I

Decision and Estimation

Es irrt der Mensch so lang er strebt.

Goethe, *Faust*

Information about the world is acquired by observation and measurement, the results of which are subject to error. One would like to think it could be eliminated if only one built elaborate enough instruments and took sufficient pains. All efforts to be free of error will, however, finally reach a bound set by nature's underlying chaos, which infects all observations of physical phenomena with an innate uncertainty. We shall seek out that ultimate limit in a domain where observations and their inevitable errors can be analyzed with some clarity and ease, the domain of optics. Although optical astronomy, communications, and radar furnish our simplest paradigms, the basic concepts to be set forth have universal applicability.

Acquiring information about a physical entity or system involves either decision or estimation. Either one must decide which of a set of statements, or *hypotheses*, best describes the system insofar as observations permit one to judge; or one must estimate the values of certain quantities, or *parameters*, characterizing it. These hypotheses and parameters exist and signify in the context of some theory about the system. That theory also describes the sources

of the uncertainty or error that corrupts the observations. The irreducibly minimum component of error in decisions and estimates is discovered by analyzing the decision or estimation procedures that minimize some convenient measure of the average amount of error. In the optical domain there are two basic sources of uncertainty: the atomic constitution of matter and the quantum-mechanical nature of light.

Let us begin with an example. Inquiring whether there is a star at a certain point of the sky, you direct a telescope toward it, photograph the point and its surroundings, develop the photographic plate, and look to see whether there is a stellar image at the point in question. If a bright star is there, all well and good; but if the star is faint, or absent altogether, it will be difficult to decide whether one is there or not. Examining the plate under magnification, you perceive that many of the grains in the emulsion have been blackened by stray light from the sky. Is there a sufficient additional concentration of developed grains in the expected pattern to justify deciding that a star is indeed present? Your decision, whether "yes" or "no," is liable to error. How can you make it so that the chance of error is as small as possible? That is the subject of statistical decision theory, or as it is called in contexts like this, *detection theory*.

If your telescope had a larger aperture, or if you could expose the photographic plate for a longer time, collecting more light from that hypothetical star, your decision could be made with greater assurance. Let us presume in what follows that the size of the aperture and the duration of the observation interval are fixed. If committed to photography, you must adopt some procedure for analyzing the distribution of developed grains in the plate. The procedure may even involve elaborate calculations by a digital computer, but it will ultimately lead to a definite choice between two hypotheses: H_0, "no star exists at the point in question," and H_1, "a star does exist there." Such a procedure is called a *strategy*. It can suffer two kinds of error, erroneously deciding that a star is present, or failing to report a star that is really there. Imagine repeating the strategy many times under the same general conditions, with the same background illuminance, photographic materials, and processing. The relative frequencies or probabilities of the two kinds of error will depend on the nature of the strategy, the magnitude of the star, the optical organization of the telescope, and the statistical properties of the light and of the granularity in the plate. In principle those error probabilities can be calculated. Fixing the probability of the first kind of error, or "false alarm," at some tolerable value, you may seek a decision strategy for which the probability of the second kind of error, or "false dismissal," is as small as possible. Statistical decision theory, which we shall introduce in Chapter II, shows how to go about this.

Perhaps there is a better way than photography. You might try a photoelectric cell instead, focusing the light from the hypothetical star on a surface that emits photoelectrons, and counting the total number emitted during

the observation interval. If a large enough number is counted, you decide that a star is really there. This procedure too is subject to error, for background light also ejects photoelectrons, and their emission is a random process. Perhaps if you also determined from what point of the image plane each photoelectron came, for instance by dividing the surface into a mosaic of tiny isolated elements, and if you effectively utilized the known distribution of illuminance in a stellar image, you could reduce the probability of error.

Whatever technique is used to record the light and analyze the result, there is a limit to the detectability of a star or other luminous object. It is set by the random nature of both the light to be detected and the interfering background light. Whatever is done to the light entering the telescope – its passage through lenses and its reflection from mirrors, its incidence on photosensitive materials, the subsequent analysis of their response – all this is really a processing of the electromagnetic field as it exists at the aperture of the instrument during the observation interval. To determine the fundamental limitation set by nature on the detectability of the object, we look for the best way of processing that aperture field.

In radio communications and radar the need to ferret weak signals out of random background noise called forth a theory of signal detection. At first, a receiver was judged on the basis of the signal-to-noise ratio at its output. During World War II, designers of radar and communication systems perceived that the basic function of a receiver is not to produce a large signal-to-noise ratio, but to decide as reliably as possible about the presence or absence of a signal of a certain kind, or to identify one of a class of possible information-bearing signals [LaU 50]. They realized that these are decisions under conditions of uncertainty and can be treated by the theory of statistical decisions, or "hypothesis testing," developed during the thirties by Neyman and Pearson [NeP 33a,b], Wald [Wal 39], and others.

