-\ell(\boldsymbol{\theta}_\lambda)} = Z(\boldsymbol{\theta}_\lambda) e^{-\langle \boldsymbol{\theta}_\lambda, \mathbf{t}_0 \rangle} = \int_{\mathcal{X}} h(\mathbf{x}) e^{\lambda \langle \boldsymbol{u}, \mathbf{t}(\mathbf{x}) - \mathbf{t}_0 \rangle} e^{\langle \boldsymbol{\theta}_0, \mathbf{t}(\mathbf{x}) - \mathbf{t}_0 \rangle} \mu(\mathrm{d}\mathbf{x}).$$

Denote the integrand by $I(\lambda)$. Define

$$\begin{aligned} A_+ &= \{\mathbf{x} \in \mathcal{X} : \langle \boldsymbol{u}, \mathbf{t}(\mathbf{x}) - \mathbf{t}_0 \rangle > 0\}, \\ A_0 &= \{\mathbf{x} \in \mathcal{X} : \langle \boldsymbol{u}, \mathbf{t}(\mathbf{x}) - \mathbf{t}_0 \rangle = 0\}, \\ A_- &= \{\mathbf{x} \in \mathcal{X} : \langle \boldsymbol{u}, \mathbf{t}(\mathbf{x}) - \mathbf{t}_0 \rangle < 0\}, \end{aligned}$$

and note that A_+, A_0, A_- form a partition of $\mathcal{X}$. The integrand $I(\lambda)$ is always non-negative and it is increasing in λ on A_+, constant on A_0, and decreasing on A_-. The *monotone convergence theorem* assures that we can pass with the limit of λ inside the integral, namely

$$\lim_{\lambda \to \infty} \int_{A_-} I(\lambda) \mu(\mathrm{d}\mathbf{x}) = 0,$$

$$\lim_{\lambda \to \infty} \int_{A_0} I(\lambda) \mu(\mathrm{d}\mathbf{x}) = \int_{A_0} e^{\langle \boldsymbol{\theta}_0, \mathbf{t}(\mathbf{x}) - \mathbf{t}_0 \rangle} \mu(\mathrm{d}\mathbf{x}) < \infty,$$

and

$$\lim_{\lambda \to \infty} \int_{A_+} I(\lambda) \mu(\mathrm{d}\mathbf{x}) = \infty.$$

Unless A_+ has measure zero, we conclude that $\lim_{\lambda \to \infty} e^{-\ell(\boldsymbol{\theta}_\lambda)} = \infty$ or equivalently $\lim_{\lambda \to \infty} -\ell(\boldsymbol{\theta}_\lambda) = \infty$.

To conclude the proof, it remains to show that A_+ has positive measure. If the measure is zero, $\langle \boldsymbol{u}, \mathbf{t}(\mathbf{x}) \rangle \leq \langle \boldsymbol{u}, \mathbf{t}_0 \rangle$ for all $\mathbf{x} \in \mathcal{X}$. We will again show that there must be equality, which contradicts minimality. This is where we use the fact that $\mathbf{t}_0$ is an interior point of $\mathcal{K}$. Let $\mathbf{x}_1 \in \mathcal{X}$ and consider the half-line from $\mathbf{t}(\mathbf{x}_1)$ through $\mathbf{t}_0$. If this half-line crosses $\mathbf{t}(\mathcal{X})$ at some other point $\mathbf{t}(\mathbf{x}_2)$ after crossing $\mathbf{t}_0$, we can write $\mathbf{t}_0$ as a convex combination of $\mathbf{t}(\mathbf{x}_1)$ and $\mathbf{t}(\mathbf{x}_2)$, which implies that $\langle \boldsymbol{u}, \mathbf{t}(\mathbf{x}_1) \rangle = \langle \boldsymbol{u}, \mathbf{t}(\mathbf{x}_2) \rangle = \langle \boldsymbol{u}, \mathbf{t}_0 \rangle$. If this half-line does not contain any other point in $\mathbf{t}(\mathcal{X})$, this whole half-line must be contained in $\mathrm{int}(\mathcal{K})$. Take any other point $\mathbf{t}(\mathbf{x}_2) \in \mathrm{int}(\mathcal{K})$ on the half-line after crossing $\mathbf{t}_0$. By definition $\mathbf{t}(\mathbf{x}_2)$ is a convex combination of

some finitely many points in $\mathbf{t}(\mathcal{X})$, which allows as to write $\mathbf{t}_0$ as a convex combination of $\mathbf{t}(\mathbf{x}_1)$ and some other points in $\mathbf{t}(\mathcal{X})$. We again conclude that $\langle \boldsymbol{u}, \mathbf{t}(\mathbf{x}_1)\rangle = \langle \boldsymbol{u}, \mathbf{t}_0\rangle$. In this way we showed that for an arbitrary $\mathbf{x} \in \mathcal{X}$, $\langle \boldsymbol{u}, \mathbf{t}(\mathbf{x})\rangle = \langle \boldsymbol{u}, \mathbf{t}_0\rangle$, which contradicts the minimality of our exponential family. □

Proposition 1.3.6 shows that $\boldsymbol{\mu}$ can be used as an alternative parametrization of the exponential family. For example, in the mean zero Gaussian distribution the mean parametrization is given by the covariance matrix Σ (or more precisely by $-\frac{1}{2}\Sigma$). In the next section, we discuss a whole range of suitable parametrizations.

1.4 *Marginal and conditional distributions**

We will consider partitioning of the sufficient statistics $\mathbf{t}$ into $\boldsymbol{u}$ and $\boldsymbol{v}$, $\mathbf{t} = (\boldsymbol{u}, \boldsymbol{v})$ with the corresponding partition of $\boldsymbol{\theta} = (\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_u, \boldsymbol{\mu}_v)$. We consider two basic examples.

Example 1.4.1. *We have shown that $X \sim N(\mu, \sigma)$ forms an exponential family with $\mathbf{t}(x) = (x, -x^2/2)$, $\boldsymbol{\theta} = (\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2})$ and $\boldsymbol{\mu} = (\mu, -\frac{1}{2}(\mu^2 + \sigma^2))$. Given a sample x_i for $i - 1, \ldots, n$, we get an exponential family with the same canonical parameter and the sufficient statistics $\mathbf{t}(\mathbf{x}_{1:n}) = (\sum_i x_i, -\frac{1}{2}\sum_i x_i^2)$ and the mean parameter $(n\mu, -\frac{n}{2}(\mu^2 + \sigma^2))$. Here we could take $\boldsymbol{u}(x) = \sum_i x_i$ and $\boldsymbol{v}(x) = -\frac{1}{2}\sum_i x_i^2$ or the other way around.*

Example 1.4.2. *In the multivariate Gaussian case we have shown that $\mathbf{t}(\mathbf{x}) = -\frac{1}{2}\mathbf{x}\mathbf{x}^\top$, $\boldsymbol{\theta} = K$, $\boldsymbol{\mu} = -\frac{1}{2}\Sigma$. Fix any subset $E \subset \{(i,i) : i = 1, \ldots, m\} \cup \{(i,j) : 1 \leq i < j \leq m\}$. This corresponds to fixing some entries of a symmetric matrix. We could take $\boldsymbol{u} = -\frac{1}{2}(x_i x_j)_{ij \in E}$ and $\boldsymbol{v} = -\frac{1}{2}(x_i x_j)_{ij \notin E}$.*

Recall the formula for the distribution of $\mathbf{t}(\mathbf{X})$ as given in Proposition 1.2.7.

Proposition 1.4.3 (Marginal distribution)**.** *In a regular exponential family with $\mathbf{t} = (\boldsymbol{u}, \boldsymbol{v})$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)$ the marginal model for $\boldsymbol{u}$ is a regular exponential family for each given $\boldsymbol{\theta}_v$, depending on $\boldsymbol{\theta}_v$ but with the same parameter space for its mean value parameter $\boldsymbol{\mu}_u$.*

Proof. The marginal distribution for $\boldsymbol{u}$ is obtained by integrating $\boldsymbol{v}$ out:

$$\begin{aligned} f(\boldsymbol{u}; \boldsymbol{\theta}) &= \int g(\boldsymbol{u}, \boldsymbol{v}) \exp\left\{\langle \boldsymbol{\theta}_u, \boldsymbol{u}\rangle + \langle \boldsymbol{\theta}_v, \boldsymbol{v}\rangle - A(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)\right\} \mathrm{d}\boldsymbol{v} \\ &= \exp\left\{\langle \boldsymbol{\theta}_u, \boldsymbol{u}\rangle - A(\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)\right\} \left(\int g(\boldsymbol{u}, \boldsymbol{v}) \exp\{\langle \boldsymbol{\theta}_v, \boldsymbol{v}\rangle\} \mathrm{d}\boldsymbol{v}\right). \end{aligned}$$

For any fixed $\boldsymbol{\theta}_v$ this has the form of a regular exponential family. This exponential family has canonical parameter $\boldsymbol{\theta}_u$ but the space of

canonical parameters will typically depend on $\boldsymbol{\theta}_v$ (it is an intersection of Θ with $\boldsymbol{\theta}_v = $ fixed). However, by Proposition 1.3.6, the mean parameter space is always the same and equal to the interior of the convex hull of $\boldsymbol{u}(\mathcal{X})$, which is equal to the projection of $\mu(\Theta)$ on the $\boldsymbol{u}$ coordinates (i.e. $\boldsymbol{\mu}_u$). □

Proposition 1.4.4 (Conditional distribution)**.** *With the same setting as in Proposition 1.4.3, the conditional model for $\mathbf{x}$ given $\boldsymbol{u}$ (and thus also for $\boldsymbol{v}$ given $\boldsymbol{u}$) is a regular exponential family with canonical statistics $\boldsymbol{v}$. The conditional model depends on $\boldsymbol{u}$ but with one and the same canonical parameter $\boldsymbol{\theta}_v$ as in the joint model.*

Proof. We have

$$f(\mathbf{x}|\boldsymbol{u};\boldsymbol{\theta}) = \frac{f(\boldsymbol{u},\mathbf{x};\boldsymbol{\theta})}{f(\boldsymbol{u};\boldsymbol{\theta})} = \frac{h(\mathbf{x})\exp\{\langle\boldsymbol{\theta}_v,\boldsymbol{v}(\mathbf{x})\}}{\int g(\boldsymbol{u},\boldsymbol{v})\exp\{\langle\boldsymbol{\theta}_v,\boldsymbol{v}(\mathbf{x})\}\mathrm{d}\boldsymbol{v}}.$$

For the fixed value of $\boldsymbol{u}$, the expression in the denominator does not depend on $\mathbf{x}$ but only on $\boldsymbol{\theta}_v$ and so it represents the normalizing constant of this distribution. Note $f(\mathbf{x}|\boldsymbol{u};\boldsymbol{\theta})$ is defined only for those $\mathbf{x}$ for which $\mathbf{t}(\mathbf{x}) = (\boldsymbol{u}(\mathbf{x}),\boldsymbol{v}(\mathbf{x})) \in \mathbf{t}(\mathcal{X})$, and thus the space of sufficient statistics depends on $\boldsymbol{u}$. However, the canonical parameter is the same, $\boldsymbol{\theta}_v$ in the projection Θ_v of Θ on the coordinates $\boldsymbol{\theta}_v$. To get $f(\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\theta})$ we only substitute $h(\mathbf{x})$ for $g(\boldsymbol{u},\boldsymbol{v})$ above but otherwise the argument is the same. □

Explicit calculations with these marginal and conditional distributions are typically hard but the two results above are important in guiding our analysis.

Example 1.4.5. *Consider the univariate Gaussian example discussed in Example 1.4.1 and the induced distribution of the sample $x_{1:n}$. Let $u = \sum_i x_i$ and $v = -\frac{1}{2}\sum_i x_i^2$ and recall that the distribution has canonical parameters $(\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2})$. The marginal distribution of $\sum_i x_i^2$ is proportional to a noncentral χ^2 and does not in general form an exponential family unless we fix the value of $\frac{\mu}{\sigma^2}$. Next, consider instead its conditional distribution given $\sum_i x_i = n\bar{x}$. This may appear quite complicated but Proposition 1.4.4 suggests that it may be still tractable. Given $\bar{x}$, $\sum_i x_i^2$ differs only by an additive constant from $\sum x_i^2 - n\bar{x}^2 = (n-1)s^2$, where we used the standard notation*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1.13}$$

Thus, it is enough to characterize the distribution of $(n-1)s^2$ given $\bar{x}$. It is well-known that s^2 is independent of $\bar{x}$, and that the distribution of $(n-1)s^2$ is proportional (by σ^2) to a (central) χ^2_{n-1}. From the explicit form of a χ^2 it is easily seen that the conditional distribution forms an exponential family.

← Exercise 1.9.8

← Exercise 1.9.8

As a corollary of the above two results we obtain a useful result on a range of possible alternative parametrizations. For a given split $\mathbf{t} = (\boldsymbol{u}, \boldsymbol{v})$ consider the vector $(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)$ with $\boldsymbol{\mu}_u \in \mu_u(\Theta)$ and $\boldsymbol{\theta}_v \in \Theta_v$. Here by Θ_v we denote the projection of Θ on the coordinates $\boldsymbol{\theta}_v$ and by $\mu_u(\Theta)$ we mean the projection of $\mu(\Theta)$ on the coordinates $\boldsymbol{\mu}_u$.

Proposition 1.4.6. *The mixed parametrization $(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)$ is a valid parametrization with the parameter space $\mu_u(\Theta) \times \Theta_v$ (variational independence!). The Fisher information for $(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)$ is*

$$I(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v) = \begin{bmatrix} (\Sigma_{\boldsymbol{uu}})^{-1} & \mathbf{0} \\ \mathbf{0} & ((\Sigma^{-1})_{\boldsymbol{vv}})^{-1} \end{bmatrix},$$

where $\Sigma = \operatorname{var}(\mathbf{t})$ and $\Sigma_{\boldsymbol{uu}} = \operatorname{var}(\boldsymbol{u})$. The same formula holds for the observed information in the MLE, $J(\overline{\boldsymbol{\mu}}_u, \overline{\boldsymbol{\theta}}_v)$.

Proof. Fix an exponential family with canonical statistics $\mathbf{t}(\mathbf{x})$ and canonical parameter $\boldsymbol{\theta} \in \Theta$. By Proposition 1.2.7, the distribution of $\mathbf{t}(\mathbf{X})$ is an exponential family with the same canonical parameter. This distribution is uniquely defined by the marginal distribution of $\boldsymbol{u}$ and the conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{u}$. The latter forms an exponential family with canonical parameter $\boldsymbol{\theta}_v \in \Theta_v$ by Proposition 1.4.4. Now fix $\boldsymbol{\theta}_v$, which corresponds to fixing the conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{u}$. By Proposition 1.4.3, the marginal distribution of $\boldsymbol{u}$ is an exponential family with the mean parameter $\boldsymbol{\mu}_u$. By Proposition 1.3.6 the range of this mean parameter is the interior of the convex hull of $\boldsymbol{u}(\mathcal{X})$ (independent on $\boldsymbol{\theta}_v$). This is precisely the projection of $\boldsymbol{\mu}(\Theta)$ on $\boldsymbol{\mu}_u$ and this shows that the map $\boldsymbol{\theta} \mapsto (\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)$ is one-to-one with range $\mu_u(\Theta) \times \Theta_v$. For the proof of the second statement see [9]. □

[9] Rolf Sundberg. *Statistical modelling by exponential families*, volume 12 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, Cambridge, 2019

← Exercise 1.9.9

← Exercise 1.9.10

Example 1.4.7. *Consider the multivariate Gaussian distribution in Example 1.1.6 with $m = 2$. We have $K = \Sigma^{-1}$, that is,*

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{bmatrix} = \frac{1}{\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2} \begin{bmatrix} \Sigma_{22} & -\Sigma_{12} \\ -\Sigma_{12} & \Sigma_{11} \end{bmatrix}.$$

The canonical parameters are (K_{11}, K_{22}, K_{12}) and the mean parameters are $-\frac{1}{2}(\Sigma_{11}, \Sigma_{22}, \Sigma_{12})$. The constraints on the canonical parameters are $K_{11} > 0$, $K_{22} > 0$, $K_{11}K_{22} > K_{12}^2$ (namely, K is positive definite). The constraints on the mean parameter follow from the constraints on Σ: $\Sigma_{11} > 0$, $\Sigma_{22} > 0$, $\Sigma_{11}\Sigma_{22} > \Sigma_{12}^2$ (Σ is positive definite). Consider a mixed parametrization $(-\frac{1}{2}\Sigma_{11}, -\frac{1}{2}\Sigma_{22}, K_{12})$. The projection $\mu_u(\Theta)$ is simply $(-\infty, 0)^2$. The projection, Θ_v is the whole real line $\mathbb{R}$ (irrespective of the value of K_{12} we can set K_{11}, K_{22} big enough for K to be positive definite).

By Proposition 1.4.6, for every choice of $\Sigma_{11} > 0, \Sigma_{22} > 0$ *and* $K_{12} \in R$, *there will be a unique positive definite matrix* Σ *with the prescribed diagonal entries* Σ_{11}, Σ_{22} *and such that* $-\frac{1}{\Sigma_{11}\Sigma_{22}-\Sigma_{12}^2}\Sigma_{12} = K_{12}$. *In this case, this can be checked directly. To compute the corresponding* Σ_{12}, *we need to solve the quadratic equation*

$$K_{12}\Sigma_{12}^2 - \Sigma_{12} - K_{12}\Sigma_{11}\Sigma_{22} = 0.$$

There are two real solutions but only one of them, namely,

$$\Sigma_{12} = \frac{1 - \sqrt{1 + 2K_{12}^2\Sigma_{11}\Sigma_{22}}}{2K_{12}}$$

results in a positive definite Σ.

1.5 *Conditional inference for canonical parameter**

Suppose $\boldsymbol{\psi}$ is the parameter of interest, where $(\boldsymbol{\lambda}, \boldsymbol{\psi})$ is a transformation of $\boldsymbol{\theta}$. For simplicity we focus on the case when $\boldsymbol{\psi} = \boldsymbol{\theta}_v$ and $\boldsymbol{\lambda} = \boldsymbol{\mu}_u$ is regarded as nuisance parameter. As shown in Proposition 1.4.6, $\boldsymbol{\lambda} = \boldsymbol{\mu}_u = \mathbb{E}_{\boldsymbol{\theta}}(\boldsymbol{u})$ is the preferable nuisance parameter (rather than $\boldsymbol{\theta}_u$), since $\boldsymbol{\theta}_v$ and $\boldsymbol{\mu}_u$ are variation independent and information orthogonal.

Proposition 1.5.1 (Conditionality principle for full families)**.** *Statistical inference about the canonical parameter component* $\boldsymbol{\theta}_v$ *in presence of the nuisance parameter* $\boldsymbol{\lambda} = \boldsymbol{\mu}_u = \mathbb{E}_{\boldsymbol{\theta}}(\boldsymbol{u})$ *should be made conditional on* $\boldsymbol{u}$, *that is, the conditional model for* $\mathbf{x}$ *or* $\boldsymbol{v}$ *given* $\boldsymbol{u}$.

This is only a recommendation so rather than providing a formal proof we motivate this statement informally.

Motivation. The likelihood for $(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)$ factorizes as

$$L(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v; \boldsymbol{t}) = L_1(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v; \boldsymbol{u}) L_2(\boldsymbol{\theta}_v; \boldsymbol{v}|\boldsymbol{u}) \tag{1.14}$$

where the two parameters are variation independent. In some cases L_1 depends only on $\boldsymbol{\mu}_u$, in which case

$$L(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v; \boldsymbol{t}) = L_1(\boldsymbol{\mu}_u; \boldsymbol{u}) L_2(\boldsymbol{\theta}_v; \boldsymbol{v}|\boldsymbol{u}).$$

Then it is clear that there is no information about $\boldsymbol{\theta}_v$ in the first factor L_1 and the argument for the principle is compelling.

However, even when L_1 depends on $\boldsymbol{\theta}_v$, there is really no information about $\boldsymbol{\theta}_v$ in $\boldsymbol{u}$. Indeed, note first that $\boldsymbol{u}$ and $\boldsymbol{\mu}_u$ have the same dimension, and that $\boldsymbol{u}$ serves as an estimator (the MLE) of $\boldsymbol{\mu}_u$, whatever is the value of $\boldsymbol{\theta}_v$. This means that the information in $\boldsymbol{u}$ about$(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)$

is totally consumed in the estimation of $\boldsymbol{\mu}_u$. Furthermore, the estimated value of $\boldsymbol{\mu}_u$ does not provide any information about $\boldsymbol{\theta}_v$, and $\boldsymbol{\mu}_u$ would not do so even if it were known, due to the variation independence between $\boldsymbol{\mu}_u$ and $\boldsymbol{\theta}_v$. Thus, the first factor L_1 contributes only information about $\boldsymbol{\mu}_u$. □

Example 1.5.2 (Conditional independence for a Gaussian sample)**.** *Suppose we want to make inference about σ^2, or σ. Then we are led to consider the conditional distribution of $\sum_i x_i^2$, given $\bar{x}$, that depends on σ alone (c.f. Example 1.4.5 and Proposition 1.4.4). The marginal distribution of $\bar{x}$ depends on both μ and σ^2, so the joint and conditional likelihoods are different functions of σ^2.*

As we have seen in Example 1.4.5, in the conditional approach $n\bar{x}^2$ is a constant, and after subtraction of this empirical constant from $\sum_i x_i^2$ we are led to the use of the statistic s^2. Now, we already know that $\bar{x}$ and s^2 are independent, so the even simpler result is that the inference should be based on the marginal model for $(n-1)s^2/\sigma^2$, with its χ^2_{n-1}-distribution. In particular this leads to the conditional and marginal ML estimator $\hat{\sigma}^2 = s^2$, which differs by the factor $n/(n-1)$ from the MLE in the joint model (with denominator n).

1.6 Kullback-Leibler divergence

The *Fenchel conjugate* of the cumulant function A is the function

$$A^*(\mathbf{t}) \;=\; \sup\{\langle \boldsymbol{\theta}, \mathbf{t}\rangle - A(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^d\}.$$

For regular exponential families $A^*(\mathbf{t}) < \infty$ if and only if $\mathbf{t} \in \mu(\Theta)$. The function A^* is convex as a supremum of linear functions[10] and, in fact, strictly convex. If $\mathbf{t} \in \mu(\Theta)$, the unique optimizer of the log-likelihood is $\theta(\mathbf{t})$, where $\theta : M \to \Theta$ is the inverse of the map $\mu : \Theta \to M$; see Proposition 1.3.6. It follows that, for $\mathbf{t} \in \mu(\Theta)$,

[10] see Proposition B.2.5.

$$A^*(\mathbf{t}) \;=\; \langle \theta(\mathbf{t}), \mathbf{t}\rangle - A(\theta(\mathbf{t})) \tag{1.15}$$

or alternatively, for $\boldsymbol{\theta} \in \Theta$,

$$A^*(\mu(\boldsymbol{\theta})) \;=\; \langle \boldsymbol{\theta}, \mu(\boldsymbol{\theta})\rangle - A(\boldsymbol{\theta}) \tag{1.16}$$

implying in particular that A^* is smooth, since μ and A are both smooth. By composite differentiation in (1.15), since $\mu(\theta(\mathbf{t})) = \mathbf{t}$, we obtain

$$\nabla A^*(\mathbf{t}) = \theta(\mathbf{t}) + \nabla\theta(\mathbf{t}) \cdot \mathbf{t} - \nabla\theta(\mathbf{t}) \cdot \mu(\theta(\mathbf{t})) = \theta(\mathbf{t}),$$

where the gradient is taken with respect to $\mathbf{t}$.

For two distributions over some state-space $\mathcal{X}$ with densities p, q with respect to some measure μ, the **Kullback-Leibler divergence** $\mathrm{K}(p, q)$ is defined as

$$\mathrm{K}(p,q) \;:=\; \int_{\mathcal{X}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \mu(\mathrm{d}\mathbf{x}).$$

The following result is well-known and it is a direct application of the Jensen's inequality stated formally in Theorem B.2.3.

Proposition 1.6.1. *We have $K(p,q) \geq 0$ with equality if and only if $p = q$ almost surely.*

Proof. We use the fact that $-\log y$ is a strictly convex function. By Theorem B.2.3,

$$0 \;=\; -\log \mathbb{E}_p \frac{q(X)}{p(X)} \;\leq\; -\mathbb{E}_p \log \frac{q(X)}{p(X)} \;=\; \mathbb{E}_p \log \frac{p(X)}{q(X)} \;=\; \mathrm{K}(p,q)$$

with equality if and only if $q(X)/p(X)$ is constant almost surely. Since p, q are both densities, this is possible if and only if they are equal almost surely. □

Given two distributions $\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}$ in the given exponential family we write the corresponding Kullback-Leibled divergence as $K(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1)$. We easily check that

$$K(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \;=\; A(\boldsymbol{\theta}_2) - A(\boldsymbol{\theta}_1) - \langle \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \boldsymbol{\mu}_1 \rangle. \tag{1.17}$$

Since $\boldsymbol{\mu}_1 = \nabla A(\boldsymbol{\theta}_1)$, (1.17) has another interpretation as the Bregman divergence (defined by the function A between $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_1$. Look this up!

Proposition 1.6.2. *Consider two distributions in the exponential family (1.1), one with the mean parameter $\boldsymbol{\mu}_1 \in \mu(\Theta)$ and the other with canonical parameter $\boldsymbol{\theta}_2 \in \Theta$. If this exponential family is regular, the* Kullback–Leibler divergence *between these two distributions is*

$$\mathrm{K}(\boldsymbol{\mu}_1, \boldsymbol{\theta}_2) \;=\; -\langle \boldsymbol{\mu}_1, \boldsymbol{\theta}_2 \rangle + A^*(\boldsymbol{\mu}_1) + A(\boldsymbol{\theta}_2). \tag{1.18}$$

The Kullback–Leibler divergence is well defined and nonnegative over $\mu(\Theta) \times \Theta$. Moreover, $\boldsymbol{K}(\boldsymbol{\mu}_1, \boldsymbol{\theta}_2) = 0$ if and only if $\boldsymbol{\mu}_1 = \mu(\boldsymbol{\theta}_2)$.

Proof. We leave it as an exercise. □ ← Exercise 1.9.11

The reason to express the Kullback–Leibler distance in terms of $\boldsymbol{\mu}_1$ and $\boldsymbol{\theta}_2$ rather than $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ (as usually done in the literature) is that we wish to exploit the following basic result.

Proposition 1.6.3. *The Kullback–Leibler divergence $\mathrm{K}(\boldsymbol{\mu}_1, \boldsymbol{\theta}_2)$ is strictly convex both in $\boldsymbol{\mu}_1$ and in $\boldsymbol{\theta}_2$.*

Proof. This follows directly from (1.18) and the fact that both $A(\boldsymbol{\theta})$ and $A^*(\boldsymbol{\mu})$ are strictly convex functions □

This set-up has been exploited in various places. See, for example, Section 5 in [11]. Another important application is in situations when a statistical submodel is given by affine restrictions on the mean parameter. Note that the MLE can be equivalently defined as the minimizer of $\mathrm{K}(\overline{\boldsymbol{\mu}}_n, \boldsymbol{\theta})$. We have a parallel definition, when the dual MLE is given as the minimizer of $K(\boldsymbol{\mu}, \theta(\overline{\boldsymbol{\mu}}_n))$.

[11] Martin J Wainwright and Michael I. Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008

Example 1.6.4 (Behrens-Fisher problem)**.** *The Behrens-Fisher problem is concerned with testing the difference between the means of two normally distributed populations when the variances of the two populations are not assumed to be equal, based on two independent samples. Since the hypothesis is linear in the mean parameter, this problem can be addressed with the dual MLE; see* [12] *for details.*

[12] E Susanne Christensen. Statistical properties of I-projections within exponential families. *Scandinavian Journal of Statistics*, pages 307–318, 1989

We finish this chapter with one of the most fundamental results motivating exponential families. The maximum entropy principle states that under uncertainty, one should take a model which maximizes the entropy subject to constraints on the known features about the system. We show that the exponential family arises naturally if the constraint is given by the expected value of some statistics.

Recall that for a distribution $\mathbb{P}$ that admits a density function $p(x)$ with respect to the base measure μ, the entropy $\mathrm{H}_{\mathbb{P}}$ of $\mathbb{P}$ is

$$\mathrm{H}_{\mathbb{P}} \;=\; -\mathbb{E}_{\mathbb{P}} \log p(X) \;=\; -\int \log p(x) p(x) \mu(\mathrm{d}x).$$

For notational simplicity, in what follows consider the exponential family (1.2), where the function $h(\mathbf{x})$ has been incorporated into the base measure μ. In this case each distribution in this family has the same support, which is equal to the support of μ. For such an exponential family $\mathbb{P}_\theta$ we have

$$\mathrm{H}_{\mathbb{P}_{\boldsymbol{\theta}}} \;=\; -\langle \boldsymbol{\theta}, \mu(\boldsymbol{\theta}) \rangle + A(\boldsymbol{\theta}) = -A^*(\mu(\boldsymbol{\theta})), \tag{1.19}$$

where the second equality follows from (1.16).

Consider the problem of maximizing the entropy for *all* distributions $\mathbb{P}$ that admit a density function absolutely continuous with respect to μ and $\mathbb{E}_{\mathbb{P}}(\boldsymbol{t}(X)) = \boldsymbol{t}_0$

$$\text{maximize } \mathrm{H}_{\mathbb{P}} \quad \text{s.t. } \mathbb{P} \sim \mu, \quad \mathbb{E}_{\mathbb{P}} \boldsymbol{t}(X) = \boldsymbol{t}_0. \tag{1.20}$$

The following result provides an important characterization of exponential families.

Theorem 1.6.5. *Consider the exponential family* (1.2)*. Suppose there exists* $\boldsymbol{\theta}_0 \in \Theta$ *such that* $\nabla A(\boldsymbol{\theta}_0) = \boldsymbol{t}_0$*. Then for any distribution* $\mathbb{P}$*, which satisfies condition of* (1.20)*, we have*

$$H_{\mathbb{P}_{\boldsymbol{\theta}_0}} - H_{\mathbb{P}} \;=\; \mathrm{K}(\mathbb{P}, \mathbb{P}_{\boldsymbol{\theta}_0}).$$

Thus, $\mathbb{P}_{\boldsymbol{\theta}_0}$ *is the unique solution to* (1.20)*.*

Proof. Note that $\nabla A(\boldsymbol{\theta}_0) = \boldsymbol{t}_0$ is equivalent to $\mu(\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\mathbf{t}(X) = \mathbf{t}_0$. Let $p(\mathbf{x})$ be the density of $\mathbb{P}$ with respect to μ. Consider

$$\begin{aligned}
\mathrm{K}(\mathbb{P}, \mathbb{P}_{\boldsymbol{\theta}_0}) &= \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{f(\mathbf{x};\boldsymbol{\theta}_0)} \mu(\mathrm{d}\mathbf{x}) \\
&= -H_{\mathbb{P}} - \int \Big(\langle \boldsymbol{\theta}_0, \mathbf{t}(\mathbf{x}) \rangle - A(\boldsymbol{\theta}_0) \Big) p(x) \mu(\mathrm{d}\mathbf{x}) \\
&= -H_{\mathbb{P}} - \langle \boldsymbol{\theta}_0, \mathbf{t}_0 \rangle + A(\boldsymbol{\theta}_0) \\
&\overset{(1.19)}{=} H_{\mathbb{P}_{\boldsymbol{\theta}_0}} - H_{\mathbb{P}}.
\end{aligned}$$

By Proposition 1.6.1, this shows that $H_{\mathbb{P}} \leq H_{\mathbb{P}_{\boldsymbol{\theta}_0}}$ with equality if and only if $\mathbb{P} = \mathbb{P}_{\boldsymbol{\theta}_0}$. This concludes the proof. □

Example 1.6.6. *Consider all distributions with the support $\mathbb{R}^m$ and with the property that $\mathbb{E}X = \mu$, $\mathbb{E}(XX^\top) = \Sigma + \mu\mu^\top$. Among all such distributions, the multivariate normal $N(\mu, \Sigma)$ is the one that maximizes the entropy.*

1.7 Generalized Linear Models

The generalized linear models are formulated based on the construction of exponential families. Here we provide only a basic treatment that explains the origin of the construction and the most important examples.

Consider the pairs $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, where the input $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ are considered fixed and the outputs $y_1, \ldots, y_n$ are independent observations each from density

$$f(y_i; \mathbf{x}_i, \boldsymbol{w}, \sigma^2) = h(y_i, \sigma^2) \exp\left\{ \tfrac{1}{\sigma^2} \left(y_i \theta_i(\boldsymbol{w}) - A(\theta_i(\boldsymbol{w})) \right) \right\}.$$

where $\theta_i(\boldsymbol{w}) = \mathbf{x}_i^\top \boldsymbol{w}$ and σ^2 is called the dispersion term. From the standard theory, we get immediately that

$$\mu(\theta_i) := \mathbb{E}[Y_i | \mathbf{x}_i, \boldsymbol{w}, \sigma^2] = A'(\theta_i).$$

Indeed, the standard result show that for a fixed σ^2, $\mathbb{E}(\frac{1}{\sigma^2} Y) = \frac{1}{\sigma^2} A'(\boldsymbol{\theta})$. In the same way, we argue that

$$V(\theta, \sigma^2) := \mathrm{var}_\theta(Y_i | \mathbf{x}_i, \boldsymbol{w}, \sigma^2) = \sigma^2 A''(\theta_i).$$

From now on we fix σ^2 and with no loss of generality we take $\sigma^2 = 1$. The log-likelihood function is

$$\ell_n(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^n \log f(y_i; \mathbf{x}_i, \boldsymbol{w}) = (\tfrac{1}{n} \sum_{i=1}^n y_i \mathbf{x}_i)^\top \boldsymbol{w} - \tfrac{1}{n} \sum_{i=1}^n A(\mathbf{x}_i^\top \boldsymbol{w}) + \text{const.}$$

Since a composition of a convex and a linear function is convex, we conclude that $\ell_n(\boldsymbol{w})$ is a concave function.

Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the data matrix with rows $\mathbf{x}_i$ and let $\mathbf{y}$ be the vectors with entries y_i. It can be shown that the unique optimizer of $\ell_n(\boldsymbol{w})$ (if it exists), must satisfy the likelihood equations

← Exercise 1.9.13

$$\mathbf{X}^\top y \;=\; \mathbf{X}^\top A'(\mathbf{X}\widehat{\boldsymbol{w}}), \tag{1.21}$$

where the function A' is applied elementwise to the vector $\mathbf{X}\widehat{\boldsymbol{w}}$. In machine learning it is customary to call A' in this context an activation function.

We now discuss a bunch of basic examples.

Example 1.7.1 (Linear regression)**.** *Consider the univariate Gaussian distribution in Example 1.1.4. Suppose now that σ^2 is fixed and the model is parametrized only by the mean μ. We can rewrite the density as*

$$f(y;\mu,\sigma^2) \;=\; \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} \exp\{\tfrac{1}{\sigma^2}(y\mu - \tfrac{\mu^2}{2})\} \;=\; h(y,\sigma^2)\exp\{\tfrac{1}{\sigma^2}(y\mu - A(\mu))\},$$

where $h(y,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}$ and $A(\mu) = \mu^2/2$. If we model the mean μ as a linear function of a vector $\mathbf{x}$, $\mu = \boldsymbol{w}^\top\mathbf{x}$ we get the standard Gaussian linear regression

$$f(y|\mathbf{x},\boldsymbol{w},\sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\tfrac{1}{2\sigma^2}(y-\boldsymbol{w}^\top\mathbf{x})^2\right\}.$$

We easily see that this is a generalized linear model as defined above. As $A'(\mu) = \mu$ and $A''(\mu) = 1$, we get $\mathbb{E}(Y|\mathbf{x},\boldsymbol{w},\sigma^2) = \mu = \boldsymbol{w}^\top\mathbf{x}$, $\operatorname{var}(Y|\mathbf{x},\boldsymbol{w},\sigma^2) = \sigma^2$ and, by (1.21), the MLE equations are $\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\boldsymbol{w}$, which are identical to the OLS estimating equations.

Example 1.7.2 (Binomial regression)**.** *The binomial distribution for the number of successes in n trials, $y \in \{0,\dots,n\}$ has the probability mass function*

$$\operatorname{Bin}(y;n,p) \;=\; \binom{n}{y} p^y(1-p)^{n-y} \;=\; \binom{n}{y}\exp\{y\log\tfrac{p}{1-p} + n\log(1-p)\}.$$

If $\theta = \log\frac{p}{1-p}$ then $p = \frac{e^\theta}{1+e^\theta}$, which is typically called a sigmoid function and denoted by $\sigma(\theta)$. Moreover, $A(\theta) = n\log(1+e^\theta)$ and so $A'(\theta) = n\sigma(\theta)$. Now we use this distribution for the GLM setup. If the response variable is the number of successes in n trials, $y \in \{0,\dots,n\}$, we can use binomial regression, which is defined by

$$f(y;\mathbf{x},N,\boldsymbol{w}) \;=\; \operatorname{Bin}(y;n,\sigma(\boldsymbol{w}^\top\mathbf{x})),$$

where the logistic regression becomes a special case with $n = 1$. This clearly has the right form and $\mathbb{E}(Y) = n\theta$, $\operatorname{var}(Y) = np(1-p)$. By Exercise 1.9.13 the likelihood equations are

$$\mathbf{X}^\top\mathbf{y} \;=\; \mathbf{X}^\top\sigma(\mathbf{X}\widehat{\boldsymbol{w}}).$$

Example 1.7.3 (Poisson regression)**.** *The Poisson distribution is a distribution over $\mathcal{X} = \{0, 1, 2, \ldots\}$ with the probability mass function*

$$f(y; \mu) \;=\; \frac{\mu^y}{y!} e^{-\mu} \qquad y \in \mathcal{X}.$$

This is an exponential family with mean parameter μ ($\mathbb{E}Y = \mu$) and canonical parameter $\theta = \log(\mu)$. In Poisson regression we take $\theta = \boldsymbol{w}^\top \boldsymbol{x}$.

Modelling the canonical parameters as a linear function of external variables is not the only choice. For example, for the Bernoulli distribution in Example 1.7.2 gives the logistic regression. An alternative approach is to model $p = \Phi(\boldsymbol{w}^\top \boldsymbol{x})$, which gives the probit regression. Generalized linear models with non-canonical link functions correspond to curved exponential families.

1.8 Diaconis-Ylvisaker conjugate priors

An important advantage of exponential families over more general classes of distributions is that they admit explicit conjugate prior for Bayesian computations. The conjugate prior measure for the exponential family (1.1) is given by the density (w.r.t. the Lebesgue measure) of the form

$$\pi(\boldsymbol{\theta}) \;=\; C \exp\{\langle \boldsymbol{\theta}, \tau \rangle - n_0 A(\boldsymbol{\theta})\}, \qquad \tau \in \mathbb{R}^d, n_0 \geq 0. \tag{1.22}$$

Note that $\pi(\boldsymbol{\theta}) \equiv 0$ outside of Θ because there $A(\boldsymbol{\theta}) \equiv +\infty$. It can be also shown that the distribution is normalizable if $n_0 > 0$ and $\tau / n_0 \in \mathcal{K} = \operatorname{conv}(T(\mathcal{X}))$[13].

[13] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, pages 269–281, 1979

Proposition 1.8.1. *For a regular exponential family, consider the conjugate prior in (1.22). If $\boldsymbol{\mu}$ is the mean parameter then we have $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\mu}] = \frac{\tau}{n_0}$.*

Proof sketch. We use the fact that $\boldsymbol{\mu} = \nabla A(\boldsymbol{\theta})$. We have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[\tau - n_0 \nabla A(\boldsymbol{\theta})] &= \int_\Theta (\tau - n_0 \nabla A(\boldsymbol{\theta})) C \exp\{\langle \boldsymbol{\theta}, \tau \rangle - n_0 A(\boldsymbol{\theta})\} \mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \nabla (C \exp\{\langle \boldsymbol{\theta}, \tau \rangle - n_0 A(\boldsymbol{\theta})\}) \mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \nabla \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}
\end{aligned}$$

It can be shown that the latter integral is equal to zero. This result is a consequence of Green's theorem (a general form of the fundamental theorem of calculus). A rigorous proof of this result was first presented in Diaconis and Ylvisaker (1979); for a simplified proof see [14]. □

[14] Lawrence D. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1986

If the prior is of the form (1.22) then the posterior $\pi(\boldsymbol{\theta}|\mathbf{x}_{1:n})$ satisfies

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{x}_{1:n}) &\propto \pi(\boldsymbol{\theta})\prod_{i=1}^{n} f(\mathbf{x}_i;\boldsymbol{\theta}) \\ &= C\prod_{i=1}^{n} h(\mathbf{x}_i)\exp\{\langle\boldsymbol{\theta}, \tau + \sum_{i=1}^{n}\mathbf{t}(\mathbf{x}_i)\rangle - (n+n_0)A(\boldsymbol{\theta})\}.\end{aligned}$$

And so it has the same general form as the prior, where (τ, n_0) is replaced with $(\tau + n\overline{\mu}_n, n_0 + n)$. By Proposition 1.8.1, the Bayes estimator of the mean parameter is

$$\mathbb{E}[\boldsymbol{\mu}|\mathbf{x}_{1:n}] = \frac{\tau + n\overline{\mu}_n}{n_0 + n} = \lambda\frac{\tau}{n_0} + (1-\lambda)\overline{\mu}_n,$$

where $\lambda = \frac{n_0}{n_0+n}$ and so the Bayes estimator is a convex combination of the MLE $\overline{\mu}_n$ and the prior mean τ/n_0.

We can easily define conjugate priors over non-canonical parameter. For example, to get a conjugate prior over the mean parameter we simply change $\boldsymbol{\theta}$ with $\theta(\boldsymbol{\mu})$ in (1.22) and change the definition of C so that the corresponding expression integrates to 1. Note that this is not the same as the density obtained through the change of variable formula.

Example 1.8.2. *If we do it for the Bernoulli distribution, the conjugate prior for the mean parameter is the Beta distribution.*

1.9 Exercises

Exercise 1.9.1. *Prove formally, using (1.4), that the space of canonical parameters in the centered multivariate Gaussian distribution is $\mathbb{S}^m_+$.*

Exercise 1.9.2. *Consider a multivariate Gaussian distribution with general mean vector $\boldsymbol{\mu} \in \mathbb{R}^m$. Show that the canonical parameter space is $\mathbb{R}^m \times \mathbb{S}^m_+$ with canonical parameters $(K\boldsymbol{\mu}, K)$ and that the sufficient statistics is $(\boldsymbol{x}, -\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^\intercal)$.*

Exercise 1.9.3. *Show that the distribution of $\boldsymbol{x}_{1:n}$ in Proposition 1.1.7 is of exponential type with the same canonical space Θ. What is the sufficient statistics? (c.f. Example 1.1.4)*

Exercise 1.9.4 (Ising model on a bipartite graph)**.** *Consider the Ising model on the bipartite graph G with m nodes $X_1, \ldots, X_m$ and n nodes $Y_1, \ldots, Y_n$ such that G has mn edges connecting each X_i with each Y_j. Show that to compute the conditional distribution of $Y = (Y_1, \ldots, Y_n)$ given $X = (X_1, \ldots, X_m)$ we essentially need to: (i) apply a linear function to X, (ii) apply an activation function to each element of the resulting vector. (Does it ring a bell?)*

Exercise 1.9.5 (Gaussian graphical models)**.** *Let G be a graph over m nodes representing m random variables $X_i \in \mathbb{R}$ for $i = 1, \dots, m$ with mean zero. Describe the distribution satisfying*

$$f(\mathbf{x}; \boldsymbol{\theta}) \;\propto\; \exp\{\sum_{i=1}^{m} \theta_{ii} x_i^2 + \sum_{ij \in G} \theta_{ij} x_i x_j\}.$$

Exercise 1.9.6. *Consider an exponential family with sufficient statistics $\boldsymbol{t}(\mathbf{x})$ and canonical parameter $\boldsymbol{\theta} \in \Theta$. Consider now a model whose parameter space is $\mathcal{L} \cap \Theta$ for some affine subspace $\mathcal{L} \subseteq \mathbb{R}^d$, such that $\mathcal{L} \cap \Theta \neq \emptyset$. Show that it forms an exponential family with $\Theta' = \Theta \cap \mathcal{L}$ and sufficient statistics which is a linear transformation of $\boldsymbol{t}(\mathbf{x})$.*

Exercise 1.9.7. *Suppose the state-space $\mathcal{X}$ is finite. Show that every exponential family over $\mathcal{X}$ is regular.*

Exercise 1.9.8. *Suppose that $\mathbf{X} = (X_1, \dots, X_n)$ has i.i.d. components that are $N(\mu, \sigma^2)$. Show that the vector $\mathbf{X} - \overline{X}\mathbf{1}$ is independent of $\overline{X} = \frac{1}{n}\mathbf{1}^\top X$. Use this fact to conclude that s^2 defined in (1.13) is independent of $\bar{x}$. Hint: This result follows from basic matrix algebra. Let $A = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Note that $A\mathbf{X} = \mathbf{X} - \overline{X}\mathbf{1}$ and that $\operatorname{cov}(\mathbf{X}) = \sigma^2 I_n$. Finally, note that $(n-1)s^2 = \operatorname{tr}(A\boldsymbol{x}\boldsymbol{x}^\top A^\top)$.*

Exercise 1.9.9. *Let E be any set of pairs of elements of $\{1, \dots, m\}$. Use Proposition 1.4.6 to show that for any two $A, B \in \mathbb{S}_+^m$ there exists a unique $X \in \mathbb{S}_+^m$ such that $X_{ij} = A_{ij}$ for $ij \in E$ and $(X^{-1})_{ij} = B_{ij}$ for $ij \notin E$.*

Exercise 1.9.10. *Consider the Gaussian distribution $N(\mu, \sigma^2)$. From the first principles, provide the two mixed parametrizations. Discuss their set of parameters and the corresponding Fisher information matrices.*

Exercise 1.9.11. *Prove Proposition 1.6.2.*

Exercise 1.9.12. *In the zero-mean Gaussian distribution $N(0, \Sigma)$ with $K = \Sigma^{-1}$, the log-likelihood function is*

$$\ell(K; S_n) \;=\; \log\det K - \langle K, S_n \rangle,$$

where $S_n = \frac{1}{n}\sum_i x_i x_i^\top$ is the sample covariance matrix. Following Section 1.6 of the notes show that the dual log-likelihood, up to some additive constants, is

$$\check{\ell}(\Sigma; S_n) \;=\; \log\det\Sigma - \langle \Sigma, S_n^{-1} \rangle.$$

Consider the bivariate Gaussian distribution with mean zero and covariance

$$\Sigma = \begin{bmatrix} a & b \\ b & a \end{bmatrix}.$$

Compare in simulations the maximum likelihood estimator $(\hat{a}, \hat{b})$ of (a, b) with the dual maximum likelihood estimator $(\check{a}, \check{b})$. Based on your simulations, what is the asymptotic behaviour of $\sqrt{n}(\hat{a} - \check{a})$?

Exercise 1.9.13. *Consider a generalized linear model with canonical link function. Argue that the maximum likelihood estimator of the parameter* $\boldsymbol{w}$*, for data* $(y_i, \mathbf{x}_i)$ *for* $i = 1, \ldots, n$*, leads to a convex optimization problem. Show that the MLE satisfies*

$$\mathbf{X}^\top \mathbf{y} \;=\; \mathbf{X}^\top A'(\mathbf{X}\boldsymbol{w}),$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ *is the data matrix,* $\mathbf{y} = (y_1, \ldots, y_n)$*, and in* $A'(\mathbf{X}\boldsymbol{w})$ *the function* A' *is applied elementwise to the vector* $\mathbf{X}\boldsymbol{w}$*.*

2
Statistical Decision Theory (1 week)

Statistical decision theory was covered in the first semester. Here I recall some of this material but in class we focus entirely on admissibility.

*2.1 Decision theoretic framework**

Statistical decision theory, developed by Abraham Wald, Jerzy Neyman and others during the mid-20th century, provides an abstraction that allows for comparison of statistical procedures. Our decision theoretic framework is made up from the following ingredients:
• **A family of probabilistic models with parameterization** $\theta \in \Theta$**.** We can think of this as a mapping from the parameter space Θ to a family of probability distributions $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. We assume that all distributions in $\mathcal{P}$ are absolutely continuous with respect to some underlying measure μ. Examples of parameter spaces and models include the following:

- $\Theta \subseteq \mathbb{R}^d$: exponential, Gaussian, other parametric models.
- $\Theta \subseteq$ (a function space): in nonparametric settings, e.g. the set of all twice-differentiable functions.

• **A decision procedure** In general this is a recipe that defines what action to take, given a set of observations $(\mathcal{X}, \mu)$. The set of possible actions is denoted $\mathcal{A}$. Examples of possible actions might include:

- Accepting or rejecting a null hypothesis, $\mathcal{A} = \{0, 1\}$.
- Estimating a value for some model parameter θ, $\mathcal{A} = \Theta$.
- Selecting one family of models as "superior" to other models (model selection).

Formally, a (non-randomized) **decision rule** is a measurable function $\delta : \mathcal{X} \to \mathcal{A}$ or $\delta : \mathcal{X}^n \to \mathcal{A}$. An estimator δ is a particular kind

of decision procedure by which we estimate the value of θ from the observations. In this case $\mathcal{A} = \Theta$. Denote by $\mathcal{D}_0$ the set of all (non-randomized) decision rules.

• **A loss function**, which tells us how to evaluate different decision procedures. This is an extended mapping $L : \Theta \times \mathcal{A} \to [0, \infty]$, in which $L(\theta, a)$ represents the loss incurred by deciding a when θ is "true". In many applications $L(\theta, a)$ is convex both in θ and in a (or jointly) but there are important examples where there is no convexity. Canonical examples of the loss functions are the convex loss functions $L(\theta, a) = \|\theta - a\|^2$, $L(\theta, a) = \|\theta - a\|$, and the $0/1$-loss $L(\theta, a) = \mathbb{1}(\theta \neq a)$. Depending on the particular application we also consider hinge loss, Kullback-Leibler divergence, and many others.

We assume that $L(\theta, a)$ lower semicontinuous in a. This means that for every $\theta \in \Theta$ and every $t \in \mathbb{R}$ the set $\{a \in \mathcal{A} : L(\theta, a) \leq t\}$ is closed. By Exercise 2.7.1, such $L(\theta, a)$ is then also measurable in a[1]. In particular, if X has distribution $\mathbb{P}_\theta$ then $L(\theta, \delta(X))$ is $\mathbb{P}_\theta$ measurable.

← Exercise 2.7.1

[1] We think about $\mathcal{A} \subseteq \mathbb{R}^d$ as a measurable space with the underlying σ-field of Borel sets (generated by the open subsets). A function $f : \mathcal{A} \to \mathbb{R}$ is then measurable if the sets $\{a : f(a) \leq t\}$ are measurable.

We define the **risk function** as:

$$R(\theta, \delta) \; = \; \mathbb{E}_\theta L(\theta, \delta(X)).$$

When the sample space is continuous, a quadratic loss function is frequently used. In this case, the risk is simply the mean squared error.

Example 2.1.1. *Suppose a coin is being tossed and you are interested in the probability of getting heads. We can model this using a family of Bernoulli distributions on the binary outcome space: $X_i \sim \text{Ber}(\theta)$ with $\theta \in \Theta = (0, 1)$. Suppose we observe n replications with i.i.d results $X = (X_1, \ldots, X_n) \in \{0, 1\}^n$ from $\mathbb{P}_\theta$. We consider various rules $\delta_i(X)$ for estimating θ. In this case $\mathcal{A} = [0, 1]$. We use the quadratic loss $L(\theta, \delta(X)) = (\theta - \delta(X))^2$. and evaluate the associated risk function for each estimator:*

- *The first estimator is the sample mean $\delta_1(X) = \frac{1}{n} \sum_{i=1}^n X_i$ with risk*

$$R(\theta, \delta_1) = \frac{1}{n} \theta(1 - \theta).$$

- *The second estimator is given by the constant value $\delta_2(X) = \frac{1}{2}$ with risk*

$$R(\theta, \delta_2) = (\theta - \tfrac{1}{2})^2.$$

- *The third estimator*

$$\delta_3(X) = \frac{\sum_{i=1}^n X_i + 3}{n + 6}$$

with risk

$$R(\theta, \delta_3) = \frac{9 + (n - 36)\theta - (n - 36)\theta^2}{(n + 6)^2}.$$

The risks functions for the first and the last estimator are plotted in Figure 2.1. Intuitively, it seems clear that if we strongly believe that θ is close to 0.5, δ_3 outperforms δ_1. Later we learn how to formalize that.

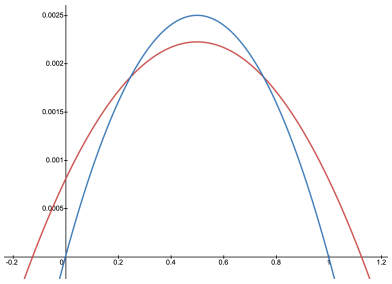


Figure 2.1: Risks functions for δ_1 (blue) and δ_3 (red) shown for $n = 100$.

Example 2.1.2 (Benefits of bias). *Suppose $X_i \sim U(0,\theta)$ are drawn i.i.d. $i = 1,\dots,n$. Consider the statistic*

$$\delta(X) = \max\{X_1,\dots,X_n\}.$$

Its distribution is easily found to be

$$\mathbb{P}(\delta(X) \le t) \;=\; \begin{cases} 0 & \text{it } t < 0 \\ \left(\frac{t}{\theta}\right)^n & \text{if } 0 \le t \le \theta \\ 1 & \text{if } t > \theta \end{cases}$$

with density $f(t) = \frac{n}{\theta^n} t^{n-1}$ for $t \in [0,\theta]$. We thus have

$$\mathbb{E}_\theta(\delta(X)) \;=\; \frac{n}{\theta^n}\int_0^\theta t^n \mathrm{d}t \;=\; \tfrac{n}{n+1}\theta.$$

Similarly,

$$\mathbb{E}_\theta(\delta^2(X)) \;=\; \frac{n}{\theta^n}\int_0^\theta t^{n+1} \mathrm{d}t \;=\; \frac{n}{n+2}\theta^2.$$

Clearly $\frac{n+1}{n}\delta(X)$ is an unbiased estimator. But instead consider all estimators of the form $a\delta(X)$, $a \in \mathbb{R}$. Find the value of a that gives the minimal risk with the quadratic loss function. We have

$$\begin{aligned} R(\theta, a\delta) \;&=\; \mathbb{E}_\theta(\theta - a\delta(X))^2 \;=\; \theta^2 - 2a\mathbb{E}_\theta(\delta(X)) + a^2\mathbb{E}_\theta\delta(X)^2 \\ &=\; \theta^2 - 2a\frac{n}{n+1}\theta^2 + a^2\frac{n}{n+2}\theta^2. \end{aligned}$$

This is a convex function optimum does not depend on θ and the minimum is easily found to be $a^* = \frac{n+2}{n+1}$. We have

$$R(\theta, \tfrac{n+1}{n}\delta) = \frac{\theta^2}{n(n+2)}, \qquad R(\theta, \tfrac{n+2}{n+1}\delta) = \frac{\theta^2}{(n+1)^2}.$$

This shows that unbiased estimators are not always best if the goal is to minimize risk.

There are situations where certain decision rules can be disregarded. A decision rule $\delta(X)$ is **inadmissible** if there is some competing procedure δ' has uniformly lower risk, meaning:

(a) $R(\theta,\delta') \le R(\theta,\delta)$ for all $\theta \in \Theta$,

(b) $R(\theta,\delta') < R(\theta,\delta)$ for at least one $\theta \in \Theta$.

Otherwise the decision rule is **admissible**. We treat admissibility in more detail in Section 2.4. In practice however, we often need to compare two admissible procedures and then there is no one obvious way to say that one is better than the other; more on that in Section 2.3.

Example 2.1.2 shows that a natural unbiased estimator may not be admissible. A slightly more artificial example follows.

Example 2.1.3. *Continuing Example 2.1.1, consider* $\delta_4(X) = X_1$*, a rather silly estimator that uses only the first observation. Then* $R(\theta, \delta_4) = \mathbb{E}_\theta(\theta - X_1)^2 = \theta(1-\theta)$ *which is always greater than* $R(\theta, \delta_1)$ *for all* $\theta \in (0,1)$ *(unless* $n = 1$ *of course). Therefore* δ_4 *is inadmissible.*

Given a sample $X = (X_1, \ldots, X_n)$, define the **empirical risk** as

$$\widehat{R}_n(\theta, \delta) = \frac{1}{n} \sum_{i=1}^{n} L(\theta, \delta(X_i)). \tag{2.1}$$

A popular way of constructing an estimator is by minimizing the empirical risk, which is directly related to M-estimation discussed later in Section 8.3.

← Exercise 2.7.2

Remark 2.1.4. *In the supervised learning set-up we have the training data* $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$*. Where* $X_i \in \mathcal{X}$ *are called features and* $Y_i \in \mathcal{Y} \subset \mathbb{R}$ *are called labels. Here the loss function* $L(y, \delta(x))$ *is defined on* $\mathcal{Y} \times \mathcal{X}$*. The (expected) risk is then defined through a double integral over* $\mathcal{X} \times \mathcal{Y}$*. In this context the empirical risk becomes*

$$\widehat{R}_n(\delta) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \delta(X_i)).$$

2.2 *Randomized decision rules**

It can often be useful to consider randomized decision rules.

Definition 2.2.1. *A* ***randomized decision rule*** δ *is a measurable mapping from* $\mathcal{X}$ *to probability measures on* $\mathcal{A}$*,* $x \mapsto \delta_x$*. By this we mean that the function* $x \mapsto \delta_x(B) \in [0,1]$ *is measurable for any fixed Borel subset* $B \subseteq \mathcal{A}$*. (such objects are also called probability kernels)*

The idea is that for a fixed $X = x$ a random action A will be drawn from δ_x, $A|X = x \sim \delta_x$, that is, $\mathbb{P}(A \in B | X = x) = \delta_x(B)$ for every Borel set $B \subseteq \mathcal{A}$. The definition makes sure that the joint distribution probabilities of the the following form are well defined:

$$\mathbb{P}(A \in B, X \in U) = \int_U \int_B \delta_x(\mathrm{d}a)\mathbb{P}(\mathrm{d}x) = \int_U \delta_x(B)\mathbb{P}(\mathrm{d}x),$$

which they are because $\delta_x(B)$ is measurable in x.

Note that the marginal distribution of X and the conditional distribution of A given X naturally specify the joint distribution of (A, X) and so also the marginal distribution of A. For a randomized decision rule, we define the risk function as

$$R(\theta, \delta) = \mathbb{E}_\theta L(\theta, A) = \mathbb{E}_\theta \left(\mathbb{E}[L(\theta, A)|X]\right) = \iint L(\theta, a)\delta_x(\mathrm{d}a)P_\theta(\mathrm{d}x). \tag{2.2}$$

Here the conditional expectation $\mathbb{E}(\cdot|X = x)$ is computed with respect to the conditional distribution δ_x of A given $X = x$. Denote by $\mathcal{D}$ the set of all randomized decision rules such that $R(\theta, \delta) < +\infty$ for all $\theta \in \Theta$. The set $\mathcal{D}$ contains all non-randomized decision rules $\mathcal{D}_0$ in which case to each x we assign the point mass at this point.

Example 2.2.2 (Statistical testing)**.** *Consider the problem of testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. Here $\mathcal{A} = \{0, 1\}$. We consider here a randomized testing procedure where based on data $x \in \mathcal{X}$ we compute $\varphi(x) \in [0, 1]$ and let δ_x be the Bernoulli distribution with the success probability $\varphi(x)$. In testing problems we typically use the $0/1$-loss:*

$$L(\theta, a) = \mathbb{1}(a = 1, \theta \in \Theta_0) + \mathbb{1}(a = 0, \theta \in \Theta_1).$$

*The **power function** assigns to θ the probability or rejecting H_0:*

$$\beta(\theta) := \mathbb{P}_\theta(A = 1) = \mathbb{E}[\mathbb{P}_\theta(A = 1|X)] = \mathbb{E}\varphi(X). \tag{2.3}$$

The risk function for the ranomized rule defined above is

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta\mathbb{E}[L(\theta, A)|X] = \mathbb{E}_\theta\mathbb{P}(A = 1, \theta \in \Theta_0|X) + \mathbb{E}_\theta\mathbb{P}(A = 0, \theta \in \Theta_1|X) \\ &= \begin{cases} \mathbb{E}_\theta\varphi(X) & \text{if } \theta \in \Theta_0 \\ \mathbb{E}_\theta(1 - \varphi(X)) & \text{if } \theta \in \Theta_1 \end{cases} \\ &= \begin{cases} \beta(\theta) & \text{if } \theta \in \Theta_0 \\ 1 - \beta(\theta) & \text{if } \theta \in \Theta_1 \end{cases}. \end{aligned}$$

In standard statistical theory, randomized decision rules appear mostly in the context of statistical testing. It is important however to discuss briefly general advantages of considering them. For two randomized decision rules δ, δ' and $\lambda \in (0, 1)$ the convex combination $(1 - \lambda)\delta + \lambda\delta'$ is the randomized decision rule defined by

$$(1 - \lambda)\delta + \lambda\delta' = \begin{cases} \delta & \text{with probability } 1 - \lambda, \\ \delta' & \text{with probability } \lambda. \end{cases} \tag{2.4}$$

With this definition the set of randomized decision rules $\mathcal{D}$ forms a convex set. The following proposition suggests that convexity will play an important role in our analysis; see also Remark 2.3.1 below.

