AN INTRODUCTION TO ESTIMATION AND DETECTION THEORY[1]

Armand M. Makowski [2] and Prakash L.K. Narayan [3]

[2]Department of Electrical and Computer Engineering, and Institute for Systems Research, University of Maryland, College Park, MD 20742. E-mail: armand@isr.umd.edu. Phone: (301) 405-6844

[3]Department of Electrical and Computer Engineering, and Institute for Systems Research, University of Maryland, College Park, MD 20742. E-mail: prakash@isr.umd.edu. Phone: (301) 405-3661

Notation and conventions

Throughout, we use $\mathbb{R}$ to denote the set of all real numbers, or equivalently, the real line $(-\infty, \infty)$. The set of all non-negative real numbers is denoted by $\mathbb{R}_+$. The set of $\{0, 1, \ldots\}$ of all non-negative integers is denoted $\mathbb{N}$. The notation $\mathbb{N}_0$ will be used for the set $\{1, 2, \ldots\}$ of all positive integers.

With p a positive integer, let $\mathbb{R}^p$ denote the p^{th} cartesian product of $\mathbb{R}$. An element $\boldsymbol{x}$ in $\mathbb{R}^p$, whose p components are denoted $x_1, \ldots, x_p$, is always interpreted as a column vector $(x_1, \ldots, x_p)'$ (with $'$ denoting transpose).

Part I

DETECTION THEORY

Chapter 1

Simple binary hypothesis testing

A decision has to be made as to which of two hypotheses (or states of nature) is the correct one. The states of nature are encoded in a rv H and a decision has to be made on the basis of an observation Y which is statistically related to H.

1.1 Motivating examples

Control process A machine produces circuit boards. It is either fully functioning ($H = 1$) or worn out ($H = 0$). Checking the state of the machine is not feasible as it would require that the production be stopped, incurring a loss of revenue for the manufacturer if the machine were indeed shown to be fully functionally. Instead, a batch of circuits is collected and tested for a number of performance parameters, say $Y_1, \ldots, Y_k$. It is known that

A simple communication example

Testing means

1.2 The probabilistic model

These examples can be cast as *binary* hypothesis testing problems: Nature is in either of two states, say $H = 0$ or $H = 1$ for sake of concreteness, and the observations are organized into an $\mathbb{R}^k$-valued rv $\boldsymbol{Y}$. We assume given two probability distribution functions $F_0, F_1 : \mathbb{R}^k \to [0, 1]$ on $\mathbb{R}^k$; they will act as conditional

probability distribution of $\boldsymbol{Y}$ given $H = 0$ and $H = 1$, respectively. This situation is summarized by

$$\begin{array}{ll} H_1: & Y \sim F_1 \\ H_0: & Y \sim F_0. \end{array} \tag{1.1}$$

In the statistical literature the hypothesis H_0 is called the *null hypothesis* and hypothesis H_1 is referred to as the *non-null hypothesis* or the *alternative*.

Probabilistically, the symbolic statement (1.1) is understood as follows: Given some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ (whose existence is discussed shortly), consider rvs $H : \Omega \to \{0, 1\}$ and $\boldsymbol{Y} : \Omega \to \mathbb{R}^k$. The probability distribution functions F_0 and F_1 being interpreted as conditional probability distribution of $\boldsymbol{Y}$ given $H = 0$ and $H = 1$, respectively, we must have

$$F_h(\boldsymbol{y}) = \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y} | H = h\right], \qquad \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^k, \\ h = 0, 1. \end{array}$$

The probability distribution of the rv H is specified by p in $[0, 1]$ with

$$p = \mathbb{P}\left[H = 1\right] = 1 - \mathbb{P}\left[H = 0\right].$$

We refer to the pmf $(1 - p, p)$ on $\{0, 1\}$, or just to p, as the *prior*.

Because

$$\begin{aligned} \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y}, H = h\right] &= \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y} | H = h\right] \mathbb{P}\left[H = h\right] \\ &= \begin{cases} (1 - p) F_0(\boldsymbol{y}) & \text{if } h = 0, \boldsymbol{y} \in \mathbb{R}^k \\ \\ p F_1(\boldsymbol{y}) & \text{if } h = 1, \boldsymbol{y} \in \mathbb{R}^k, \end{cases} \end{aligned} \tag{1.2}$$

the law of total probability shows that

$$\begin{aligned} \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y}\right] &= \sum_{h=0}^{1} \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y} | H = h\right] \mathbb{P}\left[H = h\right] \\ &= p F_1(\boldsymbol{y}) + (1 - p) F_0(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k. \end{aligned} \tag{1.3}$$

In other words, the conditional probability distributions of the observations given the hypothesis *and* the probability distribution of H completely specify the *joint* distribution of the rvs H and $\boldsymbol{Y}$.

1.3 A construction

The existence of the model described in Section 1.2 can be guaranteed through the following construction: Take $\Omega = \{0,1\} \times \mathbb{R}^k$ with generic element $\omega = (h, \boldsymbol{y})$ with $h = 0, 1$ and $\boldsymbol{y}$ an arbitrary element of $\mathbb{R}^k$. We endow Ω with the σ-field $\mathcal{F}$ given by

$$\mathcal{F} = \sigma\left(\mathcal{P}(\{0,1\}) \times \mathcal{B}(\mathbb{R}^k)\right)$$

where $\mathcal{P}(\{0,1\})$ is the power set of $\{0,1\}$, and $\mathcal{B}(\mathbb{R}^k)$ is the Borel σ-field on $\mathbb{R}^k$.

We define the mappings $H : \Omega \to \mathbb{R}$ and $\boldsymbol{Y} : \Omega \to \mathbb{R}^k$ by

$$H(\omega) = h \quad \text{and} \quad \boldsymbol{Y}(\omega) = \boldsymbol{y}, \quad \omega = (h, \boldsymbol{y}) \in \Omega.$$

Both projection mappings are Borel measurable, and therefore define rvs.

If $\mathbb{P}$ is any probability measure on the σ-field $\mathcal{F}$, then by *construction* of the rvs H and $\boldsymbol{Y}$ just given, the joint probability distribution of the pair $(H, \boldsymbol{Y})$ is necessarily given by

$$\begin{aligned} \mathbb{P}\left[H = h, \boldsymbol{Y} \leq \boldsymbol{y}\right] &= \mathbb{P}\left[\{\omega \in \Omega : \ H(\omega) = h, \boldsymbol{Y}(\omega) \leq \boldsymbol{y}\}\right] \\ &= \mathbb{P}\left[\{h\} \times (-\infty, \boldsymbol{y}]\right], \quad \begin{matrix} h = 0, 1 \\ \boldsymbol{y} \in \mathbb{R}^k \end{matrix} \end{aligned} \tag{1.4}$$

since

$$\{\omega \in \Omega : \ H(\omega) = h, \boldsymbol{Y}(\omega) \leq \boldsymbol{y}\} = \{h\} \times (-\infty, \boldsymbol{y}].$$

On the way to identify *a* probability $\mathbb{P}$ on $\mathcal{F}$ under which the joint probability distribution of the pair $(H, \boldsymbol{Y})$ satisfies (1.2), we readily conclude from (1.4) that $\mathbb{P}$ is necessarily determined on certain rectangles, namely

$$\mathbb{P}\left[\{h\} \times (-\infty, \boldsymbol{y}]\right] = \begin{cases} (1-p)F_0(\boldsymbol{y}) & \text{if } h = 0 \\ \\ pF_1(\boldsymbol{y}) & \text{if } h = 1 \end{cases} \tag{1.5}$$

for every $\boldsymbol{y}$ in $\mathbb{R}^k$. At this point we recall the following fact from Measure Theory: Any probability measure on the σ-field $\mathcal{F}$ carried by the product space $\{0,1\} \times \mathbb{R}^k$ is uniquely determined on the entire σ-field $\mathcal{F}$ by its values on the rectangle sets of the form

$$\{h\} \times (-\infty, \boldsymbol{y}], \quad \begin{matrix} h = 0, 1 \\ \boldsymbol{y} \in \mathbb{R}^k. \end{matrix}$$

Therefore, by virtue of (1.5) there exists a *unique* probability measure $\mathbb{P}$ on $\mathcal{F}$ such that (1.4) holds. More generally, it is also the case that

$$\mathbb{P}\left[\{h\} \times (-\infty, \boldsymbol{y}]\right] = \begin{cases} (1-p)\int_B dF_0(\boldsymbol{y}) & \text{if } h = 0 \\ \\ p\int_B dF_1(\boldsymbol{y}) & \text{if } h = 1 \end{cases} \tag{1.6}$$

for every Borel set B in $\mathbb{R}^k$ as a result of the fact that

$$\sigma\left((-\infty, \boldsymbol{y}],\ \boldsymbol{y} \in \mathbb{R}^k\right) = \mathcal{B}(\mathbb{R}^k).$$

Finally, under this probability measure $\mathbb{P}$ it is plain (1.5) immediately implies

$$\mathbb{P}\left[H = h\right] = \mathbb{P}\left[\{h\}\right] = \begin{cases} (1-p) & \text{if } h = 0 \\ \\ p & \text{if } h = 1 \end{cases} \tag{1.7}$$

and

$$\begin{aligned} \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y} | H = h\right] &= \frac{\mathbb{P}\left[H = h, \boldsymbol{Y} \leq \boldsymbol{y}\right]}{\mathbb{P}\left[H = h\right]} \\ &= F_h(\boldsymbol{y}) \end{aligned} \tag{1.8}$$

for every $\boldsymbol{y}$ in $\mathbb{R}^k$, as required.

1.4 Basic assumptions

During the discussion, several assumptions will be enforced on the probability distributions F_0 and F_1. The assumptions that will be most often encountered are denoted by **(A.1)** and **(A.2)** for sake of convenience. They are stated and discussed in some details below.

Condition **(A.1)**: The probability distributions F_0 and F_1 on $\mathbb{R}^k$ are both *absolutely continuous* with respect to some distribution F on $\mathbb{R}^k$ – In general F may not be a probability distribution.

Condition **(A.1)** is equivalent to saying that there exist Borel mappings $f_0, f_1 : \mathbb{R}^k \to \mathbb{R}_+$ such that

$$F_h(\boldsymbol{y}) = \int_{-\boldsymbol{\infty}}^{\boldsymbol{y}} f_h(\boldsymbol{\eta}) dF(\boldsymbol{\eta}), \qquad \begin{matrix} \boldsymbol{y} \in \mathbb{R}^k, \\ h = 0, 1. \end{matrix} \tag{1.9}$$

In some basic sense, this condition is hardly constraining since we can always take F to be the average of the two probability distributions F_0 and F_1. i.e.,

$$F(\boldsymbol{y}) \equiv \frac{1}{2}F_0(\boldsymbol{y}) + \frac{1}{2}F_1(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k. \tag{1.10}$$

in which case F is also a probability distribution. This choice for F is usually not operationally convenient and therefore discarded. However, the most often encountered situations arise when F is either Lebesgue measure on $\mathbb{R}^k$ or a counting measure on some countable subset of $\mathbb{R}^k$, in which case F is not a probability distribution.

When F is Lebesgue measure on $\mathbb{R}^k$, the Borel mappings $f_0, f_1 : \mathbb{R}^k \to \mathbb{R}_+$ are just the probability density functions induced by F_0 and F_1 in the usual sense. When F is counting measure on a countable subset $S \subseteq \mathbb{R}^k$, then the Borel mappings $f_0, f_1 : \mathbb{R}^k \to \mathbb{R}_+$ are best thought as *probability mass functions* (pdfs) $\boldsymbol{f}_0 = \{f_0(\boldsymbol{y}),\ \boldsymbol{y} \in S\}$ and $\boldsymbol{f}_1 = \{f_1(\boldsymbol{y}),\ \boldsymbol{y} \in S\}$, i.e.,

$$0 \leq f_h(\boldsymbol{y}) \leq 1, \qquad \begin{array}{c} \boldsymbol{y} \in S, \\ h = 0, 1. \end{array}$$

and

$$\sum_{\boldsymbol{y} \in S} f_h(\boldsymbol{y}) = 1, \quad h = 0, 1.$$

The condition (1.9) now takes the form

$$\mathbb{P}\left[\boldsymbol{Y} \in B | H = h\right] = \sum_{\boldsymbol{\eta} \in S \cap B} f_h(\boldsymbol{\eta}), \qquad \begin{array}{c} B \in \mathcal{B}(\mathbb{R}^k) \\ h = 0, 1. \end{array}$$

Condition **(A.2)**: The probability distribution F_1 is *absolutely continuous* with respect to the probability distribution F_0.

Under Condition **(A.1)**, with the notation introduced earlier, this is equivalent to requiring

$$f_0(\boldsymbol{y}) = 0 \quad \text{implies} \quad f_1(\boldsymbol{y}) = 0. \tag{1.11}$$

1.5 Admissible tests

Decisions as to which state of nature occurred are taken on the basis of observations; this is formalized through the following definition.

An *admissible* decision rule (or test) is any *Borel* mapping $d : \mathbb{R}^k \to \{0, 1\}$. The collection of all admissible rules is denoted by $\mathcal{D}$.

The measurability requirement entering the definition of admissibility is imposed to guarantee that the mapping $d(\boldsymbol{Y}) : \Omega \to \{0, 1\} : \omega \to d(Y(\omega))$ is indeed a rv, i. e., $[\omega \in \Omega : d(Y(\omega)) = h]$ is an event in $\mathcal{F}$ for all $h = 0, 1$. The need for this technical condition will become apparent in subsequent chapters.

The next fact will prove useful in some of the discussion

Lemma 1.5.1 *The set $\mathcal{D}$ of admissible decision rules is in one-to-one correspondence with $\mathcal{B}(\mathbb{R}^k)$.*

Proof. By definition of admissibility every test d in $\mathcal{D}$ is completely specified by the *Borel* subset $C(d)$ defined by

$$C(d) \equiv \{\boldsymbol{y} \in \mathbb{R}^k : d(\boldsymbol{y}) = 0\}. \tag{1.12}$$

Conversely, any Borel measurable subset C of $\mathbb{R}^k$ uniquely determines an admissible rule d_C in $\mathcal{D}$ through

$$d_C(\boldsymbol{y}) = \begin{cases} 1 & \text{if } \boldsymbol{y} \notin C \\ 0 & \text{if } \boldsymbol{y} \in C. \end{cases}$$

We note that $C(d_C) = C$ as expected. ■

Any admissible rule d in $\mathcal{D}$ induces *two* types of error: Upon observing $\boldsymbol{Y}$, either $H = 0$ is true and $d(\boldsymbol{Y}) = 1$ *or* $H = 1$ is true and $d(\boldsymbol{Y}) = 0$.

These two possibilities are the so–called errors of the *first* and *second* type associated with the decision rule d; they are quantified by

$$\alpha(d) \equiv \mathbb{P}\left[d(\boldsymbol{Y}) = 1 | H = 0\right] \tag{1.13}$$

and

$$\beta(d) \equiv \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 1\right], \tag{1.14}$$

respectively.

The quantity $\alpha(d)$ is sometimes called the *size* of the test d. In radar parlance, these probabilities are referred to as probabilities of *false alarm* and *miss*, respectively, with alternate notation

$$P_{\mathrm{F}}(d) \equiv \mathbb{P}\left[d(\boldsymbol{Y}) = 1 | H = 0\right] \tag{1.15}$$

and

$$P_{\mathrm{M}}(d) \equiv \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 1\right]. \tag{1.16}$$

Throughout we shall use this terminology. Sometimes, it is convenient to consider the so–called probability of *detection* given by

$$P_{\mathrm{D}}(d) \equiv \mathbb{P}\left[d(\boldsymbol{Y}) = 1 | H = 1\right] = 1 - P_{\mathrm{M}}(d). \tag{1.17}$$

1.6 Likelihood ratio tests

In subsequent chapters we shall consider several formulations for the binary hypothesis problem. In all cases the tests of interest are related to tests in the class of admissible tests $\{d_\eta, \eta \geq 0\}$ which we now introduce.

For each $\eta \geq 0$, the mapping $d_\eta : \mathbb{R}^k \to \{0, 1\}$ is defined by

$$d_\eta(\boldsymbol{y}) = 0 \quad \text{iff} \quad f_1(\boldsymbol{y}) < \eta f_0(\boldsymbol{y}). \tag{1.18}$$

It is plain from the definition (1.18) (with $\eta = 0$) that d_0 is simply the test that always selects the non-null hypothesis $H = 1$, i.e., $d_0(\boldsymbol{y}) = 1$ for every $\boldsymbol{y}$ in $\mathbb{R}^k$. On the other hand, formally substituting $\eta = \infty$ in (1.18) will be problematic at observation points where $f_0(\boldsymbol{y}) = 0$. However, by *convention* we shall interpret d_∞ as the test that always selects the null hypothesis $H = 0$, i.e., $d_0(\boldsymbol{y}) = 0$ for every $\boldsymbol{y}$ in $\mathbb{R}^k$.

Such tests take an even simpler form under the additional Condition **(A.2)** as will be seen shortly: Note that (1.18) can be rewritten as

$$d_\eta(\boldsymbol{y}) = 0 \quad \text{if} \quad \frac{f_1(\boldsymbol{y})}{f_0(\boldsymbol{y})} < \eta \quad \text{whenever } f_0(\boldsymbol{y}) > 0.$$

Taking our cue from this last statement, we define the *likelihood ratio* as any Borel mapping $L : \mathbb{R}^k \to \mathbb{R}$ of the form

$$L(\boldsymbol{y}) \equiv \begin{cases} \frac{f_1(\boldsymbol{y})}{f_0(\boldsymbol{y})} & \text{if } f_0(\boldsymbol{y}) > 0 \\ \\ \Lambda(\boldsymbol{y}) & \text{if } f_0(\boldsymbol{y}) = 0 \end{cases} \tag{1.19}$$

for some arbitrary Borel mapping $\Lambda : \mathbb{R}^k \to \mathbb{R}_+$. Different choices of this arbitrary non-negative function produce different versions of the likelihood ratio function.

Given a version of the likelihood ratio function in (1.19), we define the *likelihood ratio* test with *threshold* $\eta \geq 0$ to be the admissible decision rule $Lrt_\eta : \mathbb{R}^k \to \{0, 1\}$ given by

$$Lrt_\eta(\boldsymbol{y}) \equiv \begin{cases} 1 & \text{if } L(\boldsymbol{y}) \geq \eta \\ \\ 0 & \text{if } L(\boldsymbol{y}) < \eta. \end{cases} \tag{1.20}$$

With

$$B_h = \left\{ \boldsymbol{y} \in \mathbb{R}^k : \ f_h(\boldsymbol{y}) = 0 \right\}, \quad h = 0, 1, \tag{1.21}$$

we note that

$$\mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = h \right] = \int_{B_0} f_h(\boldsymbol{y}) dF(\boldsymbol{y}), \quad h = 0, 1. \tag{1.22}$$

Under **(A.2)**, the inclusion $B_0 \subseteq B_1$ holds and we conclude that

$$\mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = h \right] = 0, \quad h = 0, 1.$$

For any value η of the threshold it is plain that the tests d_η and Lrt_η coincide on the set $\{\boldsymbol{y} \in \mathbb{R}^k : f_0(\boldsymbol{y}) > 0\}$ (while possibly disagreeing on the complement B_0). Thus, for each $h = 0, 1$, we find that

$$\begin{aligned} & \mathbb{P}\left[d_\eta(\boldsymbol{Y}) = 0 | H = h \right] \\ = \ & \mathbb{P}\left[d_\eta(\boldsymbol{Y}) = 0, f_0(\boldsymbol{Y}) > 0 | H = h \right] + \mathbb{P}\left[d_\eta(\boldsymbol{Y}) = 0, f_0(\boldsymbol{Y}) = 0 | H = h \right] \\ = \ & \mathbb{P}\left[Lrt_\eta(\boldsymbol{Y}) = 0, f_0(\boldsymbol{Y}) > 0 | H = h \right] \\ = \ & \mathbb{P}\left[Lrt_\eta(\boldsymbol{Y}) = 0, f_0(\boldsymbol{Y}) > 0 | H = h \right] + \mathbb{P}\left[Lrt_\eta(\boldsymbol{Y}) = 0, f_0(\boldsymbol{Y}) = 0 | H = h \right] \\ = \ & \mathbb{P}\left[Lrt_\eta(\boldsymbol{Y}) = 0 | H = h \right]. \end{aligned}$$

This discussion leads to the following fact.

Lemma 1.6.1 *Assume the absolute continuity conditions* **(A.1)–(A.2)** *to hold. For each* $\eta \geq 0$*, the tests* d_η *and* Lrt_η *are* equivalent *in the sense that* $P_M(d_\eta) = P_M(Lrt_\eta)$ *and* $P_F(d_\eta) = P_F(Lrt_\eta)$*.*

1.7 Exercises

1.8 References

Chapter 2

The Bayesian formulation

The Bayesian formulation assumes *knowledge* of the conditional distributions F_1 and F_0, and of the prior distribution p of the rv H. Two other formulations, namely the Minimax formulation and the Neyman-Pearson formulation, will be studied in Chapters 4 and 5, respectively.

2.1 The Bayesian optimization problem

The cost incurred for making decisions is quantified by the mapping $C : \{0,1\} \times \{0,1\} \to \mathbb{R}$ with the interpretation that

$$C(h,d) = \begin{array}{c}\text{Cost incurred for deciding } d \\ \text{when } H = h\end{array}, \quad d, h = 0, 1.$$

As the sample ω in Ω is realized, the observation $Y(\omega)$ is recorded and the use of the admissible rule d in $\mathcal{D}$ incurs a cost $C(H(\omega), d(\boldsymbol{Y}(\omega)))$. Although it is tempting to seek to minimize this quantity, this is not possible. Indeed, the rv $\boldsymbol{Y}$ is observed, whence $d(\boldsymbol{Y})$ is known once the test d has been specified, but the state of nature H is *not* directly observable. Consequently, the value of the cost $C(H, d(\boldsymbol{Y}))$ is not available. To remedy to this difficulty, we introduce the *expected cost function* $J : \mathcal{D} \to \mathbb{R}$ given by

$$J(d) \equiv \mathbb{E}\left[C(H, d(\boldsymbol{Y}))\right], \quad d \in \mathcal{D}.$$

The *Bayesian* Problem $\mathcal{P}_{\mathrm{B}}$ is the minimization problem

$$\mathcal{P}_{\mathrm{B}}: \quad \text{Minimize } J(d) \text{ over } d \text{ in } \mathcal{D}.$$

This amounts to finding an admissible test $d^\star : \mathbb{R}^k \to \{0, 1, \ldots, M-1\}$ in $\mathcal{D}$ such that

$$J(d^\star) \leq J(d), \quad d \in \mathcal{D}. \tag{2.1}$$

Any admissible test $d^\star$ which satisfies (2.1) is called a Bayesian test, and the value

$$J(d^\star) = \inf_{d \in \mathcal{D}} J(d) = \min_{d \in \mathcal{D}} J(d) \tag{2.2}$$

is known as the *Bayesian* cost.

The solution to the Bayesian problem $\mathcal{P}_{\mathrm{B}}$ is developed with the help of an auxiliary result concerning the form of the Bayesian cost. This representation result will be useful in several places and is given here for sake of easy reference: Introduce the *relative* costs Γ_0 and Γ_1 given by

$$\Gamma_h \equiv C(h, 1-h) - C(h, h), \quad h = 0, 1 \tag{2.3}$$

and define the auxiliary expected cost function $\widehat{J} : \mathcal{D} \to \mathbb{R}$ to be

$$\widehat{J}(d) = \mathbb{E}\left[\mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right] \Gamma_H\right], \quad d \in \mathcal{D}. \tag{2.4}$$

Lemma 2.1.1 *For any admissible rule d in $\mathcal{D}$, the relation*

$$J(d) = \mathbb{E}\left[C(H, H)\right] + \widehat{J}(d) \tag{2.5}$$

holds with

$$\widehat{J}(d) = \Gamma_0(1-p) \cdot P_F(d) + \Gamma_1 p \cdot P_M(d). \tag{2.6}$$

Proof. Fix d in $\mathcal{D}$. Recall that the rvs H and $d(\boldsymbol{Y})$ are $\{0, 1\}$-valued rvs, and that the events $[d(\boldsymbol{Y}) = H]$ and $[d(\boldsymbol{Y}) \neq H]$ form a partition of Ω, i.e.,

$$\mathbf{1}\left[d(\boldsymbol{Y}) = H\right] + \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right] = \mathbf{1}\left[\Omega\right] = 1.$$

It readily follows that

$$\begin{aligned} C(H, d(\boldsymbol{Y})) &= \mathbf{1}\left[d(\boldsymbol{Y}) = H\right] C(H, H) + \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right] C(H, 1-H) \\ &= \left(1 - \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right]\right) C(H, H) + \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right] C(H, 1-H) \\ &= C(H, H) + \left(C(H, 1-H) - C(H, H)\right) \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right] \\ &= C(H, H) + \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right] \Gamma_H \end{aligned} \tag{2.7}$$

with the relative costs Γ_0 and Γ_1 given by (2.3). Taking expectations on both sides of (2.7) we obtain (2.5).

The law of total probabilities gives

$$\begin{aligned} \widehat{J}(d) &= \mathbb{E}\left[\Gamma_0 \mathbf{1}\left[d(\boldsymbol{Y}) \neq 0\right] \mathbf{1}\left[H = 0\right] + \Gamma_1 \mathbf{1}\left[d(\boldsymbol{Y}) \neq 1\right] \mathbf{1}\left[H = 1\right]\right] \\ &= \Gamma_0 (1-p) \cdot \mathbb{P}\left[d(\boldsymbol{Y}) \neq 0 | H = 0\right] + \Gamma_1 p \cdot \mathbb{P}\left[d(\boldsymbol{Y}) \neq 1 | H = 1\right] \\ &= \Gamma_0 (1-p) \cdot \mathbb{P}\left[d(\boldsymbol{Y}) = 1 | H = 0\right] + \Gamma_1 p \cdot \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 1\right], \end{aligned}$$

and the desired expression (2.6) is obtained. ■

The Bayesian cost under a given decision rule is completely determined by its probabilities of false alarm and of miss. We also note that

$$\begin{aligned} \widehat{J}(d) &= \Gamma_0 (1-p) + \Gamma_1 p \cdot \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 1\right] \\ &\quad - \Gamma_0 (1-p) \cdot \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 0\right], \quad d \in \mathcal{D} \end{aligned} \tag{2.8}$$

as an immediate consequence of (2.6).

Therefore, by Lemma 1.6.1 it follows from (2.5)-(2.6) that $J(d_\eta) = J(Lrt_\eta)$ *regardless* of the cost function $C : \{0,1\} \times \{0,1\} \to \mathbb{R}$. The same argument also shows that any two versions of the likelihood ratio function will generate likelihood ratio tests which are equivalent.

2.2 Solving the Bayesian problem $\mathcal{P}_B$

It follows from (2.6) that solving $\mathcal{P}_B$ is *equivalent* to solving the auxiliary problem $\widehat{\mathcal{P}}_B$ where

$$\widehat{\mathcal{P}}_B : \quad \text{Minimize } \widehat{J}(d) \text{ over } d \text{ in } \mathcal{D}.$$

To do so, it will be necessary to assume that the probability distributions F_0 and F_1 satisfy the absolute continuity condition **(A1)** given earlier, namely that

there exists a single distribution F on $\mathbb{R}^k$ with respect to which both F_0 and F_1 are absolutely continuous. For any test d in $\mathcal{D}$, we get

$$\begin{aligned}\mathbb{P}\left[d(\boldsymbol{Y})=0|H=h\right] &= \int_{C(d)} dF_h(\boldsymbol{y}) \\ &= \int_{C(d)} f_h(\boldsymbol{y})dF(\boldsymbol{y}), \quad h=0,1 \end{aligned} \tag{2.9}$$

with $C(d)$ defined at (1.12). It is now easy to see from (2.8) that

$$\widehat{J}(d) = \Gamma_0(1-p) + \int_{C(d)} h(\boldsymbol{y})dF(\boldsymbol{y}) \tag{2.10}$$

where the mapping $h : \mathbb{R}^k \to \mathbb{R}$ is given by

$$h(\boldsymbol{y}) \equiv \Gamma_1 p \cdot f_1(\boldsymbol{y}) - \Gamma_0(1-p)\cdot f_0(\boldsymbol{y}), \quad \boldsymbol{y}\in\mathbb{R}^k. \tag{2.11}$$

Theorem 2.2.1 *Assume the absolute continuity condition* **(A.1)** *to hold. Define the Borel set $C^\star$ by*

$$C^\star \equiv \{y \in \mathbb{R}^k : h(\boldsymbol{y}) < 0\} \tag{2.12}$$

with $h : \mathbb{R}^k \to \mathbb{R}$ given by (2.11). The decision rule $d^\star : \mathbb{R}^k \to \{0,1\}$ induced by $C^\star$ is given by

$$d^\star(\boldsymbol{y}) = \begin{cases} 1 & \text{if } x \notin C^\star \\ \\ 0 & \text{if } x \in C^\star; \end{cases} \tag{2.13}$$

it is admissible and solves the Problem $\widehat{\mathcal{P}}_B$, hence solves the Bayesian Problem $\mathcal{P}_B$.

Proof. The set $C^\star$ is a Borel subset of $\mathbb{R}^k$ due to the fact that the functions $f_0, f_1 : \mathbb{R}^k \to \mathbb{R}_+$ are themselves Borel measurable. The test $d^\star$ is therefore an admissible decision rule in $\mathcal{D}$ since $C(d^\star) = C^\star$. We now show that $d^\star$ satisfies

$$\widehat{J}(d^\star) \le \widehat{J}(d), \quad d\in\mathcal{D}. \tag{2.14}$$

Indeed, for every test d in $\mathcal{D}$, we see from (2.10) that

$$\widehat{J}(d) = \Gamma_0(1-p) + \int_{C(d)\setminus C^\star} h(\boldsymbol{y})dF(\boldsymbol{y}) + \int_{C(d)\cap C^\star} h(\boldsymbol{y})dF(\boldsymbol{y})$$

and

$$\widehat{J}(d^\star) = \Gamma_0(1-p) + \int_{C^\star \setminus C(d)} h(\boldsymbol{y}) dF(\boldsymbol{y}) + \int_{C(d)\cap C^\star} h(\boldsymbol{y}) dF(\boldsymbol{y}).$$

Therefore,

$$\widehat{J}(d) - \widehat{J}(d^\star) = \int_{C(d)\setminus C^\star} h(\boldsymbol{y}) dF(\boldsymbol{y}) + \int_{C^\star \setminus C(d)} (-h(\boldsymbol{y}))\, dF(\boldsymbol{y}) \geq 0$$

since

$$\int_{C(d)\setminus C^\star} h(\boldsymbol{y}) dF(\boldsymbol{y}) \geq 0 \quad \text{and} \qquad \int_{C^\star \setminus C(d)} h(\boldsymbol{y}) dF(\boldsymbol{y}) \leq 0$$

by the very definition of $C^\star$. The problem $\widehat{\mathcal{P}}_B$ is therefore solved by the test $d^\star$ defined at (2.13). ■

Uniqueness The solution to the Bayesian problem is *not* unique: It should be plain that $C^\star$ could be replaced by

$$C^{\star\star} \equiv \{y \in \mathbb{R}^k : h(\boldsymbol{y}) \leq 0\}$$

(with corresponding test $d^{\star\star}$) without affecting the conclusion of optimality since

$$\int_{\{\boldsymbol{y}\in\mathbb{R}^k:\ h(\boldsymbol{y})=0\}} h(\boldsymbol{y}) dF(\boldsymbol{y}) = 0.$$

While it is true that $J(d^\star) = J(d^{\star\star})$, it is not necessarily the case that the equalities $P_F(d^\star) = P_F(d^{\star\star})$ or $P_M(d^\star) = P_M(d^{\star\star})$ hold.

Implementation using likelihood ratio test Assume that $0 < p < 1$ to avoid trivial situations, and that the relative costs satisfy the conditions

$$\Gamma_h > 0, \quad h = 0, 1, \tag{2.15}$$

i.e., the cost of making an incorrect decision is greater than the cost of making a correct decision. This is of course a most reasonable assumption which always

holds in applications. Under this condition, the Bayesian decision rule $d^\star$ given in Theorem 2.2.1 takes the equivalent form

$$d^\star(\boldsymbol{y}) = 0 \quad \text{iff} \quad f_1(\boldsymbol{y}) < \frac{\Gamma_0(1-p)}{\Gamma_1 p} f_0(\boldsymbol{y}). \tag{2.16}$$

In view of the definition (1.18), the Bayesian test $d^\star$ is indeed a test d_η with η given by

$$\eta \equiv \frac{\Gamma_0(1-p)}{\Gamma_1 p}.$$

Equipped with Lemma 1.6.1 we can now restate Theorem 2.2.1.

Theorem 2.2.2 *Assume the absolute continuity conditions* **(A.1)–(A.2)** *to hold. Whenever* $\Gamma_h > 0$ *for* $h = 0, 1$*, the Bayesian decision rule* $d^\star$ *identified in Theorem 2.1 is equivalent to the likelihood ratio test* $Lrt_{\eta^\star}$ *where*

$$\eta^\star \equiv \frac{\Gamma_0(1-p)}{\Gamma_1 p} = \frac{C(0,1) - C(0,0)}{C(1,0) - C(1,1)} \cdot \frac{1-p}{p}.$$

2.3 The probability of error criterion

A special case of great interest is obtained when the cost function C takes the form

$$C(h,d) = \mathbf{1}\left[h \neq d\right], \quad h, d = 0, 1.$$

The corresponding expected cost then reduces to the probability of making an incorrect decision, namely the *probability of error*, and is given by

$$P_E(d) \equiv \mathbb{P}\left[d(\boldsymbol{Y}) \neq H\right], \quad d \in \mathcal{D}.$$

We check that

$$\Gamma_h = C(h, 1-h) - C(h,h) = 1, \quad h = 0, 1,$$

and the relations (2.5)-(2.6) yield

$$\begin{aligned} P_E(d) &= (1-p)\cdot P_F(d) + p\cdot P_M(d) \\ &= p + (1-p)\cdot P_F(d) - p\cdot P_D(d), \quad d\in\mathcal{D}. \end{aligned} \tag{2.17}$$

For the probability of error criterion, the threshold $\eta^\star$ appearing in Theorem 2.2.2 has the simpler form

$$\eta^\star = \frac{1-p}{p}.$$

The optimal decision rule $d^\star$, as described at (2.16), can now be rewritten as

$$d^\star(\boldsymbol{y}) = 0 \quad \text{iff} \quad f_1(\boldsymbol{y}) < \frac{1-p}{p} f_0(\boldsymbol{y}). \tag{2.18}$$

The ML test In the uniform prior case, i.e., $p = \frac{1}{2}$, the Bayesian test (2.18) becomes

$$d^\star(\boldsymbol{y}) = 0 \quad \text{iff} \quad f_1(\boldsymbol{y}) < f_0(\boldsymbol{y}). \tag{2.19}$$

In other words, the optimal decision is to select that hypothesis whose likelihood is largest given the observation $\boldsymbol{y}$. We refer to this strategy as the *Maximum Likelihood* (ML) test.

The MAP computer Finally, (2.18) can also be rewritten as

$$d^\star(\boldsymbol{y}) = 0 \quad \text{iff} \quad \mathbb{P}\left[H=1|\boldsymbol{Y}=\boldsymbol{y}\right] < \mathbb{P}\left[H=0|\boldsymbol{Y}=\boldsymbol{y}\right] \tag{2.20}$$

since for each $\boldsymbol{y}$ in $\mathbb{R}^k$, we have

$$\mathbb{P}\left[H=1|\boldsymbol{Y}=\boldsymbol{y}\right] = \frac{p f_1(\boldsymbol{y})}{p f_1(\boldsymbol{y}) + (1-p) f_0(\boldsymbol{y})}$$

and

$$\mathbb{P}\left[H=0|\boldsymbol{Y}=\boldsymbol{y}\right] = \frac{(1-p) f_0(\boldsymbol{y})}{p f_1(\boldsymbol{y}) + (1-p) f_0(\boldsymbol{y})}$$

by Bayes' Theorem. For each $h = 0, 1$, the conditional probability $\mathbb{P}\left[H=h|\boldsymbol{Y}=\boldsymbol{y}\right]$ is known as the *posterior* probability that $H = h$ occurs given the observation $\boldsymbol{y}$. Put differently, the optimal test (2.20) compares these posterior probabilities given the observation $\boldsymbol{y}$, and selects the hypothesis with the largest posterior probability, hence the terminology *Maximum A Posteriori* (MAP) computer.

2.4 The Gaussian case

Assume that the observation rv $\boldsymbol{Y}$ is conditionally Gaussian given H, i.e.,

$$\begin{array}{ll} H_1: & \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{m}_1, \boldsymbol{R}_1) \\ H_0: & \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{m}_0, \boldsymbol{R}_0) \end{array}$$

where $\boldsymbol{m}_1$ and $\boldsymbol{m}_0$ are elements in $\mathbb{R}^k$, and the $k \times k$ *symmetric* matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_0$ are *positive definite* (thus *invertible*). Throughout the pairs $(\boldsymbol{m}_0, \boldsymbol{R}_0)$ and $(\boldsymbol{m}_1, \boldsymbol{R}_1)$ are distinct so that the probability density functions $f_0, f_1 : \mathbb{R}^k \to \mathbb{R}_+$ are distinct since

$$f_h(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^k \det \boldsymbol{R}_h}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{m}_h)' \boldsymbol{R}_h^{-1} (\boldsymbol{y}-\boldsymbol{m}_h)}, \qquad \begin{array}{c} \boldsymbol{y} \in \mathbb{R}^k \\ h = 0, 1. \end{array}$$

Both conditions **(A.1)** and **(A.2)** obviously hold, and for each $\eta > 0$, the test d_η and Lrt_η coincide.