In radio and radar the basic data are the values of the voltage $v(t)$ at the terminals of an antenna during a certain interval $(0, T)$ of observation. This voltage is a random process described by probability distributions that embody what is known about the signals to be detected and about the statistical properties of the noise that corrupts them. Statistical decision theory shows how best to process that input voltage $v(t)$ in order to decide among the several hypotheses about what signals it contains. It prescribes certain kinds of filtering of $v(t)$, subsequent rectification, and appropriate logical operations that end in decisions. The probabilities of error that these incur reveal the ultimate limits on the detectability or distinguishability of the signals. The communications or radar engineer then sets about implementing the inferred procedure.

Instead of a single antenna, an array of many electromagnetically sensitive elements might be constructed in order to obtain a number of input voltages $\{v_j(t)\}$ that could be so combined as to detect most effectively signals coming

from a specific direction in the presence of interference coming from all directions. The array senses the electromagnetic field over a certain area A during each observation interval, and the theory, by working with the probability distributions of the field values at points in A and at times in $(0, T)$, determines the best way of processing it. The effective area A of the array corresponds to the aperture of our optical instrument.

If those probability distributions are to be meaningful, it must be possible, in principle, to determine the values of all components of the electric and magnetic intensities at all points in A and all times in $(0, T)$. Classical physics, which is adequate to describe electromagnetic fields at radio and radar frequencies, permits this exhaustive specification. Light, on the other hand, must be treated by quantum mechanics, which reveals an unavoidable mutual interference among measurements of the components of the electromagnetic field at various points and times. Quantum mechanics constricts the set of measurements that can be made on an electromagnetic field, or on a quantum system in general, and it renders meaningless the probability distributions upon which ordinary statistical decision theory is based. In analyzing the detectability of light, we must determine not only how to process the outcomes of our observations, but even what to observe in the first place. A new form of decision theory is required.

Classical physics represents the states of a system as points in a multidimensional phase space; quantum mechanics represents them as vectors in a Hilbert space. The vectors in a Hilbert space are transformed by linear operators. Statistically uncertain states, described classically by a probability distribution on the phase space, are accounted for in quantum mechanics by a particular kind of linear operator known as a *density operator*. Ordinary decision theory shows how to decide which of a set of probability distributions best describes a classical physical system; quantum decision theory seeks the best means of choosing among a number of potential density operators.

A decision procedure, as we shall learn in Chapter II, is represented in conventional decision theory by a probability distribution on the space of observed data. In quantum decision theory it is expressed by a set of linear operators subject to certain constraints and known as a *probability-operator measure*. By combining the density operators and the elements of the probability-operator measure, one can calculate the probabilities of the errors incurred by the decision strategy.

The concept of a probability-operator measure generalizes von Neumann's treatment of the "properties" (*Eigenschaften*) of quantum systems in terms of sets of commuting projection operators [Von 32, pp. 130–134; Von 55, pp. 247–254]. This broader formulation, due to Davies and Lewis [DaL 70], does not alter the physical implications of quantum mechanics, but clarifies the roles of observation and measurement and provides a framework for decision

theory. We introduce it in Chapter III in the course of a summary of the mathematical language and the basic principles of the quantum theory.

Once arbitrary quantum observation and decision strategies have been formulated mathematically, we can set up a means for determining the strategy for which the average probability of error, or more generally, the average cost of applying it, is minimum. This optimization theory closely parallels the conventional decision theory of Chapter II, but utilizes operators in Hilbert space instead of functions taking on numerical values. It is explained in Chapter IV and illustrated by some examples not requiring knowledge of quantum field theory.

In order to apply quantum decision theory to the detection of light, it is necessary to understand how the electromagnetic field is treated in quantum mechanics. Chapter V introduces this subject and presents a convenient calculus, developed by R. J. Glauber [Gla 63b], for working with the quantum field in a way that brings out its analogies with the fields and oscillations of classical physics. We show how to define density operators for the field at the aperture of an optical instrument during an observation interval by relating it to the field at a later instant of time in a lossless cavity placed behind the aperture and serving as a conceptually ideal receiver. We utilize these methods in Chapters VI and VII to analyze the detection of coherent and incoherent light, with applications to optical communications and the resolution of close point sources.

A receiver, instead of choosing among various potential signals, or deciding whether a certain signal is present or absent at its input, may be required to estimate one or more parameters of an incoming signal. A radar device, for instance, measures the arrival time of an electromagnetic echo pulse in order to determine the distance to a target. The amplitude and the phase of a modulated laser signal may convey information that is to be extracted by estimating their values. When a telescope is used to fix the position of a star, it can be considered as estimating two parameters of the spatial coherence function of the light at its aperture. Such estimates are subject to error because of interfering background light and because of the quantum nature of the signal itself.

Conventional statistical estimation theory shows how to estimate the parameters of the probability density function of a set of data in such a way that the average cost of errors in the estimates is minimum. In the quantum theory it is a matter of estimating parameters of the density operator of a system, and one seeks the observational strategy yielding minimum average cost. Again the strategy is characterized in greatest generality by a probability-operator measure. Chapter VIII develops this quantum estimation theory and applies it to estimates of the complex amplitude of a coherent light wave, the arrival time and carrier frequency of a coherent optical pulse, the intensity and frequency of light from

a natural, incoherent source, and the coordinates of the position of a star. We shall see how the minimum attainable mean square errors in these estimates depend on the strengths and other properties of signal and background.