Proposition 2.2.3. *The set $\mathcal{D}$ of all randomized decision rules is convex. The function $\delta \mapsto R(\theta, \delta)$ is a linear function of $\delta \in \mathcal{D}$ for any fixed θ, that is,*

$$R(\theta, (1-\lambda)\delta + \lambda\delta') = (1-\lambda)R(\theta, \delta) + \lambda R(\theta, \delta')$$

for every $\lambda \in (0,1)$ and $\delta, \delta' \in \mathcal{D}$.

Proof. The first statement is clear. For the second statement let A_0, A_1 be the random variables representing the randomized decisions δ and δ'. Their mixture is A and so A is equal to A_0 with probability $(1-\lambda)$ and it is equal to A_1 with probability λ. We thus have

$$R(\theta, (1-\lambda)\delta + \lambda\delta) = \mathbb{E}_\theta L(\theta, A) = (1-\lambda)\mathbb{E}_\theta L(\theta, A_0) + \lambda \mathbb{E}_\theta L(\theta, A_1),$$

which is equal to $(1-\lambda)R(\theta, \delta) + \lambda R(\theta, \delta')$ as claimed. □

It may be surprising that linearity in Proposition 2.2.3 holds irrespective of the loss function. The following remark may help.

Remark 2.2.4. *It is important to note a subtlety in the above discussion. If δ, δ' are non-randomized decision rules then $(1-\lambda)\delta + \lambda\delta'$ could mean two different objects. It could be either a randomized decision rule as defined in (2.4), but it could be also a non-randomized decision rule*

$$((1-\lambda)\delta + \lambda\delta')(x) = (1-\lambda)\delta(x) + \lambda\delta'(x).$$

For the last definition, there is no equivalent of Proposition 2.2.3 although convexity of $\mathcal{D}_0$ still holds as long as $\mathcal{A}$ is convex. In the hypothesis testing case, $\mathcal{A} = \{0, 1\}$ is not convex.

2.3 *Bayesian and minimax rules**

In the ideal situation we would be able to choose the best decision rule δ as the one that leads to the lowest risk. However the risk $R(\theta, \delta)$ of δ depends on θ and, in general, there is no total ordering on the risk functions for different θ; c.f Figure 2.1. The only realistic solution is to define a functional on the space of risk functions and evaluate decision rules according to the value of this functional.

2.3.1 *Definitions and basic properties*

Two natural choices for functionals on the risk functions are the maximum risk and the Bayes risk. The maximum risk of δ is defined as the number

$$\overline{R}(\delta) = \sup_{\theta \in \Theta} R(\theta, \delta).$$

For a given prior distribution $\pi(\theta)$ on Θ, the corresponding **Bayes risk** is

$$r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)\mathrm{d}\theta.$$

We say that a decision rule δ^* is a (randomized) **minimax rule** if

$$\overline{R}(\delta^*) = \inf_{\delta \in \mathcal{D}} \overline{R}(\delta).$$

A (randomized) **Bayes rule** with respect to prior π is any decision rule δ^* that satisfies

$$r(\pi, \delta^*) = \inf_{\delta \in \mathcal{D}} r(\pi, \delta).$$

When the infima are over $\mathcal{D}_0$ we call the corresponding decision procedures **non-randomized Bayesian** and **non-randomized minimax** rules respectively.

In practice we prefer non-randomized decision rules. The reason for considering randomized versions can be explained as follows.

Remark 2.3.1. *By Proposition 2.2.3, $R(\theta, \delta)$ is linear (and so convex) in δ for any fixed θ. By Proposition B.2.5, $\overline{R}(\delta)$ is convex and $r(\pi, \delta)$ is linear in δ. It follows that both minimizing $\overline{R}(\delta)$ and $r(\pi, \delta)$ over $\mathcal{D}$ is a (infinite dimentional) convex optimization problem. For non-randomized rules such favorable properties will hold only under special assumptions on the loss function.*

The following result shows that there is no difference between randomized and non-randomized Bayes rules.

Theorem 2.3.2. *Every Bayes rule δ^* satisfies $r(\pi, \delta^*) = \inf_{\delta \in \mathcal{D}_0} r(\pi, \delta)$.*

Proof. We omit the formal proof. See Section 1.8 in [2]. □

[2] Thomas S. Ferguson. *Mathematical statistics: A decision theoretic approach*. Probability and Mathematical Statistics, Vol. 1. Academic Press, New York-London, 1967

2.3.2 *Simple geometric insights*

It is good to briefly discuss the geometric picture that drives our intuition. Consider the set of all finite mixtures of all non-randomized decision rules. It is natural to denote this set as $\text{conv}(\mathcal{D}_0)$. Although in general $\mathcal{D} \neq \text{conv}(\mathcal{D}_0)$ from the point of view of risk analysis often we can use these two sets interchangably obtaining the same results; see Section 1.6 in [3] for a more careful discussion and further references. This insight makes Theorem 2.3.2 very natural to conjecture because $r(\pi, \delta)$ is a linear function of δ and we optimize it (essentially) over $\text{conv}(\mathcal{D}_0)$.

[3] Thomas S. Ferguson. *Mathematical statistics: A decision theoretic approach*. Probability and Mathematical Statistics, Vol. 1. Academic Press, New York-London, 1967

We next discuss a trivial instance of this when things can be explicitly depicted in two dimensions. Take $\Theta = \{0, 1\}$ the risk function $R(\theta, \delta)$ can be represented by a point $(R(0, \delta), R(1, \delta)) \in \mathbb{R}^2$. Define the risk set

$$\mathcal{R} = \{(R(0, \delta), R(1, \delta)) : \ \delta \in \text{conv}(\mathcal{D}_0)\}.$$

By Proposition 2.2.3, $\mathcal{R}$ is a convex set, namely, if $y, y' \in \mathcal{R}$ (with the underlying δ, δ') then $(1 - \lambda)y + \lambda y'$ lies in $\mathcal{R}$ as it is realized by the randomized rule $(1 - \lambda)\delta + \lambda\delta'$.

Suppose that the considered set of non-randomized decision rules $\mathcal{D}_0$ is finite. In this case $\mathcal{R}$ forms a polytope and the geometry of the Bayes rule is quite straightforward. A prior distribution π is simply a point $(\pi(0), \pi(1))$ with nonnegative coordinates that sum to 1. Moreover,

$$r(\pi, \delta) = \pi(0)R(0, \delta) + \pi(1)R(1, \delta).$$

Thus a Bayes rule will be obtained as one of the minimizers of a given linear function on $\mathcal{R}$. The maximizers may be unique but it is not always the case. In this case, Theorem 2.3.2 holds trivially.

For the minimax rule note that

$$\overline{R}(\delta) = \max\{R(0, \delta), R(1, \delta)\}.$$

Draw the line $R(0, \delta) = R(1, \delta)$ and consider how the minimax risk can be improved above or below the line.

2.3.3 *Finding the Bayes rule*

We next explain how to find a Bayes rule. By Theorem 2.3.2, it is enough to minimize $r(\pi, \delta)$ over $\delta \in \mathcal{D}_0$. For a prior distribution $\pi(\theta)$ let $\pi(\theta|x)$ denote the posterior distribution. We write $p(x|\theta)$ be the density of the distribution $\mathbb{P}_\theta$. By the Bayes theorem

$$\pi(\theta|x) \;=\; \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)\mathrm{d}\theta} \;=\; \frac{p(x|\theta)\pi(\theta)}{m(x)}, \tag{2.5}$$

where $m(x) = \int p(x|\theta)\pi(\theta)\mathrm{d}\theta$. The **posterior risk** is defined as

$$r(\delta|x) \;:=\; \int L(\theta, \delta(x))\pi(\theta|x)\mathrm{d}\theta. \tag{2.6}$$

One of the main result on Bayesian rules is the following proposition.

Proposition 2.3.3. *The Bayes risk $r(\pi, \delta)$ for $\delta \in \mathcal{D}_0$ satisfies*

$$r(\pi, \delta) \;=\; \int r(\delta|x)m(x)\mathrm{d}x.$$

Define δ^* pointwise, for every $x \in \mathcal{X}$, as

$$\delta^*(x) \;=\; \delta_0(x), \quad \textit{where} \quad r(\delta_0|x) = \inf_{\delta \in \mathcal{D}_0} r(\delta|x). \tag{2.7}$$

If $\delta^* \in \mathcal{D}_0$ (i.e. if δ^* is measurable) then, by construction, $r(\delta^*|x) = \inf_{\delta \in \mathcal{D}_0} r(\delta|x)$ for every $x \in \mathcal{X}$ and δ^* is a Bayes rule.

Proof. Using the Fubini's theorem we get

$$\begin{aligned} r(\pi, \delta) &= \int R(\theta, \delta)\pi(\theta)\mathrm{d}\theta = \iint L(\theta, \delta(x))p(x|\theta)\pi(\theta)\mathrm{d}x\mathrm{d}\theta \\ &\overset{(2.5)}{=} \iint L(\theta, \delta(x))\pi(\theta|x)m(x)\mathrm{d}x\mathrm{d}\theta \\ &= \int \left(\int L(\theta, \delta(x))\pi(\theta|x)\mathrm{d}\theta \right) m(x)\mathrm{d}x \\ &= \int r(\delta|x)m(x)\mathrm{d}x. \end{aligned}$$

It is also clear that δ^* minimizes the above integral over $\mathcal{D}_0$:

$$\inf_{\delta\in\mathcal{D}_0} r(\pi,\delta) = \inf_{\delta\in\mathcal{D}_0} \int r(\delta|x)m(x)\mathrm{d}x \geq \int (\inf_{\delta\in\mathcal{D}_0} r(\delta|x))m(x)\mathrm{d}x = r(\pi,\delta^*).$$

If $\delta^* \in \mathcal{D}_0$ (c.f. Remark 2.3.4), the inequality becomes equality. By Theorem 2.3.2, minimizing over $\mathcal{D}_0$ is equivalent to minimizing over $\mathcal{D}$ showing that δ^* is a Bayes rule. □

Remark 2.3.4. *The function δ^* defined in Proposition 2.3.3 will be however measurable in all practical situations. For more details see* [4].

[4] Lawrence D Brown and Roger Purves. Measurable selections of extrema. *The Annals of Statistics*, pages 902–912, 1973

In some situations the Bayes rule can be found explicitly.

Proposition 2.3.5. *If $\Theta, \mathcal{A} \subseteq \mathbb{R}^d$ and $L(\theta,a) = \|\theta - a\|^2$ then the Bayes rule is*

$$\delta^*(x) = \int \theta\pi(\theta|x)\mathrm{d}\theta = \mathbb{E}(\theta|X=x).$$

Proof. By Proposition 2.3.3, the decision rule $\delta^* = \inf_{\delta\in\mathcal{D}_0} r(\delta|x)$ is a Bayes rule. In our case

$$r(\delta|x) = \int \|\theta - \delta(x)\|^2\pi(\theta|x)\mathrm{d}\theta.$$

For a fixed x, write $a = \delta(x)$ and minimize the above expression with respect to a. We have

$$\nabla_a \int \|\theta - a\|^2\pi(\theta|x)\mathrm{d}\theta = -2\int(\theta - a)\pi(\theta|x)\mathrm{d}\theta,$$

which is equal to the zero vector if and only if $a = \int \theta\pi(\theta|x)\mathrm{d}\theta = \mathbb{E}(\theta|X=x)$. □

Example 2.3.6. *Given a sample $X_1,\dots,X_n$ from $N(\theta,1)$ consider the prior $\pi \sim N(0,\tau^2)$. The posterior is*

$$N\left(\frac{n\tau^2}{n\tau^2+1}\bar{x}_n, \frac{\tau^2}{n\tau^2+1}\right).$$

Thus, for this prior and with the quadratic loss, the Bayes rule is

$$\delta^*(x_1,\dots,x_n) = \frac{n\tau^2}{n\tau^2+1}\bar{x}_n.$$

The risk function is

$$R(\theta,\delta^*) = \mathbb{E}_\theta(\theta - \tfrac{n\tau^2}{n\tau^2+1}\bar{X}_n)^2 = \left(\frac{1}{n\tau^2+1}\right)^2\theta^2 + \frac{1}{n}\left(\frac{n\tau^2}{n\tau^2+1}\right)^2.$$

Since $\mathbb{E}\theta^2 = \tau^2$, the Bayes risk is

$$r(\pi,\delta^*) = \mathbb{E}_\theta(\theta - \tfrac{n\tau^2}{n\tau^2+1}\bar{X}_n)^2 = \left(\frac{1}{n\tau^2+1}\right)^2\tau^2 + \frac{1}{n}\left(\frac{n\tau^2}{n\tau^2+1}\right)^2.$$

For comparison, consider $\delta_0(x_1,\dots,x_n) = \bar{x}_n$. We have $R(\theta,\delta_0) = \frac{1}{n}$ and so $r(\pi,\delta_0) = \frac{1}{n}$.

← Exercise 2.7.3
← Exercise 2.7.4
← Exercise 2.7.5

2.3.4 *The link between minimax and Bayesian rules*

The next result gives a connection between minimax and Bayesian rules.

Theorem 2.3.7. *Let δ^* be a Bayes rule for some prior π. Suppose that*

$$R(\theta, \delta^*) \leq r(\pi, \delta^*) \qquad \text{for all} \quad \theta \in \Theta. \tag{2.8}$$

Then δ^* is minimax.

Here probably the most interesting case when this happens is when $R(\theta, \delta^*)$ is constant in θ. Compare this with the geometric picture presented in Section 2.3.2.

Proof. We prove the contrapositive statement. If δ^* is not minimax then there exists δ_0 such that $\overline{R}(\delta_0) < \overline{R}(\delta^*)$. However, the fact that $r(\pi, \delta_0) \leq \sup_\theta R(\theta, \delta_0)$ implies that

$$r(\pi, \delta_0) \;\leq\; \overline{R}(\delta_0) \;<\; \overline{R}(\delta^*) \;\overset{(2.8)}{\leq}\; r(\pi, \delta^*).$$

But this contradicts the fact that δ^* is a Bayesian rule. □

To illustrate the utility of this theorem in actually finding minimax estimators consider the following example.

Example 2.3.8 (Bernoulli distribution)**.** *Suppose $X \sim \text{Bern}(\theta)$, $\theta \in \Theta = [0, 1]$. Given any prior π over $[0, 1]$, we define $m_1 = \mathbb{E}(\theta)$ and $m_2 = \mathbb{E}(\theta^2)$. The frequentist risk for the squared loss is*

$$R(\theta, \delta) = \theta^2(1 + 2(\delta_0 - \delta_1)) + \theta(\delta_1^2 - \delta_0^2 - 2\delta_0) + \delta_0^2,$$

whereas the Bayes risk is

$$r(\pi, \delta) = m_2(1 + 2(\delta_0 - \delta_1)) + m_1(\delta_1^2 - \delta_0^2 - 2\delta_0) + \delta_0^2.$$

Note that it depends on π only through the first two moments! The Bayes decision rule is found minimizing with respect to δ_0, δ_1:

$$\delta_0 = \frac{m_1 - m_2}{1 - m_1}, \qquad \delta_1 = \frac{m_2}{m_1}. \tag{2.9}$$

To satisfy the constant risk property, we note that $R(\theta, \delta)$ does not depend on θ if and only if $\delta_1 - \delta_0 = \frac{1}{2}$ and $\delta_1^2 - \delta_0^2 - 2\delta_0 = 0$, or equivalently $\delta_0 = \frac{1}{4}$, $\delta_1 = \frac{3}{4}$. Solving for m_1, m_2 in (2.9) yields the solution

$$m_1^* = \frac{1}{2}, \qquad m_2^* = \frac{3}{8}.$$

This, if π is such that $\mathbb{E}\theta = \frac{1}{2}$ and $\mathbb{E}\theta^2 = \frac{3}{8}$ and $\delta_0^* = \frac{1}{4}$, $\delta_1^* = \frac{3}{4}$, then δ^* is a Bayesian rule with respect to π and inequality (2.8) holds. By Theorem 2.3.7, δ^* is also minimax.

Note that if $\pi = \text{Beta}(1/2, 1/2)$ then θ has these moments. The associated Bayes (and hence minimax) risk is $\frac{1}{16}$. These calculations can be generalized to the case of n i.i.d. samples $X_1, \ldots, X_n$ from the Bernoulli distribution. We leave it as an exercise.

Consider the following slight generalization of the minimax rule: allow nature to choose a distribution π over Θ. In this case our goal is to minimize

$$\bar{r}(\delta) := \sup_{\pi} r(\pi, \delta).$$

In this context it is useful to consider the notion of a least favourable prior. A prior π^* is called **least favorable** if

$$\inf_{\delta \in \mathcal{D}} r(\pi^*, \delta) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta).$$

← Exercise 2.7.6

2.3.5 *Minimax theorem**

The minimax decision problem has a natural game-theoretic interpretation. In particular, a two-player game is played between nature and the statistician, with nature picking the prior π, and the statistician choosing a (possibly randomized) decision rule δ. Then the statistician pays to nature the amount $r(\pi, \delta)$. Note that nature gains the same amount the statistician loses, so that the game is zero-sum.The following two quantities are important

$$L^* = \sup_{\pi} \inf_{\delta} r(\pi, \delta)$$

$$U^* = \inf_{\delta} \sup_{\pi} r(\pi, \delta)$$

called respectively the lower and the upper values of the game. These quantities have the following interpretation: U^* is the amount the player pays when he is told what distribution nature choses before he choses δ. Conversely, L^* is the amount the player pays when nature is told the player's strategy δ before it chooses π.

← Exercise 2.7.7

Theorem 2.3.9 (von Neumann)**.** *Suppose that*

1. *The parameter space $\Theta = \{\theta_1, \ldots, \theta_k\}$ is finite, and*
2. *The risk set*

$$\mathcal{R} = \{y \in \mathbb{R}^k : \ y_i = R(\theta_i, \delta) \text{ for some } \delta \in \mathcal{D}\}$$

is closed and lies in the nonnegative orthant.

Then

(i) The game has the value $L^* = U^*$.

(ii) There exists a probability vector $\pi \in \mathbb{R}^k$ that is a least favourable prior.

Proof. In Exercise 2.7.7 we show that $L^* \leq U^*$ so in order to establish (i) we need to prove that $L^* \geq U^*$. Given $\alpha \in \mathbb{R}$, define the lower-rectangular set of the form

$$B_\alpha = \{y \in \mathbb{R}^k : y_i \leq \alpha, \forall i = 1, \ldots, k\}.$$

In addition, define

$$\gamma \;=\; \arg\inf_{\alpha \in \mathbb{R}} \{B_\alpha \cap \mathcal{R} \neq \emptyset\}.$$

By definition of γ, for each $n \in \mathbb{N}$ there exists a (randomized) decision rule δ_n, such that

$$R(\theta_j, \delta_n) \leq \gamma + \frac{1}{n}, \qquad \text{for all } j = 1, \ldots, k.$$

Therefore, for any prior π on $\Theta = \{\theta_1, \ldots, \theta_k\}$, we have

$$r(\pi, \delta_n) \;\leq\; \gamma + \frac{1}{n}.$$

Taking a supremum over the choice of priors yields that $\sup_\pi r(\pi, \delta_n) \leq \gamma + \frac{1}{n}$ and hence for $n \in \mathbb{N}$

$$\inf_\delta \sup_\pi r(\pi, \delta) \;\leq\; \gamma + \frac{1}{n}.$$

This inequality holds for every $n \in \mathbb{N}$ and taking the limit gives $U^* \leq \gamma$.

Now we are going to use the Separating Hyperplane Theorem B.1.3 to construct a vector $\pi = (\pi_1, \ldots, \pi_k)$ that can be viewed as a least favourable prior, therefore establishing (ii). It will also show that $L^* \geq \gamma$ establishing part (i).

Consider again the lower rectangle B_γ. Observe that its interior $\text{int}(B_\gamma)$ and the risk set $\mathcal{R}$ are two disjoint convex sets in $\mathbb{R}^k$. Consequently, the separating hyperplane theorem guarantees existence of some non-zero vector $\pi \in \mathbb{R}^k$ and constant c such that

$$\langle \pi, y \rangle \;\geq\; c \qquad \text{for all } y \in \mathcal{R},$$

$$\langle \pi, y \rangle \;\leq\; c \qquad \text{for all } y \in \text{int}(B_\gamma).$$

We claim that $\pi \geq 0$. This can be proven by contradiction. Suppose $\pi_i < 0$ for some i. Directly by definition of the set B_γ, we can construct a sequence of y vectors such that $y_i \to -\infty$ and $y_j = 0$ for $j \neq i$, while still staying in $\text{int}(B_\gamma)$. However, this yields a sequence such that $\langle \pi, y \rangle$ becomes indefinitely large contradicting the separation statement (these sequence ends up on the wrong side of the hyperplane $\langle \pi, y \rangle = c$). Thus we must have $\pi \geq 0$. Since $\pi \neq 0$ we can normalize it to sum to one, so it can be interpreted as a valid prior.

Consider the vector $x^* := \gamma \mathbf{1}$. Since x^* lies in the closure of $\text{int}(B_\gamma)$, we must have

$$\langle \pi, x^* \rangle = \gamma \leq c,$$

where we have used normalization property of π. Now letting δ be an arbitrary decision rule with risk vector $z \in \mathbb{R}^k$ such that $z_i = R(\theta_i, \delta)$ for all i we have

$$r(\pi, \delta) = \langle \pi, z \rangle \geq c \geq \gamma.$$

Since δ was arbitrary, $\inf_\delta r(\pi, \delta) \geq \gamma$ and, in consequence, $L^* \geq \gamma$, which completes the proof of part (i). Furthermore, the vector π that we constructed is the least favourable prior of part (ii). □

2.4 *Admissibility and Rao-Blackwell*

It is clear that we can discard inadmissible rules from our analysis. Thus it is useful to know that the procedure we analyse is admissible. It should be clear that minimax procedures are admissible.

Theorem 2.4.1. *If a Bayes rule for π is essentially unique (in the sense of measurable functions) then δ is admissible.*

Proof. Suppose there exists δ_0 such that $R(\theta, \delta_0) \leq R(\theta, \delta)$ for all $\theta \in \Theta$. Then $r(\pi, \delta_0) \leq r(\pi, \delta)$ and so, if δ is Bayes, then δ_0 is also Bayes. By our assumption, $\delta = \delta_0$ almost surely. In consequence, for every θ

$$\mathbb{E}_\theta L(\theta, \delta(X)) = \mathbb{E}_\theta L(\theta, \delta_0(X))$$

and so the risk functions are equal implying that δ_0 does not strictly dominate δ. □

← Exercise 2.7.8
← Exercise 2.7.9
← Exercise 2.7.10

In case, we cannot assure uniqueness, it is still useful to provide some sufficient conditions on admissibility.

We now revisit the Rao-Blackwell theorem in the language of statistical decision theory. It allows for a simple general construction of decision rules that dominate a given decision rule δ. Recall that a statistic $T = T(X)$ is sufficient for θ if the conditional distribution of X given T does not depend on θ. Given any rule, we can define

$$\eta(T) = \mathbb{E}[\delta(X)|T].$$

Since T is sufficient, $\mathbb{E}[\delta(X)|T]$ does not depend on θ and so, $\eta(T)$ is a valid statistic. The following classical result explains why $\eta(T)$ may be preferred over $\delta(X)$.

Theorem 2.4.2 (Rao-Blackwell)**.** *Let T be a sufficient statistic for $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, let δ be a decision rule, and define $\eta(T) = \mathbb{E}[\delta(X)|T]$. If $\theta \in \Theta$, $R(\theta, \delta) < +\infty$, and $L(\theta, a)$ is convex in a, then*

$$R(\theta, \eta) \leq R(\theta, \delta).$$

Proof. Using the Jensen's inequality (c.f Theorem B.2.3) for the convex function $f(a) = L(\theta, a)$ and with the conditional expectation $\mathbb{E}(\cdot|T = t)$ we get

$$L(\theta, \eta(T)) = L(\theta, \mathbb{E}(\delta(X)|T)) \leq \mathbb{E}\Big(L(\theta, \delta(X))\Big|T\Big). \tag{2.10}$$

Taking expectations on both sides we conclude $R(\theta, \eta(T)) \leq R(\theta, \delta(X))$. □

This theorem states that, as long as the loss function is convex, every estimator that depends on the data not through the sufficient statistics can be improved. In consequence, in our study of optimal procedures, we can always focus on procedures that are based on the sufficient statistics.

Remark 2.4.3. *In Theorem 2.4.2 we can also show that if $L(\theta, a)$ is strictly convex in a, the inequality will be strict unless $\delta(X) = \eta(T)$ a.e.. A rough argument goes as follows: If f is strictly convex then, by Theorem B.2.3, the inequality in (2.10) becomes strict unless $\delta(X)$ becomes constant a.s. after conditioning on the event $\{T = t\}$ (for every t). This implies that $\delta(X)$ must be a function of T and so $\delta(X) = \mathbb{E}(\delta(X)|T)$ a.s. If this inequality is strict then it remains strict after taking the expectation.*

← Exercise 2.7.11

Example 2.4.4. *Suppose $X_i \sim U(0, \theta)$, $i = 1, \ldots, n$. Consider $X_{(n)} := \max\{X_1, \ldots, X_n\}$. The fact that $X_{(n)}$ is a sufficient statistics follows from the Fisher-Neyman factorization theorem. Indeed, denote $X_{(1)} := \min\{X_1, \ldots, X_n\}$, then the density of the sample $\mathbf{x} = (x_1, \ldots, x_n)$ is*

$$f_\theta(\mathbf{x}) = \frac{1}{\theta^n}\mathbb{1}\{x_{(1)} \geq 0\}\mathbb{1}\{x_{(n)} \leq \theta\},$$

which is equal to $h(\mathbf{x})g_\theta(x_{(n)})$ with $h(\mathbf{x}) = \mathbb{1}\{x_{(1)} \geq 0\}$ and $g_\theta(x_{(n)}) = \frac{1}{\theta^n}\mathbb{1}\{x_{(n)} \leq \theta\}$.

Consider now an unbiased estimator of θ given as

$$\delta(X) = \frac{2}{n}\sum_{i=1}^{n} X_i.$$

Under the quadratic loss

$$R(\theta, \delta) = \mathrm{var}_\theta(\delta(X)) = \frac{4}{n}\mathrm{var}_\theta(X_1) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Consider now the rao-blackwellized version of δ defined as $\eta(X_{(n)}) = \mathbb{E}[\delta(X)|X_{(n)}]$. By Theorem 2.4.2 it dominates δ. To compute this new estimator explicitly, fix i and note that, for every $t > 0$

$$\mathbb{E}[X_i|X_{(n)} = t, X_i < t] = \mathbb{E}[X_i|X_i < t] = \frac{t}{2}.$$

The first equation follows because $X_{(n)}$ is independent of X_i conditionally on the event $\{X_i < X_{(n)}\}$ (in this case the maximum is a function of the remaining variables). The second equality follows simply because the distribution of X_i conditionally on $X_i < t$ is $U(0,t)$, which follows by standard calculations. We have shown that

$$\mathbb{E}[X_i | X_{(n)}, X_i < X_{(n)}] = \frac{X_{(n)}}{2}.$$

Trivially,

$$\mathbb{E}[X_i | X_{(n)}, X_i = X_{(n)}] = X_{(n)}.$$

The probability of the event $\{X_i = X_{(n)}\}$ is $\frac{1}{n}$ and thus

$$\mathbb{E}[X_i | X_{(n)}] = \frac{1}{n}\mathbb{E}[X_i | X_{(n)}, X_i = X_{(n)}] + \frac{n-1}{n}\mathbb{E}[X_i | X_{(n)}, X_i < X_{(n)}] = \frac{n+1}{2n}X_{(n)},$$

which gives

$$\eta(X_{(n)}) = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[X_i | X_{(n)}] = \frac{2n}{n}\frac{n+1}{2n}X_{(n)} = \frac{n+1}{n}X_{(n)}.$$

In Example 2.1.2 we showed that this is an unbiased estimator of θ. We showed that its risk satisfies $R(\theta,\eta) = \frac{\theta^2}{n(n+2)}$. If $n \geq 2$ we then indeed have $R(\theta,\eta) \leq R(\theta,\delta)$ confirming the Rao-Blackwell theorem. Note however that Example 2.1.2 also showed that $\eta(X_{(n)})$ is not admissible!

← Exercise 2.7.11

Theorem 2.4.5. *Let $X_1, \ldots, X_n$ be a random sample from $N(\theta, 1)$. The sample mean $\bar{X}_n$ is an admissible estimator of θ under the squared loss.*

We are not going to prove this result as it is quite technical. The idea follows from the fact that $\bar{X}_n$ is a limit of Bayes rules for priors $\pi_k = N(0,k)$. Indeed, in this case, by Example 2.3.6, the posterior is $N(\frac{k^2}{k^2+1}\bar{x}_n, \frac{k}{k^2+1})$ and so it concentrates around $\bar{x}_n$ for large k. The Bayes rule is $\frac{k^2}{k^2+1}\bar{x}_n$. Since each of these is admissible, with a bit of more care we can argue that the same holds in the limit. More generally, it can be show that every admissible procedure is a limit of Bayes procedures; see Theorem 3.40 [5].

[5] Mark J. Schervish. *Theory of statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1995

2.5 *Stein's paradox*

Consider a sample $X^{(1)}, \ldots, X^{(n)} \sim N_d(\mu, \Sigma)$ with Σ known. Here we assume $\Sigma = \sigma^2 I_d$. In this case $\epsilon := X - \mu \sim N_d(0, \sigma^2 I_d)$. Using the square loss $L(\mu, a) = \|\mu - a\|^2$ we easily show that the risk of any estimator $\widehat{\mu}_n$ admits the following decomposition

This is called the Gaussian sequence model and will reappear in this lecture.

$$R(\mu, \widehat{\mu}_n) = \mathbb{E}\|\widehat{\mu}_n - \mathbb{E}[\widehat{\mu}_n]\|^2 + \|\mathbb{E}[\widehat{\mu}_n] - \mu\|^2.$$

The first term is the variance of $\widehat{\mu}_n$ and the second term is related to its bias. The MLE estimator of μ is the sample average $\bar{X}_n$. Its bias is

zero and its variance is

$$\mathbb{E}\|\tfrac{1}{n}\sum_i (X^{(i)} - \mu)\|^2 \;=\; \tfrac{1}{n}\mathbb{E}\|\epsilon\|^2 \;=\; \sigma^2\frac{d}{n}. \tag{2.11}$$

For simplicity take $\sigma^2 = 1$. In the special case when $n = 1$, the MLE is X which is an unbiased estimator with large variance d. Consider first as an alternative a simple linear estimator $\widehat{\mu}_C = CX$ with $C = \text{diag}(c)$ diagonal and $c = (c_1, \ldots, c_d)$. In this case

$$R(\mu, \widehat{\mu}_C) \;=\; \sum_{i=1}^d (1 - c_i)^2\mu_i^2 + \sum_{i=1}^d c_i^2.$$

Suppose we restrict ourselves to the hyperrectangular model class $|\mu_i| \leq \tau_i$. In this case we can easily find that the minimax risk is

$$\inf_c \sup_{-\tau \leq \mu \leq \tau} R(\mu, \widehat{\mu}_C) \;=\; \inf_c \sum_{i=1}^d (1 - c_i)^2\tau_i^2 + \sum_{i=1}^d c_i^2 \;=\; \sum_{i=1}^d \frac{\tau_i^2}{1 + \tau_i^2} \;<\; d.$$

This computation shows that for sparse model classes diagonal estimators $\widehat{\mu}_C$ may strictly dominate the MLE if c is chosen carefully.

In this section we show a famous surprising result that the observation $X = (X_1, \ldots, X_d)$ is not an admissible estimator for the parameter $\mu = (\mu_1, \ldots, \mu_d)$ unless $d \leq 2$. We start introducing the Stein's Unbiased Risk Estimates (SURE).

2.5.1 Stein's Unbiased Risk Estimates (SURE)

For a function $h : \mathbb{R}^d \to \mathbb{R}^d$ denote by $Jh(\boldsymbol{x})$ the Jacobian of h at $\boldsymbol{x} \in \mathbb{R}^d$, that is, the matrix of partial derivatives with

$$(Jh(\boldsymbol{x}))_{ij} = \frac{\partial h_i}{\partial x_j}.$$

The goal of this section is to prove the following result.

Proposition 2.5.1. *[Stein's Unbiased Risk Estimates (SURE)] Let $X_1, \ldots, X_d$ be independent, $X_i \sim N(\mu_i, 1)$. Consider an estimator $\widehat{\mu}(X)$ of $\mu = (\mu_1, \ldots, \mu_d)$ and let*

$$h(x) = x - \widehat{\mu}(x). \tag{2.12}$$

Suppose that $h(x)$ satisfies

(i) h is differentiable,

(ii) $\mathbb{E}_\mu\|Jh(X)\| < \infty$.

Define

$$\hat{R} \;=\; d + \|h(X)\|^2 - 2\text{tr}(Jh(X)).$$

Then

$$R(\mu, \widehat{\mu}) = \mathbb{E}_\mu\|\widehat{\mu}(X) - \mu\|^2 = \mathbb{E}_\mu\hat{R}.$$

Remark 2.5.2. *Proposition 2.5.1 is true if condition (i) is replaced with (i') h is weakly differentiable. A primary application being soft-thresholding when*

$$h_i(x) \;=\; h_i(x_i) \;=\; x_i - \begin{cases} x_i - \lambda & x_i > \lambda \\ 0 & |x_i| \le \lambda \\ x_i + \lambda & x_i < -\lambda. \end{cases}$$

Here $h_i(x)$ is weakly differentiable in the sense that whenever $-\infty < a \le b < \infty$ there exists $h_i'(x_i)$ such that

$$\int_a^b h_i'(x_i)\mathrm{d}x_i \;=\; h_i(b) - h_i(a).$$

To prove Proposition 2.5.1, we start with the Stein's lemma:

Lemma 2.5.3 (Stein's lemma)**.** *Let $X \sim N(\mu, \sigma^2)$ and let $h : \mathbb{R} \to \mathbb{R}$ be differentiable with $\mathbb{E}|h'(X)| < \infty$. Then*

$$E[(X-\mu)h(X)] \;=\; \sigma^2 \mathbb{E}h'(X).$$

Proof. First assume $\mu = 0, \sigma^2 = 1$. In this case we equivalently show that $\mathbb{E}(Xh(X)) = \mathbb{E}h'(X)$. Without loss of generality assume $h(0) = 0$[6]. The proof is an application of integration by parts. We have

$$(h(x)e^{-x^2/2})' = h'(x)e^{-x^2/2} - xh(x)e^{-x^2/2}$$

and so

$$0 = \mathbb{E}h'(X) - \mathbb{E}(Xh(X)),$$

which establishes the result. The fact that we get zero on the left follows from the fact that $\lim_{x\to\pm\infty} h(x)e^{-x^2/2} = 0$, which can be argued since $\mathbb{E}|h'(X)| < +\infty$. Indeed, first note that

$$|h'(t)|e^{-x^2/2}\mathbb{1}_{[0,x]}(t) \;\le\; |h'(t)|e^{-t^2/2}$$

and the dominating function on the right is integrable by assumption. We also note that the smaller function goes to zero as $x \to \pm\infty$. Thus, we can use the dominating convergence theorem

$$\begin{aligned} \lim_{x\to\pm\infty} h(x)e^{-x^2/2} &= \lim_{x\to\pm\infty} \int_0^x h'(t)e^{-x^2/2}\mathrm{d}t = \lim_{x\to\pm\infty} \int_{-\infty}^{\infty} h'(t)e^{-x^2/2}\mathbb{1}_{[0,x]}(t)\mathrm{d}t \\ &= \int_{-\infty}^{\infty} \lim_{x\to\pm\infty} h'(t)e^{-x^2/2}\mathbb{1}_{[0,x]}(t)\mathrm{d}t \;=\; 0. \end{aligned}$$

This establishes the result in the standard normal case. For general μ and σ define $Z = (X-\mu)/\sigma \sim N(0,1)$. Define $\tilde{h}(z) = h(\mu + \sigma z)$. We have

$$\begin{aligned} \mathbb{E}[(X-\mu)h(X)] &= \sigma\mathbb{E}(Z\,h(\mu+\sigma Z)) = \sigma\mathbb{E}(Z\,\tilde{h}(Z)) \\ &= \sigma\mathbb{E}\tilde{h}'(Z) = \sigma^2\mathbb{E}h'(\mu+\sigma Z) = \sigma^2\mathbb{E}h'(X), \end{aligned}$$

where moving from the first to the second line we used the case $\mu = 0, \sigma = 1$ proved earlier. □

[6] Both sides will not change if we replace h with $h - h(0)$.

We need another technical lemma.

Lemma 2.5.4. *Let $X = (X_1, \ldots, X_d)$ be a random vector with independent entries, $X_i \sim N(\mu_i, 1)$. If $h : \mathbb{R}^d \to \mathbb{R}^d$ satisfiesx $\mathbb{E}\|Jh(X)\| < \infty$. Then*

$$\mathbb{E}((X-\mu)^\top h(X)) \;=\; \mathbb{E}\mathrm{tr}(Jh(X))$$

Proof. Denote by $X_{\setminus i}$ the vector X with X_i removed. Using Lemma 2.5.3, for every $i = 1, \ldots, d$

$$\begin{aligned}\mathbb{E}((X_i-\mu_i)h_i(X)) &= \mathbb{E}\Big(\mathbb{E}[(X_i-\mu_i)h_i(X)|X_{\setminus i}]\Big)\\ &= \mathbb{E}\left(\mathbb{E}\left[\frac{\partial h_i(X)}{\partial x_i}|X_{\setminus i}\right]\right) = \mathbb{E}\left(\frac{\partial h_i(X)}{\partial x_i}\right).\end{aligned}$$

Summing over i we get the result. □

Now Proposition 2.5.1 follows easily.

Proof of Proposition 2.5.1. We have

$$\begin{aligned}R(\mu,\widehat{\mu}) &= \mathbb{E}\left((\widehat{\mu}(X)-\mu)^\intercal(\widehat{\mu}(X)-\mu)\right.\\ &= \mathbb{E}\left((\widehat{\mu}(X)-X+X-\mu)^\intercal(\widehat{\mu}(X)-X+X-\mu)\right.\\ &= \mathbb{E}\|h(X)\|^2 - 2\mathbb{E}((X-\mu)^\intercal h(X)) + \mathbb{E}\|X-\mu\|^2\\ &\overset{Lem\ 2.5.4}{=} \mathbb{E}\|h(X)\|^2 - 2\mathbb{E}(\mathrm{tr}(Jh(X))) + \mathbb{E}\|X-\mu\|^2\\ &= \mathbb{E}\hat{R},\end{aligned}$$

where the last equality follows because $\mathbb{E}\|X-\mu\|^2 \;=\; d$, which follows because $X-\mu$ is standard normal. □

A natural estimator of μ is X and it has constant risk $R(\mu, X) = \mathbb{E}\|X-\mu\|^2 = d$. Although it is unbiased, the variance is large if d is large. The James-Stein estimator of $\mu = (\mu_1, \ldots, \mu_d)$ is defined as

$$\delta_{JS}(X) \;=\; \left(1 - \frac{d-2}{\|X\|^2}\right) X.$$

Theorem 2.5.5. *The risk of the James-Stein estimator is*

$$R(\mu, \delta_{\mathrm{JS}}) = d - \mathbb{E}_\mu\left(\frac{d-2}{\|X\|}\right)^2.$$

In particular, the natural estimator is not admissible if $d \geq 3$.

Proof. For the James-Stein estimator the function h in Proposition 2.5.1 is

$$h(x) = \frac{d-2}{\|x\|^2}x$$

and so

$$\hat{R} = d + \frac{(d-2)^2}{\|X\|^2} - 2\mathrm{tr}(Jh(X)).$$

The diagonal entries of the Jacobian have a simple form

$$\frac{\partial h_i}{\partial x_i} = \frac{d-2}{\|x\|^2} - 2\frac{d-2}{\|x\|^4}x_i^2$$

giving

$$\operatorname{tr}(\mathrm{J}h(X)) = \left(\frac{d-2}{\|X\|}\right)^2 \quad \text{and} \quad \hat{R} = d - \left(\frac{d-2}{\|X\|}\right)^2.$$

By Proposition 2.5.1, if $d \geq 3$

$$R(\mu, \delta_{\mathrm{JS}}) \;=\; \mathbb{E}_\mu \hat{R} \;=\; d - \mathbb{E}_\theta \left(\frac{d-2}{\|X\|}\right)^2 \;<\; d \;=\; R(\mu, X)$$

proving inadmissibility. □

← Exercise 2.7.12

It is interesting to study the risk of a general linear estimator $\widehat{\mu}_C = CX$ for an arbitrary square matrix C. To use SURE note that $h(X) = (I_d - C)X$ and so $\mathrm{J}h(x) = (I_d - C)$ giving $\operatorname{tr}(\mathrm{J}h(x)) = d - \operatorname{tr}(C)$. Therefore

$$R(\mu, \widehat{\mu}_C) \;=\; \mathbb{E}_\mu[d + \|(I_d - C)X\|^2 - 2d + 2\operatorname{tr}(C)] \;=\; \mathbb{E}\|(C - I_d)X\|^2 - d + 2\operatorname{tr}(C). \tag{2.13}$$

We formulate the following result without a proof.

Proposition 2.5.6. *The linear estimator $\widehat{\mu}_C = CX$ is admissible if and only if*

(i) C is symmetric,

(ii) the eigenvalues satisfy $0 \leq \rho_i(C) \leq 1$,

(iii) $\rho_i(C) = 1$ for at most two i.

← Exercise 2.7.13

In many situations in the expression (2.13) $C = C(\lambda)$ depends on a regularization parameter λ so one could find optimal λ^* by minimizing the MSE over λ. One example is given by ridge regression. In this case for some $\boldsymbol{Z} \in \mathbb{R}^{d \times p}$

$$C \;=\; \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda I_d)^{-1}\boldsymbol{Z}^\top \tag{2.14}$$

and the corresponding estimator is obtained as $\widehat{\mu}_Z = \boldsymbol{Z}\widehat{\beta}$ where $\widehat{\beta} \in \mathbb{R}^p$ is obtained by solving ← Exercise 2.7.14

$$\min_\beta \frac{1}{2}\|X - \boldsymbol{Z}\beta\|^2 + \frac{\lambda}{2}\|\beta\|^2, \qquad \lambda > 0.$$

2.6 *Minimizing risk under constraints*

Earlier we saw that finding a decision rule δ that minimizes the risk $R(\theta, \delta)$ uniformly over θ is impossible. In Section 2.3 we saw two common strategies for introducing a global risk and finding optimal decision rules. An alternative approach is to introduce a reasonable constraint on the set of decision rules one is willing to consider. In this case, it can be possible to find a (constrained) δ for which $R(\theta, \delta)$ is minimized uniformly over θ. It is not so surprising that the interesting constraints are often convex. We review some of the most popular.

2.6.1 *Unbiasedness constraints*

An estimator is called unbiased if $\mathbb{E}_\theta \delta(X) = \theta$ for all θ. Sometimes the following definition is used instead.

Definition 2.6.1. *For a loss $L(\theta, a)$ a decision rule δ is **unbiased with respect to** L if*

$$\mathbb{E}_\theta(L(\theta', \delta(X))) \geq \mathbb{E}_\theta(L(\theta, \delta(X))) = R(\theta, \delta) \qquad \text{for all } \theta, \theta' \in \Theta.$$

(Note that on both sides the expectation is taken with respect to $\mathbb{P}_\theta$.)

Exercise 2.7.15 explores the connection between unbiased decision rules and unbiased estimation. Exercise 2.7.17 partially motivates the more general definition.

← Exercise 2.7.15
← Exercise 2.7.17

Proposition 2.6.2. *The set of all (randomized) unbiased decision rules is convex.*

Proof. In exactly the same way as in the proof of Theorem 2.2.3 we can show that $\mathbb{E}_\theta L(\theta', \delta)$ is linear in δ for any θ'. Let $\lambda \in (0, 1)$ and suppose that both δ and δ' are unbiased. It follows that

$$\begin{aligned} \mathbb{E}_\theta L(\theta', (1-\lambda)\delta + \lambda\delta') &= (1-\lambda)\mathbb{E}_\theta L(\theta', \delta) + \lambda \mathbb{E}_\theta L(\theta', \delta') \\ &\geq (1-\lambda)\mathbb{E}_\theta L(\theta, \delta) + \lambda \mathbb{E}_\theta L(\theta, \delta') \\ &= \mathbb{E}_\theta L(\theta, (1-\lambda)\delta + \lambda\delta') \end{aligned}$$

proving that $(1-\lambda)\delta + \lambda\delta'$ is unbiased. □

Though unbiasedness is more general, we shall focus here on the estimation problem. The next result says we need not look at Bayes estimators in this context, because (except in weird cases) they cannot be unbiased.

Proposition 2.6.3. *No unbiased estimator $\delta(X)$ of $\theta \in \Theta \subseteq \mathbb{R}^d$ can be a Bayes estimator (under the square loss) unless the prior π satisfies $\pi(\{\theta : R(\theta, \delta) = 0\}) = 1$.*

Proof. Suppose δ is a Bayes rule (under square loss with respect to π) and is unbiased. By Proposition 2.3.5 and the fact that δ is unbiased:

$$\delta(X) = \mathbb{E}(\theta|X) \qquad \text{and} \qquad \theta = \mathbb{E}(\delta(X)|\theta).$$

Then depending on the order in which we condition, we get

$$\mathbb{E}(\theta^\top \delta(X)) \;=\; \begin{cases} \mathbb{E}[\theta^\top \mathbb{E}(\delta(X)|\theta)] = \mathbb{E}(\theta^\top \theta) \\ \mathbb{E}[\delta(X)^\top \mathbb{E}(\theta|X)] = \mathbb{E}(\delta(X)^\top \delta(X)) \end{cases}.$$

Therefore $\mathbb{E}(\theta^\top \delta(X)) = \mathbb{E}(\theta^\top \theta) = \mathbb{E}(\delta(X)^\top \delta(X))$ and

$$\mathbb{E}\|\delta(X) - \theta\|^2 = \mathbb{E}(\delta(X)^\top \delta(X)) - 2\mathbb{E}(\theta^\top \delta(X)) + \mathbb{E}(\theta^\top \theta) \;=\; 0.$$

Since $r(\pi, \delta) = \mathbb{E}\|\delta(X) - \theta\|^2$ (with the expectation both with respect to X and θ) we get that $r(\pi, \delta) = 0$. But the Bayes risk also satisfies $r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)\mathrm{d}\theta$. Since $R(\theta, \delta) \geq 0$ for all θ, the only way the π-integral can be zero is if π assigns probability 1 to the set of θ where $R(\theta, \delta)$ vanishes. This proves the claim. □

Next we formulate a very powerful result, which states that there is a rule that uniformly minimizes risk over the unbiased estimators and, moreover, gives easily verifiable sufficient conditions to identify this best estimator. We first recall the notion of completness.

Definition 2.6.4. *A statistic T is called **complete** for the model $\mathcal{P} = \{\mathbb{P}_\theta\}$, if for every measurable function g, if $\mathbb{E}_\theta g(T) = 0$ for all θ then $g(T) = 0$ almost surely.*

Although we did not specify this explicitly, a sufficient statistics in a regular exponential family is always complete.

Theorem 2.6.5 (Lehmann-Scheffe)**.** *Let $X \sim \mathbb{P}_\theta$ and suppose that T is a complete sufficient statistic. Suppose the goal is to estimate θ under convex loss, and that an unbiased estimator exists[7]. Then there exists an essentially unique unbiased estimator that is a function of T and uniformly minimizes the risk.*

[7] Note that in this result the estimator is unbiased in the classical sense but the loss function is general.

Proof. We first show that if δ is an unbiased estimator that uniformly minimizes risk then, without loss of generality, we can assume it is a function of T. Let δ be an unbiased estimator and define its rao-blackwellized version $\eta(T) = \mathbb{E}(\delta(X)|T)$ as in Theorem 2.4.2. Unbiasedness of δ gives

$$\theta = \mathbb{E}_\theta \delta(X) = \mathbb{E}_\theta[\mathbb{E}(\delta|T)] = \mathbb{E}_\theta[\eta(T)]$$

and so $\eta(T)$ is unbiased too. Moreover, by Theorem 2.4.2, $R(\theta, \eta) \leq R(\theta, \delta)$ for all θ.

Now we will show that the estimator $\eta(T)$ is essentially unique. Suppose $\eta^*(T)$ is also unbiased. Then

$$\mathbb{E}_\theta(\eta(T) - \eta^*(T)) \;=\; 0 \qquad \text{for all } \theta \in \Theta$$

and by completness $\eta(T) \;=\; \eta^*(T)$ a.s. showing that $\eta(T)$ is essentially unique. □

Note that the proof also shows how such an estimator can be obtained. We simply start with an unbiased estimator and rao-blackwellize it. In Example 2.4.4 we essentially followed this construction. The estimator $\delta(X) = \frac{2}{n}\sum_{i=1}^n X_i$ is unbiased. Here $X_{(n)} = \max\{X_1, \ldots, X_n\}$ is a minimal sufficient statistics and can be showed to be complete. The rao-blackwellized version of this estimator is $\frac{n+1}{n}X_{(n)}$.

Remark 2.6.6. *We may have a situation when no unbiased estimator exists; e.g. estimating θ in a* $\mathrm{Bin}(n, \frac{1}{\theta})$. *Indeed, we would require*

$$\theta \;=\; \mathbb{E}\delta(X) \;=\; \sum_{k=0}^{n} \delta(k)\binom{n}{n}\tfrac{1}{\theta^k}(1-\tfrac{1}{\theta})^{n-k}.$$

After multiplying by θ^n we get that

$$\theta^{n+1} = \sum_{k=0} \delta(k)^n \binom{n}{k}(\theta-1)^{n-k},$$

which is impossible to hold for all $\theta \in (0,1)$ irrespective of δ because on the right we have a polynomial of order n.

In the special case when $\theta \in \mathbb{R}$ with the square loss function, if δ is unbiased then $R(\theta, \delta) = \mathrm{var}(\delta(X))$. In this case δ in Theorem 2.6.5 is called the **unbiased estimator with uniformly minimum variance (UMVU)**.

Example 2.6.7 (Finding UMVU)**.** *Let $X_1, \ldots, X_n$ be i.i.d.* $\mathrm{Exp}(\mu, 1)$. *So*

$$p(x;\mu) = \begin{cases} \exp(-(x-\mu)) & x \geq \mu \\ 0 & x < \mu. \end{cases}$$

Both $T_1 \;=\; X_{(1)} - \frac{1}{n}$ and $T_2 \;=\; \frac{1}{n}\sum_i X_i - 1$ are possible estimators and both are unbiased. Moreover, $\mathrm{var}(T_1) \;=\; \frac{1}{n^2} \;\ll\; \mathrm{var}(T_2) \;=\; \frac{1}{n}$. Note that T_1 is a function of a minimal sufficient statistics. To show that $T = X_{(1)}$ is a sufficient statistics we use the Fisher-Neyman factorization theorem and note that the distribution of the data satisfies

$$\prod_{i=1}^{n} p(x;\mu) = e^{-\sum_i x_i} e^{n\mu} \mathbb{1}\{T \geq \mu\}$$

and so indeed T is sufficient. To show that T is complete, note that the density of T is

$$p_T(t;\mu) = \begin{cases} n\exp(n(\mu - t)) & t \geq \mu \\ 0 & t < \mu. \end{cases}$$

Let $g(T)$ be a measurable function such that $\mathbb{E}_\mu g(T) = 0$ for all $\mu > 0$. Equivalently, for all $\mu > 0$

$$G(\mu) = e^{n\mu} \int_\mu^\infty e^{-nt} g(t) \mathrm{d}t = 0.$$

In particular, G is almost everywhere differentiable with $G'(\mu) = 0$ By the fundamental theorem of calculus,

$$0 = G'(\mu) = ne^{\mu n} \int_\mu^\infty e^{-nt} g(t) \mathrm{d}t + e^{n\mu}(-e^{-n\mu} g(\mu)) = g(\mu)$$

implying that T must be complete. Thus, T_1 is actually the UMVU.

2.6.2 *Equivariance constraints**

Equivariance is a classical topic in statistics. Recently it also got a lot of attention in machine learning as equivariant convolutional networks grew popular. The idea is very simple and we will explain it first on a concrete example. Suppose the exercise is to classify pictures into one of the categories. Pictures may get rotated and we do not want the label to depend on this rotation. In other words, we want the procedure to be **invariant** under picture rotations. Suppose now that before running the classification exercise, we first reduce the quality of the images in order to save space. In that case, we may want to require that the compression algorithm on a rotated image outputs rotation of the compression of the original image. In that case, we say that the procedure is **equivariant**.

In order to generalize this simple example, we formalize the set-up. Recall that a **group** is a set G with a binary operation "$\cdot$" and identity element $e \in G$ such that:

(i) $g \cdot h \in G$ for all $g, h \in G$,

(ii) $(g \cdot h) \cdot k = g \cdot (h \cdot k)$ for all $g, h, k \in G$. (associativity)

(iii) $g \cdot e = e \cdot g = g$ for all $g \in G$,

(iv) for every $g \in G$ there exists $h \in G$ such that $g \cdot h = h \cdot g = e$, we write $h = g^{-1}$.

The following examples will be important in the sequel.

Example 2.6.8. *The set of real numbers* $\mathbb{R}$ *with addition forms a group. Its identity element is* 0*. We denote this group by* $(\mathbb{R}, +)$*.*

Example 2.6.9. *The group of permutations of the set* $\{1, \ldots, m\}$*.*

Example 2.6.10. *The set of real invertible* $m \times m$ *matrices forms a group under the matrix multiplication with the identity element given by the identity matrix. This group is sometimes called the generalized linear group denoted* $\mathrm{GL}_m(\mathbb{R})$*.*

Example 2.6.11. *The group* $\mathrm{SO}(m)$ *of rotations in* $\mathbb{R}^m$*. The group is a subgroup of* $\mathrm{GL}_m(\mathbb{R})$*.*

The groups in this section will act on $\mathcal{X}$, that is, each $g \in G$ defines a function $g : \mathcal{X} \to \mathcal{X}$, which (perhaps abusing notation a bit) we denote in the same way, and:

1. $e(x) = x$ for all $x \in \mathcal{X}$,
2. $(gh)(x) = g(h(x))$ for all $g, h \in G$ and $x \in \mathcal{X}$.

Following the standard algebraic notation, we write $g \cdot x$ for this action. For example, the group $\mathrm{GL}_m(\mathbb{R})$ acts on $\mathbb{R}^m$ by the matrix multiplication $\mathbf{x} \mapsto g \cdot \mathbf{x}$ for a matrix $g \in \mathrm{GL}_m(\mathbb{R})$. The group $(\mathbb{R}, +)$ also acts on $\mathbb{R}^m$ by translations $\mathbf{x} \mapsto \mathbf{x} + c\mathbf{1}$ for $c \in (\mathbb{R}, +)$. The group of permutations acts on $\mathbb{R}^m$ by permuting the coordinates. If σ is a permutation of $\{1, \ldots, m\}$ then the corresponding transformation of $\mathbb{R}^m$ is $\mathbf{x} \mapsto (x_{\sigma(1)}, \ldots, x_{\sigma(m)})$.

The same abstract group can act on different sets. Fix a sample space $\mathcal{X} \subseteq \mathbb{R}^m$ and a group G that acts on it. Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a model for a random variable $X \in \mathcal{X}$ with the identifiable parameter θ (i.e. $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ implies $\theta = \theta'$). We say that $\mathcal{P}$ is a **group transformation model with respect to** G if it holds that:

$$X \sim \mathbb{P}_\theta \qquad \text{then} \qquad gX \sim P_{\theta'} \quad \text{for some } \theta' \in \Theta.$$

(the model is invariant under the group action) This particular θ' then determined by θ and the transformation g. In other words, the transformations g also act on θ. We will use the same notation for the action on $\mathcal{X}$ and on Θ:

$$X \sim \mathbb{P}_\theta \qquad \Longleftrightarrow \qquad g \cdot X \sim P_{g \cdot \theta}.$$

Example 2.6.12. *Let* $\mathbb{P}_0$ *be a probability measure with symmetric density* p_0 *with respect to the Lebesgue measure on* $\mathbb{R}$ *(the mean of this distribution is zero if it exists). For* $X \sim \mathbb{P}_0$ *and* $\theta \in \mathbb{R}$*, let* $\mathbb{P}_\theta$ *be the distribution of* $X + \theta$*. Doing this for all* θ *generates the family* $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$*. The parameter* θ *is called the* ***location parameter****. The normal family* $N(\theta, 1)$ *is a special case. Here the group is* $(\mathbb{R}, +)$ *and we have*

$$X \sim \mathbb{P}_\theta \qquad \Longleftrightarrow \qquad X + c \sim P_{\theta + c}.$$

Example 2.6.13. *Extending the previous example, suppose that $\mathbb{P}_{0,1}$ is a distribution of a univariate random variable X with mean zero and variance 1. Denote by $\mathbb{P}_{\mu,\sigma}$ the distribution of $\sigma X + \mu$ for $\mu \in \mathbb{R}$ and $\sigma > 0$. Then (μ, σ) is called the* ***location-scale parameter*** *for X. This setup is generalized to a random vector X in $\mathbb{R}^m$ by considering the distributions of $\sigma X + \mu\mathbf{1}$. In the vector case it is more suitable to consider a location-scale family by taking $\boldsymbol{\mu} \in \mathbb{R}^m$ and $U \in \mathbb{S}_+^m$ and considering the induced distributions of $UX + \boldsymbol{\mu}$. What is the associated group?*

Example 2.6.14. *Suppose X has a m-variate zero-mean Gaussian distribution with covariance matrix Σ. The group $\mathrm{GL}_m(\mathbb{R})$ or $m \times m$ invertible matrices acts on $\mathcal{X} = \mathbb{R}^m$, $\mathbf{x} \mapsto A\mathbf{x}$, $A \in \mathrm{GL}_m(\mathbb{R})$. If X is Gaussian with covariance Σ, then AX is Gaussian with covariance $A\Sigma A^\intercal$ and so the action of $\mathrm{GL}_m(\mathbb{R})$ on $\Theta = \mathbb{S}_+^m$ is $A \cdot \Sigma = A\Sigma A^\top$. We have*

$$X \sim \mathbb{P}_\Sigma \qquad \Leftrightarrow \qquad AX \sim \mathbb{P}_{A\Sigma A^\top}.$$

Example 2.6.15. *A slightly more involved example considers a simple Gaussian graphical model given by three-dimensional centered Gaussian distributions with the inverse covariance matrix satisfying $K_{13} = 0$. The group is given by all invertible matrices of the form $g \cdot h$, where*

$$h \in \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \right\}, \qquad g = \begin{bmatrix} * & * & 0 \\ 0 & * & 0 \\ 0 & * & * \end{bmatrix}.$$

The action on the parameter space is the same as in the previous example. (verify the details)

Consider a decision procedure in the situation when there is a group acting on $\mathcal{X}$ and Θ as described above. As it was noted in the introductory example, if we have a group acting on the sample space, we may want some statistical procedures like testing or classification to be invariant with respect to this group action. A decision rule δ is called **invariant** if $\delta(g \cdot x) = \delta(x)$ for all $x \in \mathcal{X}$ and all $g \in G$.

Example 2.6.16. *Suppose $X \sim \mathbb{P}_\theta$ and the goal is to test hypothesis $\theta \in \Theta_0$ versus $\theta \in \Theta$, $\mathcal{A} = \{0, 1\}$. Suppose that a group G acts on $\mathcal{X}$ and on Θ_0, that is, $X \sim \mathbb{P}_\theta$ for $\theta \in \Theta_0$ then $g \cdot X \sim \mathbb{P}_{\theta'}$ for $\theta' \in \Theta_0$. Intuitively, if for data X we accept/reject the null, we should also accept/reject it for data $g \cdot X$. We will be then interest in test procedures $\delta : \mathcal{X} \to \{0, 1\}$ that are invariant.*

Example 2.6.17. *Consider a classification function $\delta : \mathcal{X} \to \{1, \ldots, c\}$. In many applications, there is a group of transformations of $\mathcal{X}$ and we want the classification procedure to classify each example the same way as its transformed version (think images and their rotated versions). We again will requite that δ is invariant.*

In estimation of group transformation models we have $\mathcal{A} = \Theta$ and in this case G acts on $\mathcal{A}$ in the same way as it acts on Θ. In this case we may want to restrict to procedures that are equivariant.

Definition 2.6.18. *A function $\delta : \mathcal{X} \to \Theta$ is **equivariant** if $\delta(g \cdot x) = g \cdot \delta(x)$.*

Example 2.6.19. *Let $\mathcal{X} = \mathbb{R}^m$ and define maps $g_c(\mathbf{x}) = \mathbf{x} + c\mathbf{1}$, the location shifts. These transformations form a group. The function $\beta(\mathbf{x}) = \bar{\mathbf{x}}\mathbf{1}$ is equivariant where $\bar{\mathbf{x}}$ is the average of entries of $\mathbf{x}$. The function $\alpha(\mathbf{x}) = \mathbf{x} - \bar{\mathbf{x}}\mathbf{1}$ is invariant.*

Although this plays no role here, we note that equivariance is also a convex constraint under a minor condition. ← Exercise 2.7.18

The last ingredient in designing good equivariant procedures is a loss function that is also amenable to this group action setting.