The likelihood ratio and the likelihood ratio tests For this example, the likelihood ratio function is given by

$$L(\boldsymbol{y}) = \sqrt{\frac{\det(\boldsymbol{R}_0)}{\det(\boldsymbol{R}_1)}} \cdot e^{\frac{1}{2}Q(\boldsymbol{y})}, \quad \boldsymbol{y} \in \mathbb{R}^k$$

where we have used the notation

$$Q(\boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{m}_0)' \boldsymbol{R}_0^{-1} (\boldsymbol{y} - \boldsymbol{m}_0) - (\boldsymbol{y} - \boldsymbol{m}_1)' \boldsymbol{R}_1^{-1} (\boldsymbol{y} - \boldsymbol{m}_1).$$

Fix $\eta > 0$. By direct substitution, we conclude that

$$Lrt_\eta(\boldsymbol{y}) = 0 \quad \text{iff} \qquad e^{\frac{1}{2}Q(\boldsymbol{y})} < \sqrt{\eta^2 \cdot \frac{\det \boldsymbol{R}_1}{\det \boldsymbol{R}_0}},$$

and a simple logarithmic transformation yields

$$Lrt_\eta(\boldsymbol{y}) = 0 \quad \text{iff} \quad Q(\boldsymbol{y}) < \log\left(\eta^2 \frac{\det \boldsymbol{R}_1}{\det \boldsymbol{R}_0}\right).$$

The equal covariance case If the covariances are identical under both hypotheses, i.e.,

$$\boldsymbol{R}_0 = \boldsymbol{R}_1 \equiv \boldsymbol{R},$$

with $\boldsymbol{m}_1 \neq \boldsymbol{m}_0$, then

$$\begin{aligned} Q(\boldsymbol{y}) &= (\boldsymbol{y}-\boldsymbol{m}_0)'\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{m}_0) - (\boldsymbol{y}-\boldsymbol{m}_1)'\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{m}_1) \\ &= 2\boldsymbol{y}'\boldsymbol{R}^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_0) - \left(\boldsymbol{m}_1'\boldsymbol{R}^{-1}\boldsymbol{m}_1 - \boldsymbol{m}_0'\boldsymbol{R}^{-1}\boldsymbol{m}_0\right). \end{aligned} \quad (2.21)$$

The form of Lrt_η simplifies even further to read

$$Lrt_\eta(\boldsymbol{y}) = 0 \quad \text{iff} \quad \boldsymbol{y}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m} < \tau(\eta)$$

where we have set

$$\Delta\boldsymbol{m} \equiv \boldsymbol{m}_1 - \boldsymbol{m}_0 \quad (2.22)$$

and

$$\tau(\eta) \equiv \frac{1}{2}\left(\boldsymbol{m}_1'\boldsymbol{R}^{-1}\boldsymbol{m}_1 - \boldsymbol{m}_0'\boldsymbol{R}^{-1}\boldsymbol{m}_0\right) + \log\eta. \quad (2.23)$$

Evaluating probabilities We will now evaluate the probabilities of false alarm and miss under Lrt_η. It is plain that

$$\begin{aligned} P_F(Lrt_\eta) &= \mathbb{P}\left[Lrt_\eta(\boldsymbol{Y}) = 1 | H = 0\right] \\ &= \mathbb{P}\left[L(\boldsymbol{Y}) \geq \eta | H = 0\right] \\ &= \mathbb{P}\left[\boldsymbol{Y}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m} \geq \tau(\eta) \mid H = 0\right] \end{aligned} \quad (2.24)$$

and

$$\begin{aligned} P_M(Lrt_\eta) &= \mathbb{P}\left[Lrt_\eta(\boldsymbol{Y}) = 0 | H = 1\right] \\ &= \mathbb{P}\left[L(\boldsymbol{Y}) < \eta | H = 1\right] \\ &= \mathbb{P}\left[\boldsymbol{Y}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m} < \tau(\eta) \mid H = 1\right] \\ &= 1 - \mathbb{P}\left[\boldsymbol{Y}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m} \geq \tau(\eta) \mid H = 1\right]. \end{aligned} \quad (2.25)$$

To carry out the calculations further, recall that for each $h = 0, 1$, given $H = h$, the rv $\boldsymbol{Y}$ is conditionally Gaussian with mean vector $\boldsymbol{m}_h$ and covariance matrix $\boldsymbol{R}$. Therefore, the scalar rv $\boldsymbol{Y}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m}$ is also conditionally Gaussian with mean and variance given by

$$\mathbb{E}\left[\boldsymbol{Y}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m} | H = h\right] = \boldsymbol{m}_h'\boldsymbol{R}^{-1}\Delta\boldsymbol{m}$$

and

$$\begin{aligned}\mathrm{Var}\left[\boldsymbol{Y}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m}|H=h\right] &= \left(\boldsymbol{R}^{-1}\Delta\boldsymbol{m}\right)'\mathrm{Cov}\left[\boldsymbol{Y}|H=h\right]\left(\boldsymbol{R}^{-1}\Delta\boldsymbol{m}\right)\\ &= \left(\boldsymbol{R}^{-1}\Delta\boldsymbol{m}\right)'\boldsymbol{R}\left(\boldsymbol{R}^{-1}\Delta\boldsymbol{m}\right)\\ &= \Delta\boldsymbol{m}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m}, \end{aligned} \tag{2.26}$$

respectively. In obtaining this last relation we have used the fact that

$$\boldsymbol{Y}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m} = (\boldsymbol{R}^{-1}\Delta\boldsymbol{m})'\boldsymbol{Y}.$$

Consequently, for all $h = 0, 1$,

$$\begin{aligned}&\mathbb{P}\left[\boldsymbol{Y}'R^{-1}\Delta\boldsymbol{m} \geq \tau(\eta)|H=h\right]\\ &= \mathbb{P}\left[\boldsymbol{m}_h'\boldsymbol{R}^{-1}\Delta\boldsymbol{m} + \sqrt{\Delta\boldsymbol{m}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m}}\cdot Z \geq \tau(\eta)\right]\\ &= \mathbb{P}\left[Z \geq \frac{\tau(\eta) - \boldsymbol{m}_h'\boldsymbol{R}^{-1}\Delta\boldsymbol{m}}{\sqrt{\Delta\boldsymbol{m}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m}}}\right] \end{aligned} \tag{2.27}$$

where $Z \sim \mathrm{N}(0,1)$.

For the sake of convenience, pose

$$d^2 \equiv \Delta\boldsymbol{m}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m}, \tag{2.28}$$

and note that

$$\tau(\eta) - \boldsymbol{m}_h'\boldsymbol{R}^{-1}\Delta\boldsymbol{m} = \begin{cases} \log\eta - \frac{1}{2}d^2 & \text{if } h=1\\ \\ \log\eta + \frac{1}{2}d^2 & \text{if } h=0. \end{cases}$$

It is now clear that

$$P_F(Lrt_\eta) = 1 - \Phi\left(\frac{\log\eta + \frac{1}{2}d^2}{d}\right)$$

and

$$P_M(Lrt_\eta) = \Phi\left(\frac{\log\eta - \frac{1}{2}d^2}{d}\right).$$

We finally obtain

$$P_D(Lrt_\eta) = 1 - \Phi\left(\frac{\log\eta - \frac{1}{2}d^2}{d}\right).$$

The ML test The ML test corresponds to $\eta = 1$, in which case these expressions become

$$P_F(d_{\mathrm{ML}}) = 1 - \Phi\left(\frac{d}{2}\right) = Q\left(\frac{d}{2}\right)$$

and

$$P_M(d_{\mathrm{ML}}) = \Phi\left(-\frac{d}{2}\right) = Q\left(\frac{d}{2}\right),$$

whence

$$P_E(d_{\mathrm{ML}}) = (1-p)P_F(d_{\mathrm{ML}}) + pP_M(d_{\mathrm{ML}}) = Q\left(\frac{d}{2}\right)$$

regardless of the prior p.

2.5 The Bernoulli case

Consider the binary hypothesis testing problem

$$\begin{array}{ll} H_1: & Y \sim \mathrm{Ber}(a_1) \\ H_0: & Y \sim \mathrm{Ber}(a_0) \end{array}$$

with $a_1 < a_0$ in $(0,1)$. The case $a_0 < a_1$ is left as an exercise. Thus,

$$\mathbb{P}[Y = 1|H = h] = a_h = 1 - \mathbb{P}[Y = 0|H = h], \quad h = 0, 1$$

and Conditions **(A.1)** and **(A.2)** obviously hold with respect to counting measure F on $\{0,1\}$. The likelihood rate function is given by

$$L(y) = \left(\frac{1-a_1}{1-a_0}\right)^{1-y}\left(\frac{a_1}{a_0}\right)^{y}, \quad y \in \mathbb{R}.$$

For each $\eta > 0$, the test d_η takes the following form

$$\begin{aligned} d_\eta(y) = 0 \qquad & \text{iff} \qquad \left(\frac{1-a_1}{1-a_0}\right)^{1-y} \cdot \left(\frac{a_1}{a_0}\right)^{y} < \eta \\ & \text{iff} \qquad \left(\frac{1-a_0}{1-a_1} \cdot \frac{a_1}{a_0}\right)^{y} < \eta \cdot \frac{1-a_0}{1-a_1}, \quad y \in \mathbb{R}. \end{aligned} \tag{2.29}$$

Therefore,

$$\begin{aligned}
P_F(d_\eta) &= \mathbb{P}\left[d_\eta(Y)=1|H=0\right] \\
&= \mathbb{P}\left[\left(\frac{1-a_0}{1-a_1}\cdot\frac{a_1}{a_0}\right)^Y \geq \eta\cdot\frac{1-a_0}{1-a_1}\Big|H=0\right] \\
&= \mathbb{P}\left[Y=1,\ \left(\frac{1-a_0}{1-a_1}\cdot\frac{a_1}{a_0}\right)^Y \geq \eta\cdot\frac{1-a_0}{1-a_1}\Big|H=0\right] \\
&\quad + \mathbb{P}\left[Y=0,\ \left(\frac{1-a_0}{1-a_1}\cdot\frac{a_1}{a_0}\right)^Y \geq \eta\cdot\frac{1-a_0}{1-a_1}\Big|H=0\right] \\
&= a_0\mathbf{1}\left[\eta\cdot\frac{1-a_0}{1-a_1}\leq\frac{1-a_0}{1-a_1}\cdot\frac{a_1}{a_0}\right]+(1-a_0)\mathbf{1}\left[\eta\cdot\frac{1-a_0}{1-a_1}\leq 1\right] \\
&= a_0\mathbf{1}\left[\eta\leq\frac{a_1}{a_0}\right]+(1-a_0)\mathbf{1}\left[\eta\frac{1-a_0}{1-a_1}\leq 1\right] \\
&= a_0\mathbf{1}\left[\eta\leq\frac{a_1}{a_0}\right]+(1-a_0)\mathbf{1}\left[\eta\leq\frac{1-a_1}{1-a_0}\right]. \qquad (2.30)
\end{aligned}$$

Similarly, we get

$$\begin{aligned}
P_M(d_\eta) &= \mathbb{P}\left[d_\eta(Y)=0|H=1\right] \\
&= \mathbb{P}\left[\left(\frac{1-a_0}{1-a_1}\cdot\frac{a_1}{a_0}\right)^Y < \eta\cdot\frac{1-a_0}{1-a_1}\Big|H=1\right] \\
&= \mathbb{P}\left[Y=1,\ \left(\frac{1-a_0}{1-a_1}\cdot\frac{a_1}{a_0}\right)^Y < \eta\cdot\frac{1-a_0}{1-a_1}\Big|H=1\right] \\
&\quad + \mathbb{P}\left[Y=0,\ \left(\frac{1-a_0}{1-a_1}\cdot\frac{a_1}{a_0}\right)^Y < \eta\cdot\frac{1-a_0}{1-a_1}\Big|H=1\right] \\
&= a_1\mathbf{1}\left[\frac{1-a_0}{1-a_1}\cdot\frac{a_1}{a_0}<\eta\cdot\frac{1-a_0}{1-a_1}\right]+(1-a_1)\mathbf{1}\left[1<\eta\cdot\frac{1-a_0}{1-a_1}\right] \\
&= a_1\mathbf{1}\left[\frac{a_1}{a_0}<\eta\right]+(1-a_1)\mathbf{1}\left[1<\eta\cdot\frac{1-a_0}{1-a_1}\right] \\
&= a_1\mathbf{1}\left[\frac{a_1}{a_0}<\eta\right]+(1-a_1)\mathbf{1}\left[\frac{1-a_1}{1-a_0}<\eta\right]. \qquad (2.31)
\end{aligned}$$

2.6 Additional examples

We now present several examples where Conditions **(A.1)** or **(A.2)** fail. In all cases we assume $\Gamma_0 > 0$ and $\Gamma_1 > 0$.

An example where absolute continuity (A.2) fails Here, the observation is the scalar rv Y with F_0 and F_0 admitting probability density functions $f_0, f_1 : \mathbb{R} \to \mathbb{R}_+$ with respect to Lebesgue measure given by

$$f_0(y) = \begin{cases} 1 - \mid y \mid & \text{if } |y| \leq 1 \\ \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_1(y) = \begin{cases} \frac{1}{3} & \text{if } -1 \leq y \leq 2 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Condition **(A.1)** holds (with Lebesgue measure) but the absolute continuity condition **(A.2)** is clearly not satisfied. However, simple substitution reveals that

$$\begin{aligned} h(y) &= \Gamma_1 p \cdot f_1(y) - \Gamma_0(1-p) \cdot f_0(y) \\ &= \begin{cases} 0 & \text{if } y < -1 \\ \\ \frac{1}{3}\Gamma_1 p - \Gamma_0(1-p)(1-|y|) & \text{if } |y| \leq 1 \\ \\ \frac{1}{3}\Gamma_1 p & \text{if } 1 < y \leq 2 \\ \\ 0 & \text{if } 2 < y. \end{cases} \end{aligned} \tag{2.32}$$

The Bayesian test $d^\star$ is simply

$$d^\star(y) = 0 \quad \text{iff} \quad |y| < 1 - \frac{\frac{1}{3}\Gamma_1 p}{\Gamma_0(1-p)}.$$

Another example where absolute continuity (A.2) fails The observation is the scalar rv Y with F_0 and F_0 admitting probability density functions $f_0, f_1 : \mathbb{R} \to \mathbb{R}_+$ with respect to Lebesgue measure given by

$$f_0(y) = \begin{cases} 1 - \mid y \mid & \text{if } |y| \leq 1 \\ \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_1(y) = \begin{cases} \frac{1}{3} & \text{if } 0 \leq y \leq 3 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Condition **(A.1)** holds (with Lebesgue measure) but **(A.2)** fails. Simple substitution reveals that

$$\begin{aligned} h(y) &= \Gamma_1 p \cdot f_1(y) - \Gamma_0(1-p) \cdot f_0(y) \\ &= \begin{cases} 0 & \text{if } y < -1 \\ -\Gamma_0(1-p)(1+y) & \text{if } -1 \leq y \leq 0 \\ \frac{1}{3}\Gamma_1 p - \Gamma_0(1-p)(1-y) & \text{if } 0 < y \leq 1 \\ \frac{1}{3}\Gamma_1 p & \text{if } 1 < y \leq 3 \\ 0 & \text{if } 3 < y, \end{cases} \end{aligned} \tag{2.33}$$

and it is straightforward to check that the Bayesian test $d^\star$ is simply

$$d^\star(y) = 0 \quad \text{iff} \quad \begin{array}{c} -1 < y \leq 0 \\ \text{or} \\ 0 < y \leq 1,\ y < 1 - \frac{\frac{1}{3}\Gamma_1 p}{\Gamma_0(1-p)}. \end{array}$$

Equivalently, $d^\star$ can be described as

$$d^\star(y) = 0 \quad \text{iff} \quad y \in \left(-1, \left(1 - \frac{\Gamma_1 p}{3\Gamma_0(1-p)}\right)^+\right).$$

A final example Consider the binary hypothesis testing problem

$$\begin{aligned} H_1 : &\quad Y \sim F_1 \\ H_0 : &\quad Y \sim F_0 \end{aligned}$$

where F_0 is the discrete uniform distribution on $\{0, 1\}$, and F_1 is uniform on the interval $(0, 1)$. Thus, F_1 admits a probability density function $f_1 : \mathbb{R} \to \mathbb{R}_+$ with respect to Lebesgue measure given by

$$f_1(y) = \begin{cases} 1 & \text{if } y \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathbb{P}\left[Y = 0 | H = 0\right] = \mathbb{P}\left[Y = 1 | H = 0\right] = \frac{1}{2}.$$

In this example F cannot be taken to be either the distribution associated with Lebesgue measure on $\mathbb{R}$ or with the counting measure on $\{0,1\}$. In principle we could use F given by (1.10) but this would yield complicated expressions for the density functions $f_0, f_1 : \mathbb{R} \to \mathbb{R}^+$. Instead of applying Theorem 2.2.1 with that choice, we provide a direct optimization of the auxiliary expected cost function (2.4): For each test d in $\mathcal{D}$ we recall that we have

$$\begin{aligned} & \widehat{J}(d) \\ = \; & \Gamma_0(1-p) + \Gamma_1 p \cdot \mathbb{P}\left[d(Y)=0|H=1\right] - \Gamma_0(1-p) \cdot \mathbb{P}\left[d(Y)=0|H=0\right] \end{aligned}$$

with

$$\mathbb{P}\left[d(Y)=0|H=0\right] = \begin{cases} \frac{1}{2} & \text{if } 0 \in C(d), 1 \notin C(d) \\ \frac{1}{2} & \text{if } 1 \in C(d), 0 \notin C(d) \\ 1 & \text{if } 0 \in C(d), 1 \in C(d) \\ 0 & \text{if } 0 \notin C(d), 1 \notin C(d) \end{cases}$$

and

$$\mathbb{P}\left[d(Y)=0|H=1\right] = \int_{C(d)} f_1(y)dy = |C(d) \cap [0,1]|.$$

Adding or deleting a finite number of points from $C(d)$ will *not* affect the value of $\mathbb{P}\left[d(Y)=0|H=1\right]$, but it may change the value of $\mathbb{P}\left[d(Y)=0|H=0\right]$. Therefore, with $C(d)$ given, modify it, if needed, by adding both points 0 and 1. If C' denotes this Borel subset of $\mathbb{R}$, then $C' = C(d) \cup \{0,1\}$; if d' denotes the corresponding test, then $C(d') = C'$. Obviously

$$\mathbb{P}\left[d(Y)=0|H=1\right] = \mathbb{P}\left[d'(Y)=0|H=1\right] = |C(d') \cap [0,1]|$$

since $|C(d') \cap [0,1]| = |C(d) \cap [0,1]|$, while

$$\mathbb{P}\left[d(Y)=0|H=0\right] \leq \mathbb{P}\left[d'(Y)=0|H=0\right] = 1.$$

We can now conclude that

$$\begin{aligned} & \widehat{J}(d) \\ = \; & \Gamma_0(1-p) + \Gamma_1 p \cdot \mathbb{P}\left[d(Y)=0|H=1\right] - \Gamma_0(1-p) \cdot \mathbb{P}\left[d(Y)=0|H=0\right] \\ \geq \; & \Gamma_0(1-p) + \Gamma_1 p \cdot \mathbb{P}\left[d'(Y)=0|H=1\right] - \Gamma_0(1-p) \cdot \mathbb{P}\left[d'(Y)=0|H=0\right] \\ = \; & \Gamma_0(1-p) + \Gamma_1 p \cdot |C(d') \cap [0,1]| - \Gamma_0(1-p) \\ = \; & \Gamma_1 p \cdot |C(d') \cap [0,1]| \geq 0. \end{aligned} \tag{2.34}$$

Consider the test $d^\star : \mathbb{R} \to \{0, 1\}$ given by

$$d^\star(y) = \mathbf{1}\left[\{0, 1\}\right](y), \quad y \in \mathbb{R}.$$

The arguments leading to (2.34) also show that

$$\widehat{J}(d^\star) = \Gamma_1 p \cdot |C(d^\star) \cap [0, 1]| = 0,$$

and the test $d^\star$ is therefore a Bayesian decision rule.

2.7 Exercises

2.8 References

Chapter 3

Randomized tests

As we shall see shortly, a solution cannot always be found to the Minimax and Neyman–Pearson formulations of the hypothesis testing problem if the search is restricted to the class of decision rules $\mathcal{D}$ as done for the Bayesian set–up. In some very real sense this class $\mathcal{D}$ of tests is not always large enough to guarantee a solution; to remedy this difficulty we enlarge $\mathcal{D}$ by considering the class of *randomized* tests or decision rules.

3.1 Randomized tests

We start with a definition.

A *randomized* test δ is a Borel mapping $\delta : \mathbb{R}^k \to [0,1]$ with the following interpretation as conditional probability: Having observed $\boldsymbol{Y} = \boldsymbol{y}$, it is decided that the state of nature is 1 (resp. 0) with probability $\delta(\boldsymbol{y})$ (resp. $1 - \delta(\boldsymbol{y})$). The collection of all randomized tests will be denoted by $\mathcal{D}^\star$.

Obviously, any test d in $\mathcal{D}$ can be mechanized as a randomized test, say $\delta_d : \mathbb{R}^k \to [0,1]$, given by

$$\delta_d(\boldsymbol{y}) \equiv d(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k.$$

A test in $\mathcal{D}$ is often referred to as a *pure* strategy.

A natural question then arises as to how such randomization mechanisms can be incorporated into the probabilistic framework introduced earlier in Section 1.2: The model data is unchanged as we are given two probability distributions F_0 and F_1 on $\mathbb{R}^k$ and a prior p in $[0,1]$. We still consider a sample space Ω equipped with

a σ-field of events $\mathcal{F}$, and on it we now define the three rvs H, $\boldsymbol{Y}$ and D which take values in $\{0,1\}$, $\mathbb{R}^k$ and $\{0,1\}$, respectively. The rvs H and $\boldsymbol{Y}$ have the same interpretation as before, as state of nature and observation, respectively, while the rv D now encodes the decision to be taken on the basis of the observation $\boldsymbol{Y}$.

With each decision rule δ in $\mathcal{D}^\star$ we associate a probability measure $\mathbb{P}_\delta$ on $\mathcal{F}$ such that the following constraints are satisfied: As before, this time under $\mathbb{P}_\delta$, we still have

$$\mathbb{P}_\delta\left[\boldsymbol{Y} \leq \boldsymbol{y} | H = h\right] = F_h(\boldsymbol{y}), \qquad \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^k, \\ h = 0, 1 \end{array}$$

and

$$p = \mathbb{P}_\delta\left[H = 1\right] = 1 - \mathbb{P}_\delta\left[H = 0\right].$$

Therefore, under $\mathbb{P}_\delta$ the probability distribution of the pair $(H, \boldsymbol{Y})$ does not dependent on δ with

$$\mathbb{P}_\delta\left[H = h, \boldsymbol{Y} \leq \boldsymbol{y}\right] = \mathbb{P}_\delta\left[H = h\right] F_h(\boldsymbol{y}), \qquad \begin{array}{l} h = 0, 1, \\ \boldsymbol{y} \in \mathbb{R}^k \end{array} \tag{3.1}$$

as expected. In addition, for $h = 0, 1$ and $\boldsymbol{y}$ in $\mathbb{R}^k$, we now require that

$$\begin{aligned} \mathbb{P}_\delta\left[D = d | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] &= \begin{cases} 1 - \delta(\boldsymbol{y}) & \text{if } d = 0 \\ \\ \delta(\boldsymbol{y}) & \text{if } d = 1 \end{cases} \\ &= d\delta(\boldsymbol{y}) + (1 - d)\left(1 - \delta(\boldsymbol{y})\right). \end{aligned} \tag{3.2}$$

The joint probability distribution of the rvs H, D and $\boldsymbol{Y}$ (under $\mathbb{P}_\delta$) can now be completely specified: With $h, d = 0, 1$ and a Borel subset B of $\mathbb{R}^k$, a preconditioning argument gives

$$\begin{aligned} &\mathbb{P}_\delta\left[H = h, D = d, \boldsymbol{Y} \in B\right] \\ &\quad = \mathbb{E}_\delta\left[\mathbf{1}\left[H = h, \boldsymbol{Y} \in B\right] \mathbb{P}_\delta\left[D = d | H, \boldsymbol{Y}\right]\right] \\ &\quad = \mathbb{E}_\delta\left[\mathbf{1}\left[H = h, \boldsymbol{Y} \in B\right] \left(d\delta(\boldsymbol{Y}) + (1 - d)\left(1 - \delta(\boldsymbol{Y})\right)\right)\right] \\ &\quad = \mathbb{P}_\delta\left[H = h\right] \cdot \int_B \left(d\delta(\boldsymbol{y}) + (1 - d)\left(1 - \delta(\boldsymbol{y})\right)\right) dF_h(\boldsymbol{y}) \\ &\quad = \begin{cases} \mathbb{P}_\delta\left[H = h\right] \cdot \int_B (1 - \delta(\boldsymbol{y})) dF_h(\boldsymbol{y}) & \text{if } d = 0 \\ \\ \mathbb{P}_\delta\left[H = h\right] \cdot \int_B \delta(\boldsymbol{y}) dF_h(\boldsymbol{y}) & \text{if } d = 1. \end{cases} \end{aligned} \tag{3.3}$$

3.2 An alternate framework

The class $\mathcal{D}^\star$ of randomized strategies gives rise to a *collection* of probability triples, namely

$$\left\{(\Omega, \mathcal{F}, \mathbb{P}_\delta)\,,\ \delta \in \mathcal{D}^\star\right\}.$$

It is however possible to provide an equivalent probabilistic framework using a *single* probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. To see how this can be done, imagine that the original probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ is sufficiently rich that there exists on it a rv $U : \Omega \to [0,1]$ which is uniformly distributed on $(0,1)$, and independent of the pair of rvs H and $\boldsymbol{Y}$, This amounts to

$$\mathbb{P}\left[U \le t, H = h, \boldsymbol{Y} \le \boldsymbol{y}\right] = \mathbb{P}\left[U \le t\right]\mathbb{P}\left[H = h, \boldsymbol{Y} \le \boldsymbol{y}\right], \qquad \begin{array}{c} t \in \mathbb{R} \\ h = 0, 1, \\ \boldsymbol{y} \in \mathbb{R}^k \end{array}$$

with

$$\mathbb{P}\left[U \le t\right] = \begin{cases} 0 & \text{if } t \le 0 \\ \\ \min(t, 1) & \text{if } t \ge 0, \end{cases}$$

$$\mathbb{P}\left[H = h, \boldsymbol{Y} \le \boldsymbol{y}\right] = \mathbb{P}\left[H = h\right] F_h(\boldsymbol{y}), \qquad \begin{array}{c} h = 0, 1, \\ \boldsymbol{y} \in \mathbb{R}^k \end{array}$$

and

$$\mathbb{P}\left[H = 1\right] = p = 1 - \mathbb{P}\left[H = 0\right].$$

Now, for each decision rule δ in $\mathcal{D}^\star$, define the $\{0,1\}$-valued rv D_δ given by

$$D_\delta = \mathbf{1}\left[U \le \delta(\boldsymbol{Y})\right].$$

Note that

$$\begin{aligned} \mathbb{P}\left[D_\delta = 1 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] &= \mathbb{E}\left[\mathbf{1}\left[U \le \delta(\boldsymbol{Y})\right] | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[U \le \delta(\boldsymbol{y})\right] | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] \\ &= \mathbb{P}\left[U \le \delta(\boldsymbol{y})\right] \\ &= \delta(\boldsymbol{y}) \end{aligned} \tag{3.4}$$

under the enforced independence assumptions. Similarly it follows that

$$\begin{aligned} \mathbb{P}\left[D_\delta = 0 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] &= 1 - \mathbb{P}\left[D_\delta = 1 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] \\ &= 1 - \delta(\boldsymbol{y}). \end{aligned} \tag{3.5}$$

Therefore, the conditional distribution of D_δ (under $\mathbb{P}$) given H and $\boldsymbol{Y}$ coincides with the conditional distribution of D (under $\mathbb{P}_\delta$) given H and $\boldsymbol{Y}$, and the two formalisms are probabilistically equivalent.

3.3 Evaluating error probabilities

Evaluating error probabilities under randomized tests can be done rather easily: Consider a randomized test δ in $\mathcal{D}^\star$. In analogy with (1.15) and (1.16), we evaluate the probabilities of false alarm and miss under δ as

$$P_F(\delta) \equiv \mathbb{P}_\delta\left[D = 1 | H = 0\right] \tag{3.6}$$

and

$$P_M(\delta) \equiv \mathbb{P}_\delta\left[D = 0 | H = 1\right]. \tag{3.7}$$

It is also convenient to consider the so–called probability of *detection* given by

$$P_D(\delta) \equiv \mathbb{P}_\delta\left[D = 1 | H = 1\right] = 1 - P_M(\delta). \tag{3.8}$$

Because

$$\mathbb{P}_\delta\left[D = h | H\right] = \mathbb{E}_\delta\left[\mathbb{P}_\delta\left[D = h | H, \boldsymbol{Y}\right] | H\right], \quad h = 0, 1$$

we readily conclude that

$$P_F(\delta) = \int_{\mathbb{R}^k} \delta(\boldsymbol{y}) dF_0(\boldsymbol{y}) \tag{3.9}$$

and

$$P_M(\delta) = \int_{\mathbb{R}^k} \left(1 - \delta(\boldsymbol{y})\right) dF_1(\boldsymbol{y}), \tag{3.10}$$

so that

$$P_D(\delta) = \int_{\mathbb{R}^k} \delta(\boldsymbol{y}) dF_1(\boldsymbol{y}), \tag{3.11}$$

3.4 The Bayesian problem revisited

Assuming the cost function $C : \{0, 1\} \times \{0, 1\} \to \mathbb{R}$ introduced in Section 2.1, we define the expected cost function $J^\star : \mathcal{D}^\star \to \mathbb{R}$ given by

$$J^\star(\delta) = \mathbb{E}_\delta\left[C(H, D)\right], \quad \delta \in \mathcal{D}^\star.$$

When considering randomized decision rules, the original Bayesian Problem $\mathcal{P}_B$ is now reformulated as the minimization problem

$$\mathcal{P}_B^\star: \quad \text{Minimize } J^\star(\delta) \text{ over } \delta \text{ in } \mathcal{D}^\star.$$

This amounts to finding an admissible test $\delta^\star$ in $\mathcal{D}^\star$ such that

$$J^\star(\delta^\star) \leq J^\star(\delta), \quad \delta \in \mathcal{D}^\star. \tag{3.12}$$

Any admissible test $\delta^\star$ which satisfies (3.12) is called a randomized Bayesian test, and the value

$$J^\star(\delta^\star) = \inf_{\delta \in \mathcal{D}^\star} J^\star(\delta)) \tag{3.13}$$

is sometimes referred to as the randomized *Bayesian* cost.

Obviously, since $\mathcal{D} \subset \mathcal{D}^\star$ (with a slight abuse of notation) with

$$J^\star(\delta_d) = J(d), \quad d \in \mathcal{D},$$

it is plain that

$$\inf_{\delta \in \mathcal{D}^\star} J^\star(\delta) \leq \inf_{d \in \mathcal{D}} J(d).$$

While in principle this last inequality could be strict, we now show that it is not so and that the Bayesian problem is not affected by considering the larger set of randomized decision rules; the proof is available in Section 3.6.

Theorem 3.4.1 *Under the absolute continuity condition* **(A.1)**, *it holds that*

$$\inf_{\delta \in \mathcal{D}^\star} J^\star(\delta) = \inf_{d \in \mathcal{D}} J(d). \tag{3.14}$$

It follows from Theorem 2.13 that (3.14) is equivalent to

$$\min_{\delta \in \mathcal{D}^\star} J^\star(\delta) = \min_{d \in \mathcal{D}} J(d) = J(d^\star) \tag{3.15}$$

where the deterministic test $d^\star : \mathbb{R}^k \to \{0, 1\}$ is given by (2.13).

For easy reference we close with the following analog of Lemma 2.1.1 for randomized tests; the proof is left as an exercise.

Lemma 3.4.1 *For any admissible rule δ in $\mathcal{D}^\star$, the relation*

$$J^\star(\delta) = \mathbb{E}\left[C(H,H)\right] + \widehat{J}^\star(\delta) \tag{3.16}$$

holds with

$$\widehat{J}^\star(\delta) = \Gamma_0(1-p)\cdot P_F(\delta) + \Gamma_1 p \cdot P_M(\delta). \tag{3.17}$$

3.5 Randomizing between two pure decision rules

Consider two pure strategies d_1 and d_2 in $\mathcal{D}$. With a in $(0,1)$, we introduce a randomized policy δ_a in $\mathcal{D}^\star$ which first selects the pure strategy d_1 (resp. d_2) with probability a (resp. $1-a$), and then uses the pure policy that was selected. Formally, this amounts to defining $\delta_a : \mathbb{R}^k \to [0,1]$ by

$$\delta_a(\boldsymbol{y}) = a d_1(\boldsymbol{y}) + (1-a) d_2(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k.$$

Applying the expressions (3.9) and (3.10) with the randomized test δ_a we get

$$\begin{aligned} P_F(\delta_a) &= \int_{\mathbb{R}^k} \delta_a(\boldsymbol{y}) dF_0(\boldsymbol{y}) \\ &= \int_{\mathbb{R}^k} \left(a d_1(\boldsymbol{y}) + (1-a) d_2(\boldsymbol{y})\right) dF_0(\boldsymbol{y}) \\ &= a \int_{\mathbb{R}^k} d_1(\boldsymbol{y}) dF_0(\boldsymbol{y}) + (1-a) \int_{\mathbb{R}^k} d_2(\boldsymbol{y}) dF_0(\boldsymbol{y}) \\ &= a P_F(d_1) + (1-a) P_F(d_2). \end{aligned} \tag{3.18}$$

Similarly we find that

$$\begin{aligned} P_M(\delta_a) &= \int_{\mathbb{R}^k} \left(1 - \delta_a(\boldsymbol{y})\right) dF_1(\boldsymbol{y}) \\ &= \int_{\mathbb{R}^k} \left(1 - a d_1(\boldsymbol{y}) - (1-a) d_2(\boldsymbol{y})\right) dF_1(\boldsymbol{y}) \\ &= a \int_{\mathbb{R}^k} (1 - d_1(\boldsymbol{y})) dF_1(\boldsymbol{y}) + (1-a) \int_{\mathbb{R}^k} (1 - d_2(\boldsymbol{y})) dF_1(\boldsymbol{y}) \\ &= a P_M(d_1) + (1-a) P_M(d_2). \end{aligned} \tag{3.19}$$

It immediately follows from (3.16) and (3.17) that

$$J_p^\star(\delta_a) = aJ_p(d_1) + (1-a)J_p(d_2). \tag{3.20}$$

as we use the relations (3.18) and (3.19).

One very concrete way to implement the randomized policy δ_a on the original triple $(\Omega, \mathcal{F}, \mathbb{P})$ proceeds as follows: Consider the original probabilistic framework introduced in Section 1.2 and assume it to be sufficiently rich to carry an additional $\mathbb{R}$-valued rv V which is independent of the rvs H and $\boldsymbol{Y}$ (under $\mathbb{P}$), and is uniformly distributed on the interval $[0,1]$. Define the $\{0,1\}$-valued rv B_a given by

$$B_a = \mathbf{1}\left[V \leq a\right].$$

It is plain that the rv B_a is independent of the rvs H and $\boldsymbol{Y}$ (under $\mathbb{P}$), with

$$\mathbb{P}\left[B_a = 1\right] = a = 1 - \mathbb{P}\left[B_a = 0\right].$$

Define the decision rv D_a given by

$$D_a = B_a d_1(\boldsymbol{Y}) + (1 - B_a) d_2(\boldsymbol{Y}).$$

It is easy to check that

$$\begin{aligned}
& \mathbb{P}\left[D_a = 1 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] \\
= \;& \mathbb{P}\left[B_a d_1(\boldsymbol{Y}) + (1 - B_a) d_2(\boldsymbol{Y}) = 1 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] \\
= \;& \mathbb{P}\left[B_a d_1(\boldsymbol{y}) + (1 - B_a) d_2(\boldsymbol{y}) = 1 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] \\
= \;& \mathbb{P}\left[B_a = 1, d_1(\boldsymbol{y}) = 1 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] + \mathbb{P}\left[B_a = 0, d_2(\boldsymbol{y}) = 1 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] \\
= \;& d_1(\boldsymbol{y}) \mathbb{P}\left[B_a = 1 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] + d_2(\boldsymbol{y}) \mathbb{P}\left[B_a = 0 | H = h, \boldsymbol{Y} = \boldsymbol{y}\right] \\
= \;& d_1(\boldsymbol{y}) \mathbb{P}\left[B_a = 1\right] + d_2(\boldsymbol{y}) \mathbb{P}\left[B_a = 0\right] \\
= \;& a d_1(\boldsymbol{y}) + (1-a) d_2(\boldsymbol{y}), \qquad \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^k, \\ h = 0, 1 \end{array}
\end{aligned} \tag{3.21}$$

as desired.

3.6 A proof of Theorem 3.4.1

Pick an arbitrary test δ in $\mathcal{D}^\star$. A simple preconditioning argument shows that

$$J^\star(\delta) \;=\; \mathbb{E}_\delta\left[C(H, D)\right]$$

$$\begin{aligned} &= \mathbb{E}_\delta\left[\mathbb{E}_\delta\left[C(H,D)|H,\boldsymbol{Y}\right]\right] \\ &= \mathbb{E}_\delta\left[C(H,1)\mathbb{P}_\delta\left[D=1|H,\boldsymbol{Y}\right]+C(H,0)\mathbb{P}_\delta\left[D=0|H,\boldsymbol{Y}\right]\right] \\ &= \mathbb{E}_\delta\left[C(H,1)\cdot\delta(\boldsymbol{Y})+C(H,0)\cdot(1-\delta(\boldsymbol{Y}))\right] \\ &= \mathbb{E}_\delta\left[C(H,0)\right]+\mathbb{E}_\delta\left[(C(H,1)-C(H,0))\cdot\delta(\boldsymbol{Y})\right] \end{aligned} \tag{3.22}$$

with

$$\begin{aligned} &\mathbb{E}_\delta\left[(C(H,1)-C(H,0))\cdot\delta(\boldsymbol{Y})\right] \\ &\quad= \mathbb{E}_\delta\left[(C(H,1)-C(H,0))\cdot\mathbb{E}_\delta\left[\delta(\boldsymbol{Y})|H\right]\right] \\ &\quad= (C(1,1)-C(1,0))\,\mathbb{E}_\delta\left[\delta(\boldsymbol{Y})|H=1\right]\mathbb{P}_\delta\left[H=1\right] \\ &\qquad+(C(0,1)-C(0,0))\,\mathbb{E}_\delta\left[\delta(\boldsymbol{Y})|H=0\right]\mathbb{P}_\delta\left[H=0\right] \\ &\quad= -\Gamma_1 p\cdot\mathbb{E}_\delta\left[\delta(\boldsymbol{Y})|H=1\right]+\Gamma_0(1-p)\cdot\mathbb{E}_\delta\left[\delta(\boldsymbol{Y})|H=0\right]. \end{aligned} \tag{3.23}$$

Using the absolute continuity condition **(A.1)** we can now write

$$\mathbb{E}_\delta\left[\delta(\boldsymbol{Y})|H=h\right]=\int_{\mathbb{R}^k}\delta(\boldsymbol{y})dF_h(\boldsymbol{y})=\int_{\mathbb{R}^k}\delta(\boldsymbol{y})f_h(\boldsymbol{y})dF(\boldsymbol{y}),\quad h=0,1$$

so that

$$\begin{aligned} &J^\star(\delta)-\mathbb{E}_\delta\left[C(H,0)\right] \\ &\quad= -\Gamma_1 p\cdot\int_{\mathbb{R}^k}\delta(\boldsymbol{y})f_1(\boldsymbol{y})dF(\boldsymbol{y})+\Gamma_0(1-p)\cdot\int_{\mathbb{R}^k}\delta(\boldsymbol{y})f_0(\boldsymbol{y})dF(\boldsymbol{y}) \\ &\quad= \int_{\mathbb{R}^k}\left(-\Gamma_1 p f_1(\boldsymbol{y})+\Gamma_0(1-p)f_0(\boldsymbol{y})\right)\delta(\boldsymbol{y})dF(\boldsymbol{y}) \\ &\quad= -\int_{\mathbb{R}^k}h(\boldsymbol{y})\delta(\boldsymbol{y})dF(\boldsymbol{y}) \end{aligned} \tag{3.24}$$

where the mapping $h:\mathbb{R}^k\to\mathbb{R}$ is given by (2.11). Note that the term $\mathbb{E}_\delta\left[C(H,0)\right]$ does not depend on the randomized test δ being used.