All these quantities are parameters of the density operator of the electromagnetic field at the aperture of the observing optical instrument, the aperture field being regarded as a quantum system. The estimation of any of those parameters is in a certain sense a measurement to be performed on the system. Now a measurement is ordinarily associated in quantum theory with a self-adjoint operator on the Hilbert space, and the outcome of the measurement is required to be a number in the spectrum of that operator. In signal-parameter estimation, however, the operators associated with the quantities to be estimated are not apparent and may not even exist; and some extension of the quantum-mechanical treatment of measurement is necessary. The concept of a probability-operator measure appears to be just what is needed. We are then tempted to inquire whether the measurement of quantum-dynamical variables, such as the position and momentum of a particle, might not instructively be viewed as estimation of parameters of a density operator. We shall find that this approach resolves certain conceptual difficulties with quantum measurement, such as that attached to "simultaneous measurement" of position and momentum, and it provides a new way of interpreting the uncertainty principle. Perhaps quantum decision and estimation theory can in this way clarify that most problematical of all sources of observational uncertainty, the quantum-mechanical nature of physical reality.

II

Classical Detection and Estimation Theory

. . . (reason also is choice) . . .

Milton, *Paradise Lost* (III, 108)

1. HYPOTHESIS TESTING

Our summary of conventional detection and estimation theory, necessarily brief, will be confined to its simplest aspects and organized in such a way as to bring out most clearly its parallels to the quantum-mechanical theory, of which it is really a special case. The examples, chosen from elementary signal-detection theory, will perhaps relieve the bleak formality of the treatment, as well as provide a basis for comparison with the quantum results. Books in which the conventional theory is expounded and applied at great length can readily be obtained [Leh 59, Mid 60, Hel 68d, VaT 68], and further developments of the theory are recorded in such journals as *Information & Control* and the *IEEE Transactions on Information Theory, Aerospace & Electronic Systems*, and *Communication Technology*. In analogy to the term "classical physics" for the domain in which the conventional theory is valid, and with apologies to those in

whom the word "classical" evokes a calm vision of antiquity, we call this subject *classical detection and estimation theory*.

A certain system is observed in such a way as to obtain n numbers v_1, $v_2, \ldots, v_n$, on the basis of which a decision is to be made about its state. It might, for example, be the antenna of a radio-communication receiver, at whose terminals the voltage $v(t)$ is sampled n times during an observation interval $(0, T)$, and one is to decide which of a number of possible signals reached the antenna during $(0, T)$. The system may be in any one of M states, and the proposition, "The system is in state j," we call hypothesis H_j, $j = 1, 2, \ldots, M$. For the receiving antenna, hypothesis H_j asserts that

$$v(t) = s_j(t) + n(t),$$

where $s_j(t)$ is the jth signal and $n(t)$ stands for the random noise. The data $(v_1, v_2, \ldots, v_n) = \mathbf{v}$ are random variables whose joint probability density function (p.d.f.) is $p_j(\mathbf{v}) = p_j(v_1, v_2, \ldots, v_n)$ when the system is in state j. From past experience it is known that the system is in state j with a relative frequency ζ_j,

$$\sum_{j=1}^{M} \zeta_j = 1. \tag{1.1}$$

The numbers ζ_j are the *prior* probabilities of the several hypotheses.

As a consequence of each decision, certain actions are taken depending on which hypothesis is selected; if nothing were to be done, there would be no point to making the decisions. These actions entail certain costs that also depend on the actual state of the system. Let C_{ij} be the cost incurred by choosing hypothesis H_i when H_j is true. These numbers C_{ij} assign relative weights to the various possible errors and correct decisions. The decision procedure is to be repeated over and over under the same general circumstances, and one desires a procedure whose average cost is minimum.

A decision procedure or *strategy* must prescribe which hypothesis is to be chosen for each possible set $\mathbf{v}$ of data. One can imagine making the decisions in a way that involves a random element such that there is a probability $\pi_i(\mathbf{v})$ that hypothesis H_i is chosen when the set $\mathbf{v}$ of data is observed, $i = 1, 2, \ldots, M$. These probabilities are subject to the conditions

$$0 \leqslant \pi_i(\mathbf{v}) \leqslant 1, \qquad \sum_{i=1}^{M} \pi_i(\mathbf{v}) = 1. \tag{1.2}$$

We say that the functions $\{\pi_i(\mathbf{v})\}$ specify a *randomized strategy*. The choice of a hypothesis might be made by a roulette wheel so designed that the probabilities π_i of its stopping at the numbers $1, 2, \ldots, M$ depend on the outcomes v_1, $v_2, \ldots, v_n$ of the observation. Pure guessing among the M hypotheses corresponds to $\pi_i(\mathbf{v}) \equiv M^{-1}$ for all i. Decision theory was formulated in this general manner by Wald [Wal 39, Wal 50].