Definition 2.6.20. *The loss function $L : \Theta \times \mathcal{A} \to [0, \infty]$ is called invariant (with respect to G) if, for each $g \in G$, $\theta \in \Theta$, and $a \in \mathcal{A}$, $L(g \cdot \theta, g \cdot a) = L(\theta, a)$.*

Example 2.6.21. *[Example 2.6.14 continued] Let*

$$L(\Sigma, S) = -\log\det(S\Sigma^{-1}) + \operatorname{tr}(S\Sigma^{-1} - I_m)$$

be the Kullback-Leibler divergence between two mean-zero Gaussian distributions with covariances S and Σ. Basic matrix algebra gives that

$$L(g\Sigma g^\top, gSg^\top) = L(\Sigma, S).$$

Thus, in the problem of estimating Σ we have $\mathcal{A} = \mathbb{S}_+^m$ and $L(\Sigma, S)$ is a valid invariant loss function.

← Exercise 2.7.19

Definition 2.6.22. *An invariant decision problem is when: $\mathcal{P}$ is a group transformation model, δ is equivariant, and L is invariant.*

We view the insistence that the decision rule be equivariant as a constraint on the possible decision rules, just like unbiasedness is a constraint. Then the question is if there is an equivariant rule that uniformly minimizes the risk. The first result is a step in this direction.

Theorem 2.6.23. *In an invariant decision problem, the risk function $R(\theta, \delta)$ of an equivariant decision rule δ is an invariant function on Θ.*

Proof. We have

$$\begin{aligned} R(g \cdot \theta, \delta) &= \mathbb{E}_{g\cdot\theta}(L(g \cdot \theta, \delta(X))) = \mathbb{E}_{g\cdot\theta}(L(\theta, g^{-1} \cdot \delta(X))) \\ &= \mathbb{E}_{g\cdot\theta}(L(\theta, \delta(g^{-1} \cdot X))) = \mathbb{E}_{\theta}(L(\theta, \delta(X))), \end{aligned}$$

where the last equation follows from the fact that $X \sim \mathbb{P}_{g\cdot\theta}$ if and only if $g^{-1} \cdot X \sim \mathbb{P}_\theta$. □

This theorem really shows that $R(\theta, \delta)$ is constant on G-orbits in Θ, that is, the sets

$$G \cdot \theta = \{\theta' \in \Theta : \theta' = g \cdot \theta \text{ for some } g \in G\}.$$

This gives a plausible justification for restricting attention to equivariant decision rules. Since the risk function is constant on orbits in Θ when the loss is invariant, this makes it easier to compare equivariant rules by means of their risk functions. In particular, if the group acts transitively on the parameter space (i.e. there is only one orbit), then the problem of noncomparability of risk functions disappears altogether as the risk function becomes constant on δ. This happens in some interesting situations, e.g. for the location model.

Theorem 2.6.24 (Pitman's estimator)**.** *Suppose that* $Y = (X_1 - X_n, \ldots, X_{n-1} - X_n)$ *and* $L(\theta, a) = (\theta - a)^2$*. Suppose that* δ_0 *is a location equivariant estimator with finite risk. Then, the equivariant estimator with smallest risk is* $\delta_0(X) - \mathbb{E}_0[\delta_0(X)|Y]$*.*

Proof. If δ_0 is a location equivariant estimator with finite risk then all other equivariant estimators have the form $\delta_0(X) - v(Y)$. Indeed, δ is equivariant if and only if $\delta - \delta_0$ is invariant. So it remains to show that every invariant function f is a function of Y. This follows because

$$f(\mathbf{x}) = f((\mathbf{y}, 0) + x_n \mathbf{1}) = f(\mathbf{y}, 0) = v(\mathbf{y}).$$

By Theorem 2.6.23, the risk function is invariant in θ for an equivariant δ, and so

$$R(\theta, \delta) = R(0, \delta) = \mathbb{E}_0[\delta_0(X) - v(Y)]^2 = \mathbb{E}_0(\mathbb{E}_0[(\delta_0(X) - v(Y))^2|Y]),$$

which is minimized by minimizing $\mathbb{E}_0[(\delta_0(X) - v(Y))^2|Y = y]$ uniformly in y. This is accomplished by choosing $v(y) = \mathbb{E}_0[\delta_0(X)|Y = y]$. □

Like with the Lehmann-Scheffe theorem, we get an explicit procedure of obtaining the best equivariant estimator by improving any given equivariant estimator.

← Exercise 2.7.20

It can be shown that the Pitman's estimator is the generalized Bayes rule with respect to the uniform "prior" distribution. This fact and Theorem 2.6.24 can be both generalized; see Section 6.2.3 in [8]; we will not provide any more details.

[8] Mark J. Schervish. *Theory of statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1995

Theorem 2.6.25. *Consider an invariant decision problem. Under some assumptions, if the formal Bayes rule with respect to the right invariant Haar prior on G exists, then it is the minimum risk equivariant rule.*

The following case of the truncated normal distribution shows the power of this result. Using Theorem 2.6.24 directly is rather tedious. Computing the generalized Bayes rule $\mathbb{E}(\theta|X)$ in this case is straightforward.

← Exercise 2.7.21

2.6.3 *Type I error constraints*

Recall the testing problem in Example 2.2.2, which is a particular instance of a decision problem with $\mathcal{A} = \{0, 1\}$ and o/1-loss

$$L(\theta, a) \;=\; \mathbb{1}(a = 1, \theta \in \Theta_0) + \mathbb{1}(a = 0, \theta \in \Theta_1) \tag{2.15}$$

which is associated to testing two competing hypotheses $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. The data are $X \sim \mathbb{P}_\theta$ for $\theta \in \Theta_0 \cup \Theta_1$.

It is usual to find tests that bound the **type I error**, defined as $\sup_{\theta\in\Theta_0} \beta(\theta)$, on some level α, where the power function $\beta(\cdot)$ is defined in (2.3). By the risk calculation in Example 2.2.2,

$$\sup_{\theta\in\Theta_0} \beta(\theta) \;=\; \sup_{\theta\in\Theta_0} R(\theta, \delta).$$

On the other hand

$$\sup_{\theta\in\Theta_1} R(\theta, \delta) \;=\; \sup_{\theta\in\Theta_1} (1 - \beta(\theta)) \;=\; 1 - \inf_{\theta\in\Theta_1} \beta(\theta),$$

which is precisely the **type II error**. It follows that

$$\overline{R}(\delta) \;=\; \sup_{\theta\in\Theta_0\cup\Theta_1} R(\theta, \delta) \;=\; \max\{\sup_{\theta\in\Theta_0} R(\theta, \delta), \sup_{\theta\in\Theta_1} R(\theta, \delta)\}.$$

By "bounding the type I error" we mean optimizing $\overline{R}(\delta)$ by restricting to test procedures δ with an explicit constraint on the type I error:

$$\sup_{\theta\in\Theta_0} R(\theta, \delta) \;\le\; \alpha.$$

← Exercise 2.7.22

Definition 2.6.26. *A test φ^* with level α is called **uniformly most powerful (UMP)** if*

$$\mathbb{E}_\theta \varphi^* \;\ge\; \mathbb{E}_\theta \varphi, \qquad \text{for all } \theta \in \Theta_1$$

for all φ with level at most α.

As we will see, test of this form may appear only in very special situations and mostly in the univariate case. A particularly famous instance is discussed in Section 3.1.

2.7 *Exercises*

Exercise 2.7.1. *Show that if $f : \mathcal{A} \to [-\infty, \infty]$ for $\mathcal{A} \subseteq \mathbb{R}^d$ is lower semicontinuous then it is measurable. Hint: Use Exercise A.2.8.*

Exercise 2.7.2. *Suppose that $x_1, \ldots, x_n \in \mathbb{R}$. Show that the minimizer of $f(\theta) = \frac{1}{n}\sum_{i=1}^n (\theta - x_i)^2$ is the average $\frac{1}{n}\sum_{i=1}^n x_i$. Moreover, show that the median of the collection $x_1, \ldots, x_n$ minimizes $f(\theta) = \frac{1}{n}\sum_{i=1}^n |\theta - x_i|$.*

Exercise 2.7.3. *Show that the estimator δ_3 in Example 2.1.1 is a Bayes rule for the quadratic loss and a beta distribution.*

The next two exercises allow you to explore a similar situation in the case when the loss function is not differentiable.

Exercise 2.7.4. *In the one-dimensional case, show that if $L(\theta, a) = |\theta - a|$ then the Bayes estimator is the median of the posterior distribution $\pi(\theta|x)$ (c.f. Exercise 2.7.2).*

Exercise 2.7.5. *Show that if $L(\theta, a)$ is the $0/1$-loss then the Bayes estimator is the mode of the posterior distribution $\pi(\theta|x)$.*

Exercise 2.7.6. *In the setting of Theorem 2.3.7 show that π is least favourable.*

Exercise 2.7.7. *Show that in Section 2.3.5 we always have $L^* \leq U^*$.*

Exercise 2.7.8. *Let $(X_1, \ldots, X_n)$ be a random sample of binary random variables with $X_i \sim \text{Bern}(\theta)$ with $\theta \in (0, 1)$.*

(i) Show that the sample mean $\bar{X}_n$ is an admissible estimator of θ under the loss function $L(\theta, a) = (a - \theta)^2/[\theta(1 - \theta)]$.

(ii) Show that $\overline{X}$ is an admissible estimator of θ under the squared error loss.

Exercise 2.7.9. *Show that if a minimax rule is essentially unique then it is admissible.*

Exercise 2.7.10. *Show the following result: If risk functions for all decision rules are continuous in θ, if δ is Bayes for π and has finite integrated risk $r(\pi, \delta) < +\infty$, and if the support of π is the whole parameter space, then δ is admissible. Hint: This looks much more complicated than it actually is.*

Exercise 2.7.11. *Show that if $\delta(X)$ is an unbiased estimator, $\mathbb{E}_\theta \delta(X) = \theta$, then $\eta(T)$ is also unbiased.*

Exercise 2.7.12. *Consider the special case when $\theta = 0$. In this case $\|X\|^2 \sim \chi^2_d$. Show that $\mathbb{E}_0 \frac{1}{\|X\|^2} = \frac{1}{d-2}$ and so*

$$R(0, \delta_{\text{JS}}) \;=\; 2 \;\ll\; d.$$

Exercise 2.7.13. *Suppose that $\boldsymbol{Z} \in \mathbb{R}^{d \times p}$ is a fixed matrix of full column rank (in particular $p \leq d$). Consider a linear estimator of μ given by $\widehat{\mu} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top X$. Compute the risk of this estimator. Discuss conditions when the risk of this estimator can be significantly smaller given $p << d$.*

Exercise 2.7.14. *Show that the linear estimator $\widehat{\mu}_C$ with C defined in (2.14) satisfies the conditions of Proposition 2.5.6 and so it is admissible.*

Exercise 2.7.15. *Show that the decision rules that are unbiased with respect to the square loss $L(\theta, \delta) = \|\theta - \delta(X)\|^2$ are the rules satisfying $\mathbb{E}_\theta \delta(X) = \theta$.*

Exercise 2.7.16. *Suppose $Y \sim N_n(\mu, I_n)$. Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be a fixed matrix with full column rank. In this exercise we consider estimators of the form $\hat{\mu} = \boldsymbol{X}\hat{\beta}$ for some estimator $\hat{\beta}$. More specifically, we study the least squares estimator $\hat{\beta}^{\mathrm{LS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top Y$.*

(i) Find the formula for the risk $R(\mu, \hat{\mu}_C)$ of a general linear estimator $\hat{\mu}_C = CY$.

(ii) Let $\widehat{\mu}^{\mathrm{LS}} = \boldsymbol{X}\widehat{\beta}^{\mathrm{LS}}$ be the corresponding estimator of the mean of Y. Find the risk of this estimator directly in terms of $C_0 = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, μ, n, and p only.

(iii) Suppose $\mu = \boldsymbol{X}\beta^*$ for some $\beta^* \in \mathbb{R}^p$. Show that $R(\mu, \hat{\mu}^{\mathrm{LS}}) = p$.

(iv) Consider the ridge estimator $\hat{\mu}^{\mathrm{ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \delta I_p)^{-1}\boldsymbol{X}^\top Y$ and the corresponding matrix $C_\delta = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \delta I_p)^{-1}\boldsymbol{X}^\top$, where $\delta \geq 0$. Show that $R(\mu, \hat{\mu}^{\mathrm{ridge}}) = \mathrm{tr}(C_\delta^2) + \|(I_n - C_\delta)\mu\|^2$.

(v) Suppose that $\boldsymbol{X}^\top \boldsymbol{X} = I_p$ and that $\mu = \boldsymbol{X}\beta^*$ for some β^*. Show that for every $\beta^* \neq 0$ there exists $\delta > 0$ such that $R(\boldsymbol{X}\beta^*, \hat{\mu}^{\mathrm{ridge}}) < R(\boldsymbol{X}\beta^*, \hat{\mu}^{\mathrm{LS}})$.

Exercise 2.7.17. *Consider estimation in a regular exponential family and let the action space $\mathcal{A}$ be its set of mean parameters. Consider the loss function given by the Kullback-Leibler divergence in (1.18):*

$$L(\boldsymbol{\theta}_2, \boldsymbol{\mu}_1) = \mathrm{K}(\boldsymbol{\mu}_1, \boldsymbol{\theta}_2) = A^*(\boldsymbol{\mu}_1) + A(\boldsymbol{\theta}_2) - \langle \boldsymbol{\theta}_2, \boldsymbol{\mu}_1 \rangle.$$

Given a sample $X_1, \ldots, X_n$, consider the sample average of the sufficient statistics $\delta(X_1, \ldots, X_n) = \overline{\boldsymbol{\mu}}_n$. The minimizer $\widehat{\boldsymbol{\theta}}_n$ of $L(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n)$ is the MLE of the canonical parameter. Although $\overline{\boldsymbol{\mu}}_n$ is an unbiased estimator (in the usual sense) of the true mean parameter $\boldsymbol{\mu}$, this is typically not true for $\widehat{\boldsymbol{\theta}}_n$. Show that $\overline{\boldsymbol{\mu}}_n$ is an unbiased decision rule in the sense of Definition 2.6.1.

Exercise 2.7.18. *Show that the set of equivariant randomized decision rules is convex (with convex combination defined as in (2.4)). Discuss some natural conditions under which the set of non-randomized equivariant procedures also forms a convex set.*

Exercise 2.7.19. *Show that the squared loss function is invariant for the translation family in Example 2.6.12.*

Exercise 2.7.20. *Show that for the Gaussian model $N(\theta, 1)$ the Pitman's estimator is the sample mean.*

Exercise 2.7.21. *Let $(X_1, \ldots, X_n)$ be a random sample of random variables with the Lebesgue density*

$$f_\theta(x) = \begin{cases} \sqrt{\frac{2}{\pi}} e^{-(x-\theta)^2/2} & \text{if } x \geq \theta, \\ 0 & \text{otherwise} \end{cases}$$

where $\theta \in \mathbb{R}$ is unknown. Find the minimum risk location equivariant estimator of θ under the squared loss.

Exercise 2.7.22. *Show that bounding the type I error in Section 2.6.3 gives a convex constraint on $\delta \in \mathcal{D}$.*

3
Hypothesis testing and multiple testing (2 weeks)

Our goal in this section is not to give an extensive treatment of hypothesis testing procedures. We trust that much of it was covered in earlier courses. This includes the basic philosophy of constructing statistical tests (rejection regions, p-values etc).

Given data $X \sim \mathbb{P}_\theta$ and a model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, consider two competing hypothesis: $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$. Unless otherwise stated we assume that $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$. In this context, $\mathcal{A} = \{0, 1\}$ and we are interested in statistical testing procedures $\delta : \mathcal{X} \to \{0, 1\}$. In consequence, every **nonrandomized test** can be associated with a measurable set $S \subseteq \mathcal{X}$ such that the decision rule becomes: accept H_1 if $X \in S$ and accept H_0 when $X \notin S$. From the theoretical perspective it is also useful to allow for **randomized tests**, in which case the **critical function** φ to every $x \in \mathcal{X}$ is assigns the probability of rejecting H_0 given $X = x$

$$\varphi(x) \;=\; \mathbb{P}(\text{reject } H_0 | X = x),$$

where φ is assumed to be a measurable function from $\mathcal{X}$ to $[0, 1]$. A nonrandomized test is a randomized test with $\varphi(X) = \mathbb{1}(X \in S)$.

Recall from Example 2.2.2 that the **power function** is the function $\beta : \Theta \to [0, 1]$ defined by

$$\beta(\theta) := \mathbb{P}_\theta(\text{reject } H_0) \;=\; \int \varphi(x) \mathrm{d}\mathbb{P}_\theta \;=\; \mathbb{E}_\theta \varphi(X).$$

Using the 0/1-loss we see that the risk function satisfies

$$R(\theta, \delta) \;=\; \begin{cases} \beta(\theta) & \text{if } \theta \in \Theta_0 \\ 1 - \beta(\theta) & \text{if } \theta \in \Theta_1 \end{cases}.$$

In the idealised situation we would have $\beta(\theta) = 0$ if $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for $\theta \in \Theta_1$. This is obviously not possible in most of the cases. For example, when all $\mathbb{P}_\theta$ have the same support. However, a good test would try to be close to that. We define the **significance**

level of a test as

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \in \Theta_0} \mathbb{E}_\theta \varphi(X).$$

The main technical convenience of randomized tests comes from the fact that the set of critical functions is a convex set.

3.1 *Neyman-Pearson lemma**

In this section we discuss a classical result in the case when H_0 and H_1 are both simple with the underlying density functions p_0 and p_1 with respect to some underlying measure μ. The type I error is

$$\mathbb{E}_0 \varphi(X) \; = \; \int \varphi(x) p_0(x) \mathrm{d}\mu(x)$$

and another quantity of interest is

$$\mathbb{E}_1 \varphi(X) = \int \varphi(x) p_1(x) \mathrm{d}\mu(x).$$

A good test will result in small type I error (small $\mathbb{E}_0 \varphi$) and small type II error (big $\mathbb{E}_1 \varphi$). We want to explore to what extend this can be achived.

As discussed in Section 2.6.3, a natural approach is to consider the optimization problem:

$$\text{maximize} \quad \mathbb{E}_1 \varphi(X) \qquad \text{subject to } \mathbb{E}_0 \varphi(X) \leq \alpha.$$

We call such test the most powerful.

Consider the **likelihood ratio statistic**

$$\mathrm{LR}(x) \; = \; \frac{p_1(x)}{p_0(x)},$$

where we define $\mathrm{LR}(x) = +\infty$ if $p_1(x) > 0$, $p_0(x) = 0$, and $\mathrm{LR}(x) = 0$ if $p_0(x) = p_1(x) = 0$. For any $\lambda \geq 0$ we consider the critical function

$$\varphi_\lambda(x) \; := \; \begin{cases} 1 & \text{if } \mathrm{LR}(x) > \lambda \\ \gamma & \text{if } \mathrm{LR}(x) = \lambda \\ 0 & \text{if } \mathrm{LR}(x) < \lambda. \end{cases} \tag{3.1}$$

Note that, if $\gamma \in (0,1)$, the corresponding test is randomized, as we need to toss a γ-coin when we observe $\mathrm{LR}(x) = \lambda$. Our freedom to choose γ will add extra flexibility in a second.

Constructing a test of size α is not a problem. It is enough to toss an α-coin to decide on the decision. However, this procedure does not depend on the data so it is clear that its power will be low.

Theorem 3.1.1 (Neyman-Pearson Lemma). *Given any level $\alpha \in [0,1]$ there exists a likelihood ratio test φ with level α. Any likelihood ratio test with level α maximizes $\mathbb{E}_1\varphi$ among all tests with level $\leq \alpha$.*

Proof. To prove the first part note that for $\alpha = 0$ we can take $\lambda = +\infty$, and for $\alpha = 1$ we can take $\gamma = 1$, $\lambda = 0$ (check carefully). Now take $\alpha \in (0,1)$. By construction, $\mathbb{P}_0(\mathrm{LR} = \infty) = 0$ and so LR is finite $\mathbb{P}_0$-a.s. We claim that this implies that

$$\exists \lambda < \infty \quad \text{such that} \qquad \mathbb{P}_0(\mathrm{LR} > \lambda) \leq \alpha \quad \text{and} \quad \mathbb{P}_0(\mathrm{LR} \geq \lambda) \geq \alpha. \tag{3.2}$$

Before we continue, take a look at Figure 3.1. As we showed in Proposition C.1.2, the survival function $G(\lambda) = \mathbb{P}_0(\mathrm{LR} > \lambda)$ is right-continuous. Thus taking

$$\lambda := \inf_{t \geq 0}\{t : \mathbb{P}_0(\mathrm{LR} > t) \leq \alpha\}$$

we get $\mathbb{P}_0(\mathrm{LR} > \lambda) \leq \alpha$. By Remark C.1.3, the function $\mathbb{P}_0(\mathrm{LR} \geq \lambda)$ is left-continuous in λ. Since, for every $\epsilon > 0$

$$\mathbb{P}_0(\mathrm{LR} \geq \lambda - \epsilon) \geq \mathbb{P}_0(\mathrm{LR} > \lambda - \epsilon) > \alpha,$$

we get that $\lim_{t \to \lambda^-} \mathbb{P}_0(\mathrm{LR} \geq t) = \mathbb{P}_0(\mathrm{LR} \geq \lambda) \geq \alpha$ confirming (3.2).

By properly randomizing our procedure, we can obtain level exactly α. If $\mathbb{P}_0(\mathrm{LR} = \lambda) = 0$ then $\mathbb{P}_0(\mathrm{LR} > \lambda) = \mathbb{P}_0(\mathrm{LR} \geq \lambda) = \alpha$ and so the likelihood ratio test for this λ has level exactly α. If $\mathbb{P}_0(\mathrm{LR} = \lambda) > 0$, we define

$$\gamma = \frac{\alpha - \mathbb{P}_0(\mathrm{LR} > \lambda)}{\mathbb{P}_0(\mathrm{LR} \geq \lambda) - \mathbb{P}_0(\mathrm{LR} > \lambda)} \in (0,1)$$

and it is straightforward to check that $\mathbb{E}_0\varphi_\lambda(X) = \alpha$ with this choice of γ.

By the first part, there exists $\lambda \geq 0$ (and $\gamma \in [0,1]$) such that φ_λ has size α. Let φ be any test with level $\leq \alpha$. We have

$$\begin{aligned}
\mathbb{E}_1\varphi(X) &\leq \mathbb{E}_1\varphi(X) - \lambda(\mathbb{E}_0\varphi - \alpha) = \int \varphi(x)(p_1(x) - \lambda p_0(x))\mu(\mathrm{d}x) + \lambda\alpha \\
&\leq \int \varphi_\lambda(x)(p_1(x) - \lambda p_0(x))\mu(\mathrm{d}x) + \lambda\alpha \\
&\leq \mathbb{E}_1\varphi_\lambda(X) - \lambda(\mathbb{E}_0\varphi_\lambda - \alpha) = \mathbb{E}_1\varphi_\lambda(X).
\end{aligned}$$

□

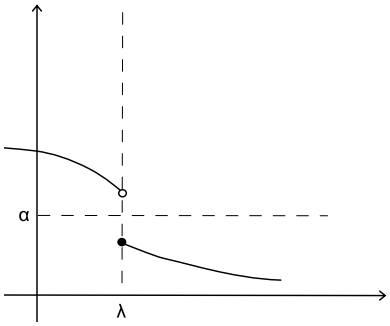


Figure 3.1: The claim (3.2) is clear if the CDF $\mathbb{P}(\mathrm{LR} \leq \lambda)$ is continuous. If it is not and α accurs at one of the jumps, we use the fact that the CDF, and so also the function $\mathbb{P}_0(\mathrm{LR} > \lambda)$ is right-continuous as on the picture.

← Exercise 3.7.1

The following corollary will be useful.

Lemma 3.1.2. *In the likelihood ratio test with level $\alpha = \mathbb{E}_0\varphi_\lambda(X)$ we have $\mathbb{E}_1\varphi_\lambda(X) > \alpha$ with equality if and only if $p_0 \equiv p_1$.*

Proof. A silly test δ, where we decide based on an α-coin flip regardless the data ($\varphi(x) = \alpha$ for all x) satisfies $\alpha = \mathbb{E}_0\delta(X)$ but also $\mathbb{E}_1\delta(X) = \alpha$. By the Neyman-Pearson lemma we conclude that $\mathbb{E}_1\varphi_\lambda(X) \geq \alpha$. By Lemma 3.7.1, we can conclude that the inequality is strict, unless the silly coin-flip is equivalent with the likelihood ratio test, which happens if and only if $p_0(x) = p_1(x)$ for all x. □

Example 3.1.3. *Suppose $p_\theta(x) = \theta e^{-\theta x}$ for $x > 0$. Consider the test $H_0 : \theta = 1$ versus $H_1 : \theta = \theta_1$ for a fixed $\theta_1 > 1$. A likelihood ratio test is given by the rejection region*

$$\mathrm{LR}(x) = \frac{p_1(x)}{p_0(x)} = \theta_1 e^{(1-\theta_1)X} > \lambda$$

or equivalently if

$$x < \frac{\log(\theta_1/\lambda)}{\theta_1 - 1} = \lambda'.$$

The level is

$$\alpha = \mathbb{P}_0(X < \lambda') = \int_0^{\lambda'} e^{-x}\mathrm{d}x = 1 - e^{-\lambda'}.$$

Solving $\lambda' = -\log(1-\alpha)$ gives a test with level α. By Proposition 3.1.1 the test constructed in this way maximizes $\mathbb{E}_1\varphi$ over all tests with level at most α. Note that the test does not depend on θ_1! Hence, this test is the UMP test for $H_0 : \theta = 1$ versus $H_1 : \theta_1 > 1$. In the next section we provide some general theory explaining this phenomenon.

Often, instead of $\mathrm{LR}(x)$ we work with the log-likelihood ratio $\lambda(x) = \log \mathrm{LR}(x)$ (for essentially the same reason as we prefer log-likelihoods). If p_0, p_1 both lie in the same exponential family with parameters $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$ then

$$\lambda(\mathbf{x}) = \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0, \boldsymbol{t}(\mathbf{x}) \rangle - (A(\boldsymbol{\theta}_1) - A(\boldsymbol{\theta}_0))$$

and so thresholding $\mathrm{LR}(\mathbf{x})$ is equivalent to thresholding a linear function of $\mathbf{t}(\mathbf{x})$. The following simple example plays an important role later.

Example 3.1.4 (Linear discriminant analysis)**.** *Consider two classes that are distributed as multivariate Gaussians, say $N(\mu_0, \Sigma)$ and $N(\mu_1, \Sigma)$, respectively, differing only in their mean vectors. For each observation we want to decide from which of the two classes it comes. In this case, the log-likelihood ratio reduces to the linear statistic*

$$\lambda(\mathbf{x}) = \left\langle \mu_1 - \mu_0, \Sigma^{-1}\left(x - \frac{\mu_0 + \mu_1}{2}\right)\right\rangle. \tag{3.3}$$

(Note that $\lambda(\mu_1) > 0$ and $\lambda(\mu_0) < 0$)

By the Neyman-Pearson lemma, the optimal decision rule is based on thresholding this statistic. Concretely, if the two classes are equally likely, then the corresponding Bayes risk of our procedure is given by

$$\mathrm{Err}(\lambda) := \frac{1}{2}\mathbb{P}_0(\lambda(X) \geq \lambda) + \frac{1}{2}\mathbb{P}_1(\lambda(X) \leq \lambda) \tag{3.4}$$

and we use it to evaluate the quality of this decision rule.

← Exercise 3.7.2

Given our Gaussian assumptions, some algebra shows that the error probability can be written in terms of the Gaussian cumulative distribution function Φ as

$$\mathrm{Err}(0) = \Phi(-\gamma/2) \qquad \textit{where } \gamma = \|\mu_1 - \mu_0\|_\Sigma, \tag{3.5}$$

where $\|x\|_\Sigma = \sqrt{x^\intercal \Sigma^{-1} x}$.

3.1.1 *Derivation from the first principles**

In our calculations above we used a guess that the optimal test should be based on the likelihood ratio and then it was straightforward to show that such a test must be optimal. In this subsection, we argue how we could come up with this guess. Note that our optimization problem is equivalent to maximizing

$$f(\varphi) = \begin{cases} \mathbb{E}_1\varphi & \text{if } \mathbb{E}_0\varphi \leq \alpha, \\ -\infty & \text{otherwise.} \end{cases}$$

Denote by φ^* an optimizer of this function and note that

$$\inf_{\lambda \geq 0} \Big\{ \mathbb{E}_1\varphi - \lambda(\mathbb{E}_0\varphi - \alpha) \Big\} = f(\varphi). \tag{3.6}$$

The function $L(\varphi, \lambda) = \mathbb{E}_1\varphi - \lambda(\mathbb{E}_0\varphi - \alpha)$ is called the Lagrangian and (3.6) shows that

$$\sup_\varphi f(\varphi) = \sup_\varphi \inf_{\lambda \geq 0} L(\varphi, \lambda), \tag{3.7}$$

where the supremum over φ is unrestricted and it runs over all critical functions. It is clear that

$$\sup_\varphi \inf_{\lambda \geq 0} L(\varphi, \lambda) \leq \inf_{\lambda \geq 0} \sup_\varphi L(\varphi, \lambda). \tag{3.8}$$

The next result states that the inequality in (3.8) is actually an equality. The proof is similar to the proof of Theorem 2.3.9.

Proposition 3.1.5. *We have*

$$\sup_\varphi \inf_{\lambda \geq 0} L(\varphi, \lambda) = \inf_{\lambda \geq 0} \sup_\varphi L(\varphi, \lambda). \tag{3.9}$$

Proof. This is a standard strategy in proving strong duality. Let

$$\mathcal{A} = \{(u,t) : \mathbb{E}_0\varphi - \alpha \leq u, \mathbb{E}_1\varphi \geq t \text{ for some } \varphi\}$$

so that $\mathbb{E}_1\varphi^* = \sup\{t : (0,t) \in \mathcal{A}\}$. Consider also a set

$$\mathcal{B} = \{(0,s) : s > \mathbb{E}_1\varphi^*\}.$$

The sets $\mathcal{A}$, $\mathcal{B}$ are both convex and disjoint subsets of $\mathbb{R}^2$ (make sure you agree). By the separating hyperplane theorem (c.f. Theorem B.1.3) there exist real numbers $\tilde{\lambda}, \mu, \beta$ such that

$$\tilde{\lambda}u - \mu t \geq \beta \qquad \text{for all } (u,t) \in \mathcal{A}, \tag{3.10}$$

$$\tilde{\lambda}u - \mu t \leq \beta \qquad \text{for all } (u,t) \in \mathcal{B}. \tag{3.11}$$

If $(u,t) \in \mathcal{A}$ then $(u',t') \in \mathcal{A}$ for every $u' > u$ and $t' < t$. Hence, (3.10) implies that $\tilde{\lambda} \geq 0$ and $\mu \geq 0$ for otherwise this expression could not be bounded below. Since $u = 0$ in $\mathcal{B}$, (3.11) states that $-\mu t \leq \beta$ for every $t > \mathbb{E}_1\varphi^*$. But then also $-\mu\mathbb{E}_1\varphi^* \leq \beta$. It follows that for every φ

$$\tilde{\lambda}(\mathbb{E}_0\varphi - \alpha) - \mu\mathbb{E}_1\varphi \;\geq\; \beta \;\geq\; -\mu\mathbb{E}_1\varphi^*.$$

or equivalently, denoting $\lambda^* = \frac{\tilde{\lambda}}{\mu}$,

$$\lambda^*(\mathbb{E}_0\varphi - \alpha) - \mathbb{E}_1\varphi \;\geq\; \tfrac{\beta}{\mu} \;\geq\; -\mathbb{E}_1\varphi^*.$$

This gives that

$$\sup_{\varphi}\inf_{\lambda\geq 0} L(\varphi,\lambda) = \mathbb{E}_1\varphi^* \geq \sup_{\varphi} L(\varphi,\lambda^*) \geq \inf_{\lambda\geq 0}\sup_{\varphi} L(\varphi,\lambda).$$

This proves the reverse of (3.8) and thus the equality in (3.9). □

By equation (3.7) and Proposition 3.1.5, in order to maximize f we can first optimize $L(\varphi,\lambda)$ over φ and then over λ. For any fixed $\lambda \geq 0$ we have

$$L(\varphi,\lambda) = \mathbb{E}_1\varphi - \lambda(\mathbb{E}_0\varphi - \alpha) = \int_{\mathcal{X}} \varphi(x)(p_1(x) - \lambda p_0(x))\mathrm{d}x + \lambda\alpha.$$

As we also argued in the previous section, irrespective of α, it is clear that φ that optimizes $L(\varphi,\lambda)$ satisfies

$$\varphi(x) = \begin{cases} 1 & \text{if } p_1(x) - \lambda p_0(x) > 0 \\ 0 & \text{if } p_1(x) - \lambda p_0(x) < 0 \end{cases}. \tag{3.12}$$

If $p_1(x) - \lambda p_0(x) = 0$, $\varphi(x)$ can take any value. So the optimum φ is given by a likelihood ratio test.

3.2 *Some constructions of UMP tests**

If the hypotheses are not both simple the situation is in general much more complicated and a uniformly most powerful test (as defined in Section 2.6.3) may be hard or impossible to obtain. In this section we discuss some of the special situations when a UMP can be obtained. We also provide alternative approaches to find a test with good properties.

3.2.1 *Monotone likelihood ratios*

A one-dimensional family of densities $p_\theta(x)$, $\theta \in \Theta \subseteq \mathbb{R}$ has **monotone likelihood ratios (MLR) in** $T(x)$ if, whenever $\theta < \theta'$, the likelihood ratio $p_{\theta'}(x)/p_\theta(x) = h(T(x))$ for a nondecreasing function h. A canonical example of such a situation is a one-dimensional exponential family with

$$p_\theta(x) = h(x)\exp\{\eta(\theta)T(x) - A(\theta)\},$$

with the canonical parameter η being a strictly increasing function of the parameter of interest θ. In this case, if $\theta' > \theta$, then

$$\frac{p_{\theta'}(x)}{p_\theta(x)} \;=\; \exp\{(\eta(\theta') - \eta(\theta))T(x) + A(\theta) - A(\theta')\},$$

which is increasing in $T(x)$.

In this section we will be interested in testing

$$H_0 : \theta \le \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

Consider the test

$$\varphi_t(x) = \begin{cases} 1 & \text{if } T(x) > t, \\ \gamma & \text{if } T(x) = t, \\ 0 & \text{if } T(x) < t \end{cases} \tag{3.13}$$

Note that for MLR families the corresponding test is equivalent to the likelihood ratio test. In particular, for any $\theta_1 > \theta_0$, this statistic gives a most powerful test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ at level $\alpha = \mathbb{E}_{\theta_0}\varphi_t(X)$. The following result allows us to get a more general statement.

Proposition 3.2.1. *If $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ has monotone likelihood ratios and $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$ for $\theta \neq \theta'$. Then, for every $t > 0$, the power function $\beta(\theta) = \mathbb{E}_\theta\varphi_t(X)$ is strictly increasing in θ.*

Proof. Note that for any $\theta_1 < \theta_2$, $\varphi_t(X)$ is equivalent to the likelihood ratio test and so it is the most powerful test for testing $H_0 : \theta = \theta_1$ versus $H_1 : \theta = \theta_2$ at level $\mathbb{E}_{\theta_1}\varphi_t(X)$. By Lemma 3.1.2, we conclude that $\mathbb{E}_{\theta_1}\varphi_t(X) \le \mathbb{E}_{\theta_2}\varphi_t(X)$ with equality if and only if $p_{\theta_1}(x) = p_{\theta_2}(x)$. □

As a corollary from this result we get that

$$\sup_{\theta \le \theta_0} \mathbb{E}_\theta \varphi_t(X) \;=\; \mathbb{E}_{\theta_0} \varphi_t(X).$$

In particular, the same test statistic can be used to get a most powerful test of size $\alpha = \mathbb{E}_{\theta_0}\varphi_t(X)$ for testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta = \theta_1$ for any $\theta_1 > \theta_0$. This reasoning gives us the following result (read carefully Definition 2.6.26 again).

Theorem 3.2.2. *Suppose the family of densities has monotone likelihood ratios. Then the test φ_t in (3.13) is uniformly most powerful for $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ and has level $\alpha = \mathbb{E}_{\theta_0}\varphi_t$. Any $\alpha \in (0,1)$ is possible.*

We discuss an example that is not an exponential family.

Example 3.2.3. *Suppose the data $X_1, \dots, X_n$ are i.i.d. from the uniform distribution on $[0,\theta]$. The joint density $p_\theta(x)$ is positive if and only if $x_i \in [0,\theta]$ for $i = 1,\dots,n$ and this happens if and only if $M(x) = \min\{x_1,\dots,x_n\} \ge 0$ and $T(x) = \max\{x_1,\dots,x_n\} \le \theta$. Thus*

$$p_\theta(x) = \begin{cases} 1/\theta^n & \text{if } M(x) \ge 0, T(x) \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $\theta_2 > \theta_1$, $M(x) \ge 0$, and $T(x) \le \theta_2$. Then

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = \begin{cases} (\theta_1/\theta_2)^n & \text{if } T(x) \le \theta_1 \\ +\infty & \text{otherwise.} \end{cases}$$

This shows that the family of joint densities has monotone likelihood ratios. If we are interested in testing $\theta \le 1$ versus $H_1 : \theta > 1$, the test function $\varphi_t(x)$ gives the UMP. This test has level

$$\mathbb{E}_1 \varphi_t(X) = \mathbb{P}_1(T \ge t) = 1 - t^n$$

and a specified level α can be achieved taking $t = (1-\alpha)^{1/n}$. The power of this test is

$$\beta(\theta) = \mathbb{P}_\theta(T \ge t) = \begin{cases} 0 & \text{if } \theta < t \\ 1 - \frac{1-\alpha}{\theta^n} & \text{if } \theta \ge t. \end{cases}$$

3.2.2 *Affine submodels**

Consider an exponential family as in (1.5). Consider a p-dimensional linear space $\mathcal{L} \subseteq \mathbb{R}^d$ and the corresponding $\Theta_0 = \Theta \cap \mathcal{L}$. There exists a simple affine change of coordinates form $\boldsymbol{\theta}$ to $(\boldsymbol{\lambda}, \boldsymbol{\psi})$ such that $\Theta_0 = \{(\boldsymbol{\lambda}, \boldsymbol{\psi}) : \boldsymbol{\psi} = \mathbf{0}\}$. If the basis of $\mathcal{L}$ is given by the columns of $\mathbb{R}^{d\times p}$ then $\boldsymbol{\theta} = A\boldsymbol{\lambda}$ parametrizes Θ_0. Thus, for $\boldsymbol{\theta} \in \Theta_0$ we have

$$\langle \boldsymbol{\theta}, \boldsymbol{t} \rangle = \langle A\boldsymbol{\lambda}, \boldsymbol{t} \rangle = \langle \boldsymbol{\lambda}, A^\intercal \boldsymbol{t} \rangle = \langle \boldsymbol{\lambda}, \boldsymbol{u} \rangle,$$

where $\boldsymbol{u} = A^\intercal \boldsymbol{t}$ is the sufficient statistics of the p-dimensional model parametrized by Θ_0.

Thus, without loss of generality we assume $\mathbf{t} = (\boldsymbol{u}, \boldsymbol{v})$ with $\boldsymbol{\psi} = \boldsymbol{\theta}_v$ and $\boldsymbol{\lambda} = \boldsymbol{\mu}_u$. We want to test the q-dimensional hypothesis ($q = d - p$) $\boldsymbol{\psi}_0 = \mathbf{0}$:

$$H_0 : \quad \boldsymbol{\psi} = \mathbf{0}, \qquad \text{versus} \quad \boldsymbol{\psi} \neq \mathbf{0}.$$

A simple example is that we have a regression model and want to delete a certain set of its regressors to reduce the dimension of the model. Alternatively, the primary model might be the smaller p-dimensional model with canonical statistics $\boldsymbol{u}$. We want to test if this model fits data, by embedding it in a wider model. As an example, consider the problem of testing a Gaussian graphical model with respect to graph G_0 (see Example 1.9.5) versus a Gaussian graphical model with respect to graph $G \supset G_0$.

We first consider a model reduction of the canonical statistics from $\boldsymbol{t} = (\boldsymbol{u}, \boldsymbol{v})$ to $\boldsymbol{u}$. In Section 1.5 we argued that any inference on $\boldsymbol{\psi}$ should be done conditionally on $\boldsymbol{u}$. The argument was that $\boldsymbol{u}$ provides no information about $\boldsymbol{\psi}$ as expressed in the likelihood factorisation (1.14) valid whether H_0 is true or not. All information provided by $\boldsymbol{u}$ is consumed in estimating $\boldsymbol{\mu}_u$ and this parameter has no information about $\boldsymbol{\psi}$ by variational inference.

Recall the form of the distribution of $\mathbf{t} = (\boldsymbol{u}, \boldsymbol{v})$ in (1.6) and consider the conditional distribution for $\boldsymbol{v}$ given $\boldsymbol{u}$. Inserting $\boldsymbol{\psi} = \mathbf{0}$ simplifies this to

$$f_0(\boldsymbol{v}|\boldsymbol{u}) = \frac{g(\boldsymbol{u}, \boldsymbol{v})}{g_0(\boldsymbol{u})}, \tag{3.14}$$

where index 0 indicates distribution under H_0, and in particular $g_0(\boldsymbol{u}) = \int g(\boldsymbol{u}, \boldsymbol{v}) \mathrm{d}\boldsymbol{v}$ is the structure function in the marginal exponential family for $\boldsymbol{u}$ under H_0. Note that, under the null, the conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{u}$ is parameter free.

We propose the following test for $H_0 : \boldsymbol{\psi} = \mathbf{0}$ versus $H_1 : \boldsymbol{\psi} \neq \mathbf{0}$:

(i) Use $f_0(\boldsymbol{v}|\overline{\boldsymbol{u}})$ as the test statistic, which, under the null, is equal to the conditional density of $\boldsymbol{v}$ given $\overline{\boldsymbol{u}}$.

(ii) Reject H_0 if $f_0(\overline{\boldsymbol{v}}|\overline{\boldsymbol{u}})$ is too small, and calculate the p-value as

$$\mathbb{P}\Big(f_0(\boldsymbol{v}|\overline{\boldsymbol{u}}) \leq f_0(\overline{\boldsymbol{v}}|\overline{\boldsymbol{u}})\Big) = \int_{f_0(\boldsymbol{v}|\overline{\boldsymbol{u}}) \leq f_0(\overline{\boldsymbol{v}}|\overline{\boldsymbol{u}})} f_0(\boldsymbol{v}|\overline{\boldsymbol{u}}) \mathrm{d}\boldsymbol{v},$$

where $\overline{\boldsymbol{u}}$, $\overline{\boldsymbol{v}}$ denote the observed quantities.

These test follow the Fisher's principle of **exact tests**. "Exactness" comes from the fact that the test achieves exactly the desired size α (as opposed to approximate tests). One obvious question is about the power of such a procedure.

Consider a special case of the above setting when $\mathbf{t} = (\boldsymbol{u}, v)$ with v one-dimensional. In this case θ_v is one-dimensional too and we can obtain a uniform most powerful test for $H_0 : \psi = 0$ versus $H_1 : \psi \neq 0$. Recall that, under the null, the conditional distribution of v given $\boldsymbol{u}$ is parameter free. We will then use directly v as the test statistic with $\boldsymbol{u}$ fixed and consider the test $\varphi^*(x)$ defined as

$$\varphi^*(x) = \begin{cases} 1 & \text{if } v > c^*(\boldsymbol{u}), \\ 1 & \text{if } v < c_*(\boldsymbol{u}), \\ \gamma^*(\boldsymbol{u}) & \text{if } v = c^*(\boldsymbol{u}), \\ \gamma_*(\boldsymbol{u}) & \text{if } v = c_*(\boldsymbol{u}), \\ 0 & \text{if } v \in (c_*(\boldsymbol{u}), c^*(\boldsymbol{u})), \end{cases}$$

with $c(\cdot)$ and $\gamma(\cdot)$ adjusted so that the test has exactly level α.

Theorem 3.2.4. *If the exponential family is regular and θ_v is one dimensional, then φ^* is a uniformly most powerful unbiased test of $H_0 : \theta_v = 0$ versus $H_1 : \theta_v \neq 0$.*

Proof. See Theorem 13.6 in [1]. □

[1] Robert W. Keener. *Theoretical statistics*. Springer Texts in Statistics. Springer, New York, 2010. Topics for a core course

Example 3.2.5 (Mean value test for a normal distribution)**.** *Consider testing $H_0 : \mu = 0$ for a sample of size n from $N(\mu, \sigma^2)$. In Example 1.1.4 we note that the canonical parameters of this family are $(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ so equivalently we test $\theta_1 = 0$ with $v = \sum_i x_i$. The exact test should be conditional, given $u = \sum_i x_i^2 = \|x\|^2$. Thus we are free to eliminate the scale parameter by considering $v/\sqrt{u}$ instead of v. Indeed, use Exercise 3.7.3 to conclude that, under the null, $u \perp\!\!\!\perp v/\sqrt{u}$ and the vector $\frac{1}{\sqrt{u}}(x_1, \ldots, x_n)$ is uniformly distributed on the $(n-1)$-dimensional unit sphere and so the distribution of $v/\sqrt{u}$ is also parameter-free. Exercise 3.7.3 also allows to conclude that $v/\sqrt{u}$ is independent of u. Thus we can forget about conditioning on u, that is, we may restrict attention to the marginal distribution of $v/\sqrt{u}$.*

← Exercise 3.7.3

The last step is to show that this test is equivalent to the t-test. Let $\tau = \sqrt{n}\bar{x}/s$ be the t-test statistic. Let us rewrite $v/\sqrt{u}$ as

$$\frac{n\bar{x}}{\sqrt{\sum_i x_i^2}} = \frac{n\bar{x}}{\sqrt{(n-1)s^2 + n\bar{x}^2}} = \frac{n\bar{x}/s}{\sqrt{(n-1) + (\sqrt{n}\bar{x}/s)^2}} = \frac{\sqrt{n}\tau}{\sqrt{(n-1)+\tau^2}}.$$

The right-hand side is seen to be an odd function and monotone function of τ. Thus, a single tail or a symmetric pair of tails in $u/\sqrt{v}$ is equivalent to a single or symmetric pair of tails in the usual t-test.

← Exercise 3.7.4

3.3 *Sequential testing**

The aim of this section is to give a brief introduction into sequential testing. For simplicity we focus on the **sequential probability ratio**

test. This test was suggested by Wald for simple versus simple testing with *i.i.d* observations with optional stopping. Let $X_1, X_2, \ldots$ be *iid* from a distribution with density p_k, $k = 0, 1$, and consider testing $H_0 : k = 0$ versus $H_1 : k = 1$. Define

$$\mathrm{LR}_n = \mathrm{LR}_n(X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{p_1(X_i)}{p_0(X_i)}, \tag{3.15}$$

the likelihood ratio for the first n observations. By convention, $\mathrm{LR}_0 = 1$. From the Neyman-Pearson theorem we know that for a fixed sample size the best test rejects H_0 according to the size of LR_n. In the **Sequential Probability Ratio Test** (SPRT) at each step the researcher has three options: stop and accept H_0, stop and accept H_1, or continue sampling. For the SPRT these options are resolved by comparing the likelihood ratio with two critical values $\gamma_0 < 1 < \gamma_1$ in the following manner:

1. If $\mathrm{LR}_n \in (\gamma_0, \gamma_1)$, take another observation.
2. If $\mathrm{LR}_n \geq \gamma_1$, reject H_0.
3. If $\mathrm{LR}_n \leq \gamma_0$, accept H_0.

Formally, the sample size for this SPRT is then a random variable defined as

$$N := \inf\{n : \mathrm{LR}_n \notin (\gamma_0, \gamma_1)\}.$$

We will set the thresholds to provide desired power

$$\beta = \mathbb{P}_1(\mathrm{LR}_N \geq \gamma_1)$$

and size

$$\alpha = \mathbb{P}_0(\mathrm{LR}_N \geq \gamma_1).$$

Note that both quantities involve N, which makes the analysis more subtle. To simplify the notation, for a fixed $n \in \mathbb{N}$, let $\mathbf{x} := (x_1, \ldots, x_n)$ and write $p_k(\mathbf{x}) := \prod_{i=1}^{n} p_k(x_i)$, $k = 0, 1$. Let $R_1 = \{x : \mathrm{LR}_n \geq \gamma_1\}$ be the rejection region. Note that

$$\mathbb{P}_1(\mathrm{LR}_n \geq \gamma_1) = \int_{R_1} p_1(\mathbf{x}) d\mathbf{x} = \int_{R_1} \mathrm{LR}_n p_0(\mathbf{x}) d\mathbf{x} \geq \gamma_1 \mathbb{P}_0(\mathrm{LR}_n \geq \gamma_1). \tag{3.16}$$

Similarly, denoting $R_0 = \{\mathrm{LR}_n \leq \gamma_0\}$,

$$\mathbb{P}_0(\mathrm{LR}_n \leq \gamma_0) = \int_{R_0} p_0(\mathbf{x}) d\mathbf{x} = \int_{R_0} \mathrm{LR}_n^{-1} p_1(\mathbf{x}) d\mathbf{x} \geq \gamma_0^{-1} \mathbb{P}_1(\mathrm{LR}_n \leq \gamma_0). \tag{3.17}$$

With a bit more careful treatment (using Wald's likelihood ratio identity, Theorem 2.3.3 and Section 3.1.1.1 in [2]) we can replace n with the

[2] Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014

stopping time N both in (3.16) and (3.17). It follows that $\beta \geq \gamma_1 \alpha$ and $(1-\alpha) \geq \gamma_0^{-1}(1-\beta)$ and so

$$\gamma_1 \leq \frac{\beta}{\alpha} \qquad \text{and} \qquad \gamma_0 \geq \frac{1-\beta}{1-\alpha}. \tag{3.18}$$

The relation between (α, β) and (γ_0, γ_1) is complicated. The above inequalities lead to an approximate analysis. Suppose (α^*, β^*) are some desired size and the power. Using the proposal of Wald, we set $\gamma_1^* = \frac{\beta^*}{\alpha^*}$ and $\gamma_0^* = \frac{1-\beta^*}{1-\alpha^*}$. Note that α^*, β^* are typically not equal to the real size and power. The bounds in (3.18) guarantee however that is α^* is small and β^* is large then $\alpha \approx \alpha^*$ and $\beta \approx \beta^*$; see Figure 3.2.

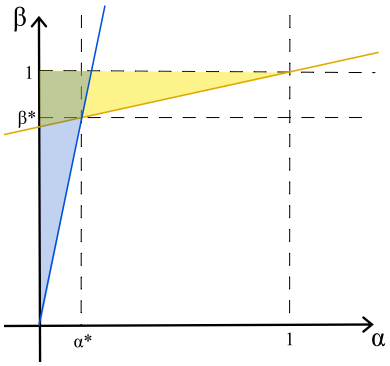


Figure 3.2: Bounds in (3.18) with $\gamma_0 = \frac{1-\beta^*}{1-\alpha^*}$ and $\gamma_1 = \frac{\beta^*}{\alpha^*}$.

Expected stopping time of SPRT. To gain insight into this issue, let us consider the expected stopping time. We can calculate the expected value of N as follows. First observe that, for any fixed time n,

$$\mathbb{E}_k(\log \mathrm{LR}_n) = \sum_{i=1}^{n} \mathbb{E}_k \left(\log \frac{p_1(X_i)}{p_0(X_i)} \right) = \begin{cases} n\mathrm{D}(p_1 \| p_0) & \text{if } k = 1 \\ -n\mathrm{D}(p_0 \| p_1) & \text{if } k = 0, \end{cases}$$

where $\mathrm{D}(p_0 \| p_1)$ is the KL-divergence between p_0 and p_1.

Proposition 3.3.1 (Wald's identity). *Let $Y_1, Y_2, \ldots$ be independent and identically distributed random variables with mean μ and suppose $\mathbb{E}|Y_i| < C$ for some C. Let N be any integer-valued random variable such that $E[N] < \infty$ and $\{N = n\} \in \sigma(Y_1, \ldots, Y_n)$. Then $\mathbb{E}[\sum_{i=1}^{N} Y_i] = \mu \mathbb{E}[N]$.*

Proof. Start by noting that the event $\{N \geq i\} = (\bigcup_{j=1}^{i-1} \{N = j\})^c$. Thus, the event is independent of $Y_i, Y_{i+1}, \ldots$ (since it is determined by $Y_1, \ldots, Y_{i-1}$). From this we see that

$$\sum_{i=1}^{\infty} \mathbb{E}(|Y_i| \mathbb{1}(N \geq i)) = \sum_{i=1}^{\infty} \mathbb{E}(|Y_i|) \mathbb{P}(N \geq i) \leq C\, \mathbb{E}\mathbb{N} < \infty. \tag{3.19}$$

Write

$$\mathbb{E}(\sum_{i=1}^{N} Y_i) = \mathbb{E}(\sum_{i=1}^{\infty} \mathbb{1}(N \geq i) Y_i) = \sum_{i=1}^{\infty} \mathbb{E}(\mathbb{1}(N \geq i) Y_i).$$

(the interchange of expectation and summation is justified by the dominated convergence theorem and (3.19)). Therefore,

$$\sum_{i=1}^{\infty} \mathbb{E}[\mathbb{1}(N \geq i) Y_i] = \mathbb{E}(Y_1) \sum_{i=1}^{\infty} \mathbb{E}(\mathbb{1}(N \geq i)) = \mu \sum_{i=1}^{\infty} \mathbb{P}(N \geq i) = \mu \mathbb{E}(N).$$

□

So, by Wald's Identity we have

$$\mathbb{E}_k(\log \mathrm{LR}_N) = \begin{cases} \mathbb{E}_1(N) \mathrm{D}(p_1 \| p_0) & \text{if } k = 1 \\ -\mathbb{E}_0(N) \mathrm{D}(p_0 \| p_1) & \text{if } k = 0. \end{cases} \tag{3.20}$$

Now to obtain an expression for $\mathbb{E}_k(N)$ we will derive another formula for $\mathbb{E}_k[\log \mathrm{LR}_N]$. Let us assume the value of the likelihood ratio is approximately equal to a threshold level when the SPRT terminates. The value of the likelihood ratio will typically be just slightly greater/lower than the upper/lower threshold level. Using this approximation and smoothing (note that $\mathbb{P}_k(\mathrm{LR}_N \in (\gamma_0, \gamma_1)) = 0$) we can write

$$\begin{aligned} \mathbb{E}_0(\log \mathrm{LR}_N) &\approx \alpha \log(\gamma_1) + (1-\alpha)\log(\gamma_0) \\ &\approx \alpha \log \frac{\beta}{\alpha} + (1-\alpha)\log\frac{1-\beta}{1-\alpha}, \end{aligned}$$

where we used the fact that $\gamma_1 = \frac{\beta}{\alpha}$ and $\gamma_0 = \frac{1-\beta}{1-\alpha}$. Similarly,

$$\mathbb{E}_1(\log \mathrm{LR}_N) \approx \beta \log \frac{\beta}{\alpha} + (1-\beta)\log\frac{1-\beta}{1-\alpha}. \tag{3.21}$$

Denoting by $\pi_\alpha \sim \mathrm{Bern}(\alpha)$ and $\pi_\beta \sim \mathrm{Bern}(\beta)$, we get $\mathbb{E}_0(\log \mathrm{LR}_N) \approx -\mathrm{D}(\pi_\alpha \| \pi_\beta)$ and $\mathbb{E}_1(\log \mathrm{LR}_N) \approx \mathrm{D}(\pi_\beta \| \pi_\alpha)$. With these approximations, using (3.20), we obtain expressions for $\mathbb{E}_k(N)$:

$$\mathbb{E}_0(N) \approx \frac{\mathrm{D}(\pi_\alpha \| \pi_\beta)}{\mathrm{D}(p_0 \| p_1)}$$

$$\mathbb{E}_1(N) \approx \frac{\mathrm{D}(\pi_\beta \| \pi_\alpha)}{\mathrm{D}(p_1 \| p_0)}.$$

Note that the expected stopping times increase as the KL divergences decreases (as the two densities become less distinguishable). Increasing β or decreasing α also increases the expected stopping time.

Optimality of SPRT The expected stopping time of the SPRT that we determined above is optimal. No other test can achieve the same β and α with a smaller expected number of samples, under either hypothesis, as the following result shows.

Lemma 3.3.2 (Lower bound on expected stopping time of any testing procedure (Wald&Wolfowitz 1948))**.** *Let α and β be given and consider any sequential test with size $\leq \alpha$ and power $\geq \beta$. Then the expected stopping times N' for the test satisfy $\mathbb{E}_k N' \geq \mathbb{E}_k N$ for $k = 0, 1$, where N is the stopping time of the corresponding SPRT with size α and power β.*

The lemma shows that if no other test can have error levels as small or smaller than the SPRT and have expected stopping times less than the values computed above for the SPRT.

Example 3.3.3 (Sequential testing in Gaussian case)**.** *Let $X_1, X_2, \ldots$ be an i.i.d. sequence of normal variables $N(\mu, 1)$. Consider the simple binary testing problem: $H_0 : \mu = 0$, $H_1 : \mu = \mu_0 > 0$. For simplicity, let us specify equal probabilities of error, that is, $\alpha = 1 - \beta \ll \frac{1}{2}$. In this case the*

optimal cut-off for the (non-sequential) likelihood ratio test is to reject the null if $\log \mathrm{LR}_n \geq 0$. *It also easily follows that*

$$\alpha = \Phi\left(\frac{\sqrt{n}\mu_0}{2}\right),$$

where Φ *is the CDF of the standard normal variable. So the number of samples required for a specified* α *is*

$$n = \frac{2(\Phi^{-1}(\alpha))^2}{\mu_0}.$$

Since $\mathrm{D}(p_0\|p_1) = \mu^2/2$ *and* $\mathrm{D}(\pi_\alpha\|\pi_\beta) = \mathrm{D}(\pi_\beta\|\pi_\alpha) = (1 - 2\alpha)\log\frac{1-\alpha}{\alpha}$, *the expected stopping time of the SPRT in this case is approximately*

$$\mathbb{E}_0(N) = \mathbb{E}_1(N) = \frac{2(1-2\alpha)}{\mu_0^2}\log\frac{1-\alpha}{\alpha}.$$

Compare the two quantities to see that the sequential sample requirement is indeed preferred.

3.4 *Motivating multiple testing*

Selective inference means searching for interesting patterns in data, with statistical guarantees that account for the search process. It encompasses multiple testing, post-selection inference, and adaptive or interactive inference. There are two main situations for considering simultaneous inference. One situation is when we test multiple hypothesis. Another is when we try to obtains simultaneous coverage for confidence intervals for a multiple parameters. In this section we will focus on multiple testing.

As a general motivation, consider the general problem of simultaneously testing a finite number of null hypotheses H_0^i for $i = 1, \ldots, m$.

Example 3.4.1. *Suppose that we have m genes and data about expression levels for each gene among healthy individuals and those with lung cancer.*

	Healthy (k patients)	*Lung cancer (l patients)*
Expression Level of Gene i	$x_{ij}^{(0)}$, $1 \leq j \leq k$	$x_{ij}^{(1)}$, $1 \leq j \leq l$

The i-th null hypothesis, denoted H_0^i, would state that the mean expression level of the i-th gene is the same in both groups of patients.

We assume that the tests for the individual hypotheses are available (T_i test statistic, R_i rejection region, P_i the associated p-value, i.e. the smallest α leading to rejection) and the problem is how to combine them into a simultaneous test procedure. The easiest yet extremely naive approach (as illustrated in the webcomic xkcd[3]) is to disregard the multiplicity and simply test each hypothesis at level α. However, with such a procedure the probability of one or more false rejections rapidly increases with n. For example, if all test are independent of size α and all null hypotheses are true then

[3] See https://xkcd.com/882/

$$\mathbb{P}_0(\bigcap\{T_i \notin R_i\}) \;=\; \prod_{i=1}^{m} \mathbb{P}_0(T_i \notin R_i) = (1-\alpha)^m.$$

In this sense the claim that the procedure controls the probability of false rejections at level α is clearly misleading. A similar situation emerges when constructing a confidence region for a parameter vector using individual confidence intervals for each component.

3.5 *Family-wise error rate*

Let $\mathcal{H}_0 \subseteq \{1, \ldots, m\}$ be the index set of the true hypotheses and let $\mathcal{R} \subseteq \{1, \ldots, m\}$ be the set of rejected hypotheses. Denote $m_0 = |\mathcal{H}_0|$. The **family-wise error rate (FWER)** is

$$\text{FWER} \;=\; \mathbb{P}(|\mathcal{H}_0 \cap \mathcal{R}| \geq 1). \tag{3.22}$$

A natural approach it to replace the usual condition for testing a single hypothesis, that the probability of false rejection not exceed α, by the requirement

$$\text{FWER} \ \leq \ \alpha$$

for all possible combinations of true and false hypotheses. Methods that control the FWER are often described by the p-values of individual tests.

3.5.1 *Example: Gaussian sequence model*

For simplicity of the discussion, for most of this section, we restrict our discussion to the important example given by the **Gaussian sequence model**. Consider a model $Y_i = \mu_i + \varepsilon_i$ for $i = 1, \ldots, m$, where $\epsilon_i \sim N(0,1)$. For now we will not assume that ε_i are independent[4]. Also the case, when variance of the noise is a general $\sigma^2 > 0$ but known can be easily covered. The typical question that is asked about this model is how we can test, which elements of $\mu = (\mu_1, \ldots, \mu_m)$ are zero, or perhaps, which elements of μ are equal to each other. There are many other questions that can be phrased as linear equalities or inequalities in the vector μ.