From Theorem 2.2.1 recall that the Bayesian rule which solves Problem $\mathcal{P}_B$ is the test $d^\star:\mathbb{R}^k\to\{0,1\}$ in $\mathcal{D}$ given by (2.13). Note that $d^\star$ can also be interpreted as the randomized rule $\delta^\star:\mathbb{R}^k\to[0,1]$ given by

$$\delta^\star(\boldsymbol{y})=\begin{cases}0 & \text{if } h(\boldsymbol{y})<0\\ \\ 1 & \text{if } h(\boldsymbol{y})\geq 0\end{cases}=\begin{cases}0 & \text{if } \boldsymbol{y}\in C^\star\\ \\ 1 & \text{if } \boldsymbol{y}\notin C^\star\end{cases}$$

where $C^\star$ is the Borel subset of $\mathbb{R}^k$ given by (2.12).

The desired result will be established if we show that

$$J^\star(\delta^\star) \leq J^\star(\delta), \quad \delta \in \mathcal{D}^\star.$$

The approach we take is reminiscent of the one used in the proof of Theorem 2.2.1: For an arbitrary δ in $\mathcal{D}^\star$, earlier calculations (3.24) show that

$$\begin{aligned} J^\star(\delta) - J^\star(\delta^\star) &= -\int_{\mathbb{R}^k} h(\boldsymbol{y})\delta(\boldsymbol{y})dF(\boldsymbol{y}) + \int_{\mathbb{R}^k} h(\boldsymbol{y})\delta^\star(\boldsymbol{y})dF(\boldsymbol{y}) \\ &= \int_{\mathbb{R}^k} h(\boldsymbol{y})\left(\delta^\star(\boldsymbol{y}) - \delta(\boldsymbol{y})\right)dF(\boldsymbol{y}) \\ &= \int_{C^\star} (-h(\boldsymbol{y}))\delta(\boldsymbol{y})dF(\boldsymbol{y}) + \int_{\mathbb{R}^k \setminus C^\star} \left(1 - \delta(\boldsymbol{y})\right) h(\boldsymbol{y})dF(\boldsymbol{y}) \\ &\geq 0 \end{aligned}$$

as desired since

$$\int_{C^\star} (-h(\boldsymbol{y}))dF(\boldsymbol{y}) \geq 0 \quad \text{and} \quad \int_{\mathbb{R}^k \setminus C^\star} \left(1 - \delta(\boldsymbol{y})\right) h(\boldsymbol{y})dF(\boldsymbol{y}) \geq 0$$

by the very definition of the set $C^\star$ and of the mapping $h : \mathbb{R}^k \to \mathbb{R}$. ■

3.7 Exercises

3.8 References

Chapter 4

The Minimax formulation

The Bayesian formulation *implicitly* assumes knowledge of the prior distribution on the hypothesis rv H. In many situations, this assumption cannot be adequately justified, and the Bayesian formulation has to abandoned for the so–called *Minimax formulation* discussed in this chapter.

4.1 Keeping track of the prior

To facilitate the discussion, we augment the notation introduced in Chapter 1 and Chapter 3 by explicitly indicating the dependence on the prior probability distribution: As before we are given two distinct probability distributions $F_0, F_1 : \mathbb{R}^k \to [0,1]$ which act as conditional probability distributions for the observation given the state of nature. As in Chapter 1, we can always construct a collection $\{(\Omega, \mathcal{F}, \mathbb{P}_p),\ p \in [0,1]\}$ of probability triples, and rvs H and $\boldsymbol{Y}$ defined on Ω which take values in $\{0,1\}$ and $\mathbb{R}^k$, respectively, such that for each p in $[0,1]$,

$$F_h(\boldsymbol{y}) = \mathbb{P}_p\left[\boldsymbol{Y} \leq \boldsymbol{y} | H = h\right], \qquad \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^k, \\ h = 0,1 \end{array}$$

and

$$p = \mathbb{P}_p\left[H = 1\right] = 1 - \mathbb{P}_p\left[H = 0\right].$$

One possible construction was given in Section 1.3: Take $\Omega = \{0,1\} \times \mathbb{R}^k$ with generic element $\omega = (h, \boldsymbol{y})$ with $h = 0, 1$ and $\boldsymbol{y}$ an arbitrary element of $\mathbb{R}^k$. We endow Ω with the σ-field $\mathcal{F}$ given by

$$\mathcal{F} = \sigma\left(\mathcal{P}(\{0,1\}) \times \mathcal{B}(\mathbb{R}^k)\right)$$

where $\mathcal{P}(\{0,1\})$ is the power set of $\{0,1\}$, and $\mathcal{B}(\mathbb{R}^k)$ is the Borel σ-field on $\mathbb{R}^k$. We define the mappings $H : \Omega \to \mathbb{R}$ and $\boldsymbol{Y} : \Omega \to \mathbb{R}^k$ by

$$H(\omega) = h \quad \text{and} \quad \boldsymbol{Y}(\omega) = \boldsymbol{y}, \quad \omega = (h, \boldsymbol{y}) \in \Omega.$$

Both projection mappings are Borel measurable, and therefore define rvs.

As before, it is plain that

$$\begin{aligned} \mathbb{P}_p\left[\boldsymbol{Y} \le y, H = h\right] &= \mathbb{P}_p\left[\boldsymbol{Y} \le \boldsymbol{y} | H = h\right] \mathbb{P}_p\left[H = h\right] \\ &= \begin{cases} (1-p)F_0(\boldsymbol{y}) & \text{if } h = 0, \boldsymbol{y} \in \mathbb{R}^k \\ \\ pF_1(\boldsymbol{y}) & \text{if } h = 1, \boldsymbol{y} \in \mathbb{R}^k. \end{cases} \end{aligned} \tag{4.1}$$

Let $\mathbb{E}_p\left[\cdot\right]$ denote expectation under $\mathbb{P}_p$.

When dealing with randomized strategies we further augment the notation $\mathbb{P}_\delta$ to read $\mathbb{P}_{\delta,p}$ when using the randomized strategy δ in $\mathcal{D}^\star$ with prior p; see Section 3.1 for details on the probabilistic framework to be used.. In that case let $\mathbb{E}_{\delta,p}\left[\cdot\right]$ denote expectation under $\mathbb{P}_{\delta,p}$.

4.2 The Bayesian problems

Fix p in $[0,1]$. Let $J_p(d)$ denote the expected cost associated with the admissible decision rule d in $\mathcal{D}$ when the prior on H is p, i.e.,

$$J_p(d) \equiv \mathbb{E}_p\left[C(H, d(\boldsymbol{Y}))\right].$$

Similarly, let $J_p^\star(\delta)$ denote the expected cost associated under the randomized decision rule δ in $\mathcal{D}^\star$ when the prior on H is p, i.e.,

$$J_p^\star(\delta) \equiv \mathbb{E}_{\delta,p}\left[C(H, D)\right].$$

The Bayesian problems introduced in Chapters 2 and 3 now read

$$\mathcal{P}_{p,\mathrm{B}}: \quad \text{Minimize } J_p(d) \text{ over } d \text{ in } \mathcal{D}$$

and

$$\mathcal{P}^\star_{p,\mathrm{B}}: \quad \text{Minimize } J_p^\star(\delta) \text{ over } \delta \text{ in } \mathcal{D}^\star.$$

The corresponding Bayesian costs will be denoted by

$$V(p) \equiv \inf_{d \in \mathcal{D}} J_p(d) \tag{4.2}$$

and

$$V^\star(p) \equiv \inf_{\delta \in \mathcal{D}^\star} J_p^\star(\delta). \tag{4.3}$$

As shown in Chapter 2, under Condition **(A.1)**, for each p in $[0,1]$ the problem $\mathcal{P}_{p,\mathrm{B}}$ has a solution which we denote $d^\star(p)$ to indicate its dependence on the prior p. Clearly, any such solution satisfies

$$J_p(d^\star(p)) \leq J_p(d), \quad d \in \mathcal{D} \tag{4.4}$$

and the equality

$$V(p) = J_p(d^\star(p)) \tag{4.5}$$

holds. Under the same condition, Theorem 3.4.1 further shows that

$$J_p^\star(\delta_{d^\star(p)}) \leq J_p(\delta), \quad \delta \in \mathcal{D}^\star$$

so that

$$V^\star(p) = V(p). \tag{4.6}$$

The following properties of the value function $V : [0,1] \to \mathbb{R}$ will be useful in the forthcoming discussion. Conditions **(A.1)** and **(A.2)** are not needed for the results to hold.

Lemma 4.2.1 *Assume $\Gamma_h > 0$ for $h = 0, 1$. The value function $V : [0,1] \to \mathbb{R}$ is concave and continuous on the closed interval $[0,1]$ with boundary values $V(0) = C(0,0)$ and $V(1) = C(1,1)$. Moreover, its right-derivative (resp. left-derivative) exists and is finite on $[0,1)$ (resp. $(0,1]$))*

The proof can be omitted in a first reading, and can be found in Section 4.10. For easy reference, recall that for each p in $[0,1]$ the expressions

$$\begin{aligned} J_p(d) &= pC(1,1) + (1-p)C(0,0) \\ &\quad + \Gamma_0(1-p) \cdot P_F(d) + \Gamma_1 p \cdot P_M(d), \quad d \in \mathcal{D} \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} J_p^\star(\delta) &= pC(1,1) + (1-p)C(0,0) \\ &\quad + \Gamma_0(1-p) \cdot P_F(\delta) + \Gamma_1 p \cdot P_M(\delta), \quad \delta \in \mathcal{D}^\star \end{aligned} \tag{4.8}$$

hold. The relationships were given in Lemma 2.1.1 and Lemma 3.4.1, respectively.

4.3 The minimax formulation

Since the exact value of the prior p is not available, the Bayesian criterion has to be modified. Two different approaches are possible; each in its own way seeks to *compensate* for the uncertainty in the modeling assumptions.

Minmax On possible approach is to introduce a *worst-case* cost associated with the original cost, and then use it as the new criterion to be minimized. With this in mind, define

$$J_{\mathrm{Max}}(d) \equiv \sup_{p\in[0,1]} J_p(d), \quad d \in \mathcal{D}. \tag{4.9}$$

We are then lead to consider the minimization problem

$$\mathcal{P}_{\mathrm{Max}}: \quad \text{Minimize } J_{\mathrm{Max}}(d) \text{ over } d \text{ in } \mathcal{D}.$$

Solving $\mathcal{P}_{\mathrm{Max}}$ amounts to finding an admissible test $d^\star_{\mathrm{m}}$ in $\mathcal{D}$ such that

$$J_{\mathrm{Max}}(d^\star_{\mathrm{m}}) \leq J_{\mathrm{Max}}(d), \quad d \in \mathcal{D}. \tag{4.10}$$

When it exists, the test $d^\star_{\mathrm{m}}$ is known as a *minimax* test.

A priori there is no guarantee that a test in $\mathcal{D}$ exists which satisfies (4.10) (even under Condition **(A.1)**) – It is not clear that a cost $\widetilde{C} : \{0,1\} \times \{0,1\} \to \mathbb{R}$ (likely related to the original cost $C : \{0,1\} \times \{0,1\} \to \mathbb{R}$) and $\tilde{p}$ in $[0,1]$ can be found such that

$$J_{\mathrm{Max}}(d) = \mathbb{E}_{\tilde{p}}\left[\widetilde{C}(H, d(\boldsymbol{Y}))\right], \quad d \in \mathcal{D}.$$

If that were indeed the case, then Theorem 2.2.1 would guarantee the existence of a minimizer.

For technical reasons to become shortly apparent we also introduce the worst-case cost under randomized strategies, namely

$$J^\star_{\mathrm{Max}}(\delta) \equiv \sup_{p\in[0,1]} J^\star_p(\delta), \quad \delta \in \mathcal{D}^\star. \tag{4.11}$$

The minimization problem of interest here is now defined as

$$\mathcal{P}^\star_{\mathrm{Max}}: \quad \text{Minimize } J^\star_{\mathrm{Max}}(\delta) \text{ over } \delta \text{ in } \mathcal{D}^\star.$$

Solving $\mathcal{P}^\star_{\rm Max}$ amounts to finding a randomized strategy $\delta^\star_{\rm m}$ in $\mathcal{D}^\star$ such that

$$J^\star_{\rm Max}(\delta^\star_{\rm m}) \leq J^\star_{\rm Max}(\delta), \quad \delta \in \mathcal{D}^\star. \tag{4.12}$$

Again a priori there is no guarantee that there exists a test in $\mathcal{D}^\star$ satisfying (4.12) (even under Condition **(A.1)**). When it exists, the test $\delta^\star_{\rm m}$ is also known as a *minimax* test.

It is natural to wonder whether the tests $d^\star_{\rm m}$ and $\delta^\star_{\rm m}$ exist, (possibly under additional conditions), whether they are different, and if not, whether $J^\star_{\rm Max}(\delta^\star_m) = J_{\rm Max}(d^\star_{\rm m})$.

Maxmin Another reasonable way to proceed consists in using the Bayesian test for that value of p which yields the *largest* Bayesian cost (4.2): With the notation introduced earlier, let $p_{\rm m}$ in $[0,1]$ such that

$$V(p_{\rm m}) = \max_{p\in[0,1]} V(p), \tag{4.13}$$

and use the Bayesian rule $d^\star(p_{\rm m})$ – The existence of $p_{\rm m}$ is guaranteed by the fact that the mapping $V : [0,1] \to \mathbb{R}$ is continuous on the closed bounded interval $[0,1]$ by Lemma 4.2.1, hence achieves its maximum value on $[0,1]$.

The value $p_{\rm m}$ satisfying (4.13) is known as the *least favorable* prior. Although the terminology is not standard, we shall refer to $d^\star(p_{\rm m})$ as a *maximin* test.

4.4 Preliminary facts

In view of the two competing approaches outlined in Section 4.3, several questions arise: (i) How does one characterize the minimax strategy $d^\star_{\rm m}$ and develop ways find it; (ii) How does one characterize the least-favorable prior $p_{\rm m}$ and develop ways find it; (iii) Is there a simple relationship between the solutions proposed by two approaches, and in particular, whether is $d^\star(p_{\rm m})$ is a candidate for $d^\star_{\rm m}$.

To frame the discussion of these issues we start with a couple of preliminary remarks.

The minimax inequalities As a first step towards understanding how the two approaches may be related to each other, consider the following arguments: From the definitions it always holds that

$$V(p) \leq J_p(d) \leq J_{\text{Max}}(d), \qquad \begin{matrix} p \in [0,1] \\ d \in \mathcal{D}. \end{matrix} \tag{4.14}$$

It is now immediate that

$$V(p) \leq \inf_{d \in \mathcal{D}} J_{\text{Max}}(d), \quad p \in [0,1]$$

since $V(p)$ does not depend on d, whence

$$\sup_{p \in [0,1]} V(p) \leq \inf_{d \in \mathcal{D}} J_{\text{Max}}(d).$$

This last inequality can be rewritten as the *minimax inequality*

$$\sup_{p \in [0,1]} \left(\inf_{d \in \mathcal{D}} J_p(d) \right) \leq \inf_{d \in \mathcal{D}} \left(\sup_{p \in [0,1]} J_p(d) \right) \tag{4.15}$$

(in pure policies)

If we were to consider randomized strategies, it is also the case that

$$V^\star(p) \leq J_p^\star(\delta) \leq J_{\text{Max}}^\star(\delta), \qquad \begin{matrix} p \in [0,1] \\ \delta \in \mathcal{D}^\star \end{matrix} \tag{4.16}$$

and arguments similar to the ones leading to (4.15) yield the minimax inequality

$$\sup_{p \in [0,1]} \left(\inf_{\delta \in \mathcal{D}^\star} J_p^\star(\delta) \right) \leq \inf_{\delta \in \mathcal{D}^\star} \left(\sup_{p \in [0,1]} J_p^\star(\delta) \right) \tag{4.17}$$

in randomized strategies.

Toward minimax equalities As we contrast the inequalities (4.15) and (4.17), it is natural to wonder whether these inequalities ever hold as *equalities*, namely

$$\sup_{p \in [0,1]} \left(\inf_{d \in \mathcal{D}} J_p(d) \right) = \inf_{d \in \mathcal{D}} \left(\sup_{p \in [0,1]} J_p(d) \right) \tag{4.18}$$

and

$$\sup_{p\in[0,1]}\left(\inf_{\delta\in\mathcal{D}^\star} J_p^\star(\delta)\right) = \inf_{\delta\in\mathcal{D}^\star}\left(\sup_{p\in[0,1]} J_p^\star(\delta)\right). \tag{4.19}$$

When this occurs we shall then say that the minimax property holds in pure and randomized policies, respectively.

It is worth pointing out that the equalities

$$\inf_{\delta\in\mathcal{D}^\star}\left(\sup_{p\in[0,1]} J_p^\star(\delta)\right) \leq \inf_{d\in\mathcal{D}}\left(\sup_{p\in[0,1]} J_p(d)\right) \tag{4.20}$$

and

$$\sup_{p\in[0,1]}\left(\inf_{\delta\in\mathcal{D}^\star} J_p^\star(\delta)\right) = \sup_{p\in[0,1]}\left(\inf_{d\in\mathcal{D}} J_p(d)\right) \tag{4.21}$$

always hold; the latter is a rewrite of (4.6) and is a simple consequence of Theorem 3.4.1. As we combine these observations with (4.17) we conclude that

$$\sup_{p\in[0,1]}\left(\inf_{d\in\mathcal{D}} J_p(d)\right) \leq \inf_{\delta\in\mathcal{D}^\star}\left(\sup_{p\in[0,1]} J_p^\star(\delta)\right) \leq \inf_{d\in\mathcal{D}}\left(\sup_{p\in[0,1]} J_p(d)\right). \tag{4.22}$$

Thus, if (4.18) happens to be true, then (4.19) necessarily holds – Put differently, the minimax property in pure policies is more difficult to achieve than the minimax property in randomized strategies. This disparity will become apparent in the discussion of the Minimax Theorem given in Section 4.5, opening the possibility that we may have to resort to randomized tests (at least in some situations) in order to achieve the minimax equality.

The structure of the worst-case costs (4.9) and (4.11) A little more can be said concerning the auxiliary costs (4.9) and (4.11): For each test d in $\mathcal{D}$, we note from (4.7) and (4.8) that

$$\sup_{p\in[0,1]} J_p(d) = \max_{p\in[0,1]} J_p(d) = \max\{J_0(d), J_1(d)\} \tag{4.23}$$

with the supremum achieved at either $p = 0$ or $p = 1$. Also, $J_0(d)$ and $J_1(d)$ can be given probabilistic interpretations as the conditional interpretations

$$J_0(d) = \mathbb{E}_p\left[C(H, d(\boldsymbol{Y})|H = 0\right] \tag{4.24}$$

and

$$J_1(d) = \mathbb{E}_p\left[C(H, d(\boldsymbol{Y})|H = 1\right] \tag{4.25}$$

with p arbitrary in $[0, 1]$. Similarly, for each randomized strategy δ in $\mathcal{D}^\star$, we have

$$\sup_{p\in[0,1]} J_p^\star(\delta) = \max_{p\in[0,1]} J_p^\star(\delta) = \max\{J_0^\star(\delta), J_1^\star(\delta)\} \tag{4.26}$$

with the supremum achieved at either $p = 0$ or $p = 1$ with probabilistic interpretations

$$J_0^\star(\delta) = \mathbb{E}_{\delta,p}\left[C(H, D)|H = 0\right] \tag{4.27}$$

and

$$J_1^\star(\delta) = \mathbb{E}_{\delta,p}\left[C(H, D)|H = 1\right] \tag{4.28}$$

with p arbitrary in $[0, 1]$.

4.5 The minimax equality

The main result concerning the minimax formulation for the binary hypothesis testing problem is summarized in the following special case of the Minimax Theorem from Statistical Decision Theory; see [**?**, Thm. 1, p. 82] for a discussion in a more general setting.

Theorem 4.5.1 *Assume $\Gamma_h > 0$ for all $h = 0, 1$. Under Condition (**A.1**), the minimax equality*

$$\sup_{p\in[0,1]}\left(\inf_{\delta\in\mathcal{D}^\star} J_p^\star(\delta)\right) = \inf_{\delta\in\mathcal{D}^\star}\left(\sup_{p\in[0,1]} J_p^\star(\delta)\right) \tag{4.29}$$

holds in randomized strategies.

In Section 4.6 we present an analysis of the minimax equality which exploits the specific structure of the binary hypothesis problem as reflected through the properties of the value function: As pointed out earlier, there always exists $p_{\rm m}$ in $[0, 1]$ such that (4.13) holds. From the concavity of the value function it follows that the set of maximizers

$$I_{\rm m} \equiv \left\{p_{\rm m} \in [0, 1] :\ V(p_{\rm m}) = \max_{p\in[0,1]} V(p)\right\}$$

is a closed interval in $[0,1]$. The set $I_{\rm m}$ will often be reduced to a singleton, in which case the value function admits a unique (isolated) maximizer. Four situations can occur depending on the location of $I_{\rm m}$ and on the smoothness of $p \to V(p)$ at the maximum. In each case we establish a minimax equality and identify the minimax strategy. Throughout we still use $d^\star(p_{\rm m})$ to denote the Bayesian test for the selected value $p_{\rm m}$ in $I_{\rm m}$, so that

$$V(p_{\rm m}) = J_{p_{\rm m}}(d^\star(p_{\rm m})) = \min_{d\in\mathcal{D}} J_{p_{\rm m}}(d). \tag{4.30}$$

From the discussion of Section 4.4 we see that (4.29) will hold if we can establish the reverse inequality to (4.17), namely

$$\inf_{\delta\in\mathcal{D}^\star}\left(\sup_{p\in[0,1]} J_p^\star(\delta)\right) \leq \sup_{p\in[0,1]}\left(\inf_{\delta\in\mathcal{D}^\star} J_p^\star(\delta)\right). \tag{4.31}$$

Recall that (4.31) will automatically hold if we show the stronger inequality

$$\inf_{d\in\mathcal{D}}\left(\sup_{p\in[0,1]} J_p(d)\right) \leq \sup_{p\in[0,1]}\left(\inf_{d\in\mathcal{D}} J_p(d)\right). \tag{4.32}$$

In the first three cases we show in effect that

$$\inf_{d\in\mathcal{D}}\left(\max_{p\in[0,1]} J_p(d)\right) \leq \max_{p\in[0,1]}\left(\min_{d\in\mathcal{D}} J_p(d)\right). \tag{4.33}$$

4.6 A proof of Theorem 4.5.1

We start with the boundary cases $p_{\rm m} = 0$ and $p_{\rm m} = 1$.

Case 1: Assume $p_{\rm m} = 0$ – Thus, $\max_{p\in[0,1]} V(p) = V(0) = J_0(d^\star(0))$. By concavity we have $\frac{d^+}{dp}V(p)\Big|_{p=0} \leq 0$ with the mapping $V : [0,1] \to \mathbb{R}$ being decreasing. But the straight line $p \to J_p(d^\star(0))$ is tangent to the value function $V : [0,1] \to \mathbb{R}$ at $p = 0$, whence

$$\frac{d^+}{dp}V(p)\Big|_{p=0} = \frac{d}{dp}J_p(d^\star(0))\Big|_{p=0} \leq 0.$$

The mapping $p \to J_p(d^\star(0))$ being affine, its derivative is therefore constant with

$$\frac{d}{dp}J_p(d^\star(0)) = \frac{d}{dp}J_p(d^\star(0))\Big|_{p=0} \leq 0, \quad p \in [0,1]$$

and the mapping $p \to J_p(d^\star(0))$ is also decreasing on $[0,1]$. This leads to

$$J_0(d^\star(0)) = \max_{p\in[0,1]} J_p(d^\star(0)).$$

With this in mind we get

$$\begin{aligned}\max_{p\in[0,1]}\left(\min_{d\in\mathcal{D}} J_p(d)\right) &= \max_{p\in[0,1]} V(p)\\ &= V(0)\\ &= J_0(d^\star(0))\\ &= \max_{p\in[0,1]} J_p(d^\star(0)).\end{aligned} \tag{4.34}$$

The desired inequality (4.32) (hence (4.31)) is now immediate from (4.34) as we note that

$$\max_{p\in[0,1]} J_p(d^\star(0)) \geq \inf_{d\in\mathcal{D}}\left(\max_{p\in[0,1]} J_p(d)\right).$$

But the minimax equality being now established in pure strategies, we conclude from the discussion that

$$\max_{p\in[0,1]} J_p(d^\star(0)) = \inf_{d\in\mathcal{D}}\left(\max_{p\in[0,1]} J_p(d)\right).$$

This shows that $d^\star_{\rm m}$ can be taken to be $d^\star(0)$. ■

Case 2: Assume $p_{\rm m} = 1$ – The proof is as in Case 1 *mutatis mutandis*, and is left as an exercise. Again, the minimax equality holds in pure strategies and $d^\star_{\rm m}$ can be taken to be $d^\star(1)$. ■

We now turn to cases when $p_{\rm m}$ is selected in $(0,1)$.

Case 3: Assume that $p_{\rm m}$ in an element of $(0,1)$ and $p \to V(p)$ is differentiable at $p = p_{\rm m}$ – It is plain that $\frac{d}{dp}V(p)\Big|_{p=p_{\rm m}} = 0$ since $p_{\rm m}$ is an interior point by assumption. By concavity the mapping $p \to J_p(d^\star(p_{\rm m}))$ is tangent to the value function $V:[0,1]\to\mathbb{R}$ at $p = p_{\rm m}$, whence

$$\frac{d}{dp}V(p)\Big|_{p=p_{\rm m}} = \frac{d}{dp}J_p(d^\star(p_{\rm m}))\Big|_{p=p_{\rm m}} = 0.$$

The mapping $p \to J_p(d^\star(p_{\rm m}))$ being affine, its derivative is constant and given by

$$\frac{d}{dp} J_p(d^\star(p_{\rm m})) = \frac{d}{dp} J_p(d^\star(p_{\rm m}))\Big|_{p=0} = 0, \quad p \in [0,1].$$

Therefore, the mapping $p \to J_p(d^\star(p_{\rm m}))$ is constant on $[0,1]$, and the equality $J_0(d^\star(p_{\rm m})) = J_1(d^\star(p_{\rm m}))$ holds. It follows from the first equality in (4.30) that

$$V(p_{\rm m}) = J_p(d^\star(p_{\rm m})) = \max_{p\in[0,1]} J_p(d^\star(p_{\rm m})), \quad p \in [0,1]. \tag{4.35}$$

On the other hand, it is plain that

$$\begin{aligned} \inf_{d\in\mathcal{D}} \left(\max_{p\in[0,1]} J_p(d) \right) &\leq \max_{p\in[0,1]} J_p(d^\star(p_{\rm m})) \\ &= J_{p_{\rm m}}(d^\star(p_{\rm m})) \\ &= \min_{d\in\mathcal{D}} J_{p_{\rm m}}(d) \\ &\leq \inf_{d\in\mathcal{D}} \left(\max_{p\in[0,1]} J_p(d) \right) \end{aligned} \tag{4.36}$$

as we use the second equality in (4.35) with $p = p_{\rm m}$, and then apply the second equality in (4.30). The inequality (4.32) (hence (4.31)) is now a straightforward consequence of (4.36).

Leveraging the fact that the minimax equality is now known to hold in pure strategies, we conclude from the discussion that

$$\max_{p\in[0,1]} J_p(d^\star(p_{\rm m})) = \inf_{d\in\mathcal{D}} \left(\max_{p\in[0,1]} J_p(d) \right),$$

and $d^\star_{\rm m}$ can therefore be taken to be $d^\star(p_{\rm m})$. ■

Case 4: Assume that $I_{\rm m} = \{p_{\rm m}\} \subseteq (0,1)$ but $p \to V(p)$ is not differentiable at $p = p_{\rm m}$ – Under such assumptions we must have

$$a_+ \equiv \frac{d^+}{dp} V(p)\Big|_{p=p_{\rm m}} < \frac{d^-}{dp} V(p)\Big|_{p=p_{\rm m}} \equiv a_-$$

by concavity with either $a_+ < 0 \leq a_-$ or $a_+ \leq 0 < a_-$. We continue the discussion under the assumption $a_+ < 0 \leq a_-$; the case $a_+ \leq 0 < a_-$ proceeds along similar lines, and is therefore omitted.

Recall that $p \to V(p)$ is defined as the envelope of a family of affine functions. Thus, under the non-differentiability assumption at $p = p_{\rm m}$, concavity guarantees that there exist two pure strategies, say $d_-, d_+ : \mathbb{R}^k \to \{0, 1\}$, such that $V(p_{\rm m}) = J_{p_{\rm m}}(d_-)$ and $V(p_{\rm m}) = J_{p_{\rm m}}(d_+)$ (because $p_{\rm m}$ is a maximum) while the straight lines $p \to J_p(d_-)$ and $p \to J_p(d_+)$ are both tangent to the value function at $p = p_{\rm m}$ – These two strategies are distinct. Hence, as discussed in earlier cases, the function $p \to J_p(d_-)$ (resp. $p \to J_p(d_+)$) is an affine function with constant derivative $a_- \geq 0$ (resp. $a_+ < 0$), hence non-decreasing (resp. decreasing). It follows that $J_0(d_-) \leq J_1(d_-)$ and $J_1(d_+) < J_0(d_+)$.

Next we introduce randomized policies $\{\delta_a,\ a \in [0, 1]\}$ obtained by randomizing two pure strategies d_- and d_+. Thus, with each a in $[0, 1]$ consider the randomized policy $\delta_a : \mathbb{R}^k \to [0, 1]$ given by

$$\delta_a = ad_+ + (1 - a)d_-.$$

The relation (3.20) discussed in Section 3.5 applies, yielding

$$J^\star_p(\delta_a) = aJ_p(d_+) + (1 - a)J_p(d_-), \quad p \in [0, 1].$$

By construction we also note that

$$V(p_{\rm m}) = J^\star_{p_{\rm m}}(\delta_a), \quad a \in [0, 1]. \tag{4.37}$$

If a suitable of a, we were to have $p \to J^\star_p(\delta_a)$ *constant* over $[0, 1]$, then the test δ_a would a performance insensitive to the value of p. This requirement (on a) is equivalent to the equality $J^\star_0(\delta_a) = J^\star_1(\delta_a)$, i.e.,

$$aJ_0(d_+) + (1 - a)J_0(d_-) = aJ_1(d_+) + (1 - a)J_1(d_-).$$

Thus,

$$a\left((J_0(d_+) - J_1(d_+)) + (J_1(d_-) - J_0(d_-))\right) = J_1(d_-) - J_0(d_-)$$

and solving for a we get

$$a^\star = \frac{J_1(d_-) - J_0(d_-)}{(J_0(d_+) - J_1(d_+)) + (J_1(d_-) - J_0(d_-))}.$$

It is a simple matter to check that $a^\star$ lies in $[0,)$ since $J_1(d_-) - J_0(d_-) \geq 0$ and $J_0(d_+) - J_1(d_+) > 0$ as discussed earlier.

It is now plain that

$$V(p_{\rm m}) = J_p^\star(\delta_{a^\star}) = \max_{p\in[0,1]} J_p^\star(\delta_{a^\star}), \quad p \in [0,1]. \tag{4.38}$$

Therefore, as in the discussion for Case 3, we have

$$\begin{aligned} \inf_{\delta\in\mathcal{D}^\star}\left(\max_{p\in[0,1]} J_p^\star(\delta)\right) &\leq \max_{p\in[0,1]} J_p^\star(\delta_{a^\star}) \\ &= V(p_{\rm m}) \\ &= \inf_{\delta\in\mathcal{D}^\star} J_p^\star(\delta) \\ &\leq \max_{p\in[0,1]}\left(\inf_{\delta\in\mathcal{D}^\star} J_p^\star(\delta)\right) \end{aligned} \tag{4.39}$$

and the desired inequality (4.31) is established.

The minimax equality now holds in mixed strategies, whence

$$\max_{p\in[0,1]} J_p^\star(\delta_{a^\star}) = \inf_{\delta\in\mathcal{D}^\star}\left(\max_{p\in[0,1]} J_p^\star(\delta)\right)$$

by virtue of (4.39). The minimax strategy is a randomized strategy $\delta_{\rm m}^\star$ which is identified as $\delta^\star(p_{\rm m})$. Note that $\delta^\star(p_{\rm m})$ is also a (randomized) Bayesian policy for the least favorable prior. ■

We summarize these findings in the following corollary to Theorem 4.5.1.

Corollary 4.6.1 *Assume $\Gamma_h > 0$ for all $h = 0, 1$. Under Condition (**A.1**), the minimax equality (4.29) holds in randomized strategies. Moreover, the minimax strategy always exists and can be interpreted as a (possibly randomized) Bayesian test under the least favorable prior $p_{\rm m}$.*

4.7 The minimax equation

The discussion of Section 4.5 shows that finding minimax tests passes through the evaluation of the value function $p \to V(p)$ and its maximizing set $I_{\rm m}$. As

simple examples already suggest in later sections, this evaluation may not always be straightforward to carry. Moreover, once an expression for the value function becomes available, finding its maximizers may turn out to be rather cumbersome. However, this two-step approach can be bypassed when $I_{\rm m}$ contains an interior point $p_{\rm m}$ at which the value function is differentiable, in which case the minimax test is given by the Bayesian test $d^\star(p_{\rm m})$. Instead a simple characterization of $p_{\rm m}$ is achieved through the so-called *Minimax Equation*.

Lemma 4.7.1 *Assume $\Gamma_h > 0$ for all $h = 0, 1$, and that $p_{\rm m}$ is an element of $(0, 1)$ and $p \to V(p)$ is differentiable at $p = p_{\rm m}$. Under Condition (**A.1**), $p_{\rm m}$ can be characterized through the Minimax Equation*

$$C(1,1) - C(0,0) = \Gamma_0 \cdot P_F(d^\star(p_{\rm m})) - \Gamma_1 \cdot P_M(d^\star(p_{\rm m})). \tag{4.40}$$

For the probability of error criterion, the Minimax Equation takes the simpler form

$$P_F(d^\star(p_{\rm m})) = P_M(d^\star(p_{\rm m})). \tag{4.41}$$

Proof. Fix p in $[0, 1]$. Upon specializing (4.7) to the test $d^\star(p)$, we get

$$\begin{aligned} J_\alpha(d^\star(p)) &= \alpha C(1,1) + (1-\alpha)C(0,0) \\ &\quad +\Gamma_0(1-\alpha) \cdot P_F(d^\star(p)) + \Gamma_1 \alpha \cdot P_M(d^\star(p)) \end{aligned} \tag{4.42}$$

with α in $[0, 1]$ and the mapping $\alpha \to J_\alpha(d^\star(p))$ is therefore affine in the variable α on the interval $[0, 1]$ Therefore, the graph of the mapping $\alpha \to J_\alpha(d^\star(p))$ is a straight line; its slope is given by

$$\frac{d}{d\alpha} J_\alpha(d^\star(p)) = C(1,1) - C(0,0) + \Gamma_1 \cdot P_M(d^\star(p)) - \Gamma_0 \cdot P_F(d^\star(p)). \tag{4.43}$$

By its definition, the Bayesian cost satisfies

$$V(\alpha) \leq J_\alpha(d), \qquad \begin{matrix} d \in \mathcal{D} \\ \alpha \in [0, \end{matrix}$$

with strict inequality for most tests. With $d = d^\star(p)$ this inequality becomes an equality when $\alpha = p$, namely

$$V(p) = J_p(d^\star(p))$$

while

$$V(\alpha) \leq J_\alpha(d^\star(p)), \quad \alpha \in [0,1].$$

With p in $(0,1)$, if the concave mapping $\alpha \to V(\alpha)$ is *differentiable* at $\alpha = p$, then the straight line $\alpha \to J_\alpha(d^\star(p))$ will be a tangent to the mapping $\alpha \to V(\alpha)$ at $\alpha = p$ – This is a consequence of the concavity established in Lemma 4.2.1. Thus,

$$\frac{d}{d\alpha}V(\alpha)\Big|_{\alpha=p} = \frac{d}{d\alpha}J_\alpha(d^\star(p))\Big|_{\alpha=p}. \tag{4.44}$$

In particular, if $p_{\rm m}$ is an element of $(0,1)$ and the mapping $\alpha \to V(\alpha)$ is differentiable at $\alpha = p_{\rm m}$, then

$$\frac{d}{d\alpha}V(\alpha)\Big|_{\alpha=p_{\rm m}} = \frac{d}{d\alpha}J_\alpha(d^\star(p_{\rm m}))\Big|_{\alpha=p_{\rm m}}. \tag{4.45}$$

But the interior point $p_{\rm m}$ being a maximum for the function $\alpha \to V(\alpha)$, we must have

$$\frac{d}{d\alpha}V(\alpha)\Big|_{\alpha=p_{\rm m}} = 0,$$

whence

$$\frac{d}{d\alpha}J_\alpha(d^\star(p_{\rm m}))\Big|_{\alpha=p_{\rm m}} = 0.$$

The equation (4.40) now follows from (4.43). ■

Obviously this analysis does not cover the cases when (i) $p_{\rm m} = 0$, (ii) $p_{\rm m} = 1$ and (iii) $p_{\rm m}$ is an element of $(0,1)$ but the mapping $\alpha \to V(\alpha)$ is not differentiable at $\alpha = p_{\rm m}$.

