Chapter 7

Statistical estimation

7.1 Parametric distribution estimation

7.1.1 Maximum likelihood estimation

We consider a family of probability distributions on $\mathbf{R}^m$, indexed by a vector $x \in \mathbf{R}^n$, with densities $p_x(\cdot)$. When considered as a function of x, for fixed $y \in \mathbf{R}^m$, the function $p_x(y)$ is called the *likelihood function*. It is more convenient to work with its logarithm, which is called the *log-likelihood function*, and denoted l:

$$l(x) = \log p_x(y).$$

There are often constraints on the values of the parameter x, which can represent prior knowledge about x, or the domain of the likelihood function. These constraints can be explicitly given, or incorporated into the likelihood function by assigning $p_x(y) = 0$ (for all y) whenever x does not satisfy the prior information constraints. (Thus, the log-likelihood function can be assigned the value $-\infty$ for parameters x that violate the prior information constraints.)

Now consider the problem of estimating the value of the parameter x, based on observing one sample y from the distribution. A widely used method, called *maximum likelihood (ML) estimation*, is to estimate x as

$$\hat{x}_{\mathrm{ml}} = \mathrm{argmax}_x p_x(y) = \mathrm{argmax}_x l(x),$$

i.e., to choose as our estimate a value of the parameter that maximizes the likelihood (or log-likelihood) function for the observed value of y. If we have prior information about x, such as $x \in C \subseteq \mathbf{R}^n$, we can add the constraint $x \in C$ explicitly, or impose it implicitly, by redefining $p_x(y)$ to be zero for $x \notin C$.

The problem of finding a maximum likelihood estimate of the parameter vector x can be expressed as

$$\begin{array}{ll} \text{maximize} & l(x) = \log p_x(y) \\ \text{subject to} & x \in C, \end{array} \qquad (7.1)$$

where $x \in C$ gives the prior information or other constraints on the parameter vector x. In this optimization problem, the vector $x \in \mathbf{R}^n$ (which is the parameter in the probability density) is the variable, and the vector $y \in \mathbf{R}^m$ (which is the observed sample) is a problem parameter.

The maximum likelihood estimation problem (7.1) is a convex optimization problem if the log-likelihood function l is concave for each value of y, and the set C can be described by a set of linear equality and convex inequality constraints, a situation which occurs in many estimation problems. For these problems we can compute an ML estimate using convex optimization.

Linear measurements with IID noise

We consider a linear measurement model,

$$y_i = a_i^T x + v_i, \quad i = 1, \ldots, m,$$

where $x \in \mathbf{R}^n$ is a vector of parameters to be estimated, $y_i \in \mathbf{R}$ are the measured or observed quantities, and v_i are the measurement errors or noise. We assume that v_i are independent, identically distributed (IID), with density p on $\mathbf{R}$. The likelihood function is then

$$p_x(y) = \prod_{i=1}^m p(y_i - a_i^T x),$$

so the log-likelihood function is

$$l(x) = \log p_x(y) = \sum_{i=1}^m \log p(y_i - a_i^T x).$$

The ML estimate is any optimal point for the problem

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^m \log p(y_i - a_i^T x), \end{array} \tag{7.2}$$

with variable x. If the density p is log-concave, this problem is convex, and has the form of a penalty approximation problem ((6.2), page 294), with penalty function $-\log p$.

Example 7.1 *ML estimation for some common noise densities.*

- *Gaussian noise.* When v_i are Gaussian with zero mean and variance σ^2, the density is $p(z) = (2\pi\sigma^2)^{-1/2} e^{-z^2/2\sigma^2}$, and the log-likelihood function is

  $$l(x) = -(m/2)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|Ax - y\|_2^2,$$

  where A is the matrix with rows $a_1^T, \ldots, a_m^T$. Therefore the ML estimate of x is $x_{\rm ml} = \operatorname{argmin}_x \|Ax - y\|_2^2$, the solution of a least-squares approximation problem.

- *Laplacian noise.* When v_i are Laplacian, *i.e.*, have density $p(z) = (1/2a)e^{-|z|/a}$ (where $a > 0$), the ML estimate is $\hat{x} = \operatorname{argmin}_x \|Ax - y\|_1$, the solution of the ℓ_1-norm approximation problem.

- *Uniform noise.* When v_i are uniformly distributed on $[-a, a]$, we have $p(z) = 1/(2a)$ on $[-a, a]$, and an ML estimate is any x satisfying $\|Ax - y\|_\infty \le a$.

