1 Introduction

In this chapter we introduce the problems of detection and estimation, and cast them in the general framework of *statistical decision theory*, using the notions of states, observations, actions, costs, and optimal decision making. We then introduce the Bayesian approach as the optimal approach in the case of random states. For nonrandom states, the minimax approach and other non-Bayesian approaches are introduced.

1.1 Background

This book is intended to be accessible to graduate students and researchers that have completed a first-year graduate course in probability and random processes. Familiarity with the notions of convergence of random sequences (in probability, almost surely, and in the mean square sense) is needed in the sections of this text that deal with asymptotics. Elementary knowledge of measure theory is helpful but not required. An excellent reference for this background material is Hajek's book [1]. Familiarity with basic tools from matrix analysis [2] and optimization [3] is also assumed.

Textbooks on detection and estimation include the classic book by Van Trees [4], and the more recent books by Poor [5], Kay [6], and Levy [7]. For a thorough treatment of the subject, the reader is referred to the classic books by Lehman [8] and Lehman and Casella [9]. While these books treat detection and estimation as distinct but loosely related topics, we stress in this text the tight connection between the underlying concepts, via Wald's statistical decision theory [10, 11, 12].

The subject matter is also related to statistical learning theory and to information theory. Statistical learning theory deals with unknown probability distributions and with the construction of decision rules whose performance approaches that of rules that know the underlying distributions. Performance analysis of the latter rules is of central interest in this book, and many of the analytic tools are also useful in learning theory [13]. Finally, the connection to information theory appears via the role of Kullback–Leibler divergence between probability distributions, Fisher information, sufficient statistics, and information geometry [14].

1.2 Notation

The following notation is used throughout this book. Random variables are denoted by uppercase letters (e.g., Y), their individual values by lowercase letters (e.g., y), and the

set of values by script letters (e.g., $\mathcal{Y}$). We use *sans serif* notation for matrices (e.g., A); the identity matrix is denoted by I. The indicator function of a set $\mathcal{A}$ is denoted by $\mathbb{1}_{\mathcal{A}}$, i.e.,

$$\mathbb{1}_{\mathcal{A}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{A} \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

For simplicity of the exposition, we shall mostly be interested in discrete random variables and in continuous random variables over Euclidean spaces. In some cases where we need to distinguish between scalar and vector variables, we use boldface for vectors (e.g., $\boldsymbol{Y} = [Y_1 \; Y_2]^\top$). The reader is referred to [1] for a thorough introduction to the topics in this section.

1.2.1 Probability Distributions

Consider a random variable Y taking its values in a set $\mathcal{Y}$ endowed with a σ-algebra $\mathscr{F}$. A probability measure P is a real-valued function on $\mathscr{F}$ that satisfies Kolmogorov's three axioms of probability: nonnegativity, unit measure, and σ-additivity.

If $\mathcal{Y}$ is (finitely or infinitely) countable, $\mathscr{F}$ is the collection of all subsets of $\mathcal{Y}$, and we denote by $\{p(y),\ y \in \mathcal{Y}\}$ a probability mass function (pmf) on $\mathcal{Y}$. The probability of a set $\mathcal{A} \subset \mathcal{Y}$ (hence $\mathcal{A} \in \mathscr{F}$) is $\mathrm{P}(\mathcal{A}) = \sum_{y \in \mathcal{A}} p(y)$. If $\mathcal{Y}$ is the n-dimensional Euclidean space $\mathbb{R}^n$ (with $n \geq 1$), we choose $\mathscr{F} = \mathscr{B}(\mathbb{R}^n)$, the Borel σ-algebra which contains all n-dimensional rectangles and all finite or infinite unions of such rectangles. We then assume that Y is a continuous random variable, i.e., it has a probability density function (pdf) which we denote by $\{p(y),\ y \in \mathcal{Y}\}$. The probability of a Borel set $\mathcal{A} \in \mathscr{B}(\mathbb{R}^n)$ is then $\mathrm{P}(\mathcal{A}) = \int_{\mathcal{A}} p(y)\, dy$. The cumulative distribution function (cdf) of Y is given by