The probability that hypothesis H_i is chosen when hypothesis H_j is true is

$$\Pr\{i \mid j\} = \int_R \pi_i(\mathbf{v}) p_j(\mathbf{v})\, d^n\mathbf{v}, \tag{1.3}$$

where $d^n\mathbf{v} = dv_1\, dv_2 \cdots dv_n$ is an elementary volume in the n-dimensional space R of the data $\mathbf{v}$. Upon this event the cost C_{ij} is incurred. As H_j is true with probability ζ_j a priori, the average cost of the strategy is

$$\begin{aligned}\bar{C} = \bar{C}[\{\pi_i\}] &= \sum_{i=1}^{M} \sum_{j=1}^{M} \zeta_j C_{ij} \Pr\{i \mid j\} \\ &= \sum_{i=1}^{M} \sum_{j=1}^{M} \zeta_j C_{ij} \int_R \pi_i(\mathbf{v}) p_j(\mathbf{v})\, d^n\mathbf{v}. \end{aligned} \tag{1.4}$$

Defining for each hypothesis H_i the "risk function"

$$W_i(\mathbf{v}) = \sum_{j=1}^{M} \zeta_j C_{ij} p_j(\mathbf{v}), \tag{1.5}$$

we write the average cost as

$$\bar{C} = \int_R \sum_{i=1}^{M} W_i(\mathbf{v}) \pi_i(\mathbf{v})\, d^n\mathbf{v}, \tag{1.6}$$

and we seek the M functions $\pi_i(\mathbf{v})$ that both satisfy (1.2) and make $\bar{C}$ as small as possible.

The value of the integral in (1.6) will be least if the integrand is made as small as possible at each point $\mathbf{v} \in R$. The average cost $\bar{C}$ will therefore be minimum if at each point $\mathbf{v}$ we choose the hypothesis for which the risk $W_i(\mathbf{v})$ is smallest; that is, we set

$$\pi_j(\mathbf{v}) = 1, \qquad \pi_i(\mathbf{v}) \equiv 0, \qquad \forall i \neq j, \tag{1.7}$$

at all points $\mathbf{v} \in R$ for which

$$W_j(\mathbf{v}) < W_i(\mathbf{v}), \qquad \forall i \neq j. \tag{1.8}$$

If there is a tie among two or more numbers $W_i(\mathbf{v})$ as the smallest of the set, it does not matter whether one always picks a particular one of these or chooses among them at random. Our strategy really does not require randomization at all. In effect it divides the data space R into M regions $R_1, R_2, \ldots, R_M$, specified by (1.8), and when the data point $\mathbf{v}$ falls into region R_j, hypothesis H_j is chosen. We included the possibility of randomization, however, in order that the optimum procedure could be developed in a manner as nearly like the derivation in quantum decision theory as possible.

In order to facilitate the comparison, we introduce the function

$$\Upsilon(\mathbf{v}) = \min_j W_j(\mathbf{v}). \tag{1.9}$$

Then (1.7)–(1.9) can be expressed as

$$[W_i(\mathbf{v}) - \Upsilon(\mathbf{v})]\,\pi_i(\mathbf{v}) = 0, \tag{1.10}$$

$$W_i(\mathbf{v}) - \Upsilon(\mathbf{v}) \geqslant 0, \tag{1.11}$$

for all data sets $\mathbf{v}$ and all hypotheses H_i. Furthermore, from (1.2) it follows by summing (1.10) over i that

$$\Upsilon(\mathbf{v}) = \sum_{i=1}^{M} W_i(\mathbf{v})\pi_i(\mathbf{v}), \tag{1.12}$$

and by (1.6) the minimum average cost is

$$\bar{C}_{\min} = \int_R \Upsilon(\mathbf{v})\, d^n\mathbf{v}. \tag{1.13}$$

Let the functions $\{\pi_i'(\mathbf{v})\}$ define some other strategy than the optimum; they must obey the conditions in (1.2). The difference between the cost $\bar{C}'$ that this other strategy incurs and the cost $\bar{C}_{\min}$ defined by (1.10)–(1.13) is, by (1.6),

$$\bar{C}' - \bar{C}_{\min} = \sum_{i=1}^{M} \int_R [W_i(\mathbf{v}) - \Upsilon]\,\pi_i'(\mathbf{v})\, d^n\mathbf{v}. \tag{1.14}$$

Because of (1.11) and because the probabilities $\pi_i'(\mathbf{v})$ must be nonnegative,

$$\bar{C}' - \bar{C}_{\min} \geqslant 0, \tag{1.15}$$

and (1.10) further shows that the minimum cost is attained by the set $\{\pi_i(\mathbf{v})\}$. Because of (1.10), $\pi_i(\mathbf{v})$ must vanish for all hypotheses H_i for which $W_i(\mathbf{v}) > \Upsilon(\mathbf{v})$. For those hypotheses H_j for which $W_j(\mathbf{v}) = \Upsilon(\mathbf{v})$, the probabilities $\pi_j(\mathbf{v})$ can be set equal to any nonnegative numbers summing to 1. Usually there is only one such hypothesis, and we obtain the equivalent prescription in (1.7) and (1.8).