4.8 The Gaussian Case

The setting is that of Section 2.4 to which we refer the reader for the notation. As shown there, for every $\eta > 0$ we have

$$P_F(Lrt_\eta) = 1 - \Phi\left(\frac{\log\eta + \frac{1}{2}d^2}{d}\right)$$

and

$$P_M(Lrt_\eta) = \Phi\left(\frac{\log\eta - \frac{1}{2}d^2}{d}\right).$$

For each p in $(0, 1]$, with

$$\eta(p) = \frac{1-p}{p} \cdot \frac{\Gamma_0}{\Gamma_1},$$

we have $d^\star(p) = Lrt_{\eta(p)}$ and the expression (4.7) yields

$$\begin{aligned} V(p) &= J_p(d^\star(p)) \\ &= pC(1,1) + (1-p)C(0,0) \\ &\quad + \Gamma_0(1-p) \cdot \left(1 - \Phi\left(\frac{\log \eta(p) + \frac{1}{2}d^2}{d}\right)\right) \\ &\quad + \Gamma_1 p \cdot \Phi\left(\frac{\log \eta(p) - \frac{1}{2}d^2}{d}\right). \end{aligned} \tag{4.46}$$

The boundary cases $p = 0$ is easily recovered upon formally substituting this value in the expression (4.46). The Minimax Equation (4.40) takes the form

$$\begin{aligned} &C(1,1) - C(0,0) \\ &= \Gamma_1 \Phi\left(\frac{\log \eta(p_{\rm m}) - \frac{1}{2}d^2}{d}\right) - \Gamma_0\left(1 - \Phi\left(\frac{\log \eta(p_{\rm m}) + \frac{1}{2}d^2}{d}\right)\right). \end{aligned} \tag{4.47}$$

Probability of error – Simplifications occur since $C(0,0) = C(1,1) = 0$ and $\Gamma_0 = \Gamma_1 = 1$: The expression (4.46) becomes

$$V(p) = (1-p) \cdot \left(1 - \Phi\left(\frac{\frac{1-p}{p} + \frac{1}{2}d^2}{d}\right)\right) + p \cdot \Phi\left(\frac{\log \frac{1-p}{p} - \frac{1}{2}d^2}{d}\right),$$

and the Minimax Equation (4.47) reduces to

$$\Phi\left(\frac{\log \eta(p_{\rm m}) - \frac{1}{2}d^2}{d}\right) + \Phi\left(\frac{\log \eta(p_{\rm m}) + \frac{1}{2}d^2}{d}\right) = 1.$$

It is easy to see that this requires $\log \eta(p_{\rm m}) = 0$ so that $p_{\rm m} = \frac{1}{2}$ (indeed in $(0,1)$), an intuitively satisfying conclusion! Moreover, the minimax test is given by $d^\star_{\rm m} = d^{(\frac{1}{2})}$.

4.9 The Bernoulli case

The setting is that of Section 2.5 to which we refer the reader for the notation. We discuss only the case $a_1 < a_0$, and leave the case $a_0 < a_1$ as an exercise for the interested reader.

Note that the condition $a_1 < a_0$ is equivalent to $1 < \frac{1-a_1}{1-a_0}$, so that the expressions (2.30) and (2.31) for the probabilities $P_F(d_\eta)$ and $P_M(d_\eta)$, respectively, are *piecewise* constant functions of η with different constant values on the intervals $(0, \frac{a_1}{a_0}]$, $(\frac{a_1}{a_0}, \frac{1-a_1}{1-a_0}]$ and $(\frac{1-a_1}{1-a_0}, \infty)$: Direct inspection of the expression (2.30) yields

$$P_F(d_\eta) = \begin{cases} 1 & \text{if } 0 < \eta \leq \frac{a_1}{a_0} \\ 1 - a_0 & \text{if } \frac{a_1}{a_0} < \eta \leq \frac{1-a_1}{1-a_0} \\ 0 & \text{if } \frac{1-a_1}{1-a_0} < \eta. \end{cases} \tag{4.48}$$

Similarly, using (2.31) we find

$$P_M(d_\eta) = \begin{cases} 0 & \text{if } 0 < \eta \leq \frac{a_1}{a_0} \\ a_1 & \text{if } \frac{a_1}{a_0} < \eta \leq \frac{1-a_1}{1-a_0} \\ 1 & \text{if } \frac{1-a_1}{1-a_0} < \eta. \end{cases} \tag{4.49}$$

Thus, for each p in $[0, 1]$, we see from (4.7) that the cost $J_p(d_\eta)$ takes a different value on each of the intervals $(0, \frac{a_1}{a_0}]$, $(\frac{a_1}{a_0}, \frac{1-a_1}{1-a_0}]$ and $(\frac{1-a_1}{1-a_0}, \infty)$: Specifically, we have:

On $(0, \frac{a_1}{a_0}]$,

$$\begin{aligned} J_p(d_\eta) &= pC(1,1) + (1-p)C(0,0) + \Gamma_0(1-p) \\ &= pC(1,1) + (1-p)C(0,1). \end{aligned} \tag{4.50}$$

On $(\frac{a_1}{a_0}, \frac{1-a_1}{1-a_0}]$,

$$\begin{aligned} &J_p(d_\eta) \\ &= pC(1,1) + (1-p)C(0,0) + \Gamma_0(1-p) \cdot (1-a_0) + \Gamma_1 p \cdot a_1 \\ &= pC(1,1) + (1-p)C(0,1) + \Gamma_1 p \cdot a_1 - \Gamma_0(1-p) \cdot a_0 \\ &= p\left(C(1,1) + \Gamma_1 a_1\right) + (1-p)\left(C(0,1) - \Gamma_0 a_0\right). \end{aligned} \tag{4.51}$$

On $(\frac{1-a_1}{1-a_0}, \infty)$,

$$\begin{aligned} J_p(d_\eta) &= pC(1,1) + (1-p)C(0,0) + \Gamma_1 p \\ &= pC(1,0) + (1-p)C(0,0). \end{aligned} \tag{4.52}$$

Recall that

$$V(p) = J_p(d_{\eta(p)}) \text{ with } \eta(p) = \tfrac{\Gamma_0(1-p)}{\Gamma_1 p}, \quad 0 < p \leq 1.$$

As the mapping $p : (0,1] \to \mathbb{R}_+ : p \to \eta(p)$ is strictly decreasing, each of the equations

$$\eta(p) = \frac{1-a_1}{1-a_0}, \quad 0 < p \leq 1$$

and

$$\eta(p) = \frac{a_1}{a_0}, \quad 0 < p \leq 1$$

has a unique solution in $(0,1)$. These solutions, denoted p_- and p_+, respectively, are given by

$$p_- = \frac{\Gamma_0(1-a_0)}{\Gamma_1(1-a_1) + \Gamma_0(1-a_0)}$$

and

$$p_+ = \frac{\Gamma_0 a_0}{\Gamma_1 a_1 + \Gamma_0 a_0}.$$

As expected $p_- < \frac{1}{2} < p_+$.

Earlier expressions can now be used, and yield

$$V(p) = \begin{cases} pC(1,0) + (1-p)C(0,0) & \text{if } p \in (0, p_-] \\ p\left(C(1,1) + \Gamma_1 a_1\right) + (1-p)\left(C(0,1) - \Gamma_0 a_0\right) & \text{if } p \in (p_-, p_+] \\ pC(1,1) + (1-p)C(0,1) & \text{if } p \in (p_+, 1). \end{cases}$$

It is plain that the function $V : [0,1] \to \mathbb{R}$ is piecewise linear with three distinct segments, namely $(0, p_-]$, $(p_-, p_+]$ and $(p_+, 1]$. There are two kinks at $p = p_-$ and $p = p_+$, respectively. That the function is concave can be seen by computing the left and right-derivatives at these points. The function $V : [0,1] \to \mathbb{R}$ is differentiable everywhere except at these kinks. However the maximum occurs at one of these points so that $p_{\mathrm{m}} \in \{p_-, p_+\}$.

Probability of error – In that case we find that

$$V(p) = \begin{cases} p & \text{if } p \in (0, p_-] \\ pa_1 + (1-p)(1-a_0) & \text{if } p \in (p_-, p_+] \\ 1-p & \text{if } p \in (p_+, 1) \end{cases} \tag{4.53}$$

with

$$p_- = \frac{1-a_0}{(1-a_1)+(1-a_0)}$$

and

$$p_+ = \frac{a_0}{a_1+a_0}.$$

It is a simple matter to check that $V(p_\pm -) = V(p_\pm +)$, establishing continuity at the kink points. As we compare $V(p_-)$ and $V(p_+)$, we readily conclude that $p_{\rm m} = p_-$ (resp. $p_{\rm m} = p_+$) iff $1 - p_+ < p_-$ (resp. $p_- < 1 - p_+$) iff $a_0 + a_1 < 1$ (resp. $1 < a_0 + a_1$). The minimax cost is then given by

$$V_{\rm m} = \begin{cases} p_- = \frac{1-a_0}{(1-a_1)+(1-a_0)} & \text{if } a_0 + a_1 < 1 \\ 1 - p_+ = \frac{a_1}{a_1+a_0} & \text{if } 1 < a_0 + a_1 \end{cases}$$

Minimax strategy is necessarily randomized and is given by

$$\delta_a = a d_+ + (1-a) d_-$$

with the pure tests $d_-, d_+ : \mathbb{R} \to \{0, 1\}$ given by

4.10 A proof of Lemma 4.2.1

The proof proceeds in several stages. We start with the fact that

$$V(p) = \inf_{d \in \mathcal{D}} J_p(d), \quad p \in [0, 1].$$

Values at the boundary points – Consider a test d in $\mathcal{D}$. With $p = 0$ and $p = 1$ in (4.7) we get

$$J_0(d) = C(0,0) + \Gamma_0 P_F(d)$$

and

$$J_1(d) = C(1,1) + \Gamma_1 P_M(d).$$

Using the conditions $\Gamma_0 > 0$ and $\Gamma_1 > 0$, we conclude that

$$V(0) = \inf_{d\in\mathcal{D}} J_0(d) = C(0,0) + \Gamma_0 \cdot \inf_{d\in\mathcal{D}} P_F(d)$$

and

$$V(1) = \inf_{d\in\mathcal{D}} J_0(d) = C(1,1) + \Gamma_1 \cdot \inf_{d\in\mathcal{D}} P_M(d).$$

However, $P_F(d_F) = 0$ for the test $d_F : \mathbb{R}^k \to \{0,1\}$ which always selects the null hypothesis ($H = 0$) while $P_M(d_M) = 0$ for the test $d_M : \mathbb{R}^k \to \{0,1\}$ which always selects the alternative ($H = 1$). It follows that $\inf_{d\in\mathcal{D}} P_F(d) = 0$ and $\inf_{d\in\mathcal{D}} P_M(d) = 0$, whence $V(0) = C(0,0)$ and $V(1) = C(1,1)$. ■

Concavity on $[0,1]$ and continuity on $(0,1)$ – Once the test d is selected, the probabilities $P_F(d)$ and $P_M(d)$ appearing in (4.7) do *not* depend on p, and are determined only through F_0 and F_1. Thus, the mapping $p \to J_p(d)$ is affine, hence concave in p. As a result, the mapping $V : [0,1] \to \mathbb{R}$ is concave on the closed interval $[0,1]$, being the infimum of the family $\{J_p(d),\ d \in \mathcal{D}\}$ of concave functions. Because a concave function defined on an open interval is necessarily continuous on that open interval, the mapping $V : [0,1] \to \mathbb{R}$ is continuous on $(0,1)$ by virtue of Fact 9.4.2. ■

Continuity at the boundary points – We now turn to showing that the mapping $V : [0,1] \to \mathbb{R}$ is also continuous at the boundary points $p = 0$ and $p = 1$. We discuss only the case $p = 0$; the case $p = 1$ can be handled *mutatis mutandis* and is left to the interested reader as an exercise.

For notational convenience here and below we write

$$\Delta(p) \equiv \inf_{d\in\mathcal{D}} \left(\Gamma_0(1-p) \cdot P_F(d) + \Gamma_1 p \cdot P_M(d)\right), \quad p \in (0,1].$$

Recall that $V(0) = C(0,0)$ by the first part of the proof. Thus, for each p in $(0,1]$ we get from the definition of $V(p)$ that

$$V(p) - V(0) \;=\; p\,(C(1,1) - C(0,0)) + \Delta(p) \tag{4.54}$$

by virtue of (4.7). The continuity of the mapping $V : [0,1] \to \mathbb{R}$ at $p = 0$ is therefore equivalent to

$$\lim_{p \to 0} \Delta(p) = 0. \tag{4.55}$$

For any fixed p in $(0,1]$, the conditions $\Gamma_0 > 0$ and $\Gamma_1 > 0$ yield the inequalities

$$0 \leq \Delta(p) \leq \Gamma_1 p \tag{4.56}$$

since under the test d_F (introduced earlier in the proof) we have $P_F(d_F) = 0$ and $P_M(d_F) = 1$. The conclusion (4.55) is now immediate. ■

Differentiability – The existence *and* finiteness of the right-derivative and left-derivative on the open interval $(0,1)$ are simple consequences of Fact 9.4.4. The same argument also shows that the right-derivative (resp. left-derivative) does exist at $p = 0$ (resp. $p = 1$); however it may not necessarily be finite.

Instead, we provide a direct argument to show the existence and initeness of the right-derivative (resp. left-derivative) at $p = 0$ (resp. $p = 1$). We carry out the discussion only for $p = 0$ as the case $p = 1$ is similar: For each p in $(0,1]$, we note that

$$\frac{V(p) - V(0)}{p} \;=\; C(1,1) - C(0,0) + \frac{\Delta(p)}{p} \tag{4.57}$$

with

$$\frac{\Delta(p)}{p} = \inf_{d \in \mathcal{D}} \left(\Gamma_0 \left(\frac{1}{p} - 1 \right) \cdot P_F(d) + \Gamma_1 \cdot P_M(d) \right).$$

This last expression shows that $p \to \frac{\Delta(p)}{p}$ is decreasing on $(0,1]$, whence the limit $\lim_{p \downarrow 0} \frac{\Delta(p)}{p}$ always exists. This limit is finite by virtue of the bounds

$$0 \leq \frac{\Delta(p)}{p} \leq \Gamma_1, \quad p \in (0,1]$$

which are inherited from the earlier bounds (4.56). This shows the existence of a finite right-derivative at $p = 0$. ■

4.11 Exercises

4.1.

Let I denote an interval of $\mathbb{R}$, not necessarily finite, closed or open, and let A be an arbitrary index set. For each α in A, let $f_\alpha : I \to \mathbb{R}$ be a concave function. With the function $g : I \to \mathbb{R}$ defined by

$$g(x) = \inf \left(f_\alpha(x) : \ \alpha \in A \right), \quad x \in I$$

show that the mapping $g : I \to \mathbb{R}$ is concave.

4.2.

With $h > 0$ show that the equation

$$\Phi(x - h) + \Phi(x + h) = 1, \quad x \in \mathbb{R}$$

has a unique solution $x = 0$.

4.3.

4.12 References

Chapter 5

The Neyman-Pearson formulation

In many situations, not only is the prior probability p not available but it is quite difficult to make meaningful cost assignments. This is typically the case in radar applications – After all, what is the real cost of failing to detect an incoming missile? While it is tempting to seek to minimize *both* the probabilities of miss and false alarm, these are (usually) conflicting objectives and a *constrained* optimization problem is considered instead. The Neyman-Pearson formulation of the binary hypothesis problem given next constitutes an approach to handle such situations.

5.1 A constrained optimization problem

Fix α in $(0, 1)$ (the limiting case $\alpha = 0$ being of little practical interest). Let $\mathcal{D}_\alpha$ denote the collection of admissible tests in $\mathcal{D}$ of *size* at most α, namely

$$\mathcal{D}_\alpha = \{d \in \mathcal{D} : \ P_{\mathrm{F}}(d) \leq \alpha\}.$$

The Neyman-Pearson formulation is based on solving the constrained optimization problem NP_α where

$$\mathrm{NP}_\alpha : \quad \text{Maximize } P_{\mathrm{D}}(d) \text{ over } d \text{ in } \mathcal{D}_\alpha.$$

Solving NP_α amounts to finding a test $d_{\mathrm{NP}}(\alpha)$ in $\mathcal{D}_\alpha$ with the property that

$$P_{\mathrm{D}}(d) \leq P_{\mathrm{D}}(d_{\mathrm{NP}}(\alpha)), \quad d \in \mathcal{D}_\alpha.$$

Such a test $d_{\mathrm{NP}}(\alpha)$, when it exists, is called a *Neyman–Pearson* test of size α, or

alternatively, an α-level Neyman–Pearson decision rule. Such decision rules may not be unique. Following the accepted terminology, its *power* $\beta(\alpha)$ is given by

$$\beta(\alpha) \equiv P_{\mathrm{D}}(d_{\mathrm{NP}}(\alpha)) = \sup_{d \in \mathcal{D}_\alpha} P_{\mathrm{D}}(d).$$

When reformulated as

$$\mathrm{NP}_\alpha : \quad \text{Minimize } P_{\mathrm{M}}(d) \text{ over } d \text{ in } \mathcal{D}_\alpha,$$

the constrained optimization problem NP_α can be solved by standard Lagrangian arguments which are outlined in the next section. Throughout we assume that Condition (**A.1**) holds.

5.2 The Lagrangian arguments

Fix α in $(0, 1)$. For each $\lambda \geq 0$ consider the Lagrangian functional $J_\lambda : \mathcal{D} \to \mathbb{R}$ given by

$$J_\lambda(d) = P_{\mathrm{M}}(d) + \lambda \left(P_{\mathrm{F}}(d) - \alpha \right), \quad d \in \mathcal{D}.$$

The *Lagrangian* problem LP_λ is now defined as the *unconstrained* minimization problem

$$\mathrm{LP}_\lambda : \quad \text{Minimize } J_\lambda(d) \text{ over } d \text{ in } \mathcal{D}.$$

Solving LP_λ amounts to finding a test $d^\star_\lambda$ in $\mathcal{D}$ such that

$$J_\lambda(d^\star_\lambda) \leq J_\lambda(d), \quad d \in \mathcal{D}.$$

Solving the Lagrangian problem LP_λ Fix $\lambda > 0$. For any test d in $\mathcal{D}$, we note that

$$\begin{aligned} J_\lambda(d) &= \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 1\right] + \lambda \left(\mathbb{P}\left[d(\boldsymbol{Y}) = 1 | H = 0\right] - \alpha \right) \\ &= \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 1\right] + \lambda \left(1 - \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 0\right] - \alpha \right) \\ &= \lambda(1 - \alpha) + \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 1\right] - \lambda \mathbb{P}\left[d(\boldsymbol{Y}) = 0 | H = 0\right] \\ &= \lambda(1 - \alpha) + \int_{C(d)} h_\lambda(\boldsymbol{y}) dF(\boldsymbol{y}) \end{aligned} \tag{5.1}$$

with $h_\lambda : \mathbb{R}^k \to \mathbb{R}$ given by

$$h_\lambda(\boldsymbol{y}) = f_1(\boldsymbol{y}) - \lambda f_0(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k.$$

By the comparison arguments used in the proof of Theorem 2.2.1, the Lagrangian problem LP_λ is easily seen to be solved by the test $d^\star_\lambda : \mathbb{R}^k \to \{0,1\}$ given by

$$d^\star_\lambda(\boldsymbol{y}) = 0 \quad \text{iff} \quad h_\lambda(\boldsymbol{y}) < 0, \tag{5.2}$$

or equivalently,

$$d^\star_\lambda(\boldsymbol{y}) = 0 \quad \text{iff} \quad f_1(\boldsymbol{y}) < \lambda f_0(\boldsymbol{y}). \tag{5.3}$$

Note that in the notation associated with the definition (**??**) we have $d^\star_\lambda = d_\lambda$. ■

Meeting the constraint The next step consists in finding some value $\lambda(\alpha) > 0$ of the Lagrangian multiplier such that the test $d_{\lambda(\alpha)}$ meets the constraint, i.e.,

$$P_{\mathrm{F}}(d_{\lambda(\alpha)}) = \alpha. \tag{5.4}$$

If such value $\lambda(\alpha)$ were to exist, then the optimality $d_{\lambda(\alpha)}$ implies

$$J_{\lambda(\alpha)}(d_{\lambda(\alpha)}) \leq J_{\lambda(\alpha)}(d), \quad d \in \mathcal{D},$$

or equivalently,

$$P_{\mathrm{M}}(d_{\lambda(\alpha)}) \leq P_{\mathrm{M}}(d) + \lambda(\alpha)\left(P_{\mathrm{F}}(d) - \alpha\right), \quad d \in \mathcal{D}.$$

Consequently, for every test d *in* $\mathcal{D}_\alpha$ (and not merely in $\mathcal{D}$), it follows that

$$P_{\mathrm{M}}(d_{\lambda(\alpha)}) \leq P_M(d)$$

since then $P_{\mathrm{M}}(d) \leq \alpha$. The test $d_{\lambda(\alpha)}$ is a test in $\mathcal{D}_\alpha$ by virtue of (5.4), hence it solves NP_α – In other words, $d_{\mathrm{NP}}(\alpha)$ can be taken to be $d_{\lambda(\alpha)}$. ■

A difficulty The Lagrangian argument hinges upon the possibility of finding a value $\lambda(\alpha)$ of the Lagrange multiplier such that $P_{\mathrm{F}}(d_{\lambda(\alpha)}) = \alpha$. Unfortunately, this may not be always possible, unless additional assumptions are imposed. To see how this may indeed happen, note that

$$\begin{aligned} P_{\mathrm{F}}(d_\lambda) &= \mathbb{P}\left[d_\lambda(\boldsymbol{Y}) = 1 | H = 0\right] \\ &= \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \lambda f_0(\boldsymbol{Y}) | H = 0\right], \quad \lambda > 0. \end{aligned} \tag{5.5}$$

The mapping $\mathbb{R}_+ \to [0,1] : \lambda \to P_{\rm F}(d_\lambda)$ is clearly *monotone non-increasing*. However, the constraint $P_{\rm F}(d_\lambda) = \alpha$ may fail to hold for some α in $(0,1]$ because the set of values $\{P_{\rm F}(d_\lambda),\ \lambda \geq 0\}$ need not contain α. This will occur if the mapping $\lambda \to P_{\rm F}(d_\lambda)$ is not continuous at some point, say $\lambda^\star > 0$, with

$$\lim_{\lambda \uparrow \lambda^\star} P_{\rm F}(d_\lambda) < \alpha < \lim_{\lambda \downarrow \lambda^\star} P_{\rm F}(d_\lambda).$$

In Section **??** we illustrate such situations on simple examples that involve discrete rvs. Randomized policies are introduced to solve this difficulty. There are however situations where this can be avoided because each one of the problems NP_α (properly defined over randomized strategies) has a solution within the set of non-randomized policies $\mathcal{D}$.

5.3 The Neyman-Pearson Lemma

The discussion of Section 5.2 suggests the need to consider an extended version of the Neyman-Pearson formulation where randomized strategies are allowed.

Fix α in $(0,1]$. Let $\mathcal{D}^\star_\alpha$ denote the collection of all randomized tests in $\mathcal{D}^\star$ of *size* at most α, namely

$$\mathcal{D}^\star_\alpha = \{\delta \in \mathcal{D}^\star :\ P_{\rm F}(\delta) \leq \alpha\}.$$

The constrained optimization problem $\mathrm{NP}^\star$ is now replaced by the following constrained optimization problem $\mathrm{NP}^\star_\alpha$ where

$$\mathrm{NP}^\star_\alpha : \quad \text{Maximize } P_{\rm D}(\delta) \text{ over } \delta \text{ in } \mathcal{D}^\star_\alpha.$$

Solving $\mathrm{NP}^\star_\alpha$ amounts to finding a test $\delta_{\rm NP}(\alpha)$ in $\mathcal{D}^\star_\alpha$ with the property that

$$P_{\rm D}(\delta) \leq P_{\rm D}(\delta_{\rm NP}(\alpha)), \quad \delta \in \mathcal{D}^\star_\alpha.$$

Such a test $\delta_{\rm NP}(\alpha)$, when it exists, is also called a *Neyman–Pearson* test of size α, or alternatively, an α-level Neyman–Pearson decision rule. It may not be unique.

The existence of the Neyman–Pearson test $\delta_{\rm NP}(\alpha)$ of size α, its characterization and uniqueness are discussed below through three separate lemmas, known collectively as the Neyman-Pearson Lemma. Proofs are delayed until Section 5.4.

First a definition: With $\eta \geq 0$ and Borel mapping $\gamma : \mathbb{R}^k \to [0,1]$ (to be selected shortly), define the randomized test $\delta^\star : \mathbb{R}^k \to [0,1]$ in $\mathcal{D}^\star$ given by

$$\delta^\star(\boldsymbol{y}) = \begin{cases} 1 & \text{if } \eta f_0(\boldsymbol{y}) < f_1(\boldsymbol{y}) \\ \gamma(\boldsymbol{y}) & \text{if } f_1(\boldsymbol{y}) = \eta f_0(\boldsymbol{y}) \\ 0 & \text{if } f_1(\boldsymbol{y}) < \eta f_0(\boldsymbol{y}). \end{cases} \tag{5.6}$$

The inequality discussed next lays the groundwork for identifying the Neyman–Pearson test $\delta_{\mathrm{NP}}(\alpha)$.

Lemma 5.3.1 *For any test $\delta : \mathbb{R}^k \to [0,1]$ in $\mathcal{D}^\star$, the inequality*

$$P_{\mathrm{D}}(\delta^\star) - P_{\mathrm{D}}(\delta) \geq \eta \left(P_{\mathrm{F}}(\delta^\star) - P_{\mathrm{F}}(\delta) \right) \tag{5.7}$$

holds where the randomized test $\delta^\star : \mathbb{R}^k \to [0,1]$ in $\mathcal{D}^\star$ is given by (5.7).

If we select $\eta \geq 0$ and $\gamma : \mathbb{R}^k \to [0,1]$ so that $\delta^\star$ satisfies the equality

$$P_{\mathrm{F}}(\delta^\star) = \alpha, \tag{5.8}$$

then the inequality (5.7) reads

$$P_{\mathrm{D}}(\delta^\star) - P_{\mathrm{D}}(\delta) \geq \eta \left(\alpha - P_{\mathrm{F}}(\delta) \right), \quad \delta \in \mathcal{D}^\star. \tag{5.9}$$

For any test $\delta : \mathbb{R}^k \to [0,1]$ in $\mathcal{D}^\star_\alpha$, we then conclude that

$$P_{\mathrm{D}}(\delta^\star) - P_{\mathrm{D}}(\delta) \geq \eta \left(\alpha - P_{\mathrm{F}}(\delta) \right) \geq 0 \tag{5.10}$$

since $P_{\mathrm{F}}(\delta) \leq \alpha$. In other words,

$$P_{\mathrm{D}}(\delta) \leq P_{\mathrm{D}}(\delta^\star), \quad \delta \in \mathcal{D}^\star_\alpha$$

and the test $\delta^\star$ solves the constrained problem $\mathrm{NP}^\star_\alpha$.

We now show that the parameter $\eta \geq 0$ and the Borel mapping $\gamma : \mathbb{R}^k \to [0,1]$ can indeed be selected so that a test $\delta^\star$ of the form (5.7) indeed satisfies (5.8).

Lemma 5.3.2 *For every α in $(0,1]$ it is always possible to select $\eta \geq 0$ and a Borel mapping $\gamma : \mathbb{R}^k \to [0,1]$ in (5.6) so that (5.8) holds.*

Finally uniqueness is shown to hold in the following sense.

Lemma 5.3.3 *For every α in $(0,1]$, if $\delta_{\mathrm{NP}}(\alpha)$ is a Neyman–Pearson test (possibly in $\mathcal{D}^\star$) of size α, then it necessarily holds that*

$$\mathbb{P}\left[\delta_{\mathrm{NP}}(\alpha)(\boldsymbol{Y}) = \delta^\star(\boldsymbol{Y}) | H = h \right] = 1, \quad h = 0, 1 \tag{5.11}$$

where the test $\delta^\star$ is given by (5.6) with $\eta \geq 0$ and Borel mapping $\gamma : \mathbb{R}^k \to [0,1]$ selected so that (5.8) holds.

5.4 Proofs

Throughout the discussion α is given in $(0, 1]$ and held fixed.

A proof of Lemma 5.3.1 Let $\delta : \mathbb{R}^k \to [0, 1]$ be an arbitrary test in $\mathcal{D}^\star$. As discussed in Section 3.3 recall that

$$P_{\mathrm{F}}(\delta) = \int_{\mathbb{R}^k} \delta(\boldsymbol{y}) f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \quad \text{and} \quad P_{\mathrm{F}}(\delta^\star) = \int_{\mathbb{R}^k} \delta^\star(\boldsymbol{y}) f_0(\boldsymbol{y}) dF(\boldsymbol{y}),$$

while

$$P_{\mathrm{D}}(\delta) = \int_{\mathbb{R}^k} \delta(\boldsymbol{y}) f_1(\boldsymbol{y}) dF(\boldsymbol{y}) \quad \text{and} \quad P_{\mathrm{D}}(\delta^\star) = \int_{\mathbb{R}^k} \delta^\star(\boldsymbol{y}) f_1(\boldsymbol{y}) dF(\boldsymbol{y}).$$

It follows that

$$\begin{aligned} & P_{\mathrm{D}}(\delta^\star) - P_{\mathrm{D}}(\delta) - \eta \left(P_{\mathrm{F}}(\delta^\star) - P_{\mathrm{F}}(\delta) \right) \\ &= \int_{\mathbb{R}^k} \left(\delta^\star(\boldsymbol{y}) - \delta(\boldsymbol{y}) \right) f_1(\boldsymbol{y}) dF(\boldsymbol{y}) - \eta \int_{\mathbb{R}^k} \left(\delta^\star(\boldsymbol{y}) - \delta(\boldsymbol{y}) \right) f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \\ &= \int_{\mathbb{R}^k} \left(\delta^\star(\boldsymbol{y}) - \delta(\boldsymbol{y}) \right) \left(f_1(\boldsymbol{y}) - \eta f_0(\boldsymbol{y}) \right) dF(\boldsymbol{y}) \\ &= \int_{\mathbb{R}^k} P_\eta(\boldsymbol{y}) dF(\boldsymbol{y}) \end{aligned} \tag{5.12}$$

where we have set

$$P_\eta(\boldsymbol{y}) \equiv \left(\delta^\star(\boldsymbol{y}) - \delta(\boldsymbol{y}) \right) \left(f_1(\boldsymbol{y}) - \eta f_0(\boldsymbol{y}) \right), \qquad \boldsymbol{y} \in \mathbb{R}^k.$$

Direct inspection shows that we always have

$$P_\eta(\boldsymbol{y}) \geq 0, \quad \boldsymbol{y} \in \mathbb{R}^k. \tag{5.13}$$

Obviously, we have $P_\eta(\boldsymbol{y}) = 0$ when $f_1(\boldsymbol{y}) = \eta f_0(\boldsymbol{y})$. When $\eta f_0(\boldsymbol{y}) < f_1(\boldsymbol{y})$, then

$$P_\eta(\boldsymbol{y}) = \left(1 - \delta(\boldsymbol{y}) \right) \left(f_1(\boldsymbol{y}) - \eta f_0(\boldsymbol{y}) \right) \geq 0,$$

while when $f_1(\boldsymbol{y}) < \eta f_0(\boldsymbol{y})$, then

$$P_\eta(\boldsymbol{y}) = -\delta(\boldsymbol{y}) \left(f_1(\boldsymbol{y}) - \eta f_0(\boldsymbol{y}) \right) \geq 0.$$

It is now plain from (5.12) and (5.13) that

$$P_{\mathrm{D}}(\delta^\star) - P_{\mathrm{D}}(\delta) - \eta \left(P_{\mathrm{F}}(\delta^\star) - P_{\mathrm{F}}(\delta) \right) \geq 0 \tag{5.14}$$

and the inequality (5.7) follows. ■

A proof of Lemma 5.3.2 Using the definition (5.6) of the randomized test $\delta^\star$, we note that

$$\begin{aligned} & P_{\mathrm{F}}(\delta^\star) \\ = & \int_{\mathbb{R}^k} \delta^\star(\boldsymbol{y}) f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \\ = & \int_{\left\{\boldsymbol{y}\in\mathbb{R}^k : f_1(\boldsymbol{y})=\eta f_0(\boldsymbol{y})\right\}} \gamma(\boldsymbol{y}) f_0(\boldsymbol{y}) dF(\boldsymbol{y}) + \int_{\left\{\boldsymbol{y}\in\mathbb{R}^k : f_1(\boldsymbol{y})>\eta f_0(\boldsymbol{y})\right\}} f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \\ = & \int_{\left\{\boldsymbol{y}\in\mathbb{R}^k : f_1(\boldsymbol{y})=\eta f_0(\boldsymbol{y})\right\}} \gamma(\boldsymbol{y}) f_0(\boldsymbol{y}) dF(\boldsymbol{y}) + \mathbb{P}\left[f_1(\boldsymbol{Y}) > \eta f_0(\boldsymbol{Y}) | H=0\right]. \end{aligned}$$

As we seek to satisfy (5.8), we need to select $\eta \geq 0$ and a Borel mapping $\gamma : \mathbb{R}^k \to [0,1]$ such that

$$\alpha - \mathbb{P}\left[f_1(\boldsymbol{Y}) > \eta f_0(\boldsymbol{Y}) | H=0\right] = \int_{\left\{\boldsymbol{y}\in\mathbb{R}^k : f_1(\boldsymbol{y})=\eta f_0(\boldsymbol{y})\right\}} \gamma(\boldsymbol{y}) f_0(\boldsymbol{y}) dF(\boldsymbol{y}).$$

This last relation suggests introducing the quantity $\eta(\alpha)$ defined by

$$\eta(\alpha) = \inf\left\{\eta \geq 0 : \ \mathbb{P}\left[f_1(\boldsymbol{Y}) > \eta f_0(\boldsymbol{Y}) | H=0\right] < \alpha\right\}.$$

The definition of $\eta(\alpha)$ is well posed since $\eta \to \mathbb{P}\left[f_1(\boldsymbol{Y}) > \eta f_0(\boldsymbol{Y}) | H=0\right]$ is non-increasing (and right-continuous) on $(0,\infty)$.

Two cases are possible: If

$$\mathbb{P}\left[f_1(\boldsymbol{Y}) > \eta(\alpha) f_0(\boldsymbol{Y}) | H=0\right] < \alpha,$$

then take $\gamma : \mathbb{R}^k \to [0,1]$ to be constant, say

$$\gamma(\boldsymbol{y}) = \gamma(\alpha), \quad \boldsymbol{y} \in \mathbb{R}^k.$$

In that case, the constant $\gamma(\alpha)$ satisfies

$$\alpha - \mathbb{P}\left[f_1(\boldsymbol{Y}) > \eta(\alpha) f_0(\boldsymbol{Y}) | H=0\right] = \gamma(\alpha) \int_{\left\{\boldsymbol{y}\in\mathbb{R}^k : f_1(\boldsymbol{y})=\eta(\alpha) f_0(\boldsymbol{y})\right\}} f_0(\boldsymbol{y}) dF(\boldsymbol{y}),$$

whence

$$\gamma(\alpha) = \frac{\alpha - \mathbb{P}\left[f_1(\boldsymbol{Y}) > \eta(\alpha) f_0(\boldsymbol{Y}) | H=0\right]}{\mathbb{P}\left[f_1(\boldsymbol{Y}) = \eta(\alpha) f_0(\boldsymbol{Y}) | H=0\right]}. \tag{5.15}$$

If

$$\mathbb{P}\left[f_1(\boldsymbol{Y}) > \eta(\alpha) f_0(\boldsymbol{Y}) | H = 0\right] = \alpha,$$

then the mapping $\gamma : \mathbb{R}^k \to [0,1]$ must selected so that

$$\int_{\left\{\boldsymbol{y}\in\mathbb{R}^k : f_1(\boldsymbol{y}) = \eta f_0(\boldsymbol{y})\right\}} \gamma(\boldsymbol{y}) f_0(\boldsymbol{y}) dF(\boldsymbol{y}) = 0.$$

This can be achieved by taking the constant mapping given by

$$\gamma(\boldsymbol{y}) = 0, \quad \boldsymbol{y} \in \mathbb{R}^k.$$

■

A proof of Lemma 5.3.3 The test $\delta_{\mathrm{NP}}(\alpha)$ being a Neyman–Pearson test of size α, the equality

$$P_{\mathrm{D}}(\delta_{\mathrm{NP}}(\alpha)) = P_{\mathrm{D}}(\delta^\star)$$

must hold where the test $\delta^\star$ is given by (5.6) with $\eta > 0$ and Borel mapping $\gamma : \mathbb{R}^k \to [0,1]$ selected so that (5.8) holds. This a consequence of the fact that both $\delta_{\mathrm{NP}}(\alpha)$ and $\delta^\star$ solve the problem $\mathrm{NP}^\star_\alpha$.

It then follows from (5.7) that

$$0 = P_{\mathrm{D}}(\delta^\star) - P_{\mathrm{D}}(\delta_{\mathrm{NP}}(\alpha)) \geq \eta\left(\alpha - P_{\mathrm{F}}(\delta_{\mathrm{NP}}(\alpha))\right) \geq 0 \tag{5.16}$$

since $P_{\mathrm{F}}(\delta^\star) = \alpha$ under the choice of $\eta > 0$ and the Borel mapping $\gamma : \mathbb{R}^k \to [0,1]$, whence $P_{\mathrm{F}}(\delta_{\mathrm{NP}}(\alpha)) = \alpha$.