[4] With independence, this model already appeared in our discussion of the Stein's paradox.

In the question of testing which μ_i are zero, we already mentioned the naive approach: take $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. Test $H_0^i : \mu_i = 0$ using the test statistic $T_i = |Y_i|$ and rejection rule $\mathbb{1}\{|Y_i| > z_{\alpha/2}\}$. One classical fix is given by the Bonferroni correction, which is to use α/m instead of α. Now, we test $H_0^i : \mu_i = 0$ with the test $\mathbb{1}\{|Y_i| > z_{\alpha/2m}\}$. Assuming all the nulls are true, we obtain

$$\text{FWER} \ = \ \mathbb{P}_0(\exists i \ \ |\varepsilon_i| > z_{\alpha/2m}) \ \leq \ \sum_{i=1}^{m} \mathbb{P}_0(|\varepsilon_i| > z_{\alpha/2m}) \ = \ m\frac{\alpha}{m} \ = \ \alpha,$$

where in the inequality we used the union bound. It is clear that the union bound can be conservative, which implies that the Bonferroni bound can be conservative too. In the special case when all ε_i are independent, the events $\{|\varepsilon_i| > z_{\alpha/2m}\}$ are independent and

$$\begin{aligned} \text{FWER} \ &= \ \mathbb{P}_0(\exists i \ \ |\varepsilon_i| > z_{\alpha/2m}) \ = \ 1 - \prod_{i=1}^{m} \mathbb{P}_0(|\varepsilon_i| \leq z_{\alpha/2m}) \\ &= \ 1 - (1 - \frac{\alpha}{m})^m \ \approx \ 1 - e^{-\alpha} \ \approx \ \alpha. \end{aligned}$$

Therefore, in this case, the Bonferroni procedure provides a good control over the family-wise error rate. This also tells us that if we have many hypotheses (e.g. $m = 10{,}000$ genes in the biological example) then Bonferroni's test has size approximately $1 - e^{-\alpha}$, which for small α is approximately α. For example, if $\alpha = 0.05$, then $1 - e^{-\alpha} = 0.04877\ldots$ So to get a test of size 0.05, we could test each hypothesis at level $0.0512/m$.

Remark 3.5.1 (Šidák correction)**.** *The above calculation shows that we could use $z_{\tilde{\alpha}/2}$ with $\tilde{\alpha} = 1-(1-\alpha)^{1/m}$ in order to get the error rate precisely α under independence.*

The Bonferroni correction can become overly conservative in the case when ε_i are dependent. In the extreme situation, when they are all equal we get

$$\text{FWER} = \mathbb{P}(\exists i \ |\varepsilon_i| > z_{\alpha/2m}) = \mathbb{P}(|\varepsilon_1| > z_{\alpha/2m}) = \frac{\alpha}{m}.$$

As we said, one of the problems of selective inference can be to find guarantees for simultaneous coverage. In what follows we are going to assume full independence of the errors ε_i. Consider a problem in which for all $v \in \mathbb{R}^m$ we want to build a confidence interval CI_v to cover the parameter $v^\top \mu$ with the property that

$$\mathbb{P}\left(v^\top \mu \in \mathsf{CI}_v \text{ for all } v \in \mathbb{R}^m\right) \geq 1-\alpha.$$

Since there are infinitely many vectors v, using the Bonferroni correction is not possible. One way to solve this problem is to use the **Scheffé's method**. First, it is clear that the problem depends only on the direction of v and not on its norm. Assume then that $\|v\| = 1$. The confidence interval CI_v will be centered around $v^\top Y$. To get its length, we need to bound $|v^\top Y - v^\top \mu|$ for all v such that $\|v\| = 1$. We use

$$\chi_m(\alpha) = \text{ the } (1-\alpha)\text{-quantile of the } \chi_m \text{ distribution}$$

and define

$$\mathsf{CI}_v = (v^\top Y - \chi_m(\alpha), v^\top Y + \chi_m(\alpha)).$$

We have

$$\begin{aligned}
\mathbb{P}\left(v^\top \mu \in \mathsf{CI}_v \text{ for all } \|v\| = 1\right) &= \mathbb{P}\left(|v^\top Y - v^\top \mu| < \chi_m(\alpha) \text{ for all } \|v\| = 1\right) \\
&= \mathbb{P}\left(\|Y - \mu\| < \chi_m(\alpha)\right) = \mathbb{P}\left(\|\varepsilon\| < \chi_m(\alpha)\right) \\
&= \alpha.
\end{aligned}$$

Remark 3.5.2. *In this section, we always assumed that $\sigma^2 = 1$ or equivalently that σ^2 is known. If it is unknown, it can be still often estimated using replicates. So suppose $Y_{ij} = \mu_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \overset{i.i.d.}{\sim} N(0,\sigma^2)$ for $i = 1,\dots,m$ and $j = 1,\dots,n$. Let $\overline{Y}_i = \frac{1}{n}\sum_j Y_{ij}$, $\bar{\varepsilon}_i = \frac{1}{n}\sum_j \varepsilon_{ij}$. Then $\overline{Y}_i = \mu_i + \bar{\varepsilon}_i$ with $\bar{\varepsilon}_i \overset{i.i.d.}{\sim} N(0,\sigma^2/n)$. Using ideas in Example 1.4.5, we can then simply construct an estimate $\hat{\sigma}^2$ of σ^2, where $\hat{\sigma}^2 \perp\!\!\!\perp \bar{\varepsilon}_1,\dots,\bar{\varepsilon}_m$. Moreover,*

$$\frac{\hat{\sigma}^2}{\sigma^2}(n-1)m \sim \chi^2_{(n-1)m}.$$

Now the correct modification of Bonferonni bounds is to use the quantiles of $t_{(n-1)m}$ instead of $N(0,1)$. In the Scheffé's method we use $F_{m,(n-1)m}$ in place of χ_m.

3.5.2 *Bonferroni and Holm*

First, recall a basic fact of statistical tests. Suppose a null hypothesis H_0 is true, and we perform a statistical test of H_0 and obtain a p-value P. What is the distribution of P? Recall that the p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct. If our test statistic T has a continuous distribution under H_0 with CDF F, and the rejection regions are of the form $\{X \in \mathcal{X} : T(X) \leq t_\alpha\}$, then the p-value is just the lower tail probability $P = F(T)$. Thus, for any $u \in (0,1)$

$$\mathbb{P}_0(P \leq u) = \mathbb{P}_0(F(T) \leq u) = \mathbb{P}_0(T \leq F^{-1}(u)) = F(F^{-1}(u)) = u.$$

So $P \sim U(0,1)$ under the null. Similarly, $P \sim U(0,1)$ if we reject for large T, or both large and small T.

If T has a discrete distribution under H_0, then so does P, so the null distribution of P would not be exactly uniform. However, we still have that

$$\mathbb{P}(P \leq u) \leq u \qquad \text{for all } u \in (0,1). \tag{3.23}$$

To show this define

$$G(u) \;=\; \sup\{y : F(y) \leq u\}. \tag{3.24}$$

Note that, by definition, $F(t) \leq u$ implies that $t \leq G(u)$. Since F is non-decreasing and right-continuous we have $F(G(u)) \geq u$[5]. Suppose that u is such that $F(G(u)) = u$, then

$$\mathbb{P}_0(P \leq u) \;=\; \mathbb{P}_0(F(T) \leq u) \;\leq\; \mathbb{P}(T \leq G(u)) \;=\; F(G(u)) \;=\; u.$$

If $F(G(u)) > u$, F had a jump at $\lambda = G(u)$

$$\lim_{y \to \lambda^-} F(y) \leq u, \qquad \lim_{y \to \lambda^+} F(y) > u.$$

By Remark C.1.3 the function $\mathbb{P}(T < \lambda)$ is left-continuous in λ. Thus, we have

$$\mathbb{P}_0(P \leq u) \;=\; \mathbb{P}_0(F(T) \leq u) \;=\; \mathbb{P}(T < \lambda) \;=\; \lim_{t \to \lambda^-} \mathbb{P}(T < y) \;\leq\; \lim_{y \to \lambda^-} F(y) \;\leq\; u.$$

This gives us one simple way of dealing with multiple testing problem.

Theorem 3.5.3 (Bonferroni Procedure)**.** *If, for $i = 1, \ldots, m$, hypothesis H_0^i is rejected when $P_i \leq \alpha/m$, then the FWER for the simultaneous testing of $H_0^1, \ldots, H_0^m$ satisfies* FWER $\leq \alpha$.

[5] Indeed, if $t > G(u)$ then $F(t) > u$. By right-continuity, if converging to $G(u)$ from the right gives $F(G(u)) \geq u$.

Proof. Suppose hypotheses H_0^i with $i \in \mathcal{H}_0$ are true and the other are false. From the union bound it follows that

$$\begin{aligned} \text{FWER} &= \mathbb{P}\Big(\text{reject any } H_0^i \text{ with } i \in \mathcal{H}_0\Big) \leq \sum_{i \in \mathcal{H}_0} \mathbb{P}(\text{reject } H_0^i) \\ &= \sum_{i \in \mathcal{H}_0} \mathbb{P}(P_i \leq \frac{\alpha}{m}) \overset{(3.23)}{\leq} \sum_{i \in \mathcal{H}_0} \frac{\alpha}{m} \leq \frac{m_0}{m}\alpha \leq \alpha. \end{aligned}$$

□

From the proof, it is clear that, although this procedure controls FWER, it is generally too conservative to be useful in detecting H_0^i that are false. If m_0 is smaller than m, using α/m_0 would be preferable. The problem, of course, is that m_0 is not known. Note however that is one H_0^i is false then $m_0 \leq m-1$ and we could use $\alpha/(m-1)$ as the threshold. If this H_0^i was true and we rejected it then we already made a mistake so we can do whatever we want and there is no harm in using $\alpha/(m-1)$ for the other hypotheses (if we make a mistake, it does not matter how many).

The Holm procedure tries to make use of the above observations. It can conveniently be stated in terms of the p-values $P_1, \ldots, P_n$ of the n individual tests. The procedure starts by sorting the p-values. Call them $P_{(1)} \leq \cdots \leq P_{(m)}$. Then it goes as follows:

1. If $P_{(1)} \leq \frac{\alpha}{m}$, reject $H_0^{(1)}$ and continue. Else stop.
2. If $P_{(2)} \leq \frac{\alpha}{m-1}$, reject $H_0^{(2)}$ and continue. Else stop.

... ...

m. If $P_{(m)} \leq \alpha$, reject $H_0^{(m)}$.

In other words, the procedure finds the smallest r such that $P_{(r)} > \frac{1}{m-(r-1)}\alpha$ and it rejects the null hypotheses $H_0^{(1)}, \ldots, H_0^{(r-1)}$.

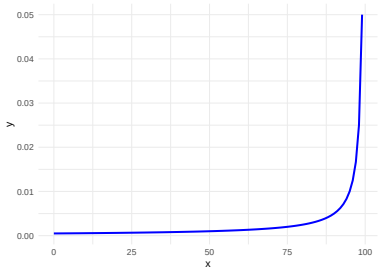


Figure 3.3: The plot of the function $\frac{\alpha}{m-x}$ with $m = 100$ and α. The ordered p-values $P_{(1)}, \ldots, P_{(r)}$ will lie below this curve.

Theorem 3.5.4. *The Holm procedure satisfies* FWER $\leq \alpha$.

Proof. Suppose $\mathcal{H}_0$ is the set of true hypotheses. Order the P-values as above $P_{(1)} \leq \cdots \leq P_{(m)}$. Ideally, if $i \in \mathcal{H}_0$ then P_i appears later in this order sequence. Let j be the smallest (random) index satisfying

$$P_{(j)} = \min_{i \in \mathcal{H}_0} P_i.$$

Note that, by construction, $j \leq m - m_0 + 1$ (there are at least $m_0 - 1$ indices following (j)). Now, the Holm procedure commits a false rejection if $r \geq j+1$, or in other words,

$$P_{(1)} \leq \frac{\alpha}{m}, \quad P_{(2)} \leq \frac{\alpha}{m-1}, \quad \ldots, \quad P_{(j)} \leq \frac{\alpha}{m-j+1}$$

which implies that

$$\min_{i\in\mathcal{H}_0} P_i \;=\; P_{(j)} \;\leq\; \frac{\alpha}{m-j+1} \;\leq\; \frac{\alpha}{m_0}.$$

(This may be confusing because you may think that we could make a mistake by rejecting some other null. Note however that, if $P_{(j)} > \frac{\alpha}{m-j+1}$ then the procedure stops and no other mistakes are done!) We thus get

$$\begin{aligned}\mathbb{P}(|\mathcal{H}_0\cap\mathcal{R}|\geq 1) &= \mathbb{P}(P_{(1)}\leq\frac{\alpha}{m},\ldots,P_{(j)}\leq\frac{\alpha}{m-j+1})\\ &\leq \mathbb{P}(\min_{i\in\mathcal{H}_0} P_i\leq\frac{\alpha}{m-j+1}) \;\leq\; \mathbb{P}(\min_{i\in\mathcal{H}_0} P_i\leq\frac{\alpha}{m_0}).\end{aligned}$$

By the union bound, the probability of a false rejection is bounded above by

$$\mathbb{P}\left(\min_{i\in\mathcal{H}_0} P_i\leq\frac{\alpha}{m_0}\right) \;\leq\; \sum_{i\in\mathcal{H}_0}\mathbb{P}(P_i\leq\frac{\alpha}{m_0}) \;\leq\; \alpha.$$

□

This means that the Holm procedure is strictly more powerful than Bonferroni without any extra assumptions. However, Figure 3.3 also suggest that, when m is very large, the Holm procedure may not substantially differ from the Bonferroni correction.

3.6 *False discovery rate*

Controlling the FWER may be too conservative and greatly reduce our power to detect real effects, especially when m is large. In many modern "large-scale testing" applications, focus has shifted from FWER in (3.22) to the false-discovery proportion (FDP)

$$\text{FDP} \;=\; \frac{|\mathcal{H}_0\cap\mathcal{R}|}{|\mathcal{R}|\vee 1}$$

and on procedures that control its expected value $\mathbb{E}(\text{FDP})$, called the **false-discovery rate (FDR)**. The FDR can be interpreted as follows: if FDR ≤ 0.1, we expect around 90% of the discoveries to be true.

Controlling FDR is a shift in paradigm - we are willing to tolerate some type I errors (false discoveries), as long as most of the discoveries we make are still true. It has been argued that in applications where the statistical test is thought of as providing a "definitive answer" for whether an effect is real, FWER control is still the correct objective. In contrast, for applications where the statistical test identifies candidate effects that are likely to be real and which merit further study, it may be better to target FDR control.

3.6.1 *Benjamini-Hochberg procedure*

It is clear that the false discovery rate depends heavily on the number of true hypotheses. Thus, any procedure that controls FDR should be adaptive. As we showed earlier, if $i \in \mathcal{H}_0$ then $\mathbb{P}(P_i \leq u) \leq u$ for all $u \in (0,1)$. The Benjamini-Hochberg (BH) procedure compares the sorted p-values to a diagonal cutoff line, finds the largest p-value that still falls below this line, and rejects the null hypotheses for the p-values up to and including this one. Formally, the BH procedure at level α is defined as follows:

1. Sort the p-values. Call them $P_{(1)} \leq \cdots \leq P_{(m)}$ as before.
2. Find the largest r such that $P_{(r)} \leq \frac{r}{m}\alpha$.
3. Reject the null hypotheses $H_0^{(1)}, \ldots, H_0^{(r)}$.

Remark 3.6.1. *Just to avoid confusion with the Holm procedure, note that we do not require that $P_{(j)} \leq \frac{j}{m}\alpha$ for $j < r$.*

It is useful to observe the following.

Lemma 3.6.2. *Suppose the B-H procedure rejects exactly r hypotheses. Then $i \in \mathcal{R}$ if and only if $P_i \leq \frac{r}{m}\alpha$.*

Proof. If $i \in \mathcal{R}$ then $P_i \leq P_{(r)} \leq \frac{r}{m}\alpha$, which proves the right implication. For the left implication we argue using a contrapositive statement. Suppose $i \notin \mathcal{R}$, let s be such that $P_i = P_{(s)}$. Then $s > r$ and $P_{(s)} > \frac{s}{m}\alpha > \frac{r}{m}\alpha$. □

Theorem 3.6.3 (Benjamini and Hochberg)**.** *Consider tests of m null hypotheses. If the test statistics (or equivalently, p-values) of these tests are independent, then the FDR of the above procedure satisfies*

$$\text{FDR} \;\leq\; \alpha\frac{m_0}{m} \;\leq\; \alpha.$$

Proof. We have

$$\text{FDR} \;=\; \mathbb{E}\left(\frac{|\mathcal{H}_0 \cap \mathcal{R}|}{|\mathcal{R}| \vee 1}\right) \;=\; \mathbb{E}\left(\frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}\{i \in \mathcal{R}\}}{|\mathcal{R}| \vee 1}\right) \;=\; \sum_{i\in\mathcal{H}_0} \mathbb{E}\left(\frac{\mathbb{1}\{i \in \mathcal{R}\}}{|\mathcal{R}| \vee 1}\right).$$

Let $R = |\mathcal{R}|$ and let $R_{\setminus i}$ be the number of rejections if we replace P_i with zero and run BH_α on $P_1, \ldots, P_{i-1}, 0, P_{i+1}, \ldots, P_m$. Note that $P_i \perp\!\!\!\perp R_{\setminus i}$ and $R_{\setminus i} \geq 1$. Consider the following:

Claim: $i \in \mathcal{R} \quad\Leftrightarrow\quad R = R_{\setminus i} \quad\Leftrightarrow\quad P_i \leq \frac{\alpha R_{\setminus i}}{m}$.

Before we prove the claim, we show how it helps us to prove the

theorem. Assuming that the claim holds

$$\begin{aligned}\text{FDR} &= \sum_{i\in\mathcal{H}_0}\mathbb{E}\left(\frac{\mathbb{1}\{P_i \le \frac{\alpha R_{\searrow i}}{m}\}}{R_{\searrow i}}\right) = \sum_{i\in\mathcal{H}_0}\mathbb{E}\left(\frac{\mathbb{P}(P_i \le \frac{\alpha R_{\searrow i}}{m}|R_{\searrow i})}{R_{\searrow i}}\right)\\ &\overset{P_i \perp\!\!\!\perp R_{\searrow i}}{\le} \sum_{i\in\mathcal{H}_0}\mathbb{E}\left(\frac{\frac{\alpha R_{\searrow i}}{m}}{R_{\searrow i}}\right) = \alpha\frac{m_0}{m} \le \alpha.\end{aligned}$$

It remains to prove the claim. We do it in several steps.

[$i \in \mathcal{R} \Rightarrow P_i \le \frac{\alpha R_{\searrow i}}{m}$]: First note that BH_α is monotone in the p-values: If $P_i \ge P_i'$ for all then $R \le R'$. From this it follows that $R \le R_{\searrow i}$. Suppose $i \in \mathcal{R}$. Then

$$P_i \le \frac{\alpha R}{m} \le \frac{\alpha R_{\searrow i}}{m}.$$

[$P_i \le \frac{\alpha R_{\searrow i}}{m} \Rightarrow R = R_{\searrow i}$]: By definition of $R_{\searrow i}$, we must have $R_{\searrow i}$ many values from $P_1, \ldots, P_{i-1}, 0, P_{i+1}, \ldots, P_m$ which are $\le \frac{\alpha R_{\searrow i}}{m}$. Since $P_i \le \frac{\alpha R_{\searrow i}}{m}$, it follows that there are $R_{\searrow i}$ many values from $P_1, \ldots, P_{i-1}, P_i, P_{i+1}, \ldots, P_m$ which are $\le \frac{\alpha R_{\searrow i}}{m}$. In particular, $R \ge R_{\searrow i}$. The other inequality was shown above and so we get equality.

[$R = R_{\searrow i} \Rightarrow i \in \mathcal{R}$] We prove this implication by contradiction. Suppose $R = R_{\searrow i}$ but $i \notin \mathcal{R}$. If $i \notin \mathcal{R}$ then $P_i > \frac{\alpha R}{m}$ and there are R many P_j's with $P_j \le \frac{\alpha R}{m}$. Now run BH_α on $P_1 \ldots, P_{i-1}, 0, P_{i+1}, \ldots, P_m$. Then there are $R+1$ many values $\le \frac{\alpha R}{m} \le \frac{\alpha(R+1)}{m}$. Thus, we must have $R_{\searrow i} \ge R+1 > R$, which gives contradiction.

□

← Exercise 3.7.5

3.6.2 *Benjamini-Yekutieli justification*

The main problem with the validation of the Benjamini-Hochberg procedure is that it requires that the hypotheses are independent. Handling uniformly all dependence structure between the hypotheses is hard. One popular setting where the B-H procedure can be still validated is when the dependence between the P-values satisfies a form of positive dependence. We will now discuss this in more detail.

A set $S \subset \mathbb{R}^n$ is nondecreasing if $\boldsymbol{x} \in S$ and $\boldsymbol{y} \ge \boldsymbol{x}$ implies that $\boldsymbol{y} \in S$.

Definition 3.6.4. *A random vector $\boldsymbol{X} = (X_1, \ldots, X_m)$ is positively regression dependent on $I \subseteq \{1, \ldots, m\}$ (PRDS) if $\mathbb{P}(\boldsymbol{X} \in S|X_i = x_i)$ is non-decreasing in x_i for every nondecreasing set S and any $i \in I$.*

If $(X_1, \ldots, X_m)$ is PDRS on I and $Y_i := f_i(X_i)$ for all $1 \le i \le m$ with f_i strictly increasing or decreasing, then $(Y_1, \ldots, Y_m)$ is PRDS on I as well. Tranformation of this form are called co-monotone transformations. Thus PRDS property is preserved under co-monotone

transformations. It follows that $P_i = \mathbb{P}(T_i \leq X_i)$ is PDRS as well (at least in the continuous case).

Theorem 3.6.5. *If the joint distribution of $\boldsymbol{P} = (P_1, \ldots, P_m)$ is PRDS on the subset $\mathcal{H}_0$, the Benjamini-Hochberg procedure controls the FDR at level $\frac{m_0}{m}\alpha$.*

The proof of this result relies on the following proposition.

Proposition 3.6.6. *If $\boldsymbol{P}$ is PRDS on the set of true nulls, then the function $\mathbb{P}(\boldsymbol{P} \in S | P_i \leq t)$ for $i \in \mathcal{H}_0$ is non-decreasing in t for S a non-decreasing set.*

Proof. For any t, $\mathbb{P}(\boldsymbol{P} \in S | P_i \leq t) = \frac{\mathbb{P}(\boldsymbol{P} \in S, P_i \leq t)}{\mathbb{P}(P_i \leq t)}$. For $t > t'$, we have that

$$\mathbb{P}(\boldsymbol{P} \in S | P_i \leq t') = \frac{\mathbb{P}(\boldsymbol{P} \in S, P_i \leq t) + \mathbb{P}(\boldsymbol{P} \in S, P_i \in (t, t'])}{\mathbb{P}(P_i \leq t) + \mathbb{P}(P_i \in (t, t'])}$$

To show that $\mathbb{P}(\boldsymbol{P} \in S | P_i \leq t') \leq \mathbb{P}(\boldsymbol{P} \in S | P_i \leq t)$, it suffices to show that

$$\frac{\mathbb{P}(\boldsymbol{P} \in S, P_i \leq t)}{\mathbb{P}(P_i \leq t)} \leq \frac{\mathbb{P}(\boldsymbol{P} \in S, P_i \in (t, t'])}{\mathbb{P}(P_i \in (t, t'])}.$$

The last statement is because for any positive number a, b, c, d, $\frac{a}{b} \leq \frac{a+c}{b+d}$ if and only if $\frac{a}{b} \leq \frac{c}{d}$. If F_i denotes the CDF of P_i then

$$\begin{aligned}
\mathbb{P}(\boldsymbol{P} \in S, P_i \in (t, t']) &= \mathbb{E}\mathbb{1}\{\boldsymbol{P} \in S, P_i \in (t, t']\} \\
&= \mathbb{E}[\mathbb{1}\{P_i \in (t, t']\}\mathbb{E}(\mathbb{1}\{\boldsymbol{P} \in S\} | P_i)] \\
&= \mathbb{E}[\mathbb{1}\{P_i \in (t, t']\}\mathbb{P}(\boldsymbol{P} \in S | P_i)] \\
&= \int_t^{t'} \mathbb{P}(\boldsymbol{P} \in S | P_i = s) \mathrm{d}F_i(s) \\
&\overset{(PRDS)}{\geq} \int_t^{t'} \mathbb{P}(\boldsymbol{P} \in S | P_i = t) \mathrm{d}F_i(s) \\
&= \mathbb{P}(\boldsymbol{P} \in S | P_i = t)\mathbb{P}(P_i \in (t, t']).
\end{aligned}$$

Similarly we have

$$\begin{aligned}
\mathbb{P}(\boldsymbol{P} \in S, P_i \leq t) &= \int_0^{t} \mathbb{P}(\boldsymbol{P} \in S | P_i = s) \mathrm{d}F_i(s) \\
&\overset{(PRDS)}{\leq} \int_0^{t} \mathbb{P}(\boldsymbol{P} \in S | P_i = t) \mathrm{d}F_i(s) \\
&= \mathbb{P}(\boldsymbol{P} \in S | P_i = t)\mathbb{P}(P_i \leq t).
\end{aligned}$$

We have just shown that

$$\frac{\mathbb{P}(\boldsymbol{P} \in S, P_i \leq t)}{\mathbb{P}(P_i \leq t)} \leq \mathbb{P}(\boldsymbol{P} \in S | P_i = t) \leq \frac{\mathbb{P}(\boldsymbol{P} \in S, P_i \in (t, t'])}{\mathbb{P}(P_i \in (t, t'])}$$

as claimed. □

Proof of Theorem 3.6.5. In the proof of Theorem 3.6.3 we noted that

$$\text{FDR} \;=\; \sum_{i\in\mathcal{H}_0} \mathbb{E}\left(\frac{\mathbb{1}\{i\in\mathcal{R}\}}{|\mathcal{R}|\vee 1}\right)$$

Note that it is enough to show that $\mathbb{E}\left(\frac{\mathbb{1}\{i\in\mathcal{R}\}}{|\mathcal{R}|\vee 1}\right) \le \alpha/m$ for $i\in\mathcal{H}_0$.

By Lemma 3.6.2, if r rejections are made then H_0^i is rejected if and only if $P_i \le \frac{\alpha r}{m}$. Hence, $\mathbb{1}\{i\in\mathcal{R}\} = \mathbb{1}\{P_i \le \frac{\alpha r}{m}\}$. This gives

$$\mathbb{E}\left(\frac{\mathbb{1}\{i\in\mathcal{R}\}}{R\vee 1}\right) \;=\; \mathbb{E}\left(\mathbb{E}\left[\frac{\mathbb{1}\{i\in\mathcal{R}\}}{R\vee 1}|R\right]\right) \;=\; \mathbb{E}\left(\frac{1}{R\vee 1}\mathbb{P}[i\in\mathcal{R}|R]\right) \;=\; \sum_{r=1}^{m}\frac{1}{r}\mathbb{P}(P_i\le\tfrac{\alpha r}{m}, R=r).$$

For any true null we now have

$$\begin{aligned}\sum_{r=1}^{m}\frac{1}{r}\mathbb{P}(P_i\le\tfrac{\alpha r}{m}, R=r) &= \sum_{r=1}^{m}\frac{1}{r}\mathbb{P}(P_i\le\tfrac{\alpha r}{m})\mathbb{P}(|\mathcal{R}|=r|P_i\le\tfrac{\alpha r}{m})\\ &\le \frac{\alpha}{m}\sum_{r=1}^{m}\mathbb{P}(R=r|P_i\le\tfrac{\alpha r}{m}).\end{aligned}$$

Hence, it suffices to show that $\sum_{r=1}^{m}\mathbb{P}(R=r|P_i\le\frac{\alpha r}{m})\le 1$. Observe that $\{R\le r\}$ is an increasing event, that is, it can be written as $\{\boldsymbol{P}\in S\}$ for some non-decreasing set S. This is because increasing all p-values increases the p-value at each rank. Hence, any ranked p-value above the threshold remains above it's threshold, that is, we accept at least as many as before and hence, do not reject more hypotheses. Using this, we get that

$$\begin{aligned}\sum_{r=1}^{m}\mathbb{P}(R=r|P_i\le\tfrac{\alpha r}{m}) &= \left(\mathbb{P}(R\le m|P_i\le\alpha)-\mathbb{P}(R\le 0|P_i\le\tfrac{\alpha}{m})\right)\\ &+ \sum_{r=1}^{m-1}\left(\mathbb{P}(R\le r|P_i\le\tfrac{\alpha r}{m})-\mathbb{P}(R\le r|P_i\le\tfrac{\alpha(r+1)}{m})\right).\end{aligned}$$

Note that each summand in the second line is non-positive since $\mathbb{P}(R\le r|P_i\le x)$ is increasing in x. Also $\mathbb{P}(R\le 0|P_i\le\frac{\alpha}{m})\ge 0$, which implies that

$$\sum_{k=1}^{m}\mathbb{P}(R=k|P_i\le\tfrac{\alpha k}{m}) \;\le\; \mathbb{P}(R\le m|P_i\le\alpha) \;\le\; 1.$$

As stated before, this proves the upper bound on the FDR. □

3.7 Exercises

Exercise 3.7.1. *Show that if an α-level test maximizes $\mathbb{E}_1\varphi$ then it must be essentially equal to the likelihood ratio test. (Hint: Use the proof of Theorem 3.1.1)*

Exercise 3.7.2. *Show that the threshold $\lambda = 0$ in Example 3.1.4 is optimal to optimize the Bayes risk with equal weights on each class. Hint: Note that $Z = \Sigma^{-1}(X - \mu_0)$ is standard normal if X comes from the first class.*

Exercise 3.7.3. *Show that if Z is d-dimensional standard Gaussian vector then $D = \|Z\|$ and $U = Z/\|Z\|$ are independent. Conclude that every d-dimensional vector $X \sim N_d(\mu, \Sigma)$ admits a stochastic representation $X = \mu + D\Sigma^{1/2}U$, where $D \perp\!\!\!\perp U$ with $D^2 \sim \chi^2_d$ and U being uniformly distributed on the unit sphere.*

Exercise 3.7.4 (Correlation test)**.** *Given a sample of size $n > 2$ from a bivariate normal distribution, use the same type of procedure as in Example 3.2.5 to derive an exact test of the hypothesis that the two variates are uncorrelated. That is: specify u and v, find a function of them that is parameter-free under H_0, conclude independence, and go over to the marginal distribution. Finally transform to a test statistic of known distribution. Hint: $\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$ is exactly t_{n-2}-distributed under H_0.*

Exercise 3.7.5. *Suppose we perform* 10 *tests (e.g. test the association between* 10 *different outcomes and a potential prognostic factor) and obtain the p-values* 0.0140, 0.2960, 0.9530, 0.0031, 0.1050, 0.6410, 0.7810, 0.9010, 0.0053, 0.4500*.*

1. *Which hypotheses are rejected after Bonferroni correction?*
2. *Which hypotheses are rejected after Holm correction?*
3. *Which hypotheses are rejected using the Benjamini and Hochberg procedure?*

Exercise 3.7.6. *Consider $2m$ independent coins $X_i \sim \text{Bern}(\theta_i)$ $i = 1, \ldots, m$. Suppose that $\theta_i = \frac{1}{2} + \frac{i}{2m+1}$ for $i = 1, \ldots, m$ and $\theta_i = \frac{1}{2}$ for $i = m+1, \ldots, 2m$. For each of the coins we make n independent tosses.*

(a) For each i find the most powerful test for testing $\theta = \frac{1}{2}$ against $\theta > \frac{1}{2}$ for level $\alpha \leq 0.05$ (finding the test of size exactly α may be too hard).

(b) Consider now the multiple testing problem for all m coins. We proved that Bonferroni and the Holm procedures both control the FWER. Describe how Bonferroni will look in this case. We want FWER≤ 0.05.

(c) Take a look at the power of Bonferroni. Denote by $\mathbb{P}_i$ the distribution $\text{Bern}(\theta_i)$ for $i = 1, \ldots, m$ (these are the coins for which the null does not hold). Try to find some sufficient conditions to bound the probability of the type II error for each of the false nulls by, say, 0.05*.*

(d) Provide some simulations to see the difference in power between the Bonferroni and Holm procedures. For fixed m consider $n = m, 10m, 100m$ to see how the answer depends on the ratio between n and m.

(some parts of this problem can be approached in various ways so there is no one correct solution)

Part II

Statistical Learning Theory: An Empirical Process Perspective

4
Motivation and examples (1 week)

This part of the lecture focuses on statistical learning theory taking the empirical process perspective. These techniques have become now a standard toolbox for studying modern statistical scenarios. We will introduce basic tools in emprirical processes and apply those tools in mathematical statistics and machine learning. We will focus on the non-asymptotic perspective[1].

We start by introducing the main objects of this theory and some motivating examples[2].

[1] Good references:
Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge, 2019; and A. W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes—with applications to statistics*. Springer Series in Statistics. Springer, Cham, 2023. Second edition

[2] Thanks go to Qiang Sun for inspiring this. Part of the material was discussed by Wenlong in the Fall semester.

4.1 Uniform law of large numbers

Suppose $X, X_1, \dots, X_n$ are independent and identically distributed random variables taking values in $\mathcal{X} \subseteq \mathbb{R}$. Let the CDF of the underlying distribution be $F(t) = \mathbb{P}(X \le t)$ for $t \in \mathbb{R}$. The empirical CDF $\widehat{F}_n$ is

$$\widehat{F}_n(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le t\}.$$

As $\mathbb{E}\mathbb{1}\{X \le t\} = \mathbb{P}(X \le t)$, it is clear that, for every $t \in \mathbb{R}$, $\mathbb{E}\widehat{F}_n(t) = F(t)$. By the strong law of large numbers, for every t, $\widehat{F}_n(t) \overset{\text{a.s.}}{\to} F(t)$ as $n \to \infty$. The following stronger result is well-known.

Theorem 4.1.1 (Glivenko-Cantelli)**.** *For any distribution, the empirical CDF $\widehat{F}_n$ is a strongly consistent estimator of the population CDF in the uniform norm, meaning that*

$$\|\widehat{F}_n - F\|_\infty \;:=\; \sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - F(t)| \;\overset{\text{a.s.}}{\to} 0.$$

We will provide the proof of this result using a more general theory in Section 6.3; c.f. Proposition 6.3.6.

One reason, this result can be useful is because in many situations we want to estimate some functional $\gamma(F)$ of the population CDF. For example $\gamma_g(F) := \int g(x)\mathrm{d}F(x)$ is the expectation $\mathbb{E}g(X)$. Also, for

any $\alpha \in [0,1]$, the quantile functional Q_α is given

$$Q_\alpha(F) \;:=\; \inf\{t \in \mathbb{R}; F(t) \geq \alpha\}.$$

If we have access to F only through the sample CFD $\widehat{F}_n$. A natural way to estimate $\gamma(F)$ is using the plug-in estimator $\gamma(\widehat{F}_n)$. We could define convergence $\widehat{F}_n \to F$ in the sup-norm, as $\|\widehat{F}_n - F\|_\infty \overset{\text{a.s.}}{\to} 0$. Then, we get that $\gamma(\widehat{F}_n)$ is almost surely consistent for $\gamma(F)$ as long as the functional γ is continuous with respect to the sup-norm. Indeed, by continuity in the sup-norm

$$\{\omega : \|\hat{F}_n - F\|_\infty \to 0\} \;\subseteq\; \{\omega : \|\gamma(\hat{F}_n) - \gamma(F)\|_\infty \to 0\}.$$

By Theorem 4.1.1, the set on the left has measure 1 and so the set on the right also has measure one proving $\gamma(\hat{F}_n) \overset{\text{a.s.}}{\to} \gamma(F)$.

Recall that a function $f : \mathcal{X} \to \mathbb{R}$ on a metric space $(\mathcal{X}, d)$ is continuous if $p_n \to p$ in $\mathcal{X}$ implies that $f(p_n) \to f(p)$ in $\mathbb{R}$. Also recall that $p_n \to p$ is equivalent to $d(p_n, p) \to 0$.

We are also interested in the following generalization of Theorem 4.1.1. Let $\mathcal{F}$ denote a class of integrable real-valued functions on $\mathcal{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution $\mathbb{P}$ over $\mathcal{X}$. We want to analyze the following quantity

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \;:=\; \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X) \right|, \tag{4.1}$$

Note that there may be measurability concerns associated with this definition. We will skip the details; see Section 4.4 in Wainwright's book for some details.

which measures the absolute deviation between the sample average $\frac{1}{n}\sum f(X_i)$ and the population average $\mathbb{E}[f(X)]$, uniformly over class $\mathcal{F}$.

Definition 4.1.2. *We say that $\mathcal{F}$ is a **Glivenko-Cantelli class** for $\mathbb{P}$ if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \to 0$ almost surely as $n \to \infty$. If the convergence in probability holds, we say that $\mathcal{F}$ is a weak Glivenko-Cantelli class for $\mathbb{P}$.*

If $\mathcal{F}$ is the family of indicator functions $f(x) = \mathbb{1}\{x \leq t\}$ for some $t \in \mathbb{R}$, we recover the CDF example in Theorem 4.1.1. In Chapter 6 we discuss some other examples of **Glivenko-Cantelli classes**. In general, the condition is that the class cannot be too rich as illustrated by the example below.

Example 4.1.3. *Suppose that $\mathcal{F}$ is the set of all indicator functions $\mathbb{1}_A(x)$ for all measurable sets $A \subseteq \mathbb{R}$ and suppose that $\mathbb{P}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. We have $\mathbb{E}\mathbb{1}_A(X) = \mathbb{P}(X \in A)$. For any $x_1, \ldots, x_n \in \mathbb{R}$ the set $A = \mathbb{R} \setminus \{x_1, \ldots, x_n\}$ is measurable and $\mathbb{P}(X \in A) = 1$. We thus get that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = 1$ for all n, and so, this class of indicator functions is not Glivenko-Cantelli.*

Even if $\mathcal{F}$ is not a Glivenko-Cantelli class, we still want to control the size (4.1) by anserwing the following questions:

1. Does the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ in (4.1) concentrate around its expectation? What is the size of this expectation?

2. Provide finite-sample bounds on $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$, that is, bounds that hold for every n, in terms of the class function and the distribution $\mathbb{P}$ of X.

The strategy for controlling (4.1) is to first show that this quantity concentrates around its mean. Given the techniques described in Chapter 5, this is often easy[3]. Second, we need to control the mean. This is often more complicated and requires more advanced techniques that we overview in Chapter 6. Before we discuss more technical aspects, we first discuss one motivating example in machine learning and one in statistics.

[3] For instance, for the class in Example 4.1.3, we show concentration around the mean in Example 5.3.6.

4.1.1 *Example: Binary classification*

Consider a random vector (X, Y) with values in $\mathcal{X} \times \{-1, 1\}$. A classifier is a function $g : \mathcal{X} \to \{-1, 1\}$. The error of the classifier is given by

$$R(g) := \mathbb{P}(g(X) \neq Y).$$

In the language of the statistical decision theory we first define the loss $L(g) = \mathbb{1}\{g(X) \neq Y\}$. Then the error $R(g)$ is simply the risk of the classifier g.
← Exercise 4.3.1

The goal of binary classification is to construct a classifier with small error based on n *i.i.d.* observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ having the same distribution as (X, Y). The empirical error of the classifier g is

$$R_n(g) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(g(X_i) \neq Y_i).$$

A natural strategy for classification is to pick a class of classifiers $\mathcal{C}$ and then to choose the classifier in $\mathcal{C}$ which has the smallest **training error**

$$\widehat{g}_n := \arg\min_{g \in \mathcal{C}} R_n(g).$$

It is natural to associate each classifier with the set $\{x : g(x) = 1\}$. Thus, the class $\mathcal{C}$ can be identified with a set of measurable subsets of $\mathcal{X}$.

Here one possible choice is to use the logistic regression in Example 1.7.2. A good classifier should have a small **out of sample test error**

$$R(\widehat{g}_n) = \mathbb{P}(\widehat{g}_n(X) \neq Y | (X_1, Y_1), \ldots, (X_n, Y_n)).$$

One possible question concerns how close is $\widehat{g}_n$ to the optimal classifier $g^* = \arg\min_{g \in \mathcal{C}} R(g)$. To compare $R(\widehat{g}_n)$ with $R(g^*)$ note that

$$\begin{aligned} R(\widehat{g}_n) &= R(g^*) + R(\widehat{g}_n) - R_n(\widehat{g}_n) + R_n(\widehat{g}_n) - R(g^*) \\ &\leq R(g^*) + R(\widehat{g}_n) - R_n(\widehat{g}_n) + R_n(g^*) - R(g^*) \\ &\leq R(g^*) + 2 \sup_{g \in \mathcal{C}} |R_n(g) - R(g)|. \end{aligned} \tag{4.2}$$

Another question could be about comparing the training arror and the test error, which amounts to comparing $R(\widehat{g}_n)$ with $R_n(\widehat{g}_n)$. Here

we have

$$R(\widehat{g}_n) = R_n(\widehat{g}_n) + R(\widehat{g}_n) - R_n(\widehat{g}_n) \tag{4.3}$$
$$\leq R_n(\widehat{g}_n) + \sup_{g\in\mathcal{C}} |R_n(g) - R(g)|. \tag{4.4}$$

With the analogous bound on $R_n(\widehat{g}_n)$ we conclude that $|R_n(\widehat{g}_n) - R(\widehat{g}_n)| \leq \sup_{g\in\mathcal{C}} |R_n(g) - R(g)|$. The key quantity in both cases is

$$\sup_{g\in\mathcal{C}} |R_n(g) - R(g)|,$$

which is a special case of (4.1), when $\mathcal{F}$ is taken to be the class of all functions $\mathbb{1}(g(x) \neq y)$ as g varies over $\mathcal{C}$, where the data are (X_i, Y_i) instead of X_i. We provide more details in Section 6.3.1 after developing necessary theory.

4.1.2 *Example: M-Estimation*

Let $X, X_1, \ldots, X_n$ are i.i.d. from $\mathbb{P}$ where $\mathbb{P} \in \mathcal{P}$, with $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ for $\Theta \subseteq \mathbb{R}^d$. A popular method of finding an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is to minimize a criterion function of the form

$$\theta \mapsto M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i).$$

Here $m_\theta : \mathcal{X} \mapsto \mathbb{R} \cup \{+\infty\}$ are known functions. An estimator maximizing $M_n(\theta)$ is called an M-estimator. Often the minimizer is sought by setting a derivative equal to zero. Therefore, the name M-estimator is also used for estimators satisfying systems of equations of the type

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i) = \mathbf{0}. \tag{4.5}$$

Here $\psi_\theta : \mathcal{X} \to \mathbb{R}^d$ are known vector-valued maps.

Example 4.1.4 (Maximum likelihood estimators)**.** *Suppose $X, X_1, \ldots, X_n$ have a common density p_θ. Then the maximum likelihood estimator maximizes the likelihood $\prod_i p_\theta(X_i)$, or equivalently*

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i).$$

This is an expectation of $\log p(X)$ with respect to the sample distribution.

Thus a maximum likelihood estimator is an M-estimator with $m_\theta = -\log p_\theta$. If the density is differentiable with respect to θ for each fixed x, then the maximum likelihood estimator also solves an equation of type (4.5) with $\psi_\theta = \nabla_\theta \log p_\theta$, the score function of the model.

Other simple examples of M-estimators include the sample median (c.f. Exercise 2.7.2), the least squares estimator, and more generally, the estimators obtained by minimizing the empirical risk (2.1) in

which case $m_\theta(x) = L(\theta, \delta(x))$. The general theory of M-estimators originated in robust statistics.

Example 4.1.5 (Huber Robust loss)**.** *Consider a regression problem. Let* $\mathbf{y} \in \mathbb{R}^n$ *and* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *be the design matrix with rows* $\mathbf{x}_i, \ldots, \mathbf{x}_n$*. Let* $L(y, a) = \frac{1}{2}(y-a)^2$*. In the standard least squares approach, the vector of coefficient is estimated by minimizing the empirical risk*

$$m(\beta) \;=\; \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} L(y_i, \beta^T\mathbf{x}_i).$$

Consider instead the loss

Note that $L_\delta(y, a)$ is differentiable everywhere.

$$L_\delta(y, a) \;=\; \begin{cases} \frac{1}{2}(y-a)^2 & \text{if } |y-a| \le \delta \\ \delta(|y-a| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}.$$

The corresponding M-estimator is the minimizer of

$$m_\delta(\beta) \;=\; \frac{1}{n}\sum_{i=1}^{n} L_\delta(y_i, \beta^T\mathbf{x}_i).$$

Let $M(\theta) = \mathbb{E}m_\theta(X)$. In the theory of M-estimation, the target quantity for the estimator $\widehat{\theta}_n$ is

$$\theta^* \;:=\; \arg\min_{\theta\in\Theta} M(\theta).$$

For example, suppose data come from the distribution $\mathbb{P}$ with density q. Let $m_\theta(X) = -\log p_\theta(X)$. Then

$$\arg\min_\theta -\mathbb{E}\log p_\theta(X) \;=\; \arg\min_\theta \mathbb{E}\log\frac{q(X)}{p_\theta(X)},$$

which is simply the Kullback-Leibler divergence between the distribution $\mathbb{P}$ and $\mathbb{P}_\theta$. In particular, if $\mathbb{P} \in \mathcal{P}$ then it is the unique minimizer.

The main question of interest while studying M-estimators concerns accuracy of $\widehat{\theta}_n$ for estimating θ^*. In the asymptotic regime $n \to \infty$, the two key questions are:

1. Is $\widehat{\theta}_n$ consistent for estimating θ^*, equivalently, does $d(\widehat{\theta}_n, \theta^*) \xrightarrow{p} 0$; see Section 8.1 for relevant definitions and basic results.

2. If yes, what is the rate of convergence of $d(\widehat{\theta}_n, \theta^*)$ to zero? The usual rate of convergence is $O_p(n^{-1/2})$, so that $\sqrt{n}(\widehat{\theta}_n - \theta^*)$ converges in distribution[4].

3. How shall we do inference on θ^* using $\widehat{\theta}_n$. For this we may need to understand better how $\widehat{\theta}_n$ concentrates around θ^*.

[4] For a simple example see Exercise 8.6.5.

Some basic asymptotic results for M-estimators are provided in Appendix 8.3. Here, we complement this with the link to the uniform law of large numbers. For the first question we investigate closeness of $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$ to $M(\theta) = \mathbb{E}m_\theta(X)$ in some sort of uniform sense over θ, which leads to investigation of (4.1) for

$$\mathcal{F} = \{m_\theta : \theta \in \Theta\}$$

and then translate the result back to the result on $d(\widehat{\theta}_n, \theta^*)$.

This is how this can be done: Let $D_\epsilon = \{\theta \in \Theta : d(\theta, \theta^*) \geq \epsilon\}$. We can bound $d(\widehat{\theta}_n, \theta^*)$ as

$$\begin{aligned}
\mathbb{P}(d(\widehat{\theta}_n, \theta^*) \geq \epsilon) &\leq \mathbb{P}(\sup_{\theta \in D_\epsilon}(M_n(\theta^*) - M_n(\theta)) \geq 0) \\
&\leq \mathbb{P}\left(\sup_{\theta \in D_\epsilon}\{(M_n(\theta^*) - M(\theta^*)) - (M_n(\theta) - M(\theta))\} \geq -\sup_{\theta \in D_\epsilon}\{M(\theta^*) - M(\theta)\}\right) \\
&\leq \mathbb{P}\left(2\sup_{\theta \in D_\epsilon}\{|M_n(\theta) - M(\theta)|\} \geq \inf_{\theta \in D_\epsilon}\{M(\theta) - M(\theta^*)\}\right),
\end{aligned}$$

where going from the first to the second line we used the fact that $\sup_\theta f(\theta) \leq \sup_\theta g(\theta) + \sup_\theta(f(\theta) - g(\theta))$. Note that the bound above again relates to the random quantity (4.1). If θ^* is uniquely and globally identifiable, that is, there exists $\eta > 0$ depending on ϵ such that $\inf_{\theta \in D_\epsilon}(M(\theta) - M(\theta^*)) > \eta$. Thus, with proper control of the size of $\sup_{\theta \in D_\epsilon}|M_n(\theta) - M(\theta)|$, the right-hand side in the above display diminishes to zero[5].

[5] We get $\mathbb{P}(d(\widehat{\theta}, \theta^*) \geq \epsilon) \leq \mathbb{P}(\sup_{\theta \in D_\epsilon}|M_n(\theta) - M(\theta)| \geq \eta/2)$.

4.2 *The Uniform Central Limit Theory*

The classical central limit theorem (CLT) studies the following type of results: for an *i.i.d.* sequence $X, X_1, \ldots, X_n$

$$\mathbb{G}_n(f) := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right) \rightsquigarrow N(0, \text{var}(f(X))). \quad (4.6)$$

The uniform central limit theorem studies the above convergence in distribution uniformly over f in the class $\mathcal{F}$. To illustrate this idea consider the following example.

4.2.1 *Example: Uniform empirical process*

Suppose $X, X_1, \ldots, X_n$ are i.i.d. uniform on $[0, 1]$. Let $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]}(x) : t \in \mathbb{R}\}$. Define

$$U_n(t) = \sqrt{n}(\widehat{F}_n(t) - F(t)), \qquad t \in \mathbb{R}, \quad (4.7)$$

where $\widehat{F}_n(t) = \frac{1}{n}\sum_i \mathbb{1}\{X_i \geq t\}$ and $F(t) = \mathbb{P}(X \leq t) = t$.

This defines a stochastic process $\{U_n(t) : t \in \mathbb{R}\}$ which is typically referred to as an **empirical process** or, more specifically in this case, a uniform empirical process.

How do we find a candidate for the limit? The CLT states that, for each $t \in [0,1]$, $U_n(t) \rightsquigarrow N(0, t(1-t))$ as $n \to \infty$[6]. Moreover, for every fixed $t_1, \ldots, t_k$, the multivariate CLT states that

$$(U_n(t_1), \ldots, U_n(t_k)) \rightsquigarrow N(0, \Sigma),$$

where $\Sigma = (\Sigma_{ij})$, $\Sigma_{ij} = t_i \wedge t_j - t_i t_j$. This follows simply because we can rewrite $U_n(t) = \frac{1}{\sqrt{n}} \sum_i (\mathbb{1}(X_i \leq t) - t)$ and, for any t, t',

$$\begin{aligned} \mathbb{E}U_n(t)U_n(t') &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(\mathbb{1}(X_i \leq t) - t)(\mathbb{1}(X_i \leq t') - t')] \\ &= \frac{1}{n}\sum_{i=1}^{n} (t \wedge t' - tt') \;=\; t \wedge t' - tt'. \end{aligned}$$

In consequence, the candidate for the limiting process is the Gaussian process with kernel $\kappa(s,t) = s \wedge t - st$.

We will be interested in the underlying limiting process called a Brownian bridge.

Definition 4.2.1 (Brownian bridge)**.** *Brownian bridge $\{U(t) : 0 \leq t \leq 1\}$ is a stochastic process satisfying the following conditions*

1. *Every realization is continuous in $[0,1]$ with $U(0) = U(1) = 0$.*
2. *For every fixed $t_1, \ldots, t_k$, the vector $(U(t_1), \ldots, U(t_k)) \sim N(0, \Sigma)$, where $\sigma_{ij} = t_i \wedge t_j - t_i t_j$.*

Thus, the finite dimensional representation of $\{U_n(t) : t \in [0,1]\}$ converges in distribution to that of $\{U(t) : t \in [0,1]\}$. It is then natural to ask whether the entire process $\{U_n(t)\}$ converges in distribution to the process $\{U(t)\}$.

Convergence of stochastic processes can be defined as follows. First recall the definition of convergence in distribution given in Section 8.1. Equivalently, by Portmanteau Lemma 8.1.1, we say that a random sequence (Z_n) with values in $\mathbb{R}^k$ converges in distribution to Z if and only if $\mathbb{E}h(Z_n) \to \mathbb{E}h(Z)$ for every $h \in \mathcal{C}_b(\mathbb{R}^k)$, where $\mathcal{C}_b(\mathbb{R}^k)$ denotes the class of real-valued bounded functions on $\mathbb{R}^k$. This equivalent definition can now be generalized to stochastic processes.

Definition 4.2.2. *For any set $\mathcal{F}$ by $\ell^\infty(\mathcal{F})$ denote the set of bounded real-valued functions on $\mathcal{F}$ with the sup-norm.*

In particular, $\ell^\infty([0,1])$ is the space of all bounded functions on $[0,1]$ with metric $\rho(f,g) := \sup_{t\in[0,1]} |f(t) - g(t)|$.

[6] We have $\mathbb{E}U_n(t) = 0$ and $\mathrm{var}(U_n(t)) = \mathrm{var}(\mathbb{1}\{X \leq t\})$

A Brownian bridge is an example of a Gaussian process on $[0,1]$ with the underlying kernel function $\kappa(s,t) = s \wedge t - st$.

Definition 4.2.3. *The process $\{U_n(t)\}$ converges to $\{U(t)\}$ in distribution if*

$$\mathbb{E}h(U_n) \to \mathbb{E}h(U) \qquad \text{as } n \to \infty$$

for every bounded and continuous real-valued function $h : \ell^\infty[0,1] \to \mathbb{R}$.

There is one measure-theoretic issue with this definition as for some h, $h(U_n)$ could not be measurable. A technical remedy is to replace expectation by the outer expectation $\mathbb{E}^* h(U_n)$ induced by the outer measure. For details see Chapter 1 in [7].

[7] A. W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes—with applications to statistics*. Springer Series in Statistics. Springer, Cham, 2023. Second edition

4.2.2 *The general case*

Consider the general empirical process $\mathbb{G}_n(f)$ defined in (4.6), where $f \in \mathcal{F}$. Under the assumption that $\sup_{f\in\mathcal{F}} |f(x)| < \infty$ for every $x \in \mathcal{X}$, the function $f \mapsto \mathbb{G}_n(f)$ belongs to $\ell^\infty(\mathcal{F})$.

Definition 4.2.4. *We say that $\mathcal{F}$ is a **Donsker class** (or $\mathbb{P}$-Donsker) if $\mathbb{G}_n(\mathcal{F}) = \{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ converges in distribution in $\ell^\infty(\mathcal{F})$ to some $\mathbb{G}(\mathcal{F}) = \{\mathbb{G}(f) : f \in \mathcal{F}\}$ as $n \to \infty$.*

The limiting process $\mathbb{G}(\mathcal{F})$ is a Gaussian process: for every $f_1, \dots, f_k$ the vector $(G(f_1), \dots, G(f_k))$ is a multivariate Gaussian distribution.

This all looks abstract but the following example should illustrate importance of these considerations in statistics.

Example 4.2.5 (A goodness-of-fit statistics)**.** *Consider $X, X_1, \dots, X_n$* i.i.d. *from a distribution $\mathbb{P}$ on $\mathbb{R}$ with CDF F. Suppose we want to test $H_0 : F = F_0$ versus $H_1 : F \neq F_0$. Kolmogorov proposed the following test statistics*

$$D_n := \sqrt{n} \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_0(x)|. \tag{4.8}$$

It turns out that, under H_0, the distribution of D_n does not depend on F_0; see Exercise 4.3.2. If F_0 is strictly increasing and continuous we can then with no loss of generality assume that F_0 is a uniform distribution on $[0,1]$; $F_0(t) = t$ for $t \in [0,1]$. We obtain that under the null

← Exercise 4.3.2

$$D_n = \sup_{0\le t\le 1} |U_n(t)| = \|U_n\|_\infty \to \|U\|_\infty,$$

where $U_n(t)$ was defined in (4.7). Thus[8]

[8] $\mathbb{P}_n(D_n \le x) = \mathbb{E}\mathbb{1}\{D_n \le x\} = \mathbb{E}\mathbb{1}\{\|U_n\|_\infty \le x\}$

$$\lim_{n\to\infty} \mathbb{P}(D_n \le x) = \mathbb{P}(\sup_{0\le t\le 1} |U(t)| \le x).$$

Example 4.2.6 (Asymptotics of MLE)**.** *Suppose $X, X_1, \dots, X_n$ are iid from $\mathbb{P}_{\theta_0}$. The maximum likelihood estimator is*

$$\widehat{\theta}_n := \arg\max_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^n \log p(x_i; \theta).$$

A classical result is that, under second order or third order smoothness conditions $\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow N(0, \boldsymbol{I}(\theta_0)^{-1})$*, where* $\boldsymbol{I}(\theta_0)$ *is the Fisher information matrix defined as*

See Section 8.3.2 for some details.

$$\boldsymbol{I}(\theta_0) = \mathbb{E}[\nabla \log p_\theta(X) \nabla^\top \log p_\theta(X)]\Big|_{\theta=\theta_0}.$$

What is the minimal smoothness asumption that is needed for the above result to hold? It turns out that if we use UCLT together with the notion of differentiability in quadratic mean (DQM), first-order smoothness will be enough. This is useful in some examples that involve the Laplace density.

Example 4.2.7 (Asymptotics of M-estimators)**.** *The UCLT can also be used to derive asymptotic distributions for M-estimators. We consider two representative examples of location estimators.*

1. *The sample median is defined as*

$$\widehat{\theta}_n = \arg\min_{\theta \in \mathbb{R}} \tfrac{1}{n} \sum_{i=1}^n |X_i - \theta|.$$

Assume the CDF F *is differentiable around its median* θ_0 *with positive derivative* $F'(\theta_0) =: p(\theta_0)$*. Using UCLT, we can prove that*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow N(0, \tfrac{1}{4p^2(\theta_0)}).$$

2. *A form of mode estimator can be defined as*

$$\widehat{\theta}_n = \arg\max \tfrac{1}{n} \sum_{i=1}^n m_\theta(X_i),$$

where $m_\theta(x) = \mathbb{1}(|x - \theta| \le 1)$*. For this estimator, the asymptotic distribution is very complicated as we have*

This is θ that contains the most of X_i's in its unit neighborhood.

$$n^{1/3}(\widehat{\theta}_n - \theta) \rightsquigarrow \arg\max_{h \in \mathbb{R}} \{aZ(h) - bh^2\},$$

where Z *is a standard two-sided Brownian motion starting from* 0*, and*

$$a^2 = p(\theta_0 + 1) - p(\theta_0 - 1) \quad \textit{and} \quad b = \frac{1}{2}(p'(\theta_0 + 1) - p'(\theta_0 - 1)),$$

where p *is the density function, unimodal, and symmetric.*

4.3 Exercises

Exercise 4.3.1. *In the binary classification problem in Section 4.1.1, show that the Bayes classifier* g_0 *given by*

$$g_0(x) = \begin{cases} 1 & \textit{if} \quad \mathbb{P}(Y = 1 | X = x) \ge \frac{1}{2}, \\ -1 & \textit{otherwise} \end{cases}$$

minimizes the misclassification probability $R(g)$ *over all measurable functions* g*.*

Exercise 4.3.2. *Show that the the statistics* (4.8) *does not depend on* F_0. *Hint: First assume that the CDF is continuous and strictly increasing. For the general result use* (3.24).

5
Concentration of measure (3 weeks)

In this chapter we are interested in bounding random fluctuations of functions of many independent random variables. Variables $X, X_1, \ldots, X_n$ are independent and take values in some $\mathcal{X}$. Let $g : \mathcal{X}^n \to \mathbb{R}$ and

$$Z \;=\; g(X_1, \ldots, X_n).$$

The function g can be quite complex. For example, as motivated in the previous chapter, we could have

$$Z \;=\; \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \;=\; \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right|.$$

How large are "typical" deviations of Z from $\mathbb{E}Z$? In particular, we seek upper bounds for

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \qquad \text{and} \qquad \mathbb{P}(Z \leq \mathbb{E}Z - t)$$

for $t > 0$. When $\mathbb{E}Z$ is unknown, to obtain direct bounds on the concentration of Z, explicit bounds on $\mathbb{E}Z$ will also be needed, which will lead to more advanced considerations in Chapter 6.

There are various methods that include: martingales, information theoretic and transportation methods, Talagrand's induction method, and logarithmic Sobolev inequalities. We will not get into too many details here. We refer to two excellent books [1] and [2] for a thorough overview of the theory.

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013

[2] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge, 2019

5.1 *Basic inequalities*

In this section we set up the scene for more advanced considerations by recalling some basic probability inequalities. The simplest inequality is the Markov's inequality.

Proposition 5.1.1 (Markov's inequality)**.** *If $Z \geq 0$ and $t > 0$ then*

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}Z}{t}.$$

← Exercise 5.5.1

Proof. We have

$$t\mathbb{P}(Z \geq t) \;=\; \mathbb{E}[t\mathbb{1}(Z \geq t)] \;\leq\; \mathbb{E}[Z\mathbb{1}(Z \geq t)] \;\leq\; \mathbb{E}Z,$$

where the last inequality follows because $Z \geq 0$. □

Markov's inequality with its elementary proof looks very innocent. It is then surprising to see how many powerful results can be obtained from it. We first show that, for distributions for which the second moment exists, the Markov's inequality implies the Chebyshev's inequality.