ML interpretation of penalty function approximation

Conversely, we can interpret any penalty function approximation problem

$$\text{minimize} \quad \textstyle\sum_{i=1}^m \phi(b_i - a_i^T x)$$

as a maximum likelihood estimation problem, with noise density

$$p(z) = \frac{e^{-\phi(z)}}{\int e^{-\phi(u)}\, du},$$

and measurements b. This observation gives a statistical interpretation of the penalty function approximation problem. Suppose, for example, that the penalty function ϕ grows very rapidly for large values, which means that we attach a very large cost or penalty to large residuals. The corresponding noise density function p will have very small tails, and the ML estimator will avoid (if possible) estimates with any large residuals because these correspond to very unlikely events.

We can also understand the robustness of ℓ_1-norm approximation to large errors in terms of maximum likelihood estimation. We interpret ℓ_1-norm approximation as maximum likelihood estimation with a noise density that is Laplacian; ℓ_2-norm approximation is maximum likelihood estimation with a Gaussian noise density. The Laplacian density has larger tails than the Gaussian, *i.e.*, the probability of a very large v_i is far larger with a Laplacian than a Gaussian density. As a result, the associated maximum likelihood method expects to see greater numbers of large residuals.

Counting problems with Poisson distribution

In a wide variety of problems the random variable y is nonnegative integer valued, with a Poisson distribution with mean $\mu > 0$:

$$\mathbf{prob}(y = k) = \frac{e^{-\mu}\mu^k}{k!}.$$

Often y represents the count or number of events (such as photon arrivals, traffic accidents, etc.) of a Poisson process over some period of time.

In a simple statistical model, the mean μ is modeled as an affine function of a vector $u \in \mathbf{R}^n$:

$$\mu = a^T u + b.$$

Here u is called the vector of *explanatory variables*, and the vector $a \in \mathbf{R}^n$ and number $b \in \mathbf{R}$ are called the *model parameters*. For example, if y is the number of traffic accidents in some region over some period, u_1 might be the total traffic flow through the region during the period, u_2 the rainfall in the region during the period, and so on.

We are given a number of observations which consist of pairs (u_i, y_i), $i = 1, \ldots, m$, where y_i is the observed value of y for which the value of the explanatory variable is $u_i \in \mathbf{R}^n$. Our job is to find a maximum likelihood estimate of the model parameters $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$ from these data.

The likelihood function has the form

$$\prod_{i=1}^{m} \frac{(a^T u_i + b)^{y_i} \exp(-(a^T u_i + b))}{y_i!},$$

so the log-likelihood function is

$$l(a, b) = \sum_{i=1}^{m} (y_i \log(a^T u_i + b) - (a^T u_i + b) - \log(y_i!)).$$

We can find an ML estimate of a and b by solving the convex optimization problem

$$\text{maximize} \quad \sum_{i=1}^{m} (y_i \log(a^T u_i + b) - (a^T u_i + b)),$$

where the variables are a and b.

Logistic regression

We consider a random variable $y \in \{0, 1\}$, with

$$\mathbf{prob}(y = 1) = p, \qquad \mathbf{prob}(y = 0) = 1 - p,$$

where $p \in [0, 1]$, and is assumed to depend on a vector of explanatory variables $u \in \mathbf{R}^n$. For example, $y = 1$ might mean that an individual in a population acquires a certain disease. The probability of acquiring the disease is p, which is modeled as a function of some explanatory variables u, which might represent weight, age, height, blood pressure, and other medically relevant variables.

The *logistic model* has the form

$$p = \frac{\exp(a^T u + b)}{1 + \exp(a^T u + b)}, \tag{7.3}$$

where $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$ are the model parameters that determine how the probability p varies as a function of the explanatory variable u.

Now suppose we are given some data consisting of a set of values of the explanatory variables $u_1, \ldots, u_m \in \mathbf{R}^n$ along with the corresponding outcomes $y_1, \ldots, y_m \in \{0, 1\}$. Our job is to find a maximum likelihood estimate of the model parameters $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$. Finding an ML estimate of a and b is sometimes called *logistic regression*.

We can re-order the data so for $u_1, \ldots, u_q$, the outcome is $y = 1$, and for $u_{q+1}, \ldots, u_m$ the outcome is $y = 0$. The likelihood function then has the form

$$\prod_{i=1}^{q} p_i \prod_{i=q+1}^{m} (1 - p_i),$$

where p_i is given by the logistic model with explanatory variable u_i. The log-likelihood function has the form