In other words, $P_{\mathrm{D}}(\delta_{\mathrm{NP}}(\alpha)) = P_{\mathrm{D}}(\delta^\star)$ and $P_{\mathrm{F}}(\delta_{\mathrm{NP}}(\alpha)) = P_{\mathrm{F}}(\delta^\star)$. Using these facts in the expression (5.12) (with the strategy $\delta_{\mathrm{NP}}(\alpha)$) we find that

$$\begin{aligned} 0 &= P_{\mathrm{D}}(\delta^\star) - P_{\mathrm{D}}(\delta_{\mathrm{NP}}(\alpha)) - \eta\left(P_{\mathrm{F}}(\delta^\star) - P_{\mathrm{F}}(\delta_{\mathrm{NP}}(\alpha))\right) \\ &= \int_{\mathbb{R}^k} \left(\delta^\star(\boldsymbol{y}) - \delta_{\mathrm{NP}}(\alpha)(\boldsymbol{y})\right)\left(f_1(\boldsymbol{y}) - \eta f_0(\boldsymbol{y})\right) dF(\boldsymbol{y}) \end{aligned} \tag{5.17}$$

with

$$\left(\delta^\star(\boldsymbol{y}) - \delta_{\mathrm{NP}}(\alpha)(\boldsymbol{y})\right)\left(f_1(\boldsymbol{y}) - \eta f_0(\boldsymbol{y})\right) \geq 0, \quad \boldsymbol{y} \in \mathbb{R}^k$$

by virtue of (5.13). It immediately follows that

$$\left(\delta^\star(\boldsymbol{y}) - \delta_{\mathrm{NP}}(\alpha)(\boldsymbol{y})\right)\left(f_1(\boldsymbol{y}) - \eta f_0(\boldsymbol{y})\right) = 0 \quad F - a.e. \tag{5.18}$$

on $\mathbb{R}^k$. Therefore,

$$\delta_{\rm NP}(\alpha)(\boldsymbol{y}) = \delta^\star(\boldsymbol{y}) \quad F - a.e. \tag{5.19}$$

on $\{\boldsymbol{y} \in \mathbb{R}^k : \ f_1(\boldsymbol{y}) \neq \eta f_0(\boldsymbol{y})\}$. ■

5.5 Examples

The Gaussian case Consider again the situation discussed in Section 2.6 where the observation rv $\boldsymbol{Y}$ is conditionally Gaussian given H, i.e.,

$$\begin{array}{ll} H_1: & \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{m}_1, \boldsymbol{R}) \\ H_0: & \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{m}_0, \boldsymbol{R}) \end{array}$$

where $\boldsymbol{m}_1$ and $\boldsymbol{m}_0$ are distinct elements in $\mathbb{R}^k$, and the $k \times k$ *symmetric* matrix $\boldsymbol{R}$ is positive definite (thus invertible). From the discussion given in Section 2.6, it follows for each $\lambda > 0$ the test d_λ takes the form

$$d_\lambda(\boldsymbol{y}) = 0 \quad \text{iff} \quad \boldsymbol{y}'\boldsymbol{R}^{-1}\Delta\boldsymbol{m} > \phi(\lambda)$$

with $\Delta\boldsymbol{m}$ and $\phi(\lambda)$ given by (2.22) and (2.23), respectively. We also have

$$P_{\rm F}(d_\lambda) = 1 - \Phi\left(\frac{\log\lambda + \frac{1}{2}d^2}{d}\right).$$

where d^2 is given by (2.28) – It is plain that the function $\lambda \to P_{\rm F}(d_\lambda)$ is continuous on $\mathbb{R}_+$ with $\{P_{\rm F}(d_\lambda),\ \lambda > 0\} = (0,1)$. Given α in the unit interval $(0,1)$, the value $\lambda(\alpha)$ is *uniquely* determined through the relation

$$1 - \alpha = \Phi\left(\frac{\log\lambda + \frac{1}{2}d^2}{d}\right).$$

This is equivalent to

$$\lambda(\alpha) = e^{d\cdot x_{1-\alpha} - \frac{1}{2}d^2}.$$

where for t in $(0,1)$, let x_t denote the only solution to the equation

$$\Phi(x) = t, \quad x \in \mathbb{R}.$$

Discontinuity with Bernoulli rvs The setting is that of Section 2.5 to which we refer the reader for the notation. We discuss only the case $a_1 < a_0$, and leave the case $a_0 < a_1$ as an exercise for the interested reader. We have shown that

$$P_{\mathrm{F}}(d_\lambda) = \begin{cases} 1 & \text{if } 0 < \lambda \leq \frac{a_1}{a_0} \\ 1 - a_0 & \text{if } \frac{a_1}{a_0} < \lambda \leq \frac{1-a_1}{1-a_0} \\ 0 & \text{if } \frac{1-a_1}{1-a_0} < \lambda \end{cases} \tag{5.20}$$

as λ ranges over $(0, \infty)$.

Note that $\lambda \to P_{\mathrm{F}}(d_\lambda)$ is left-continuous but not continuous with

$$\{P_{\mathrm{F}}(d_\lambda), \quad \lambda > 0\} = \{0, 1 - a_0, 1\}\,.$$

Discontinuity with Poisson rvs With $\mathcal{P}(m)$ denoting the Poisson pmf on $\mathbb{N}$ with parameter $m > 0$, consider the following simple binary hypothesis testing problem

$$\begin{array}{ll} H_1: & Y \sim \mathcal{P}(m_1) \\ H_0: & Y \sim \mathcal{P}(m_0) \end{array}$$

where $m_1 \neq m_0$ in $(0, \infty)$, Thus,

$$\mathbb{P}\left[Y = k | H = h\right] = \frac{(m_h)^k}{k!} e^{-m_h}, \qquad \begin{array}{c} h = 0, 1 \\ k = 0, 1, \ldots \end{array}$$

In this example, we take F to be the counting measure on $\mathbb{N}$, and for every $\lambda \geq 0$, the definition of d_λ reduces to

$$\begin{aligned} d_\lambda(k) = 0 \quad \text{iff} \quad & \frac{(m_1)^k}{k!} e^{-m_1} < \lambda \frac{(m_0)^k}{k!} e^{-m_0} \\ \text{iff} \quad & \left(\frac{m_1}{m_0}\right)^k < \lambda e^{-(m_0 - m_1)} \end{aligned} \tag{5.21}$$

with $k = 0, 1, \ldots$.

If $m_0 < m_1$, then

$$\begin{aligned} d_\lambda(k) = 0 \quad \text{iff} \quad & \frac{(m_1)^k}{k!} e^{-m_1} < \lambda \frac{(m_0)^k}{k!} e^{-m_0} \\ \text{iff} \quad & \left(\frac{m_1}{m_0}\right)^k < \lambda e^{-(m_0 - m_1)} \\ \text{iff} \quad & k < \eta(\lambda) \end{aligned} \tag{5.22}$$

with $k = 0, 1, \ldots$, where

$$\eta(\lambda) = \frac{\log \lambda e^{-(m_0 - m_1)}}{\log\left(\frac{m_1}{m_0}\right)}.$$

It follows that

$$\begin{aligned} P_{\mathrm{F}}(d_\lambda) &= \mathbb{P}\left[d_\lambda(Y) = 1 | H = 0\right] \\ &= \mathbb{P}\left[Y \geq \eta(\lambda) | H = 0\right] \\ &= \sum_{k=0:\ \eta(\lambda) \leq k}^{\infty} \frac{(m_0)^k}{k!} e^{-m_0}. \end{aligned} \tag{5.23}$$

In this last expression only the integer ceiling $\lceil \eta(\lambda) \rceil$ of $\eta(\lambda)$ matters, where $\lceil \eta(\lambda) \rceil = \inf\{k \in \mathbb{N} : \eta(\lambda) \leq k\}$, whence

$$P_{\mathrm{F}}(d_\lambda) = \sum_{k=\lceil \eta(\lambda) \rceil}^{\infty} \frac{(m_0)^k}{k!} e^{-m_0}.$$

As a result, the mapping $\lambda \to P_{\mathrm{F}}(d_\lambda)$ is easily seen to be a *left-continuous piecewise constant* mapping with

$$P_{\mathrm{F}}(d_\lambda) = P_{\mathrm{F}}(d_{\lambda_n}), \qquad \begin{array}{l} \lambda_n < \lambda \leq \lambda_{n+1} \\ n = 0, 1, \ldots \end{array}$$

where $\{\lambda_n,\ n = 1, 2, \ldots\}$ is a strictly monotone increasing sequence determined by the relation

$$n = \frac{\log \lambda_n e^{-(m_0 - m_1)}}{\log\left(\frac{m_1}{m_0}\right)}. \quad n = 1, 2, \ldots$$

or equivalently,

$$\lambda_n = \left(\frac{m_1}{m_0}\right)^n e^{-(m_1 - m_0)}, \quad n = 1, 2, \ldots$$

It is now plain that whenever α is chosen in $[0, 1]$ such that

for some integer $n = 0, 1, \ldots$ then the requirement that $P_{\mathrm{F}}(d_{\lambda(\alpha)}) = \alpha$ cannot be met. This difficulty is circumvented by enlarging $\mathcal{D}$ with *randomized* policies; see Section **??**.

5.6 Exercises

5.7 References

Chapter 6

The receiver operating characteristics

In this chapter we investigate various properties of the mappings $\eta \to P_{\rm F}(d_\eta)$ and $\eta \to P_{\rm F}(d_\eta)$ as η ranges over $\mathbb{R}_+$. This leads to defining the *receiver operating characteristic* curve, and to developing it into a handy operational tool to solve the various versions of the binary hypothesis problem discussed so far.

6.1 A basic limiting result

We start with a basic observation.

Lemma 6.1.1 *Assume Condition* **(A.1)** *to hold. For each* $h = 0, 1$*, the mapping*

$$\mathbb{R}_+ \to [0,1] : \eta \to \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}) | H = h\right]$$

is monotone non-decreasing. Left (resp. right) limits exist at all points on $(0, \infty)$ *(resp.* $[0, \infty)$*), with*

$$\begin{aligned} & \lim_{\eta \downarrow \lambda} \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}) | H = h\right] \qquad (6.1) \\ = \;& \mathbb{P}\left[f_1(\boldsymbol{Y}) > \lambda f_0(\boldsymbol{Y}), f_0(\boldsymbol{Y}) > 0 | H = h\right] + \mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = h\right], \quad \lambda \geq 0 \end{aligned}$$

and

$$\begin{aligned} & \lim_{\eta \uparrow \lambda} \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}) | H = h\right] \\ = \;& \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \lambda f_0(\boldsymbol{Y}) | H = h\right], \quad \lambda > 0. \qquad (6.2) \end{aligned}$$

For future reference, for each $\eta \geq 0$ define the Borel subset $R(\eta)$ of $\mathbb{R}^k$ by

$$R(\eta) \equiv \left\{ \boldsymbol{y} \in \mathbb{R}^k : \ f_1(\boldsymbol{y}) \geq \eta f_0(\boldsymbol{y}) \right\}. \tag{6.3}$$

Note that

$$\mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}) | H = h \right] = \mathbb{P}\left[\boldsymbol{Y} \in R(\eta) | H = h \right], \quad h = 0, 1, . \tag{6.4}$$

Proof. The asserted monotonicity property is a consequence of the inclusion $R(\eta_2) \subseteq R(\eta_1)$ holding whenever $\eta_1 < \eta_2$. The existence of left (resp. right) limits at all points on $(0, \infty)$ (resp. $[0, \infty)$) immediately follows.

Consider $h = 0, 1$. Fix $\lambda \geq 0$. By standard continuity facts from measure theory it follows that

$$\begin{aligned} \lim_{\eta \downarrow \lambda} \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}) | H = h \right] &= \lim_{\eta \downarrow \lambda} \mathbb{P}\left[\boldsymbol{Y} \in R(\eta) | H = h \right] \\ &= \mathbb{P}\left[\cup_{\lambda < \eta} \left[\boldsymbol{Y} \in R(\eta) \right] | H = h \right] \\ &= \mathbb{P}\left[\boldsymbol{Y} \in \cup_{\eta > \lambda} R(\eta) | H = h \right], \end{aligned}$$

and we obtain (6.1) as we note that

$$\cup_{\eta > \lambda} R(\eta) = \left\{ \boldsymbol{y} \in \mathbb{R}^k : \begin{array}{c} f_1(\boldsymbol{y}) > \lambda f_0(\boldsymbol{y}), \\ f_0(\boldsymbol{y}) > 0 \end{array} \right\} \cup \left\{ \boldsymbol{y} \in \mathbb{R}^k : \ f_0(\boldsymbol{y}) = 0 \right\}.$$

In a very similar way, with $\lambda > 0$ we get

$$\begin{aligned} \lim_{\eta \uparrow \lambda} \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}) | H = h \right] &= \lim_{\eta \uparrow \lambda} \mathbb{P}\left[\boldsymbol{Y} \in R(\eta) | H = h \right] \\ &= \mathbb{P}\left[\cap_{\eta < \lambda} \left[\boldsymbol{Y} \in R(\eta) \right] | H = h \right] \\ &= \mathbb{P}\left[\boldsymbol{Y} \in \cap_{\eta < \lambda} R(\eta) | H = h \right] \\ &= \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \lambda f_0(\boldsymbol{Y}) | H = h \right] \end{aligned}$$

because

$$\cap_{\eta < \lambda} R(\eta) = \left\{ \boldsymbol{y} \in \mathbb{R}^k : \ f_1(\boldsymbol{y}) \geq \lambda f_0(\boldsymbol{y}) \right\} = R(\lambda).$$

This establishes (6.2). ■

6.2 Continuity properties

We shall specialize Lemma 6.1.1 for $h = 0$ and $h = 1$.

Lemma 6.2.1 *The mapping $\eta \to P_{\mathrm{F}}(d_\eta)$ is monotone non-decreasing on $\mathbb{R}_+$, with the left (resp. right) limit existing at all points in $(0, \infty)$ (resp. $[0, \infty)$). Under Conditions* **(A.1)**, *it holds that In particular, it holds that*

$$\lim_{\eta \uparrow \lambda} P_{\mathrm{F}}(d_\eta) = \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \lambda f_0(\boldsymbol{Y}) | H = 0\right] = P_{\mathrm{F}}(d_\lambda), \quad \lambda > 0 \tag{6.5}$$

while

$$\lim_{\eta \downarrow \lambda} P_{\mathrm{F}}(d_\eta) = \mathbb{P}\left[f_1(\boldsymbol{Y}) > \lambda f_0(\boldsymbol{Y}) | H = 0\right], \quad \lambda \geq 0. \tag{6.6}$$

Proof. Applying (6.1) and (6.2) with $h = 0$ readily yields the desired conclusions as we recall that

$$\mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = 0\right] = 0 \tag{6.7}$$

under Condition **(A.1)**. ■

Thus, the mapping $\eta \to P_{\mathrm{F}}(d_\eta)$ is left-continuous on $(0, \infty)$, but not necessarily right-continuous on $[0, \infty)$ as we note that

$$\lim_{\eta \downarrow \lambda} P_{\mathrm{F}}(d_\eta) = P_{\mathrm{F}}(d_\lambda) - \mathbb{P}\left[f_1(\boldsymbol{Y}) = \lambda f_0(\boldsymbol{Y}) | H = 0\right] \geq 0. \tag{6.8}$$

Lemma 6.2.2 *The mapping $\eta \to P_{\mathrm{D}}(d_\eta)$ is monotone non-decreasing on $\mathbb{R}_+$, with the left (resp. right) limit existing at all points on $(0, \infty)$ (resp. $[0, \infty)$). Under Conditions* **(A.1)** *and* **(A.2)**, *it holds that*

$$\lim_{\eta \uparrow \lambda} P_{\mathrm{D}}(d_\eta) = \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \lambda f_0(\boldsymbol{Y}) | H = 1\right] = P_{\mathrm{D}}(d_\lambda), \quad \lambda > 0 \tag{6.9}$$

while

$$\lim_{\eta \downarrow \lambda} P_{\mathrm{D}}(d_\eta) = \mathbb{P}\left[f_1(\boldsymbol{Y}) > \lambda f_0(\boldsymbol{Y}) | H = 1\right], \quad \lambda \geq 0. \tag{6.10}$$

Proof. As discussed in Section **??**, under Conditions **(A.1)** and **(A.2)** it also holds that

$$\mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = 1\right] = 0. \tag{6.11}$$

Applying Lemma 6.1.1 with $h = 1$ and using this last fact we readily get the result. ■

Again, the mapping $\eta \to P_{\mathrm{D}}(d_\eta)$ is left-continuous on $(0, \infty)$ but not necessarily right-continuous on $[0, \infty)$ as we note that

$$\lim_{\eta \downarrow \lambda} P_{\mathrm{D}}(d_\eta) = P_{\mathrm{D}}(d_\lambda) - \mathbb{P}\left[f_1(\boldsymbol{Y}) = \lambda f_0(\boldsymbol{Y}) | H = 1\right] \geq 0. \tag{6.12}$$

We close with the behavior at $\eta = 0$ and at $\eta = \infty$; proofs are available in Section 6.8.

Lemma 6.2.3 *Under Condition* **(A.1)** *we have*

$$\lim_{\eta \to 0} P_{\mathrm{F}}(d_\eta) = \mathbb{P}\left[f_1(\boldsymbol{Y}) > 0 | H = 0\right], \tag{6.13}$$

and if Condition **(A.2)** *also holds, then*

$$\lim_{\eta \to 0} P_{\mathrm{D}}(d_\eta) = 1. \tag{6.14}$$

In principle, it is possible that $\mathbb{P}\left[f_1(\boldsymbol{Y}) > 0 | H = 0\right] < 1$. However, with the notation (**??**) introduced in Section **??**, we see that (6.13) becomes

$$\lim_{\eta \to 0} P_{\mathrm{F}}(d_\eta) = 1 \tag{6.15}$$

if $B_0 \subseteq B_1$ since then $\mathbb{P}\left[f_1(\boldsymbol{Y}) > 0 | H = 0\right] = 1$ (or equivalently, (6.7)) under Condition **(A.1)**.

Note that $B_0 \subseteq B_1$ implies $B_0 = B_1$ under Conditions **(A.1)** and **(A.2)**. In that case we have "continuity" at the origin because $P_{\mathrm{F}}(d_0) = 1$ and $P_{\mathrm{D}}(d_0) = 1$ (under the convention used for d_0 in Section **??**).

Lemma 6.2.4 *Under Condition* **(A.1)** *we always have*

$$\lim_{\eta\to\infty} P_{\mathrm{F}}(d_\eta) = 0, \tag{6.16}$$

and if Condition **(A.2)** *also holds, then*

$$\lim_{\eta\to\infty} P_{\mathrm{D}}(d_\eta) = 0. \tag{6.17}$$

Note that $P_{\mathrm{F}}(d_\infty) = 0$ and $P_{\mathrm{D}}(d_\infty) = 0$ (under the convention used for d_∞ introduced Section **??**), implying "continuity" at infinity.

6.3 The receiver operating characteristic (ROC)

A careful inspection of the solutions to the three formulations discussed so far shows that sometimes under mild assumptions, the test of interest takes the form

$$d_\eta(\boldsymbol{y}) = 0 \quad \text{iff} \quad f_1(\boldsymbol{y}) < \eta f_0(\boldsymbol{y}) \tag{6.18}$$

for some $\eta > 0$ – It is only the value of the threshold η that varies with the problem formulation. With the notation used earlier, we have

In the Bayesian formulation,

$$\eta_B = \frac{\Gamma_0(1-p)}{\Gamma_1 p}$$

In the minimax formulation,

$$\eta_m = \frac{\Gamma_0(1-p_{\mathrm{m}})}{\Gamma_1 p_{\mathrm{m}}}$$

with p_{m} such that

$$V(p_{\mathrm{m}}) = \max\left(V(p) : \ p \in [0,1]\right).$$

When p_{m} is a point in $(0,1)$ at which $V : [0,1] \to \mathbb{R}$ is differentiable, then p_{m} can be characterized through the Minimax Equation.

In the Neyman–Pearson formulation,

$$\eta_{\mathrm{NP}}(\alpha) = \lambda(\alpha).$$

with $\lambda(\alpha)$ satisfying the constraint (5.4).

In view of this, it seems natural to analyze in some details the performance of the tests (6.18). This is done by considering how their probabilities of detection and of false alarm, namely $P_{\rm F}(d_\eta)$ and $P_{\rm D}(d_\eta)$, vary in relation to each other as η ranges from $\eta = 0$ to $\eta = +\infty$. This is best understood by plotting the graph (Γ) of the detection probability against the corresponding probability of false alarm. Such a graph is analogous to a *phase portrait* for two-dimensional non-linear ODEs, and is called a *receiver operating characteristic* (ROC) curve. Its *parametric* representation is given by

$$\mathbb{R}_+ \to [0,1] \times [0,1] : \eta \to (P_{\rm F}(d_\eta), (P_{\rm D}(d_\eta)),$$

whence

$$(\Gamma): \quad \{(P_{\rm F}(d_\eta), (P_{\rm D}(d_\eta)),\ \eta \geq 0\}\,.$$

This graph is *completely* determined by the probability distributions F_0 and F_1 of the observation rv $\boldsymbol{Y}$ under the two hypotheses (through the densities f_0 and f_1 with respect to the underlying distribution F) and not by cost assignments or the prior probabilities. A typical ROC curve is drawn below.

6.4 Geometric properties of the ROC curve

The following geometric properties of the ROC curve are key to its operational usefulness.

Theorem 6.4.1 *Assume that Conditions* **(A.1)** *and* **(A.2)** *hold.*

(i): Both mappings $\mathbb{R}_+ \to [0,1] : \eta \to P_{\rm F}(d\eta)$ *and* $\mathbb{R}_+ \to [0,1] : \eta \to P_{\rm D}(d\eta)$ *are monotone non-increasing, with* $\lim_{\eta\to 0} P_{\rm F}(d_\eta) \equiv P_{\rm F}^\star \leq P_{\rm F}(d_0) = 1$ *and* $\lim_{\eta\to 0} P_{\rm D}(d_\eta) = P_{\rm D}(d_\infty) = 1$, *and* $\lim_{\eta\uparrow\infty} P_{\rm F}(d_\eta) = P_{\rm F}(d_\infty) = 0$ *and* $\lim_{\eta\uparrow\infty} P_{\rm D}(d_\eta) = P_{\rm D}(d_\infty) = 0$.

(ii): If the right-derivative of $\eta \to P_{\rm F}(d_\eta)$ *exists at* $\eta = \lambda$ *for some* $\lambda \geq 0$, *then the right-derivative of* $\eta \to P_{\rm D}(d_\eta)$ *also exists at* $\eta = \lambda$, *and the relation*

$$\frac{d^+}{d\eta} P_{\rm D}(d_\eta)\Big|_{\eta=\lambda} = \lambda \cdot \frac{d^+}{d\eta} P_{\rm F}(d_\eta)\Big|_{\eta=\lambda} \tag{6.19}$$

holds.

(iii): If the left-derivative of $\eta \to P_{\mathrm{F}}(d_\eta)$ exists at $\eta = \lambda$ for some $\lambda > 0$, then the left-derivative of $\eta \to P_{\mathrm{D}}(d_\eta)$ also exists at $\eta = \lambda$, and the relation

$$\frac{d^-}{d\eta} P_{\mathrm{D}}(d_\eta)\Big|_{\eta=\lambda} = \lambda \cdot \frac{d^-}{d\eta} P_{\mathrm{F}}(d_\eta)\Big|_{\eta=\lambda} \tag{6.20}$$

holds.

A proof of Theorem 6.4.1 is available in Section 6.7. It follows from this last result that whenever the mapping $\eta \to P_{\mathrm{F}}(d_\eta)$ is *strictly decreasing* and *differentiable*, then the mapping $\eta \to P_{\mathrm{D}}(d_\eta)$ is also *strictly decreasing* and *differentiable*, whence the curve (Γ) can be represented as the graph of a function $\Gamma : [0,1] \to [0,1] : P_{\mathrm{F}} \to P_{\mathrm{D}} = \Gamma(P_{\mathrm{F}})$, namely

$$P_{\mathrm{D}}(d_\eta) = \Gamma(P_{\mathrm{F}}(d_\eta)), \quad \eta \geq 0. \tag{6.21}$$

In such circumstance, Theorem 6.4.1 yields the the following information concerning this mapping. We consider only the case when $P_F^\star = 1$ in Theorem 6.4.1; the situation when $P_F^\star < 1$ can be handled in a similar way with details left to the interested reader.

Corollary 6.4.1 *Assume Conditions* **(A.1)** *and* **(A.2)** *to hold. Whenever the mapping $\mathbb{R}_+ \to [0,1] : \eta \to P_{\mathrm{F}}(d_\eta)$ is differentiable and strictly decreasing, so is the mapping $\mathbb{R}_+ \to [0,1] : \eta \to P_{\mathrm{D}}(d_\eta)$. In that case, there exists a mapping $\Gamma : [0,1] \to [0,1]$ satisfying (6.21) which is differentiable, strictly increasing and concave with*

$$\frac{d\Gamma}{dP_{\mathrm{F}}}\left(P_{\mathrm{F}}(d_\eta)\right) = \eta, \quad \eta \geq 0. \tag{6.22}$$

Proof. By Part (i) of Theorem 6.4.1 we see that the mapping $\mathbb{R}_+ \to [0,1] : \eta \to P_{\mathrm{F}}(d_\eta)$ maps $\mathbb{R}_+$ onto $(0,1]$. Being strictly decreasing and differentiable, this mapping admits an inverse, denoted $(0,1] \to \mathbb{R}_+ : P_{\mathrm{F}} \to \eta(P_{\mathrm{F}})$, with the property that

$$P_{\mathrm{F}}(d_{\eta(P_{\mathrm{F}})}) = P_{\mathrm{F}}, \quad P_{\mathrm{F}} \in (0,1].$$

By the Implicit Function Theorem this inverse mapping $P_{\mathrm{F}} \to \eta(P_{\mathrm{F}})$ is differentiable; it is also strictly decreasing.

Define the mapping $\Gamma : [0,1] \to [0,1]$ by setting

$$\Gamma(P_{\mathrm{F}}) \equiv P_{\mathrm{D}}(d_{\eta(P_{\mathrm{F}})}), \quad P_{\mathrm{F}} \in [0,1] \tag{6.23}$$

with the understanding that $\Gamma(1) = 1$.

By Theorem 6.4.1 the differentiability of the mapping $\eta \to P_{\mathrm{F}}(d_\eta)$ on $\mathbb{R}_+$ implies that of the mapping $\eta \to P_{\mathrm{D}}(d_\eta)$ on $\mathbb{R}_+$, and the mapping $P_{\mathrm{F}} \to \Gamma(P_{\mathrm{F}})$ is therefore also differentiable. It is simple matter to check that this mapping is strictly decreasing on $[0,1]$.

By the very definition of the function Γ, the identity (6.21) must hold. Differentiating both sides of (6.21) with respect of η we find

$$\begin{aligned} \frac{d}{d\eta} P_{\mathrm{D}}(d_\eta) &= \frac{d}{d\eta} \Gamma(P_{\mathrm{F}}(d_\eta)) \\ &= \frac{d\Gamma}{dP_{\mathrm{F}}}(P_{\mathrm{F}}(d_\eta)) \cdot \frac{d}{d\eta} P_{\mathrm{F}}(d_\eta) \end{aligned} \tag{6.24}$$

as we use the Chain Rule. But Theorem 6.4.1 implies also that

$$\frac{d}{d\eta} P_{\mathrm{D}}(d_\eta) = \eta \frac{d}{d\eta} P_{\mathrm{F}}(d_\eta).$$

Combining these facts we conclude that

$$\eta \frac{d}{d\eta} P_{\mathrm{F}}(d_\eta) = \frac{d\Gamma}{dP_{\mathrm{F}}}(P_{\mathrm{F}}(d_\eta)) \cdot \frac{d}{d\eta} P_{\mathrm{F}}(d_\eta).$$

The mapping $\mathbb{R}_+ \to [0,1] : \eta \to P_{\mathrm{F}}(d_\eta)$ being assumed differentiable and strictly decreasing, we have $\frac{d}{d\eta} P_{\mathrm{F}}(d_\eta) < 0$. Dividing by $\frac{d}{d\eta} P_{\mathrm{F}}(d_\eta)$ we get (6.22). The other properties follow readily. ■

6.5 Operating the ROC

These results are most useful for operationally using the ROC curve:

For the Neyman–Pearson test of size α, consider the point on the ROC curve with abscissa α. It is determined by the threshold value $\eta(\alpha)$ with the property that $P_{\mathrm{F}}(d_{\eta(\alpha)}) = \alpha$, and $d_{NP}(\alpha)$ is simply $d_{\eta(\alpha)}$. Note that $\eta(\alpha)$ is the ***slope*** of the tangent to the ROC curve at the point with abscissa α and the power $\beta(\alpha)$ of the test is simply the ordinate of that point.

For the Bayesian problem, η is determined by the cost assignment and the prior distribution of the rv H. The values of $P_{\rm D}(d_\eta)$ and $P_{\rm F}(d_\eta)$ can be easily determined by finding the point on the ROC where the tangent has slope η.

The Minimax Equation takes the form

$$C(1,1) - C(0,0) + \Gamma_1 P_{\rm M}(d_\eta) - \Gamma_0 P_{\rm F}(d_\eta) = 0,$$

or equivalently

$$C(1,1) - C(0,0) + \Gamma_1 = \Gamma_1 P_{\rm D}(d_\eta) + \Gamma_0 P_{\rm F}(d_\eta).$$

This shows that the minimax rule $d^\star_{\rm m}$ is obtained as follows. Consider the straight line (L) in the $(P_{\rm F}, P_{\rm D})$-plane with equation

$$(L) \quad C(1,1) - C(0,0) + \Gamma_1 = \Gamma_1 P_{\rm D} + \Gamma_0 P_{\rm F}.$$

Let $(P^\star_{\rm F}, P^\star_{\rm D})$ be the point of intersection of the straight line (L) wth the ROC curve (Γ), and let $\eta^\star$ be the corresponding threshold value, i.e., $P^\star_{\rm F} = P_{\rm F}(d_{\eta^\star})$ and $P^\star_{\rm D} = P_{\rm D}(d_{\eta^\star})$. It is now clear that $d^\star_{\rm m} = d_{\eta^\star}$.

6.6 Examples

Building on material developed earlier we now discuss the ROC in the Gaussian and Bernoulli cases, respectively.

The Gaussian case The setting is that of Section 2.4 to which we refer the reader for the notation. As shown there, for any $\eta > 0$ we have

$$P_{\rm F}(d_\eta) = 1 - \Phi\left(\frac{\log\eta + \frac{1}{2}d^2}{d}\right)$$

and

$$P_{\rm D}(d_\eta) = 1 - \Phi\left(\frac{\log\eta - \frac{1}{2}d^2}{d}\right).$$

To find the ROC curve, note that

$$d\Phi^{-1}\left(1 - P_{\rm F}(d_\eta)\right) - \frac{d^2}{2} = \log\eta,$$

while we must have

$$d\Phi^{-1}\left(1-P_{\mathrm{D}}(d_\eta)\right)+\frac{d^2}{2}=\log\eta,$$

whence

$$d\Phi^{-1}\left(1-P_{\mathrm{F}}(d_\eta)\right)-\frac{d^2}{2}=d\Phi^{-1}\left(1-P_{\mathrm{D}}(d_\eta)\right)+\frac{d^2}{2}.$$

It follows that

$$\Phi^{-1}\left(1-P_{\mathrm{D}}(d_\eta)\right)=\Phi^{-1}\left(1-P_{\mathrm{F}}(d_\eta)\right)-d$$

so that

$$1-P_{\mathrm{D}}(d_\eta)=\Phi\left(\Phi^{-1}\left(1-P_{\mathrm{F}}(d_\eta)\right)-d\right)$$

This shows that here the mapping $\Gamma:[0,1]\to[0,1]$ is well defined and given by

$$P_{\mathrm{D}}=1-\Phi\left(\Phi^{-1}\left(1-P_{\mathrm{F}}\right)-d\right),\quad P_{\mathrm{F}}\in[0,1].$$

The Bernoulli case The setting is that of Section 2.5 to which we refer the reader for the notation. We discuss only the case $a_1<a_0$, and leave the case $a_0<a_1$ as an exercise for the interested reader. It was shown that

$$P_{\mathrm{F}}(d_\eta) \;=\; \begin{cases} 1 & \text{if } 0<\eta\le\frac{a_1}{a_0} \\ 1-a_0 & \text{if } \frac{a_1}{a_0}<\eta\le\frac{1-a_1}{1-a_0} \\ 0 & \text{if } \frac{1-a_1}{1-a_0}<\eta \end{cases} \tag{6.25}$$

and

$$P_{\mathrm{D}}(d_\eta) \;=\; \begin{cases} 1 & \text{if } 0<\eta\le\frac{a_1}{a_0} \\ 1-a_1 & \text{if } \frac{a_1}{a_0}<\eta\le\frac{1-a_1}{1-a_0} \\ 0 & \text{if } \frac{1-a_1}{1-a_0}<\eta. \end{cases} \tag{6.26}$$

Therefore,

$$\{(P_{\mathrm{F}}(d_\eta),P_{\mathrm{D}}(d_\eta)),\quad \eta\ge 0\}=\{(0,0),(1-a_0,1-a_1),(1,1)\}.$$

Strictly speaking, in this case the ROC is not a "curve" as it comprises only three points. However, the points on the two segments $[(0,0),(1-a_0,1-a_1)]$ and $[(1-a_0,1-a_1),(1,1)]$ are achievable through randomization.

6.7 A proof of Theorem 6.4.1

We start with some facts that prove useful in discussing Theorem 6.4.1: With $\lambda > 0$, recall the set $R(\lambda)$ defined in Section 6.2 as

$$R(\lambda) \equiv \{\boldsymbol{y} \in \mathbb{R}^k : \ f_1(\boldsymbol{y}) \geq \lambda f_0(\boldsymbol{y})\}.$$

Noting that

$$d_\lambda(\boldsymbol{y}) = 1 \quad \text{iff} \quad \boldsymbol{y} \in R(\lambda),$$

it is plain that

$$\begin{aligned} P_{\mathrm{F}}(d_\lambda) &= \mathbb{P}\left[d_\lambda(\boldsymbol{Y}) = 1 | H = 0\right] \\ &= \mathbb{P}\left[\boldsymbol{Y} \in R(\lambda) | H = 0\right] = \int_{R(\lambda)} f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \end{aligned}$$

and

$$P_{\mathrm{D}}(d_\lambda) = \mathbb{P}\left[d_\lambda(\boldsymbol{Y}) = 1 | H = 1\right] = \int_{R(\lambda)} f_1(\boldsymbol{y}) dF(\boldsymbol{y}).$$

For each $\Delta\lambda > 0$, easy algebra now leads to

$$\begin{aligned} P_{\mathrm{F}}(d_{\lambda+\Delta\lambda}) - P_{\mathrm{F}}(d_\lambda) &= \int_{R(\lambda+\Delta\lambda)} f_0(\boldsymbol{y}) dF(\boldsymbol{y}) - \int_{R(\lambda)} f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \\ &= -\int_{R_+(\lambda;\Delta\lambda)} f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \end{aligned}$$

where

$$R_+(\lambda;\Delta\lambda) \equiv \{\boldsymbol{y} \in \mathbb{R}^k : \ \lambda f_0(\boldsymbol{y}) \leq f_1(\boldsymbol{y}) < (\lambda + \Delta\lambda) f_0(\boldsymbol{y})\}.$$

Similarly, we have

$$\begin{aligned} P_{\mathrm{F}}(d_{\lambda-\Delta\lambda}) - P_{\mathrm{F}}(d_\lambda) &= \int_{R(\lambda-\Delta\lambda)} f_0(\boldsymbol{y}) dF(\boldsymbol{y}) - \int_{R(\lambda)} f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \\ &= \int_{R_-(\lambda;\Delta\lambda)} f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \end{aligned}$$

where

$$R_-(\lambda;\Delta\lambda) \equiv \{\boldsymbol{y} \in \mathbb{R}^k : \ (\lambda - \Delta\lambda) f_0(\boldsymbol{y}) \leq f_1(\boldsymbol{y}) < \lambda f_0(\boldsymbol{y})\}.$$

We can now turn to the proof of Theorem 6.4.1: Part (i) is already covered in Section 6.2. We shall discuss only Part (ii) as Part (iii) can be established by similar arguments *mutatis mutandi*. This is left to the interested reader.