$$F(y) = \mathrm{P}(\mathcal{A}_y), \tag{1.2}$$

where $\mathcal{A}_y = \{y' \in \mathcal{Y} \;:\; y'_i \leq y_i,\ 1 \leq i \leq n\}$ is an n-dimensional orthant. To summarize:

$$\mathrm{P}(\mathcal{A}) = \begin{cases} \sum_{y \in \mathcal{A}} p(y) & : \text{discrete } \mathcal{Y} \\ \int_{\mathcal{A}} p(y)\, dy & : \text{continuous } \mathcal{Y}. \end{cases} \tag{1.3}$$

1.2.2 Conditional Probability Distributions

A conditional pmf of a random variable Y is a collection of pmfs $\{p_x(y),\ y \in \mathcal{Y}\}$ indexed by a conditioning variable $x \in \mathcal{X}$. Note that x is not necessarily interpreted as a realization of random variable X, and in that sense, $p_x(y)$ is different from the more traditional conditional distribution of a random variable Y, given another random variable X. Similarly, a conditional pdf is a collection of pdfs $\{p_x(y),\ y \in \mathcal{Y}\}$ indexed by a conditioning variable $x \in \mathcal{X}$. The conditional probability of a Borel set $\mathcal{A} \in \mathscr{B}(\mathbb{R}^n)$ is $\mathrm{P}_x(\mathcal{A}) = \int_{\mathcal{A}} p_x(y)\, dy$. The conditional cdf is denoted by $F_x(y) = \mathrm{P}_x(\mathcal{A}_y)$ where $\mathcal{A}_y$ is the orthant defined in Section 1.2.1. For both the discrete and the continuous cases, we may write $p(y|x)$ instead of $p_x(y)$, especially in the case that x is a realization of a random variable X.

1.2.3 Expectations and Conditional Expectations

The expectation of a function $g(Y)$ of a random variable Y is given by

$$\mathbb{E}[g(Y)] = \begin{cases} \sum_{y\in\mathcal{Y}} p(y)g(y) & : \text{discrete } \mathcal{Y} \\ \int_{\mathcal{Y}} p(y)g(y)\,dy & : \text{continuous } \mathcal{Y}. \end{cases} \tag{1.4}$$

The expectation may be viewed as a linear function of the pmf p in the discrete case, and a linear functional of the pdf p in the continuous case.[1] The conditional expectation of $g(Y)$ given x is denoted by $\mathbb{E}_x[g(Y)]$ and takes the form

$$\mathbb{E}_x[g(Y)] = \begin{cases} \sum_{y\in\mathcal{Y}} p_x(y)g(y) & : \text{discrete } \mathcal{Y} \\ \int_{\mathcal{Y}} p_x(y)g(y)\,dy & : \text{continuous } \mathcal{Y}. \end{cases} \tag{1.5}$$

1.2.4 Unified Notation

The cases of discrete and continuous Y can be handled in a unified way, avoiding duplication of formulas. Indeed the probability of a set $\mathcal{A} \in \mathscr{F}$ may be written as the Lebesgue–Stieltjes integral

$$\mathrm{P}(\mathcal{A}) = \int_{\mathcal{A}} p(y)\,d\mu(y), \tag{1.6}$$

where μ is a finite measure on $\mathscr{F}$, equal to the Lebesgue measure in the continuous case ($d\mu(y) = dy$, or equivalently $\mu(\mathcal{A}) = \int_{\mathcal{A}} dy$ for all $\mathcal{A} \in \mathscr{F}$) and to the counting measure in the discrete case.[2] The expectation of a function $g(Y)$ is likewise given by

$$\mathbb{E}[g(Y)] = \int_{\mathcal{Y}} p(y)g(y)\,d\mu(y). \tag{1.7}$$

For conditional expectations we have

$$\mathbb{E}_x[g(Y)] = \int_{\mathcal{Y}} p_x(y)g(y)\,d\mu(y), \quad \forall x \in \mathcal{X}. \tag{1.8}$$