The risk functions $W_i(\mathbf{v})$ figuring in this analysis are proportional to the posterior risks $r_i(\mathbf{v})$ of the hypotheses in view of the data $\mathbf{v}$ observed,

$$r_i(\mathbf{v}) = \sum_{k=1}^{M} C_{ik}\, \Pr\{H_k \mid \mathbf{v}\} = W_i(\mathbf{v})/p(\mathbf{v}), \tag{1.16}$$

where

$$\Pr\{H_k \mid \mathbf{v}\} = \zeta_k p_k(\mathbf{v})/p(\mathbf{v}), \tag{1.17}$$

is the posterior probability of hypothesis H_k,

$$p(\mathbf{v}) = \sum_{j=1}^{M} \zeta_j p_j(\mathbf{v})$$

being the overall joint p.d.f. of the data. The best strategy picks the hypothesis with smallest posterior risk.

In "An Essay Toward Solving a Problem in the Doctrine of Chances," which appeared in the *Philosophical Transactions of the Royal Society of London* in 1763, the Reverend Thomas Bayes treated the problem of making decisions under conditions of uncertainty, and he recommended selecting that hypothesis for which the posterior probability is greatest [Bay 58]. The formula in (1.17) for calculating it is called Bayes's rule. Bayes's prescription results from an assignment of costs

$$C_{ij} \equiv 1, \qquad i \neq j; \qquad C_{ii} \equiv 0, \tag{1.18}$$

which penalizes all errors equally, or from one that equally rewards all correct decisions,

$$C_{ii} \equiv -1; \qquad C_{ij} \equiv 0, \qquad i \neq j. \tag{1.19}$$

This strategy of Bayes's minimizes the average probability of error

$$P_e = 1 - \sum_{j=1}^{M} \zeta_j \Pr\{j \mid j\} = \sum_{j=1}^{M} \zeta_j \sum_{k \neq j} \Pr\{k \mid j\}. \tag{1.20}$$

Nevertheless, current literature calls a strategy minimizing the average cost for any assignment $\{C_{ij}\}$ of costs a *Bayes strategy*, and this general formulation of hypothesis testing is often termed "Bayesian."

The data $\mathbf{v}$ have been considered here as points in a continuum and assigned probability density functions $p_j(\mathbf{v})$, $j = 1, 2, \ldots, M$. When on the other hand they are discrete random variables, the p.d.f.'s $p_j(\mathbf{v})$ must be replaced throughout our derivation by the probabilities $\Pr\{\mathbf{v} \mid H_j\}$ of obtaining the data under the M hypotheses. Integrals over the data space R are replaced by summations. With these changes, the Bayes strategy continues to be specified by (1.7)–(1.8) or by (1.10)–(1.12). Ties among the minimum values of $W_i(\mathbf{v})$ must now be anticipated, but can be resolved arbitrarily.

In detection theory the set of available data is often infinitely large; one can imagine sampling the voltage $v(t)$ at the antenna terminals at all instants of time during the observation interval. Difficulty then arises in defining the probability density functions $p_j(\mathbf{v})$. The most direct approach is to start with a finite set $\mathbf{v}$ of n data and pass to the limit $n \to \infty$. The limit process is simplest if carried out on the posterior probabilities $\Pr\{H_k \mid \mathbf{v}\}$ defined in (1.17). One sets up a dummy hypothesis H_0 under which the p.d.f. of the data is $p_0(\mathbf{v})$; often this is taken as

their joint p.d.f. when no signals at all are present and the data represent samples of the noise alone. The posterior probabilities can then be expressed as

$$\Pr\{H_k | \mathbf{v}\} = \zeta_k \Lambda_k(\mathbf{v}) \Big/ \sum_{j=1}^{M} \zeta_j \Lambda_j(\mathbf{v}), \tag{1.21}$$

where

$$\Lambda_k(\mathbf{v}) = p_k(\mathbf{v})/p_0(\mathbf{v}), \qquad k = 1, 2, \ldots, M, \tag{1.22}$$

are the likelihood ratios. They go into functionals $\Lambda_k[v(t)]$ of the observed voltage in the limit $n \to \infty$ of infinitely dense sampling. Examples will be presented in the following sections.

2. BINARY DECISIONS

(a) The Bayesian Formulation

When there are only two hypotheses between which to choose, they are customarily labeled H_0 and H_1. In signal detection, hypothesis H_0 usually asserts that the input $v(t)$ at the antenna terminals to the receiver contains only noise,

$$H_0: \quad v(t) = n(t);$$

hypothesis H_1 asserts that one of a specified class of signals $s(t)$ is on hand as well,

$$H_1: \quad v(t) = s(t) + n(t).$$

Statisticians term H_0 the "null hypothesis" and H_1 the "alternative" [Leh 59].