Proposition 5.1.2 (Chebyshev's inequality)**.** *For every $t > 0$ we have*

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \;=\; \mathbb{P}((Z - \mathbb{E}Z)^2 \geq t^2) \;\leq\; \frac{\mathrm{var}(Z)}{t^2}.$$

More generally, if the corresponding k-th moment exists for some $k > 0$ then

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq \frac{\mathbb{E}(|Z-\mathbb{E}Z|^k)}{t^k}$$

A particularly important instance is when Z is a sum of i.i.d. random variables, $Z = \sum_{i=1}^n X_i$. Denote $\mu = \mathbb{E}X$, $\sigma^2 = \mathrm{var}(X)$. In this case $\mathrm{var}(Z) = n\sigma^2$ and we get

$$\mathbb{P}(|\sum_{i=1}^n X_i - n\mu| \geq t) \;\leq\; \frac{n\sigma^2}{t^2}$$

or equivalently, denoting $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$,

$$\mathbb{P}(|\sqrt{n}(\bar{X}_n - \mu)| \geq t) \;\leq\; \frac{\sigma^2}{t^2}.$$

This bound is however not very tight. Let $U \sim N(0,1)$ and note that, by the Central Limit Theorem,

$$\lim_{n\to\infty} \mathbb{P}(|\sqrt{n}(\bar{X}_n - \mu)| \geq t) \;=\; 2\mathbb{P}(U \geq \tfrac{t}{\sigma}) \;\leq\; \frac{\sigma}{t}e^{-t^2/(2\sigma^2)},$$

where the last inequality is part of Exercise 5.5.2. In particular, at least for very large n, we expect an exponential decrease in t^2/σ^2.

← Exercise 5.5.2

The trick to get the expected rate of decrease is to use the Markov's inequality in a more clever way.

Proposition 5.1.3 (Chernoff bounds)**.** *Suppose that the moment generating function of Z exists in some neighborhood $(-b, b)$ of zero. Then for every $\lambda > 0$ in this neighborhood*

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \;=\; \mathbb{P}(e^{\lambda(Z-\mathbb{E}Z)} \geq e^{\lambda t}) \;\leq\; \frac{\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}}{e^{\lambda t}}.$$

In consequence,

$$\log \mathbb{P}(Z - \mathbb{E}Z \geq t) \;\leq\; \inf_{\lambda\in[0,b)} \{\log \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} - \lambda t\}.$$

Note that if $M(\lambda) = \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}$ is the moment generating function of $Z - \mathbb{E}Z$ then $K(\lambda) = \log M(\lambda)$ is the cumulant generating function. We get

$$\log \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} - \lambda t \;=\; K(\lambda) - \lambda t.$$

By Hölder's inequality, $K(\lambda)$ is a convex function function with the same proof as in Theorem 1.3.2. We will later discuss some techniques to bound the moment generating function.

The convex conjugate to $K(\lambda)$ is $K^*(t) = \inf_{\lambda\in\mathbb{R}}\{K(\lambda) - \lambda t\}$ and we get $\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq e^{K^*(t)}$.
← Exercise 5.5.3

A useful observation is that, if $X, X_1, X_2, \ldots, X_n$ are i.i.d. then

$$\mathbb{P}(\bar{X}_n - \mu \geq t) \;\leq\; \frac{\mathbb{E}e^{\lambda(\bar{X}_n-\mu)}}{e^{\lambda t}} \;=\; \left(\frac{\mathbb{E}e^{\frac{\lambda}{n}(X-\mathbb{E}X)}}{e^{\frac{\lambda}{n}t}}\right)^n.$$

Optimizing over λ is equivalent to optimizing over λ/n and so a Chernoff bound on a single X, directly gives a Chernoff bound for an average of its independent copies.

Example 5.1.4. *Let $Y, Y_1, \ldots, Y_n$ be an i.i.d. sample such that $Y = X^2$ with $X \sim N(0,1)$. First note that*

$$\mathbb{E}e^{\lambda Y} \;=\; \begin{cases} \sqrt{\frac{1}{1-2\lambda}} & \text{for } \lambda < \frac{1}{2}, \\ +\infty & \text{otherwise.} \end{cases}$$

Thus, for $\lambda < 1/2$,

$$\mathbb{P}(Y - 1 \geq t) \;\leq\; \sqrt{\frac{1}{1-2\lambda}}\frac{1}{e^{\lambda(t+1)}}.$$

The optimal $\lambda^* = \frac{1}{2}\frac{t}{t+1}$, which gives the Chernoff bound[3]

$$\mathbb{P}(Y - 1 \geq t) \;\leq\; \sqrt{t+1}e^{-t/2}$$

[3] Plot this to see that this bound is actually not that great for $t \in (0, 5)$.

and consequently

$$\mathbb{P}(\bar{Y}_n - 1 \geq t) \;\leq\; (t+1)^{n/2}e^{-nt/2}.$$

It is convenient to have an explicit exponential bound for small t. In this case we can use the fact that $\log(1+t) - t \leq -t^2/4$ for $t \in [0,1]$[4] *to conclude that, for every $t \in [0,1]$,*

$$\mathbb{P}(\bar{Y}_n - 1 \geq t) \;\leq\; (t+1)^{n/2}e^{-nt/2} \;=\; e^{\frac{n}{2}(\log(t+1)-t)} \;\leq\; e^{-nt^2/8}.$$

[4] To show that $-t^2/4 - \log(1+t) + t \geq 0$ for $t \in [0,1]$ we note that the left hand side is zero for $t = 0$ and its derivative is nonnegative for $t \in [0,1]$.

← Exercise 5.5.4

A simple yet powerful illustration of the Chernoff bounds in Proposition 5.1.3 is given by following example.

Example 5.1.5 (Johnson-Lindenstrauss lemma)**.** *Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^D$ and $\epsilon > 0$. We are looking for a function $f : \mathbb{R}^D \to \mathbb{R}^d$ with $d < D$ such that*

$$(1-\epsilon)\|\boldsymbol{a}_i - \boldsymbol{a}_j\|^2 \;<\; \|f(\boldsymbol{a}_i) - f(\boldsymbol{a}_j)\|^2 \;<\; (1+\epsilon)\|\boldsymbol{a}_i - \boldsymbol{a}_j\|^2 \qquad (5.1)$$

for all $i, j = 1, \ldots, n$. Johnson-Lindenstrauss lemma states such an embedding f exists if

$$d \;>\; \frac{16}{\epsilon^2} \log n. \tag{5.2}$$

(Note that the bound does not depend on D!) We present a probabilistic proof of this fact and show that f can be taken as a linear function. In fact, if we choose f at random, it works with positive probability. Let $W \in \mathbb{R}^{d \times D}$ be a random matrix such that the W_{ij} are independent $N(0, 1/d)$.

Take $f(\boldsymbol{a}) = W\boldsymbol{a}$. Then, denoting $\boldsymbol{b}_{ij} := (\boldsymbol{a}_i - \boldsymbol{a}_j)/\|\boldsymbol{a}_i - \boldsymbol{a}_j\|$, we can rewrite (5.1) as

$$1 - \epsilon \;<\; \|W\boldsymbol{b}_{ij}\|^2 < 1 + \epsilon.$$

or equivalently

$$\max_{i,j} \Big| \|W\boldsymbol{b}_{ij}\|^2 - 1 \Big| \;<\; \epsilon. \tag{5.3}$$

We first show that $\mathbb{E}\|W\boldsymbol{b}_{ij}\|^2 = 1$. Indeed, for any $\boldsymbol{b} \in \mathbb{R}^D$,

$$\mathbb{E}\|W\boldsymbol{b}\|^2 \;=\; \mathbb{E}(\boldsymbol{b}^\intercal W^\intercal W \boldsymbol{b}) \;=\; \boldsymbol{b}^\intercal \mathbb{E}(W^\intercal W)\boldsymbol{b} \;=\; \boldsymbol{b}^\intercal \boldsymbol{b} \;=\; \|\boldsymbol{b}\|^2. \tag{5.4}$$

Since $\|\boldsymbol{b}_{ij}\| \;=\; 1$ we get $\mathbb{E}\|W\boldsymbol{b}_{ij}\|^2 \;=\; 1$. To show (5.3) with positive probability, it is equivalent to show that

$$\mathbb{P}(\max_{i,j} \Big| \|W\boldsymbol{b}_{ij}\|^2 - 1 \Big| \geq \epsilon) < 1.$$

We next show that $\|W\boldsymbol{b}_{ij}\|^2$ can be written as $\frac{1}{d}\sum_{i=1}^d Z_i^2$ with $Z_i \;\sim$ $N(0,1)$. Indeed, for every $\boldsymbol{b} \in \mathbb{R}^D$ such that $\|\boldsymbol{b}\| = 1$

$$\|W\boldsymbol{b}\|^2 \;=\; (W\boldsymbol{b})^\intercal (W\boldsymbol{b}) \;=\; \sum_{i=1}^d (\boldsymbol{w}_i^\intercal \boldsymbol{b})^2,$$

where $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d$ are the rows of W. Note that $\mathbb{E}(\boldsymbol{w}_i^\intercal \boldsymbol{b}) = 0$ and

$$\mathbb{E}((\boldsymbol{w}_i^\intercal \boldsymbol{b})^2) = \boldsymbol{b}^\top \mathbb{E}(\boldsymbol{w}_i \boldsymbol{w}_i)\boldsymbol{b} \;=\; \frac{1}{d}\boldsymbol{b}^\top \boldsymbol{b} \;=\; \frac{1}{d}.$$

Thus, $\boldsymbol{w}_i^\intercal \boldsymbol{b}$ for $i = 1, \ldots, d$ are i.i.d. $N(0, 1/d)$, we get

$$\|W\boldsymbol{b}\|^2 - 1 \;=\; \frac{1}{d}\sum_{i=1}^d (Z_i^2 - 1),$$

where Z_i are i.i.d. $N(0,1)$. By the calculations in Example 5.1.4 and by (5.10), for every $\epsilon \leq 1$ we conclude that

$$\mathbb{P}(|\|W\boldsymbol{b}\|^2 - 1| \geq \epsilon) \;\leq\; 2e^{-\epsilon^2 d/8}.$$

By the union bound

$$\mathbb{P}(\max_{i,j} |\|W\boldsymbol{b}_{ij}\|^2 - 1| > \epsilon) \leq \binom{n}{2} \mathbb{P}(|\|W\boldsymbol{b}\|^2 - 1| > \epsilon) \leq \binom{n}{2} 2e^{-\epsilon^2 d/8}.$$

It remains to check that if (5.2) *holds then the right hand side is smaller than 1. But this is clear. Rewrite* (5.2) *as* $d\epsilon^2/8 > \log n^2$ *and get*

$$\binom{n}{2} 2e^{-\epsilon^2 d/8} \;=\; e^{\log(n(n-1))-\epsilon^2 d/8} < e^{\log(n(n-1))-\log n^2} = \frac{n-1}{n} < 1.$$

Note that the proof is not constructive, as the linear map satisfying (5.1) *is not given explicitly.*

5.2 *Sub-gaussian and sub-exponential random variables*

It is clear that if the values of Z concentrate, its variance must be small. To establish concentration using the Chebyshev's inequality in Proposition 5.1.2 it is important to obtain good bounds on the variance of Z. Similarly, in order to utilize the Chernoff bounds in Proposition 5.1.3 we need to obtain good bounds for the moment generating function. In this and in the next section we will focus on this last task.

To motivate the next definition note that if $X \sim N(\mu, \sigma^2)$ then

$$\mathbb{E}e^{\lambda(X-\mu)} \;=\; e^{\sigma^2\lambda^2/2} \qquad \text{for all } \lambda \in \mathbb{R}.$$

Substituting this into the Chernoff bound in Proposition 5.1.3 we get

$$\inf_{\lambda \geq 0} \left\{ \log \mathbb{E}e^{\lambda(X-\mu)} - \lambda t \right\} \;=\; \inf_{\lambda \geq 0} \left\{ \frac{\lambda^2\sigma^2}{2} - \lambda t \right\} \;=\; -\frac{t^2}{2\sigma^2},$$

which gives

$$\mathbb{P}(X - \mu \geq t) \;\leq\; e^{-t^2/2\sigma^2} \qquad \text{for all } t \geq 0. \tag{5.5}$$

In other words, for $\delta \in (0,1)$,

$$\mathbb{P}\left(X - \mu \geq \sigma\sqrt{2\log(\tfrac{1}{\delta})} \right) \leq \delta.$$

The same concentration bounds hold in much greater generality for so called sub-Gaussian variables.

← Exercise 5.5.5

Definition 5.2.1. *A variable X is sub-Gaussian (or σ-sub-Gaussian) if there is $\sigma > 0$ such that*

$$\mathbb{E}e^{\lambda(X-\mu)} \;\leq\; e^{\sigma^2\lambda^2/2} \qquad \textit{for all } \lambda \in \mathbb{R}$$

or equivalently

$$K_X(\lambda) - \lambda\mu \;\leq\; \sigma^2\lambda^2/2 \qquad \textit{for all } \lambda \in \mathbb{R}, \tag{5.6}$$

where K_X is the cumulant generating function of X.

The following proposition shows that, if X takes values in a bounded interval $[a, b]$, then X is sub-Gaussian with $\sigma = (b-a)/2$.

Proposition 5.2.2. *Suppose that X has values in some bounded interval $[a, b]$. Then $\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq e^{\lambda^2(b-a)^2/8}$, that is, X is sub-Gaussian with $\sigma = (b-a)/2$.*

In the proof we will use the following simple result.

Lemma 5.2.3 (Popoviciu's inequality)**.** *If $X \in [a, b]$ then* $\operatorname{var}(X) \leq \left(\frac{b-a}{2}\right)^2$.

Proof. For every $u \in \mathbb{R}$, $\mathbb{E}(X-u)^2 = \operatorname{var}(X) + (u-\mu)^2$ and so $\operatorname{var}(X) \leq \mathbb{E}(X-u)^2$. Take $u = \frac{a+b}{2}$ then

$$\mathbb{E}(X-u)^2 = \frac{1}{4}\mathbb{E}(X-a+X-b)^2 \leq \frac{1}{4}\mathbb{E}(X-a+b-X)^2 = \left(\frac{b-a}{2}\right)^2.$$

□

Proof of Proposition 5.2.2. Suppose f is the density of X with respect to the underlying measure μ. Consider the cumulant generating function $K(\lambda) = \log \mathbb{E}e^{\lambda X}$ and note that $K(0) = 0$, $K'(0) = \mathbb{E}X$, and $K''(0) = \operatorname{var}(X)$. Directly by definition of $K(\lambda)$

$$f(x;\lambda) = f(x)e^{\lambda x - K(\lambda)}$$

defines a valid density function. Direct calculations (or Proposition 1.3.4) show that the second derivative of $K(\lambda)$ is the variance of the distribution with density $e^{\lambda x - K(\lambda)}f(x)$[5]. In particular, $K''(\lambda) = \mathbb{E}_\lambda[X^2] - (\mathbb{E}_\lambda[X])^2$, where $\mathbb{E}_\lambda[g(X)] = \mathbb{E}[g(X)\frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}] = \int_{\mathbb{R}} g(x)e^{\lambda x - K(\lambda)}f(x)\mathrm{d}x$. By Lemma 5.2.3,

$$0 \leq K''(\lambda) = \mathbb{E}_\lambda[X^2] - (\mathbb{E}_\lambda[X])^2 \leq \left(\frac{b-a}{2}\right)^2.$$

By the Taylor's theorem, for every $\lambda \in \mathbb{R}$

$$K(\lambda) = K(0) + K'(0)\lambda + K''(\theta)\frac{\lambda^2}{2}$$

for some θ between 0 and λ. Using what we know, we conclude that

$$K(\lambda) - \lambda\mathbb{E}X \leq \left(\frac{b-a}{2}\right)^2 \frac{\lambda^2}{2} \qquad \text{for all } \lambda \in \mathbb{R},$$

which is equivalent to the claimed inequality. □

[5] Note that λ is a parameter describing an exponential family where f corresponds to $\lambda = 0$.

← Exercise 5.5.6

The following elementary result will be useful.

Lemma 5.2.4. *Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random vector such that X_i is σ_i-sub-Gaussian for $i = 1, \ldots, n$. If X_i are independent and $\boldsymbol{u} \in \mathbb{R}^n$, then the random variable $\boldsymbol{u}^\intercal \boldsymbol{X}$ is sub-Gaussian with parameter $\sigma = \sqrt{\sigma_1^2 u_1^2 + \ldots + \sigma_n^2 u_n^2}$.*

Proof. Denote by $\boldsymbol{\mu}$ the mean vector of $\mathbb{E}\boldsymbol{X}$. Equivalently we must show $K_{\boldsymbol{u}^\intercal \boldsymbol{X}}(\lambda) - \lambda \boldsymbol{u}^\intercal \boldsymbol{\mu} \leq \sigma^2\lambda^2/2$. We have

$$K_{\boldsymbol{u}^\intercal \boldsymbol{X}}(\lambda) - \lambda \boldsymbol{u}^\intercal \boldsymbol{\mu} = \sum_{i=1}^n \left(K_{X_i}(\lambda u_i) - \lambda u_i \mu_i\right) \leq \sum_{i=1}^n \frac{\sigma_i^2 \lambda^2 u_i^2}{2} = (\sum_{i=1}^n \sigma_i^2 u_i^2)\frac{\lambda^2}{2},$$

which completes the proof. □

We list a couple of useful observations that follow directly from Lemma 5.2.4. They show that the σ parameter behaves similarly to the standard deviation.

Corollary 5.2.5. *For a collection of independent σ-sub-Gaussian random variables $X_1, \ldots, X_n$, their average $\bar{X}_n$ is sub-Gaussian with parameter $\sigma/\sqrt{n}$.*

Corollary 5.2.6. *If X is σ-sub-Gaussian, then $aX + b$ ($a \in \mathbb{R}, b \in \mathbb{R}$) is sub-Gaussian with parameter $|a|\sigma$. In particular, $-X$ is σ-sub-Gaussian.*

Corollary 5.2.7. *If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a vector of independent random variables such that each X_i is σ-sub-Gaussian, and $\|\boldsymbol{u}\| = 1$, then $\boldsymbol{u}^T\boldsymbol{X}$ is σ-sub-Gaussian.*

Suppose that X is σ-sub-Gaussian. By Corollary 5.2.6, $-X$ is also σ-sub-Gaussian. Thus, on the top of (5.5) we also have

$$\mathbb{P}(X - \mu \leq -t) \;\leq\; e^{-t^2/2\sigma^2} \qquad \text{for all } t \geq 0.$$

Using the union bound, we conclude the following result.

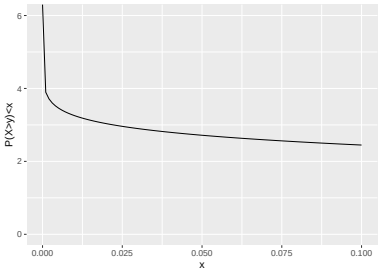
Corollary 5.2.8. *For a σ-sub-Gaussian variable*

$$\mathbb{P}(|X - \mu| \geq t) \;\leq\; 2e^{-t^2/2\sigma^2} \qquad \textit{for all } t \geq 0.$$

In other words, for $\delta \in (0,1)$,

$$\mathbb{P}\left(|X - \mu| \geq \sigma\sqrt{2\log(\tfrac{2}{\delta})}\right) \leq \delta.$$

The graph below shows the price we are paying by decreasing $\delta \in (0, 0.1)$ (the x-axis) with respect to the guaranteed bound (y-axis); $\sigma = 1$.



The following result applies Chernoff bounds to a sequence of sub-Gaussian variables.

Proposition 5.2.9 (Hoeffding inequality)**.** *If X_i are independent σ_i-sub-Gaussian then for all $t \geq 0$ we have*

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| \geq t\right) \;\leq\; 2e^{-t^2n^2/(2\sum_i \sigma_i^2)}$$

Proof. By Lemma 5.2.4, $\bar{X}_n$ is sub-Gaussian with parameter

$$\sigma = \frac{1}{n}\sqrt{\sum_{i=1}^n \sigma_i^2}.$$

Thus, the result follows by Corollary 5.2.8. □

Remark 5.2.10. *If $\sigma_i = \sigma$ for all i we also note that the Hoeffding inequality implies*

$$\mathbb{P}\left(\left|\sqrt{n}(\bar{X}_n - \mathbb{E}(\bar{X}_n))\right| \geq t\right) \;\leq\; 2e^{-t^2/(2\sigma^2)}.$$

Example 5.2.11 (Sub-Gaussian sequence model)**.** *Consider independent $Y_1, \dots, Y_n$ that are σ-sub-Gaussian with means μ_i, where σ^2 is known. Let $S \subset \{1, \dots, n\}$ be the support of the vector $\mu = (\mu_1, \dots, \mu_n)$, that is, $\mu_i \neq 0$ for $i \in S$ and $\mu_i = 0$ for $i \notin S$. One possible way of finding the zero μ_i's is by considering the LASSO estimator*

$$\hat{\mu} \;=\; \arg\min_{\mu} \left\{ \tfrac{1}{2}\|\mathbf{Y} - \mu\|^2 + \lambda \sum_{i=1}^{n} |\mu_i| \right\}$$

for a fixed $\lambda \geq 0$. It is a simple exercise[6] *to show that*

$$\hat{\mu}_i \;=\; \begin{cases} Y_i - \lambda & \text{if } Y_i > \lambda, \\ 0 & \text{if } |Y_i| \leq \lambda, \\ Y_i + \lambda & \text{if } Y_i < -\lambda. \end{cases}.$$

[6] Check this!
In fact, our analysis applies to any procedure based on thresholding.

To analyse this procedure we consider two types of errors

1. *Type I: $|Y_i| > \lambda$ for $i \notin S$, and*
2. *Type II: $|Y_i| \leq \lambda$ for $i \in S$.*

We would like to set the threshold λ to control the Family-wise Error Rate. In other words, $Z := \max_{i \notin S} |Y_i| \leq \lambda$ with high probability. Let s be the number of the nonzeros entries in μ then $m_0 := n - s$ is the number of zeros. By the union bound and using the fact that each Y_i for $i \notin S$ is mean zero σ-sub-Gaussian

$$\mathbb{P}(Z \geq \lambda) \leq \sum_{i \notin S} \mathbb{P}(|Y_i| \geq \lambda) \leq 2m_0 e^{-\lambda^2/2\sigma^2}.$$

Suppose that $\lambda = \sigma\sqrt{2\log(\frac{2m_0}{\alpha})}$ for some $\alpha \in (0,1)$ then $\mathbb{P}(Z \geq \lambda) \leq \alpha$, which binds the family-wise error rate. Moreover, if $\lambda_n = 2\sigma\sqrt{\log(m_0)}$, then

$$\mathbb{P}(Z \geq \lambda_n) \leq \frac{2}{m_0},$$

which goes to zero as m_0 grows to infinity (with n)[7]*. Of course, in practice, we also want λ to be as small as possible to reduce the type II error. We will not discuss this issue here*[8]*.*

[7] We could compare this procedure with the Bonferroni and the Holm procedures discussed earlier. Note that the latter two have no theoretical guarantees if the underlying distribution is not Gaussian.
[8] Exercise: Analyze the type II error as a function of $\mu^* = \min_{i \in S} |\mu_i|$

Another general type of bound on the moment generating function is the following.

Definition 5.2.12. *A random variable with mean $\mu = \mathbb{E}X$ is sub-exponential if there are non-negative parameters ν, α such that*

$$\mathbb{E}(e^{\lambda(X-\mu)}) \;\leq\; e^{\nu^2\lambda^2/2} \qquad \text{for all } |\lambda| < \frac{1}{\alpha}.$$

This condition is rather mild and it is essentially equivalent with the existence of the cumulant generating function in a neighbourhood of zero.

Example 5.2.13. *Let $Z \sim N(0,1)$ and let $X = Z^2$. For $\lambda < 1/2$, we have*

$$\mathbb{E}(e^{\lambda(X-1)}) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\lambda(z^2-1)} e^{-z^2/2} dz = \begin{cases} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} & \text{if } \lambda < \frac{1}{2}, \\ +\infty & \text{if } \lambda \geq \frac{1}{2}. \end{cases}$$

In particular, since the moment generating function is not defined everywhere, X is not sub-Gaussian. With a bit of calculus we see however that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{4\lambda^2/2} \qquad \text{for all } |\lambda| < \frac{1}{4}$$

and so X is sub-exponential with parameters $(\nu, \alpha) = (2, 4)$.

With essentially the same proof, Lemma 5.2.4 generalizes to sub-exponential variables.

Lemma 5.2.14. *Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random vector such that X_i is (ν_i, α_i)-sub-exponential for $i = 1, \ldots, n$. If X_i are independent and $\boldsymbol{u} \in \mathbb{R}^n$, then the random variable $\boldsymbol{u}^\intercal \boldsymbol{X}$ is sub-exponential with parameters $\nu = \sqrt{\nu_1^2 u_1^2 + \ldots + \nu_n^2 u_n^2}$ and $\alpha = \max_i |u_i| \alpha_i$.*

We obtain simple concentration inequalities for sub-exponential variables.

Proposition 5.2.15. *Suppose that X is sub-exponential with parameters (ν, α) and $t \geq 0$. Then*

$$\mathbb{P}(X - \mu \geq t) \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{\alpha}, \\ e^{-\frac{t}{2\alpha}} & \text{if } t > \frac{\nu^2}{\alpha}. \end{cases}$$

and

$$\mathbb{P}(|X - \mu| \geq t) \leq \begin{cases} 2e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{\alpha}, \\ 2e^{-\frac{t}{2\alpha}} & \text{if } t > \frac{\nu^2}{\alpha}. \end{cases}$$

Proof. Using the Chernoff bound we know that

$$\mathbb{P}(X - \mu \geq t) \leq \inf_{\lambda \in [0, \frac{1}{\alpha})} \{\mathbb{E} e^{\lambda(X - \mathbb{E}X)} e^{-\lambda t}\} \leq \inf_{\lambda \in [0, \frac{1}{\alpha})} \{e^{\nu^2\lambda^2/2 - \lambda t}\}.$$

The global optimum of $e^{\nu^2\lambda^2/2 - \lambda t}$ is $\lambda^* = t/\nu^2 \geq 0$. Thus, if $t/\nu^2 < 1/\alpha$ (equiv. $\nu^2 > t\alpha$), we get the sub-Gaussian bound $e^{-t^2/(2\nu^2)}$. Otherwise, if $\nu^2 \leq t\alpha$, the infimum is obtained on the boundary $\lambda^* = 1/\alpha$ and it is equal to $e^{\nu^2/(2\alpha^2) - t/\alpha}$. Note however that the fact that $\nu^2 \leq t\alpha$, allows to bound this by $e^{-t/(2\alpha)}$. □

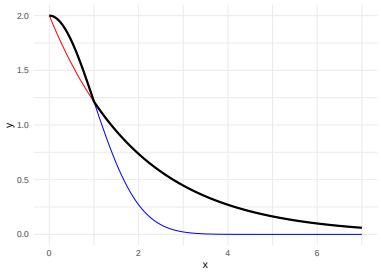


Figure 5.1: Illustration of the subexponential bounds for $\nu = \alpha = 1$. The blue line represents $2e^{-t^2/2}$. The red line represents $2e^{-t/2}$. In black their pointwise maximum.

Lemma 5.2.14 and Proposition 5.2.15 give now a handful of useful results. For example, if X_i are independent sub-exponential with parameters (ν, α) then $\bar{X}_n$ is sub-exponential with parameters $(\frac{\nu}{\sqrt{n}}, \frac{\alpha}{n})$ and so

$$\mathbb{P}(|\bar{X}_n - \mu| \geq t) \leq \begin{cases} 2e^{-\frac{t^2 n}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{\alpha}, \\ 2e^{-\frac{tn}{2\alpha}} & \text{if } t > \frac{\nu^2}{\alpha}. \end{cases}$$

5.3 *Martingale-based methods*

We will now briefly discuss techniques related to martingale representations. Recall the basic set-up of the martingale theory in Appendix C.2. Let $X_1, \ldots, X_n$ be independent random variables and $Z = f(X_1, \ldots, X_n)$. Denote

$$\mathbb{E}_i[\cdot] \;=\; \mathbb{E}[\cdot | X_1, \ldots, X_i].$$

Thus $\mathbb{E}_0 Z = \mathbb{E}Z$ and $\mathbb{E}_n Z = Z$. Moreover, if $i < j$ then $\mathbb{E}_i \mathbb{E}_j Z = \mathbb{E}_j \mathbb{E}_i Z = \mathbb{E}_i Z$. It is easy to see that $Y_k = \mathbb{E}_k Z$ forms a martingale sequence if $\mathbb{E}|Z| < \infty$. Indeed, by Jensen's inequality

$$\mathbb{E}(|\mathbb{E}_k Z|) \leq \mathbb{E}\mathbb{E}_k |Z| = \mathbb{E}|Z| < \infty.$$

This is called the Doob's martingale.

Writing

$$\Delta_i \;=\; \mathbb{E}_i Z - \mathbb{E}_{i-1} Z$$

we have

$$Z - \mathbb{E}Z \;=\; \sum_{i=1}^{n} \Delta_i. \tag{5.7}$$

This is the Doob martingale representation of Z. In particular, we write $Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i$, where Δ_i for the corresponding Doob martingale difference sequence.

Theorem 5.3.1. *Let $\{\Delta_i, \mathcal{F}_i\}$ be a martingale difference sequence as above and suppose that $\mathbb{E}_{k-1}[e^{\lambda \Delta_k}] \leq e^{\lambda^2 \nu_k^2 / 2}$ almost surely for any $|\lambda| < 1/\alpha_k$ (a form of a sub-exponential condition). Then $Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i$ is subexponential with parameters $(\nu_*, \alpha_*) = (\|\boldsymbol{\nu}\|, \|\boldsymbol{\alpha}\|_\infty)$. In particular,*

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq \begin{cases} 2e^{-\frac{t^2}{2\nu_*^2}} & \text{if } 0 \leq t \leq \frac{\nu_*^2}{\alpha_*}, \\ 2e^{-\frac{t}{2\alpha_*}} & \text{if } t > \frac{\nu_*^2}{\alpha_*}. \end{cases}$$

Proof. The first part can be directly shown by recursive conditioning. Indeed, we first write

$$\mathbb{E}e^{\lambda(Z - \mathbb{E}Z)} = \mathbb{E}e^{\lambda \sum_{i=1}^n \Delta_i} = \mathbb{E}\mathbb{E}_{n-1} e^{\lambda \sum_{i=1}^n \Delta_i} = \mathbb{E}e^{\lambda \sum_{i=1}^{n-1} \Delta_i} \mathbb{E}_{n-1} e^{\lambda \Delta_n}$$

and this is bounded for all $|\lambda| < 1/\alpha_n$ by

$$e^{\lambda^2 \nu_n^2 / 2} \mathbb{E}e^{\lambda \sum_{i=1}^{n-1} \Delta_i}.$$

Proceeding in a similar fashion, we get the conclusion. The second of the theorem part follows by the first part and Proposition 5.2.15. □

A very useful version of this result is when each Δ_i is bounded, $\Delta_k \in [a_k, b_k]$, in which case it is also sub-Gaussian (also conditionally on $\mathcal{F}_{k-1}$).

Definition 5.3.2. *A function $f : \mathcal{X}^n \to \mathbb{R}$ satisfies **bounded differences inequality**, if there exist constants $L_1, \dots, L_n$ such that*

$$|f(x) - f(x')| \le L_k,$$

whenever x, x' differ only in the k-th coordinate.

A canonical example of such a function is

$$f(x) \;=\; \sum_{i=1}^{n} f_i(x_i),$$

where all f_i are bounded functions.

Proposition 5.3.3 (Bounded differences inequality)**.** *Suppose that $Z = f(X_1, \dots, X_n)$, where f satisfies the bounded differences inequality with parameters $L_1, \dots, L_n$ and such that the random vector $X = (X_1, \dots, X_n)$ has independent components. Then Z is sub-Gaussian with parameter*

$$\sigma = \tfrac{1}{2}\sqrt{L_1^2 + \dots + L_n^2}.$$

In particular, for all $t \ge 0$

$$\mathbb{P}(|Z - \mathbb{E}Z| \ge t) \;\le\; 2e^{-\frac{t^2}{2\sigma^2}} \;=\; 2e^{-\frac{2t^2}{\sum_{i=1}^n L_i^2}}.$$

Proof. Fix i and take a look at Δ_i. Let

$$a_i := \inf_x \mathbb{E}[Z|X_1, \dots, X_{i-1}, x] - \mathbb{E}_{i-1}Z$$

and

$$b_i := \sup_x \mathbb{E}[Z|X_1, \dots, X_{i-1}, x] - \mathbb{E}_{i-1}Z.$$

It is clear that $a_i \le \Delta_i \le b_i$ almost surely. To use Proposition 5.2.2 we need to bound $b_i - a_i$. For that, note first that

$$\mathbb{E}[f(X_1, \dots, X_i, \dots, X_n)|X_1, \dots, X_{i-1}, X_i = x] = \mathbb{E}[f(X_1, \dots, x, \dots, X_n)|X_1, \dots, X_{i-1}],$$

which follows by independence of all X_i's (in both cases the integral is with respect to the marginal distribution of $(X_{i+1}, \dots, X_n)$).

We have

$$\begin{aligned}
b_i - a_i &= \sup_x \mathbb{E}[Z|X_1, \dots, X_{i-1}, x] - \inf_x \mathbb{E}[Z|X_1, \dots, X_{i-1}, x] \\
&\le \sup_{x,y} \Big|\mathbb{E}[Z|X_1, \dots, X_{i-1}, x] - \mathbb{E}[Z|X_1, \dots, X_{i-1}, y]\Big| \\
&= \sup_{x,y} \Big|\mathbb{E}_{i-1}[f(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n) - f(X_1, \dots, X_{i-1}, y, X_{i+1}, \dots, X_n)]\Big| \\
&\le L_i.
\end{aligned}$$

Since $\Delta_i \in [a_i, b_i]$ and $b_i - a_i \le L_i$, by Proposition 5.2.2,

$$\mathbb{E}_{i-1}e^{\lambda\Delta_i} \;\le\; e^{\lambda^2\nu_i^2/2}, \quad \text{for } \nu_i = L_i/2 \text{ and all } \lambda \in \mathbb{R}. \tag{5.8}$$

Now we can use Theorem 5.3.1 to conclude that $Z - \mathbb{E}Z$ (and so also Z) is sub-Gaussian with parameter $\frac{1}{2}\sqrt{L_1^2 + \ldots + L_n^2}$. The probability bound then follows from (5.5). □

Remark 5.3.4. *Note that there was an alternative route to simply observe that $|f(x) - f(y)| \leq \sum_k L_k$ for all $x, y \in \mathcal{X}^n$. In other words, Z is bounded and so sub-Gaussian with parameter $\sigma = \sum_k L_k$ even without assuming independence of $X_1, \ldots, X_n$. However, $\frac{1}{2}\sqrt{\sum_k L_k^2} < \sum_k L_k$ and this difference can be critical.*

Example 5.3.5 (Kernel density estimation)**.** *Let $X_1, \ldots, X_n$ be i.i.d. real samples drawn according to some density ϕ. The kernel density estimate is*

$$\phi_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $h > 0$, and K is a nonnegative "kernel" $\int K = 1$. The L_1-error is

$$Z = f(X_1, \ldots, X_n) = \int_{\mathcal{X}} |\phi_n(x) - \phi(x)| \mathrm{d}x.$$

It is easy to see that, if X, X' differ only in the i-th coordinate then

$$|f(X) - f(X')| \leq \frac{1}{nh}\int \left| K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x - X_i'}{h}\right)\right| \mathrm{d}x \leq \frac{2}{n}.$$

By Proposition 5.3.3,

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2e^{-\frac{t^2 n}{2}} \qquad \text{for all } t \geq 0.$$

Note however that to analyze the quality of the kernel density estimator, we need bounds on Z directly. For that we need to separately study $\mathbb{E}Z$. [9]

[9] This is well studies in the literature. See for example Section II.4.2 in .

P. P. B. Eggermont and V. N. LaRiccia. *Maximum penalized likelihood estimation. Vol. I.* Springer Series in Statistics. Springer-Verlag, New York, 2001

Example 5.3.6 (Uniform deviations)**.** *Let $\mathcal{A}$ be a collection of measurable subsets of $\mathcal{X}$ and let $X, X_1, \ldots, X_n$ be random points drawn from $\mathcal{X}$ i.i.d. Let*

$$\mathbb{P}(A) = \mathbb{P}\{X \in A\} \qquad \text{and} \qquad \mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}.$$

Let $Z = f(X_1, \ldots, X_n) = \sup_{A \in \mathcal{A}} |\mathbb{P}(A) - \mathbb{P}_n(A)|$. If X, X' differ only in the i-th coordinate then, for every $A \in \mathcal{A}$,

$$\left||\mathbb{P}(A) - \mathbb{P}_n(A)| - |\mathbb{P}(A) - \mathbb{P}_n'(A)|\right| \leq |\mathbb{P}_n(A) - \mathbb{P}_n'(A)| = \frac{1}{n}|\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X_i' \in A\}| \leq \frac{1}{n}.$$

We conclude that $|f(X) - f(X')| \leq 1/n$. By Proposition 5.3.3,

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2e^{-2t^2 n} \qquad \text{for all } t \geq 0.$$

This holds irrespective of the underlying distribution or richness of $\mathcal{A}$. As in the previous example, a major issue can be to understand better $\mathbb{E}Z$.

Example 5.3.7 (U-statistics). *Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function. Given an i.i.d. sequence X_i of random variables, the quantity*

$$Z := \frac{1}{\binom{n}{2}} \sum_{i<j} g(X_i, X_j).$$

is known as a pairwise U-statistic. For instance, if $g(s,t) = |s-t|$, then U is an unbiased estimator of the mean absolute pairwise deviation $\mathbb{E}(|X_1 - X_2|)$. Suppose g is bounded, say $\|g\|_\infty \le b$. Writing $Z =: f(X_1, \dots, X_n)$. For any two $x, x' \in \mathbb{R}^n$ that differ only in the i-th coordinate we have

$$|f(x) - f(x')| \le \frac{1}{\binom{n}{2}} \sum_{j \neq i} |g(x_i, x_j) - g(x_i', x_j)| \le \frac{(n-1)2b}{\binom{n}{2}} = \frac{4b}{n}.$$

By Proposition 5.3.3,

$$\mathbb{P}(|Z - \mathbb{E}Z| \ge t) \le 2e^{-\frac{t^2 n}{8b^2}} \qquad \text{for all } t \ge 0.$$

In particular, Z is a consistent estimator of $\mathbb{E}Z = \mathbb{E}g(X_1, X_2)$.

The following example will become important later in the lecture. This is also our first example for which the constants $L_1, \dots, L_n$ may be different.

Example 5.3.8 (Rademacher complexity). *Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ be a vector of independent Rademacher random variables (values ± 1 with equal probability). Given a collection of vectors $\mathcal{A} \subseteq \mathbb{R}^n$, define the random variable*

$$Z(\mathcal{A}) := \sup_{a \in \mathcal{A}} \langle a, \varepsilon \rangle.$$

Although we do not use it here, note that $Z(\mathcal{A})$ is a convex function of ε.

*The random variable $Z = Z(\mathcal{A})$ measures the size of $\mathcal{A}$ in a certain sense, and its expectation $\mathcal{R}(\mathcal{A})$ is known as the **Rademacher complexity** of the set $\mathcal{A}$. Suppose $\mathcal{A}$ is bounded. We will use Proposition 5.3.3 to show that $Z = f(\varepsilon_1, \dots, \varepsilon_n)$ is sub-Gaussian. Note that for every $a \in \mathcal{A}$, if $\varepsilon, \varepsilon'$ differ only in the i-th coordinate, then*

$$\langle a, \varepsilon \rangle - \sup_{b \in \mathcal{A}} \langle b, \varepsilon' \rangle = \inf_{b \in \mathcal{A}} (\langle a, \varepsilon \rangle - \langle b, \varepsilon' \rangle) \le \langle a, \varepsilon \rangle - \langle a, \varepsilon' \rangle = a_i(\varepsilon_i - \varepsilon_i') \le 2|a_i|.$$

Thus, we can take the supremum over $a \in \mathcal{A}$ to conclude that

$$f(\varepsilon) - f(\varepsilon') = \sup_{a \in \mathcal{A}} \langle a, \varepsilon \rangle - \sup_{b \in \mathcal{A}} \langle b, \varepsilon' \rangle \le 2 \sup_{a \in \mathcal{A}} |a_i| =: L_i.$$

The argument is symmetric if we swap ε and ε' and so f satisfies the bounded differences inequality. By Proposition 5.3.3, $Z(\mathcal{A})$ is then sub-Gaussian with parameter $2\sqrt{\sum_i \sup_{a \in \mathcal{A}} a_i^2}$. We remark that, using different techniques, this sub-Gaussianity parameter can be reduced to $\sup_{a \in \mathcal{A}} \|a\|$.

An important application of this theory is for the variable $Z = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$, as defined in (4.1), at least in the case when $\mathcal{F}$ is uniformly bounded.

Theorem 5.3.9. *Assume $\mathcal{F}$ is uniformly b-bounded, that is, $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$. We have*

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp\left\{-\frac{nt^2}{2b^2}\right\}.$$

Proof. The function $g(x_1, \ldots, x_n) = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f(X)\right|$ satisfies the bounded difference property with $L_i = \frac{2b}{n}$. □

5.4 *Lipschitz functions of Gaussian variables*

Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is **Lipschitz with parameter** L if

$$|f(x) - f(y)| \leq L\|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^n.$$

Theorem 5.4.1. *Let $X = (X_1, \ldots, X_n) \sim N_n(0, I_n)$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz with parameter L. Then the variable $f(X)$ is L-sub-Gaussian, and hence*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2e^{-\frac{t^2}{2L^2}} \qquad \textit{for all } t \geq 0.$$

Note that the sub-Gaussian parameter is completely dimension-free! For this reason this result is sometimes referred to as "Dimension Free Concentration Inequality".

We will prove this result with a slightly worse constant and assuming differentiability (Lipschitz functions are differentiable almost everywhere though). In this case $\|\nabla f(x)\| \leq L$ for all x (easy exercise, e.g. consider the directional derivative of $f(x)$ in the direction $\nabla f(x)$).

Proof. We have $f(X) - \mathbb{E}f(X) = f(X) - \mathbb{E}_{X'}f(X')$, where X' is an independent copy of X. By the Jensen's inequality

$$\mathbb{E}_X e^{\lambda(f(X) - \mathbb{E}_{X'}f(X'))} \leq \mathbb{E}_{X,X'} e^{\lambda(f(X) - f(X'))}. \tag{5.9}$$

Suppose f is differentiable then, by the fundamental theorem of calculus,

$$\begin{aligned} f(X) - f(X') &= \int_0^{\pi/2} \tfrac{\mathrm{d}}{\mathrm{d}\theta} f(X\sin\theta + X'\cos\theta)\mathrm{d}\theta \\ &= \int_0^{\pi/2} \Big\langle \nabla f(X\sin\theta + X'\cos\theta), X\cos\theta - X'\sin\theta \Big\rangle \mathrm{d}\theta \end{aligned}$$

Note that the variables $X_\theta := X\sin\theta + X'\cos\theta$ and $X'_\theta := X\cos\theta - X'\sin\theta$ are independent standard normal and so the distribution of (X, X') is the same as the distribution of (X_θ, X'_θ). The right-hand side in (5.9) can be rewritten as

$$\mathbb{E}_{X,X'} e^{\lambda \int_0^{\pi/2} \langle \nabla f(X_\theta), X'_\theta\rangle \mathrm{d}\theta}.$$

If $\boldsymbol{\theta}$ is a random variable uniformly distributed on $(0, \frac{\pi}{2})$ then

$$\int_0^{\pi/2} \langle \nabla f(X_\theta), X'_\theta \rangle \mathrm{d}\theta = \frac{\pi}{2} \mathbb{E}_{\boldsymbol{\theta}} \langle \nabla f(X_{\boldsymbol{\theta}}), X'_{\boldsymbol{\theta}} \rangle.$$

Using the Jensen's inequality again (and Fubini's theorem), we get

$$\mathbb{E}_{X,X'} e^{\lambda \int_0^{\pi/2} \langle \nabla f(X_\theta), X'_\theta \rangle \mathrm{d}\theta} \;=\; \mathbb{E}_{X,X'} e^{\lambda \frac{\pi}{2} \mathbb{E}_{\boldsymbol{\theta}} \langle \nabla f(X_{\boldsymbol{\theta}}), X'_{\boldsymbol{\theta}} \rangle} \;\leq\; \mathbb{E}_{\boldsymbol{\theta}} \mathbb{E}_{X,X'} e^{\lambda \frac{\pi}{2} \langle \nabla f(X_{\boldsymbol{\theta}}), X'_{\boldsymbol{\theta}} \rangle}.$$

For any fixed θ, (X_θ, X'_θ) has the same distribution as (X, X'). It then follows that

$$\mathbb{E}_{\boldsymbol{\theta}} \mathbb{E}_{X,X'} e^{\lambda \frac{\pi}{2} \langle \nabla f(X_{\boldsymbol{\theta}}), X'_{\boldsymbol{\theta}} \rangle} \;=\; \mathbb{E}_{\boldsymbol{\theta}} \mathbb{E}_{X,X'} e^{\lambda \frac{\pi}{2} \langle \nabla f(X), X' \rangle} = \mathbb{E}_{X,X'} e^{\lambda \frac{\pi}{2} \langle \nabla f(X), X' \rangle}.$$

Note that for fixed X, the variable $\langle \nabla f(X), X' \rangle$ is Gaussian with mean zero and variance $\|\nabla f(X)\|^2$ and hence

$$\mathbb{E}_{X,X'} e^{\lambda \frac{\pi}{2} \langle \nabla f(X), X' \rangle} \;\leq\; \mathbb{E}_X e^{\lambda^2 \pi^2 \|\nabla f(X)\|^2 / 8} \;\leq\; e^{\lambda^2 \pi^2 L^2 / 8}.$$

This calculation shows that

$$\mathbb{E}_X e^{\lambda (f(X) - \mathbb{E}_{X'} f(X'))} \leq e^{\lambda^2 \pi^2 L^2 / 8}$$

or, in other words, that $f(X) - \mathbb{E} f(X)$ is sub-Gaussian with parameter $\sigma = \frac{\pi L}{2}$, which is slightly more than the claimed L. □

This result is useful for a broad range of problems.

Example 5.4.2 (χ^2 concentration)**.** *For a given vector $Z = (Z_1, \ldots, Z_n)$ of i.i.d. standard normal variables, we have $Y := \sum_{i=1}^n Z_i^2 = \|Z\|^2 \sim \chi^2_n$. The most direct way to obtain tail bounds on Y was given in Example 5.1.4. Alternatively we can use Theorem 5.4.1. Define $V = \sqrt{Y/n} = \|Z\|/\sqrt{n}$. Since the Euclidean norm is a 1-Lipschitz function, we get*

$$\mathbb{P}(V - \mathbb{E}V \geq t) \;\leq\; e^{-nt^2/2}.$$

To obtain bounds on $Y - \mathbb{E}Y$ note that, by Jensen's inequality,

$$\mathbb{E}[V] \leq \sqrt{\mathbb{E}[V^2]} = 1.$$

Thus

$$\mathbb{P}(V - \mathbb{E}V \geq t) \;\geq\; \mathbb{P}(V - 1 \geq t) \;=\; \mathbb{P}(Y/n \geq (t+1)^2).$$

Putting everything together we conclude

$$\mathbb{P}(Y/n \geq (1+t)^2) \;\leq\; e^{-nt^2/2} \qquad \textit{for all } t \geq 0.$$

Example 5.4.3 (Order statistics)**.** *Given a random vector $(X_1, \ldots, X_n)$, its order statistics are obtained by reordering its entries in a non-decreasing manner as*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}.$$

As particular cases, we have $X_{(n)} = \max_i X_i$ and $X_{(1)} = \min_i X_i$. Given another random vector $(Y_1, \dots, Y_n)$, it can be shown that $|X_{(k)} - Y_{(k)}| \leq \|X - Y\|$ for all $k = 1, \dots, n$[10], so that each order statistic is a 1-Lipschitz function. Consequently, when X is a standard Gaussian random vector, Theorem 5.4.1 implies that

$$\mathbb{P}(|X_{(i)} - \mathbb{E}X_{(i)}| \geq t) \;\leq\; 2e^{-t^2/2} \qquad \text{for all } t \geq 0.$$

[10] The fact that $|X_k - Y_k| \leq \|X - Y\|$ for every k is trivial. But this statement is more subtle!

Example 5.4.4 (Singular values of Gaussian random matrices)**.** *For $n \geq d$, consider the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, and denote by*

$$\sigma_1(\boldsymbol{X}) \geq \sigma_2(\boldsymbol{X}) \geq \cdots \geq \sigma_d(\boldsymbol{X}) \geq 0$$

the singular values of $\boldsymbol{X}$. By Weyl's theorem

$$\max_{i=1,\dots,d} |\sigma_i(\boldsymbol{X}) - \sigma_i(\boldsymbol{Y})| \;\leq\; \|\boldsymbol{X} - \boldsymbol{Y}\| \;\leq\; \|\boldsymbol{X} - \boldsymbol{Y}\|_F.$$

In other words, each singular value $\sigma_i(\boldsymbol{X})$ is a 1-Lipschitz function of $\boldsymbol{X}$. Suppose now that $\boldsymbol{W}$ is random with independent standard normal entries. Theorem 5.4.1 implies that

$$\mathbb{P}(|\sigma_i(\boldsymbol{W}) - \mathbb{E}\sigma_i(\boldsymbol{W})| \geq t) \;\leq\; 2e^{-t^2/2} \qquad \text{for all } t \geq 0,$$

or in other words

$$\mathbb{P}(|\sigma_i(\tfrac{1}{\sqrt{n}}\boldsymbol{W}) - \mathbb{E}\sigma_i(\tfrac{1}{\sqrt{n}}\boldsymbol{W})| \geq t) \;\leq\; 2e^{-t^2 n/2} \qquad \text{for all } t \geq 0.$$

If $\mathbf{X}$ has i.i.d. rows from $N_d(0, \Sigma)$ then $\mathbf{X} = \boldsymbol{W}\sqrt{\Sigma}$ and the sample covariance satisfies

$$\widehat{\Sigma} = \tfrac{1}{n}\mathbf{X}^\intercal\mathbf{X} = \sqrt{\Sigma}(\tfrac{1}{\sqrt{n}}\boldsymbol{W})^\intercal(\tfrac{1}{\sqrt{n}}\boldsymbol{W})\sqrt{\Sigma}.$$

We can exploit this to obtain some bounds on $\|\widehat{\Sigma} - \Sigma\|$ (see Chapter 6 in [11]).

[11] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge, 2019

Example 5.4.5 (Gaussian width)**.** *Let $W = (W_1, \dots, W_n)$ be the n-dimensional standard Gaussian vector. Given a collection of vectors $\mathcal{A} \subset \mathbb{R}^n$, define the random variable*

$$Z = Z(\mathcal{A}) := \sup_{a \in \mathcal{A}} \langle a, W \rangle.$$

The variable Z is one way of measuring the size of the set $\mathcal{A}$, which is called the Gaussian width. We view Z as a function $(w_1, \dots, w_n) \mapsto f(w_1, \dots, w_n)$. Fixing a, we get

$$\langle a, w \rangle - \langle a, w' \rangle \;\leq\; \|a\|\|w - w'\| \;\leq\; \sup_{a \in \mathcal{A}} \|a\|\|w - w'\|.$$

Then

$$\sup_{a \in \mathcal{A}} \langle a, w \rangle - \sup_{a \in \mathcal{A}} \langle a, w' \rangle \;\leq\; \|a\|\|w - w'\|,$$

and similarly

$$\sup_{a\in\mathcal{A}}\langle a,w'\rangle - \sup_{a\in\mathcal{A}}\langle a,w\rangle \;\leq\; \sup_{a\in\mathcal{A}}\|a\|\|w-w'\|.$$

Let $D(\mathcal{A}) := \sup_{a\in\mathcal{A}}\|a\|$. *Thus* $f(w)$ *is* $D(\mathcal{A})-$ *Lipschitz and this*

$$\mathbb{P}(|Z-\mathbb{E}Z|\geq t) \;\leq\; 2e^{-\frac{t^2}{2D^2(\mathcal{A})}}.$$

5.5 Exercises

Exercise 5.5.1. *Provide a non-negative random variable for this the Markov inequality holds as equality (for a fixed* t*). Is there a random variable for which it holds as equality for every* $t\geq 0$*?*

Exercise 5.5.2. *Show that for* $U\sim N(0,1)$ *it holds that*

$$\mathbb{P}(U\geq t) \;\leq\; \frac{1}{\sqrt{2\pi}}\frac{1}{t}e^{-t^2/2}.$$

Hint: If $\phi(u)$ *is the density of* U*, show* $\phi'(u)+u\phi(u)=0$ *for all* $u\in\mathbb{R}$*. Conclude that* $\int_t^\infty \phi(u)\mathrm{d}u \leq \frac{1}{t}\phi(t)$.

Exercise 5.5.3. *Show that in Proposition 5.1.3 the same way we get*

$$\log\mathbb{P}(Z-\mathbb{E}Z\leq -t) \;\leq\; \inf_{\lambda\in(-b,0]}\{\log\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}+\lambda t\}.$$

Conclude bounds on $\mathbb{P}(|Z-\mathbb{E}Z|\geq t)$.

Exercise 5.5.4. *Use Exercise 5.5.3 to show that in the Example 5.1.4*

$$\mathbb{P}(\bar{Y}_n-1\leq -t)\leq (1-t)^{n/2}e^{nt/2}.$$

Conclude for $t\in[0,1]$ *that*

$$\mathbb{P}(|\bar{Y}_n-1|\geq t)\leq 2e^{-nt^2/8}. \tag{5.10}$$

Exercise 5.5.5. *Use Exercise 5.5.3 to show that for all* $t\geq 0$

$$\mathbb{P}(X-\mu\leq -t) \;\leq\; e^{-t^2/2\sigma^2} \qquad \textit{and} \qquad \mathbb{P}(|X-\mu|\geq t) \;\leq\; 2e^{-t^2/2\sigma^2}.$$

Exercise 5.5.6. *Suppose* X *is* σ*-sub-Gaussian with mean* μ*. Show that* $\mathrm{var}(X)\leq\sigma^2$.

Exercise 5.5.7. *Suppose* X_i, $i=1,\ldots,N$ *are zero-mean* σ*-sub-gaussian random variables (not neccessarily independent). Use Proposition 6.1.1 to show that*

$$\mathbb{P}(\max_i X_i - \mathbb{E}\max_i X_i\geq t) \;\leq\; e^{-t^2/\sigma^2} \qquad \textit{for } t<2\sigma\sqrt{2\log N}.$$

6
More advanced techniques (1-2 weeks)

6.1 Maximal inequalities

Lemma 5.2.4 and Lemma 5.2.14 provide concentration bounds on linear combinations of sub-Gaussian or sub-exponential variables. In many instances, we will be interested in controlling the maximum over the parameters of such linear combinations. The main motivation is in empirical risk minimization (see (2.1)) but many other applications exist. The purpose of this section is to present such results.

We begin by the simplest case possible: the maximum over a finite set.

Proposition 6.1.1. *Suppose $X_1, \dots, X_N$ are zero-mean σ-sub-Gaussian then*

$$\mathbb{E} \max_{i=1,\dots,N} X_i \le \sigma\sqrt{2\log N} \quad \textit{and} \quad \mathbb{E} \max_{i=1,\dots,N} |X_i| \le \sigma\sqrt{2\log(2N)}.$$

Moreover, for any $t > 0$

$$\begin{aligned} \mathbb{P}(\max_{i=1,\dots,N} X_i \ge t) &\le N e^{-t^2/(2\sigma^2)} \\ \mathbb{P}(\max_{i=1,\dots,N} |X_i| \ge t) &\le 2N e^{-t^2/(2\sigma^2)}. \end{aligned} \tag{6.1}$$

In other words, for $\delta \in (0,1)$,

$$\mathbb{P}\left(\max_i X_i \ge \sigma\sqrt{2\log(\tfrac{N}{\delta})}\right) \le \delta.$$

Proof. Take $\lambda > 0$. Then we get

$$\begin{aligned} e^{\lambda \mathbb{E}\max_i X_i} &\overset{\text{Jensen}}{\le} \mathbb{E} e^{\lambda \max_i X_i} = \mathbb{E}\max_i e^{\lambda X_i} \\ &\le \sum_{i=1}^N \mathbb{E} e^{\lambda X_i} \le N e^{\sigma^2\lambda^2/2}. \end{aligned}$$

By taking logs and optimizing with respect to λ we get that the optimal $\lambda = \sqrt{2\log N}/\sigma$. Plugging this back, we obtain

$$\mathbb{E}\max_i X_i \;\le\; \frac{\log N}{\lambda} + \frac{\sigma^2\lambda}{2} \;\le\; \sigma\sqrt{2\log N}.$$

The analogous bound for the absolute values follows from the fact that $\max_i |X_i| = \max\{X_1, -X_1, \dots, X_N, -X_N\}$. Finally, the probability bounds follow directly by the union bound and σ-sub-Gaussianity of X_i. □

Figure 6.1 offers an illustration of Proposition 6.1.1. The blue curve depicts a Monte Carlo estimate of the maximum of N standard normal variables ($\sigma = 1$) and its theoretical bound $\sqrt{2\log(N)}$ is plotted in red. The bound reveals the right rate. Note also that in Proposition 6.1.1 we did not require that X_i are independent!

Let $\boldsymbol{X} = (X_1, \dots, X_N)$ be a vector of independent σ-sub-Gaussian variables. It is often important to analyze suprema of random functions over a general set $A \subseteq \mathbb{R}^N$:

$$\sup_{\boldsymbol{u} \in A} \boldsymbol{u}^\top \boldsymbol{X} \tag{6.2}$$

By Corollary 5.2.7, each $\boldsymbol{u}^\top \boldsymbol{X}$ is $\sigma\|\boldsymbol{u}\|$-sub-Gaussian. If A if finite, we can use Proposition 6.1.1 to get

$$\mathbb{E} \max_{\boldsymbol{u} \in A} \boldsymbol{u}^\top \boldsymbol{X} \;\le\; \sigma \operatorname{diam}(A) \sqrt{2 \log |A|},$$

where $\operatorname{diam}(A) = \max_{\boldsymbol{u} \in A} \|\boldsymbol{u}\|$.

Sometimes, even if A is not finite, there is a trivial reduction to the finite case. Suppose that $P \subseteq \mathbb{R}^d$ is a polytope, that is, a set of convex combinations of some fixed points $v_1, \dots, v_N$. In this case we have

$$\sup_{\boldsymbol{u} \in P} \boldsymbol{u}^\intercal \mathbf{X} \;=\; \max_{i=1,\dots,N} \boldsymbol{v}_i^\intercal \mathbf{X}$$

and now we can use Proposition 6.1.1 to obtain bounds on $\mathbb{E} \sup_{\boldsymbol{u} \in P} \boldsymbol{u}^\intercal \mathbf{X}$ and $\mathbb{P}(\sup_{\boldsymbol{u} \in P} \boldsymbol{u}^\intercal \mathbf{X} \ge t)$. Of particular interest are polytopes that have a small number of vertices. A primary example is the ℓ_1 ball of $\mathbb{R}^d$ defined by

$$\mathbb{B}_1 \;=\; \{x \in \mathbb{R}^d : \sum_{i=1}^d |x_i| \le 1\},$$

which has exactly $2d$ vertices.

If A is bounded, we still have a principled way to study the supremum in (6.2). To simplify the discussion we first present the main ideas in the specific case of the ℓ_2 ball

$$\mathbb{B}_2 \;=\; \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i^2 \le 1\}.$$

The general idea is to cover $\mathbb{B}_2$ with a finite set of points such that the maximum over this finite set is of the same order as the maximum over the entire ball.

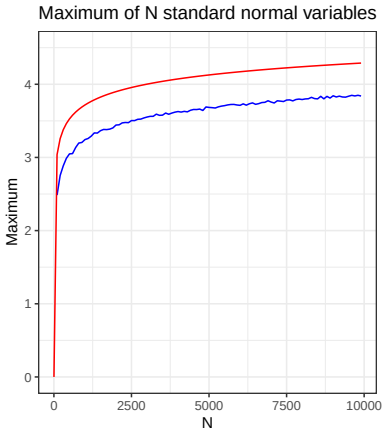


Figure 6.1: Illustration of the bound in Proposition 6.1.1.

← Exercise 5.5.7

Definition 6.1.2. *Fix $K \subset \mathbb{R}^d$ and $\epsilon > 0$. A set $\mathcal{N}$ is called an **ϵ-cover of** K, if $\mathcal{N} \subseteq K$ and for any $z \in K$, there exists $x \in \mathcal{N}$ such that $\|x - z\| \leq \epsilon$.*

Definition 6.1.3. *The ϵ-covering number of K is*

$$N(\epsilon, K) \;:=\; \inf\{N \in \mathbb{N} : \exists \text{ an } \epsilon\text{-cover of } K \text{ of size } N\}$$

The following lemma gives an upper bound on the size of the smallest ϵ-cover of $\mathbb{B}_2$.

Lemma 6.1.4. *Fix $\epsilon \in (0, 1)$. Then the unit Euclidean ball $\mathbb{B}_2$ has an ϵ-cover $\mathcal{N}$ of cardinality $|\mathcal{N}| \leq (3/\epsilon)^d$.*

Proof. Consider the following iterative construction of the ϵ-cover. Choose $x_1 = 0$. For any $i \geq 2$, take x_i to be any point in $\mathbb{B}_2$ such that $\|x_i - x_j\| > \epsilon$ for all $j < i$. If no such point exists, stop the procedure[1]. This will create an ϵ-cover. We now control its size.