1.2.5 General Random Variables

The expressions defined in the previous sections can be generalized to mixed discrete-continuous random variables using the Lebesgue–Stieltjes formalism. Indeed let Y be a random variable over $\mathbb{R}^n$ with cdf F defined in (1.2). Then the probability of a Borel set $\mathcal{A}$ is

$$\mathrm{P}(\mathcal{A}) = \int_{\mathcal{A}} dF(y) \tag{1.9}$$

(the Lebesgue–Stieltjes integral) and the expectation of a function $g(Y)$ is

$$\mathbb{E}[g(Y)] = \int_{\mathcal{Y}} g(y)dF(y). \tag{1.10}$$

[1] Typical functions $g(Y)$ of interest are polynomials of Y (the expected values thereof are moments of the distribution) and indicator functions of sets (the expected values thereof are probabilities of said sets).

[2] The counting measure on a set of points $y_1, y_2, \ldots$ is defined as $\mu(\mathcal{A}) = \sum_i \mathbb{1}\{y_i \in \mathcal{A}\}$.

The random variable Y has a density with respect to a measure μ which is the sum of Lebesgue measure and a counting measure, in which case the expressions (1.6) and (1.7) hold and allow simple calculations.[3] The same concept applies to conditional probabilities and conditional expectations, with F replaced by the conditional cdf F_x. Finally, if $n \geq 2$ and the distribution P assigns positive probabilities to subsets of several dimensions, then Y is not of the mixed discrete-continuous type but (1.9) and (1.10) hold, and so do the expressions (1.6) and (1.7) using a suitable μ.

For any pair of random variables (X, Y), assuming $f(y) = \mathbb{E}[X|Y = y]$ exists for each $y \in \mathcal{Y}$, let $\mathbb{E}[X|Y] \triangleq f(Y)$ which is a function of Y. The law of iterated expectations states that $\mathbb{E}[X] = \mathbb{E}(\mathbb{E}[X|Y])$, which will be very convenient throughout this text. In particular, this law may be applied to evaluate the expectation of a function $g(X, Y)$ as $\mathbb{E}[g(X, Y)] = \mathbb{E}(\mathbb{E}[g(X, Y)|Y])$.

The (scalar) Gaussian random variable with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$. The Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix C is denoted by $\mathcal{N}(\boldsymbol{\mu}, \mathsf{C})$.

1.3 Statistical Inference

The detection and estimation problems treated in this course are instances of the following general statistical inference problem. Given *observations* y taking their value in some data space $\mathcal{Y}$, infer (estimate) some unknown *state* x taking its value in a state space $\mathcal{X}$. The observations are noisy. More precisely, they are stochastically related to the state via a *conditional probability distribution* $\{\mathrm{P}_x, x \in \mathcal{X}\}$, i.e., a collection of distributions on $\mathcal{Y}$ indexed by $x \in \mathcal{X}$. Note that no assumption of randomness of the state is made at this point. The key problems are to construct a *good* estimator and to characterize its performance. This raises the following central issues:

- What performance criteria should be used?
- Can we derive an estimator that is *optimal* under some desirable performance criterion?
- Is such an estimator computationally tractable?
- If not, can the optimization be restricted to a useful class of tractable estimators?

The theory applies to a remarkable variety of statistical inference problems covering applications in diverse areas (communications, signal processing, control, life sciences, economics, etc.).

[3] For instance, let $\mathcal{Y}$ be the interval $[0, 1]$ and consider the cdf $F_0(y) = y\mathbb{1}\{0 \leq y \leq 1\}$ corresponding to the uniform distribution on $\mathcal{Y}$, the cdf $F_1(y) = \mathbb{1}\{\frac{1}{3} \leq y \leq 1\}$ corresponding to the degenerate distribution that assigns all mass at the point $y = \frac{1}{3}$, and the cdf $F_2(y) = \frac{1}{2}[F_0(y) + F_1(y)]$ corresponding to the mixture of the continuous and discrete random variables above. The means μ_0 and μ_1 of the continuous and discrete random variables are equal to $\frac{1}{2}$ and $\frac{1}{3}$, respectively. The mean of the mixture distribution is $\mu_3 = \frac{1}{2}[\mu_1 + \mu_2] = \frac{5}{12}$.