Under these hypotheses the data $\mathbf{v} = (v_1, v_2, \ldots, v_n)$, which may be samples of $v(t)$, have the joint probability density functions $p_j(\mathbf{v})$, $j = 0, 1$. The prior probabilities of the two hypotheses are ζ_0 and ζ_1, $\zeta_0 + \zeta_1 = 1$. The cost of choosing hypothesis H_i when H_j is true is again designated by C_{ij}. The Bayes strategy selects hypothesis H_1 when the data $\mathbf{v}$ are such that $W_1(\mathbf{v}) < W_0(\mathbf{v})$, these risk functions defined as in (1.5). The reader can easily show that this condition is equivalent to

$$\Lambda(\mathbf{v}) = p_1(\mathbf{v})/p_0(\mathbf{v}) > \Lambda_0, \tag{2.1}$$

where the decision level Λ_0 is given by

$$\Lambda_0 = \zeta_0(C_{10} - C_{00})/\zeta_1(C_{01} - C_{11}). \tag{2.2}$$

The function $\Lambda(\mathbf{v})$ is called the likelihood ratio, and in the limit of an infinite number of data it goes into the likelihood functional $\Lambda[v(t)]$. This is sometimes

called the "Radon–Nikodym derivative" of the probability measure of the data $v(t)$ under H_1 with respect to their probability measure under H_0.

(b) The Neyman–Pearson Criterion

Many statisticians spurn the Bayesian approach, averring that in practice the costs C_{ij} are difficult to ascertain and the prior probabilities ξ_j unreliable if not meaningless. The panoply of costs and prior probabilities is most easily eliminated from binary hypothesis testing, for which an alternative viewpoint was proposed by Neyman and Pearson [NeP 33a, 33b]. The probability

$$Q_0 = \Pr\{H_1 \mid H_0\}$$

of choosing hypothesis H_1 when H_0 is true is fixed at or below some tolerable level, and the optimum strategy maximizes the probability

$$Q_d = \Pr\{H_1 \mid H_1\}$$

of correctly choosing H_1 when H_1 is true. Statisticians call Q_0 the "size" and Q_d the "power" of a statistical test. In detection theory they are called the *false-alarm probability* and the *detection probability*, respectively. A test or strategy for which Q_d is maximum for a preassigned value of Q_0 is said to satisfy the Neyman–Pearson criterion. We shall see that it too bases its decision on the value of the likelihood ratio $\Lambda(\mathbf{v})$.

Allowing for the possibility that the data may be discrete, and adopting a line of reasoning parallel to what will be followed in the quantum-mechanical counterpart of this problem, we consider a randomized strategy. The probability that hypothesis H_1 is chosen when the set $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ is observed is denoted by $\pi(\mathbf{v})$, $0 \leqslant \pi(\mathbf{v}) \leqslant 1$; hypothesis H_0 is chosen with probability $1 - \pi(\mathbf{v})$. The false-alarm and detection probabilities are then

$$Q_0[\pi] = \int_R \pi(\mathbf{v}) p_0(\mathbf{v})\, d^n\mathbf{v} \tag{2.3}$$

and

$$Q_d[\pi] = \int_R \pi(\mathbf{v}) p_1(\mathbf{v})\, d^n\mathbf{v} = \int_R \pi(\mathbf{v})\Lambda(\mathbf{v}) p_0(\mathbf{v})\, d^n\mathbf{v}, \tag{2.4}$$

where $\Lambda(\mathbf{v})$ is the likelihood ratio defined in (2.1).

For each possible function $\pi(\mathbf{v})$ taking values in the closed interval $[0,1]$, plot the point $(Q_0[\pi], Q_d[\pi])$ in the (Q_0, Q_d) plane. It lies somewhere in a region D within the square $0 \leqslant Q_0 \leqslant 1$, $0 \leqslant Q_d \leqslant 1$. As shown in Fig. 2.1, this region D is convex. The reasons for this are as follows:

Suppose that the strategies represented by functions $\pi'(\mathbf{v})$ and $\pi''(\mathbf{v})$ determine the two points A: (Q_0', Q_d') and B: (Q_0'', Q_d''), and consider the decision

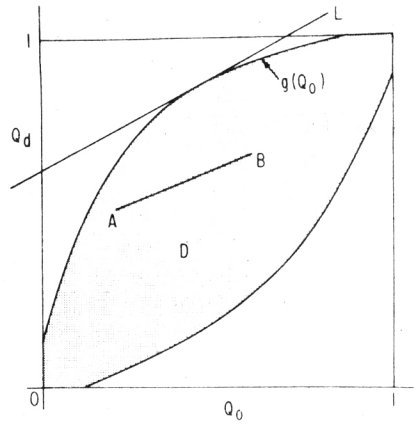

Fig. 2.1. False-alarm and detection probabilities in binary decisions.

probability function

$$\pi(\mathbf{v}) = \alpha\pi'(\mathbf{v}) + \beta\pi''(\mathbf{v}), \qquad 0 < \alpha < 1, \quad \beta = 1 - \alpha,$$

which corresponds to picking one of the two strategies at random, the first with probability α, the second with probability β. It leads to the false-alarm and detection probabilities

$$Q_0 = \alpha Q_0' + \beta Q_0'', \qquad Q_d = \alpha Q_d' + \beta Q_d'',$$

and the point (Q_0, Q_d) lies on the straight-line segment connecting A and B. Thus for any pair of points A and B in D, all points on the line joining them must also lie in D, and D is therefore convex. Taking $\pi(\mathbf{v}) \equiv 0$ and $\pi(\mathbf{v}) \equiv 1$, respectively, shows that D includes the origin $(0, 0)$ and point $(1, 1)$ at the opposite corner of the square.