Since $\|x - y\| > \epsilon$ for all $x, y \in \mathcal{N}$, the Euclidean balls of radius $\epsilon/2$ and centered at points of $\mathcal{N}$ are disjoint. Moreover,

$$\bigcup_{z \in \mathcal{N}} \{z + \tfrac{\epsilon}{2}\mathbb{B}_2\} \;\subset\; (1 + \tfrac{\epsilon}{2})\mathbb{B}_2.$$

Thus, measuring volumes, we get

$$\mathrm{vol}((1 + \tfrac{\epsilon}{2})\mathbb{B}_2) \;\geq\; \mathrm{vol}\Big(\bigcup_{z \in \mathcal{N}} \{z + \tfrac{\epsilon}{2}\mathbb{B}_2\}\Big) \;=\; |\mathcal{N}|\mathrm{vol}(\tfrac{\epsilon}{2}\mathbb{B}_2)$$

or equivalently

$$(1 + \tfrac{\epsilon}{2})^d \;\geq\; |\mathcal{N}|(\tfrac{\epsilon}{2})^d,$$

which gives the following bound

$$|\mathcal{N}| \;\leq\; (1 + \tfrac{2}{\epsilon})^d \;\leq\; (\tfrac{3}{\epsilon})^d$$

□

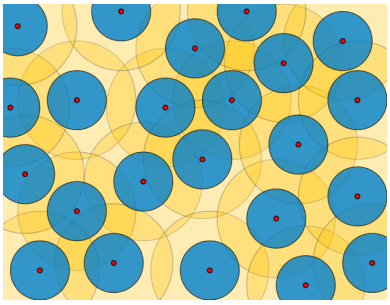


Figure 6.2: ϵ-cover of a square. Here in addition the points are at least ϵ apart from each other. Source: Wikipedia.

[1] The procedure needs to stop by compactness of $\mathbb{B}_2$. If not, we would have a sequence (x_n) in $\mathbb{B}_2$, which then would need to have a convergent subsequence. This would contradict the assumption that the elements of this sequence are all at least ϵ apart.

Recall from Corollary 5.2.7 that if $\mathbf{X} = (X_1, \ldots, X_n)$ is a vector of independent σ-sub-Gaussian random variables then for any unit vector $\boldsymbol{u}$ the variable $\boldsymbol{u}^\intercal \mathbf{X}$ is σ-sub-Gaussian. We can then introduce the following general definition of sub-Gaussian vectors.

Definition 6.1.5. *A random vector $\mathbf{X}$ is called σ-sub-Gaussian if $\boldsymbol{u}^\intercal \mathbf{X}$ is σ-sub-Gaussian for any unit vector $\boldsymbol{u}$.*

In this definition we do not require the components of $\mathbf{X}$ to be independent

Theorem 6.1.6. *Let $\mathbf{X} \in \mathbb{R}^d$ be a mean-zero σ-sub-Gaussian vector. Then*

$$\mathbb{E}\|\mathbf{X}\| \;=\; \mathbb{E}(\max_{\boldsymbol{u} \in \mathbb{B}_2} \boldsymbol{u}^\intercal \mathbf{X}) \;\leq\; 4\sigma\sqrt{d}.$$

Moreover, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$\|\mathbf{X}\| \;\leq\; 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}.$$

Proof. Let N be a $1/2$-cover of $\mathbb{B}_2$ that satisfies $|\mathcal{N}| \leq 6^d$. Next, observe that for every $\boldsymbol{u} \in \mathbb{B}_2$, there exists $\mathbf{y} \in \mathcal{N}$ and $\boldsymbol{h}$ such that $\|\boldsymbol{h}\| \leq 1/2$ and $\boldsymbol{u} = \mathbf{y} + \boldsymbol{h}$. Therefore,

$$\max_{\boldsymbol{u}\in\mathbb{B}_2} \boldsymbol{u}^\mathsf{T}\mathbf{X} \;\leq\; \max_{\mathbf{y}\in\mathcal{N}} \mathbf{y}^\mathsf{T}\mathbf{X} \;+\; \max_{\boldsymbol{h}\in\frac{1}{2}\mathbb{B}_2} \boldsymbol{h}^\mathsf{T}\mathbf{X} \;=\; \max_{\mathbf{y}\in\mathcal{N}} \mathbf{y}^\mathsf{T}\mathbf{X} \;+\; \frac{1}{2}\max_{\boldsymbol{h}\in\mathbb{B}_2} \boldsymbol{h}^\mathsf{T}\mathbf{X}.$$

Therefore,

$$\max_{\boldsymbol{u}\in\mathbb{B}_2} \boldsymbol{u}^\mathsf{T}\mathbf{X} \;\leq\; 2\max_{\mathbf{y}\in\mathcal{N}} y^\mathsf{T}\mathbf{X}, \tag{6.3}$$

and so, using Proposition 6.1.1,

$$\mathbb{E}\max_{\boldsymbol{u}\in\mathbb{B}_2} \boldsymbol{u}^\mathsf{T}\mathbf{X} \;\leq\; 2\mathbb{E}\max_{\mathbf{y}\in\mathcal{N}} y^\mathsf{T}\mathbf{X} \;\leq\; 2\sigma\sqrt{2\log|\mathcal{N}|} \;\leq\; 4\sigma\sqrt{d}.$$

The bound with high probability then follows because, using (6.3),

$$\mathbb{P}(\max_{\boldsymbol{u}\in\mathbb{B}_2} \boldsymbol{u}^\mathsf{T}\mathbf{X} \geq t) \;\leq\; \mathbb{P}(2\max_{\mathbf{y}\in\mathcal{N}} \mathbf{y}^\mathsf{T}\mathbf{X} \geq t) \;\leq\; |\mathcal{N}|e^{-t^2/(8\sigma^2)} \;\leq\; 6^d e^{-t^2/(8\sigma^2)}.$$

Plugging $t = 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$ we verify the bound. □

← Exercise 6.5.1

6.2 *Rademacher complexity and bounds on suprema*

In this section we are again concerned with the variable

$$Z_n = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X) \right| = \|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F}.$$

In Theorem 5.3.9, we showed that if $\mathcal{F}$ is uniformly b-bounded, we get

$$\mathbb{P}(Z_n - \mathbb{E}Z_n \geq t) \;\leq\; \exp\{-\tfrac{nt^2}{2b^2}\}. \tag{6.4}$$

By setting $\delta = \exp\{-\frac{nt^2}{2b^2}\}$, we get $t = \sqrt{\frac{2b^2}{n}\log(\frac{1}{\delta})}$. In other words, with probability at least $1-\delta$, we have

$$0 \;\leq\; Z_n \;\leq\; \mathbb{E}Z_n + \sqrt{\frac{2b^2}{n}\log(\frac{1}{\delta})}. \tag{6.5}$$

The second term on the right diminishes to 0 at the order $1/\sqrt{n}$. Since $\mathbb{E}Z$ is unknown, we need to bound it. The three main tools to construct such bounds are: symmetrization, discretization, and chaining. We now discuss them in more detail.

Definition 6.2.1 (Rademacher Complexity)**.** *For a fixed collection* $x_1^n := \{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathcal{X}$, *consider*

$$\mathcal{F}(x_1^n) = \{(f(x_1), \ldots, f(x_n)) : \; f \in \mathcal{F}\} \;\subseteq \mathbb{R}^n. \tag{6.6}$$

The empirical Rademacher complexity is defined as

$$\mathcal{R}_n(\mathcal{F}(x_1^n)/n) := \mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right].$$

Taking expectation w.r.t. $X_1, X_2, \ldots, X_n$, we obtain the Rademacher complexity as

$$\mathcal{R}_n(\mathcal{F}) \;:=\; \mathbb{E}_X \mathcal{R}_n(\mathcal{F}(X_1^n)/n).$$

Consider the following basic bound.

Theorem 6.2.2 (Symmetrization)**.** *For any class of measurable functions $\mathcal{F}$, we have*

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} \;\le\; 2\mathcal{R}_n(\mathcal{F}).$$

Proof. Suppose $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$. As a result, we have

$$\begin{aligned}
\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} &= \mathbb{E}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X_i)) \right| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X_i')) \right| \\
&\qquad \text{(consider the convex function } g(y_1,\ldots,y_n) = \sup_f |\tfrac{1}{n}\textstyle\sum_i (f(x_i) - y_i)|) \\
&\overset{\text{Jensen}}{\le} \mathbb{E}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n (f(X_i) - f(X_i')) \right|.
\end{aligned}$$

We have $f(X_i) - f(X_i') \overset{d}{=} \varepsilon_i(f(X_i) - f(X_i'))$ for all i. This is why we call this the symmetrization argument. Thus we obtain

$$\begin{aligned}
\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} &= \mathbb{E}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i (f(X_i) - f(X_i')) \right| \right) \\
&\overset{\triangle}{\le} 2\mathbb{E}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \right| \right) \\
&= 2\mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

□

As a corollary, using (6.5), we get the following important result.

Proposition 6.2.3. *For any uniformly b-bounded class of functions $\mathcal{F}$, any positive integer $n \ge 1$ and any scalar $t \ge 0$, we have*

$$\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} \;\le\; 2\mathcal{R}_n(\mathcal{F}) + t$$

with probability at least $1 - \exp(-\frac{nt^2}{2b^2})$. Consequently, as long as $\mathcal{R}_n(\mathcal{F}) = o(1)$, we have $\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} \to 0$ almost surely (i.e. $\mathcal{F}$ is a Glivenko-Cantelli class for any $\mathbb{P}$).

Equivalently, $\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} \;\le\; 2\mathcal{R}_n(\mathcal{F}) + b\sqrt{\frac{2}{n}\log(\frac{1}{\delta})}$ with probability $\ge 1 - \delta$.

Proof. Let $Z_n := \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$. The first part follows directly from the earlier result and Theorem 5.3.9. In particular, convergence in probability of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ to zero follows easily[2]. To get stronger almost sure convergence we employ the Borel-Cantelli lemma. By basic calculus, we have[3]

$$\{Z_n \to 0\} \;=\; \bigcap_{N=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{m \geq n} \{Z_m \leq \frac{1}{N}\}. \tag{6.7}$$

To prove almost sure convergence, we show that the complement of the set in (6.7) has measure zero. For a fixed N let $E_n = \{Z_n > \frac{1}{N}\}$. If n is large enough then $2\mathcal{R}_n(\mathcal{F}) < \frac{1}{2N}$ and so, for such n, taking $t = \frac{1}{2N}$ we get $\mathbb{P}(E_n) = \mathbb{P}(Z_n > \frac{1}{N}) \leq e^{-\frac{n}{2b^2N^2}}$. This implies that

$$\sum_{n \geq 1} \mathbb{P}(E_n) \;<\; \infty.$$

By the Borel-Cantelli lemma[4], for every $N \geq 1$, $\mathbb{P}(\bigcup_{n=1}^{\infty} \bigcap_{m \geq n} \{Z_n \leq \frac{1}{N}\}) = 1$. Thus, in (6.7), we have a countable collection of measure 1 events. This implies that their intersection has measure 1 too, proving $\mathbb{P}(Z_n \to 0) = 1$. □

[2] Recall that $Z_n \xrightarrow{p} 0$ if $\forall \epsilon > 0\ \mathbb{P}(Z_n > \epsilon) \to 0$. Moreover, $Z_n \xrightarrow{a.s.} 0$ if $\mathbb{P}(Z_n \to 0) = 1$.

[3] Recall $x_n \to 0$ if and only if $\forall N \geq 1\ \exists n \geq 1$ s.t. $\forall m \geq n\ |x_m| \leq \frac{1}{N}$. Also, $\bigcup_{\alpha \in A} E_\alpha = \{x : \exists \alpha \text{ s.t. } x \in E_\alpha\}$ and $\bigcap_{\alpha \in A} E_\alpha = \{x : \forall \alpha \text{ s.t. } x \in E_\alpha\}$, which is why quantificator statements easily translate to set operations as in (6.7).

[4] B-C states that if $\sum_{n \geq 1} \mathbb{P}(E_n) < \infty$ then $\mathbb{P}(\bigcap_{n \geq 1} \bigcup_{m \geq n} E_m) = 0$.

To bound $\mathcal{R}_n(\mathcal{F})$, we first fix $x_1, .., x_n \in \mathcal{X}$ and bound the empirical Rademacher complexity $\mathcal{R}_n(\mathcal{F}(x_1^n)/n)$. Second, if the upper bound for $\mathcal{R}_n(\mathcal{F}(x_1^n)/n)$ does not depend on $x_1, \ldots, x_n$, then it automatically becomes an upper bound for $\mathcal{R}_n(\mathcal{F})$.

We now state a simple upper bound for $\mathcal{R}_n(A) = \mathbb{E}[\sup_{a \in A} \frac{1}{n} \sum_{i=1}^n \varepsilon_i a_i]$ if $A \in \mathbb{R}^n$ is a set with finite elements. As a corollary from Proposition 6.1.1 we get the following bound.

← Exercise 6.5.2

Proposition 6.2.4 (Discretization). *Suppose A is a finite subset of $\mathbb{R}^n$ with cardinality of $|A|$. Then*

$$\mathcal{R}_n(A) \;=\; \mathbb{E} \max_{a \in A} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i a_i \right| \;\leq\; \max_{a \in A} \sqrt{\frac{1}{n} \sum_{i=1}^n a_i^2} \, \sqrt{\frac{2 \log(2|A|)}{n}}.$$

Recall that if ε_i are independent 1-sub-Gaussian then $\frac{1}{n} a^\top \varepsilon$ is $\frac{1}{n}\|a\|$-sub-Gaussian.

In Section 6.3 we show that the finite case already leads to deep and interesting results. Later, in Section 6.4 we show more general techniques.

6.3 *Polynomial discrimination and VC dimension*

In Proposition 6.2.4 we showed how to bound the Rademacher complexity for a finite set A. In this section we exploit this result.

Definition 6.3.1 (Boolean Class). *We say $\mathcal{F}$ is a Boolean (function) class if $\forall f \in \mathcal{F}$ and $\forall x \in \mathcal{X}$, $f(x) \in \{0, 1\}$.*

An important example is given by the binary classification discussed in Section 4.1.1 and the uniform empirical process.

Example 6.3.2 (Binary Classification)**.** *Consider a pair of random objects (X, Y) having some joint distribution where $X \in \mathcal{X}$, $Y \in \{0, +1\}$. A classifier is a function $g : \mathcal{X} \mapsto \{0, +1\}$ The error of the classifier is given by*

$$L(g) := \mathbb{P}(g(X) \neq Y) = \mathbb{E}\mathbb{1}(g(X) \neq Y).$$

Thus, in binary classification problems, the functions of interest are of the form $\mathbb{1}(g(X) \neq Y)$, and $\{\mathbb{1}(g(X) \neq Y) : g \in G\}$ is a Boolean class for any set of classifiers G.

Example 6.3.3 (Uniform Empirical Process)**.** *When deriving the asymptotic law of the empirical CFD, the function class of interest $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}\}$ is a Boolean class.*

Now, fix a Boolean class $\mathcal{F}$ and points $\{x_1, \cdots, x_n\}$ in $\mathcal{X}$. Then $\mathcal{F}(x_1^n)$ defined in (6.6) is a finite set, contained in $\{0, 1\}^n$, whose cardinality is then at most 2^n. Applying Proposition 6.2.4, to $A = \mathcal{F}(x_1^n)/n$ we obtain

$$\mathcal{R}_n(\mathcal{F}(x_1^n)/n) = \mathbb{E} \max_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_i \varepsilon_i f(x_i) \right| \leq \sqrt{\frac{2 \log (2|\mathcal{F}(x_1^n)|)}{n}}, \quad (6.8)$$

since the first term in Proposition 6.2.4 equals to $\sup_{f \in \mathcal{F}} \sqrt{\sum_{i=1}^n f^2(x_i)/n}$, which is less than or equal to 1.

Of course, if $|\mathcal{F}(x_1^n)| \approx 2^n$ then the bound in (6.8) is not very interesting. It becomes interesting, for example, when the cardinality of the function class grows only as a polynomial function of n, as formalized below.

Definition 6.3.4 (Polynomial Discrimination)**.** *The Boolean class $\mathcal{F}$ is said to have polynomial discrimination if there exists a polynomial $\rho(\cdot)$ such that for every positive integer n and every set of n points $\{x_1, \cdots, x_n\}$, the set $\mathcal{F}(x_1^n)$ satisfies*

$$|\mathcal{F}(x_1^n)| \leq \rho(n).$$

The significance of this property is that, together with inequality (6.8), it provides a straightforward approach to controlling the Rademacher complexity, $\mathcal{R}_n(\mathcal{F}) = o(1)$. By Proposition 6.2.3 and (6.8), we get the following result.

Proposition 6.3.5. *If a Boolean class $\mathcal{F}$ has polynomial discrimination then it is Glivenko-Cantelli class for any $\mathbb{P}$.* [5]

[5] c.f. Definition 4.1.2

The first important example is given by the empirical process.

Proposition 6.3.6. *The Boolean class $\mathcal{F} = \{\mathbb{1}\{x \leq t\} : t \in \mathbb{R}\}$ satisfies $|\mathcal{F}(x_1^n)| \leq n+1$. In particular, it has polynomial discrimination and so $\mathcal{F}$ is a Glivenko-Cantelli class for any $\mathbb{P}$.*

But how does one check if a given Boolean class $\mathcal{F}$ has polynomial discrimination in general? One of the most popular approaches is to use the Vapnik Chervonenkis dimension, or the VC dimension for short. We have the following lemma.

Definition 6.3.7 (VC Dimension)**.** *The VC dimension $\mathsf{VC}(\mathcal{F})$ of a class of Boolean functions $\mathcal{F}$ on $\mathcal{X}$ is defined as the **largest integer** D such that there exists a finite subset $\{x_1, \cdots, x_D\}$ of $\mathcal{X}$ satisfying $\mathcal{F}(x_1^D) = \{0,1\}^D$.*

Note that it always holds that $\mathcal{F}(x_1^D) \subseteq \{0,1\}^D$

Definition 6.3.8 (Shattering)**.** *A finite subset $\{x_1, \cdots, x_m\} \subset \mathcal{X}$ is said to be shattered by Boolean class $\mathcal{F}$ if $\mathcal{F}(x_1^m) = \{0,1\}^m$, i.e., $\left|\mathcal{F}(x_1^m)\right| = 2^m$.*

Remark 6.3.9. *By the definitions of VC dimension and shattering, we know the VC dimension is the largest integer n for which there is a n-point set $\{x_1, \cdots, x_n\}$ which can be shattered by $\mathcal{F}$.*

Example 6.3.10. *Let $\mathcal{F}$ be the set of indicator functions of the form $\mathbb{1}_{(-\infty,t]}(x)$ on $\mathbb{R}$. Then $\mathsf{VC}(\mathcal{F}) = 1$. Indeed, if $x_1 < x_2$ then there is no $f \in \mathcal{F}$ for which $(f(x_1), f(x_2)) = (1,0)$. By the same argument, no bigger set of points in $\mathbb{R}$ can be shuttered.*

The next lemma shows that any class with finite VC dimension has polynomial discrimination with degree of at most the VC dimension.

Lemma 6.3.11 (Sauer's Lemma)**.** *Suppose $\mathsf{VC}(\mathcal{F}) = D$, then for every $n \geq 1$ and every collection $\{x_1, \cdots, x_n\}$ of n points, we have*

$$|\mathcal{F}(x_1^n)| \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D}.$$

If $k > n$, we have $\binom{n}{k} = 0$, and if $D \leq n$, we have

$$|\mathcal{F}(x_1^n)| \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D} \leq \left(\frac{en}{D}\right)^D.$$

Proof. See Proposition 4.18 of [6] for the detailed proof. □

[6] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge, 2019

As an immediate corollary, we get the following result.

Proposition 6.3.12. *Suppose $\mathcal{F}$ is a Boolean class and $\mathsf{VC}(\mathcal{F}) \leq D$. Then for $n \geq D$, we have*

$$\mathcal{R}_n(\mathcal{F}) \;\leq\; 2\sqrt{\frac{D}{n}\log\left(\frac{en}{D}\right)}$$

and so

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \;\leq\; 4\sqrt{\frac{D}{n}\log\left(\frac{en}{D}\right)}.$$

Proof. Combined with VC lemma and our previous upper bound (6.8), we can obtain that for $n \geq D$

$$\mathcal{R}_n(\mathcal{F}) \;\leq\; \sqrt{\frac{2\log 2 + 2\log\left(\frac{en}{D}\right)^D}{n}} \;\leq\; \sqrt{\frac{2(D+1)\log\left(\frac{en}{D}\right)}{n}} \;\leq\; 2\sqrt{\frac{D\log\left(\frac{en}{D}\right)}{n}},$$

where the second inequality follows because, for $n \geq D$, we have $2 \leq \left(\frac{en}{D}\right)$ and the last uses that $D \geq 1$. □

6.3.1 *Binary classification revisited*

Consider again the classification problem discussed in Section 4.1.1. For a fixed family of classifiers $\mathcal{C}$, we study $\sup_{g\in\mathcal{C}} |R_n(g) - R(g)|$, where $R(g) = \mathbb{P}(g(X) \neq Y)$ and $R_n(g) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(g(X_i) \neq Y_i)$. The study of this quantity is well motivated by (4.2) and (4.3). This corresponds to $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$, where $\mathcal{F}$ is taken to be the class of all functions $\mathbb{1}(g(x) \neq y)$ as g varies over $\mathcal{C}$ (this class is uniformly 1-bounded), where the data are (X_i, Y_i) instead of X_i. Using the bounded differences inequality, Proposition 6.2.3, and Proposition 6.3.12, we get that with probability $\geq 1 - \delta$

$$\sup_{g\in\mathcal{C}} |R_n(g) - R(g)| \;=\; 2\sqrt{\frac{\mathsf{VC}(\mathcal{F})}{n}\log\left(\frac{en}{\mathsf{VC}(\mathcal{F})}\right)} + \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)} \;\leq\; 2\sqrt{\mathsf{VC}(\mathcal{F})\frac{\log(n)+1}{n}} + \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}.$$

To make these bounds efficient in a wide variety of examples, we formulate a bunch of results.

Lemma 6.3.13. *In the setting above,* $\mathsf{VC}(\mathcal{F}) \leq \mathsf{VC}(\mathcal{C})$.

Proof. We want to show that if $\mathcal{F}$ can shatter $(x_1, y_1), \ldots, (x_n, y_n)$, then $\mathcal{C}$ can shatter $x_1, \ldots, x_n$. For this, let $\eta_1, \ldots, \eta_n \in \{0, 1\}$. We need to obtain $g \in \mathcal{C}$ such that $g(x_i) = \eta_i$ for $i = 1, \ldots, n$. Define

$$\delta_i \;:=\; \eta_i \mathbb{1}\{y_i = 0\} + (1 - \eta_i)\mathbb{1}(y_i = 1).$$

As $\mathcal{F}$ can shatter $(x_1, y_1), \ldots, (x_n, y_n)$, there exists $f \in \mathcal{F}$, $f(x, y) = \mathbb{1}(g(x) \neq y)$ for some $g \in \mathcal{C}$ such that $f(x_i, y_i) = \delta_i$, $i = 1, \ldots, n$. Then $g(x_i) = \eta_i$, for $i = 1, \ldots, n$. This proves that $\mathcal{C}$ shatters $x_1, \ldots, x_n$. □

When the logistic function is used for this classification task, the classifier is given by

$$g(x) \;=\; \mathbb{1}\{h(x) \geq \tfrac{1}{2}\}, \qquad h(x) = \frac{1}{1 + e^{-x^\top \beta}}$$

or equivalently

$$g(x) \;=\; \mathbb{1}\{x^\top \beta \geq 0\}, \qquad \beta \in \mathbb{R}^d.$$

Proposition 6.3.14 (Finite-Dimensional Vector Spaces)**.** *Let $\mathcal{V}$ be a d-dimensional vector space of real-valued functions on $\mathcal{X}$.*[7] *Let $\mathcal{C} := \{\mathbb{1}(f \geq 0) : f \in \mathcal{V}\}$. Then $\mathsf{VC}(\mathcal{C}) \leq d$.*

[7] For example, $f = x^\intercal\beta$ with β running through $\mathbb{R}^d$.

Proof. For any fixed collection $\{x_1, \cdots, x_{d+1}\}$ of $d+1$ points. Consider $\mathcal{T} = \{(f(x_1), \cdots, f(x_{d+1})) : f \in \mathcal{V}\}$, then $\mathcal{T}$ is a linear subspace of $\mathbb{R}^{d+1}$ with dimension at most d. Therefore, there exists a nonzero vector $y \in \mathbb{R}^{d+1}$ such that

$$\sum_{i=1}^{d+1} y_i f(x_i) = 0, \quad \text{for all } f \in \mathcal{V}. \tag{6.9}$$

Without loss of generality, we assume that at least one y_k is positive, $k \in \{1, 2, \ldots, d+1\}$ (If not, we can let $y' = -y$, then y' satisfies both (6.9) and the assumption). Now suppose there exists $\{x_1, \cdots, x_{d+1}\}$ shattered by $\mathcal{C}$, meaning that each possible sign pattern of $(f(x_1), \ldots, f(x_{d+1}))$ is possible. Then we can find a $f \in \mathcal{V}$ such that

$$\begin{aligned} f(x_i) &\geq 0, \quad \text{if } y_i \leq 0, \\ f(x_i) &< 0, \quad \text{if } y_i > 0. \end{aligned}$$

Thus we get $\sum_{i=1}^{d+1} y_i f(x_i) < 0$, which contradicts with (6.9). □

Suppose now that a logistic classifier is considered and d is the dimension of the input space. Using these results, we conclude that with probability $\geq 1 - \delta$

$$\sup_{g \in \mathcal{C}} |R_n(g) - R(g)| \leq 2\sqrt{\frac{d(\log(n)+1)}{n}} + \sqrt{\frac{2}{n}\log(\frac{1}{\delta})}.$$

6.4 Chaining

The situation like described in Section 6.3 is very special. Typically we do not have that the functions $f \in \mathcal{F}$ all take values in a finite set. In general, to exploit the discretization argument in Proposition 6.2.4 we need to work harder.

Suppose now that $A \subseteq \mathbb{R}^n$ is arbitrary. Let $N = N(\delta)$ be the δ-covering number of A with respect to the norm $\|a\|_n := \|a\|/\sqrt{n}$. A naive approach for bounding $\mathcal{R}_n(A)$ is as follows. Fix a δ-covering $\mathcal{N}$ of A and for every a denote by $\pi(a)$ the closest point in $\mathcal{N}$ to a we get

$$\mathcal{R}_n(A) = \mathbb{E}\sup_{u \in A} \tfrac{1}{n}\left|\varepsilon^\top u\right| \leq \mathbb{E}[\max_{z \in \mathcal{N}} \tfrac{1}{n}|\varepsilon^\top z|] + \mathbb{E}[\sup_{u \in A} \tfrac{1}{n}|\varepsilon^\top (u - \pi(u))|] \leq \sqrt{\frac{2\log(2N(\delta))}{n}} \sup_{u \in A} \frac{\|u\|}{\sqrt{n}} + \delta,$$

where the last inequality follows by Proposition 6.2.4 and the Cauchy-Schwarz inequality. This inequality holds for every δ but $N(\delta)$ monotonically increases as δ decreases.

Get back to our problem of bounding the empirical Rademacher complexity $\mathcal{R}(\mathcal{F}(x_1^n)/n)$. In this section, we only assume that the class $\mathcal{F}$ has a finite and integrable envelope function F: a function such that $|f(x)| \le F(x) < \infty$, for every x and $f \in \mathcal{F}$.

Consider now the $L_2(\mathbb{P})$-norm, $\|f\|_{\mathbb{P},2} = \sqrt{\mathbb{P}|f|^2}$. Given a subset $\mathcal{F}$ of the $L_2(\mathbb{P})$-space, we denote the δ-covering number by $N(\delta, \mathcal{F}, L_2(\mathbb{P}))$. Note that

$$\|f\|_{\mathbb{P}_n,2} = \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(X_i)} \qquad \text{and} \qquad \sup_{f\in\mathcal{F}} \|f\|_{\mathbb{P}_n,2} \le \|F\|_{\mathbb{P}_n,2},$$

which is the same as $\|a\|_n$ with $a = (f(X_1), \dots, f(X_n))$. In other words, $N(\delta)$ for $A = \mathcal{F}(X_1^n)/n$ above is the same as $N(\delta, \mathcal{F}, L_2(\mathbb{P}_n))$. Thus,

$$\mathcal{R}_n(\mathcal{F}(X_1^n)/n) \le \sqrt{\frac{2\log N(\delta, \mathcal{F}, L_2(\mathbb{P}_n))}{n}}\|F\|_{\mathbb{P}_n,2} + \delta,$$

which implies

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}(X_1^n)/n) &\le \left(\sqrt{\frac{2\log N(\delta\|F\|_{\mathbb{P}_n,2}, \mathcal{F}, L_2(\mathbb{P}_n))}{n}} + \delta\right)\|F\|_{\mathbb{P}_n,2} \\ &\le \left(\sqrt{\frac{2\log \sup_Q N(\delta\|F\|_{Q,2}, \mathcal{F}, L_2(Q))}{n}} + \delta\right)\|F\|_{\mathbb{P}_n,2},\end{aligned}$$

where the supremum is taken over all measures supported on a finite set with n points[8]. From this, we can conclude the following result, which we state without proof.

[8] Note that this gives a trivial bound on the Rademacher complexity $\mathbb{E}\mathcal{R}(\mathcal{F}(X_1^n/n))$ because, by Jensen's inequality $\mathbb{E}\|F\|_{\mathbb{P}_n,2} \le \|F\|_{\mathbb{P},2}$ and this is the only term in the bound above that depends on X_1^n.

Theorem 6.4.1. *Let $\mathcal{F}$ be a suitably measurable class of measurable functions with $\sup_Q N(\delta\|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)) < \infty$ for every $\delta > 0$. If $\mathbb{E}F(X) < \infty$ then $\mathcal{F}$ is $\mathbb{P}$-Glivenko-Cantelli.*

A tighter bound on the Rademacher complexity can be obtained using chaining.

Theorem 6.4.2 (Chaining)**.** *Suppose $0 \in A \subseteq \mathbb{R}^n$ and let $D := \sup_{a\in A}\|a\|/\sqrt{n}$. Then*

$$\mathcal{R}_n(A) \le \frac{16}{\sqrt{n}}\int_0^{D/2}\sqrt{\log N(\delta)}\,\mathrm{d}\delta.$$

Proof. Let A_m be a $D/2^m$-covering of A with $|A_m| = N(D/2^m)$. We have $A_0 = \{0\}$. For every $m \ge 0$ let $\pi_m(a)$ denote the point in A_m that is the closest to a. Note that

$$\tfrac{1}{n}\varepsilon^\top a = \sum_{m=0}^{\infty}\tfrac{1}{n}\varepsilon^\top(\pi_{m+1}(a) - \pi_m(a)).$$

By Proposition 6.2.4, we have

$$\begin{aligned}
\mathbb{E}[\max_{a\in A} \tfrac{1}{n}\varepsilon^\top(\pi_{m+1}(a)-\pi_m(a))] &\le \sqrt{\frac{2\log(2|A_m|\cdot|A_{m+1}|)}{n}} \max_{a\in A} \frac{\|\pi_{m+1}(a)-\pi_m(a)\|}{\sqrt{n}} \\
&= \sqrt{\frac{2\log(2|A_{m+1}|^2)}{n}} \max_{a\in A} \frac{\|\pi_{m+1}(a)-a+a-\pi_m(a)\|}{\sqrt{n}} \\
&\le \frac{4}{2^m} D\sqrt{\frac{\log(|A_{m+1}|)}{n}} \\
&\le \frac{4}{2^m} D\sqrt{\frac{\log(N(D/2^{m+1}))}{n}}.
\end{aligned}$$

This assures that

$$\begin{aligned}
\mathbb{E}[\sup_{a\in A} \frac{1}{n}|\varepsilon^\top a|] &\le \sum_{m=0}^{\infty} \mathbb{E}[\max_{a\in A} \tfrac{1}{n}|\varepsilon^\top(\pi_{m+1}(a)-\pi_m(a))|] \\
&\le 4\sum_{m=0}^{\infty} \frac{D}{2^m}\sqrt{\frac{\log N(D/2^{m+1})}{n}} \\
&\le 16\sum_{m=0}^{\infty} \frac{D}{2^{m+2}}\sqrt{\frac{\log N(D/2^{m+1})}{n}} \\
&\le 16\int_0^{D/2} \sqrt{\frac{\log N(\delta)}{n}}\mathrm{d}\delta
\end{aligned}$$

as claimed. □

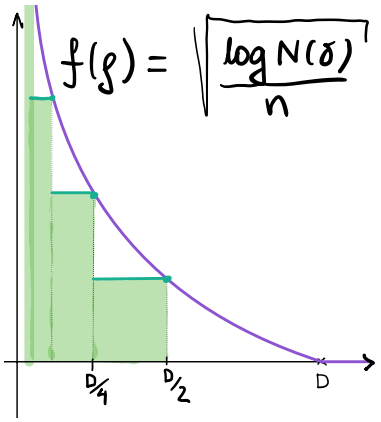


Figure 6.3: Illustration of the last bound in the proof of Theorem 6.4.2.

Theorem 6.4.3. *Suppose there exists a function F such that $|f(x)| \le F(x)$ for all $f \in \mathcal{F}$. Then*

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \le 8\sqrt{\frac{\mathbb{E}(F(X)^2)}{n}} \int_0^1 \sqrt{\log \sup_Q N(\delta\|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))}\mathrm{d}\delta,$$

where the supremum is taken over all measures supported on a finite set with n points.

Proof. Consider a fixed sample $(X_i)_{i=1}^n$ with the underlying distribution $\mathbb{P}_n$. Note that

$$\sup_{a\in\mathcal{F}(X_1^n)} \sqrt{\frac{1}{n}\sum_{i=1}^n a_i^2} = \sup_{f\in\mathcal{F}} \sqrt{\frac{1}{n}\sum_{i=1}^n f(X_i)^2} \le \sqrt{\frac{1}{n}\sum_{i=1}^n F(X_i)^2} = \sqrt{\mathbb{P}_n F^2}$$

By Theorem 6.4.2, conditionally on $(X_i)_{i=1}^n$

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| \Big| (X_i)_{i=1}^n\right] &\le \frac{16}{\sqrt{n}} \int_0^{\sqrt{\mathbb{P}_n F^2}/2} \sqrt{\log N(\delta; \mathcal{F}, L^2(\mathbb{P}_n))}\mathrm{d}\delta \\
&= \frac{8}{\sqrt{n}}\sqrt{\mathbb{P}_n F^2} \int_0^1 \sqrt{\log N(\tfrac{\delta}{2}\sqrt{\mathbb{P}_n F^2}; \mathcal{F}, L^2(\mathbb{P}_n))}\mathrm{d}\delta \\
&\le \frac{8}{\sqrt{n}}\sqrt{\mathbb{P}_n F^2} \int_0^1 \sup_Q \sqrt{\log N(\tfrac{\delta}{2}\|F\|_{L^2(Q)}; \mathcal{F}, L^2(Q))}\mathrm{d}\delta.
\end{aligned}$$

To get the desired bound we now take the expectation with respect to $(X_i)_{i=1}^n$. By the Jensen inequality we get

$$\mathbb{E}\sqrt{\frac{\mathbb{P}_n F^2}{n}} \;\leq\; \sqrt{\frac{\mathbb{E}F^2(X)}{n}},$$

which gives the desired result. □

We conclude stating the following result that shows that Proposition 6.3.12 can be strengthened.

Proposition 6.4.4. *Suppose $\mathcal{F}$ is a Boolean class and $\mathsf{VC}(\mathcal{F}) \leq D$. Then for any $\delta > 0$, we have*

$$\sup_Q N(\delta, \mathcal{F}, L^2(Q)) \;\leq\; (\frac{c_1}{\delta})^{c_2 D},$$

where c_1, c_2 are some positive (universal) constants, and the supremum is over all probability distributions over $\mathcal{X}$. Consequently,

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \;\leq\; c\sqrt{\frac{D}{n}}$$

for some universal constant c.

6.5 Exercises

Exercise 6.5.1. *Let $X \in \mathbb{R}^{n\times d}$ be a random matrix with i.i.d. entries that are σ-sub-Gaussian. Denote by S^{n-1} the unit sphere in $\mathbb{R}^n$ and by $\mathbb{B}_2^n$ the corresponding unit ball.*

(i) Show that $u^T X v$ is a σ-sub-Gaussian random variable for any $u \in \mathbb{B}_2^n$, $v \in \mathbb{B}_2^d$.

(ii) The operator norm $\|X\|$ is defined as $\|X\| := \sup_{v\neq 0} \frac{\|Xv\|}{\|v\|}$. Show that

$$\|X\| \;=\; \max_{u\in S^{n-1}} \max_{v\in S^{d-1}} u^T X v \;=\; \max_{u\in \mathbb{B}_2^n} \max_{v\in \mathbb{B}_2^d} u^T X v.$$

(iii) Using (i) and (ii), show that there exists a constant $C > 0$ such that $\mathbb{E}\|X\| \leq C(\sqrt{n} + \sqrt{d})$.

Exercise 6.5.2. *Prove Proposition 6.2.4.*

7
Applications in statistics (1-2 weeks)

Please read Chapter 1 in [1] to build up some intuition for why high-dimensional statistics is important in general. In here we only focus on a basic illustration for the results in the previous chapter. For more details, see the lecture notes of Philippe Rigollet[2], on which I based most of this chapter.

[1] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge, 2019

[2] https://math.mit.edu/~rigollet/PDFs/RigNotes17.pdf

7.1 Sub-Gaussian sequence model with sparsity

Consider the Gaussian sequence model, that is, let

$$X_i \;=\; \mu_i + \epsilon_i, \qquad i = 1, \dots, d, \tag{7.1}$$

where $\epsilon_i \sim N(0, \sigma^2)$. This is the same model as we considered in Section 2.5 in our discussion of the Stein's paradox. Occasionally, we will also relax the assumption of Gaussianity of ϵ in which case we refer to it a a sub-Gaussian sequence model.

Given a random sample $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^d$ from model (7.1), we can estimate μ by the sample mean $Y = \overline{X}_n = \frac{1}{n}\sum_{j=1}^n X^{(j)}$. Note that $\mathbb{E}Y_i = \mu_i$ and $\mathrm{var}(Y_i) = \sigma^2/n$ and thus, equivalently, we can consider a Gaussian sequence model

$$Y_i \;=\; \mu_i + \epsilon_i, \qquad i = 1, \dots, d,$$

where $\mathrm{var}(\epsilon_i) = \sigma^2/n$ or, more generally, a sub-Gaussian sequence model for which ϵ_i is $\frac{\sigma}{\sqrt{n}}$-sub-Gaussian. We use the estimator $\widehat{\mu} = (Y_1, \dots, Y_d)$ with the risk[3]

$$R(\mu, \widehat{\mu}) \;=\; \mathbb{E}[\sum_{i=1}^d \epsilon_i^2] \;\le\; \frac{\sigma^2 d}{n}.$$

[3] This formula already appeared in (2.11).

The inequality becomes equality in the Gaussian case.

To think about this as a high dimensional problem, we let d, n be both large in which case $\widehat{\mu}$ may not be a good estimator unless $n \gg d$.

In order to estimate μ in the high-dimensional setting we will require some additional structure on μ. In Section 2.5 we saw a simple

example of how explicit bounds on the entries of μ can be exploited by an estimator of the form CY with C diagonal.

Here we assume that μ is sparse or most of its entries are small. In this case, a natural estimator is given by hard thresholding. Hard thresholding gives the estimator

$$\widehat{\mu}_i^\tau \;=\; Y_i \mathbb{1}(|Y_i| \geq \tau) \qquad \text{for all } i = 1, \ldots, d.$$

Equivalently, the thresholding estimator is a solution to the problem:

$$\widehat{\mu}^\tau \;=\; \arg\min_{a \in \mathbb{R}} \left\{ \|Y - a\|^2 + \tau^2 \sum_{i=1}^{d} \mathbb{1}\{a_i \neq 0\} \right\}.$$

In each a_i the function is Y_i^2 if $a_i = 0$ and $(Y_i - a_i)^2 + \tau^2$ otherwise. The latter function is minimized at $a_i = Y_i$ with the optimal value τ^2. Thus the optimum of the whole function depends on whether $|Y_i| \geq \tau$ or not.

We will try to obtain risk bounds for hard thresholding. Note that, by (6.1), with probability at least $1 - \delta$,

$$\max_i |\epsilon_i| \;\leq\; \sigma \sqrt{\frac{2\log(2d/\delta)}{n}} \;=:\; \tau. \tag{7.2}$$

Recall: ϵ_i is $\frac{\sigma}{\sqrt{n}}$-sub-Gaussian.

The consequences of this inequality are two-fold. First, if $\mu_i = 0$ then $|Y_i| = |\epsilon_i| \leq \tau$ with high probability. Thus, if we observe $|Y_i| \gg \tau$ then it must correspond to $\mu_i \neq 0$. Second, if $|Y_i| \leq \tau$ then μ_i cannot be too large because, by the triangle inequality,

$$|\mu_i| \;\leq\; |Y_i| + |\epsilon_i| \;\leq\; 2\tau.$$

Therefore, we loose at most 2τ by taking $\widehat{\mu}_i^\tau = 0$.

Proposition 7.1.1. *Let $\widehat{\mu}^{2\tau}$ be the hard thresholding estimator with threshold 2τ, where τ is defined in (7.2) Then, (7.2) implies that*

$$\|\widehat{\mu}^{2\tau} - \mu\|^2 \;\leq\; 9 \sum_{i=1}^{d} \min\{\mu_i^2, \tau^2\}. \tag{7.3}$$

In particular, (7.3) holds with probability $\geq 1 - \delta$.[4]

[4] Recall that (7.3) depends on δ through (7.2).

Proof. We condition on the high probability event (7.2), which we call $\mathcal{E}$ in this proof. Fix index i and note that:

1. If $|\mu_i| \leq \tau$ then (given $\mathcal{E}$) $|Y_i| \leq |\mu_i| + |Y_i - \mu_i| \leq 2\tau$ and so $\widehat{\mu}_i^{2\tau} = 0$. In this case, $(\mu_i - \widehat{\mu}_i^{2\tau})^2 = \mu_i^2$.
2. If $|\mu_i| > 3\tau$ then (given $\mathcal{E}$) $|Y_i| \geq |\mu_i| - |Y_i - \mu_i| > 2\tau$ and so $\widehat{\mu}_i^{2\tau} = Y_i$. In this case $(\mu_i - \widehat{\mu}_i^{2\tau})^2 = \epsilon_i^2 \leq \tau^2$.
3. If $\tau < |\mu_i| \leq 3\tau$, then

$$(\widehat{\mu}_i^{2\tau} - \mu_i)^2 \;=\; (Y_i \mathbb{1}(|Y_i| \geq 2\tau) - \mu_i)^2 \;=\; \mu_i^2 \mathbb{1}(|Y_i| < 2\tau) + \epsilon_i^2 \mathbb{1}(|Y_i| \geq 2\tau) \;\leq\; \max\{\epsilon_i^2, \mu_i^2\} \;\leq\; 9\tau^2.$$

Putting these together, we see that (7.2) implies (7.3). □

The next theorem shows how this can be exploited. It is convenient to have the following standard notation:

$$a_n \lesssim b_n \qquad \text{means} \qquad a_n \leq C b_n \quad \text{for some } C > 0 \text{ and all } n.$$

Theorem 7.1.2. *Consider the same set-up as in Proposition 7.1.1. If (7.2) holds, then the following statements hold:*

(i) If $\|\mu\|_0 = s$ then

$$R(\mu, \widehat{\mu}^{2\tau}) \;\leq\; 9s\tau^2 \;=\; 18\sigma^2 \frac{s\log(2d/\delta)}{n} \;\lesssim\; \sigma^2 \frac{s\log(d)}{n}. \tag{7.4}$$

Note that we do not bound the risk but the conditional risk. But here the event we condition on holds with high probability $(1-\delta)$ so this is still informative.

(ii) If $\min_{i\in \mathrm{supp}(\mu)} |\mu_i| > 3\tau$, then

$$\mathrm{supp}(\widehat{\mu}^{2\tau}) \;=\; \mathrm{supp}(\mu).$$

Thus we get some guarantees on support recovery.

Equation (7.4) shows that in the sparse setting $\widehat{\mu}^{\tau}$ may be consistent as long as $\frac{s\log(d)}{n} \to 0$, where s is the number of non-zero entries of the vector μ. Quite surprisingly, this may happen even if d is exponentially larger than n.

A natural question is whether sparsity is required for such a high-dimensional consistency result. There are several examples showing that this is not the case. To get a flavour of these results suppose that $\|\mu\|_1 \leq R$ for some $R \in \mathbb{R}$. Let $\tau = \sigma\sqrt{\frac{\log(2d/\delta)}{n}}$ as defined in (7.2). If $\|\mu\|_1 \leq R$, then the number of i such that $|\mu_i| > \tau$ is at most R/τ. By Theorem 7.1.1, under the event $\mathcal{E}$ defined in (7.2),

$$R(\widehat{\mu}^{2\tau}, \mu) \;\leq\; 9\sum_{i=1}^{d} \min\{\mu_i^2, \tau^2\}$$

Thus, we can use the previous (conditional) risk bound to obtain:

$$\begin{aligned}
R(\widehat{\mu}^{2\tau}, \mu) \;&\leq\; 9\sum_{i=1}^{d} \min\{\mu_i^2, \tau^2\} \\
&=\; 9\sum_{i:|\mu_i|>\tau} \tau^2 + 9\sum_{i:|\mu_i|\leq\tau} \mu_i^2 \\
&\leq\; \frac{9R}{\tau}\tau^2 + 9\sum_{i:|\mu_i|\leq\tau} \mu_i^2 \\
&\leq\; 9R\tau + 9\tau \sum_{i:|\mu_i|\leq\tau} |\mu_i| \\
&\leq\; 18R\tau \;=\; 18R\sigma\sqrt{\frac{\log(2d/\delta)}{n}}.
\end{aligned}$$

Notice that the rate of convergence is different from the sparse case, roughly behaving as $1/\sqrt{n}$ instead of $1/n$. Nevertheless, this is still a strong, nice and surprising result.

7.2 *Fixed design linear regression*

Consider the standard linear regression problem

$$\mathbf{y} \;=\; \mathbf{X}\theta^* + \epsilon,$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a fixed matrix, $\theta^* \in \mathbb{R}^d$ is a fixed vector and ϵ is a random vector such that each ϵ_i is an independent random variable with $\mathbb{E}\epsilon_i \;=\; 0$, $\mathrm{var}(\epsilon_i) \;=\; \sigma^2$. Given the data $\mathbf{y}$, the least squares estimator of θ^* is obtained by minimizing

$$M_n(\theta) \;=\; \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\intercal\theta)^2 \;=\; \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\theta\|^2 \tag{7.5}$$

with respect to $\theta \in \mathbb{R}^d$. If $\mathrm{rank}(\mathbf{X}) \;=\; d$ (in part. $n \geq d$) then the optimum exists and is unique. It is given by the well known formula

$$\bar{\theta} \;=\; (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}.$$

If $n < d$ then the optimum is attained over an affine subspace of positive dimension and so it is not unique. It is then customary to fix

$$\bar{\theta} \;=\; (\boldsymbol{X}^\intercal\boldsymbol{X})^{+}\boldsymbol{X}^\intercal\boldsymbol{y},$$

where A^+ is the pseudo-inverse of A. In this case $\bar{\theta}$ minimizes the norm over all the optimal points.

The quality of an estimator $\widehat{\theta}$ is normally analysed in this context based on the mean square error $\mathbb{E}\|\widehat{\theta} - \theta^*\|^2$. For simplicity we will first study the mean squared prediction error

$$\mathrm{MSE}(\mathbf{X}\widehat{\theta}) \;=\; \frac{1}{n}\|\mathbf{X}(\widehat{\theta} - \theta^*)\|^2 \;=\; (\widehat{\theta} - \theta^*)^\intercal(\tfrac{1}{n}\mathbf{X}^\intercal\mathbf{X})(\widehat{\theta} - \theta^*). \tag{7.6}$$

As we will see later, properly choosing the design also ensures that $\|\widehat{\theta} - \theta^*\|^2$ is small as long as $\mathrm{MSE}(\mathbf{X}\widehat{\theta})$ is small.

We are now going to prove our first result on the finite sample performance of the least squares estimator for fixed design.

Theorem 7.2.1. *Assume that the linear model holds with ϵ which is σ-sub-Gaussian. Then the least squares estimator $\bar{\theta}$ satisfies*

$$\mathbb{E}(\mathrm{MSE}(\mathbf{X}\bar{\theta})) \;=\; \frac{1}{n}\mathbb{E}\|\mathbf{X}(\bar{\theta} - \theta^*)\|^2 \;\leq 4\sigma^2\frac{r}{n} \;\lesssim\; \sigma^2\frac{r}{n},$$

where $r = \mathrm{rank}(\mathbf{X}^\intercal\mathbf{X})$. Moreover, for any $\delta \geq 0$, with probability at least $1-\delta$, it holds

$$\mathrm{MSE}(\mathbf{X}\bar{\theta}) \leq 32\sigma^2\frac{2r + \log(1/\delta)}{n}$$

Proof. By definition

$$\|\mathbf{y} - \mathbf{X}\bar{\theta}\|^2 \;\leq\; \|\mathbf{y} - \mathbf{X}\theta^*\|^2 \;=\; \|\epsilon\|^2. \tag{7.7}$$

More generally, we could consider a feature map $\psi : \mathbb{R}^d \to \mathbb{R}^m$ and the corresponding feature matrix $\mathbf{\Psi} \in \mathbb{R}^{n\times d}$ whose rows are obtained by applying ψ to the rows of $\mathbf{X}$. The corresponding model would be

$$\mathbf{y} = \mathbf{\Psi}\theta^* + \epsilon.$$

Moreover, denoting $\bar{\Delta} = \bar{\theta} - \theta^*$,

$$\|\mathbf{y} - \mathbf{X}\bar{\theta}\|^2 \;=\; \|\mathbf{X}\theta^* + \epsilon - \mathbf{X}\bar{\theta}\|^2 \;=\; \|\mathbf{X}\bar{\Delta}\|^2 - 2\epsilon^\intercal \mathbf{X}\bar{\Delta} + \|\epsilon\|^2.$$

This, together with (7.7), allows us to conclude that

$$\|\mathbf{X}\bar{\Delta}\|^2 \;\leq\; 2\epsilon^\intercal \mathbf{X}\bar{\Delta}. \tag{7.8}$$

Using the Cauchy-Schwarz inequality directly for the right-hand side in (7.8) would be wasteful in the case when $r \ll d$. Instead, let $V = [v_1 \cdots v_r] \in \mathbb{R}^{n\times r}$ be a matrix whose r columns form an orthonormal basis of the column span of $\mathbf{X}$. In particular, there exists $\boldsymbol{u}$ such that $\mathbf{X}\bar{\Delta} = V\boldsymbol{u}$ and, denoting $\tilde{\epsilon} = V^\intercal \epsilon$, we have

$$\epsilon^\intercal \mathbf{X}\bar{\Delta} \;=\; \epsilon^\intercal V\boldsymbol{u} \;=\; \tilde{\epsilon}^\intercal \boldsymbol{u} \;\leq\; \|\tilde{\epsilon}\|\|\boldsymbol{u}\| \;=\; \|\tilde{\epsilon}\|\|\mathbf{X}\bar{\Delta}\|,$$

where we used the fact that $\|V\boldsymbol{u}\| = \|\boldsymbol{u}\|$ as $V^T V = I_r$. Using this inequality in (7.8) gives that

$$\|\mathbf{X}\bar{\Delta}\| \;\leq\; 2\|\tilde{\epsilon}\| \tag{7.9}$$

and so

$$\mathbb{E}\|\mathbf{X}(\bar{\theta} - \theta^*)\|^2 \;\leq\; 4\mathbb{E}\|\tilde{\epsilon}\|^2 \;=\; 4\sum_{i=1}^{r} \mathbb{E}\tilde{\epsilon}_i^2 \;\leq\; 4r\sigma^2,$$

where the last inequality follows from the fact that each entry $\tilde{\epsilon}_i = v_i^\intercal \epsilon$ is zero-mean σ-sub-Gaussian and Exercise 5.5.6. This concludes the proof of the bound on $\mathbb{E}(\mathrm{MSE}(\mathbf{X}\bar{\theta}))$.

For the second statement note that, by (7.9),

$$\mathbb{P}(\mathrm{MSE}(\mathbf{X}\bar{\theta}) \geq t) = \mathbb{P}(\|\mathbf{X}(\bar{\theta} - \theta^*)\|^2 \geq nt) \leq \mathbb{P}(\|\tilde{\epsilon}\|^2 \geq nt/4) \;=\; \mathbb{P}(\|\tilde{\epsilon}\| \geq \sqrt{nt}/2).$$

Since $\|\tilde{\epsilon}\| = \sup_{\boldsymbol{u}\in\mathbb{B}^2} \boldsymbol{u}^\intercal \tilde{\epsilon}$ we can use the last inequality in the proof of Theorem 6.1.6 to conclude that

$$\mathbb{P}(\|\tilde{\epsilon}\| \geq \sqrt{nt}/2) \;\leq\; 6^r e^{-\frac{nt}{32\sigma^2}}.$$

Taking $t = 32\sigma^2 \frac{2r+\log(1/\delta)}{n}$ we verify that

$$\mathbb{P}(\mathrm{MSE}(\mathbf{X}\bar{\theta}) \geq t) \;\leq\; \delta.$$

□

If $r = d \leq n$ then bounds on prediction errors give bounds on $\|\bar{\theta} - \theta^*\|^2$. In this case $B = \frac{1}{n}\mathbf{X}^\intercal\mathbf{X}$ has rank d and, using (7.6), we get

$$\|\bar{\theta} - \theta^*\|^2 \;\leq\; \frac{\mathrm{MSE}(\mathbf{X}\bar{\theta})}{\gamma_{\min}(B)}, \tag{7.10}$$

where $\gamma_{\min}(B)$ is the minimal eigenvalue of B. Theorem 7.2.1 can be therefore used to bound $\|\bar{\theta} - \theta^*\|^2$ directly. By contrast, in the high dimensional setting B is not positive definite and we need more structure.

7.3 *Constrained least squares estimator*

Let $K \subseteq \mathbb{R}^d$ be a symmetric convex set[5]. If we knew a priori that $\theta^* \in K$, we may prefer a constrained least squares estimator $\bar{\theta}_K$ defined by

$$\bar{\theta}_K \in \arg\min_{\theta \in K} \|\mathbf{y} - \mathbf{X}\theta\|^2.$$

[5] Symmetric means that $K = -K$.

The equivalent of inequality (7.8) still holds, that is, $\|\mathbf{X}(\bar{\theta}_K - \theta^*)\|^2 \leq 2\epsilon^\top \mathbf{X}(\bar{\theta}_K - \theta^*)$. Further,

$$\|\mathbf{X}(\bar{\theta}_K - \theta^*)\|^2 \;\leq\; 2\epsilon^\top \mathbf{X}(\bar{\theta}_K - \theta^*) \;\leq\; 2 \sup_{\theta \in K-K} \epsilon^\top \mathbf{X}\theta,$$

where $K - K = \{x - y : x, y \in K\}$. It is easy to see that since K is symmetric and convex $K - K = 2K$ so that

← Exercise 7.6.1

$$2 \sup_{\theta \in K-K} \epsilon^\top \mathbf{X}\theta \;=\; 4 \sup_{v \in \mathbf{X}K} \epsilon^\top v$$

where $\mathbf{X}K = \{\mathbf{X}\theta : \theta \in K\} \subseteq \mathbb{R}^n$. This is the measure of the size of $\mathbf{X}K$. If $\epsilon \sim N(0, I_d)$, the expected value of the above supremum is called the **Gaussian width** of $\mathbf{X}K$.

ℓ_1 constrained least squares Assume here that $K = \mathbb{B}_1$ is the unit ℓ_1 ball of $\mathbb{R}^d$. Recall that it has exactly $2d$ vertices $\pm e_1, \ldots, \pm e_d$, where e_i is the i-th canonical unit vector. It implies that the set $\mathbf{X}\mathbb{B}_1$ is also a polytope with at most $2d$ vertices that are contained in the set $\{-\mathbf{X}_1, \mathbf{X}_1, \ldots, \mathbf{X}_d, \mathbf{X}_d\}$, where $\mathbf{X}_i$ is the i-th column of $\mathbf{X}$.

Theorem 7.3.1. *Suppose $\theta^* \in \mathbb{B}_1$. Moreover, assume the conditions of Theorem 7.2.1 and that the columns of $\mathbf{X}$ are normalized so that $\max_i \|\mathbf{X}_i\| \leq \sqrt{n}$. Then the constrained least squares estimator $\bar{\theta}_{\mathbb{B}_1}$ satisfies*

$$\mathbb{E}[\mathrm{MSE}(\mathbf{X}\bar{\theta}_{\mathbb{B}_1})] \;=\; \frac{1}{n}\mathbb{E}\|\mathbf{X}(\bar{\theta}_{\mathbb{B}_1} - \theta^*)\|^2 \;\lesssim\; \sigma\sqrt{\frac{\log d}{n}}.$$

Moreover, for any $\delta \in (0,1)$, with probability $1 - \delta$, it holds

$$\mathrm{MSE}(\mathbf{X}\bar{\theta}_{\mathbb{B}_1}) \;\lesssim\; \sigma\sqrt{\frac{\log(d/\delta)}{n}}.$$

Proof. From the considerations preceding the theorem, we got that

$$\|\mathbf{X}(\bar{\theta}_{\mathbb{B}_1} - \theta^*)\|^2 \;\leq\; 4 \sup_{v \in \mathbf{X}\mathbb{B}_1} \epsilon^\top v.$$

Moreover, because $\mathbf{X}\mathbb{B}_1$ is a polytope, we have

$$\sup_{v \in \mathbf{X}\mathbb{B}_1} \epsilon^\top v \;=\; \max_{i=1,\ldots,d} |\epsilon^\top \mathbf{X}_i|.$$

Since ϵ is σ-sub-Gaussian[6], then for any column $\mathbf{X}_i$ such that $\|\mathbf{X}_i\| \leq$

[6] Recall this multivariate definition given before Theorem 6.1.6.

$\sqrt{n}$, the random variable $\epsilon^\top \mathbf{X}_i$ is $(\sqrt{n}\sigma)$-sub-Gaussian. Therefore, applying Proposition 6.1.1, we get

$$\mathbb{E} \sup_{v \in \mathbf{X}\mathbb{B}_1} \epsilon^\top v \;\leq\; \sigma\sqrt{2n\log(2d)},$$

which gives the claimed bound on $\mathbb{E}[\mathrm{MSE}(\mathbf{X}\bar{\theta}_{\mathbb{B}_1})]$. Again by Proposition 6.1.1 we get that, for any $t > 0$

$$\mathbb{P}(\|\mathbf{X}(\bar{\theta}_{\mathbb{B}_1} - \theta^*)\|^2 \geq 4t) \;\leq\; \mathbb{P}(\sup_{v \in \mathbf{X}\mathbb{B}_1} \epsilon^\top v \geq t) \leq 2de^{-t^2/(2\sigma^2 n)}.$$

From this we conclude[7]

$$\mathbb{P}(\mathrm{MSE}(\mathbf{X}\widehat{\theta}) \geq t) \;\leq\; 2de^{-nt^2/(32\sigma^2)}.$$

[7] Recall from (7.6) that $\mathrm{MSE}(\mathbf{X}\widehat{\theta}) = \frac{1}{n}\|\mathbf{X}(\widehat{\theta} - \theta^*)\|^2$.

To conclude the proof, we find t such that

$$2de^{-nt^2/(32\sigma^2)} \leq \delta \;\Leftrightarrow\; t^2 \geq 32\sigma^2\frac{\log(2d)}{n} + 32\sigma^2\frac{\log(1/\delta)}{n}.$$

□

Note that the proof of Theorem 7.2.1 also applies to $\bar{\theta}_{\mathbb{B}_1}$ so that $\bar{\theta}_{\mathbb{B}_1}$ benefits from the best of both rates

← Exercise 7.6.2

$$\mathbb{E}[\mathrm{MSE}(\mathbf{X}\bar{\theta}_{\mathbb{B}_1})] \;\lesssim\; \min\{\tfrac{r}{n}, \sqrt{\tfrac{\log d}{n}}\}.$$

This is called an elbow effect. The elbow takes place around $r \simeq \sqrt{n\log d}$.

7.4 *LASSO regression*

The focus of this section is on the situation when the true parameter vector θ^* is sparse. In this case, even if $n < d$ it may be possible to control the error of recovering θ^*. Building upon (7.5), we consider the regularized estimator

$$\widehat{\theta} \;:=\; \arg\min_{\theta \in \mathbb{R}^d} \left\{ \tfrac{1}{2n}\|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda\|\theta\|_1 \right\}, \tag{7.11}$$

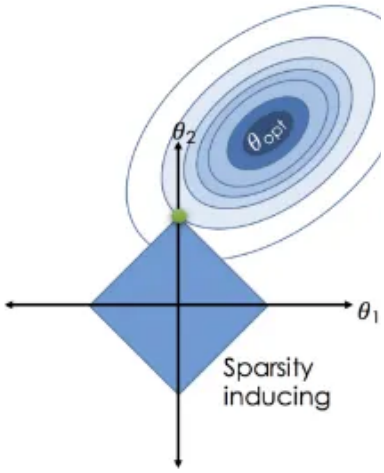
where $\lambda \geq 0$ is a fixed penalty parameter. This is called the LASSO regression problem.