1.3.1 Statistical Model

The statistical model is a triple $(\mathcal{X}, \mathcal{Y}, \{P_x, x \in \mathcal{X}\})$ consisting of a state space $\mathcal{X}$, an observation space $\mathcal{Y}$, and a family of conditional distributions P_x on $\mathcal{Y}$, indexed by $x \in \mathcal{X}$. The state space may be a continuum or a discrete set, in which cases the inference problem is conventionally called estimation or detection, respectively.

Estimation: Often the state $x \in \mathcal{X}$ is a scalar parameter, e.g., the unknown attenuation factor of a transmission system, a temperature, or some other physical quantity. More generally x could be a vector-valued parameter, e.g., x represents location in two- or three-dimensional Euclidean space, or x represents a set of individual states associated with sensors at different physical locations. Another example of vector-valued x is a discrete-time signal (speech, geophysical, etc.). Yet another version of this problem involves functions defined over some compact set, e.g., x is a continuous-time signal over a finite time window $[0, T]$. The notion of time can be extended to space, and so x could also be a two- or three-dimensional discrete-space signal, or a continuous-space signal. In all cases, the parameter space $\mathcal{X}$ is a continuum (an uncountably infinite set). In this text, we will limit our attention to the m-dimensional Euclidean space $\mathcal{X} = \mathbb{R}^m$. Many of the key insights are already apparent in the scalar case $(m = 1)$.

Detection: In other problems, the state x is a discrete quantity, e.g., a bit in a communication system $(\mathcal{X} = \{0, 1\})$, or a sequence of bits $(\mathcal{X} = \{0, 1\}^n)$ where n is the length of the binary sequence. Alternatively, x could take one of m values for a signal classification problem (e.g., x represents a word in a speech recognition system, or a person in a biometric verification system). The observed signal comes from one of m possible classes or categories.

Similarly, observations can come in a variety of formats: the data space $\mathcal{Y}$ could be:

- a finite set, e.g., the observation is a single bit $y \in \mathcal{Y} = \{0, 1\}$ in a data communication system, or a sequence of bits;
- a countably infinite set, e.g., the set of all binary sequences, $\mathcal{Y} = \{0, 1\}^*$;
- an uncountably infinite set, such as $\mathbb{R}$ for scalar observations, or $\mathbb{R}^m$ for vector-valued observations, where $m > 1$, or the space $L^2[0, T]$ of finite-energy analog signals defined over the time window $[0, T]$.

In many problems, the observations form a length-n sequence whose elements take their value in a common space $\mathcal{Y}_1$. Then $\mathcal{Y} = \mathcal{Y}_1^n$ is the n-fold product of $\mathcal{Y}_1$, and we use the boldface notation $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ to represent its elements. The sequence $\boldsymbol{Y}$ often follows the product conditional distribution $p_x(\boldsymbol{y}) = \prod_{i=1}^n p_{Y_i|X=x}(y_i)$, given the state x.

The estimation problem consists of designing a function $\hat{x} : \mathcal{Y} \to \mathcal{X}$ (*estimator*) that produces an *estimate* $\hat{x}(y) \in \mathcal{X}$, given observations $y \in \mathcal{Y}$. The detection problem admits the same formulation, the only difference being that $\mathcal{X}$ is a discrete set instead of a continuum. Constraints such as linearity can be imposed on the design of the estimator.