Designate by $Q_d = g(Q_0)$ the curve bounding the region D above. That curve is convex upward, as shown in Fig. 2.1, because of the convexity of D. The straight line L tangent to it at the point $(Q_0, g(Q_0))$ must therefore lie above D everywhere else. Let λ be the slope of the line. Then for all decision functions $\pi(\mathbf{v})$

$$g(Q_0) - \lambda Q_0 \geqslant Q_d[\pi] - \lambda Q_0[\pi] = \int_R \pi(\mathbf{v})[\Lambda(\mathbf{v}) - \lambda] p_0(\mathbf{v})\, d^n\mathbf{v}.$$

The decision function $\pi(\mathbf{v})$ attaining the maximum detection probability Q_d for the given false-alarm probability Q_0 will be the one for which the integral on the right-hand side is as large as possible, and because $p_0(\mathbf{v}) \geqslant 0$, that decision function must satisfy

$$\pi(\mathbf{v}) = 1, \quad \Lambda(\mathbf{v}) - \lambda > 0; \qquad \pi(\mathbf{v}) = 0, \quad \Lambda(\mathbf{v}) - \lambda < 0. \tag{2.5}$$

Let Z be the set of data points $\mathbf{v}$ for which $\Lambda(\mathbf{v}) = \lambda$ exactly. Then we put

$$\pi(\mathbf{v}) = f, \qquad \mathbf{v} \in Z, \quad 0 \leqslant f < 1, \tag{2.6}$$

where f and λ are such numbers that

$$Q_0 = f \Pr\{\mathbf{v} \in Z \mid H_0\} + \Pr\{\Lambda(\mathbf{v}) > \lambda \mid H_0\} \tag{2.7}$$

equals the preassigned value of the false-alarm probability. Thus when $\Lambda(\mathbf{v}) > \lambda$, hypothesis H_1 is always chosen, and when $\Lambda(\mathbf{v}) < \lambda$, H_0 is chosen; but if $\Lambda(\mathbf{v})$ equals λ exactly, hypothesis H_1 is chosen with probability f, by using some chance device such as a properly weighted coin. The probability of detection is then

$$Q_d = f \Pr\{\mathbf{v} \in Z \mid H_1\} + \Pr\{\Lambda(\mathbf{v}) > \lambda \mid H_1\}. \tag{2.8}$$

For continuous data, the region Z has in general zero probability under both hypotheses. The value of the decision level $\lambda = \Lambda_{np}$ is then selected so that the false-alarm probability

$$Q_0 = \int_R U[\Lambda(\mathbf{v}) - \lambda] p_0(\mathbf{v})\, d^n\mathbf{v} \tag{2.9}$$

takes on the preassigned value; $U(x) = 1$, $x > 0$; $U(x) = 0$, $x < 0$. The boundary curve $Q_d = g(Q_0)$ is smooth, and it is parametrized by the value of λ, which runs from 0 at (1, 1) to ∞ at (0, 0). It is called the *operating characteristic* of the binary-hypothesis test.

When the data are discrete, the curve $Q_d = g(Q_0)$ has a polygonal form. On each of its straight segments the parameter λ has a constant value, and a segment is traced from left to right as f varies from 0 to 1. The likelihood ratios $\Lambda(\mathbf{v})$ now take on only a countable set of values, which one arranges in descending order. One adds the probabilities $\Pr\{\Lambda(\mathbf{v}) \mid H_0\}$, starting with the highest value of $\Lambda(\mathbf{v})$, until the preassigned value of Q_0 is just exceeded. The decision level $\lambda = \Lambda_{np}$ is equal to the likelihood ratio $\Lambda(\mathbf{v})$ whose probability was last added in.