Fix $\eta \geq 0$. With $\Delta\eta > 0$, the very definition of $R_+(\eta; \Delta\eta)$ implies the inequalities

$$\eta \int_{R_+(\eta;\Delta\eta)} f_0(\boldsymbol{y}) dF(\boldsymbol{y}) \leq \int_{R_+(\eta;\Delta\eta)} f_1(\boldsymbol{y}) dF(\boldsymbol{y})$$

and

$$\int_{R+(\eta;\Delta\eta)} f_1(\boldsymbol{y}) dF(\boldsymbol{y}) \leq (\eta + \Delta\eta) \int_{R_+(\eta;\Delta\eta)} f_0(\boldsymbol{y}) dF(\boldsymbol{y}).$$

It then follows that

$$(\eta + \Delta\eta) \cdot \frac{P_{\mathrm{F}}(d_{\eta+\Delta\eta}) - P_{\mathrm{F}}(d_\eta)}{\Delta\eta} \leq \frac{P_{\mathrm{D}}(d_{\eta+\Delta\eta}) - P_{\mathrm{D}}(d_\eta)}{\Delta\eta}$$

and

$$\frac{P_{\mathrm{D}}(d_{\eta+\Delta\eta}) - P_{\mathrm{D}}(d_\eta)}{\Delta\eta} \leq \eta \cdot \frac{P_{\mathrm{F}}(d_{\eta+\Delta\eta}) - P_{\mathrm{F}}(d_\eta)}{\Delta\eta}.$$

If the right-derivative of $\eta \to P_{\mathrm{F}}(d_\eta)$ exists, then

$$\frac{d^+}{d\eta} P_{\mathrm{F}}(d_\eta) = \lim_{\Delta\eta \downarrow 0} \frac{P_{\mathrm{F}}(d_{\eta+\Delta\eta}) - P_{\mathrm{F}}(d_\eta)}{\Delta\eta}$$

and an easy sandwich argument shows that the limit

$$\lim_{\Delta\eta \downarrow 0} \frac{P_{\mathrm{D}}(d_{\eta+\Delta\eta}) - P_{\mathrm{D}}(d_\eta)}{\Delta\eta}$$

also exists. Therefore, the right-derivative of $\eta \to P_{\mathrm{D}}(d_\eta)$ also exists and is given by

$$\frac{d^+}{d\eta} P_{\mathrm{D}}(d_\eta) = \eta \cdot \frac{d^+}{d\eta} P_{\mathrm{F}}(d_\eta).$$

■

6.8 Proofs of Lemmas 6.2.3 and 6.2.4

A proof of Lemma 6.2.3 Applying (6.1) (with $\lambda = 0$) yields

$$\begin{aligned} & \lim_{\eta \downarrow 0} \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}) | H = h\right] \qquad (6.27) \\ = \; & \mathbb{P}\left[f_1(\boldsymbol{Y}) > 0, f_0(\boldsymbol{Y}) > 0 | H = h\right] + \mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = h\right], \quad h = 0, 1. \end{aligned}$$

Under Condition **(A.1)**, the validity of (6.7) implies (6.13) as we use (6.27) with $h = 0$. Under the additional Condition **(A.2)**, (6.11) and (6.27), this time with $h = 1$, lead to

$$\lim_{\eta \downarrow 0} \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}) | H = 1\right] = \mathbb{P}\left[f_1(\boldsymbol{Y}) > 0 | H = 1\right]$$

and (6.14) readily follows. ■

A proof of Lemma 6.2.4 Fix $h = 0, 1$. For each $\eta > 0$, we note that

$$\begin{aligned} & \mathbb{P}\left[d_\eta(\boldsymbol{Y}) = 1 | H = h\right] \\ = \; & \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}) | H = h\right] \\ = \; & \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}), f_0(\boldsymbol{Y}) > 0 | H = h\right] + \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq 0, f_0(\boldsymbol{Y}) = 0 | H = h\right] \\ = \; & \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}), f_0(\boldsymbol{Y}) > 0 | H = h\right] + \mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = h\right]. \end{aligned}$$

The usual monotonicity argument yields

$$\begin{aligned} & \lim_{\eta \to \infty} \mathbb{P}\left[d_\eta(\boldsymbol{Y}) = 1 | H = h\right] \\ = \; & \lim_{\eta \to \infty} \mathbb{P}\left[f_1(\boldsymbol{Y}) \geq \eta f_0(\boldsymbol{Y}), f_0(\boldsymbol{Y}) > 0 | H = h\right] + \mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = h\right] \\ = \; & \mathbb{P}\left[f_1(\boldsymbol{Y}) = \infty, f_0(\boldsymbol{Y}) > 0 | H = h\right] + \mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = h\right] \\ = \; & \mathbb{P}\left[f_0(\boldsymbol{Y}) = 0 | H = h\right]. \qquad (6.28) \end{aligned}$$

First use (6.28) with $h = 0$: Under Condition **(A.1)** the consequence (6.7) implies $\lim_{\eta \to \infty} \mathbb{P}\left[d_\eta(\boldsymbol{Y}) = 1 | H = 0\right] = 0$. With $h = 1$, under the additional Condition **(A.2)** we also get $\lim_{\eta \to \infty} \mathbb{P}\left[d_\eta(\boldsymbol{Y}) = 1 | H = 1\right] = 0$. This completes the proof of Lemma 6.2.4. ■

6.9 Exercises

6.10 References

Chapter 7

The M-ary hypothesis testing problem

As we shall see in this chapter and in the next one, the simple binary hypothesis testing problem of Chapter 7 admits several important generalizations. The version discussed here assumes that there are more than two hypotheses; it is of particular relevance to the design of optimal receivers in digital modulation. In this formulation, nature assumes M distinct states, labeled $0, 1, \ldots, M-1$, and is now encoded in a rv H which take values in the discrete set $\{0, 1, \ldots, M-1\}$. A decision has to be made as to which of these M hypotheses is the correct one on the basis of an observation $\boldsymbol{Y}$ which is statistically related to H.

7.1 Motivating examples

Digital communications

Manufacturing

7.2 The probabilistic model

To formulate the M-*ary* hypothesis testing problem we proceed very much as in Chapter 2: With positive integer $M \geq 2$, we are given M distinct probability distribution functions $F_0, \ldots, F_{M-1}$ on $\mathbb{R}^k$, and a pmf $\boldsymbol{p} = (p_0, \ldots, p_{M-1})$ on

$\{0, 1, \ldots, M-1\}$ with

$$0 \leq p_m \leq 1, \quad m = 0, 1, \ldots, M-1 \quad \text{and} \quad \sum_{m=0}^{M-1} p_m = 1.$$

The situation is summarized by

$$H_m : \boldsymbol{Y} \sim F_m, \quad m = 0, 1, \ldots, M-1. \tag{7.1}$$

We construct a sample space Ω equipped with a σ-field of events $\mathcal{F}$, and rvs H and $\boldsymbol{Y}$ defined on it and taking values in $\{0, 1, \ldots, M-1\}$ and $\mathbb{R}^k$, respectively. Now the probability distribution functions $F_0, \ldots, F_{M-1}$ have the interpretation that

$$F_m(\boldsymbol{y}) = \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y} | H = m\right], \qquad \begin{array}{c} \boldsymbol{y} \in \mathbb{R}^k, \\ m = 0, 1, \ldots, M-1. \end{array}$$

The probability distribution of the rv H is specified by the pmf $\boldsymbol{p} = (p_0, \ldots, p_{M-1})$ with

$$p_m = \mathbb{P}\left[H = m\right], \quad m = 0, 1, \ldots, M-1.$$

Again, the conditional probability distributions of the observations given the hypothesis *and* the probability distribution of H completely specify the *joint* distribution of the rvs H and $\boldsymbol{Y}$: Indeed, for each $m = 0, 1, \ldots, M-1$,

$$\begin{aligned} \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y}, H = m\right] &= \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y} | H = m\right] \mathbb{P}\left[H = m\right] \\ &= p_m F_m(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k. \end{aligned} \tag{7.2}$$

The *unconditional* probability distribution function of the rv $\boldsymbol{Y}$ is easily determined to be

$$\mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y}\right] = \sum_{m=0}^{M-1} \mathbb{P}\left[\boldsymbol{Y} \leq y, H = m\right], \quad \boldsymbol{y} \in \mathbb{R}^k$$

by the law of total probabilities, whence

$$\begin{aligned} G(\boldsymbol{y}) &\equiv \mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y}\right] \\ &= \sum_{m=0}^{M-1} p_m F_m(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k. \end{aligned} \tag{7.3}$$

During the discussion, several assumptions will be enforced on the probability distributions $F_0, \ldots, F_{M-1}$. The most common assumption is denoted by **(A.3)** for sake of convenience; it parallels Condition **(A.1)** made in the binary case:

Condition **(A.3)**: The probability distributions $F_0, \ldots, F_{M-1}$ on $\mathbb{R}^k$ are *absolutely continuous* with respect to some distribution F on $\mathbb{R}^k$.

This is equivalent to saying that for each $m = 0, 1, \ldots, M-1$, there exists a Borel mapping $f_m : \mathbb{R}^k \to \mathbb{R}_+$ such that

$$F_m(\boldsymbol{y}) = \int_{-\infty}^{\boldsymbol{y}} f_m(\boldsymbol{\eta}) dF(\boldsymbol{\eta}), \qquad \begin{array}{c} \boldsymbol{y} \in \mathbb{R}^k, \\ m = 0, 1, \ldots, M-1. \end{array} \tag{7.4}$$

We refer to these Borel mappings as probability density functions with respect to F.

This condition is hardly constraining since we can always take F to be the average of the M distributions $F_0, \ldots, F_{M-1}$. i.e.,

$$F(\boldsymbol{y}) \equiv \frac{1}{M} \left(\sum_{m=0}^{M-1} F_m(\boldsymbol{y}) \right), \quad \boldsymbol{y} \in \mathbb{R}^k \tag{7.5}$$

However, in most applications F is either Lebesgue measure on $\mathbb{R}^k$ or a counting measure on some countable subset of $\mathbb{R}^k$.

Under Condition **(A.3)**, the unconditional probability distribution function $G : \mathbb{R}^k \to [0, 1]$ of the rv $\boldsymbol{Y}$ is automatically absolutely continuous with respect to the distribution F on $\mathbb{R}^k$: Indeed, we see from (7.4) and (7.3) that

$$\begin{aligned} G(\boldsymbol{y}) &= \sum_{m=0}^{M-1} p_m \int_{-\infty}^{\boldsymbol{y}} f_m(\boldsymbol{\eta}) dF(\boldsymbol{\eta}) \\ &= \int_{-\infty}^{\boldsymbol{y}} \left(\sum_{m=0}^{M-1} p_m f_m(\boldsymbol{\eta}) \right) dF(\boldsymbol{\eta}) \\ &= \int_{-\infty}^{\boldsymbol{y}} g(\boldsymbol{\eta}) dF(\boldsymbol{\eta}). \quad \boldsymbol{y} \in \mathbb{R}^k \end{aligned} \tag{7.6}$$

with Borel mapping $g : \mathbb{R}^k \to \mathbb{R}_+$ given by

$$g(\boldsymbol{y}) = \sum_{m=0}^{M-1} p_m f_m(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k. \tag{7.7}$$

In other words, the unconditional probability distribution function $G : \mathbb{R}^k \to [0, 1]$ admits $g : \mathbb{R}^k \to \mathbb{R}_+$ as probability density function with respect to F.

7.3 Admissible tests

We begin with a formal definition of an admissible test in the context M-ary hypothesis testing.

An *admissible* decision rule or test is any *Borel* mapping $d : \mathbb{R}^k \to \{0, 1, \ldots, M-1\}$. The collection of all such admissible rules is still denoted by $\mathcal{D}$.

Again the measurability requirement is imposed to guarantee that the mapping $d(Y) : \Omega \to \{0, 1, \ldots, M-1\} : \omega \to d(Y(\omega))$ is indeed a rv, i.e., $\{\omega \in \Omega : d(Y(\omega)) = m\}$ is an event in $\mathcal{F}$ for all $m = 0, 1, \ldots, M-1$.

A collection $\{C_0, \ldots, C_{M-1}\}$of M subsets of $\mathbb{R}^k$ forms an *M-ary Borel partition* of $\mathbb{R}^k$ if

(i) For each $m = 0, 1, \ldots, M-1$, the set C_m is a Borel subset of $\mathbb{R}^k$;

(ii) We have

$$C_m \cap C_k = \emptyset, \qquad \begin{matrix} m \neq k \\ m, k = 0, 1, \ldots, M-1 \end{matrix}$$

and

(iii) The condition

$$\cup_{m=0}^{M-1} C_m = \mathbb{R}^k$$

holds.

The collection of all M-ary Borel partitions of $\mathbb{R}^k$ is denoted $\mathcal{P}_M(\mathbb{R}^k)$. Lemma 1.5.1 has the following analog in the context M-ary hypothesis testing.

Lemma 7.3.1 *The set $\mathcal{D}$ of admissible decision rules is in one-to-one correspondence with $\mathcal{P}_M(\mathbb{R}^k)$.*

Proof. For every test d in $\mathcal{D}$, the Borel sets defined by

$$C_m(d) \equiv \{\boldsymbol{y} \in \mathbb{R}^k : \ d(\boldsymbol{y}) = m\}, \quad m = 0, 1, \ldots, M-1$$

are determined by d, and obviously form an M-ary Borel partition of $\mathbb{R}^k$.

Conversely, consider an M-ary Borel measurable partition $\{C_0, ..., C_{M-1}\}$ in $\mathcal{P}_M(\mathbb{R}^k)$. With this partition we can associate the mapping $d_{C_0,\ldots,C_{M-1}} : \mathbb{R}^k \to \{0, \ldots, M-1\}$ given by

$$d_{C_0,\ldots,C_{M-1}}(\boldsymbol{y}) = m, \qquad \begin{matrix} \boldsymbol{y} \in C_m, \\ m = 0, \ldots, M-1. \end{matrix}$$

By construction this mapping is an admissible test in $\mathcal{D}$ as we note that

$$C_m(d_{C_0,\ldots,C_{M-1}}) = C_m, \quad m = 0, \ldots, M-1$$

by the fact that the collection $\{C_0, ..., C_{M-1}\}$ is a partition of $\mathbb{R}^k$. ■

7.4 The Bayesian formulation

The probabilistic model The Bayesian formulation assumes *knowledge* of the prior distribution $\boldsymbol{p} = (p_0, \ldots, p_{M-1})$ of the rv H, and of the conditional distributions $F_0, \ldots, F_{M-1}$ of the rv $\boldsymbol{Y}$ given H.

The optimization problem A cost is incurred for making decisions. This is captured through the mapping $C : \{0, 1, \ldots, M-1\} \times \{0, 1, \ldots, M-1\} \to \mathbb{R}$ with the interpretation that

$$C(m, \ell) = \begin{array}{c}\text{Cost incurred for deciding } \ell \\ \text{when } H = m\end{array}, \quad \ell, m = 0, 1, \ldots, M-1.$$

Using the admissible rule d in $\mathcal{D}$ incurs a cost $C(H, d(Y))$, but as for the binary hypothesis testing problem, the value of the cost $C(H, d(\boldsymbol{Y}))$ is not available, and attention focuses on the *expected cost* $J : \mathcal{D} \to \mathbb{R}$ given by

$$J(d) \equiv \mathbb{E}\left[C(H, d(\boldsymbol{Y}))\right], \quad d \in \mathcal{D}.$$

Here as well, the *Bayesian* Problem $\mathcal{P}_{\mathrm{B}}$ is the minimization problem

$$\mathcal{P}_{\mathrm{B}}: \quad \text{Minimize } J(d) \text{ over } d \text{ in } \mathcal{D}.$$

Its solution is any test $d^\star$ in $\mathcal{D}$ such that

$$J(d^\star) \leq J(d), \quad d \in \mathcal{D}. \tag{7.8}$$

Any test $d^\star$ in $\mathcal{D}$ which satisfies (7.8) is called a Bayesian test, and the value

$$J(d^\star) = \inf_{d \in \mathcal{D}} J(d) = \min_{d \in \mathcal{D}} J(d) \tag{7.9}$$

is known as the *Bayesian* cost.

The Bayesian test For each $\ell = 0, 1, \ldots, M-1$, we define the Borel mapping $h_\ell : \mathbb{R}^k \to \mathbb{R}$ given by

$$h_\ell(\boldsymbol{y}) \equiv \sum_{m=0,\ m\neq\ell}^{M-1} p_m \left(C(m,\ell) - C(m,m)\right) f_m(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k. \tag{7.10}$$

The next result identifies the Bayesian test; its proof parallels that given in Section 2.2 for the binary case but with some important differences. Details are given in Section 7.5.

Theorem 7.4.1 *Assume the absolute continuity condition* **(A.1)** *to hold. The test* $d^\star : \mathbb{R}^k \to \{0, 1, \ldots, M-1\}$ *given by*

$$d^\star(\boldsymbol{y}) = \arg\min \left(\ell = 0, \ldots, M-1 :\ h_\ell(\boldsymbol{y})\right), \quad \boldsymbol{y} \in \mathbb{R}^k \tag{7.11}$$

(with a lexicographic tiebreaker in the event of ties) is admissible and solves the Bayesian Problem $\mathcal{P}_{\mathrm{B}}$.

7.5 A proof of Theorem 7.4.1

A reduction step Fix a test d in $\mathcal{D}$. The decomposition

$$\mathbf{1}\left[d(\boldsymbol{Y}) = H\right] + \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right] = 1$$

holds so that

$$\begin{aligned} & C(H, d(\boldsymbol{Y})) \\ &= \mathbf{1}\left[d(\boldsymbol{Y}) = H\right] C(H,H) + \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right] C(H, d(\boldsymbol{Y})) \\ &= C(H,H) + \left(C(H, d(\boldsymbol{Y})) - C(H,H)\right) \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right]. \end{aligned} \tag{7.12}$$

Defining the auxiliary expected cost function $\widehat{J} : \mathcal{D} \to \mathbb{R}$ to be

$$\widehat{J}(d) = \mathbb{E}\left[\left(C(H, d(\boldsymbol{Y})) - C(H,H)\right) \mathbf{1}\left[d(\boldsymbol{Y}) \neq H\right]\right], \quad d \in \mathcal{D} \tag{7.13}$$

we again readily conclude that

$$J(d) = E[C(H,H)] + \widehat{J}(d), \quad d \in \mathcal{D}. \tag{7.14}$$

Therefore, solving $\mathcal{P}_{\mathrm{B}}$ is *equivalent* to solving the auxiliary problem $\widehat{\mathcal{P}}_{\mathrm{B}}$ where

$$\widehat{\mathcal{P}}_{\mathrm{B}} : \quad \text{Minimize } \widehat{J}(d) \text{ over } d \text{ in } \mathcal{D}.$$

Preparatory computations Fix d in $\mathcal{D}$. From (7.14) we note that

$$\begin{aligned}
\widehat{J}(d) &= \mathbb{E}\left[(C(H,d(\boldsymbol{Y}))-C(H,H))\,\mathbf{1}\left[d(\boldsymbol{Y})\neq H\right]\right] \\
&= \sum_{m=0}^{M-1}\left(\sum_{\ell=0,\ \ell\neq m}^{M-1}\mathbb{E}\left[(C(m,\ell)-C(m,m))\,\mathbf{1}\left[H=m,d(\boldsymbol{Y})=\ell\right]\right]\right) \\
&= \sum_{m=0}^{M-1}\left(\sum_{\ell=0,\ \ell\neq m}^{M-1}(C(m,\ell)-C(m,m))\,\mathbb{P}\left[H=m,d(\boldsymbol{Y})=\ell\right]\right) \\
&= \sum_{m=0}^{M-1}\left(\sum_{\ell=0,\ \ell\neq m}^{M-1}(C(m,\ell)-C(m,m))\,\mathbb{P}\left[d(\boldsymbol{Y})=\ell|H=m\right]p_m\right) \\
&= \sum_{m=0}^{M-1}\left(\sum_{\ell=0,\ \ell\neq m}^{M-1}p_m\,(C(m,\ell)-C(m,m))\int_{C_\ell(d)}f_m(\boldsymbol{y})dF(\boldsymbol{y})\right).
\end{aligned}$$

Interchanging the order of summation we get

$$\begin{aligned}
\widehat{J}(d) &= \sum_{\ell=0}^{M-1}\left(\sum_{m=0,\ m\neq\ell}^{M-1}p_m\,(C(m,\ell)-C(m,m))\int_{C_\ell(d)}f_m(\boldsymbol{y})dF(\boldsymbol{y})\right) \\
&= \sum_{\ell=0}^{M-1}\int_{C_\ell(d)}\left(\sum_{m=0,\ m\neq\ell}^{M-1}p_m\,(C(m,\ell)-C(m,m))\,f_m(\boldsymbol{y})\right)dF(\boldsymbol{y}) \\
&= \sum_{\ell=0}^{M-1}\int_{C_\ell(d)}h_\ell(\boldsymbol{y})dF(\boldsymbol{y}) \qquad (7.15)
\end{aligned}$$

where for each $\ell=0,1,\ldots,M-1$, the mapping $h_\ell:\mathbb{R}^k\to\mathbb{R}$ is given by (7.10).

Solving $\widehat{\mathcal{P}}_{\mathrm{B}}$ Define the mapping $h:\mathbb{R}^k\to\mathbb{R}$ given by

$$h(\boldsymbol{y})\equiv\min_{\ell=0,1,\ldots,M-1}h_\ell(\boldsymbol{y}),\quad \boldsymbol{y}\in\mathbb{R}^k. \qquad (7.16)$$

where for each $\ell=0,1,\ldots,M-1$, the mapping $h_\ell:\mathbb{R}^k\to\mathbb{R}$ is given by (7.10). The following facts

$$h(\boldsymbol{y})=h_{d^\star(\boldsymbol{y})}(\boldsymbol{y}),\quad \boldsymbol{y}\in\mathbb{R}^k \qquad (7.17)$$

and

$$h(\boldsymbol{y})\leq h_\ell(\boldsymbol{y}),\qquad \begin{array}{c}\ell=0,1,\ldots,M-1,\\ \boldsymbol{y}\in\mathbb{R}^k\end{array} \qquad (7.18)$$

are simple consequences of the definition (7.11) of $d^\star$.

Pick an arbitrary test in $\mathcal{D}$. Using the expression (7.15) we get

$$\begin{aligned}
&J(d) - J(d^\star) \\
&\quad = \widehat{J}(d) - \widehat{J}(d^\star) \\
&\quad = \sum_{\ell=0}^{M-1} \int_{C_\ell(d)} h_\ell(\boldsymbol{y}) dF(\boldsymbol{y}) - \sum_{\ell=0}^{M-1} \int_{C_\ell(d^\star)} h_\ell(\boldsymbol{y}) dF(\boldsymbol{y}) \\
&\quad \geq \sum_{\ell=0}^{M-1} \int_{C_\ell(d)} h(\boldsymbol{y}) dF(\boldsymbol{y}) - \sum_{\ell=0}^{M-1} \int_{C_\ell(d^\star)} h_\ell(\boldsymbol{y}) dF(\boldsymbol{y}) \\
&\quad = \sum_{\ell=0}^{M-1} \int_{C_\ell(d)} h(\boldsymbol{y}) dF(\boldsymbol{y}) - \sum_{\ell=0}^{M-1} \int_{C_\ell(d^\star)} h(\boldsymbol{y}) dF(\boldsymbol{y}) \\
&\quad = \int_{\mathbb{R}^k} h(\boldsymbol{y}) dF(\boldsymbol{y}) - \int_{\mathbb{R}^k} h(\boldsymbol{y}) dF(\boldsymbol{y}) \\
&\quad = 0. \qquad (7.19)
\end{aligned}$$

The inequality above and the equality that follows are consequences of (7.17) and (7.18), respectively. The last two steps used the fact that the collection $\{C_0(d), \ldots, C_{M-1}(d)\}$ (resp. $\{C_0(d^\star), \ldots, C_{M-1}(d^\star)\}$) forms an M-ary Borel partition of $\mathbb{R}^k$. In particular, this observation implies

$$\sum_{\ell=0}^{M-1} \int_{C_\ell(d)} h(\boldsymbol{y}) dF(\boldsymbol{y}) = \int_{\mathbb{R}^k} h(\boldsymbol{y}) dF(\boldsymbol{y})$$

and

$$\sum_{\ell=0}^{M-1} \int_{C_\ell(d^\star)} h(\boldsymbol{y}) dF(\boldsymbol{y}) = \int_{\mathbb{R}^k} h(\boldsymbol{y}) dF(\boldsymbol{y}).$$

From (7.19) it follows that

$$J(d) - J(d^\star) \geq 0, \quad d \in \mathcal{D},$$

and the optimality of $d^\star$ is now established. ■

7.6 The probability of error criterion

When C takes the form

$$C(m,k) = \mathbf{1}\left[\ell \neq m\right], \quad m,\ell = 0,1,\ldots,M-1,$$

the expected cost reduces to the *probability of error* criterion given by

$$P_E(d) = \mathbb{P}\left[d(\boldsymbol{Y}) \neq H\right], \quad d \in \mathcal{D}.$$

The Bayesian test $d^\star : \mathbb{R}^k \to \{0,1,\ldots,M-1\}$ given by (7.11) now takes the following form: For each $\ell = 0,\ldots,M-1$, the mapping $h_\ell : \mathbb{R}^k \to \mathbb{R}$ is now given by

$$h_\ell(\boldsymbol{y}) = \sum_{m=0,\ m\neq\ell}^{M-1} p_m f_m(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k.$$

But the probability distribution function $G : \mathbb{R}^k \to [0,1]$ of the observation rv $\boldsymbol{Y}$ is given by (7.3), and under condition **(A.1)** it has probability density function $g : \mathbb{R}^k \to \mathbb{R}_+$ given by (7.7). Therefore, for each $\ell = 0,\ldots,M-1$, we have

$$h_\ell(\boldsymbol{y}) = g(\boldsymbol{y}) - p_\ell f_\ell(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k$$

and the test $d^\star : \mathbb{R}^k \to \{0,1,\ldots,M-1\}$ can be rewritten more compactly as

$$d^\star(\boldsymbol{y}) = \arg\max\left(\ell = 0,\ldots,M-1 :\ p_\ell f_\ell(\boldsymbol{y})\right), \quad \boldsymbol{y} \in \mathbb{R}^k \tag{7.20}$$

with a lexicographic tiebreaker in the event of ties.

The ML test When all the hypotheses are *equally likely*, namely

$$p_0 = \ldots = p_{M-1} = \frac{1}{M},$$

then (7.20) becomes

$$d^\star(\boldsymbol{y}) = \arg\max\left(\ell = 0,\ldots,M-1 :\ f_\ell(\boldsymbol{y})\right), \quad \boldsymbol{y} \in \mathbb{R}^k \tag{7.21}$$

with a lexicographic tiebreaker in the event of ties, so that

$$d^\star(\boldsymbol{y}) = m \quad \text{iff} \quad f_m(\boldsymbol{y}) = \max\left(f_\ell(\boldsymbol{y}),\ \ell = 0,1,\ldots,M-1\right) \tag{7.22}$$

with a lexicographic tiebreaker in the event of ties.

The MAP computer Bayes's Theorem gives

$$\mathbb{P}\left[H=\ell|\boldsymbol{Y}=\boldsymbol{y}\right]=\frac{p_\ell f_\ell(\boldsymbol{y})}{\sum_{m=0}^{M-1}p_m f_m(\boldsymbol{y})},\qquad \begin{array}{c}\ell=0,1,\ldots,M-1,\\ \boldsymbol{y}\in\mathbb{R}^k.\end{array}$$

This allows rewriting the test $d^\star:\mathbb{R}^k\to\{0,1,\ldots,M-1\}$ in the more compact form

$$d^\star(\boldsymbol{y})=\arg\max\left(\ell=0,\ldots,M-1:\ \mathbb{P}\left[H=\ell|\boldsymbol{Y}=\boldsymbol{y}\right]\right),\quad \boldsymbol{y}\in\mathbb{R}^k \tag{7.23}$$

with a lexicographic tiebreaker in the event of ties. As with the binary case, we refer to this rule as the *Maximum A Posteriori* computer.

7.7 The Gaussian case

Consider the case where the rv $\boldsymbol{Y}$ is Gaussian under each hypothesis, namely

$$H_m:\ \boldsymbol{Y}\sim\mathrm{N}(\boldsymbol{a}_m,\boldsymbol{R}_m).\quad m=0,1,\ldots,M-1. \tag{7.24}$$

where $\boldsymbol{a}_0,\ldots,\boldsymbol{a}_{M-1}$ are elements in $\mathbb{R}^k$, and the $k\times k$ *symmetric* matrices $\boldsymbol{R}_0,\ldots,\boldsymbol{R}_{M-1}$ are *positive definite* (thus *invertible*). Condition **(A.3)** holds with respect to Lebesgue measure.

Throughout the M pairs $(\boldsymbol{a}_0,\boldsymbol{R}_0),\ldots,(\boldsymbol{a}_{M-1},\boldsymbol{R}_{M-1})$ are assumed to be distinct so that the probability density function $f_0,\ldots,f_{M-1}:\mathbb{R}^k\to\mathbb{R}_+$ are distinct. Indeed, for each $m=0,\ldots,M-1$, the probability density function $f_m:\mathbb{R}^k\to\mathbb{R}_+$ is given by

$$f_m(\boldsymbol{y})=\frac{1}{\sqrt{(2\pi)^k\det\boldsymbol{R}_m}}e^{-\frac{1}{2}Q_m(\boldsymbol{y})},\quad \boldsymbol{y}\in\mathbb{R}^k$$

with

$$Q_m(\boldsymbol{y})=(\boldsymbol{y}-\boldsymbol{a}_m)'\boldsymbol{R}_m^{-1}(\boldsymbol{y}-\boldsymbol{a}_m),\quad \boldsymbol{y}\in\mathbb{R}^k.$$

Thus, for each $\ell=0,\ldots,M-1$, we have

$$\begin{aligned}&h_\ell(\boldsymbol{y})\\ &\quad=\sum_{m=0,\ m\neq\ell}^{M-1}p_m\left(C(m,\ell)-C(m,m)\right)\frac{1}{\sqrt{(2\pi)^k\det\boldsymbol{R}_m}}e^{-\frac{1}{2}Q_m(\boldsymbol{y})},\quad \boldsymbol{y}\in\mathbb{R}^k.\end{aligned}$$

For the probability of error criterion, for each $\ell = 0, \ldots, M-1$, this last expression becomes

$$\begin{aligned} h_\ell(\boldsymbol{y}) &= \sum_{m=0,\ m\neq\ell}^{M-1} \frac{p_m}{\sqrt{(2\pi)^k \det \boldsymbol{R}_m}} e^{-\frac{1}{2}Q_m(\boldsymbol{y})} \\ &= g(\boldsymbol{y}) - \frac{p_\ell}{\sqrt{(2\pi)^k \det \boldsymbol{R}_\ell}} e^{-\frac{1}{2}Q_\ell(\boldsymbol{y})}, \quad \boldsymbol{y} \in \mathbb{R}^k. \end{aligned} \tag{7.25}$$

The equal covariance case When

$$\boldsymbol{R}_0 = \ldots = \boldsymbol{R}_{M-1} \equiv \boldsymbol{R},$$

additional simplifications occur: For each $\ell = 0, \ldots, M-1$, we find

$$\begin{aligned} &h_\ell(\boldsymbol{y}) \\ &= \frac{1}{\sqrt{(2\pi)^k \det \boldsymbol{R}}} \sum_{m=0,\ m\neq\ell}^{M-1} \frac{p_m \left(C(m,\ell) - C(m,m)\right)}{\sqrt{(2\pi)^k \det \boldsymbol{R}}} e^{-\frac{1}{2}Q(\boldsymbol{y}-\boldsymbol{a}_m)}, \quad \boldsymbol{y} \in \mathbb{R}^k. \end{aligned}$$

with

$$Q(\boldsymbol{y}) = \boldsymbol{y}' \boldsymbol{R}^{-1} \boldsymbol{y}, \quad \boldsymbol{y} \in \mathbb{R}^k.$$

For the probability of error criterion, for each $\ell = 0, \ldots, M-1$, we find

$$\begin{aligned} h_\ell(\boldsymbol{y}) &= \sum_{m=0,\ m\neq\ell}^{M-1} \frac{p_m}{\sqrt{(2\pi)^k \det \boldsymbol{R}}} e^{-\frac{1}{2}Q(\boldsymbol{y}-\boldsymbol{a})} \\ &= g(\boldsymbol{y}) - \frac{p_\ell}{\sqrt{(2\pi)^k \det \boldsymbol{R}}} e^{-\frac{1}{2}Q(\boldsymbol{y}-\boldsymbol{a}_\ell)}, \quad \boldsymbol{y} \in \mathbb{R}^k. \end{aligned} \tag{7.26}$$

Writing

$$d(p; \boldsymbol{y}) = \log p - \frac{1}{2} Q(\boldsymbol{y}), \qquad \begin{matrix} p \in (0,1) \\ \boldsymbol{y} \in \mathbb{R}^k, \end{matrix}$$

the Bayesian test $d^\star : \mathbb{R}^k \to \{0, \ldots, M-1\}$ now reduces to

$$d^\star(\boldsymbol{y}) = m \quad \text{iff} \quad \begin{matrix} d(p_m; \boldsymbol{y} - \boldsymbol{a}_m) \\ = \max\left(d(p_\ell; \boldsymbol{y} - \boldsymbol{a}_\ell),\ \ell = 0, 1, \ldots, M-1\right) \end{matrix}$$

with a lexicographic tiebreaker in the event of ties.

7.8 Exercises

Chapter 8

Composite hypothesis testing problems

8.1 A motivating example

Consider the following problem of deciding between two hypotheses. Under the null hypothesis H_0, the observation $\boldsymbol{Y}$ is an $\mathbb{R}^k$-valued rv which is normally distributed with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{R}$ which are both *known*. Under the alternative hypothesis H_1, the $\mathbb{R}^k$-valued rv $\boldsymbol{Y}$ is normally distributed with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{R}$ where $\boldsymbol{\theta} \neq \boldsymbol{m}$ is only known to lie in a subset Θ_1 of $\mathbb{R}^k$, and is otherwise unspecified. We assume that $\boldsymbol{m}$ is *not* an element of Θ_1.

This problem of testing for the binary hypothesis H_0 vs. H_1 can also be interpreted as one of deciding between the hypotheses

$$\begin{array}{ll} H_1: & \{H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_1\} \\ H_0: & \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{R}). \end{array} \tag{8.1}$$

where for each $\boldsymbol{\theta} \in \mathbb{R}^k$, we write

$$H_\theta:\ \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{R}).$$

In such situations, when Θ_1 is not reduced to a singleton, the alternative hypothesis can be viewed as a *composite* hypothesis $\{H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_1\}$. We emphasize that we seek to decide between H_0 and H_1, or equivalently, between H_0 and $\{H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_1\}$; the precise value of $\boldsymbol{\theta}$ is not thought

8.2 The probabilistic model

More generally consider two non-empty Borel subsets Θ_0 and Θ_1 of $\mathbb{R}^p$ for some positive integer p. Assume that

$$\Theta_0 \cap \Theta_1 = \emptyset.$$

We shall set

$$\Theta = \Theta_0 \cup \Theta_1.$$

so that the pair Θ_0 and Θ_1 forms a Borel partition of Θ.

Given is a family of probability distributions $\{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ on $\mathbb{R}^k$. For mathematical reasons, it is required that the mapping $\Theta \times \mathbb{R}^k \to [0,1] : (\boldsymbol{\theta}, \boldsymbol{y}) \to F_{\boldsymbol{\theta}}(\boldsymbol{y})$ be Borel measurable. This condition is satisfied in all applications of interest.

We are given a measurable space $(\Omega, \mathcal{F})$. The state of nature is modeled by means of a rv $\vartheta : \Omega \to \Theta$ defined on $(\Omega, \mathcal{F})$. The observation is given by an $\mathbb{R}^k$-valued rv $\boldsymbol{Y} : \Omega \to \mathbb{R}^k$ defined on the same measurable space $(\Omega, \mathcal{F})$. with the interpretation that

$$\mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y} | \vartheta = \boldsymbol{\theta}\right] = F_{\boldsymbol{\theta}}(\boldsymbol{y}), \qquad \begin{matrix} \boldsymbol{y} \in \mathbb{R}^k, \\ \boldsymbol{\theta} \in \Theta. \end{matrix} \tag{8.2}$$

The state of nature and the corresponding observation are summarized as

$$H_\theta : \ \boldsymbol{Y} \sim F_{\boldsymbol{\theta}} \tag{8.3}$$

with $\boldsymbol{\theta}$ ranging in Θ.

The *composite* binary hypothesis testing problem is then the problem of deciding between the two composite hypotheses $H_0 = \{H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_0\}$ and $H_1 = \{H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_1\}$ on the basis of the observation $\boldsymbol{Y}$.

If either Θ_0 or Θ_1 is reduced to a single element, the corresponding hypothesis is termed *simple*. Obviously the problems of simple binary hypothesis testing discussed in Chapters 1–5 obtains when each of the sets Θ_0 and Θ_1 contains exactly one element.

The composite binary hypothesis testing problem is concisely denoted by

$$\begin{matrix} H_1 : & \boldsymbol{Y} \sim F_\theta, & \theta \in \Theta_1 \\ H_0 : & \boldsymbol{Y} \sim F_\theta, & \theta \in \Theta_0 \end{matrix} \tag{8.4}$$

Sometimes, once the family of probability distributions $\{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ has been specified, the notation is simplified to read

$$\begin{matrix} H_1 : & \boldsymbol{\theta} \in \Theta_1 \\ H_0 : & \boldsymbol{\theta} \in \Theta_0 \end{matrix} \tag{8.5}$$

As in earlier chapters, the discussion will require that certain assumptions are enforced. Condition **(A.4)** given next parallels Condition **(A.1)** given in the binary case and Condition **(A.3)** given in the M-ary case.

Condition **(A.4)**: For each $\boldsymbol{\theta}$ in Θ, the probability distribution $F_{\boldsymbol{\theta}}$ on $\mathbb{R}^k$ is absolutely continuous with respect to some distribution F on $\mathbb{R}^k$.

This is equivalent to requiring that for each $\boldsymbol{\theta}$ in Θ, there exists a Borel mapping $f_{\boldsymbol{\theta}} : \mathbb{R}^k \to \mathbb{R}_+$ such that

$$F_{\boldsymbol{\theta}}(\boldsymbol{y}) = \int_{-\infty}^{\boldsymbol{y}} f_{\boldsymbol{\theta}}(\boldsymbol{\eta}) dF(\boldsymbol{\eta}), \quad \boldsymbol{y} \in \mathbb{R}^k. \tag{8.6}$$

For mathematical reasons we require that the mapping

$$\Theta \times \mathbb{R}^k \to \mathbb{R}_+ : (\theta, \boldsymbol{y}) \to f_\theta(\boldsymbol{y})$$

be Borel measurable. This condition is satisfied in all applications of interest.

In a manner reminiscent of the parameter estimation problem of Chapter 1 there are two possible cases, depending on whether or not $\boldsymbol{\theta}$ is modeled as a rv; these are the Bayesian and non-Bayesian cases, respectively.