$$l(a, b) \;=\; \sum_{i=1}^{q} \log p_i + \sum_{i=q+1}^{m} \log(1 - p_i)$$

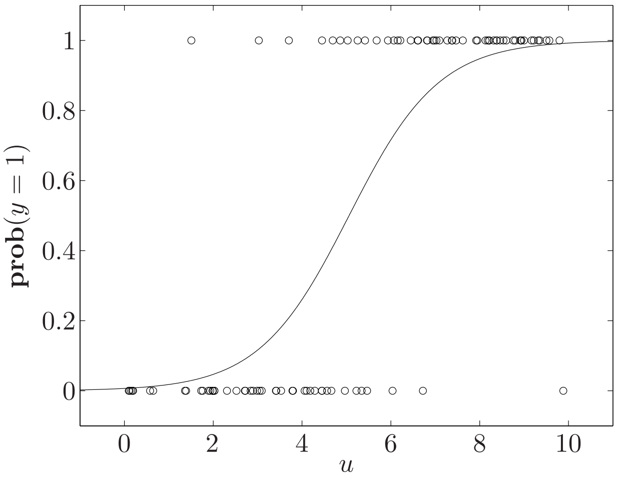

Figure 7.1 *Logistic regression.* The circles show 50 points (u_i, y_i), where $u_i \in \mathbf{R}$ is the explanatory variable, and $y_i \in \{0,1\}$ is the outcome. The data suggest that for $u < 5$ or so, the outcome is more likely to be $y = 0$, while for $u > 5$ or so, the outcome is more likely to be $y = 1$. The data also suggest that for $u < 2$ or so, the outcome is very likely to be $y = 0$, and for $u > 8$ or so, the outcome is very likely to be $y = 1$. The solid curve shows $\mathbf{prob}(y = 1) = \exp(au + b)/(1 + \exp(au + b))$ for the maximum likelihood parameters a, b. This maximum likelihood model is consistent with our informal observations about the data set.

$$
\begin{aligned}
&= \sum_{i=1}^{q} \log \frac{\exp(a^T u_i + b)}{1 + \exp(a^T u_i + b)} + \sum_{i=q+1}^{m} \log \frac{1}{1 + \exp(a^T u_i + b)} \\
&= \sum_{i=1}^{q} (a^T u_i + b) - \sum_{i=1}^{m} \log(1 + \exp(a^T u_i + b)).
\end{aligned}
$$

Since l is a concave function of a and b, the logistic regression problem can be solved as a convex optimization problem. Figure 7.1 shows an example with $u \in \mathbf{R}$.

Covariance estimation for Gaussian variables

Suppose $y \in \mathbf{R}^n$ is a Gaussian random variable with zero mean and covariance matrix $R = \mathbf{E}\, yy^T$, so its density is

$$p_R(y) = (2\pi)^{-n/2} \det(R)^{-1/2} \exp(-y^T R^{-1} y/2),$$

where $R \in \mathbf{S}^n_{++}$. We want to estimate the covariance matrix R based on N independent samples $y_1, \dots, y_N \in \mathbf{R}^n$ drawn from the distribution, and using prior knowledge about R.

The log-likelihood function has the form

$$l(R) \quad = \quad \log p_R(y_1, \dots, y_N)$$

$$\begin{aligned} &= -(Nn/2)\log(2\pi) - (N/2)\log\det R - (1/2)\sum_{k=1}^{N} y_k^T R^{-1} y_k \\ &= -(Nn/2)\log(2\pi) - (N/2)\log\det R - (N/2)\,\mathbf{tr}(R^{-1}Y), \end{aligned}$$

where

$$Y = \frac{1}{N}\sum_{k=1}^{N} y_k y_k^T$$

is the sample covariance of $y_1, \ldots, y_N$. This log-likelihood function is *not* a concave function of R (although it is concave on a subset of its domain $\mathbf{S}^n_{++}$; see exercise 7.4), but a change of variable yields a concave log-likelihood function. Let S denote the inverse of the covariance matrix, $S = R^{-1}$ (which is called the *information matrix*). Using S in place of R as a new parameter, the log-likelihood function has the form

$$l(S) = -(Nn/2)\log(2\pi) + (N/2)\log\det S - (N/2)\,\mathbf{tr}(SY),$$

which *is* a concave function of S.

Therefore the ML estimate of S (hence, R) is found by solving the problem

$$\begin{array}{ll} \text{maximize} & \log\det S - \mathbf{tr}(SY) \\ \text{subject to} & S \in \mathcal{S} \end{array} \tag{7.4}$$

where $\mathcal{S}$ is our prior knowledge of $S = R^{-1}$. (We also have the implicit constraint that $S \in \mathbf{S}^n_{++}$.) Since the objective function is concave, this is a convex problem if the set $\mathcal{S}$ can be described by a set of linear equality and convex inequality constraints.

First we examine the case in which no prior assumptions are made on R (hence, S), other than $R \succ 0$. In this case the problem (7.4) can be solved analytically. The gradient of the objective is $S^{-1} - Y$, so the optimal S satisfies $S^{-1} = Y$ if $Y \in \mathbf{S}^n_{++}$. (If $Y \notin \mathbf{S}^n_{++}$, the log-likelihood function is unbounded above.) Therefore, when we have no prior assumptions about R, the maximum likelihood estimate of the covariance is, simply, the sample covariance: $\hat{R}_{\rm ml} = Y$.