1.3.2 Some Generic Estimation Problems

1. Parameter estimation: estimate the mean μ and variance σ^2 of a Gaussian distribution from a sequence of observations Y_i, $1 \leq i \leq n$ drawn i.i.d. (independent and identically distributed) $\mathcal{N}(\mu, \sigma^2)$. Here $\mathcal{Y} = \mathbb{R}^n$ and $\mathcal{X} = \mathbb{R} \times \mathbb{R}^+$. A simple estimator of the mean and variance would be the so-called sample mean $\hat{\mu}(\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^n Y_i$ and sample variance $\hat{\sigma}^2(\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\mu}(\boldsymbol{Y}))^2$ which are respectively linear and quadratic functions of the data.
2. Probability mass function estimation: estimate a pmf $p = \{p(1), p(2), \ldots, p(k)\}$ from a sequence of observations Y_i, $1 \leq i \leq n$ drawn i.i.d. p. Here $\mathcal{Y} = \{1, \ldots, k\}^n$ and $\mathcal{X}$ is the probability simplex in $\mathbb{R}^k$. A simple estimator would be the empirical pmf $\hat{p}(l) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i = l\}$ for $l = 1, 2, \ldots, k$.
3. Estimation of signal in i.i.d. noise: $Y_i = s_i + Z_i$ for $i = 1, 2, \ldots, n$ where Z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. Here $\mathcal{Y} = \mathcal{X} = \mathbb{R}^n$. A simple estimator would be $\hat{S}_i = cY_i$, where $c \in \mathbb{R}$ is a weight to be optimized.
4. Parametric estimation of signal in i.i.d. noise: $Y_i = s_i(\theta) + Z_i$ for $i = 1, 2, \ldots, n$ where $\{s_i(\theta)\}_{i=1}^n$ is a sequence parameterized by $\theta \in \mathcal{X}$, and Z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. The unknown parameter could be estimated using the nonlinear least-squares estimator $\hat{\theta} = \text{argmin}_{\theta \in \mathcal{X}} \sum_i (Y_i - s_i(\theta))^2$.
5. Signal estimation, prediction and smoothing (Figure 1.1): $Y_i = (h \star s)_i + Z_i$ for $i = 1, 2, \ldots, n$, where Z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$, and $h \star s$ denotes the convolution of the sequence $\boldsymbol{s}$ with the impulse response of the linear system. A candidate estimator would be $\hat{S}_i = (g \star Y)_i$, where g is the impulse response of an estimation filter. In a prediction problem, a filter is designed to estimate a future sample of the signal (e.g., estimate s_{i+1}). In a smoothing problem, a filter is designed to estimate past values of the signal.
6. Image denoising: $Y_{ij} = s_{ij} + Z_{ij}$ for $i = 1, 2, \ldots, n_1$ and $j = 1, 2, \ldots, n_2$ and Z_{ij} are i.i.d. $\mathcal{N}(0, \sigma^2)$.
7. Estimation of a continuous-time signal: $Y(t) = s(t) + Z(t)$ for $0 \leq t \leq T$ where x belongs to a separable Hilbert space such as $L^2[0, T]$, and Z is stationary Gaussian noise with mean zero and covariance function $R(t)$, $t \in \mathbb{R}$.

1.3.3 Some Generic Detection Problems

1. Binary hypothesis testing: under hypothesis H_0, the observations are a sequence Y_i, $i = 1, 2, \ldots, n$ drawn i.i.d. p_0. Under the rival hypothesis H_1, the observations are drawn i.i.d. p_1.
2. Signal detection in i.i.d. Gaussian noise: $Y_i = \theta s_i + Z_i$ for $i = 1, 2, \ldots, n$ where $\boldsymbol{s} \in \mathbb{R}^n$ is a known sequence, $\theta \in \{0, 1\}$, and Z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$.