8.3 The Bayesian case

In some settings it is appropriate to imagine that the value of the parameter $\boldsymbol{\theta}$ is indeed through some randomization mechanism. Thus, assume that there exists a Θ-valued rv ϑ defined on the measurable space $(\Omega, \mathcal{F})$, and let $K : \mathbb{R}^p \to [0, 1]$ denote its probability distribution. Thus,

$$\mathbb{P}[\vartheta \leq \boldsymbol{t}] = K(\boldsymbol{t}), \quad \boldsymbol{t} \in \mathbb{R}^p.$$

The requirement that $\mathbb{P}[\vartheta \in \Theta] = 1$ is equivalent to

$$\int_{\Theta^c} dK(\boldsymbol{t}) = 0.$$

As we now show, this composite binary hypothesis testing problem can be reformulated as a simple binary hypothesis testing problem. To do so we define the $\{0, 1\}$-valued rv H given by

$$H \equiv \mathbf{1}[\vartheta \in \Theta_1].$$

Note that

$$\begin{aligned} p &= \mathbb{P}[H=1] \\ &= \mathbb{P}[\vartheta \in \Theta_1] \\ &= \int_{\Theta_1} dK(\boldsymbol{t}). \end{aligned} \tag{8.7}$$

In a similar manner, since $1 - H = \mathbf{1}[\vartheta \in \Theta_0]$, we conclude that

$$\begin{aligned} 1-p &= \mathbb{P}[H=0] \\ &= \mathbb{P}[\vartheta \in \Theta_0] \\ &= \int_{\Theta_0} dK(\boldsymbol{t}). \end{aligned} \tag{8.8}$$

For each $h = 0, 1$, the conditional probability distribution of $\boldsymbol{Y}$ given that $H = h$ can be calculated as

$$\begin{aligned} F_h(\boldsymbol{y}) &\equiv \mathbb{P}[\boldsymbol{Y} \leq \boldsymbol{y} | H = h] \\ &= \frac{\mathbb{P}[\boldsymbol{Y} \leq \boldsymbol{y}, H = h]}{\mathbb{P}[H = h]} \\ &= \frac{\mathbb{P}[\boldsymbol{Y} \leq \boldsymbol{y}, \vartheta \in \Theta_h]}{\mathbb{P}[\vartheta \in \Theta_h]} \\ &= \frac{\mathbb{E}[\mathbb{E}[\mathbf{1}[\boldsymbol{Y} \leq \boldsymbol{y}] | \vartheta] \mathbf{1}[\vartheta \in \Theta_h]]}{\mathbb{P}[\vartheta \in \Theta_h]} \\ &= \frac{\mathbb{E}[\mathbb{P}[\boldsymbol{Y} \leq \boldsymbol{y} | \vartheta] \mathbf{1}[\vartheta \in \Theta_h]]}{\mathbb{P}[\vartheta \in \Theta_h]} \\ &= \frac{\mathbb{E}[F_\vartheta(\boldsymbol{y}) \mathbf{1}[\vartheta \in \Theta_h]]}{\mathbb{P}[\vartheta \in \Theta_h]} \\ &= \frac{\int_{\Theta_h} F_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})}{\int_{\Theta_h} dK(\boldsymbol{t})}, \quad \boldsymbol{y} \in \mathbb{R}^k. \end{aligned} \tag{8.9}$$

Fix $h = 0, 1$ and $\boldsymbol{y}$ in $\mathbb{R}^k$. Under Assumption **(A.4)** we note that

$$\begin{aligned} \int_{\Theta_h} F_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t}) &= \int_{\Theta_h} \left(\int_{-\infty}^{\boldsymbol{y}} f_{\boldsymbol{t}}(\boldsymbol{\eta}) dF(\boldsymbol{\eta}) \right) dK(\boldsymbol{t}) \\ &= \int_{-\infty}^{\boldsymbol{y}} \left(\int_{\Theta_h} f_{\boldsymbol{t}}(\boldsymbol{\eta}) dK(\boldsymbol{t}) \right) dF(\boldsymbol{\eta}) \end{aligned} \tag{8.10}$$

by Tonelli's Theorem, whence

$$F_h(\boldsymbol{y}) \;=\; \int_{-\infty}^{\boldsymbol{y}} \left(\frac{\int_{\Theta_h} f_{\boldsymbol{t}}(\boldsymbol{\eta}) dK(\boldsymbol{t})}{\int_{\Theta_h} dK(\boldsymbol{t})} \right) dF(\boldsymbol{\eta}). \tag{8.11}$$

Thus, Condition **(A.1)** holds for F_0 and F_1 with respect to F with probability density functions $f_0, f_1 : \mathbb{R}^k \to \mathbb{R}_+$ given by

$$f_h(\boldsymbol{y}) \equiv \frac{\int_{\Theta_h} f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})}{\int_{\Theta_h} dK(\boldsymbol{t})}, \qquad \begin{array}{c} h = 0, 1, \\ \boldsymbol{y} \in \mathbb{R}^k. \end{array}$$

At this point the reader might expect the corresponding tests $\{d_\eta,\ \eta \geq 0\}$ to play a prominent role where for each $\eta \geq 0$, the test $d_\eta : \mathbb{R}^k \to \{0, 1\}$ is given by

$$\begin{array}{lll} d_\eta(\boldsymbol{y}) = 0 & \text{iff} & f_1(\boldsymbol{y}) < \eta f_0(\boldsymbol{y}) \\ & \text{iff} & \dfrac{\int_{\Theta_1} f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})}{\int_{\Theta_1} dK(\boldsymbol{t})} < \eta \dfrac{\int_{\Theta_0} f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})}{\int_{\Theta_0} dK(\boldsymbol{t})}. \end{array} \tag{8.12}$$

In Section 8.9 we will see that this is not the case.

8.4 The Bayesian cost problem

In the Bayesian setup described in Section 8.3, we proceed as in Chapter 2 by introducing a cost incurred for making decisions. This is quantified by the mapping $C : \Theta \times \{0, 1\} \to \mathbb{R}$ with the interpretation that

$$C(\boldsymbol{\theta},d) = \begin{array}{c} \text{Cost incurred for deciding } d \\ \text{when } \boldsymbol{\vartheta} = \boldsymbol{\theta} \end{array}, \qquad \begin{array}{c} \boldsymbol{\theta} \in \Theta \\ d = 0, 1. \end{array}$$

We require that for each $d = 0, 1$, the mapping $\Theta \to \mathbb{R} : \boldsymbol{\theta} \to C(\boldsymbol{\theta},d)$ is Borel measurable. This guarantees that for every test $d : \mathbb{R}^k \to \{0, 1\}$, $C(\boldsymbol{\vartheta},d(\boldsymbol{Y}))$ is a rv defined on $(\Omega, \mathcal{F})$. To avoid unnecessary technical difficulties (and for ease of exposition) we further assume that

$$0 \leq C(\boldsymbol{\theta},d) \leq B, \qquad \begin{array}{c} d = 0, 1 \\ \boldsymbol{\vartheta} \in \Theta \end{array}$$

for some scalar B. Together these requirements ensure that the expectation $\mathbb{E}\left[C(\boldsymbol{\vartheta},d(\boldsymbol{Y}))\right]$ is well defined and finite.

With any admissible test $d : \mathbb{R}^k \to \{0, 1\}$, we define the expected cost

$$J(d) = \mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))\right]. \tag{8.13}$$

As before, the *Bayesian* Problem $\mathcal{P}_{\mathrm{B}}$ is the minimization problem

$$\mathcal{P}_{\mathrm{B}}: \quad \text{Minimize } J(d) \text{ over } d \text{ in } \mathcal{D}.$$

This amounts to finding an admissible test $d^\star$ in $\mathcal{D}$ such that

$$J(d^\star) \leq J(d), \quad d \in \mathcal{D}. \tag{8.14}$$

Any admissible test $d^\star$ which satisfies (8.14) is called a Bayesian test, and the value

$$J(d^\star) = \inf_{d \in \mathcal{D}} J(d) = \min_{d \in \mathcal{D}} J(d) \tag{8.15}$$

is known as the *Bayesian* cost.

In Section 8.3 we have shown that the Bayesian formulation of the composite binary hypothesis problem can be recast as a simple binary hypothesis problem. However, as discussed in Section 8.9, in general it is not possible to write

$$J(d) = \mathbb{E}\left[C_{\mathrm{New}}(H, d(\boldsymbol{Y}))\right], \quad d \in \mathcal{D}$$

for some mapping $C_{\mathrm{New}} : \{0, 1\} \times \{0, 1\} \to \mathbb{R}$ (possibly derived from the original cost function $C : \Theta \times \{0, 1\} \to \mathbb{R}$). This already suggests that the Bayesian test may not belong to the class of tests $\{d_\eta,\ \eta \geq 0\}$ introduced at (8.12).

Solving the Bayesian Problem $\mathcal{P}_{\mathrm{B}}$ In view of these comments a different approach is needed. Fix d in $\mathcal{D}$. A standard preconditioning argument gives

$$\begin{aligned}
J(d) &= \mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))|\boldsymbol{Y}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left[d(\boldsymbol{Y}) = 0\right] C(\boldsymbol{\vartheta}, d(\boldsymbol{Y})) + \mathbf{1}\left[d(\boldsymbol{Y}) = 0\right] C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))|\boldsymbol{Y}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left[d(\boldsymbol{Y}) = 0\right] C(\boldsymbol{\vartheta}, 0) + \mathbf{1}\left[d(\boldsymbol{Y}) = 1\right] C(\boldsymbol{\vartheta}, 1)|\boldsymbol{Y}\right]\right] \\
&= \mathbb{E}\left[\mathbf{1}\left[d(\boldsymbol{Y}) = 0\right] \mathbb{E}\left[C(\boldsymbol{\vartheta}, 0)|\boldsymbol{Y}\right] + \mathbf{1}\left[d(\boldsymbol{Y}) = 1\right] \mathbb{E}\left[C(\boldsymbol{\vartheta}, 1)|\boldsymbol{Y}\right]\right] \\
&= \mathbb{E}\left[\mathbf{1}\left[d(\boldsymbol{Y}) = 0\right] \widehat{C}(0, \boldsymbol{Y}) + \mathbf{1}\left[d(\boldsymbol{Y}) = 1\right] \widehat{C}(1, \boldsymbol{Y})\right]
\end{aligned} \tag{8.16}$$

where we have introduced the Borel mapping $\widehat{C} : \{0,1\} \times \mathbb{R}^k \to \mathbb{R}$ given by

$$\widehat{C}(d, \boldsymbol{y}) \equiv \mathbb{E}\left[C(\boldsymbol{\vartheta}, d) | \boldsymbol{Y} = \boldsymbol{y}\right], \qquad \begin{array}{l} d = 0, 1 \\ \boldsymbol{y} \in \mathbb{R}^k. \end{array}$$

Theorem 8.4.1 *Under the foregoing assumptions, the test $d^\star : \mathbb{R}^k \to \{0,1\}$ defined by*

$$d^\star(\boldsymbol{y}) = \begin{cases} 0 & \text{if } \widehat{C}(0, \boldsymbol{y}) < \widehat{C}(1, \boldsymbol{y}) \\ \\ 1 & \text{if } \widehat{C}(1, \boldsymbol{y}) \leq \widehat{C}(0, \boldsymbol{y}) \end{cases} \tag{8.17}$$

is admissible and solves the Bayesian Problem $\mathcal{P}_{\mathrm{B}}$.

Note that the result does not even depend on Assumption (**A.4**). Moreover, when $\widehat{C}(0, \boldsymbol{y}) = \widehat{C}(1, \boldsymbol{y})$, we may select $d^\star(\boldsymbol{y}) = 0$ or $d^\star(\boldsymbol{y}) = 1$ somewhat arbitrarily as long as the resulting mapping $\mathbb{R}^k \to \{0,1\}$ is Borel measurable.

Proof. The admissibility of $d^\star$ follows from the Borel measurability of the mapping $\widehat{C} : \{0,1\} \times \mathbb{R}^k \to \mathbb{R}$. To show its optimality note that for each test d in $\mathcal{D}$ it holds that

$$\begin{aligned} J(d) &= \mathbb{E}\left[\mathbf{1}\left[d(\boldsymbol{Y}) = 0\right] \widehat{C}(0, \boldsymbol{Y}) + \mathbf{1}\left[d(\boldsymbol{Y}) = 1\right] \widehat{C}(1, \boldsymbol{Y})\right] \\ &= \mathbb{E}\left[\widehat{C}(1, \boldsymbol{Y})\right] + \mathbb{E}\left[\mathbf{1}\left[d(\boldsymbol{Y}) = 0\right] \left(\widehat{C}(0, \boldsymbol{Y}) - \widehat{C}(1, \boldsymbol{Y})\right)\right] \\ &= \mathbb{E}\left[\widehat{C}(1, \boldsymbol{Y})\right] + \mathbb{E}\left[\mathbf{1}\left[\boldsymbol{Y} \in C(d)\right] \left(\widehat{C}(0, \boldsymbol{Y}) - \widehat{C}(1, \boldsymbol{Y})\right)\right] \end{aligned}$$

where as usual the Borel set $C(d)$ is given by (1.12). The conclusion readily follows by arguments similar to the ones used for the proof of Theorem 2.2.1: Indeed, we have

$$\begin{aligned} &J(d) - J(d^\star) \\ &\quad = \mathbb{E}\left[\left(\mathbf{1}\left[\boldsymbol{Y} \in C(d)\right] - \mathbf{1}\left[\boldsymbol{Y} \in C(d^\star)\right]\right) \left(\widehat{C}(0, \boldsymbol{Y}) - \widehat{C}(1, \boldsymbol{Y})\right)\right] \\ &\quad = \mathbb{E}\left[\mathbf{1}\left[\boldsymbol{Y} \in C(d) \backslash C(d^\star)\right] \cdot \left(\widehat{C}(0, \boldsymbol{Y}) - \widehat{C}(1, \boldsymbol{Y})\right)\right] \\ &\qquad - \mathbb{E}\left[\mathbf{1}\left[\boldsymbol{Y} \in C(d^\star) \backslash C(d)\right] \cdot \left(\widehat{C}(0, \boldsymbol{Y}) - \widehat{C}(1, \boldsymbol{Y})\right)\right] \end{aligned}$$

As we note that

$$C(d^\star) \equiv \left\{ \boldsymbol{y} \in \mathbb{R}^k : \ \widehat{C}(0, \boldsymbol{y}) < \widehat{C}(1, \boldsymbol{y}) \right\}$$

it is plain that

$$\mathbb{E}\left[\mathbf{1}\left[\boldsymbol{Y} \in C(d) \backslash C(d^\star) \right] \cdot \left(\widehat{C}(0, \boldsymbol{Y}) - \widehat{C}(1, \boldsymbol{Y}) \right) \right] \geq 0$$

and

$$\mathbb{E}\left[\mathbf{1}\left[\boldsymbol{Y} \in C(d^\star) \backslash C(d) \right] \cdot \left(\widehat{C}(0, \boldsymbol{Y}) - \widehat{C}(1, \boldsymbol{Y}) \right) \right] \leq 0.$$

It follows that $J(d^\star) \leq J(d)$ and the optimality of $d^\star$ is now established. ■

8.5 The non-Bayesian case – A generalized Neyman-Pearson formulation

In this formulation θ is an unknown parameter lying in Θ, and an approach à la Neyman-Pearson seems warranted. Since composite hypotheses are now involved, earlier definitions given in Chapter 4 need to be modified.

Consider a test $d : \mathbb{R}^k \to \{0, 1\}$ in $\mathcal{D}$ that tests the null hypothesis H_0 against the alternative H_1. We define its *size* to be the quantity

$$\alpha_{\Theta_0}(d) \equiv \sup_{\boldsymbol{\theta} \in \Theta_0} \mathbb{P}_{\boldsymbol{\theta}} \left[d(\boldsymbol{Y}) = 1 \right]. \tag{8.18}$$

With $\boldsymbol{\theta}$ in Θ_0, the probability $\mathbb{P}_{\boldsymbol{\theta}} \left[d(\boldsymbol{Y}) = 0 \right]$ can be interpreted as the probability of false alarm under the test d given that the hypothesis $H_{\boldsymbol{\theta}}$ is indeed correct.

Fix α in $[0, 1]$. Let $\mathcal{D}_{\Theta_0, \alpha}$ denote the collection of all tests in $\mathcal{D}$ whose size is no greater than α, namely,

$$\mathcal{D}_{\Theta_0, \alpha} \equiv \left\{ d \in \mathcal{D} : \ \alpha_{\Theta_0}(d) \leq \alpha \right\}. \tag{8.19}$$

When Θ_0 is reduced to a singleton $\boldsymbol{\theta}_0$ we write $\mathcal{D}_{\boldsymbol{\theta}_0, \alpha}$ instead of $\mathcal{D}_{\{\boldsymbol{\theta}_0\}, \alpha}$, so that

$$\mathcal{D}_{\boldsymbol{\theta}_0, \alpha} \equiv \left\{ d \in \mathcal{D} : \ \mathbb{P}_{\theta_0} \left[d(\boldsymbol{Y}) = 1 \right] \leq \alpha \right\}.$$

The inclusion

$$\mathcal{D}_{\Theta_0,\alpha} \subseteq \mathcal{D}_{\boldsymbol{\theta}_0,\alpha}, \quad \boldsymbol{\theta}_0 \in \Theta_0 \tag{8.20}$$

always holds – It is also easy to see that

$$\mathcal{D}_{\Theta_0,\alpha} = \cap_{\boldsymbol{\theta}_0 \in \Theta_0} \mathcal{D}_{\boldsymbol{\theta}_0,\alpha}.$$

When dealing with composite hypotheses, the generalized Neyman-Pearson formulation takes the following form:

For α in $[0,1]$, find a test $d_{\rm UMP}(\alpha) : \mathbb{R}^k \to \{0,1\}$ in $\mathcal{D}_{\Theta_0,\alpha}$ which is optimal in the sense that

$$\mathbb{P}_{\boldsymbol{\theta}}\left[d(\boldsymbol{Y})=1\right] \leq \mathbb{P}_{\boldsymbol{\theta}}\left[d_{\rm UMP}(\alpha)(\boldsymbol{Y})=1\right], \qquad \begin{array}{c}\boldsymbol{\theta} \in \Theta_1, \\ d \in \mathcal{D}_{\Theta_0,\alpha}.\end{array} \tag{8.21}$$

Such a test $d_{\rm UMP}(\alpha)$, when it exists, is called a *Uniformly Most Powerful* (UMP) test of size α for testing for testing $H_0 : H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_0$ against $H_1 : H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_1$,

---

A natural question is whether UMP tests exist and when they do, how does one go about identifying them. The discussion in Section 8.6 below provides some useful pointers concerning these issues.

8.6 Searching for UMP tests

If a UMP test $d_{\rm UMP}(\alpha)$ of size α *did* exist for testing $H_0 : H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_0$ against $H_1 : H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_1$, then by definition $d_{\rm UMP}(\alpha)$ is a test in $\mathcal{D}_{\Theta_0,\alpha}$, hence in $\mathcal{D}_{\boldsymbol{\theta}_0,\alpha}$ for *each* $\boldsymbol{\theta}_0$ in Θ_0. However, in general the optimality property (8.21) does not necessarily imply

$$\mathbb{P}_{\boldsymbol{\theta}}\left[d(\boldsymbol{Y})=1\right] \leq \mathbb{P}_{\boldsymbol{\theta}}\left[d_{\rm UMP}(\alpha)(\boldsymbol{Y})=1\right], \qquad \begin{array}{c}\boldsymbol{\theta} \in \Theta_1, \\ d \in \mathcal{D}_{\boldsymbol{\theta}_0,\alpha}\end{array} \tag{8.22}$$

since *a priori* the inclusion (8.20) may be strict.

Were it the case that $\mathcal{D}_{\Theta_0,\alpha} = \mathcal{D}_{\boldsymbol{\theta}_0,\alpha}$ for some $\boldsymbol{\theta}_0$ in Θ_0, then (8.22) would hold (as it now coincides with (8.21)), and the test $d_{\rm UMP}(\alpha)$ would therefore act as a Neyman-Pearson test $d_{\rm NP}(\alpha; \boldsymbol{\theta}_0, \boldsymbol{\theta})$ of size α for testing $H_0 \equiv H_{\boldsymbol{\theta}_0}$ against $H_1 \equiv H_{\boldsymbol{\theta}}$ for *each* $\boldsymbol{\theta}$ in Θ_1.

There is of course one situation when the equality $\mathcal{D}_{\Theta_0,\alpha} = \mathcal{D}_{\boldsymbol{\theta}_0,\alpha}$ obviously holds, namely when H_0 is a simple hypothesis so that Θ_0 is reduced to a singleton,

say $\Theta_0 = \{\boldsymbol{\theta}_0\}$ for some $\boldsymbol{\theta}_0$ not in Θ_1. The discussion above then implies that any UMP test $d_{\mathrm{UMP}}(\alpha)$, if one exists, for testing $H_0 \equiv H_{\boldsymbol{\theta}_0}$ against $H_1 : H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_1$, must satisfy

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\theta}}\left[d_{\mathrm{UMP}}(\alpha)(\boldsymbol{Y}) = 1\right] \\ &\quad = \ \mathbb{P}_{\boldsymbol{\theta}}\left[d_{\mathrm{NP}}(\alpha; \boldsymbol{\theta}_0, \boldsymbol{\theta}))(\boldsymbol{Y}) = 1\right], \quad \boldsymbol{\theta} \in \Theta_1. \end{aligned} \tag{8.23}$$

This is clearly a non-trivial restriction on the problem, and already suggests that UMP tests may not always exist even H_0 is a simple hypothesis – This is discussed on an example in Section 8.8.

Nevertheless, when H_0 is a simple hypothesis, these observations do point to an obvious strategy for finding UMPs: For each $\boldsymbol{\theta}$ in Θ_1, determine the Neyman-Pearson test $d_{\mathrm{NP}}(\alpha; \boldsymbol{\theta}_0, \boldsymbol{\theta})$ of size α for testing $H_0 \equiv H_{\boldsymbol{\theta}_0}$ against $H_1 \equiv H_{\boldsymbol{\theta}}$. If its implementation does not require explicit knowledge of $\boldsymbol{\theta}$, then the set $C(d_{\mathrm{NP}}(\alpha; \boldsymbol{\theta}_0, \boldsymbol{\theta}))$ will be independent of $\boldsymbol{\theta}$ in the sense that there exists a Borel subset C of $\mathbb{R}^k$ such that $C(d_{\mathrm{NP}}(\alpha; \boldsymbol{\theta}_0, \boldsymbol{\theta})) = C$ for every $\boldsymbol{\theta}$ in Θ_1. The test $d^\star : \mathbb{R}^k \to \{0, 1\}$ defined by

$$d^\star(\boldsymbol{y}) = \begin{cases} 0 & \text{if } \boldsymbol{y} \in C \\ \\ 1 & \text{if } \boldsymbol{y} \notin C \end{cases}$$

is an admissible test in $\mathcal{D}_{\boldsymbol{\theta}_{0,\alpha}}$ since by construction

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\theta}_0}\left[d^\star(\boldsymbol{Y}) = 1\right] &= \mathbb{P}_{\boldsymbol{\theta}_0}\left[\boldsymbol{Y} \notin C\right] \\ &= \mathbb{P}_{\boldsymbol{\theta}_0}\left[\boldsymbol{Y} \notin C(d_{\mathrm{NP}}(\alpha; \boldsymbol{\theta}_0, \boldsymbol{\theta}))\right] \\ &= \mathbb{P}_{\boldsymbol{\theta}_0}\left[d_{\mathrm{NP}}(\alpha; \boldsymbol{\theta}_0, \boldsymbol{\theta})(\boldsymbol{Y}) = 1\right] = \alpha. \end{aligned} \tag{8.24}$$

By the same arguments we also conclude that

$$\mathbb{P}_{\boldsymbol{\theta}}\left[d^\star(\boldsymbol{Y}) = 1\right] = \mathbb{P}_{\boldsymbol{\theta}}\left[d_{\mathrm{NP}}(\alpha; \boldsymbol{\theta}_0, \boldsymbol{\theta})(\boldsymbol{Y}) = 1\right], \quad \boldsymbol{\theta} \in \Theta_1 \tag{8.25}$$

and we conclude that $d^\star$ is a UMP test for of size α for testing for testing $H_0 \equiv H_{\boldsymbol{\theta}_0}$ against $H_1 : H_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_1$,

When H_0 is a composite hypothesis, finding a UMP test can be quite tricky. A first natural step would consist in finding the Neyman-Pearson test $d_{\mathrm{NP}}(\alpha; \boldsymbol{\theta}_0, \boldsymbol{\theta})$ of size α to test $H_0 \equiv H_{\boldsymbol{\theta}_0}$ against $H_1 \equiv H_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ *arbitrary* in Θ_0 and Θ_1, respectively. In some cases exploring the structure of these tests may lead to the UMP test of size α.

These ideas are illustrated through an example in Section 8.7 and Section 8.8.

8.7 An example

The discussion of Section 8.5 will be illustrated in the case when $\Theta \subseteq \mathbb{R}$ and the probability distributions $\{F_\theta,\ \theta \in \mathbb{R}\}$ are unit-variance Gaussian distributions on $\mathbb{R}$: Thus, with θ arbitrary in $\mathbb{R}$,

$$H_\theta :\ Y \sim \mathrm{N}(\theta, 1)$$

so that F_θ admits the density $f_\theta : \mathbb{R} \to \mathbb{R}_+$ given by

$$f_\theta(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2}, \quad y \in \mathbb{R}.$$

With distinct θ_0 and θ_1 in $\mathbb{R}$, consider the Neyman-Pearson formulation for the binary hypothesis problem

$$\begin{array}{ll} H_1 : & Y \sim \mathrm{N}(\theta_1, 1) \\ H_0 : & Y \sim \mathrm{N}(\theta_0, 1) \end{array} \tag{8.26}$$

As shown in Section 5.5 this problem has a complete solution: Fix $\lambda > 0$. The test $d_\lambda : \mathbb{R} \to \{0, 1\}$ takes the form

$$\begin{array}{lll} d_\lambda(y) = 0 & \text{iff} & \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta_1)^2} < \lambda \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta_0)^2} \\ & \text{iff} & (y-\theta_0)^2 < 2\log\lambda + (y-\theta_1)^2 \\ & \text{iff} & 2y(\theta_1 - \theta_0) < 2\log\lambda + \theta_1^2 - \theta_0^2. \end{array} \tag{8.27}$$

For notational convenience we shall write

$$T_\lambda(\theta_0; \theta_1) \equiv \frac{\log\lambda}{\theta_1 - \theta_0} + \frac{\theta_1 + \theta_0}{2}.$$

Two cases arise: If $\theta_0 < \theta_1$, then

$$d_\lambda(y) = 0 \qquad \text{iff} \qquad y < T_\lambda(\theta_0; \theta_1) \tag{8.28}$$

and by standard arguments we get

$$\begin{aligned} \mathbb{P}_{\theta_0}\left[d_\lambda(\boldsymbol{Y}) = 1\right] &= \mathbb{P}_{\theta_0}\left[Y \geq T_\lambda(\theta_0; \theta_1)\right] \\ &= \mathbb{P}_{\theta_0}\left[Y - \theta_0 \geq \frac{\log\lambda}{\theta_1 - \theta_0} + \frac{\theta_1 - \theta_0}{2}\right] \\ &= 1 - \Phi\left(\frac{\log\lambda}{\theta_1 - \theta_0} + \frac{\theta_1 - \theta_0}{2}\right). \end{aligned} \tag{8.29}$$

If $\theta_1 < \theta_0$, then

$$d_\lambda(y) = 0 \qquad \text{iff} \qquad y > T_\lambda(\theta_0; \theta_1), \tag{8.30}$$

and this time we find

$$\begin{aligned} \mathbb{P}_{\theta_0}\left[d_\lambda(\boldsymbol{Y}) = 1\right] &= \mathbb{P}_{\theta_0}\left[Y \leq T_\lambda(\theta_0; \theta_1)\right] \\ &= \mathbb{P}_{\theta_0}\left[Y - \theta_0 \leq \frac{\log\lambda}{\theta_1 - \theta_0} + \frac{\theta_1 - \theta_0}{2}\right] \\ &= \Phi\left(\frac{\log\lambda}{\theta_1 - \theta_0} + \frac{\theta_1 - \theta_0}{2}\right). \end{aligned} \tag{8.31}$$

Fix α in $(0, 1)$. The Neyman-Pearson test of size α for testing H_{θ_1} against H_{θ_0} is the test $d_{\lambda(\theta_1,\theta_0;\alpha)}$ where $\lambda(\theta_1, \theta_0; \alpha)$ is that value of $\lambda > 0$ determined by $\mathbb{P}_{\theta_0}\left[d_\lambda(\boldsymbol{Y}) = 1\right] = \alpha$.

If $\theta_0 < \theta_1$, then

$$\frac{\log\lambda(\theta_1, \theta_0; \alpha)}{\theta_1 - \theta_0} + \frac{\theta_1 - \theta_0}{2} = \Phi^{-1}(1 - \alpha),$$

and the test $d_{\lambda(\theta_1,\theta_0;\alpha)}$ is given by

$$d_{\lambda(\theta_1,\theta_0;\alpha)}(y) = 0 \quad \text{iff} \quad y < \theta_0 + \Phi^{-1}\left(1 - \alpha\right).$$

If $\theta_1 < \theta_0$, then

$$\frac{\log\lambda(\theta_1, \theta_0; \alpha)}{\theta_1 - \theta_0} + \frac{\theta_1 - \theta_0}{2} = \Phi^{-1}(\alpha),$$

and the test $d_{\lambda(\theta_1,\theta_0;\alpha)}$ is given by

$$d_{\lambda(\theta_1,\theta_0;\alpha)}(y) = 0 \quad \text{iff} \quad y > \theta_0 + \Phi^{-1}\left(\alpha\right).$$

8.8 UMP tests for the example

We now consider four different situations, each associated with different sets Θ_0 and Θ_1.

Case I: $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = (\theta_0, \infty)$ The test $d_{\lambda(\theta_1,\theta_0;\alpha)}$ given by

$$d_{\lambda(\theta_1,\theta_0;\alpha)}(y) = 0 \quad \text{iff} \quad y < \theta_0 + \Phi^{-1}\left(1 - \alpha\right)$$

can be implemented *without* explicit knowledge of θ_1. By the discussion of Section 8.5, it is plain that a UMP test $d_{\mathrm{UMP}}(\alpha)$ of size α exists with

$$C(d_{\mathrm{UMP}}(\alpha)) = \left\{y \in \mathbb{R} :\ y < \theta_0 + \Phi^{-1}\left(1 - \alpha\right)\right\} = \left(-\infty, \theta_0 + \Phi^{-1}\left(1 - \alpha\right)\right).$$

Case II: $\Theta_0 = \{\theta_0\}$ **and** $\Theta_1 = (-\infty, \theta_0)$ Here, the test $d_{\lambda(\theta_1,\theta_0;\alpha)}$ given by

$$d_{\lambda(\theta_1,\theta_0;\alpha)}(y) = 0 \quad \text{iff} \quad y > \theta_0 + \Phi^{-1}(\alpha).$$

can also be implemented without explicit knowledge of θ_1. Again, the discussion of Section 8.5 shows that a UMP test $d_{\mathrm{UMP}}(\alpha)$ of size α exists with

$$C(d_{\mathrm{UMP}}(\alpha)) = \left\{y \in \mathbb{R} : \ y > \theta_0 + \Phi^{-1}(\alpha)\right\} = \left(\theta_0 + \Phi^{-1}(\alpha), \infty\right).$$

Case III: $\Theta_0 = \{\theta_0\}$ **and** $\Theta_1 = \mathbb{R} - \{\theta_0\}$ It is plain from the discussion in Cases I and II that a UMP test $d_{\mathrm{UMP}}(\alpha)$ does not exist.

Case IV: $\Theta_0 = (-\infty, \theta_0]$ **and** $\Theta_1 = (\theta_0, \infty)$ We begin by picking σ *arbitrary* in Θ_0, so $\sigma \leq \theta_0$. Consider now the composite hypothesis testing problem

$$\begin{array}{ll} H_1: & Y \sim \mathrm{N}(\theta, 1), \quad \theta \in \Theta_1 \\ H_0: & Y \sim \mathrm{N}(\sigma, 1) \end{array} \tag{8.32}$$

The discussion in Case I also shows that a UMP test of size α exists; it was identified as the test $d^\star(\alpha;\sigma) : \mathbb{R}^k \to \{0,1\}$ given by

$$d^\star(\alpha;\sigma)(y) = 0 \quad \text{iff} \quad y < \sigma + \Phi^{-1}(1-\alpha).$$

This means that for every θ in Θ_1, it holds that

$$\begin{array}{rcl} \mathbb{P}_\theta\left[d^\star(\alpha;\sigma)(Y) = 1\right] & = & \mathbb{P}_\theta\left[Y \geq \sigma + \Phi^{-1}(1-\alpha)\right] \\ & \geq & \mathbb{P}_\theta\left[d(Y) = 1\right] \end{array} \tag{8.33}$$

for every test d in $\mathcal{D}_{\sigma,\alpha}$, i.e., for every test in $\mathcal{D}$ such that $\mathbb{P}_\sigma\left[d(Y) = 1\right] \leq \alpha$.

Now recall that $\mathcal{D}_{\Theta_0,\alpha}$ is contained in $\mathcal{D}_{\sigma,\alpha}$ since

$$\mathcal{D}_{\Theta_0,\alpha} = \left\{d \in \mathcal{D} : \ \sup_{\sigma' \leq \theta_0} \mathbb{P}_{\sigma'}\left[d(Y) = 1\right] \leq \alpha\right\}.$$

Therefore, for every θ in Θ_1, we have

$$\mathbb{P}_\theta\left[d(Y) = 1\right] \leq \mathbb{P}_\theta\left[Y \geq \sigma + \Phi^{-1}(1-\alpha)\right], \quad d \in \mathcal{D}_{\Theta_0,\alpha}$$

by virtue of (8.33) applied to the smaller class $\mathcal{D}_{\Theta_0,\alpha}$ of tests. Thus, with $\sigma \leq \theta_0$, the test $d^\star(\alpha;\sigma)$ would be the UMP test $d_{\mathrm{UMP}}(\alpha)$ we seek if only it belonged to $\mathcal{D}_{\Theta_0,\alpha}$.

Next we show that σ can be selected in Θ_0 such that the test $d^\star(\alpha;\sigma)$ is indeed an element of $\mathcal{D}_{\Theta_0,\alpha}$, in which case it is the desired UMP test $d_{\rm UMP}(\alpha)$. We shall prove that this happened if and only if $\sigma = \theta_0$: For arbitrary σ and σ' in $\mathbb{R}$ we always have

$$\begin{aligned} F(\sigma,\sigma') &\equiv \mathbb{P}_{\sigma'}\left[Y \geq \sigma + \Phi^{-1}(1-\alpha)\right] \\ &= \mathbb{P}_{\sigma'}\left[Y \geq \sigma' + (\sigma-\sigma') + \Phi^{-1}(1-\alpha)\right] \\ &= \mathbb{P}_{\sigma'}\left[Y - \sigma' \geq (\sigma-\sigma') + \Phi^{-1}(1-\alpha)\right] \\ &= 1 - \Phi\left((\sigma-\sigma') + \Phi^{-1}(1-\alpha)\right). \end{aligned}$$

With σ in $\mathbb{R}$ *given*, we see that the mapping $\mathbb{R} \to [0,1] : \sigma' \to F(\sigma,\sigma')$ is *strictly increasing* with

$$F(\sigma,\sigma) = \alpha$$

since $\Phi\left(\Phi^{-1}(1-\alpha)\right) = 1-\alpha$, whence $F(\sigma,\sigma) < \alpha$ if $\sigma' < \sigma$ and $\alpha < F(\sigma,\sigma)$ if $\sigma < \sigma'$. It now follows that

$$\sup_{\sigma' \leq \theta_0} \mathbb{P}_{\sigma'}\left[d^\star(\alpha;\sigma)(Y) = 1\right] > \alpha \quad \text{if } \sigma < \theta_0$$

while

$$\sup_{\sigma' \leq \theta_0} \mathbb{P}_{\sigma'}\left[d^\star(\alpha;\theta_0)(Y) = 1\right] = \alpha.$$

This shows that the test $d^\star(\alpha;\theta_0)$ is indeed an element of $\mathcal{D}_{\Theta_0,\alpha}$, and therefore can be used to implement the desired UMP test $d_{\rm UMP}(\alpha)$. ■

8.9 Reformulating the Bayesian cost

Fix d in $\mathcal{D}$. By iterated conditioning we conclude that

$$\begin{aligned} J(d) &= \mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))|\boldsymbol{\vartheta}\right]|H\right]\right] \end{aligned} \tag{8.34}$$

with

$$\begin{aligned} &\mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))|\boldsymbol{\vartheta}\right] \\ &= C(\boldsymbol{\vartheta},0)\cdot\mathbb{P}\left[d(\boldsymbol{Y})=0|\boldsymbol{\vartheta}\right] + C(\boldsymbol{\vartheta},1)\cdot\mathbb{P}\left[d(\boldsymbol{Y})=1|\boldsymbol{\vartheta}\right] \\ &= C(\boldsymbol{\vartheta},0)\cdot\mathbb{P}\left[\boldsymbol{Y}\in C(d)|\boldsymbol{\vartheta}\right] + C(\boldsymbol{\vartheta},1)\cdot\mathbb{P}\left[\boldsymbol{Y}\notin C(d)|\boldsymbol{\vartheta}\right] \\ &= C(\boldsymbol{\vartheta},0)\int_{C(d)} f_{\boldsymbol{\vartheta}}(\boldsymbol{y})dF(\boldsymbol{y}) + C(\boldsymbol{\vartheta},1)\int_{C(d)^c} f_{\boldsymbol{\vartheta}}(\boldsymbol{y})dF(\boldsymbol{y}). \end{aligned} \tag{8.35}$$

On the other hand, for each $h = 0, 1$, we note as before that

$$\begin{aligned}
\mathbb{P}\left[\boldsymbol{\vartheta} \leq \boldsymbol{t} | H = h\right] &= \frac{\mathbb{P}\left[\boldsymbol{\vartheta} \leq \boldsymbol{t}, H = h\right]}{\mathbb{P}\left[H = h\right]} \\
&= \frac{\mathbb{P}\left[\boldsymbol{\vartheta} \leq \boldsymbol{t}, \boldsymbol{\vartheta} \in \Theta_h\right]}{\mathbb{P}\left[\boldsymbol{\vartheta} \in \Theta_h\right]} \\
&= \frac{\int_{(-\boldsymbol{\infty}, \boldsymbol{t}] \cap \Theta_h} dK(\boldsymbol{\tau})}{\int_{\Theta_h} dK(\boldsymbol{\tau})} \\
&= \frac{\int_{-\boldsymbol{\infty}}^{\boldsymbol{t}} \mathbf{1}\left[\boldsymbol{\tau} \in \Theta_h\right] dK(\boldsymbol{\tau})}{\int_{\Theta_h} dK(\boldsymbol{\tau})}, \quad \boldsymbol{t} \in \mathbb{R}^p. \qquad (8.36)
\end{aligned}$$

Therefore, the conditional distribution of $\boldsymbol{\vartheta}$ given $H = h$ is absolutely continuous with respect to the probability distribution $K : \mathbb{R}^p \to [0, 1]$, the corresponding probability density function $k_h : \mathbb{R}^p \to \mathbb{R}_+$ being given by

$$k_h(\boldsymbol{t}) = \frac{1}{\int_{\Theta_h} dK(\boldsymbol{\tau})} \cdot \mathbf{1}\left[\boldsymbol{t} \in \Theta_h\right], \quad \boldsymbol{t} \in \mathbb{R}^p.$$

Next, we get

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))|\boldsymbol{\vartheta}\right]|H = h\right] \\
= \;& \mathbb{E}\left[C(\boldsymbol{\vartheta}, 0)\int_{C(d)} f_{\boldsymbol{\vartheta}}(\boldsymbol{y})dF(\boldsymbol{y}) + C(\boldsymbol{\vartheta}, 1)\int_{C(d)^c} f_{\boldsymbol{\vartheta}}(\boldsymbol{y})dF(\boldsymbol{y})\Big|H = h\right] \\
= \;& \int_{\mathbb{R}^p}\left(C(\boldsymbol{t}, 0)\int_{C(d)} f_{\boldsymbol{t}}(\boldsymbol{y})dF(\boldsymbol{y}) + C(\boldsymbol{t}, 1)\int_{C(d)^c} f_{\boldsymbol{t}}(\boldsymbol{y})dF(\boldsymbol{y})\right) k_h(\boldsymbol{t})dK(\boldsymbol{t}) \\
= \;& \int_{C(d)}\left(\int_{\mathbb{R}^p} C(\boldsymbol{t}, 0) f_{\boldsymbol{t}}(\boldsymbol{y}) k_h(\boldsymbol{t})dK(\boldsymbol{t})\right) dF(\boldsymbol{y}) \\
&+ \int_{C(d)^c}\left(\int_{\mathbb{R}^p} C(\boldsymbol{t}, 1) f_{\boldsymbol{t}}(\boldsymbol{y}) k_h(\boldsymbol{t})dK(\boldsymbol{t})\right) dF(\boldsymbol{y}) \\
= \;& \int_{C(d)} \frac{\int_{\Theta_h} C(\boldsymbol{t}, 0) f_{\boldsymbol{t}}(\boldsymbol{y})dK(\boldsymbol{t})}{\int_{\Theta_h} dK(\boldsymbol{t})} dF(\boldsymbol{y}) + \int_{C(d)^c} \frac{\int_{\Theta_h} C(\boldsymbol{t}, 1) f_{\boldsymbol{t}}(\boldsymbol{y})dK(\boldsymbol{t})}{\int_{\Theta_h} dK(\boldsymbol{t})} dF(\boldsymbol{y}).
\end{aligned}$$

Next we observe that

$$\begin{aligned}
J(d) &= \mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))|\boldsymbol{\vartheta}\right]|H = 0\right]\mathbb{P}\left[H = 0\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))|\boldsymbol{\vartheta}\right]|H = 1\right]\mathbb{P}\left[H = 1\right],
\end{aligned}$$

with elementary algebra showing that

$$\begin{aligned}
J(d) &= \mathbb{E}\left[C(\boldsymbol{\vartheta}, d(\boldsymbol{Y}))\right] \\
&= \left(\int_{C(d)} \frac{\int_{\Theta_0} C(\boldsymbol{t},0) f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})}{\int_{\Theta_0} dK(\boldsymbol{t})} dF(\boldsymbol{y})\right) \cdot \mathbb{P}\left[H=0\right] \\
&\quad + \left(\int_{C(d)^c} \frac{\int_{\Theta_0} C(\boldsymbol{t},1) f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})}{\int_{\Theta_0} dK(\boldsymbol{t})} dF(\boldsymbol{y})\right) \cdot \mathbb{P}\left[H=0\right] \\
&\quad + \left(\int_{C(d)} \frac{\int_{\Theta_1} C(\boldsymbol{t},0) f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})}{\int_{\Theta_1} dK(\boldsymbol{t})} dF(\boldsymbol{y})\right) \cdot \mathbb{P}\left[H=1\right] \\
&\quad + \left(\int_{C(d)^c} \frac{\int_{\Theta_1} C(\boldsymbol{t},1) f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})}{\int_{\Theta_1} dK(\boldsymbol{t})} dF(\boldsymbol{y})\right) \cdot \mathbb{P}\left[H=1\right] \\
&= \int_{C(d)} \left(\int_{\Theta} C(\boldsymbol{t},0) f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})\right) dF(\boldsymbol{y}) \\
&\quad + \int_{C(d)^c} \left(\int_{\Theta} C(\boldsymbol{t},1) f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})\right) dF(\boldsymbol{y}) \qquad (8.37)
\end{aligned}$$

as we group like terms after noting that $\mathbb{P}\left[H=h\right] = \int_{\Theta_h} dK(\boldsymbol{t})$.