The LASSO regression problem is a convex problem but it is not differentiable. By standard convex duality theory assures that, for any given $\lambda \geq 0$, the LASSO problem (7.11) is equivalent to the constrained problem

$$\text{minimize} \quad \tfrac{1}{2n}\|\mathbf{y} - \mathbf{X}\theta\|^2 \quad \text{subject to } \|\theta\|_1 \leq R \tag{7.12}$$

for some $R \geq 0$. This observation aids some intuition behind what the LASSO estimator is actually doing (contemplate the side figure).

Although the optimum is not given in a closed form, there is a simple numerical algorithm that can be used for optimization. The algorithm relies on the observation that if $d = 1$ then the optimum is given in a closed form. Using this fact, we can run a coordinate descent algorithm where at each step we update one coordinate of θ keeping the other coordinates fixed.



There are three main problems related with the analysis of the LASSO estimator. One focuses on $\|\widehat{\theta} - \theta^*\|$ to establish high-dimensional consistency. The other focuses on the prediction performance $\|\mathbf{X}(\widehat{\theta} - \theta^*)\|$ as analysed for the least squares estimator in the previous section. Finally, we can study how the LASSO regression in recovering the true support of θ^*. Below we briefly focus on the first and the last.

7.4.1 *High-dimensional consistency*

Recall that $M_n(\theta) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\theta\|^2$. In the discussion surrounding (7.10) we argued that if the spectrum of $\frac{1}{n}\mathbf{X}^\intercal\mathbf{X}$ (and so the Hessian of $M_n(\theta)$) is bounded away from zero, a bound on the predictive error provide bounds on $\mathrm{MSE}(\bar{\theta})$. Considering the dual problem (7.12) with R satisfying $\|\theta^*\|_1 \leq R$, we could conclude something similar for the LASSO estimator. In the high-dimensional setting $\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}$ is not positive definite. However, using the fact that θ^* is sparse, we need enough of curvature of M_n only in some directions.

The Hessian matrix $\nabla^2 M_n(\theta) = \frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}$ is positive definite with minimal eigenvalue at least κ if

$$\frac{1}{n}\Delta^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\Delta = \frac{1}{n}\|\boldsymbol{X}\Delta\|^2 \geq \kappa\|\Delta\|^2 \qquad \text{for all } \Delta \in \mathbb{R}^d.$$

As we said, if $n < d$, there is no $\kappa > 0$ for which this condition holds. Instead, for any $S \subseteq \{1, \ldots, d\}$ and $\alpha > 0$ define

$$\mathbb{C}_\alpha(S) := \{\Delta \in \mathbb{R}^d : \|\Delta\|_1 \leq (1+\alpha)\|\Delta_S\|_1\}.$$

Large portion of $\|\Delta\|_1$ is due to Δ_S. In other words $\|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1$

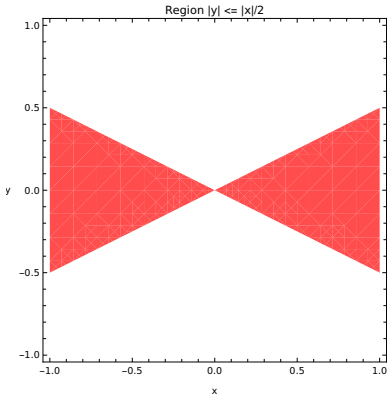


Figure 7.1: Suppose that $d = 2$, $\theta = (x, y)$, $S = \{1\}$, and $\alpha = \frac{1}{4}$. Then $\mathbb{C}_\alpha(S) = \{(x, y) : |y| \leq \frac{1}{2}|x|\}$.

We say that the matrix $\boldsymbol{X}$ satisfies the **restricted eigenvalue (RE)** condition over S with parameters (κ, α) if

$$\frac{1}{n}\|\boldsymbol{X}\Delta\|^2 \geq \kappa\|\Delta\|^2 \qquad \text{for all } \Delta \in \mathbb{C}_\alpha(S).$$

Suppose that:

(A1) The vector θ^* is supported on a subset $S \subseteq \{1, \ldots, d\}$ with $|S| = s$.

(A2) The design matrix satisfies the restricted eigenvalue condition over S with parameters $(\kappa, 3)$.

Theorem 7.4.1. *Under assumptions (A1) and (A2) any solution of* (7.11) *for* $\lambda_n \geq 2\|\frac{\boldsymbol{X}^\intercal\epsilon}{n}\|_\infty$ *we have*

$$\|\widehat{\theta}_n - \theta^*\| \leq \frac{3}{\kappa}\sqrt{s}\lambda_n.$$

Proof. We first show that, if $\lambda_n \geq 2\|\frac{X^\intercal \epsilon}{n}\|_\infty$ then the error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ belongs to $\mathbb{C}_3(S)$. Let $L(\theta; \lambda_n) = \frac{1}{2n}\|\mathbf{y} - \boldsymbol{X}\theta\|^2 + \lambda_n\|\theta\|_1$. We have

$$L(\widehat{\theta}; \lambda_n) \;\leq\; L(\theta^*; \lambda_n) \;=\; \frac{1}{2n}\|\epsilon\|^2 + \lambda_n\|\theta^*\|_1.$$

With $\widehat{\Delta} = \widehat{\theta} - \theta^*$ we get

$$\|\boldsymbol{y} - \boldsymbol{X}\widehat{\theta}\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\theta^* - \boldsymbol{X}\widehat{\Delta}\|^2 = \|\epsilon - \boldsymbol{X}\widehat{\Delta}\|^2.$$

We can use it in the previous expression to conclude

$$0 \;\leq\; \frac{1}{2n}\|\boldsymbol{X}\widehat{\Delta}\|^2 \;\leq\; \frac{1}{n}\epsilon^\intercal \boldsymbol{X}\widehat{\Delta} + \lambda_n\left(\|\theta^*\|_1 - \|\widehat{\theta}\|_1\right). \tag{7.13}$$

Since θ^* is S-sparse, we can write

$$\|\theta^*\|_1 - \|\widehat{\theta}\|_1 \;=\; \|\theta^*_S\|_1 - \|\theta^*_S + \widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1.$$

Substituting this into (7.13) gives

$$\begin{aligned} 0 \;&\leq\; \frac{1}{n}\|\boldsymbol{X}\widehat{\Delta}\|^2 \;\leq\; \frac{2}{n}\epsilon^\intercal \boldsymbol{X}\widehat{\Delta} + 2\lambda_n\left(\|\theta^*_S\|_1 - \|\theta^*_S + \widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\right) \\ &\leq\; 2\left\|\frac{\boldsymbol{X}^\intercal\epsilon}{n}\right\|_\infty \|\widehat{\Delta}\|_1 + 2\lambda_n\left(\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\right) \\ &\leq\; \lambda_n\left(3\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\right), \end{aligned}$$

where the last inequality follows from the choice of λ_n. The fact that $3\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1 \;\geq\; 0$ establishes that $\widehat{\Delta} \;\in\; \mathbb{C}_3(S)$ so that the RE condition can be applied. Doing so, we conclude that

$$\kappa\|\widehat{\Delta}\|^2 \;\leq\; \frac{1}{n}\|\mathbf{X}\widehat{\Delta}\|^2 \;\leq\; \lambda_n\left(3\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\right) \;\leq\; 3\lambda_n\|\widehat{\Delta}_S\|_1.$$

Since $\|\widehat{\Delta}_S\|_1 \leq \sqrt{s}\|\widehat{\Delta}_S\| \leq \sqrt{s}\|\widehat{\Delta}\|$, the conclusion follows. □

We now show how this result can be applied in the classical linear Gaussian model for which the noise vector ϵ has i.i.d. $N(0, \sigma^2)$ entries. More generally, the same calculation applies when ϵ is σ-sub-Gaussian.

In addition, we assume that $\mathbf{X}$ satisfies the RE condition and that it is C-column normalized, meaning that[8]

$$\max_j \|\boldsymbol{X}_j\| \leq C\sqrt{n},$$

where $\boldsymbol{X}_j$ denotes the columns of $\mathbf{X}$. With this set-up, the random variable $\|\frac{1}{n}\mathbf{X}^\intercal\epsilon\|_\infty$ corresponds to the absolute maximum of d zero mean Gaussian variables, each with variance at most $C^2\sigma^2/n$. Consequently from standard sub-Gaussian tail bounds in Proposition 6.1.1

$$\mathbb{P}\left(\|\tfrac{1}{n}\mathbf{X}^\intercal\epsilon\|_\infty \geq t\right) \;\leq\; 2de^{-\frac{nt^2}{2C^2\sigma^2}}.$$

[8] Exercise 7.6.3 partially motivates this assumption.
← Exercise 7.6.3

Plugging, $t = C\sigma\left(\sqrt{\frac{2\log d}{n}} + \delta\right)$ we easily verify that

$$\mathbb{P}\left(\|\tfrac{1}{n}\mathbf{X}^\intercal\epsilon\|_\infty \geq C\sigma\left(\sqrt{\frac{2\log d}{n}} + \delta\right)\right) \;\leq\; 2e^{-n\delta^2/2} \qquad \text{for all } \delta > 0.$$

Consequently, if we set

$$\lambda_n = 2C\sigma(\sqrt{\frac{2\log d}{n}} + \delta)$$

then Theorem 7.4.1 implies that

$$\|\widehat{\theta} - \theta^*\| \;\leq\; \frac{6C\sigma}{\kappa}\sqrt{s}\left\{\sqrt{\frac{2\log d}{n}} + \delta\right\}$$

with probability at least $1 - 2e^{-n\delta^2/2}$.

7.4.2 *Model recovery consistency*

Here we again focus on the deterministic design. For variable selection consistency the restricted eigenvalue condition is replaced by a closely related condition.

(A3) **Lower eigenvalue**: The smallest eigenvalue of the sample covariance submatrix indexed by S is bounded below

$$\gamma_{\min}(\frac{1}{n}\mathbf{X}_S^\intercal\mathbf{X}_S) \;\geq\; c_{\min} \;>\; 0.$$

(A4) **Mutual incoherence**: There exists $\alpha \in [0, 1)$ such that

$$\max_{j\in S^c} \|(\mathbf{X}_S^\intercal\mathbf{X}_S)^{-1}\mathbf{X}_S^\intercal X_j\|_1 \;\leq\; \alpha.$$

Note that condition (A3) is rather mild and it is required to get identitifiability even if S is known in advance. The second condition is more subtle and roughly it says that no variables in S^c are too correlated with the support variables. This is form of orthogonality ($(\mathbf{X}_S^\intercal\mathbf{X}_S)^{-1}\mathbf{X}_S^\intercal X_j$ is a projection of X_j on the span of $\mathbf{X}_S$), which is actually unlikely to hold in big datasets.

Recall that for a vector $w \in \mathbb{R}^d$, $\|w\|_\infty = \max_i |w_i|$. If $W \in \mathbb{R}^{k\times l}$ is a matrix then $\|W\|_\infty = \max_{i=1,\ldots,k} \|W_i\|_1$, where W_i is the i-th row of W.

Theorem 7.4.2. *Consider an S-sparse linear regression model for which the design matrix satisfies conditions (A3) and (A4). If*

$$\lambda_n \;\geq\; \frac{2}{1-\alpha}\left\|\frac{1}{n}\mathbf{X}_{S^c}^\intercal\left(I_n - \mathbf{X}_S(\mathbf{X}_S^\intercal\mathbf{X}_S)^{-1}\mathbf{X}_S^\intercal\right)\epsilon\right\|_\infty,$$

then $\widehat{\theta}$ has the following properties:

(a) Uniqueness: There is a unique optimal solution $\widehat{\theta}$.

(b) No false inclusion: This solution has its support set $\widehat{S}$ contained within the true support set S.

(c) ℓ_∞-bounds: The error $\widehat{\Delta}$ satisfies

$$\|\widehat{\Delta}_S\|_\infty \leq \|(\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1}\mathbf{X}_S^\intercal \epsilon\|_\infty + \|(\tfrac{1}{n}\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1}\|_\infty \lambda_n =: B(\lambda_n, \mathbf{X}).$$

(d) No false exclusion: The lasso includes all $i \in S$ such that $|\theta_i^*| > B(\lambda_n, \mathbf{X})$, and hence is variable selection consistent if $\min_{i\in S}|\theta_i^*| > B(\lambda_n, \mathbf{X})$.

Corollary 7.4.3. *Consider the S-sparse linear model based on a noise vector ϵ with zero-mean i.i.d. σ-sub-Gaussian entries, and a deterministic design matrix $\mathbf{X}$ that satisfies (A3) and (A4), as well as the C-column normalization condition. The LASSO estimator with*

$$\lambda_n = \frac{2C\sigma}{1-\alpha}\left\{\sqrt{\frac{2\log(d-s)}{n}} + \delta\right\}$$

for some $\delta > 0$. Then $\widehat{\theta}$ is unique with its support contained within S (no type I errors), and satisfies the ℓ_∞-error bound

$$\|\widehat{\Delta}_S\|_\infty \;\leq\; \frac{\sigma}{\sqrt{c_{\min}}}\left\{\sqrt{\frac{2\log s}{n}} + \delta\right\} + \|(\tfrac{1}{n}\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1}\|_\infty \lambda_n$$

all with probability at least $1 - 4e^{-n\delta^2/2}$.

Both results are left without a proof, see Chapter 7 in [9] for details.

[9] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge, 2019

7.5 Random matrices

7.5.1 Spectral norm of sub-Gaussian random matrices

Let $A \in \mathbb{R}^{m\times n}$, then $\|A\| = \max_{\|\mathbf{x}\|\leq 1}\|A\mathbf{x}\|$ is called the operator norm of A. Using the variational characterization of the norm, we get

$$\|A\| = \max_{\|\mathbf{x}\|\leq 1, \|\mathbf{y}\|\leq 1} \mathbf{y}^\top A\mathbf{x}.$$

Lemma 7.5.1. *Let $\mathcal{N}$ be an ϵ-covering of $\mathbb{B}_2^n$. Then*

$$\max_{\mathbf{x}\in\mathcal{N}}\|A\mathbf{x}\| \;\leq\; \|A\| \;\leq\; \frac{1}{1-\epsilon}\max_{\mathbf{x}\in\mathcal{N}}\|A\mathbf{x}\|.$$

If, in addition, $\mathcal{M}$ is an ϵ-covering of $\mathbb{B}_2^m$ then

$$\max_{\mathbf{x}\in\mathcal{N},\mathbf{y}\in\mathcal{M}} \mathbf{y}^\top A\mathbf{x} \;\leq\; \|A\| \;\leq\; \frac{1}{1-2\epsilon}\max_{\mathbf{x}\in\mathcal{N},\mathbf{y}\in\mathcal{M}} \mathbf{y}^\top A\mathbf{x}.$$

Proof. For the first statement – we used this argument already in the proof of Theorem 6.1.6. For the second statement, use the fact that $\|A\mathbf{x}\| = \max_{\|\mathbf{y}\|\leq 1} \mathbf{y}^\top A\mathbf{x}$ and that $(1-\epsilon)^2 \geq 1-2\epsilon$. □

Proposition 7.5.2. *Let $A \in \mathbb{R}^{m\times n}$ be a random matrix with independent, mean zero, σ-sub-Gaussian entries. Then, for all $t \geq 0$*

$$\|A\| \;\leq\; 5\sigma(\sqrt{m}+\sqrt{n}+t)$$

with probability $\geq 1-e^{-t^2}$.

Proof. First, consider $\frac{1}{4}$-coverings $\mathcal{M}, \mathcal{N}$ of $\mathbb{B}_2^m$ and $\mathbb{B}_2^n$. By Lemma 6.1.4, we can assume $|\mathcal{M}| \leq 12^m$ and $|\mathcal{N}| \leq 12^n$. By Lemma 7.5.1, $\|A\| \leq 2\max_{\mathbf{y}\in\mathcal{M},\mathbf{x}\in\mathcal{N}} \mathbf{y}^\top A\mathbf{x}$. By Lemma 5.2.4, each $\mathbf{y}^\top A\mathbf{x}$ is σ-sub-Gaussian. By the union bound we conclude

$$\mathbb{P}(\|A\| \geq u) \;\leq\; \mathbb{P}\Big(\max_{\mathbf{y}\in\mathcal{M},\mathbf{x}\in\mathcal{N}} \mathbf{y}^\top A\mathbf{x} \geq \tfrac{u}{2}\Big) \;\leq\; 12^{n+m} e^{-\frac{u^2}{8\sigma^2}}.$$

Take $u^* = 5\sigma(\sqrt{m}+\sqrt{n}+t)$. Then

$$\mathbb{P}(\|A\| \geq u^*) \leq 12^{m+n} e^{-\frac{25}{8}(m+n+t^2)} \;\leq\; e^{-t^2}.$$

□

We note also that $\mathbb{E}\|A\| \leq 5\sigma(\sqrt{m}+\sqrt{n})$. This follows from Proposition 6.1.1. Indeed,

$$\mathbb{E}\|A\| \;\leq\; 2\mathbb{E}\Big[\max_{\mathbf{y}\in\mathcal{M},\mathbf{x}\in\mathcal{N}} \mathbf{y}^\top A\mathbf{x}\Big] \;\leq\; 2\sigma\sqrt{2\log 12^{m+n}} \;\leq\; 5\sigma\sqrt{m+n},$$

from which the conclusion follows.

7.5.2 *Recovering communities in the stochastic block model*

This section is adapted from:

Roman Vershynin. *High-dimensional probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018

We consider a basic version of the stochastic block model. This is a model for random graphs on n nodes, which are divided into two equal-sized communities. The model comes with two parameters $0 < q < p < 1$. Each edge $i-j$ appears independently with probability p if i, j are in the same community and q if they are in two different communities.

We represent a graph with an adjacency matrix $A \in \{0,1\}^{n\times n}$, where

$$\mathbb{P}(A_{ij}=1) \;=\; \mathbb{E}A_{ij} \;=\; \begin{cases} p & \text{if } i,j \text{ lie in the same community,} \\ q & \text{if } i,j \text{ lie in different communities.} \end{cases}$$

If the vertices are ordered so that the first $n/2$ belong to the first community then $\mathbb{E}A$ has a simple block structure

$$\mathbb{E}A \;=\; \begin{bmatrix} p\mathbf{1}\mathbf{1}^\top & q\mathbf{1}\mathbf{1}^\top \\ q\mathbf{1}\mathbf{1}^\top & p\mathbf{1}\mathbf{1}^\top \end{bmatrix} \;=\; \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} p & q \\ q & p \end{bmatrix} \begin{bmatrix} \mathbf{1}^\top & \mathbf{0}^\top \\ \mathbf{0}^\top & \mathbf{1}^\top \end{bmatrix}$$

where $\mathbf{0}, \mathbf{1} \in \mathbb{R}^{n/2}$ are vector of zeros and ones. Denote

$$u_1 = \begin{bmatrix} \mathbf{1} \\ \mathbf{1} \end{bmatrix}, \qquad u_2 = \begin{bmatrix} \mathbf{1} \\ -\mathbf{1} \end{bmatrix}, \qquad v_1 = \tfrac{1}{\sqrt{n}} u_1, \qquad v_2 = \tfrac{1}{\sqrt{n}} u_2,$$

where v_1, v_2 are the normalized versions of u_1, u_2 with $\|v_1\| = \|v_2\| = 1$. It is then easy to check that

$$\mathbb{E}Av_1 \;=\; n\tfrac{p+q}{2} v_1 \qquad \text{and} \qquad \mathbb{E}Av_2 \;=\; n\tfrac{p-q}{2} v_2.$$

In other words, u_1, u_2 are the eigenvectors of $\mathbb{E}A$ with eigenvalues $\lambda_1 = n\frac{p+q}{2}$ and $\lambda_2 = n\frac{p-q}{2}$ respectively. Since $\mathbb{E}A$ has rank 2, all other eigenvalues are zero.

Consider now situation when we observe A but the community structure is unknown. If we observed $\mathbb{E}A$, we could compute the eigenvector corresponding to the second largest eigenvalue and then we could assign vertices to communities according to whether the corresponding entry in this eigenvector is positive or negative. A natural question is what happens if we do the same on the observed matrix A.

By the Weyl's theorem we know that for any two symmetric matrices $\max_i |\lambda_i(S) - \lambda_i(T)| \;\leq\; \|S - T\|$. We have a similar result for eigenvectors.

Theorem 7.5.3 (Davis-Kahan)**.** *If S, T are symmetric $n \times n$ matrices. Fix i and suppose $\min_{j \neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0$ then*

$$\sin\left(\angle\{v_i(S), v_i(T)\}\right) \;\leq\; \frac{2\|S - T\|}{\delta}.$$

Note that the sine of an angle being small means that the angle between $v_i(S)$ and $v_i(T)$ is either close to 0 or close to π. We can always find $\theta \in \{-1, 1\}$ such that the angle between $v_i(S)$ and $\theta v_i(T)$ is close to zero.

Proposition 7.5.4. *With the same assumption as in Theorem 7.5.3, there exists $\theta \in \{-1, 1\}$ such that*

$$\|v_i(S) - \theta v_i(T)\| \;\leq\; \frac{2^{3/2}\|S - T\|}{\delta}.$$

Proof. Let $u = v_i(S)$, $v = \theta v_i(T)$, where θ is fixed so that $u^\top v \geq 0$. Note that $\|u\| = \|v\| = 1$ and $\|u - v\|^2 = 2(1 - u^\top v)$. Moreover, using the formula $\cos(\angle(u, v)) = u^\top v$ we get

$$\sin^2(\angle(u, v)) \;=\; 1 - (u^\top v)^2 \;\geq\; 1 - u^\top v \;=\; \tfrac{1}{2}\|u - v\|^2.$$

By Theorem 7.5.3, we then conclude

$$\|u - v\|^2 \;\leq\; 2\sin^2(\angle(u, v)) \;\leq\; \frac{8\|S - T\|}{\delta^2},$$

from which the result follows. □

We are going to use this result with $S = \mathbb{E}A$ and $T = A$. To use the David-Kahan theorem, we first check that λ_2 (the second eigenvalue of $\mathbb{E}A$) is well-separated from the rest of the spectrum, that is, from λ_1 and 0. We have

$$\delta \;=\; \min\{\lambda_1 - \lambda_2, \lambda_2 - 0\} \;=\; n\min\{\tfrac{p-q}{2}, q\} \;=:\; n\mu.$$

In SBM A_{ij}'s are independent. Since $|A_{ij} - \mathbb{E}A_{ij}| \le 1$, it follows that $A_{ij} - \mathbb{E}A_{ij}$ are independent 1-sub-Gaussian random variables. It follows that

$$\|A - \mathbb{E}A\| \;\le\; 10(\sqrt{n} + t)$$

with probability $\ge 1 - e^{-t^2}$. By Proposition 7.5.4, there exists $\theta \in \{-1, 1\}$ such that

$$\|v_i(\mathbb{E}A) - \theta v_i(A)\| \;\le\; \frac{2^{3/2}\|A - \mathbb{E}A\|}{\delta} \;\le\; \frac{5 \cdot 2^{5/2}}{\mu}\left(\frac{t}{n} + \frac{1}{\sqrt{n}}\right)$$

with probability $\ge 1 - e^{-t^2}$. Take $t = \sqrt{n}$ to conclude that

$$\|v_i(\mathbb{E}A) - \theta v_i(A)\| \;\le\; \frac{C}{\mu\sqrt{n}} \qquad (C := 5 \cdot 2^{7/2})$$

with probability $\ge 1 - e^{-n}$.

Let $u_i \;=\; \sqrt{n}v_i$. Note that the entries of $u_i(\mathbb{E}A)$ are ± 1, so if signs of $u_i(\mathbb{E}A)$ and $\theta u_i(A)$ do not agree, they constribute $>\; 1$ to $\|u_i(\mathbb{E}A) - u_i(A)\|^2$. We know that, with high probability,

$$\|u_i(\mathbb{E}A) - \theta u_i(A)\|^2 \;\le\; \frac{C^2}{\mu^2}$$

and so there cannot be more than $\frac{C^2}{\mu^2}$ (a constant) of entries that contribute more than 1!

We conclude that, with high probability, we can correctly classify all but finite number of vertices. We formulate this as a theorem.

Theorem 7.5.5 (Spectral clustering for SBMs)**.** *Let $A \sim SBM(n, p, q)$ with $p > q$ and* $\min\{tfracp - q2, q\} =: \mu > 0$*. Then with probability $\ge 1 - e^{-n}$ the spectral clustering algorithm identifies communities of A correctly up to C^2/μ^2 misspecified vertices.*

7.5.3 *Covariance matrix estimation*

The basic problem can be formulated as follows. Suppose $\mathbf{X} \in \mathbb{R}^{n \times d}$ such that each row $\mathbf{x}_i$ is i.i.d. $N_d(0, \Sigma)$. The sample covariance matrix is defined as

$$\widehat{\Sigma}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\intercal \;=\; \frac{1}{n}\mathbf{X}^\intercal\mathbf{X}.$$

This is a random positive semi-definite matrix whose expectation satisfies

$$\mathbb{E}\widehat{\Sigma}_n \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\intercal \;=\; \Sigma.$$

The question we ask is how $\widehat{\Sigma}_n$ concentrates around Σ. Ideally, we would also like to relax the Gaussianity assumption.

Concentration can be measured in various norms. If we use the operator norm, the general idea is to use the following variational representation: If $Q \in \mathbb{S}^d$ then

$$\|Q\| \;=\; \sup_{\|v\|=1} |v^\intercal Q v|.$$

Thus

$$\|\widehat{\Sigma}_n - \Sigma\| \;=\; \sup_{\|v\|=1} |v^\intercal \widehat{\Sigma}_n v - v^\intercal \Sigma v| \;=\; \sup_{\|v\|=1} \left| \frac{1}{n}\sum_{i=1}^{n}(v^\intercal \mathbf{x}_i)^2 - v^\intercal \Sigma v \right|.$$

Let $\mathcal{F}$ be the class of functions $f(\mathbf{x}) \;=\; (v^\intercal \mathbf{x})^2$ for $\|v\| \;=\; 1$. Since $\mathbb{E}f(\mathbf{x}) = v^T \Sigma v$, we can rewrite

$$\|\widehat{\Sigma}_n - \Sigma\| \;=\; \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \mathbb{E}f(X) \right|.$$

Note that the spectrum of $\widehat{\Sigma}_n$ is directly related to the spectrum of $\mathbf{X}$. We have

$$\gamma_{\max}(\widehat{\Sigma}_n) = \gamma_{\max}(\tfrac{1}{n}\mathbf{X}^\intercal \mathbf{X}) = (\sigma_{\max}(\tfrac{1}{\sqrt{n}}\mathbf{X}))^2$$

and

$$\gamma_{\min}(\widehat{\Sigma}_n) = \gamma_{\min}(\tfrac{1}{n}\mathbf{X}^\intercal \mathbf{X}) = (\sigma_{\min}(\tfrac{1}{\sqrt{n}}\mathbf{X}))^2.$$

Since singular values are continuous functions of its matrix argument, we then expect that

$$\sigma_{\max}(\tfrac{1}{\sqrt{n}}\mathbf{X}) = \sqrt{\gamma_{\min}(\widehat{\Sigma}_n)} \approx \sqrt{\gamma_{\min}(\Sigma)} = \gamma_{\min}(\sqrt{\Sigma}).$$

Indeed, we have the following result.

Theorem 7.5.6 (The Gaussian case)**.** *Suppose $\mathbf{X} \in \mathbb{R}^{n\times d}$ such that each row $\mathbf{x}_i$ is i.i.d. from $N_d(0,\Sigma)$. Then for every $\delta > 0$*

$$\mathbb{P}\left(\sigma_{\max}(\tfrac{1}{\sqrt{n}}\mathbf{X})) \;\geq\; (1+\delta)\gamma_{\max}(\sqrt{\Sigma}) + \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} \right) \;\leq\; e^{-n\delta^2/2}$$

and, if $n \geq d$, we also have

$$\mathbb{P}\left(\sigma_{\min}(\tfrac{1}{\sqrt{n}}\mathbf{X}) \;\leq\; (1-\delta)\gamma_{\min}(\sqrt{\Sigma}) - \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} \right) \;\leq\; e^{-n\delta^2/2}$$

Proof. The proof, as many results of that form has two step. First we show that $\sigma_{\max}(\frac{1}{\sqrt{n}}\mathbf{X}))$ concentrates around its expectation. Second, we provide bounds on this expectation that allow us to conclude the given formulas.

Step I: We can write $\mathbf{X} = \boldsymbol{W}\sqrt{\Sigma}$, where $\boldsymbol{W}$ has i.i.d. standard normal rows. The first part can be shown as in Example 5.4.4. By Weyl's theorem $\boldsymbol{W} \mapsto \sigma_{\max}(\frac{1}{\sqrt{n}}\boldsymbol{W}\sqrt{\Sigma})$ is Lipschitz with parameter $L = \frac{1}{\sqrt{n}}\|\sqrt{\Sigma}\|$. By Theorem 5.4.1

$$\mathbb{P}(\sigma_{\max}(\tfrac{1}{\sqrt{n}}\mathbf{X}) \geq \mathbb{E}\sigma_{\max}(\tfrac{1}{\sqrt{n}}\mathbf{X})) + t) \;\leq\; e^{-nt^2/(2\|\Sigma\|)}.$$

Taking $t = \sqrt{\|\Sigma\|}\delta$,

$$\mathbb{P}(\sigma_{\max}(\tfrac{1}{\sqrt{n}}\mathbf{X}) \geq \mathbb{E}\sigma_{\max}(\tfrac{1}{\sqrt{n}}\mathbf{X})) + \sqrt{\|\Sigma\|}\delta) \;\leq\; e^{-n\delta^2/2}.$$

Preparing for the second part of the proof, note that if we could show that $\mathbb{E}(\sigma_{\max}(\frac{1}{\sqrt{n}}\mathbf{X})) \leq \sqrt{\|\Sigma\|} + \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}$, we would be done.

Step II: We want to show that $\mathbb{E}(\sigma_{\max}(\frac{1}{\sqrt{n}}\mathbf{X})) \leq \sqrt{\|\Sigma\|} + \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}$. We use the following variational characterization

$$\sigma_{\max}(\tfrac{1}{\sqrt{n}}\mathbf{X}) \;=\; \max_{\|\boldsymbol{u}\|=1}\max_{\|\boldsymbol{v}\|=1} \boldsymbol{u}^\intercal(\tfrac{1}{\sqrt{n}}\mathbf{X})\boldsymbol{v}.$$

Let $S^{n-1} \subset \mathbb{R}^n$ be the set of vectors with unit norm. Consider the zero-mean Gaussian process

$$Z_{\boldsymbol{u},\boldsymbol{v}} \;:=\; \boldsymbol{u}^\intercal(\tfrac{1}{\sqrt{n}}\mathbf{X})\boldsymbol{v} \qquad \text{for } (\boldsymbol{u},\boldsymbol{v}) \in S^{n-1} \times S^{d-1}.$$

Consider the induced metric on $S^{n-1} \times S^{d-1}$

$$d((\boldsymbol{u},\boldsymbol{v}),(\boldsymbol{u}',\boldsymbol{v}')) \;:=\; \sqrt{\mathbb{E}(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\boldsymbol{u}',\boldsymbol{v}'})^2}.$$

We have

$$\mathbb{E}(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\boldsymbol{u}',\boldsymbol{v}'})^2 \leq \|\Sigma\|\|\boldsymbol{u}-\boldsymbol{u}'\|^2 + \|\boldsymbol{v}-\boldsymbol{v}'\|_\Sigma^2.$$

We construct another Gaussian process $Y_{\boldsymbol{u},\boldsymbol{v}}$, whose covariance function is equal to this right-hand side. □

7.6 Exercises

Exercise 7.6.1. *Show that if $K \subset \mathbb{R}^d$ is convex and symmetric ($K = -K$) then $K - K = \{x - y : x, y \in K\}$ is equal to $2K = \{2x : x \in K\}$.*

Exercise 7.6.2. *Show that the proof of Theorem 7.2.1 applies to the estimator $\bar{\theta}_{\mathbb{B}_1}$.*

Exercise 7.6.3. *Suppose* $\mathbf{X} \in \mathbb{R}^{n \times d}$ *is a design matrix, whose rows are independent observations of a d-variate mean zero distribution for which all marginal distributions are* σ*-sub-Gaussian. Show that that, for every* j*, the* j *column* $\mathbf{X}_j$ *of* $\mathbf{X}$ *satisfies* $\|\mathbf{X}_j\| \leq 4\sigma\sqrt{n} + t$ *with probability* $\geq 1 - e^{-t^2/(8\sigma^2)}$.

Part III

Some topics from the first semester

8
Classical large sample theory

This section will be mostly based on a few chapters from [1]. We will be relying on basic definitions and results on convergence in probability theory.

[1] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998

8.1 Preliminaries

We consider vectors with values in $\mathcal{X} \subseteq \mathbb{R}^m$. The set $\mathcal{X}$ forms a metric space with the induced metric $d(x, y) = \|x - y\|$ (but any other equivalent metric is fine). The inequality $x \leq y$ is meant coordinatewise. Many of the results discussed here can be generalized to arbitrary measure spaces with the underlying Borel measure (open sets are measurable).

Recall that a sequence of random vectors X_n converges in distribution to X, denotes $X_n \rightsquigarrow X$, if

$$\mathbb{P}(X_n \leq x) \;\to\; \mathbb{P}(X \leq x),$$

for every point x for which the limit CDF $x \mapsto \mathbb{P}(X \leq x)$ is continuous. Convergence in distribution is often called the weak convergence. Moreover, X_n converges to X in probability, denoted $X_n \xrightarrow{p} X$, if for all $\epsilon > 0$

$$\mathbb{P}(d(X_n, X) > \epsilon) \;\to\; 0.$$

You have studied various equivalent formulations of weak convergence. The one useful for us is given in item (iv) in the next lemma.

Lemma 8.1.1 (Portmanteau Lemma)**.** *We have $X_n \rightsquigarrow X$ if and only if any of the following conditions holds:*

(i) $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ *for all bounded, continuous functions f,*

(ii) $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ *for all bounded, Lipschitz functions f ,*

(iii) $\liminf \mathbb{E}f(X_n) \geq \mathbb{E}f(X)$ *for all nonnegative, continuous functions f,*

(iv) $\liminf \mathbb{P}(X_n \in U) \geq \mathbb{P}(X \in U)$ *for every open set* $U \subseteq \mathcal{X}$,

(v) $\limsup \mathbb{P}(X_n \in B) \leq \mathbb{P}(X \in B)$ *for every closed set* $B \subseteq \mathcal{X}$,

Proof. See Lemma 2.2 in [2]. □

[2] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998

The following result should be also known.

Theorem 8.1.2 (Continuous mapping)**.** *Let* $g : \mathcal{X} \to \mathbb{R}^m$ *be continuous.*

(i) If $X_n \rightsquigarrow X$ *then* $g(X_n) \rightsquigarrow g(X)$.

(ii) If $X_n \xrightarrow{p} X$ *then* $g(X_n) \xrightarrow{p} g(X)$.

Proof. (i). Let $U \subseteq \mathbb{R}^m$ be an open set. Since g is continuous, $g^{-1}(U)$ is open in $\mathcal{X}$. By Portmanteau lemma,

$$\begin{aligned} \liminf \mathbb{P}(g(X_n) \in U) &= \liminf \mathbb{P}(X_n \in g^{-1}(U)) \\ &\geq \mathbb{P}(X \in g^{-1}(U)) = \mathbb{P}(g(X) \in U). \end{aligned}$$

Using the Portmanteau lemma again, we conclude that $g(X_n) \rightsquigarrow g(X)$.

(ii). Fix $\epsilon > 0$. For each $\delta > 0$ let B_δ be the set of x for which there exists y with $d(x, y) < \delta$, but $d(g(x), g(y)) > \epsilon$. If $X \notin B_\delta$ and $d(g(X_n), g(X)) > \epsilon$, then $d(X_n, X) \geq \delta$. Consequently,

$$\mathbb{P}\left(d(g(X_n), g(X)) > \epsilon\right) \leq \mathbb{P}(X \in B_\delta) + \mathbb{P}(d(X_n, X) \geq \delta).$$

The second term on the right converges to zero as $n \to \infty$ for every fixed $\delta > 0$. Because $B_\delta \cap \mathcal{X} \downarrow \emptyset$ by continuity of g, the first term converges to zero as $\delta \to 0$. □

It is important to remember basic relations between different notions of convergence.

Theorem 8.1.3. *Let* X_n, X *and* Y_n *be random vectors. Then*

(i) $X_n \xrightarrow{p} X$ *implies* $X_n \rightsquigarrow X$.

(ii) $X_n \xrightarrow{p} c$ *for* $c \in \mathbb{R}$ *if and only if* $X_n \rightsquigarrow c$.

(iii) if $X_n \rightsquigarrow X$ *and* $d(X_n, Y_n) \xrightarrow{p} 0$, *then* $Y_n \rightsquigarrow X$.

Moreover, if $X_n \rightsquigarrow X$ *and* $Y_n \to c$ *then*

(iv) $X_n + Y_n \rightsquigarrow X + c$,

(v) $X_n Y_n \rightsquigarrow cX$.

(vi) $Y_n^{-1} X_n \rightsquigarrow c^{-1} X$ *provided* $c \neq 0$.

The second part of the theorem is called the Slutsky lemma. For a proof see Theorem 2.7 and Lemma 2.8 in [3].

[3] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998

← Exercise **??**

Example 8.1.4 (t-statistic). *Say (X_n) is a series of i.i.d. random variables with $\mathbb{E}X_i = \mu$ and $\operatorname{var}(X_i) = \sigma^2$. Let $\bar{X}_n = \frac{1}{n}\sum_i X_i$ and let*

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - \bar{X}_n^2).$$

Then

$$t_{n-1} \;:=\; \frac{\bar{X}_n - \mu}{S_n}\sqrt{n-1} \;\rightsquigarrow\; N(0,1).$$

To prove this, note that by CLT we have $\sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow N(0,\sigma^2)$. Also, by the law of large numbers $\bar{X}_n \xrightarrow{p} \mu$ and $\frac{1}{n}\sum X_i^2 \xrightarrow{p} \mathbb{E}X_i^2 = \sigma^2 + \mu^2$. By the continuous mapping theorem $\bar{X}_n^2 \xrightarrow{p} \mu^2$, which together with the Slutsky lemma implies that

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \;\rightsquigarrow\; (\sigma^2 + \mu^2) - \mu^2 \;=\; \sigma^2.$$

Hence, we have

$$\sqrt{\frac{n-1}{n}}\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \;\rightsquigarrow\; \frac{1}{\sigma}N(0,\sigma^2) \;=\; N(0,1).$$

We say that X_n is **bounded in probability** if for every $\epsilon > 0$ there exists $M \in \mathbb{R}$ such that

$$\sup_n \mathbb{P}(\|X_n\| > M) \;<\; \epsilon.$$

By Prohorov's theorem if $X_n \rightsquigarrow X$ for some X then it is bounded in probability (prove it!). Conversely, if X_n is bounded in probability then for some subsequence X_{n_k} we have $X_{n_k} \to X$ for some X.

We now introduce a special notation. Write $X_n = o_P(1)$ if $X_n \xrightarrow{p} 0$ and $X_n = O_P(1)$ if X_n is bounded in probability. More generally,

← Exercise 8.6.2

$$X_n = o_P(R_n) \quad \text{means} \quad X_n = Y_n R_n \text{ and } Y_n = o_P(1),$$

$$X_n = O_P(R_n) \quad \text{means} \quad X_n = Y_n R_n \text{ and } Y_n = O_P(1).$$

Remark 8.1.5. *It is clear that $X_n = o_P(1)$ implies $X_n = O_P(1)$. Indeed, fix $\epsilon > 0$, if $X_n = o_P(1)$ then, for any $M_0 > 0$, there exist $N \in \mathbb{N}$ such that $\mathbb{P}(\|X_n\| > M_0) < \epsilon$ for all $n \neq N$. Let $M_1, \ldots, M_{N-1}$ be any numbers such that $\mathbb{P}(\|X_n\| > M_n) < \epsilon$. Taking $M := \max\{M_0, M_1, \ldots, M_{N-1}\}$ we get that $\mathbb{P}(\|X_n\| > M) < \epsilon$ for all $n \in \mathbb{N}$.*

← Exercise 8.6.3

The following result allows us to effectively work with Taylor series expansions.

← Exercise 8.6.4

Lemma 8.1.6. *Let $R : \mathcal{X} \to \mathbb{R}$ be a continuous function and let X_n be a sequence of random vectors with values in $\mathcal{X}$ and such that $X_n = o_P(1)$. Then, for every $p > 0$,*

(i) if $R(h) = o(\|h\|^p)$ *as* $h \to 0$, *then* $R(X_n) = o_P(\|X_n\|^p)$,

(ii) if $R(h) = O(\|h\|^p)$ *as* $h \to 0$, *then* $R(X_n) = O_P(\|X_n\|^p)$.

Proof. Define $g(h)$ as $g(h) = R(h)/\|h\|^p$ for $h \neq 0$ and $g(0) = 0$. Then $R(X_n) = g(X_n)\|X_n\|^p$.

(i) By assumption, g is continuous at zero, and so it is continuous everywhere. By the continuous mapping theorem, $g(X_n) \xrightarrow{p} g(0) = 0$.

(ii) By assumption, there exists M_0 and $\delta > 0$ such that $|g(h)| \leq M_0$ whenever $\|h\| \leq \delta$. Thus, for every n, $\mathbb{P}(|g(X_n)| > M_0) \leq \mathbb{P}(\|X_n\| > \delta) \to 0$, and so the sequence $g(X_n)$ is bounded in probability by exactly the same argument as we used in Remark 8.1.5.

□

8.2 Delta Method

Suppose an estimator T_n for a parameter θ is available but the quantity of interest is $\phi(\theta)$ for some known function ϕ. It is natural to use the plug-in estimator $\phi(T_n)$. For example, if T_n is the MLE for θ then $\phi(T_n)$ is the MLE for $\phi(\theta)$. But how do the asymptotic properties of $\phi(T_n)$ follow those of T_n?

The continuous-mapping theorem implies that if T_n is consistent for θ and ϕ is continuous, then $\phi(T_n)$ is consistent for $\phi(\theta)$. Here we show that a much stronger statement is true if ϕ is differentiable: if $\sqrt{n}(T_n - \theta) \rightsquigarrow T$ then $\sqrt{n}(\phi(T_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(T)$, where ϕ'_θ is the linear mapping representing the derivative of ϕ at θ (cf. Section A.3). In particular, asymptotic normality is preserved as a linear transformation of a Gaussian vector is Gaussian.

Theorem 8.2.1. *Let* $U \subset \mathbb{R}^k$ *open and let* $\phi : U \to \mathbb{R}^m$ *be a map differentiable at* $\theta \in U$. *Let* T_n *be random vectors taking values in* U. *If* $r_n(T_n - \theta) \rightsquigarrow T$ *for numbers* $r_n \to \infty$, *then* $r_n(\phi(T_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(T)$. *Moreover, the difference between* $r_n(\phi(T_n) - \phi(\theta))$ *and* $\phi'_\theta(r_n(T_n - \theta))$ *converges to zero in probability.*

Proof. Because $r_n(T_n - \theta)$ converges in distribution, we have $r_n(T_n - \theta) = O_P(1)$ and so $T_n - \theta = o_P(1)$. By differentiability of ϕ, $R(h) = \phi(\theta + h) - \phi(\theta) - \phi'(h)$ satisfies $R(h) = o(\|h\|)$ as $h \to \mathbf{0}$. Lemma 8.1.6 allows to replace the fixed h by a random sequence and gives

$$\phi(T_n) - \phi(\theta) - \phi'_\theta(T_n - \theta) \equiv R(T_n - \theta) = o_P(\|T_n - \theta\|).$$

Multiply this with r_n and note that $o_P(r_n\|T_n - \theta\|) = o_P(1)$ because $r_n(T_n - \theta) = O_P(1)$. This gives the last statement of the theorem. Because linear maps are continuous, the continuous mapping theorem

gives that $\phi'_\theta(r_n(T_n - \theta)) \rightsquigarrow \phi'_\theta(T)$. By the Theorem 8.1.3(iii) we get that $r_n(\phi(T_n) - \phi(\theta))$ has the same weak limit. □

Some basic examples of that you saw in the first semester.

Example 8.2.2. *The delta method may be useful in a wide range of scenarios. For example, suppose that we want to obtain asymptotic confidence intervals for some parameter but the asymptotic distribution depends on θ. More concretely, suppose $X_1, \dots, X_n$ are i.i.d.* Poiss(θ). *Then $\sqrt{n}(\bar{X}_n - \theta) \rightsquigarrow N(0, \theta)$. By the delta method $\sqrt{n}(f(\bar{X}_n) - f(\theta)) \rightsquigarrow N(0, (f'(\theta)^2\theta))$. If we solve for $f'(\theta) = C/\sqrt{\theta}$ we get $f(\theta) = \sqrt{\theta}$, $C = 1/2$ and then*

$$\sqrt{n}(\sqrt{\bar{X}_n} - \sqrt{\theta}) \rightsquigarrow N(0, \frac{1}{4}).$$

Therefore we can easily construct the asymptotic confidence interval for $\sqrt{\theta}$ as $\sqrt{\bar{X}_n} \pm \frac{z_{\alpha/2}}{4\sqrt{n}}$. This example easily generalizes.

← Exercise 8.6.5

As an important application of the delta method we show that the MLE in the exponential family has an asymptotically Gaussian distribution. Recall that in the exponential family (1.1) it holds that

$$\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{t}(X)) = \mu(\boldsymbol{\theta}) = \nabla A(\boldsymbol{\theta}), \qquad \mathrm{var}_{\boldsymbol{\theta}}(\mathbf{t}(X)) = V(\boldsymbol{\theta}) = \nabla^2 A(\boldsymbol{\theta}).$$

Suppose that the random sample $X_1, \dots, X_n$ is generated from $\mathbb{P}_{\boldsymbol{\theta}_0}$. Denote $\bar{\mathbf{t}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{t}(X_i)$. By the central limit theorem, it follows that

$$\sqrt{n}\nabla \ell_n(\boldsymbol{\theta}_0) = \sqrt{n}(\bar{\mathbf{t}}_n - \mu(\boldsymbol{\theta}_0)) \rightsquigarrow N_d(\mathbf{0}, V(\boldsymbol{\theta}_0)). \tag{8.1}$$

Using the delta method we conclude.

Theorem 8.2.3. *The MLE $\widehat{\boldsymbol{\theta}}_n$ in the exponential family (1.1) based on the sample $X_1, \dots, X_n$ satisfies $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightsquigarrow N(\mathbf{0}, V^{-1}(\boldsymbol{\theta}_0))$.*

Proof. As noted above $\sqrt{n}(\bar{\mathbf{t}}_n - \mu(\boldsymbol{\theta}_0)) \rightsquigarrow N(\mathbf{0}, V(\boldsymbol{\theta}_0))$. The delta method, applied with $\phi = \mu^{-1}$, implies that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ also converges in distribution to a Gaussian distribution. The fact that $\widehat{\boldsymbol{\theta}}_n = \mu^{-1}(\bar{\mathbf{t}}_n)$ is part of Proposition 1.3.6. The Jacobian of μ^{-1} at $\mu(\boldsymbol{\theta}_0)$ is the inverse of the Jacobian of μ at $\boldsymbol{\theta}_0$, which is equal to $\nabla^2 A(\boldsymbol{\theta}_0) = V(\boldsymbol{\theta}_0)$. Thus, $\phi'_{\mu(\boldsymbol{\theta}_0)}(T) = V^{-1}(\boldsymbol{\theta}_0) \cdot T$. Now it is straightforward to check that the asymptotic covariance matrix is $V^{-1}(\boldsymbol{\theta}_0)$. □

With small amount of extra work, this result can be generalized to curved exponential families for which the canonical parameter $\boldsymbol{\theta}$ is parametrized in a smooth way by a lower dimensional parameters $\boldsymbol{\tau}$. Denote by $\widehat{\boldsymbol{\tau}}$ the MLE in this smaller model and assume that the sample is generated from the parameter $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\tau}_0)$ corresponding to $\boldsymbol{\tau}_0$. Taking the first-order expansion around $\boldsymbol{\tau}_0$ gives

← Exercise 8.6.6

$$0 \;=\; \nabla_{\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}(\widehat{\boldsymbol{\tau}}_n)) \;=\; \nabla_{\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}(\boldsymbol{\tau}_0)) + \nabla^2_{\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}(\boldsymbol{\tau}_0))(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0) + o_P(\|\boldsymbol{\tau}_0 - \widehat{\boldsymbol{\tau}}_n\|).$$

Multiply now this equation by $\sqrt{n}$ and use Exercise 8.6.6 to conclude that

$$-\nabla^2_{\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}(\boldsymbol{\tau}_0)) \cdot \sqrt{n}(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0) \;=\; \sqrt{n} \nabla_{\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}(\boldsymbol{\tau}_0)) + o_P(1). \tag{8.2}$$

By the chain rule, $\nabla_{\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}(\boldsymbol{\tau}_0)) \;=\; (\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_0))^T \nabla \ell_n(\boldsymbol{\theta}(\boldsymbol{\tau}_0))$, and so, using (8.1), we get

$$\sqrt{n} \nabla_{\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}(\boldsymbol{\tau}_0)) \;\rightsquigarrow\; N\left(\mathbf{0}, (\tfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_0))^T \cdot V(\boldsymbol{\theta}_0) \cdot \tfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_0)\right).$$

By Slutsky Lemma (Theorem 8.1.3(iv)), the left-side expression in (8.2) also converges in distribution to the same Gaussian. By Exercise 8.6.7 we conclude that $\sqrt{n}(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0)$ is asymptotically normal with mean zero and the covariance matrix equal to the *inverse* of the Fisher information matrix $\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_0)^T \cdot V(\boldsymbol{\theta}_0) \cdot \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_0)$.

← Exercise 8.6.7

8.3 *M-estimators*

In this chapter $X_1, \ldots, X_n$ are i.i.d. $\mathbb{P}$ where $\mathbb{P} \in \mathcal{P}$, with $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ such that

1. $\Theta \subset \mathbb{R}^d$ open,
2. densities of $\mathbb{P}_\theta$ are $p(\cdot, \theta)$, $\theta \in \Theta$.

M-estimators were defined and discussed in Section 4.1.2. In this section we discuss basic asymptotics of this class of estimators.

8.3.1 *Consistency*

In order to develop some general asymptotic theory of M-estimators we start by discussing their consistency. Under minor conditions the law of large numbers will give us that

$$M_n(\theta) \xrightarrow{p} M(\theta) \qquad \text{for every } \theta, \tag{8.3}$$

where $M(\theta) = \mathbb{E} m_\theta(X)$. It is reasonable to expect that the sequence of maximizers $\hat{\theta}_n$ of M_n converges in probability to the maximizer of M. However, pointwise convergence in (8.3) is too weak because $\hat{\theta}_n$ depends on the whole function $\theta \mapsto M_n(\theta)$. We present approach based on the assumption of the **uniform convergence** of M_n to M: $\|M_n - M\|_\infty \xrightarrow{p} 0$, where

$$\|M_n - M\|_\infty \;:=\; \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|.$$

Theorem 8.3.1. *Let $\{M_n\}$ be a sequence of random functions, continuous on Θ. Say $\|M_n - M\|_\infty \xrightarrow{p} 0$, where M is some non-random continuous function on Θ. Then*

1. *If $T_n \xrightarrow{p} T^*$ then $M_n(T_n) \xrightarrow{p} M(T^*)$.*
2. *If $t^* = \arg\max M(t)$ uniquely and $T_n \in \arg\max M_n(t)$, then $T_n \xrightarrow{p} t^*$.*

Proof. 1. By the triangle inequality

$$|M_n(T_n) - M(T^*)| \leq |M_n(T_n) - M(T_n)| + |M(T_n) - M(T^*)|.$$

The second term on the right goes to zero in probability by the continuous mapping theorem. The first term goes to zero by the uniform converge hypothesis in the theorem because $|M_n(T_n) - M(T_n)| \leq \|M_n - M\|_\infty \xrightarrow{p} 0$. We conclude that $|M_n(T_n) - M(T^*)| \xrightarrow{p} 0$ or, equivalently $M_n(T_n) \xrightarrow{p} M(T^*)$. 2. For some $\epsilon > 0$, let $K_\epsilon = K \cap (B_\epsilon(t^*))^c$, where $B_\epsilon(t^*)$ is a ball of radius ϵ around t^*. Let $m = M(t^*)$ and let $m_\epsilon = \sup_{t \in K_\epsilon} M(t)$. Since t^* is unique $\delta := m - m_\epsilon > 0$. From uniform convergence, there is some N s.t. for all $n > N$ we have $\|M_n - M\|_\infty < \frac{\delta}{2}$. We can therefore infer the following:

$$\sup_{t \in K_\epsilon} M_n(t) \;<\; \sup_{t \in K_\epsilon} M(t) + \frac{\delta}{2} \;=\; m_\epsilon + \frac{\delta}{2} \;=\; m - \frac{\delta}{2}$$

and

$$M_n(T_n) \;=\; \max_{t \in K} M_n(t) \;\geq\; M_n(t^*) \;>\; M(t^*) - \frac{\delta}{2} \;=\; m - \frac{\delta}{2}.$$

By these two inequalities we get that $T_n \notin K_\epsilon$, and thus $T_n \in B_\epsilon(t^*)$. Hence, $\|M_n - M\|_\infty < \delta \Rightarrow \|T_n - t^*\| < \epsilon \Rightarrow \mathbb{P}(\|T_n - t^*\| > \epsilon) \leq \mathbb{P}(\|M_n - M\|_\infty) > \frac{\delta}{2}) \Rightarrow T_n \xrightarrow{p} t^*$. □

The assumption of uniform convergence holds for the log-likelihood function in exponential families but is typically too strong, especially when Θ is not compact.

Example 8.3.2. *Consider the problem of estimating the variance using the sample variance $s_n^2 = \frac{1}{n}\sum_i x_i^2$. This corresponds to optimizing the function*

$$M_n(\sigma) = \log\sigma + \frac{s_n^2}{2\sigma^2}$$

with $M(\sigma) = \log\sigma + \frac{\sigma_0^2}{2\sigma^2}$, where σ_0^2 is the variance of the sample. Note that $|M_n(\sigma) - M(\sigma)|$ is unbounded for $\sigma > 0$. Nevertheless, consistency follows directly by the law of large numbers.

Under some regularity conditions it is possible to get a fairly universal approach that does not require uniform convergence. Suppose that $m_\theta(x)$ is twice continuously differentiable with respect to θ on the open parameters set Θ and consider a Taylor series expansion of ∇M_n at the true data generating θ_0. Since Θ is open then the M-estimator $\hat{\theta}_n$ satisfies $\nabla M_n(\hat{\theta}_n) = \mathbf{0}$ and so

$$\mathbf{0} \;=\; \nabla M_n(\theta_0) + \nabla^2 M_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0), \tag{8.4}$$

where $\tilde{\theta}_n$ lies between $\hat{\theta}_n$ and θ_0. By the law of large numbers, as long as $\mathbb{E}_{\theta_0} \nabla m_{\theta_0}(X) < \infty$,

$$\nabla M_n(\theta_0) \;\xrightarrow{p}\; \mathbb{E}_{\theta_0} \nabla M_n(\theta_0).$$

To establish consistency of $\hat{\theta}_n$ using (8.4) it suffices to have that: ← Exercise 8.6.8

(i) $\mathbb{E}_{\theta_0} \nabla M_n(\theta_0) = 0$ and

(ii) $(\nabla^2 M_n(\tilde{\theta}_n))^{-1} = O_P(1)$.

In Exercise 8.6.8 we show that (i) holds when $m_\theta(x) = -\log p_\theta(x)$ (MLE) as well as $m_\theta(x) = \|\theta - \delta(X)\|^2$, where $\delta(X)$ is an unbiased estimator of the parameter θ. More generally, (i) holds if θ_0 minimizes the function $\theta \mapsto \mathbb{E}_{\theta_0} m_\theta(X)$. Thus, typically (ii) remains the main regularity assumption that needs to be checked in order to establish consistency.

8.3.2 *Asymptotic normality*

We will now discuss asymptotic normality of M-estimators assuming that $\hat{\theta}_n$ is a consistent estimator and $\Theta \subseteq \mathbb{R}^d$ is open. If ψ_θ in (4.5) is twice differentiable then there is a standard approach for proving asymptotic normality based on the Taylor series expansion. As earlier, suppose that

$$\Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi_\theta(X_i), \qquad \Psi(\theta) = \mathbb{E}\psi_\theta(X) = \mathbb{E}\Psi_n(\theta)$$

and assume that $\hat{\theta} \xrightarrow{p} \theta_0$, where

$$\Psi_n(\hat{\theta}_n) = 0, \quad \Psi(\theta_0) = 0.$$

Taylor's theorem gives

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + \nabla\Psi_n(\theta_0)(\hat{\theta}_n - \theta_0) + R_n,$$

where $R_n = o_P(\|\hat{\theta}_n - \theta_0\|)$. In other words,

$$\nabla\Psi_n(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) \;=\; -\sqrt{n}\Psi_n(\theta_0) - \sqrt{n}R_n,$$

By the central limit theorem, as long as, $\boldsymbol{B}(\theta_0) := \mathbb{E}\Psi_{\theta_0}(X)\Psi_{\theta_0}^{\intercal}(X)$ has finite entries,

$$-\sqrt{n}\Psi_n(\hat{\theta}_n) \rightsquigarrow N(0, \boldsymbol{B}(\theta_0)).$$

By the law of large numbers, as long as $\boldsymbol{x}A(\theta_0) := \mathbb{E}\nabla\psi_{\theta_0}(X)$ has finite entries, then

$$\nabla\Psi_n(\theta_0) \xrightarrow{p} \boldsymbol{A}(\theta_0)$$

Under suitable regularity conditions $\sqrt{n}R_n = o_P(1)$, in which case we can use the Slutsky lemma to conclude that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, \boldsymbol{A}(\theta_0)^{-1}\boldsymbol{B}(\theta_0)\boldsymbol{A}(\theta_0)^{-\intercal}). \tag{8.5}$$

The conditions that assure that $\sqrt{n}R_n = o_P(1)$ may be complicated in general. However, they will hold if Ψ_n is sufficiently smooth.

A special case of interest is the maximum likelihood estimator. In this case $\psi_\theta(x) = \nabla \log p_\theta(x)$ and so

$$\mathbb{E}\psi_\theta(X) = \mathbb{E}\nabla\ell_n(\theta), \qquad \mathbb{E}\nabla\psi_\theta(X) = \mathbb{E}\nabla^2\ell_n(\theta).$$

Under some regularity conditions the maximum likelihood estimator is consistent and asymptotically normal: as in (8.5), with

$$\boldsymbol{A}(\theta_0) = -\mathbb{E}_{\theta_0}[\nabla^2\ell_n(\theta_0)] \qquad \text{and} \qquad \boldsymbol{B}(\theta_0) = \mathbb{E}_{\theta_0}\left[\nabla\ell_n(\theta_0)\nabla\ell_n(\theta_0)^{\intercal}\right].$$

In this case $\boldsymbol{B}(\theta_0)$ is the variance of the score also known as the Fisher information matrix. If $\log p_\theta(x)$ is twice continuously differentiable, we can write

$$\nabla^2 \log p_\theta(X) = \frac{1}{p_\theta(X)}\nabla^2 p_\theta(X) - (\nabla \log p_\theta(X))(\nabla \log p_\theta(X))^T.$$

Since

$$\mathbb{E}_\theta \frac{1}{p_\theta(X)}\nabla^2 p_\theta(X) = \int \nabla^2 p_\theta(x)\mathrm{d}x = \nabla^2 \int p_\theta(x)\mathrm{d}x = 0,$$

we conclude that in this case $\boldsymbol{A}(\theta) = \boldsymbol{B}(\theta)$ and so the MLE $\hat{\theta}_n$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, \boldsymbol{B}(\theta_0)^{-1}). \tag{8.6}$$

It is of considerable interest to establish asymptotic normality also in the case when Ψ_n is not twice differentiable. But we will not discuss this case in more detail here.

← Exercise 8.6.9

8.4 Generalized likelihood ratio test

The tests in Chapter 3 have strong optimality properties but require conditions on the densities for the data and the form of the hypotheses that are rather special and can fail for many natural models.

By contrast, the generalized likelihood ratio test introduced in this chapter requires little structure, but it does not have exact optimality properties. Use of this test is justified by large sample theory.

Let the data $X_1, \ldots, X_n$ be i.i.d. with common density p_θ for $\theta \in \Theta$. We want to test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where $\Theta_0 \subset \Theta_1 \subseteq \Theta$. Note that the hypotheses are nested.

A sensible extension of the idea behind the likelihood ratio test, as discussed in Proposition 3.1.1, is to base a test on the log-likelihood ratio

$$\lambda_n \;=\; \lambda_n(X_1, \ldots, X_n) \;=\; \log \frac{\sup_{\theta \in \Theta_1} \prod_{i=1}^n p_\theta(X_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n p_\theta(X_i)}.$$

As before, the null hypothesis is rejected for large values of the statistic. Let $\ell_n(\theta) = \frac{1}{n} \sum_i \log p_\theta(X_i)$ be the log-likelihood function and suppose both suprema are attained. If $\hat{\theta}_n$ is the optimum over Θ_1 and $\hat{\theta}_{n,0}$ is the optimum over Θ_0 then

$$\lambda_n \;=\; n(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_{n,0})).$$

We next give an example that can be viewed as the limiting situation for which the approximation is exact:

Example 8.4.1 (The Gaussian sequence model)**.** *Let $Y_1, \ldots, Y_n$ be independent with $Y_i \sim N(\mu_i, \sigma_0^2)$ where σ_0 is known. In other words,*

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n) \;\sim\; N_n(\boldsymbol{\mu}, \sigma_0^2 I_n).$$

This model is sometimes called the Gaussian sequence model and will be one of the important examples in our discussion of high-dimensional problems. We are now interested in testing whether $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is a member of a q-dimensional linear subspace $\mathcal{L}_0 \subset \mathbb{R}^n$, versus the alternative that $\mu \in \mathcal{L} \setminus \mathcal{L}_0$ where $\mathcal{L}$ is an r-dimensional linear subspace of $\mathbb{R}^d$ and $\mathcal{L}_0 \subset \mathcal{L}$, $r > q$.