As an example of this kind of randomized strategy, let the decision between H_0 and H_1 be based on the nonnegative integral-valued datum n whose probabilities under the two hypotheses are

$$\Pr\{n \mid H_i\} = (1 - v_i) v_i^n, \qquad i = 0, 1, \quad n \geqslant 0, \tag{2.10}$$

with $v_1 > v_0$; n might be the number of photoelectrons emitted during a certain interval $(0, T)$ by a detector onto which thermal light of extremely small spectral width $W \ll T^{-1}$ is incident. The likelihood ratio

$$\Lambda(n) = \left(\frac{1 - v_1}{1 - v_0}\right)\left(\frac{v_1}{v_0}\right)^n$$

increases monotonically with n; hypothesis H_1 will always be selected when n

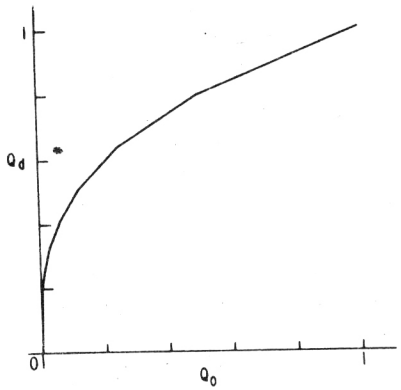

Fig. 2.2. Operating characteristic for decisions between geometric distributions: $v_0 = 0.5$, $v_1 = 0.8$.

exceeds a certain number ν, and H_0 whenever $n < \nu$. When $n = \nu$, however, it will be necessary to choose H_1 with such a probability $\pi(\nu) = f$ that

$$Q_0 = f(1 - v_0)v_0{}^{\nu} + \sum_{n=\nu+1}^{\infty} (1 - v_0)v_0{}^{n} = f(1 - v_0)v_0{}^{\nu} + v_0{}^{\nu+1} \tag{2.11}$$

equals the preassigned false-alarm probability. The decision level ν is the greatest integer in $(\ln Q_0)/(\ln v_0)$, and f is then easily calculated from (2.11). The detection probability is

$$Q_d = f(1 - v_1)v_1{}^{\nu} + v_1{}^{\nu+1}. \tag{2.12}$$

For each integer ν, Q_d is a linear function of f and a fortiori of Q_0, whence the polygonal form of curve $Q_d = g(Q_0)$. Figure 2.2 shows this operating characteristic for $v_0 = 0.5$, $v_1 = 0.8$.

(c) Detection of a Known Signal in Gaussian Noise

(i) Forming the Likelihood Ratio

In the simplest binary-detection problem a receiver is to choose between two hypotheses,

$$H_0: \quad v(t) = n(t), \qquad H_1: \quad v(t) = n(t) + s(t),$$

about the voltage $v(t)$ between the terminals of its antenna as observed during an interval $(0, T)$. Here $s(t)$ is a signal of completely known form, and $n(t)$ is a Gaussian random process with expected values zero,

$$\mathrm{E}[n(t) \mid H_i] = 0, \qquad i = 0, 1,$$

and autocovariance function

$$\mathbf{E}[n(t_1)n(t_2) \mid H_i] = \varphi(t_1, t_2), \qquad i = 0, 1. \tag{2.13}$$

We shall treat this simple detection problem as an example, referring the reader to textbooks for further details [Mid 60, Hel 68d, VaT 68].

In order to specify the probability density functions needed for forming the likelihood ratio, we must sample the input $v(t)$ in some way. A convenient and versatile kind of sampling generates the numbers

$$v_k = \int_0^T f_k(t)v(t)\,dt, \tag{2.14}$$

in which the functions $f_k(t)$ form a complete orthonormal set over $(0, T)$. When we say that $n(t)$ is a Gaussian random process, we mean that any finite number m of these samples has a joint p.d.f. of the Gaussian form,

$$p_i(\mathbf{v}) = (2\pi)^{-m/2} \mid \det \varphi \mid^{-1/2}$$
$$\times \exp\left[-\frac{1}{2}\sum_{k=1}^{m}\sum_{l=1}^{m}\mu_{kl}(v_k - \bar{v}_{ki})(v_l - \bar{v}_{li})\right], \qquad i = 0, 1, \tag{2.15}$$

in which

$$\bar{v}_{ki} = \mathbf{E}(v_k \mid H_i) = \begin{cases} 0, & i = 0, \\ s_k = \displaystyle\int_0^T f_k(t)s(t)\,dt, & i = 1, \end{cases} \tag{2.16}$$

are the expected values of the samples and

$$\varphi_{kl} = \mathbf{E}[(v_k - \bar{v}_{ki})(v_l - \bar{v}_{li}) \mid H_i]$$
$$= \int_0^T \int_0^T f_k(t_2)\varphi(t_2, t_1)f_l(t_1)\,dt_1\,dt_2 \tag{2.17}$$

are their covariances; $\varphi = \|\varphi_{kl}\|$ is the $m \times m$ matrix of these covariances, $\det \varphi$ its determinant, and

$$\mu = \varphi^{-1} = \|\mu_{ij}\| \tag{2.18}$$

its inverse. Gaussian noise arises from the fluctuations of the myriad atoms, ions, and electrons composing the receiver and its surroundings, which induce a randomly fluctuating voltage of this kind at the input terminals of the receiver.

The likelihood ratio is most simply formed by choosing the sampling functions $f_k(t)$ so that the covariance matrix φ is diagonal. To this end the functions $f_k(t)$ are taken as the eigenfunctions of the autocovariance $\varphi(t_2, t_1)$, defined by the integral equation

$$\lambda_k f_k(t_2) = \int_0^T \varphi(t_2, t_1)f_k(t_1)\,dt_1. \tag{2.19}$$