Finally we conclude that

$$\begin{aligned}
J(d) &= \int_{\mathbb{R}^k} \left(\int_{\Theta} C(\boldsymbol{t},1) f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})\right) dF(\boldsymbol{y}) \\
&\quad + \int_{C(d)} \left(\int_{\Theta} \left(C(\boldsymbol{t},0) - C(\boldsymbol{t},1)\right) f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t})\right) dF(\boldsymbol{y}) \qquad (8.38)
\end{aligned}$$

and the Bayesian test $d^\star : \mathbb{R}^k \to \{0,1\}$ is therefore given by

$$d^\star(\boldsymbol{y}) = 0 \quad \text{iff} \quad \int_{\Theta} \left(C(\boldsymbol{t},0) - C(\boldsymbol{t},1)\right) f_{\boldsymbol{t}}(\boldsymbol{y}) dK(\boldsymbol{t}) < 0. \qquad (8.39)$$

The Bayesian test $d^\star$ is *not* an element of the class of tests $\{d_\eta,\ \eta \geq 0\}$ introduced at (8.12). ■

8.10 Exercises

Part II

ESTIMATION THEORY

Part III

APENDICES

Chapter 9

Useful facts from Real Analysis

9.1 Limits in $\mathbb{R}$

We refer to any mapping $a : \mathbb{N}_0 \to \mathbb{R}$ as a ($\mathbb{R}$-valued) *sequence*; sometimes we shall also use the notation $\{a_n,\ n = 1, 2, \ldots\}$.

A sequence $a : \mathbb{N}_0 \to \mathbb{R}$ converges to $a^\star$ in $\mathbb{R}$ if for every $\varepsilon > 0$, there exists an integer $n^\star(\varepsilon)$ in $\mathbb{N}_0$ such that

$$|a_n - a^\star| \leq \varepsilon, \quad n \geq n^\star(\varepsilon). \tag{9.1}$$

We write $\lim_{n\to\infty} a_n = a^\star$, and refer to the scalar $a^\star$ as the *limit* of the sequence.

Sometimes it is desirable to make sense of situations where the values of the sequence become either unboundedly large or unboundely negative, in which case we shall write $\lim_{n\to\infty} a_n = \infty$ and $\lim_{n\to\infty} a_n = -\infty$, respectively. A precise definition of such occurences is as follows: We write $\lim_{n\to\infty} a_n = \infty$ to signify that for every $M > 0$, there exists a integer $n^\star(M)$ in $\mathbb{N}_0$ such that

$$a_n > M, \quad n \geq n^\star(M). \tag{9.2}$$

It is now natural to define $\lim_{n\to\infty} a_n = -\infty$ whenever $\lim_{n\to\infty} (-a_n) = \infty$.

If there exists $a^\star$ in $\mathbb{R} \cup \{\pm\infty\}$ such that $\lim_{n\to\infty} a_n = a^\star$, we shall simply say that the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ *converges* or *is convergent* (without any reference to its limit). Sometimes we shall also say that the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ converges (or is convergent) *in* $\mathbb{R}$ to indicate that the limit $a^\star$ is an element of $\mathbb{R}$ (thus finite).

Convergence is guaranteed under conditions of monotonicity: A sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is *monotone non-decreasing* (resp. *non-increasing*) if $a_n \leq a_{n+1}$ (resp. $a_{n+1} \leq a_n$) for all $n = 1, 2, \ldots$.

Fact 9.1.1 *A monotone sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is always convergent, although its limit may be $\pm\infty$.*

9.2 Accumulation points

Since not all sequences converge, it is important to understand how can non-convergence occur. To that end, consider a sequence $a : \mathbb{N}_0 \to \mathbb{R}$. A *subsequence* of the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is any sequence of the form $\mathbb{N}_0 \to \mathbb{R} : k \to a_{n_k}$ where

$$n_k < n_{k+1}, \quad k = 1, 2, \ldots$$

This forces $\lim_{k\to\infty} n_k = \infty$.

An *accumulation point* for the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is defined as any element $a^\star$ in $\mathbb{R} \cup \{\pm\infty\}$ such that

$$\lim_{k\to\infty} a_{n_k} = a^\star$$

for *some* subsequence $\mathbb{N}_0 \to \mathbb{R} : k \to a_{n_k}$.

A convergent sequence $a : \mathbb{N}_0 \to \mathbb{R}$ has exactly *one* accumulation point, namely its limit. if the sequence *does not* converge, it must necessarily have distinct accumulation points, in which case there is a smallest and a largest accumulation point. The next definition formalizes this observation: Given a sequence $a : \mathbb{N}_0 \to \mathbb{R}$, the quantities

$$\limsup_{n\to\infty} a_n = \inf_{n\geq 1} \left(\sup_{m\geq n} a_m \right)$$

and

$$\liminf_{n\to\infty} a_n = \sup_{n\geq 1} \left(\inf_{m\geq n} a_m \right)$$

are known as the *limsup* and *liminf* of the sequence $a : \mathbb{N}_0 \to \mathbb{R}$, respectively.

It is plain that

$$\inf_{m \geq n} a_m \leq \sup_{m \geq n} a_m, \quad n = 1, 2, \ldots$$

and that the sequences $\{\inf_{m \geq n} a_m,\ n = 1, 2, \ldots\}$ and $\{\sup_{m \geq n} a_m,\ n = 1, 2, \ldots\}$ are non-decreasing and non-increasing, respectively, with

$$\lim_{n \to \infty} \left(\inf_{m \geq n} a_m \right) = \liminf_{n \to \infty} a_n$$

and

$$\lim_{n \to \infty} \left(\sup_{m \geq n} a_m \right) = \limsup_{n \to \infty} a_n,$$

hence the terminology. A useful characterization of convergence can now be provided in terms of the limsup and liminf.

Fact 9.2.1 *Consider a sequence $a : \mathbb{N}_0 \to \mathbb{R}$: If it converges to an element $a^\star$ (in $\mathbb{R} \cup \{\pm\infty\}$), then*

$$\liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n = a^\star.$$

Conversely, if $\liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n = a^\star$ for some $a^\star$ in $\mathbb{R} \cup \{\pm\infty\}$, then the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ converges to $a^\star$.

9.3 Continuous functions

Let I denote a subset of $\mathbb{R}$. A function $g : I \to \mathbb{R}$ is said to be *left-continuous* at x in I if for any sequence $a : \mathbb{N}_0 \to I$ such that $\lim_{n \to \infty} a_n = x$ with $a_n < x$ for all $n = 1, 2, \ldots$ we have

$$\lim_{n \to \infty} g(a_n) = g(x). \tag{9.3}$$

Similarly, a function $g : I \to \mathbb{R}$ is said to be *right-continuous* at x in I if for any sequence $a : \mathbb{N}_0 \to I$ such that $\lim_{n \to \infty} a_n = x$ with $x < a_n$ for all $n = 1, 2, \ldots$ we have (9.3).

Finally, a function $g : I \to \mathbb{R}$ is said to be *continuous* at x in I if it is both left-continuous and right-continuous at x. This is equivalent to (9.3) holding for any sequence $a : \mathbb{N}_0 \to I$ such that $\lim_{n \to \infty} a_n = x$.

A function $g : I \to \mathbb{R}$ is left-continuous (resp. right-continuous, continuous) on I if the function $g : I \to \mathbb{R}$ is left-continuous (resp. right-continuous, continuous) at every point x in I.

9.4 Convex functions

Let I denote an interval in $\mathbb{R}$. A function $g : I \to \mathbb{R}$ is said to be *convex* if for every x_0 and x_1 in I, it holds that

$$g((1-\lambda)x_0 + \lambda x_1) \leq (1-\lambda)g(x_0) + \lambda g(x_1), \quad \lambda \in [0,1]. \tag{9.4}$$

A function $g : I \to \mathbb{R}$ is said to be *concave* if the function $-g$ is convex. Here are some well-known facts concerning convex functions; the analog properties for concave functions are easily obtained *mutatis mutandis*.

Fact 9.4.1 *Let $g : I \to \mathbb{R}$ be a convex function. With $x < y < z$ in I, we have the basic inequalities*

$$\frac{g(y)-g(x)}{y-x} \leq \frac{g(z)-g(x)}{z-x} \tag{9.5}$$

and

$$\frac{g(z)-g(x)}{z-x} \leq \frac{g(z)-g(y)}{z-y}. \tag{9.6}$$

Proof. With $x < y < z$ in I, write

$$y = (1-\lambda)x + \lambda z \quad \text{where} \quad \lambda = \frac{y-x}{z-x}.$$

It is plain that λ is an element of $(0,1)$, and the convexity of g implies

$$\begin{aligned} g(y) &\leq (1-\lambda)g(x) + \lambda g(z) \\ &= \frac{z-y}{z-x}g(x) + \frac{y-x}{z-x}g(z). \end{aligned} \tag{9.7}$$

Substracting $g(x)$ from both sides of this inequality we get

$$\begin{aligned} g(y) - g(x) &\leq \frac{z-y}{z-x}g(x) + \frac{y-x}{z-x}g(z) - g(x) \\ &= \left(\frac{y-x}{z-x}\right)(g(z) - g(x)). \end{aligned}$$

and this establishes (9.5).

On the other hand, subtracting $g(z)$ from (9.7) we find that

$$\begin{aligned} g(y) - g(z) &\leq \frac{z-y}{z-x} g(x) + \frac{y-x}{z-x} g(z) - g(z) \\ &= \left(\frac{z-y}{z-x} \right) (g(x) - g(z)) \end{aligned}$$

and the inequality (9.5) readily follows. ■

Fact 9.4.2 *If the mapping $g : I \to \mathbb{R}$ is convex on an interval I, then it is also continuous on the interior of I.*

Continuity may fail at the boundary points as the next example shows: With $I = [0, 1]$, take the mapping $g : I \to \mathbb{R}$ given by

$$g(x) = \begin{cases} 0 & \text{if } x = 0 \\ \\ x & \text{if } 0 < x \leq 1. \end{cases}$$

This function is clearly convex on $[0, 1]$ but it fails to be continuous at $x = 0$.

Proof. Pick x in the interior of I so that $[x - \tau, x + \tau] \subseteq I$ for some $\tau > 0$. With $0 < t < 1$ we note that $x + t\tau = t(x + \tau) + (1 - t)x$, whence $g(x + t\tau) \leq tg(x + \tau) + (1 - t)g(x)$ by the convexity of g and we conclude

$$\frac{g(x + t\tau) - g(x)}{t} \leq g(x + \tau) - g(x).$$

Similarly, since $x - t\tau = t(x - \tau) + (1 - t)x$, we have $g(x - t\tau) \leq tg(x - \tau) + (1 - t)g(x)$ by the convexity of g, so that

$$g(x) - g(x - \tau) \leq \frac{g(x) - g(x - t\tau)}{t}.$$

But $\frac{1}{2}(x - t\tau) + \frac{1}{2}(x + t\tau) = x$, and using convexity again leads to

$$g(x) - g(x - t\tau) \leq g(x + t\tau) - g(x),$$

whence

$$g(x) - g(x-\tau) \leq \frac{g(x-t\tau) - g(x)}{-t} \leq \frac{g(x+t\tau) - g(x)}{t} \leq g(x+\tau) - g(x).$$

Therefore, with

$$M \equiv \max\left(|g(x) - g(x-\tau)|, |g(x+\tau) - g(x)|\right),$$

we get

$$-M \leq \frac{g(x) - g(x-t\tau)}{-t} \leq \frac{g(x+t\tau) - g(x)}{t} \leq M, \quad 0 < t < 1.$$

Changing notation we see that

$$\left|\frac{g(x+h) - g(x)}{h}\right| \leq M, \quad |h| \leq \tau \tag{9.8}$$

and the continuity of g at x is now immediate since $\lim_{n\to\infty} g(a_n) = g(x)$ for any sequence I-valued sequence $a : \mathbb{N}_0 \to I$ such that $\lim_{n\to\infty} a_n = x$. ■

In the course of the proof of Fact 9.4.2 we saw that the property (9.8), known as local Lipschitz continuity, holds.

Fact 9.4.3 *If the mapping $g : I \to \mathbb{R}$ is convex on an interval I, then it is also locally Lipschitz at every point in the interior of I.*

This paves the way for the following differentiability result.

Fact 9.4.4 *If the mapping $g : I \to \mathbb{R}$ is convex on some interval I, then its left and right-derivatives always exist at every point of continuity in I (and whenever appropriate at the boundary points of I).*

Proof. Pick ξ in the interior of I so that $(\xi - \tau, \xi + \tau) \subseteq I$ for some $\tau > 0$. By virtue of (9.5) we see that the mapping $t \to \frac{g(\xi+t)-g(\xi)}{t}$ is monotone increasing on

the interval $(0, \tau)$ – Just use $x = \xi$, $y = \xi + t_1$ and $z = \xi + t_2$ with $0 < t_1 < t_2 < \tau$. It follows that the limit defining the right-derivative at ξ, namely

$$\frac{d^+}{dx} g(\xi) = \lim_{t \downarrow 0} \frac{g(\xi + t) - g(\xi)}{t},$$

always exists.

Similarly, by virtue of (9.6) we see that the mapping $t \to \frac{g(\xi - t) - g(\xi)}{-t}$ is monotone increasing on the interval $(0, \tau)$ – Just use $z = \xi$, $y = \xi - t_1$ and $z = \xi - t_2$ with $0 < t_1 < t_2 < \tau$. It follows that the limit defining the left-derivative at ξ, namely

$$\frac{d^-}{dx} g(\xi) = \lim_{t \downarrow 0} \frac{g(\xi - t) - g(\xi)}{-t},$$

always exists. ■

9.5 Measurable spaces

Let S denote an arbitrary non-empty set. A non-empty collection $\mathcal{S}$ of subsets of S is a *σ-field* (also known as an *σ-algebra*) on S if

(i) $\mathcal{S}$ contains the empty set $\emptyset$;

(ii) $\mathcal{S}$ is closed under complementarity: If $E \in \mathcal{S}$, then $E^c \in \mathcal{S}$ (where E^c is the complement of E in S); and

(iii) $\mathcal{S}$ is closed under countable union: With I a *countable* index set, if $E_i \in \mathcal{S}$ for each $i \in I$, then $\cup_{i \in I} E_i \in \mathcal{S}$.

The pair $(S, \mathcal{S})$ is sometimes referred to as a *measurable space*. For every non-empty set S, there at least two distinct σ-fields on S, namely the trivial σ-field $\mathcal{S}_{\text{Triv}} = \{\emptyset, S\}$ and the complete σ-field $\mathcal{P}(S)$ (where $\mathcal{P}(S)$ denotes the power set of S).

If $\mathcal{S}_1$ and $\mathcal{S}_2$ are two σ-fields on S, we say that $\mathcal{S}_1$ contains $\mathcal{S}_2$, written $\mathcal{S}_1 \subseteq \mathcal{S}_2$, if any element of $\mathcal{S}_1$ is an element of $\mathcal{S}_2$. Thus, for any σ-field $\mathcal{S}$, we have $\mathcal{S}_{\text{Triv}} \subseteq \mathcal{S} \subseteq \mathcal{P}(S)$.

If $\mathcal{G}$ is a collection of subsets of S, then $\sigma(\mathcal{G})$ is defined as the *smallest* σ-field on S which contains $\mathcal{G}$, i.e., every element of $\mathcal{G}$ is also an element of $\sigma(\mathcal{G})$. We shall refer to $\sigma(\mathcal{G})$ as the σ-field on S generated by $\mathcal{G}$.

9.6 Borel measurability

With A denoting a subset of $\mathbb{R}^p$ for some positive integer p, we define the σ-field $\mathcal{B}(A)$ to be

$$\mathcal{B}(A) \equiv \sigma\left(\mathcal{O}(A)\right)$$

where $\mathcal{O}(A)$ denotes the collection of all *open sets* contained in A. In particular,

$$\mathcal{B}(\mathbb{R}^p) \equiv \sigma\left(\mathcal{O}(\mathbb{R}^p)\right)$$

where $\mathcal{O}(\mathbb{R}^p)$ denotes the collection of all *open sets* contained in $\mathbb{R}^p$.

Consider an arbitrary set S equipped with a σ-field $\mathcal{S}$. A mapping $g : S \to \mathbb{R}^p$ is said to be a *Borel* mapping if the conditions

$$g^{-1}(B) \in \mathcal{S}, \quad B \in \mathcal{B}(\mathbb{R}^p) \tag{9.9}$$

are all satisfied where

$$g^{-1}(B) \equiv \{s \in S : \ g(s) \in B\}.$$

Fact 9.6.1 *Let $\mathcal{G}$ denote a collection of subsets of $\mathbb{R}^p$ which generates the Borel σ-field $\mathcal{B}(\mathbb{R}^p)$, i.e.,*

$$\mathcal{B}(\mathbb{R}^p) = \sigma\left(\mathcal{G}\right). \tag{9.10}$$

It holds that the mapping $g : S \to \mathbb{R}^p$ is a Borel mapping if and only if the weaker *set of conditions*

$$g^{-1}(E) \in \mathcal{S}, \quad E \in \mathcal{G} \tag{9.11}$$

holds.

There are many generators known for the Borel σ-field $\mathcal{B}(\mathbb{R}^p)$. For instance, we have (9.10) with

- $\mathcal{G} = \mathcal{R}_{\text{open}}(\mathbb{R}^p)$ where $\mathcal{R}_{\text{open}}(\mathbb{R}^p)$ is the collection of all *finite open rectangles*, i.e.,

$$\mathcal{R}_{\text{open}}(\mathbb{R}^p) \equiv \left\{ I_1 \times \ldots \times I_p, \quad \begin{array}{c} I_k \in \mathcal{I}(\mathbb{R}) \\ k = 1, \ldots, p \end{array} \right\}$$

  where

$$\mathcal{I}(\mathbb{R}) = \{(a, b) : \ a, b \in \mathbb{R}\}$$

- $\mathcal{G} = \mathcal{R}_{\rm SW}(\mathbb{R}^p)$ where $\mathcal{R}_{\rm SW}(\mathbb{R}^p)$ is the collection of all *closed Southwest rectangles*, i.e.,

$$\mathcal{R}_{\rm SW}(\mathbb{R}^p) \equiv \left\{ \begin{array}{c} I_k = (-\infty, a_k] \\ I_1 \times \ldots \times I_p, \quad a_k \in \mathbb{R} \\ k = 1, \ldots, p \end{array} \right\}.$$

It follows from the discussion above that a mapping $g : S \to \mathbb{R}^p$ is a Borel mapping if the seemingly weaker conditions

$$\left\{ s \in S : \; g(s) \in \prod_{i=1}^{p} (-\infty, a_k] \right\} \in \mathcal{S}, \quad (a_1, \ldots, a_p) \in \mathbb{R}^p$$

all hold. Equivalently, a mapping $g : S \to \mathbb{R}^p$ is a Borel mapping if

$$\{ s \in S : \; g_k(s) \leq a_k, \; k = 1, \ldots, p \} \in \mathcal{S}, \quad (a_1, \ldots, a_p) \in \mathbb{R}^p$$

where it is understood that

$$g(s) = \left(g_1(s), \ldots, g_p(s) \right), \quad s \in S.$$

It is now plain that for each $k = 1, \ldots, p$, the component mapping $g_k : S \to \mathbb{R}$ is also a Borel mapping – Just take $a_\ell = \infty$ for all $\ell = 1, \ldots, k$ different from k. Conversely, since

$$\{ s \in S : \; g_k(s) \leq a_k, \; k = 1, \ldots, p \} = \cap_{k=1}^{p} \{ s \in S : \; g_k(s) \leq a_k \}$$

for arbitrary $(a_1, \ldots, a_p)$ in $\mathbb{R}^p$, we see that the mapping $g : S \to \mathbb{R}^p$ is a Borel mapping if and only if each of the component mappings $g_1 : S \to \mathbb{R}, \ldots, g_p : S \to \mathbb{R}$ is a Borel mapping.

Most (if not all) mappings $\mathbb{R}^p \to \mathbb{R}^q$ encountered in applications are Borel mappings. Furthermore, any *continuous* mapping $\mathbb{R}^p \to \mathbb{R}^q$ can be shown to be a Borel mapping!

Chapter 10

Useful facts from Probability Theory

10.1 Probability models

Probabilistic reasoning assumes the availability of a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ where: (i) The *sample space* Ω is the collection of all outcomes (samples) generated by the random experiment $\mathcal{E}$ of interest; (ii) Events are collections of outcomes, and the collection of events whose likelihood of occurrence can be defined is a *σ-field* $\mathcal{F}$ on Ω; and (iii) The "likelihood" of occurrence to events in $\mathcal{F}$ is assigned through a *probability measure* $\mathbb{P}$ defined on $\mathcal{F}$.

With Ω an arbitrary set, a non-empty collection $\mathcal{F}$ of subsets of Ω is a *σ-field* (also known as an *σ-algebra*) on Ω if $\mathcal{F}$ (i) contains the empty set $\emptyset$; (ii) is closed under complementarity: If $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$ (iii) is closed under countable union: With I a *countable* index set, if $E_i \in \mathcal{F}$ for each $i \in I$, then $\cup_{i \in I} E_i \in \mathcal{F}$.

The "likelihood" of occurrence to events in $\mathcal{F}$ is assigned through a *probability measure* $\mathbb{P}$ defined on $\mathcal{F}$.

A *probability (measure)* $\mathbb{P}$ on the σ-field $\mathcal{F}$ (or on $(\Omega, \mathcal{F})$) is a mapping $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that (i) $\mathbb{P}[\emptyset] = 0$ and $\mathbb{P}[\Omega] = 1$; (ii) σ-additivity: With I a countable index set, if $E_i \in \mathcal{F}$ for each $i \in I$, then

$$\mathbb{P}\left[\cup_{i \in I} E_i\right] = \sum_{i \in I} \mathbb{P}\left[E_i\right]$$

whenever the subsets $\{E_i, i \in I\}$ are *pairwise disjoint*, namely

$$E_i \cap E_j = \emptyset, \qquad \begin{array}{c} i \neq j \\ i, j \in I \end{array}$$

10.2 Random variables

All random variables (rvs) can always be thought as being defined on some given probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the sample space, $\mathcal{F}$ is a σ-field of events on Ω and $\mathbb{P}$ is a probability measure on $\mathcal{F}$.

Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, a mapping $X : \Omega \to \mathbb{R}^p$ is an ($\mathbb{R}^p$-valued) *random variable* (rv) if

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}, \quad B \in \mathcal{B}(\mathbb{R}^p).$$

In other words, the mapping $X : \Omega \to \mathbb{R}^p$ is a rv if it is a Borel mapping $X : \Omega \to \mathbb{R}^p$ – Here $S = \Omega$ and $\mathcal{S} = \mathcal{F}$. We shall often write $[X \in B]$ in lieu of $X^{-1}(B)$ and $\mathbb{P}[X \in B]$ for $\mathbb{P}[[X \in B]]$.

The *probability distribution (function)* $F_X : \mathbb{R}^p \to [0, 1]$ of the rv X is defined by

$$\begin{aligned} F_X(x) &\equiv \mathbb{P}[X \in (-\infty, x_1] \times \ldots \times (-\infty, x_p]] \\ &= \mathbb{P}[X_1 \leq x_1, \ldots, X_p \leq x_p], \quad x = (x_1, \ldots, x_p) \in \mathbb{R}^p. \end{aligned} \tag{10.1}$$

with the notation $X = (X_1, \ldots, X_p)$.

It turns out that there is as much probabilistic information in the probability distribution $F_X : \mathbb{R}^p \to [0, 1]$ of the rv X as in

$$\{\mathbb{P}[X \in B], \ B \in \mathcal{B}(\mathbb{R}^p)\}$$

In fact, knowledge of $F_X : \mathbb{R}^p \to \mathbb{R}$ allows a *unique* reconstruction of

$$\{\mathbb{P}[X \in B], \ B \in \mathcal{B}(\mathbb{R}^p)\}.$$

10.3 Probability distributions

Properties of F_X (Case $p = 1$): It is easy to see that the following properties hold:

- Monotonicity:
$$F_X(x) \leq F_X(y), \quad x, y \in \mathbb{R}$$
- Right-continuous:
$$\lim_{y \downarrow x} F_X(y) = F_X(x), \quad x \in \mathbb{R}$$
- Left limit exists:
$$\lim_{y \uparrow x} F_X(y) = F_X(x-) \quad \text{with} \quad \mathbb{P}[X = x] = F_X(y) - F_X(x-), \quad x \in \mathbb{R}$$
- Behavior at infinity: Monotonically
$$\lim_{x \to -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F_X(x) = 1$$

A *probability distribution (function)* on $\mathbb{R}$ is any mapping $F : \mathbb{R} \to [0, 1]$ such that

- Monotonicity:
$$F(x) \leq F(y), \quad x, y \in \mathbb{R}$$
- Right-continuous:
$$\lim_{y \downarrow x} F(y) = F(x), \quad x \in \mathbb{R}$$
- Left limit exists:
$$\lim_{y \uparrow x} F(y) = F(x-) \quad x \in \mathbb{R}$$
- Behavior at infinity: Monotonically
$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1$$

Any rv $X : \Omega \to \mathbb{R}$ generates a probability distribution function $F_X : \mathbb{R} \to [0,1]$. Conversely, for any probability distribution function $F : \mathbb{R} \to [0,1]$, there exists a probability triple $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ and a rv $\tilde{X} : \widetilde{\Omega} \to \mathbb{R}$ defined on it such that

$$\widetilde{\mathbb{P}}\left[\tilde{X} \leq x\right] = F(x), \quad x \in \mathbb{R}$$

This is the basis of Monte-Carlo simulation. There exists a multi-dimensional analog to this fact.

Proof. Take $\widetilde{\Omega} = [0,1]$, $\widetilde{\mathcal{F}} = \mathcal{B}([0,1])$ and $\widetilde{\mathbb{P}} = \lambda$. Define the rv $\tilde{X} : \widetilde{\Omega} \to \mathbb{R}$ by setting

$$\tilde{X}(\widetilde{\omega}) = F^{-}(\widetilde{\omega}), \quad \widetilde{\omega} \in [0,1]$$

where $F^{-} : [0,1] \to [-\infty, \infty]$ is the *generalized inverse* of F given by

$$F^{-}(u) = \inf\left(x \in \mathbb{R} : \ u \leq F(x)\right), \quad 0 \leq u \leq 1.$$

with the understanding that $F^{-}(u) = \infty$ if the defining set is empty, i.e., $F(x) < u$ for all x in $\mathbb{R}$.

Discrete distributions

A rv $X : \Omega \to \mathbb{R}^p$ is a *discrete* rv if there exists a *countable* subset $S \subseteq \mathbb{R}^p$ such that

$$\mathbb{P}\left[X \in S\right] = 1.$$

Note that

$$\mathbb{P}\left[X \in B\right] = \sum\nolimits_{x \in S \cap B} \mathbb{P}\left[X = x\right], \quad B \in \mathcal{B}(\mathbb{R}^p).$$

It is often more convenient to characterize the distributional properties of the rv X through its probability mass function (pmf) of the rv X given by

$$\boldsymbol{p}_X \equiv (p_X(x), \ x \in S)$$

with

$$p_X(x) = \mathbb{P}\left[X = x\right], \quad x \in S.$$

Absolutely continuous distributions

A rv $X : \Omega \to \mathbb{R}^p$ is an (absolutely) continuous rv if there exists a Borel mapping $f_X : \mathbb{R}^p \to \mathbb{R}_+$ such that

$$\mathbb{P}\left[X_i \leq x_i, \ i = 1, \ldots, p\right] = \int_{-\infty}^{x} f_X(\xi) d\xi, \quad x = (x_1, \ldots, x_p) \in \mathbb{R}^p.$$

Properties of F_X when $p \geq 1$

- Monotonicity needs to be modified and now reads

$$\mathbb{P}\left[x_k < X_k \leq y_k\right] \geq 0, \qquad \begin{array}{c} x_k < y_k \\ x_k, y_k \in \mathbb{R} \\ k = 1, \ldots, p \end{array}$$

  with the understanding that the quantity $\mathbb{P}\left[x_k < X_k \leq y_k\right]$ is expressed solely in terms of $F_X : \mathbb{R}^p \to [0, 1]$.

- Right-continuous:

$$\lim_{y \downarrow x} F_X(y) = F_X(x), \quad x \in \mathbb{R}^p$$

  with the understanding that $y_k \downarrow x_k$ for each $k = 1, \ldots, p$.

- Left limit exists:

$$\lim_{y \uparrow x} F_X(y) = F_X(x-) \quad \text{with} \quad \mathbb{P}\left[X = x\right] = F_X(y) - F_X(x-), \quad x \in \mathbb{R}^p$$

  with the understanding that $y_k \uparrow x_k$ for each $k = 1, \ldots, p$.

- Behavior at infinity:

$$\lim_{\min(x_k,\ k=1,\ldots,p) \to -\infty} F_X(x) = 0$$

  and

$$\lim_{\min(x_k,\ k=1,\ldots,p) \to \infty} F_X(x) = 1$$

Independence of rvs

Consider a collection of rvs $\{X_i,\ i \in I\}$ which are all defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that for each i in I, the rv X_i is a $\mathbb{R}^{p_i}$-valued rv for some positive integer p_i.

With I finite, we shall say that the rvs $\{X_i,\ i \in I\}$ are *mutually independent* if for each selection of B_i in $\mathcal{B}(R^{p_i})$ for each i in I, the events

$$\{[X_i \in B_i],\ i \in I\}$$

are mutually independent. It is easy to see that this is equivalent to requiring

$$\mathbb{P}\left[\cap_{i\in I}[X_i \in B_i]\right] = \prod_{i\in I} \mathbb{P}\left[X_i \in B_i\right], \qquad \begin{matrix} B_i \in \mathcal{B}(R^{p_i}) \\ i \in I. \end{matrix}$$

More generally, with I arbitrary (and possibly uncountable), the rvs $\{X_i,\ i \in I\}$ are said to be mutually independent if for every finite subset $J \subseteq I$, the rvs $\{X_j,\ j \in J\}$ are mutually independent!

Product spaces

Some facts: Consider two arbitrary sets Ω_a and Ω_b (possibly identical). Let $\mathcal{A}$ and $\mathcal{B}$ denote non-empty collections of subsets of Ω_a and Ω_b, respectively. While the collection $\mathcal{A} \times \mathcal{B}$ is usually *not* a σ-field on $\Omega_a \times \Omega_b$, even when $\mathcal{A}$ and $\mathcal{B}$ are themselves σ-fields, it can be shown that

$$\sigma\left(\mathcal{A} \times \mathcal{B}\right) = \sigma\left(\sigma\left(\mathcal{A}\right) \times \sigma\left(\mathcal{B}\right)\right).$$

10.4 Gaussian rvs

With scalar m and $\sigma^2 > 0$, the rv X is said to be a Gaussian rv with mean m and variance σ^2, written $X \sim \mathrm{N}(m, \sigma^2)$, if its cumulative probability distribution function is given by

$$\mathbb{P}\left[X \le x\right] = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{\xi - m}{\sigma}\right)^2} d\xi, \quad x \in \mathbb{R}.$$

The zero mean unit variance Gaussian rv is often referred to as a standard Gaussian rv; its probability density function $\varphi : \mathbb{R} \to \mathbb{R}_+$ is given by

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}, \tag{10.2}$$

and its cumulative probability distribution function is then

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt, \quad x \in \mathbb{R}. \tag{10.3}$$

Obviously, if $X \sim \mathrm{N}(m, \sigma^2)$ and $Z \sim \mathrm{N}(0, 1)$, then X and $m + \sigma Z$ have the same distribution.

In the context of digital communications, it is customary to use the Q-function $Q : \mathbb{R}_+ \to [0, 1]$ given by

$$Q(x) = 1 - \Phi(x) = \int_{x}^{\infty} \varphi(t)dt, \quad x \ge 0.$$

Chapter 11

The classical limit theorems

The setting of the next four sections is as follows: The rvs $\{X_n,\ n = 1, 2, \ldots\}$ are rvs defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. With this sequence we associate the sums

$$S_n = \sum_{k=1}^{n} X_k, \quad n = 1, 2, \ldots$$

Two types of results will be discussed: The first class of results are known as Laws of Large Numbers; they deal with the convergence of the sample averages

$$\bar{S}_n = \frac{1}{n} \sum_{k=1}^{n} X_k, \quad n = 1, 2, \ldots$$

The second class of results are called Central Limit Theorems and provide a rate of convergence in the Laws Large Numbers.

11.1 Weak Laws of Large Numbers (I)

Laws of Large Numbers come in two types which are distinguished by the mode of convergence used. When convergence in probability is used, we refer to such results as weak Laws of Large Numbers. The most basic such results is given first.

Theorem 11.1.1 *Assume the rvs* $\{X, X_n,\ n = 1, 2, \ldots\}$ *to be i.i.d. rvs with* $\mathbb{E}\left[|X|^2\right] < \infty$. *Then,*

$$\frac{S_n}{n} \xrightarrow{P}_n \mathbb{E}\left[X\right]. \tag{11.1}$$

As we now show, the finiteness of the second moment of X can be dropped.

Theorem 11.1.2 *Assume the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ to be i.i.d. rvs with $\mathbb{E}[|X|] < \infty$. Then, we have*

$$\frac{S_n}{n} \xrightarrow{P}_n \mathbb{E}[X]. \tag{11.2}$$

11.2 The Strong Law of Large Numbers

Strong Laws of Large Numbers are given are convergence statements in the a.s. sense. The classical Strong Law of Large Numbers is given next.

Theorem 11.2.1 *Assume the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ to be i.i.d. rvs with $\mathbb{E}[|X|] < \infty$. Then,*

$$\lim_{n\to\infty} \frac{S_n}{n} = \mathbb{E}[X] \quad a.s. \tag{11.3}$$

11.3 The Central Limit Theorem

The Central Limit Theorem completes the Law of Large Numbers, in that it provides some indication as to the rate at which convergence takes place.

Theorem 11.3.1 *Assume the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ to be i.i.d. rvs with $\mathbb{E}[|X|^2] < \infty$. Then, we have*

$$\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}[X]\right) \Longrightarrow_n \sqrt{\mathrm{Var}[X]} \cdot U \tag{11.4}$$

where U is standard zero-mean unit-variance Gausssian rv.