Transform to canonical form by setting $\boldsymbol{U} = Q\boldsymbol{Y}$, where $Q \in \mathbb{R}^{n \times n}$ is an orthogonal matrix with the first q rows spanning $\mathcal{L}_0$ and the first r rows spanning $\mathcal{L}$. By construction, $\boldsymbol{U} \sim N_n(Q\boldsymbol{\mu}, \sigma_0^2 I_n)$, where the mean vector $\boldsymbol{\eta} = Q\boldsymbol{\mu}$ satisfies $\eta_i = 0$ for $i \geq q+1, \ldots, n$ under H_0, and $\eta_i = 0$ for $i \geq r+1, \ldots, n$ under H_1.

Set $\theta_i = \eta_i/\sigma_0$, $i = 1, \ldots, r$ and $X_i = U_i/\sigma_0$, $i = 1, \ldots, n$. Then $\boldsymbol{X} \sim N(\boldsymbol{\theta}, I_n)$. Moreover, the hypothesis H_0 is equivalent to $H_0 : \theta_i = 0$ for $i \geq q+1$, and the alternative is $H_1 : \theta_i = 0$ for $i \geq r+1$. The log-likelihood for $\boldsymbol{X} \sim N(\boldsymbol{\theta}, I_n)$ is

$$\ell(\boldsymbol{\theta}) \;=\; -\frac{1}{2}\log(2\pi) - \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\theta}\|^2,$$

and so $\arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \mathbf{x}) = \arg\min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{x} - \boldsymbol{\theta}\|^2$. *Under H_0,*

$$2\lambda \;=\; 2(\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)) \;=\; \sum_{i=q+1}^{r} X_i^2 \sim \chi^2_{r-q}.$$

It is a remarkable fact that χ^2_{r-1} holds as an approximation to the null distribution of $2\lambda_n$ quite generally when the hypothesis is a q-dimensional submanifold of an r-dimensional parameter space. Some of the sufficient regularity conditions under which this holds were briefly discussed in Section 8.3.1 (e.g. Θ open, $\log p_\theta(x)$ twice continuously differentiable with respect to θ, some conditions on the behavior of the hessian $\nabla^2 \ell_n$ around $\theta_0 \in \Theta_0$).

Theorem 8.4.2 (Wilk's theorem)**.** *Consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where $\Theta_0 \subset \Theta_1 \subseteq \Theta$ and Θ_0, Θ_1 are manifolds of dimension q and r respectively. Under the same conditions guaranteeing asymptotic normality of the two MLEs, we have under the null hypothesis,*

$$2\lambda_n \ \rightsquigarrow \ \chi^2_{r-q}.$$

Sketch of the proof. For simplicity, we assume that $\Theta_1 = \Theta$. Let $\hat{\theta}_{0,n}$ be the MLE over Θ_0 and $\hat{\theta}_n$ be the MLE over Θ. Directly by definition

$$2\lambda_n \ = \ 2n(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_{0,n})).$$

Applying the first order Taylor expansion of ℓ_n at $\hat{\theta}_n$, and using the fact that $\nabla \ell_n(\hat{\theta}_n) = 0$, we find that

$$\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_{0,n}) \ = \ -\frac{1}{2}(\hat{\theta}_n - \hat{\theta}_{0,n})^\intercal \nabla^2_\theta \ell_n(\tilde{\theta}_n)(\hat{\theta}_n - \hat{\theta}_{0,n}),$$

where $\tilde{\theta}_n =$ lies between $\hat{\theta}_{0,n}$ and $\hat{\theta}_n$. By asymptotic normality $\hat{\theta}_{0,n} \xrightarrow{p} \theta_0$ and $\hat{\theta}_n \xrightarrow{p} \theta_0$ and so also $\tilde{\theta}_n \xrightarrow{p} \theta_0$. Hence, we argue that

$$-\nabla^2_\theta \ell_n(\tilde{\theta}_n) \ \xrightarrow{p} \ -\mathbb{E}[\nabla^2_\theta \ell_n(\theta_0)] \ = \ I(\theta_0).$$

(This statement is actually a bit subtle and may require some further regularity assumptions.) Therefore

$$2n(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_{0,n})) = \sqrt{n}(\hat{\theta}_n - \hat{\theta}_{0,n})^\intercal \left(-\nabla^2_\theta \ell_n(\tilde{\theta}_n)\right) \sqrt{n}(\hat{\theta}_n - \hat{\theta}_{0,n}).$$

In the case when $r = d$, $\hat{\theta}_{0,n} = \theta_0$ since there is no maximization and then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \ \rightsquigarrow \ N(0, I(\theta_0)^{-1})$$

from the asymptotic normality of the MLE. Now putting together the pieces:

$$\sqrt{n}(\hat{\theta}_n - \theta_0)^\intercal \left(-\frac{1}{n}\nabla^2_\theta \ell_n(\tilde{\theta}_n)\right) \sqrt{n}(\hat{\theta}_n - \theta_0) \ \rightsquigarrow \ \sum_{i=1}^{d} Z_i^2 \ = \ \chi^2_d,$$

where $Z_i \sim N(0,1)$. The case when $r < d$ does not give much more insight, but a lot more algebra. □

Example 8.4.3 (Exponential families)**.** *Suppose that the observations are sampled from a density p_θ in the d-dimensional regular exponential family*

$$p_\theta(x) \;=\; h(x)e^{\langle\theta,t(x)\rangle - A(\theta)}.$$

Let $\Theta \subseteq \mathbb{R}^d$ be the canonical parameter space, and consider testing a null hypothesis $\Theta_0 \subset \Theta$ versus its complement. The log-likelihood ratio statistic is given by

$$\lambda_n = n \sup_{\theta\in\Theta} \inf_{\theta\in\Theta_0} \left(\langle\theta - \theta_0, \bar{t}_n\rangle - A(\theta) + A(\theta_0)\right).$$

Note that the Kullback-Leibler divergence of the measures $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_\theta$ is equal to

$$K(\theta,\theta_0) \;=\; \mathbb{E}_\theta \log \frac{p_\theta}{p_{\theta_0}} = \langle\theta - \theta_0, \mathbb{E}_\theta t(X)\rangle - A(\theta) + A(\theta_0).$$

If the maximum likelihood estimator $\hat{\theta} \in \Theta$ exists, then $\mathbb{E}_{\hat{\theta}}(t(X)) = \bar{t}_n$ by Proposition 1.3.6. Comparing the two preceding displays, we see that the likelihood ratio statistic can be written as

$$\lambda_n \;=\; n \inf_{\theta_0\in\Theta_0} K(\hat{\theta},\theta_0).$$

This formula can be used to study the asymptotic properties of the likelihood ratio statistic directly.

8.5 Limits of Bayesian procedures

In this section $\boldsymbol{\theta}$ will denote a random parameter rather than a vector. If $\boldsymbol{\theta}$ possesses a density π, then the density of the posterior distribution of $\boldsymbol{\theta}$ is given by Bayes' formula

$$\pi(\theta|X_1,\ldots,X_n) \;=\; \frac{\prod_{i=1}^n p(X_i|\theta)\pi(\theta)}{\int \prod_{i=1}^n p(X_i|\theta)\pi(\theta)\mathrm{d}\theta} \tag{8.7}$$

which is a random kernel in the same sense as randomized decision rules; c.f. Definition 2.2.1. This expression may define a probability density even if π is not a probability density itself. A prior distribution with infinite mass is called improper.

We pose the question, under what conditions are Bayes methods asymptotically consistent/efficient. We start with an example.

Example 8.5.1. *$X \sim \mathrm{Bin}(n,\theta)$ with prior $\boldsymbol{\theta} \sim \mathrm{Beta}(a,b)$, that is, $\pi(\theta) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\theta^{a-1}(1-\theta)^{b-1}$. Since the posterior satisfies*

$$\pi(\theta|x) \propto \theta^{a+x-1}(1-\theta)^{b+n-x-1},$$

it must be equal to $\mathrm{Beta}(a+x, b+n-x)$. Since the mean in $\mathrm{Beta}(a,b)$ is $\frac{a}{a+b}$, the Bayes estimate under the squared lossx (c.f. Proposition 2.3.5) is:

$$\delta^n_{a,b} \;=\; \mathbb{E}(\theta|x) \;=\; \frac{a+x}{a+b+n}.$$

If we assume that the true value of θ is θ_0,

$$\sqrt{n}(\delta^n_{a,b} - \theta_0) = \sqrt{n}\left(\frac{X}{n} - \theta_0\right) + \frac{\sqrt{n}}{a+b+n}\left(a - (a+b)\frac{X}{n}\right).$$

Note that $\sqrt{n}(\frac{X}{n} - \theta_0) \rightsquigarrow N(0, \theta_0(1-\theta_0))$ since $\frac{X}{n}$ is MLE of θ and

$$\tfrac{\sqrt{n}}{a+b+n}(a - (a+b)\tfrac{X}{n}) \xrightarrow{p} 0.$$

Therefore the Bayes estimate is asymptotically consistent and efficient for all a and b.

Bernstein-von Mises theorem theorem guarantees that the random kernel $\pi(\theta|X_1, \ldots, X_n)$ in (8.7) is "close" to $N(\hat{\theta}_n, (nI(\theta_0))^{-1})$, where $\hat{\theta}_n$ is the MLE and θ_0 is the true value.

Theorem 8.5.2 (Bernstein-von Mises)**.** *Assume that the prior density π is continuous and strictly positive and the standard regularity conditions for asymptotic normality of the MLE $\hat{\theta}_n$ hold. Then the conditional density of $\sqrt{n}(\boldsymbol{\theta} - \hat{\theta}_n)$ given $X_1, \ldots, X_n$ converges to the PDF of $N(0, (I(\theta_0))^{-1})$.*

Proof. (Proof sketch, see Section 5.5 in [4] for details) Condition on $X_1, \ldots, X_n$. Let $v = \sqrt{n}(\theta - \hat{\theta}_n)$ so that $\theta = \hat{\theta}_n + v/\sqrt{n}$. Denote

$$f(v) = \pi(\hat{\pi}_n + \tfrac{v}{\sqrt{n}}|X_1, \ldots, X_n)$$

so that

$$\log f(v) \;=\; \text{const} + n\ell_n(\hat{\theta}_n + \tfrac{v}{\sqrt{n}}) + \log \pi(\hat{\theta}_n + \tfrac{v}{\sqrt{n}}). \tag{8.8}$$

Using the second order expansion of $\ell_n(\hat{\theta}_n + \frac{v}{\sqrt{n}})$ around $\hat{\theta}_n$ and using the fact that $\nabla \ell_n(\hat{\theta}_n) = 0$ we get

$$n\ell_n(\hat{\theta}_n + \tfrac{v}{\sqrt{n}}) = n\ell_n(\hat{\theta}_n) + \frac{1}{2}v^T \nabla^2 \ell_n(\tilde{\theta}_n)v,$$

where $\tilde{\theta}_n = \hat{\theta}_n + \frac{tv}{\sqrt{n}}$ for $t \in (0,1)$. Plugging this to (8.8) gives

$$\log f(v) \;=\; \text{const} + \frac{1}{2}v^T \nabla^2 \ell_n(\hat{\theta}_n + \tfrac{tv}{\sqrt{n}})v + \log \pi(\hat{\theta}_n + \tfrac{v}{\sqrt{n}}). \tag{8.9}$$

This holds for any fixed $v \in \mathbb{R}^d$ and so, by (8.9), the conditional distribution of $\boldsymbol{v} = \sqrt{n}(\boldsymbol{\theta} - \hat{\theta}_n)$ is proportional to

$$\pi(\hat{\theta}_n + \tfrac{v}{\sqrt{n}}) \exp\{-\frac{1}{2}v^T(-\nabla^2 \ell_n(\hat{\theta}_n + \tfrac{tv}{\sqrt{n}}))v\}.$$

Since $\hat{\theta}_n$ converges in probability to θ_0, $v/\sqrt{n}$ converges to zero and under appropriate uniformity conditions, $-\nabla^2 \ell_n(\theta_n + tv/\sqrt{n})$ converges to $I(\theta_0)$. Therefore, expression (8.9) after normalizing converges to the density of $N(0, (I(\theta_0))^{-1})$. □

[4] Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics—basic ideas and selected topics. Vol. 1.* Texts in Statistical Science Series. CRC Press, Boca Raton, FL, second edition, 2015

8.6 Exercises

Exercise 8.6.1. *Say $X_n \rightsquigarrow X \in \mathbb{R}^d$ and $Y_n \rightsquigarrow c \in \mathbb{R}^k$ (constant) and let $f : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^m$ continuous almost everywhere. Show that $f(X_n, Y_n) \rightsquigarrow f(X, c)$.*

Exercise 8.6.2. *Argue that $X_n \xrightarrow{p} X$ if and only if $d(X_n, X) = o_P(1)$.*

Exercise 8.6.3. *Show that:*

$$\begin{aligned} o_P(1) + o_P(1) &= o_P(1) \\ o_P(1) + O_P(1) &= O_P(1) \\ o_P(1)O_P(1) &= o_P(1) \\ (1 + o_P(1))^{-1} &= O_P(1) \\ o_P(R_n) &= R_n o_P(1) \\ O_P(R_n) &= R_n O_P(1) \\ o_P(O_P(1)) &= o_P(1). \end{aligned}$$

Exercise 8.6.4. *Let (X_n) be a random sequence in $\mathbb{R}^m$. Show that $X_n = o_p(1)$ if and only if each coordinate sequence is $o_P(1)$.*

Exercise 8.6.5. *Let $X_1, \ldots, X_n$ be i.i.d. $\mathbb{E}X_i = \mu$, $\operatorname{var}(X_i) = 1$. Find constants such that $r_n(\bar{X}_n^2 - a_n)$ converges in distribution when $\mu = 0$ and when $\mu \neq 0$.*

Exercise 8.6.6. *Show that $\sqrt{n}(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0) = O_P(1)$. Conclude that $\widehat{\boldsymbol{\tau}}_n \xrightarrow{p} \boldsymbol{\tau}_0$.*

Exercise 8.6.7. *Show that in exponential families*

$$-\nabla^2_{\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}(\boldsymbol{\tau}_0)) \xrightarrow{p} \tfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_0))^T \cdot V(\boldsymbol{\theta}_0) \cdot \tfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_0).$$

Exercise 8.6.8. *Show that (i) in the bottom of Section 8.3.1 holds when $m_\theta(x) = -\log p_\theta(x)$ (MLE) as well as $m_\theta(x) = \|\theta - \delta(X)\|^2$, where $\delta(X)$ is an unbiased estimator of the parameter θ.*

Exercise 8.6.9. *Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, 1)$, where it is known that $\mu \geq 0$. Show that the MLE is not asymptotically normal under $\mu = 0$. Why does this not contradict our result on asymptotic normality of the MLE?*

Part IV

Mathematical appendix

A
Real Analysis

The Advanced Theory of Statistics is a technical subject. It is then important to actively look for high-level insights. This allows not only to understand the material better but also to see how the presented results may be generalized. Part of the goal of this lecture is to help students look for such insights.

In this appendix we briefly cover some fundamental concepts that help to understand many parts of theoretical statistics. These are: differentiation in vector spaces and convexity. This exposition assumes certain level of mathematical maturity on the level of a basic real analysis course. For more details, we refer to
`https://pzwiernik.github.io/docs/RealAnalysisNotes.pdf`.

A.1 Vector spaces

A set V is a **vector space** if (i) $\mathbf{0}$ lies in V (ii) for any two $\boldsymbol{x}, \boldsymbol{y} \in V$ also $\boldsymbol{x} + \boldsymbol{y} \in V$, (iii) if $\boldsymbol{x} \in V$ and $\lambda \in \mathbb{R}$ then $\lambda \cdot \boldsymbol{x}$ lies in V. A general abstract definition of a vector space is more complicated because it needs to explain what we mean by $\mathbf{0}$ and what we mean by the algebraic operations $+$ and $\cdot$. Here however, we always work with variations of the following three examples (Examples A.1.1-A.1.4), where all these objects are naturally defined.

Example A.1.1 (The Euclidean space $\mathbb{R}^d$)**.** *The real space $\mathbb{R}^d$ with elements $\mathbf{x} = (x_1, \ldots, x_d)$, $x_i \in \mathbb{R}$, is an example of a vector space. We define the standard inner product as*

$$\langle \mathbf{x}, \mathbf{y} \rangle \ := \ x_1 y_1 + \cdots + x_d y_d \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

The vector space $\mathbb{R}^d$ equipped with the induced norm $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ is called the Euclidean space. Every subset of $\mathbb{R}^d$ given by linear *equations also forms a vector space with the induced norm.*

In general, an inner product on a vector space V is a function from

$V \times V$ to $\mathbb{R}$ that must satisfy the following three conditions for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$

1. $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$ (symmetry)
2. $\langle \alpha \boldsymbol{x} + \beta \boldsymbol{y}, \boldsymbol{z} \rangle = \alpha \langle \boldsymbol{x}, \boldsymbol{z} \rangle + \beta \langle \boldsymbol{y}, \boldsymbol{z} \rangle$ (linearity in the first argument)
3. $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$ and is zero only if $\boldsymbol{x} = \mathbf{0}$. (positive definiteness)

The inner product space induces a norm in the standard way

$$\|\mathbf{x}\| \ := \ \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

but not all norms are obtained in this way. Every vector space with a given norm becomes a metric space with distance function satisfying

$$d(\mathbf{x}, \mathbf{y}) \ := \ \|\mathbf{x} - \mathbf{y}\| \qquad \text{for } \mathbf{x}, \mathbf{y} \in V.$$

Example A.1.2 (The space of $m \times n$ matrices)**.** *The space $\mathbb{R}^{m \times n}$ of all real $m \times n$ matrices also forms a vector space. The standard inner product is*

$$\langle A, B \rangle \ = \ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij} \ = \ \operatorname{tr}(AB^{\mathsf{T}}) \qquad \textit{for all } A, B \in \mathbb{R}^{m \times n}.$$

The induced norm is the Frobenius norm $\|A\|_F := \sqrt{\langle A, A \rangle}$ but for matrices we typically work with the operator norm instead

$$\|A\| \ := \ \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

A special case is the vector space $\mathbb{S}^m$ of all $m \times m$ symmetric matrices with the inner product induced from $\mathbb{R}^{m \times m}$:

$$\mathbb{S}^m \ = \ \{A \in \mathbb{R}^{m \times m} : \ A = A^T\}.$$

Exercise A.1.3. *Show that* $\operatorname{tr}(AB^{\mathsf{T}}) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$ *for all* $A, B \in \mathbb{R}^{m \times n}$.

All finite dimensional vector spaces are similar and behave like the Euclidean space $\mathbb{R}^d$. In particular, defining a metric space structure is straightforward. In more complicated situations we may need to work a bit extra. Consider the following important example.

Example A.1.4 (L^2 functions)**.** *The set of all measurable functions $f : \mathcal{X} \to \mathbb{R}$ on some set $\mathcal{X}$ also forms a vector space with addition defined pointwise (why? what is the zero of this vector space?). The natural candidate for an inner product, namely*

$$\langle f, g \rangle \ := \ \int_{\mathcal{X}} f(x) g(x) \mathrm{d}\mu$$

does not need to satisfy positive definitedness and it may not be a well-defined function to $\mathbb{R}$.

Instead, we will work in a smaller functional space $L^2(\mathcal{X})$ of all measurable functions satisfying $\int f^2(x) \mathrm{d}\mu < +\infty$, where two functions are identified if they differ on a subset of measure zero. In this space, $\langle f, g \rangle$ defines a valid inner product.

A.2 Continuity and semicontinuity

A.2.1 Continuity

Continuity is one of the most fundamental concepts of real analysis.

Definition A.2.1. *If $f : \mathcal{X} \to \mathcal{Y}$ is a mapping between metric spaces then f is **continuous** at $\mathbf{x} \in \mathcal{X}$ if for every sequence $(\mathbf{x}_n)$ in $\mathcal{X}$, $\mathbf{x}_n \to \mathbf{x}$ implies $f(\mathbf{x}_n) \to f(\mathbf{x})$. We say that f is continuous if it is continuous at every $\mathbf{x}$.*

Proposition A.2.2. *Composition of continuous functions is continuous.*

Proof. Follows easily from the definition. We leave the details as an exercise. □

There are various equivalent definitions of continuity and it is important to be aware of them. The sequential definition is typically the easiest both conceptually and operationally. However, in this course we will also be using two other definitions, which we discuss next.

The standard way of defining continuous functions is, so called, (ϵ, δ)-condition.

Theorem A.2.3. *A function $f : \mathcal{X} \to \mathcal{Y}$ is continuous at x if and only if*

$$\forall \epsilon > 0 \;\; \exists \delta > 0 \quad \Big(d(\mathbf{x}, \mathbf{y}) < \delta \;\; \Rightarrow \;\; d(f(\mathbf{x}), f(\mathbf{y})) < \epsilon \Big).$$

Another important reformulation of continuity builds on the concept of preimage. Let $f : \mathcal{X} \to \mathcal{Y}$ be given. The **preimage** of a set $V \subset \mathcal{Y}$ is

$$f^{-1}(V) \;:=\; \{\mathbf{x} \in \mathcal{X} : \; f(\mathbf{x}) \in V\}. \tag{A.1}$$

For example, if $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by $f(x, y) = x^2 + y^2 + 2$ then the preimage of the interval $[3, 6]$ is the annulus in the plane with inner radius 1 and outer radius 2.

Theorem A.2.4. *A function $f : \mathcal{X} \to \mathcal{Y}$ is continuous if and only if the preimage $f^{-1}(V)$ of any open set $V \subset \mathcal{Y}$ is open in $\mathcal{X}$.*

A.2.2 Semicontinuity and Fatou's lemma

A closely related notion to continuity of a function is that of semicontinuity. Recall that $\overline{\mathbb{R}} = [-\infty, +\infty]$ is the set of extended real numbers with the usual definition of algebraic operations that incorporate the two extra "numbers" $-\infty, +\infty$. Let $\mathcal{X}$ be any metric space.

Definition A.2.5. *A function $f : \mathcal{X} \to \overline{\mathbb{R}}$ is **lower semicontinuous at** p if, for $\epsilon > 0$, p is an interior point of $\{x : f(x) > f(p) - \epsilon\}$. Similarly, f is **upper semicontinuous at** $p \in \mathcal{X}$ if, for every real $\epsilon > 0$, p is an interior point of the set $\{x : f(x) < f(p) + \epsilon\}$.*

Recall that for a real sequence $(s_n)_{n\in\mathbb{N}}$ the set E of subsequential limits is the set of $s \in \mathbb{R}$ such that some subsequence of (s_n) converges to s. We define $\liminf_{n\to\infty} s_n = \inf E$ and $\limsup_{n\to\infty} s_n = \sup E$. This definition can be extended to functional limits. Consider all sequences $x_n \to p$ and the corresponding converging subsequences $f(x_{n_k})$. Then $\liminf_{x\to p} f(x)$ is the inferior of the set of all such subsequential limits. The upper functional limit $\limsup_{x\to p} f(x)$ is defined analogously.

There are important alternative ways to formulate semi-continuity.

Exercise A.2.6. *Show that $f : \mathcal{X} \to \overline{\mathbb{R}}$ is lower semicontinuous at $p \in \mathcal{X}$ if and only if $\liminf_{x\to p} f(x) \geq f(p)$. Similarly, $f : \mathcal{X} \to \overline{\mathbb{R}}$ is upper semicontinuous at $p \in \mathcal{X}$ if and only if $\limsup_{x\to p} f(x) \leq f(p)$.*

This exercise immediately gives us the following result.

Theorem A.2.7. *A function $f : \mathcal{X} \to \mathbb{R}$ is continuous at p if and only if it is both upper and lower semicontinuous at p.*

We say that a function is lower (upper) semicontinuous if it is lower (upper) semicontinuous at every point $p \in \mathcal{X}$.

Exercise A.2.8. *Show that $f : \mathcal{X} \to \overline{\mathbb{R}}$ is lower semicontinuous if and only if $\{x : f(x) > y\}$ is open for every $y \in \mathbb{R}$, or equivalently, when $\{x : f(x) \leq y\}$ is closed.*

Lower semi-continuity[1] is an essential concept in convex analysis. In probability it appears, for example, in the Fatou's lemma, which we will use in later chapters.

[1] For more on semicontinuity see https://pzwiernik.github.io/docs/RealAnalysisNotes.pdf.

Theorem A.2.9 (Fatou's Lemma)**.** *Given a measure space $(\mathcal{X}, \mathcal{F}, \mu)$ and a set $U \in \mathcal{F}$, let $\{f_n\}$ be a sequence of measurable nonnegative functions $f_n : U \to [0, +\infty]$. Define the function $f : U \to [0, \infty]$ by setting*

$$f(x) = \liminf_{n\to\infty} f_n(x) \qquad \text{for all } x \in U.$$

Then f is measurable, and also

$$\int_U f \mathrm{d}\mu \leq \liminf_{n\to\infty} \int_U f_n(x) \mathrm{d}\mu,$$

where the integrals may be infinite.

In particular, this last statement is saying that $Z(f) = \int_{\mathcal{X}} f \mathrm{d}\mu$ is lower semicontinuous. Namely, for every measurable f and a sequence f_n that converges pointwise to f ($\lim_n f_n(x) = f(x)$ for all $x \in \mathcal{X}$),

$$Z(f) \leq \liminf_n Z(f_n)$$

(c.f Exercise A.2.6).

A.3 Differentiation

Given a function $f : V \to R$, from a vector space V, we may be interested in a local behaviour of f around some point $\boldsymbol{a} \in V$. The fundamental tool is given by the directional derivative. If $\boldsymbol{u}$ is a vector in V, then the directional derivative of f at $\boldsymbol{a} \in V$ in the direction $\boldsymbol{u} \in V$ is

$$\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{a}) := \lim_{t \to 0} \frac{f(\boldsymbol{a} + t\boldsymbol{u}) - f(\boldsymbol{a})}{t}. \tag{A.2}$$

If the directional derivative $\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{a})$ exists, it depends only on the behaviour of f in a small open neighbourhood $U \subset V$ of $\boldsymbol{a}$ and so the function does not need to be defined over the whole V. The sign of the derivative has an important interpretation: if $\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{a}) > 0$ then the value of the function increases as we move infinitesimally from $\boldsymbol{a}$ in the direction $\boldsymbol{u}$.

Remark A.3.1. *In infinite dimensional spaces the directional derivative is typically called the Gateaux derivative.*

If $f : V \to \mathbb{R}$ is differentiable at $\boldsymbol{a}$ the directional derivarives exist and then

$$f(\boldsymbol{a} + t\boldsymbol{u}) = f(\boldsymbol{a}) + t\,\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{a}) + r(t\boldsymbol{u}),$$

where the remainder r satisfies $\lim_{t \to 0} \frac{r(t\boldsymbol{u})}{t} = 0$ (in other words $r(t\boldsymbol{u}) = o(t)$). This gives a simple "algebraic" way of computing the directional derivatives. Before we give some examples, we note that with a choice of an inner product on V, we get

$$\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{a}) = \langle \nabla f(\boldsymbol{a}), \boldsymbol{u} \rangle, \tag{A.3}$$

where $\nabla f(\boldsymbol{a}) \in V$ denotes the gradient of f at $\boldsymbol{a}$. '

Exercise A.3.2. *Use the Cauchy-Schwarz inequality and (A.3) to show that $\nabla f(\boldsymbol{a})$ is the direction of the steepest increase of f locally around $\boldsymbol{a}$.*

Example A.3.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be given by $f(\mathbf{x}) = \mathbf{x}^\intercal A \mathbf{x}$ for $A \in \mathbb{S}^d$. Then*

$$f(\boldsymbol{x} + t\boldsymbol{u}) - f(\boldsymbol{x}) = t(\boldsymbol{u}^\intercal A\boldsymbol{x} + \boldsymbol{x}^\intercal A\boldsymbol{u}) + o(t),$$

which gives that $\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{x}) = \boldsymbol{u}^\intercal A\boldsymbol{x} + \boldsymbol{x}^\intercal A\boldsymbol{u} = \langle 2A\boldsymbol{x}, \boldsymbol{u} \rangle$ and so $\nabla f(\boldsymbol{x}) = 2A\boldsymbol{x}$.

Example A.3.4. *Let $f : \mathbb{R}^{n \times n} \to \mathbb{R}$ be defined by $f(A) = \mathrm{tr}(A^2)$. Then*

$$f(A + tU) - f(A) = t(\mathrm{tr}(UA) + \mathrm{tr}(AU)) + o(t),$$

which gives that $\mathrm{D}_U f(A) = 2\mathrm{tr}(UA) = \langle 2A^\intercal, U \rangle$ and so $\nabla f(A) = 2A^\intercal$.

A slightly more involved matrix example is given as an exercise.

Exercise A.3.5. *Consider the function* $f(\Sigma) = \log\det(\Sigma)$ *defined for all* $\Sigma \in \mathbb{S}^m$ *that are positive definite. Find the gradient of this function (it is an element of* $\mathbb{S}^m$*). Hint: Computing* $\mathrm{D}_U f(\Sigma)$ *consider the eigenvalues of* $\Sigma^{-1/2}U\Sigma^{-1/2}$.

We also have an infinite-dimensional example.

Example A.3.6. *Let* $f : L^2(\mathbb{R}) \to \mathbb{R}$ *be defined by* $f(\varphi) = \int \varphi^2(x)\mathrm{d}x = \|\varphi\|^2$ *then, for every* $u \in L^2(\mathbb{R})$,

$$f(\varphi + tu) - f(\varphi) \;=\; 2t\int \varphi(x)u(x)\mathrm{d}x + o(t),$$

which gives that $\mathrm{D}_u f(\varphi) = 2\int \varphi(x)u(x)\mathrm{d}x = 2\langle\varphi, u\rangle$ *and so* $\nabla f(\varphi) = 2\varphi$.

The infinite dimensional case is extremely important in semiparametric and nonparametric statistics but it also appears in the theory of optimal statistical procedures. This may require a bit more careful treatment but we will introduce relevant concepts of functional analysis along the way. The message we tried to convey above is that very often it is useful to think about this inifite dimensional case as a special case of the standard analysis on $\mathbb{R}^d$.

To conclude this section we generalize (A.3) by defining the derivative as a linear map.

Definition A.3.7. *Suppose* U *is open in* $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^m$. *The function* $\boldsymbol{f}$ *is differentiable at* $\boldsymbol{a} \in U$ *with derivative* $\boldsymbol{f}'_{\boldsymbol{a}}$ *if* $\boldsymbol{f}'_{\boldsymbol{a}} : \mathbb{R}^n \to \mathbb{R}^m$ *is a linear function and*

$$\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) \;=\; \boldsymbol{f}(\boldsymbol{a}) + \boldsymbol{f}'_{\boldsymbol{a}}(\boldsymbol{h}) + \boldsymbol{r}(\boldsymbol{h}),$$

where the remainder $\boldsymbol{r}(\boldsymbol{h})$ *satisfies*

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}_n} \frac{\boldsymbol{r}(\boldsymbol{h})}{\|\boldsymbol{h}\|} \;=\; \boldsymbol{0}_m.$$

If $\boldsymbol{f}$ *is differentiable at every* $\boldsymbol{a} \in U$, *we say that* $\boldsymbol{f}$ *is differentiable in* U.

Exercise A.3.8. *Check if the above examples are differentiable.*

Proposition A.3.9. *Let* $\boldsymbol{f} : U \to \mathbb{R}^m$, *where* U *is an open subset in* $\mathbb{R}^n$, *be differentiable at* $\boldsymbol{a} \in U$. *If* $\boldsymbol{u} \in \mathbb{R}^n$ *then*

$$\mathrm{D}_{\boldsymbol{u}}\boldsymbol{f}(\boldsymbol{a}) \;=\; \boldsymbol{f}'_{\boldsymbol{a}}(\boldsymbol{u}).$$

Proof. Since $\boldsymbol{f}'_{\boldsymbol{a}}$ exists

$$f(\boldsymbol{a}+t\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{a}) \;=\; \boldsymbol{f}'_{\boldsymbol{a}}(t\boldsymbol{u}) + \boldsymbol{r}(t\boldsymbol{u})$$

with $\frac{\boldsymbol{r}(t\boldsymbol{u})}{t} \to \boldsymbol{0}$ as $t \to 0$. Dividing by t and taking the limit, we get that

$$\boldsymbol{f}'_{\boldsymbol{a}}(\boldsymbol{u}) \;=\; \lim_{t\to 0}\frac{\boldsymbol{f}(\boldsymbol{a}+t\boldsymbol{u})-\boldsymbol{f}(\boldsymbol{a})}{t} \;=\; \mathrm{D}_{\boldsymbol{u}}\boldsymbol{f}(\boldsymbol{a}).$$

□

Exercise A.3.10. *Given the same set-up as in Proposition A.3.9, show that* $\boldsymbol{f}'_{\boldsymbol{a}}(\boldsymbol{u}) = \mathrm{J}\boldsymbol{f}(\boldsymbol{a}) \cdot \boldsymbol{u}$*, where* $\mathrm{J}\boldsymbol{f}(\boldsymbol{a})$ *is the Jacobian of* $\boldsymbol{f}$ *at* $\boldsymbol{a}$*, that is* $\mathrm{J}\boldsymbol{f}(\boldsymbol{a}) \in \mathbb{R}^{m \times n}$ *and its* (i, j)*-th entry is* $\frac{\partial}{\partial x_j} f_i$*.*

B
Convex Analysis

B.1 Convexity and hyperplane separation

Let V be a vector space. Recall that a set $C \subseteq V$ is *convex* if for any two points $\boldsymbol{x}, \boldsymbol{y} \in C$ and for all $\lambda \in (0,1)$

$$\boldsymbol{z}_\lambda \;:=\; (1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y} \;\in\; C.$$

The point $\boldsymbol{z}_\lambda$ can be rewritten as $\boldsymbol{z}_\lambda = \mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})$ so, as λ varies from 0 to 1, $\boldsymbol{z}_\lambda$ moves from $\mathbf{x}$ to $\mathbf{y}$ along the segment joining $\mathbf{x}$ and $\mathbf{y}$. This gives a geometric interpretation of convex sets: for any two points in the set, the segment between these two points is contained in the set.

Given a non-empty subset $C \subset \mathbb{R}^k$ we define the **distance to** C function $d_C : \mathbb{R}^k \to \mathbb{R}$ by

$$d_C(\boldsymbol{x}) \;:=\; \inf_{\boldsymbol{y} \in C} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

This function is well defined because for every $\boldsymbol{x} \in \mathbb{R}^k$ the set $\{\|\boldsymbol{x} - \boldsymbol{y}\| \;:\; \boldsymbol{y} \in C\} \subset \mathbb{R}$ is bounded from below (by zero) and so its infimum is well-defined. The following result is fundamental for many applications of convex analysis.

Theorem B.1.1 (Minimum distance to a set)**.** *(1) Let $E, F \subset \mathbb{R}^k$. Then $d_F : E \to \mathbb{R}$ is Lipschitz continuous function with constant 1 (in particular, it is continuous).*

(2) If F is closed, then

$$\forall \boldsymbol{x} \in E \;\; \exists \boldsymbol{y} \in F \text{ such that } \;\; d_F(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{y}\|.$$

(3) If F is also convex, for every $\boldsymbol{x}$ such $\boldsymbol{y}$ is unique in F.

Proof. (1) Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in E$. By the triangle inequality, $\|\boldsymbol{x}_1 - \boldsymbol{y}\| \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \|\boldsymbol{x}_2 - \boldsymbol{y}\|$. If $\boldsymbol{y} \in F$ then $d_F(\boldsymbol{x}_1) \leq \|\boldsymbol{x}_1 - \boldsymbol{y}\|$ and so

$$d_F(\boldsymbol{x}_1) \;\leq\; \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \|\boldsymbol{x}_2 - \boldsymbol{y}\|.$$

Since the inequality holds for every $y \in F$. Take infimum over all $y \in F$ to get that $d_F(x_1) \leq \|x_1 - x_2\| + d_F(x_2)$. In the same way, starting with $\|x_2 - y\| \leq \|x_1 - x_2\| + \|x_1 - y\|$, conclude that $d_F(x_2) \leq \|x_1 - x_2\| + d_F(x_1)$. It follows that $|d_F(x_1) - d_F(x_2)| \leq \|x_1 - x_2\|$, which implies Lipschitz continuity of d_F.

(2) Let $x \in E$ and fix $y_0 \in F$. We have $d_F(x) \leq \|x - y_0\| = r$. Define $\tilde{F} = F \cap \overline{N_r(x)}$, where $\overline{N_r(x)}$ is the closed ball of radius r around $\mathbf{x}$. Since F is closed and $\overline{N_r(x)}$ is compact, $\tilde{F}$ is also compact. Since $\|x - y\|$ is a continuous function of y, there exists $y_1 \in \tilde{F}$ such that $\inf_{y \in F} \|x - y\| = \|x - y_1\|$.

(3) Let y_1 and y_2 in F be such that $\|x - y_1\| = \|x - y_2\| = d_F(x)$. Define $p = y_2 - y_1$ and $h : [0,1] \to \mathbb{R}$ by $h(\lambda) = \|x - y_1 - \lambda p\|^2$. We have $h(0) = h(1)$ and also, because F is convex, $y_1 + \lambda p \in F$ for $\lambda \in [0,1]$ and so h is *minimized* at $\lambda = 0$ and $\lambda = 1$. Since $h(\lambda)$ is a quadratic function with nonnegative coefficient $\|y_1 - y_2\|^2$ of λ^2, this is only possible if $y_1 = y_2$. □

Proposition B.1.2. *Let $E, F \subset \mathbb{R}^k$ with F closed and convex. Let $g : E \to F$ be given by $g(x) = \arg\inf_{y \in F} \|x - y\|$. Then g is a well-defined and*

$$\|g(x_2) - g(x_1)\| \leq \|x_2 - x_1\| \qquad \text{for all } x_1, x_2 \in E. \tag{B.1}$$

In particular, g is a continuous function.

Proof. Because F is closed and convex, Theorem B.1.1 assures that g is a well-defined function, that is, for each $x \in E$ there is a unique $y \in F$ such that $g(x) = y$. To show (B.1), take $p = g(x_2) - g(x_1)$;

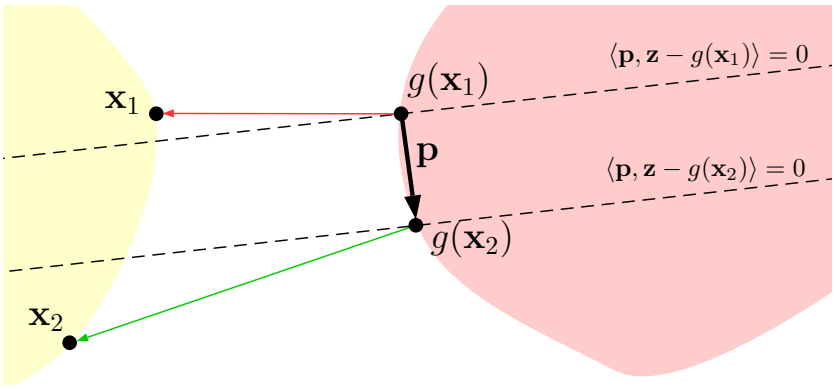


Figure B.1: Illustration of the proof of Proposition B.1.2

c.f. Figure B.1. The set of points $g(x_1) + tp$ for $t \in [0,1]$ lies in F, by convexity, and so

$$h(t) = \|x_1 - g(x_1) - tp\|^2$$

has a minimum at $t = 0$. This is a quadratic function in t with a strictly positive coefficient of t^2. The only way for such a function to

have a minimum at $t = 0$ is that its derivative at $t = 0$ is nonnegative, or, in other words, the coefficient of t is nonnegative. This coefficient is $-2\langle \boldsymbol{p}, \boldsymbol{x}_1 - g(\boldsymbol{x}_1)\rangle$, which implies that $\langle \boldsymbol{p}, \boldsymbol{x}_1 - g(\boldsymbol{x}_1)\rangle \leq 0$. In a similar way, we show that $\langle \boldsymbol{p}, \boldsymbol{x}_2 - g(\boldsymbol{x}_2)\rangle \geq 0$. But these two inequalities imply that

$$\langle \boldsymbol{p}, \boldsymbol{x}_2 - \boldsymbol{x}_1\rangle \;\geq\; \langle \boldsymbol{p}, g(\boldsymbol{x}_2) - g(\boldsymbol{x}_1)\rangle \;=\; \|\boldsymbol{p}\|^2.$$

The Cauchy-Schwarz inequality gives that $\langle \boldsymbol{p}, \boldsymbol{x}_2 - \boldsymbol{x}_1\rangle \leq \|\boldsymbol{p}\|\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|$. This implies that $\|\boldsymbol{p}\| \leq \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|$, which is precisely (B.1).

□

We say that two sets $A, B \subset \mathbb{R}^k$ are **separated by a hyperplane** if there exists $\boldsymbol{p} \neq \boldsymbol{0}_k$ and $c \in \mathbb{R}$ such that for every $\boldsymbol{x} \in A$, $\boldsymbol{y} \in B$ we have $\langle \boldsymbol{p}, \boldsymbol{x}\rangle \leq c \leq \langle \boldsymbol{p}, \boldsymbol{y}\rangle$. The separation is strict if the inequalities can be made strict. The following theorem is one of the most important results of convex geometry.

Theorem B.1.3 (Hyperplane Separation Theorem)**.** *Let C and K be disjoint non-empty convex sets in $\mathbb{R}^k$. Then they are separated by a hyperplane. Let C be closed and K compact. Then C and K are strictly separated by a hyperplane.*

Note that compactness of K in the second part of Theorem B.1.3 is necessary. For example $A = \{\boldsymbol{x} \in \mathbb{R}^2 :\; x_1 \leq 0\}$, $B = \{\boldsymbol{x} \in \mathbb{R}^2 : x_1 > 0, x_2 \geq 1/x_1\}$ are closed but not strictly separated.

Proof. We prove the second part of the theorem leaving the first part as an exercise. By Theorem B.1.1, $d_C(\boldsymbol{x})$ is a continuous real-valued function on K and so it achieves its minimum. Call it $\boldsymbol{x}_0 \in K$. By Theorem B.1.1(c) there is exactly one $\boldsymbol{y}_0 \in C$ such that $d_C(\boldsymbol{x}_0) = \|\boldsymbol{x}_0 - \boldsymbol{y}_0\|$. Set $\boldsymbol{p} = \boldsymbol{x}_0 - \boldsymbol{y}_0$. Then $\boldsymbol{p} \neq \boldsymbol{0}_k$ and $0 < \|\boldsymbol{p}\|^2 = \langle \boldsymbol{p}, \boldsymbol{x}_0 - \boldsymbol{y}_0\rangle$ so

$$\langle \boldsymbol{p}, \boldsymbol{x}_0\rangle \;>\; \langle \boldsymbol{p}, \boldsymbol{y}_0\rangle$$

so it suffices to show that $\langle \boldsymbol{p}, \boldsymbol{x}\rangle \geq \langle \boldsymbol{p}, \boldsymbol{x}_0\rangle$ for every $\boldsymbol{x} \in K$ and $\langle \boldsymbol{p}, \boldsymbol{y}\rangle \leq \langle \boldsymbol{p}, \boldsymbol{y}_0\rangle$ for all $\boldsymbol{y} \in C$. We show the second, the first is similar and follows from convexity. Let $\boldsymbol{y} \in C$ and set $\boldsymbol{y}_\lambda = (1 - \lambda)\boldsymbol{y}_0 + \lambda \boldsymbol{y}$. Then

$$\boldsymbol{x}_0 - \boldsymbol{y}_\lambda \;=\; \boldsymbol{x}_0 - \boldsymbol{y}_0 - \lambda(\boldsymbol{y} - \boldsymbol{y}_0) \;=\; \boldsymbol{p} - \lambda(\boldsymbol{y} - \boldsymbol{y}_0)$$

and so

$$\begin{aligned} \|\boldsymbol{x}_0 - \boldsymbol{y}_\lambda\|^2 &= \|\boldsymbol{p} - \lambda(\boldsymbol{y} - \boldsymbol{y}_0)\|^2 = \\ &= \lambda^2\|\boldsymbol{y} - \boldsymbol{y}_0\|^2 - 2\lambda\langle \boldsymbol{p}, \boldsymbol{y} - \boldsymbol{y}_0\rangle + \|\boldsymbol{p}\|^2. \end{aligned}$$

This is a quadratic function of λ that achieves its minimum at $\lambda = 0$ (by construction!). This implies that the derivative of this function

at zero must be nonnegative. This derivative is equal to the coefficient of λ which is $2\langle \boldsymbol{p}, \boldsymbol{y}_0 - \boldsymbol{y}\rangle$. This implies that $\langle \boldsymbol{p}, \boldsymbol{y}\rangle \leq \langle \boldsymbol{p}, \boldsymbol{y}_0\rangle$ as claimed. □

Exercise B.1.4. *Show that each closed convex set is an intersection of closed half-spaces.*

B.2 Convex functions and optimization

A function f defined on a convex set $C \subseteq V$ with values in $\overline{\mathbb{R}}$ is convex if

$$f\left((1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y}\right) \leq (1-\lambda) f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) \qquad \text{for all } \boldsymbol{x} \neq \boldsymbol{y} \in C, \lambda \in (0,1).$$

Moreover, f is strictly convex if the inequality is always strict. A function f is (strictly) concave if $-f$ is (strictly) convex.

Exercise B.2.1. *Let $\mathbb{S}_+^m$ be the set of all symmetric $m \times m$ that are positive definite. Show that the following functions are convex:*

(a) $f : \mathbb{R}^d \to \mathbb{R}$ *defined by* $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$.

(b) $f : \mathbb{S}_+^m \to \mathbb{R}$ *defined by* $f(\Sigma) = -\log\det(\Sigma)$.

(c) $L : L^2(\mathbb{R}) \to \mathbb{R}$ *defined by* $L(f) = \int_{-\infty}^{\infty} f^2(x)\mathrm{d}x$.

Are they strictly convex?

We now discuss the most important features of convex function. Note that we never assume that V is a finite-dimensional space.

Proposition B.2.2. *If $f : C \to \overline{\mathbb{R}}$ is convex and $\mathcal{L} \subset V$ is a linear subspace then f restricted to $C \cap \mathcal{L}$ is also convex.*

Proof. This follows directly from the definition. □

Theorem B.2.3 (Jensen's inequality)**.** *Suppose $f : C \to \mathbb{R}$ is a convex function and let X be a random variable with $\mathbb{P}(X \in C) = 1$ and $\mathbb{E}X < \infty$. Then $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$. If f is strictly convex then the inequality is strict unless X is constant almost surely.*

Proposition B.2.4. *If $\boldsymbol{a}$ is a local optimum of a convex function f, then f is a global optimum.*

Proof. We argue by contradiction. If $\boldsymbol{x}$, $\boldsymbol{y}$ are two local optima with $f(\boldsymbol{x}) < f(\boldsymbol{y})$ ($\boldsymbol{y}$ is local but not global) then, for every $\lambda \in (0,1)$, $f(\boldsymbol{z}_\lambda) < f(\boldsymbol{y})$. This contradicts local optimality of $\boldsymbol{y}$. □

Proposition B.2.5. *Let $g : C \times \mathcal{A} \to \mathbb{R}$ be such that $g(x, \alpha)$ is convex in $x \in C$ for every fixed α. Then the function*

$$f(x) := \sup_{\alpha \in \mathcal{A}} g(x, \alpha)$$

(defined as the pointwise supremum) is a convex function.

Proof. Exercise. □

Example B.2.6 (Fenchel conjugate). *Let $f(x)$ be a convex function on C. Then the function*

$$g(x,y) \;=\; \langle x,y\rangle - f(x)$$

is concave in x and linear (and so also convex) in y. Using Proposition B.2.5 we get

$$f^*(y) \;=\; \sup_{x\in C}\{\langle x,y\rangle - f(x)\}$$

is a convex function. The function f^* is called the Fenchel conjugate of f. Note that the Fenchel inequality follows easily:

$$f(x)+f^*(y) \;\geq\; \langle x,y\rangle \qquad \text{for all } x,y.$$

For a simple example, take $f(x)=x^2$.

Proposition B.2.4 suggests that optimizing convex functions is much easier than optimizing general functions. However, the study of local behaviour of convex functions is easy also for a different reason. Define a one-sided directional derivative

$$\mathrm{D}^+_{\boldsymbol{u}} f(\boldsymbol{a}) \;=\; \lim_{t\to 0^+}\frac{f(\boldsymbol{a}+t\boldsymbol{u})-f(\boldsymbol{a})}{t}. \tag{B.2}$$

If the directional derivative $\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{a})$ in (A.2) exists then the one-sided derivative exists and they are equal. But there are important examples when the directional derivative does not exist but the one-sided derivative does.

Exercise B.2.7. *Show that at the origin none of the directional derivatives of $f(\mathbf{x})=\|\mathbf{x}\|_1$ exists but all the one-sided derivatives do.*

The importance of one-sided directional derivatives comes from the fact that if $\mathrm{D}^+_{\boldsymbol{u}} f(\boldsymbol{a})$ is positive then the function is increasing when we move infinitesimally from $\boldsymbol{a}$ in the direction $\boldsymbol{u}$. This can be used to easily provide necessary conditions for a local optimum even in the constrained setting.

Theorem B.2.8. *Let $f: C\to\mathbb{R}\cup\{\infty\}$ be convex, and let $\mathbf{x},\mathbf{y}\in\mathrm{dom}(f)$. Then for every $\boldsymbol{z}_\lambda = (1-\lambda)\mathbf{x}+\lambda\mathbf{y}$, $\lambda\in(0,1)$, the one-sided derivative $\mathrm{D}^+_{\mathbf{x}-\mathbf{y}} f(\boldsymbol{z}_\lambda)$ exists and is an increasing function of λ.*

We claim that this result is true also over infinite-dimensional spaces. Our proof of Theorem B.2.8 starts with the following lemma.

Lemma B.2.9. *Let $\varphi:[0,1]\to\mathbb{R}$ be convex. Consider the "chords" $0\leq p<q<1$, and $0<s<t\leq 1$, with $p\leq s$ and $q\leq t$. Then*

$$\frac{\varphi(q)-\varphi(p)}{q-p}\leq\frac{\varphi(t)-\varphi(s)}{t-s}.$$

Proof. Since $p < q \leq t$, then

$$q = \frac{t-q}{t-p}p + \frac{q-p}{t-p}t$$

is a convex combination of p and t, so the convexity of φ gives

$$\varphi(q) \leq \frac{t-q}{t-p}\varphi(p) + \frac{q-p}{t-p}\varphi(t).$$

This is equivalent to

$$\frac{\varphi(q)-\varphi(p)}{q-p} \leq \frac{\varphi(t)-\varphi(p)}{t-p}. \tag{B.3}$$

Now, apply this inequality to the function $\psi(\lambda) = \varphi(1-\lambda)$ and the points $1-t < 1-s \leq 1-p$. The result may be rewritten as

$$\frac{\varphi(t)-\varphi(p)}{t-p} \leq \frac{\varphi(t)-\varphi(s)}{t-s}.$$

Together with (B.3), this is the required inequality. □

Proof of Theorem B.2.8. Let $\varphi(\lambda) = f(z_\lambda)$ and apply Lemma B.2.9 for the points $p = s = \lambda < q \leq t$ to obtain

$$\frac{\varphi(q)-\varphi(\lambda)}{q-\lambda} \leq \frac{\varphi(t)-\varphi(\lambda)}{t-\lambda}.$$

It follows that $\psi(q) = (\varphi(q)-\varphi(\lambda))/(q-\lambda)$ is an increasing function of q, $q > \lambda$. Note that

$$\psi(q) = t^{-1}\{f(z_\lambda + t(\mathbf{x}-\mathbf{y})) - f(z_\lambda)\}, \quad t = q-\lambda.$$

Applying Lemma B.2.9 to the points $p < s = q = \lambda < t$ shows that

$$\psi(p) = \frac{\varphi(\lambda)-\varphi(p)}{\lambda-p} \leq \frac{\varphi(t)-\varphi(\lambda)}{t-\lambda} = \psi(t).$$

Hence, $\psi(t)$ is increasing and bounded below to $t > \lambda$, and thus,

$$\lim_{t\to\lambda^+} \psi(t)$$

exists. Equivalently the one-sided derivative $D^+_{\mathbf{x}-\mathbf{y}}(z_\lambda)$ exists for $\lambda \in (0,1)$. By taking limits in the inequality in Lemma B.2.9, namely $q \to p^+$, $t \to s^+$, we see that $\varphi'(p) \leq \varphi'(s)$ (one-sided derivatives) for almost all p and s with $p \leq s$. Thus $D^+_{\mathbf{x}-\mathbf{y}}(z_\lambda)$ is an increasing function of λ. □

Exercise B.2.10. *Show that any convex function defined and finite on a convex set C must be continuous on its interior. Although this result is general, for simplicity, you can focus on its one-dimensional version.*

We now informally state an important result, which gives a fundamental understanding of a large family of optimization problems. Often we optimize a function over a convex set C given by bunch of linear constraints together with inequality constraints $g_i(x) \leq 0$ for $i \in \mathcal{I}$ with g_i convex. If the functions g_i are all continuously differentiable and all one-sided derivatives in (B.2) exist then $\boldsymbol{a}$ is a local minimum of f if and only if $\mathrm{D}^+_{\boldsymbol{u}} f(\boldsymbol{a}) \geq 0$ for all $\boldsymbol{u}$ such that $\mathrm{D}_{\boldsymbol{u}} g_i(\boldsymbol{a}) \leq 0$ for all i such that $g_i(\boldsymbol{a}) = 0$ (active constraints).

C
Probability

C.1 Continuity of probability

In what follows we fix a probabiliy space $(\Omega, \mathcal{B}, \mathbb{P})$. If A_n is a sequence of events then we say that (A_n) increases to A if $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$ and $A = \bigcup_{n\geq 1} A_n$.

Proposition C.1.1. *If (A_n) is a sequence of events increasing to A, then*

$$\lim_{n\to\infty} \mathbb{P}(A_n) \;=\; \mathbb{P}(A).$$

Proof. Clearly $A_n = A_n \cap A_{n+1}$ for all $n \in \mathbb{N}$. Let $A_0 = \emptyset$, $C_1 = A_1$, and define $C_{n+1} = A_{n+1} \setminus A_n$. Notice that $C_1, C_2, \ldots$ are disjoint with

$$\bigcup_{j=1}^{n} C_j = A_n \quad \text{and} \quad \bigcup_{j=1}^{\infty} C_j = \bigcup_{j=1}^{\infty} A_j = A.$$

Using the fact that $\mathbb{P}$ is countably additive, we conclude

$$\mathbb{P}(A) \;=\; \mathbb{P}(\bigcup_{j=1}^{\infty} C_j) \;=\; \sum_{j=1}^{\infty} \mathbb{P}(C_j) \;=\; \lim_{n\to\infty} \sum_{j=1}^{n} \mathbb{P}(C_j) \;=\; \lim_{n\to\infty} \mathbb{P}(\bigcup_{j=1}^{n} C_j) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

□

This immediately gives the following result.

Proposition C.1.2. *Both the CDF $F(t) = \mathbb{P}(X \leq t)$ of X and its survival function $G(t) = \mathbb{P}(X > t)$ are right-continuous.*

Proof. Let (t_n) be any monotone sequence such that $t_n > t$ and $t_n \to t$. Then the events $A_n = \{X > t_n\}$ satisfy $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$. Since $\bigcup_{n\in\mathbb{N}} A_n = \{X > t\}$ we have that (A_n) increases to $\{X > t\}$. By Proposition C.1.1,

$$\lim_{n\to\infty} \mathbb{P}(X > t_n) \;=\; P(X > t).$$

Since t_n is otherwise arbitrary, we conclude $\lim_{x\to t^+} G(x) = G(t)$ and so the survival function is right-continuous. The claim for the CDF follows immediately as

$$\lim_{x\to t^+} F(x) = \lim_{x\to t^+}(1-G(x)) = 1 - \lim_{x\to t^+} G(x) = 1 - G(t) = F(t).$$

□

Remark C.1.3. *Using the same approach we can show that the functions $\mathbb{P}(X < t)$ and $\mathbb{P}(X \geq t)$ are left-continuous. We leave it as an exercise.*

C.2 Martingales

Let $X_1, \ldots, X_n$ be independent random variables with values in $\mathcal{X}$. Let $\{\mathcal{F}_k\}_{k=1}^{\infty}$ be a sequence of σ-fields, $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$. Let $\{Y_k\}_{k=1}^{\infty}$ be a sequence of variables such that Y_k is $\mathcal{F}_k$-measurable (we say that $\{Y_k\}_{k=1}^{\infty}$ is adapted to the filtration $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$).

Definition C.2.1. *Given a sequence $\{Y_k\}_{k=1}^{\infty}$ adapted to the filtration $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, the pair $\{(Y_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ is a **martingale** if, for all $k \geq 1$,*

$$\mathbb{E}(|Y_k|) < \infty \quad \textit{and} \quad \mathbb{E}[Y_{k+1}|\mathcal{F}_k] = Y_k.$$

Example C.2.2 (Simple random walk)**.** *A particle jumps either one step to the right or one step to the left with the corresponding probabilities p and $q = 1 - p$. Assume that the subsequent moves are independent of each other. Define $S_n = X_1 + \ldots + X_n$. It is clear that $\mathbb{E}|S_n| \leq n$ and*

$$\mathbb{E}[S_{n+1}|X_1, \ldots, X_n] = S_n + (p - q),$$

and so $Y_n = S_n - n(p-q)$ defined a martingale with respect to X.

Example C.2.3 (Likelihood ratio)**.** *Let p_1, p_0 be two mutually absolutely continuous densities, and let $X_1, X_2, \ldots$ be an i.i.d. sequence from p_0. For each $k \in \mathbb{N}$ let $Y_k = \prod_{i=1}^{k} \frac{p_1(X_i)}{p_0(X_i)}$ be the likelihood ratio based on the first k samples (c.f (3.15)). Then the sequence is a martingale with respect to $\{X_k\}_{k=1}^{\infty}$. Indeed,*

$$\mathbb{E}[Y_{k+1}|\mathcal{F}_k] = \mathbb{E}\left[\frac{p_1(X_{k+1}}{p_0(X_{k+1})}\right] \prod_{i=1}^{k} \frac{p_1(X_i)}{p_0(X_i)} = Y_k,$$

using the fact that $\mathbb{E}[\frac{p_1(X_{k+1}}{p_0(X_{k+1})}] = 1$.

There are many cases of interest in which the martingale condition $\mathbb{E}(Y_{k+1|\mathcal{F}_k}) = Y_k$ does not hold, being replaced instead by an inequality: $\mathbb{E}(Y_{k+1}|\mathcal{F}_k) \geq Y_k$ for all k, or by $\mathbb{E}(Y_{k+1}|\mathcal{F}_k) \leq Y_k$ for all k. Sequences satisfying such inequalities have many of the properties of martingales. Recall $x^+ = \max\{0, x\}$ and $x^- = -\min\{0, x\}$.

Definition C.2.4. *Given a sequence $\{Y_k\}_{k=1}^{\infty}$ adapted to the filtration $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, the pair $\{(Y_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ is a* ***submartingale*** *if, for all $k \geq 1$,*

$$\mathbb{E}(Y_k^+) < \infty \quad \textit{and} \quad \mathbb{E}[Y_{k+1}|\mathcal{F}_k] \geq Y_k,$$

or a ***supermartingale*** *if, for all $k \geq 1$,*

$$\mathbb{E}(Y_k^-) < \infty \quad \textit{and} \quad \mathbb{E}[Y_{k+1}|\mathcal{F}_k] \leq Y_k,$$

We call the pair $\{(S_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ predictable if S_{k+1} is $\mathcal{F}_k$-measurable for all k. We call a predictable process $\{(S_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ increasing if $S_1 = 0$ and $\mathbb{P}(S_{k+1} \geq S_k) = 1$ for all k.

Theorem C.2.5 (Doob decomposition)**.** *A submartingale Y_k with finite means may be expressed in the form*

$$Y_k \;=\; M_k + S_k,$$

where M_k is a martingale and S_k is an increasing predictable process. This decomposition is unique.

A closely related notion to martingales is that of a **martingale difference sequence**, which is an adapted sequence $\{\Delta_k, \mathcal{F}_k\}_{k=1}^{\infty}$ such that for all $k \geq 1$,

$$\mathbb{E}|\Delta_k| \leq \infty \quad \text{and} \quad \mathbb{E}(\Delta_{k+1}|\mathcal{F}_k) = 0.$$

If $\{Y_k\}$ is a martingale then $\Delta_k \;=\; Y_k - Y_{k-1}$ is a martingale difference sequence. In our case, this easily follows from the fact that $\mathbb{E}_k(\mathbb{E}_{k+1}(Z)) = \mathbb{E}_k Z$.