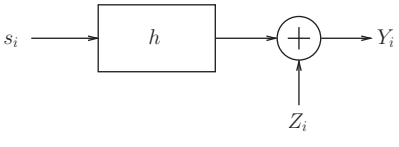

Figure 1.1 Signal prediction and smoothing

3. M-ary signal classification in i.i.d. Gaussian noise: $Y_i = s_i(\theta) + Z_i$ for $i = 1, 2, \ldots, n$ where $\boldsymbol{s}(\theta) \in \mathbb{R}^n$ is a sequence parameterized by $\theta \in \{0, 1, \ldots, M-1\}$, and Z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. This generalizes Problem #2 above. A simple detector is the correlation detector, the output $\hat{\theta}$ of which maximizes the correlation score $T(\theta) = \sum_{i=1}^n Y_i\, s_i(\theta)$ over all $\theta \in \{0, 1, \ldots, M-1\}$.
4. Composite hypothesis testing in i.i.d. Gaussian noise: $Y_i = \alpha s_i(\theta) + Z_i$ for $i = 1, 2, \ldots, n$ where the multiplier $\alpha \in \{0, 1\}$, the sequence $\boldsymbol{s}(\theta) \in \mathbb{R}^n$ is a function of some nuisance parameter $\theta \in \mathbb{R}^m$, and Z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. Here only α is of interest; it is possible (but not required) to estimate θ as a means to solve the detection problem.

1.4 Performance Analysis

For estimation problems, various metrics can be used to measure the closeness of the estimator $\hat{x}$ to the original x. For instance, let $\mathcal{X} = \mathbb{R}^m$ and consider:

- squared error $C(x, \hat{x}) = \sum_{i=1}^m (\hat{x}_i - x_i)^2$;
- absolute error $C(x, \hat{x}) = \sum_{i=1}^m |\hat{x}_i - x_i|$.

The squared-error criterion is more tractable but penalizes large errors heavily. This could be desirable in some problems but undesirable in others (e.g., in the presence of outliers in the data). In such cases, the absolute-error criterion might be preferable.

For detection problems, one may consider:

- for $\mathcal{X} = \{0, 1, \ldots, m-1\}$: zero-one loss $\mathbb{1}\{\hat{x} \neq x\}$;
- for $\mathcal{X} = \{0, 1\}^d$ (space of length-d binary sequences): Hamming distance $d_H(\hat{x}, x) = \sum_{i=1}^d \mathbb{1}\{\hat{x}_i \neq x_i\}$. The Hamming distance, normalized by d, is the bit error rate (BER) in a digital communication system.

In both cases, the performance metric can be evaluated on a specific x, on a specific y, or in some average sense. The average value of the performance metric is often used as a design criterion for the estimation (detection) system, as discussed next.

1.5 Statistical Decision Theory

Detection and estimation problems fall under the umbrella of Abraham Wald's statistical decision theory [10, 11], where the goal is to make a right (optimal) choice from a set of alternatives in a noisy environment. A general framework for statistical decision theory is depicted in Figure 1.2. The first block represents the observational model which is governed by the conditional distribution $p_x(y)$ of the observations given the state x. The second block is a decision rule δ that outputs an action $a = \delta(y)$. Each possible action a incurs a cost $C(a, x)$, to be minimized in some sense.[4]

[4] The cost is also called *loss*. Alternatively, one could take a more positive view and specify a *utility function* $U(a, x)$, to be maximized. Both problem formulations are equivalent if one specifies $U(a, x) = b - C(a, x)$ for some constant b.

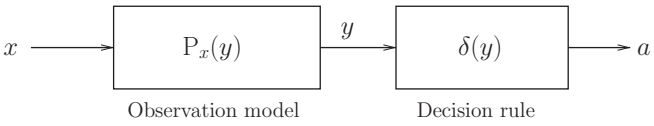

Figure 1.2 Statistical decision-making framework

Formally, there are six basic ingredients in a typical decision theory problem:

1. $\mathcal{X}$: the set of states. For detection problems, the number of states is finite, i.e., $|\mathcal{X}| = m < \infty$. For binary detection problems, we usually let $\mathcal{X} = \{0, 1\}$. We denote a typical state for detection problems by the variable $j \in \mathcal{X}$. We denote a typical state for estimation problems by the variable $x \in \mathcal{X}$.
2. $\mathcal{A}$: the set of actions, or possible decisions about the state. In most detection and estimation problems, $\mathcal{A} = \mathcal{X}$. However, in problems such as decoding with an erasure option, $\mathcal{A}$ could be larger than $\mathcal{X}$. In other problems such as composite hypothesis testing, $\mathcal{A}$ is smaller than $\mathcal{X}$. We denote a typical decision by the variable $a \in \mathcal{A}$.
3. $C(a, x)$: the cost of taking action a when the state of Nature is x. Typically $C(a, x) \geq 0$, and the cost is to be minimized in some suitable sense. Quantifying the costs incurred from decisions allows us to optimize the decision rule. An example of cost function that is relevant in many applications is the *uniform* cost function for which $|\mathcal{X}| = |\mathcal{A}| = m < \infty$ and
$$C(a, x) = \begin{cases} 0 & \text{if } a = x \\ 1 & \text{if } a \neq x, \end{cases} \qquad x \in \mathcal{X}, a \in \mathcal{A}. \tag{1.11}$$
4. $\mathcal{Y}$: the set of observations. The decision is made based on some *random* observation Y taking values in $\mathcal{Y}$. The observations could be continuous (with $\mathcal{Y} \subset \mathbb{R}^n$) or discrete.
5. $\{P_x, x \in \mathcal{X}\}$: the observational model, a family of probability distributions on $\mathcal{Y}$ conditioned on the state of Nature x.
6. $\mathcal{D}$: the set of decision rules or tests. A decision rule is a mapping $\delta : \mathcal{Y} \mapsto \mathcal{A}$ that associates an action $a = \delta(y)$ to each possible observation $y \in \mathcal{Y}$.

Detection problems are also referred to as *hypothesis testing* problems, with the understanding that each state corresponds to a hypothesis about the nature of the observations. The hypothesis corresponding to state j is denoted by H_j.

1.5.1 Conditional Risk and Optimal Decision Rules

The cost associated with a decision rule $\delta \in \mathcal{D}$ is a random quantity (because Y is random) given by $C(\delta(Y), x)$. Therefore, to *order* decision rules according to their "merit" we use the quantity
$$R_x(\delta) \triangleq \mathbb{E}_x[C(\delta(Y), x)] = \int_{\mathcal{Y}} p_x(y) C(\delta(y), x) d\mu(y), \tag{1.12}$$
which we call the *conditional risk* associated with δ when the state is x.

The conditional risk function can be used to obtain a (partial) ordering of the decision rules in $\mathcal{D}$, in the following sense.

Table 1.1 Decision rules and conditional risks for Example 1.1

δ	$y=1$	$y=2$	$y=3$	$R_0(\delta)$	$R_1(\delta)$
δ_1	0	0	0	0	1
δ_2	0	0	1	0	0.5
δ_3	0	1	0	0	0.5
$\boxed{\delta_4}$	0	1	1	$\boxed{0}$	$\boxed{0}$
δ_5	1	0	0	1	1
δ_6	1	0	1	1	0.5
δ_7	1	1	0	1	0.5
δ_8	1	1	1	1	0

DEFINITION 1.1 A decision rule δ is *better*[5] than decision rule δ' if

$$R_x(\delta) \leq R_x(\delta'), \quad \forall x \in \mathcal{X}$$

and

$$R_x(\delta) < R_x(\delta') \text{ for at least one } x \in \mathcal{X}.$$

Sometimes it may be possible to find a decision rule $\delta^\star \in \mathcal{D}$ which is better than any other $\delta \in \mathcal{D}$. In this case, the statistical decision problem is solved. Unfortunately, this usually happens only for trivial cases as in the following example.

Example 1.1 Suppose $\mathcal{X} = \mathcal{A} = \{0, 1\}$ with the uniform cost function as in (1.11). Furthermore suppose the observation Y takes values in the set $\mathcal{Y} = \{1, 2, 3\}$ and the conditional pmfs of Y are:

$$p_0(1) = 1,\ p_0(2) = p_0(3) = 0, \qquad p_1(1) = 0,\ p_1(2) = p_1(3) = 0.5.$$

Then it is easy to see that we have the conditional risks for the eight possible decision rules depicted in Table 1.1. Clearly, δ_4 is the best rule according to Definition 1.1, but this happens only because the conditional pmfs p_0 and p_1 have disjoint supports (see also Exercise 2.1).

Since conditional risks cannot be used directly in finding optimal solutions to statistical decision making problems except in trivial cases, there are two general approaches for finding optimal decision rules: *Bayesian* and *minimax*.

1.5.2 Bayesian Approach

If the state of Nature is random with known *prior* distribution $\pi(x)$, $x \in \mathcal{X}$, then one can define the *average* risk or *Bayes* risk associated with a decision rule δ, which is given by

$$r(\delta) \triangleq \mathbb{E}\,[R_X(\delta)] = \mathbb{E}[C(\delta(Y), X)]. \tag{1.13}$$

[5] If δ dominates δ' as in this definition, the decision rule δ' is sometimes said to be *inadmissible* [8] for the statistical inference problem.

More explicitly,

$$r(\delta) = \begin{cases} \int_{\mathcal{X}} \int_{\mathcal{Y}} \pi(x) p_x(y) C(\delta(y), x) dy dx & : \text{continuous } X, Y \\ \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \pi(x) p_x(y) C(\delta(y), x) & : \text{discrete } X, Y, \end{cases} \tag{1.14}$$

which may also be written using the unified notation

$$r(\delta) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \pi(x) p_x(y) C(\delta(y), x) d\mu(y) d\nu(x) \tag{1.15}$$

for appropriate measures μ and ν on $\mathcal{Y}$ and $\mathcal{X}$, respectively.

The optimal decision rule, in the Bayesian framework, is the one that minimizes the Bayes risk:

$$\delta_{\mathrm{B}} = \arg\min_{\delta \in \mathcal{D}} r(\delta), \tag{1.16}$$

and is known as the Bayesian decision rule. The minimizer in (1.16) might not be unique; all such minimizers are Bayes rules.

1.5.3 Minimax Approach

What if we are not given a prior distribution on the set $\mathcal{X}$? We could postulate a distribution on $\mathcal{X}$ (for example, a uniform distribution) and use the Bayesian approach. On the other hand, one may want to guarantee a certain level of performance for all choices of state. In this case, we use a minimax approach. To this end, we define the maximum (or worst-case) risk (see Figure 1.3(a)):

$$R_{\max}(\delta) \triangleq \max_{x \in \mathcal{X}} R_x(\delta). \tag{1.17}$$

The minimax decision rule minimizes the worst-case risk:

$$\begin{aligned} \delta_{\mathrm{m}} &\triangleq \arg\min_{\delta \in \mathcal{D}} R_{\max}(\delta) \\ &= \arg\min_{\delta \in \mathcal{D}} \max_{x \in \mathcal{X}} R_x(\delta). \end{aligned} \tag{1.18}$$

The minimizer in (1.18) might not be unique; all such minimizers are minimax rules.

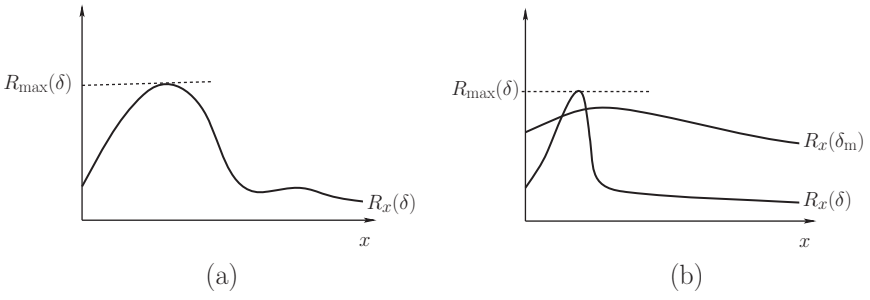

Figure 1.3 Minimax risk: (a) conditional risk and maximum risk for decision rule δ; (b) example where minimax rule is pessimistic