Now we consider some examples of constraints on R that can be expressed as convex constraints on the information matrix S. We can handle lower and upper (matrix) bounds on R, of the form

$$L \preceq R \preceq U,$$

where L and U are symmetric and positive definite, as

$$U^{-1} \preceq R^{-1} \preceq L^{-1}.$$

A condition number constraint on R,

$$\lambda_{\max}(R) \le \kappa_{\max}\lambda_{\min}(R),$$

can be expressed as

$$\lambda_{\max}(S) \le \kappa_{\max}\lambda_{\min}(S).$$

This is equivalent to the existence of $u > 0$ such that $uI \preceq S \preceq \kappa_{\max} uI$. We can therefore solve the ML problem, with the condition number constraint on R, by solving the convex problem

$$\begin{array}{ll} \text{maximize} & \log\det S - \mathbf{tr}(SY) \\ \text{subject to} & uI \preceq S \preceq \kappa_{\max} uI \end{array} \qquad (7.5)$$

where the variables are $S \in \mathbf{S}^n$ and $u \in \mathbf{R}$.

As another example, suppose we are given bounds on the variance of some linear functions of the underlying random vector y,

$$\mathbf{E}(c_i^T y)^2 \leq \alpha_i, \quad i = 1, \ldots, K.$$

These prior assumptions can be expressed as

$$\mathbf{E}(c_i^T y)^2 = c_i^T R c_i = c_i^T S^{-1} c_i \leq \alpha_i, \quad i = 1, \ldots, K.$$

Since $c_i^T S^{-1} c_i$ is a convex function of S (provided $S \succ 0$, which holds here), these bounds can be imposed in the ML problem.

7.1.2 Maximum a posteriori probability estimation

Maximum a posteriori probability (MAP) estimation can be considered a Bayesian version of maximum likelihood estimation, with a prior probability density on the underlying parameter x. We assume that x (the vector to be estimated) and y (the observation) are random variables with a joint probability density $p(x, y)$. This is in contrast to the statistical estimation setup, where x is a parameter, not a random variable.

The *prior density* of x is given by

$$p_x(x) = \int p(x, y)\, dy.$$

This density represents our prior information about what the values of the vector x might be, before we observe the vector y. Similarly, the prior density of y is given by

$$p_y(y) = \int p(x, y)\, dx.$$

This density represents the prior information about what the measurement or observation vector y will be.

The conditional density of y, given x, is given by

$$p_{y|x}(x, y) = \frac{p(x, y)}{p_x(x)}.$$

In the MAP estimation method, $p_{y|x}$ plays the role of the parameter dependent density p_x in the maximum likelihood estimation setup. The conditional density of x, given y, is given by

$$p_{x|y}(x, y) = \frac{p(x, y)}{p_y(y)} = p_{y|x}(x, y)\frac{p_x(x)}{p_y(y)}.$$

When we substitute the observed value y into $p_{x|y}$, we obtain the *posterior density* of x. It represents our knowledge of x after the observation.

In the MAP estimation method, our estimate of x, given the observation y, is given by

$$\begin{aligned} \hat{x}_{\mathrm{map}} &= \mathrm{argmax}_x p_{x|y}(x,y) \\ &= \mathrm{argmax}_x p_{y|x}(x,y)p_x(x) \\ &= \mathrm{argmax}_x p(x,y). \end{aligned}$$

In other words, we take as estimate of x the value that maximizes the conditional density of x, given the observed value of y. The only difference between this estimate and the maximum likelihood estimate is the second term, $p_x(x)$, appearing here. This term can be interpreted as taking our prior knowledge of x into account. Note that if the prior density of x is uniform over a set C, then finding the MAP estimate is the same as maximizing the likelihood function subject to $x \in C$, which is the ML estimation problem (7.1).

Taking logarithms, we can express the MAP estimate as

$$\hat{x}_{\mathrm{map}} = \mathrm{argmax}_x(\log p_{y|x}(x,y) + \log p_x(x)). \tag{7.6}$$

The first term is essentially the same as the log-likelihood function; the second term penalizes choices of x that are unlikely, according to the prior density (*i.e.*, x with $p_x(x)$ small).

Brushing aside the philosophical differences in setup, the only difference between finding the MAP estimate (via (7.6)) and the ML estimate (via (7.1)) is the presence of an extra term in the optimization problem, associated with the prior density of x. Therefore, for any maximum likelihood estimation problem with concave log-likelihood function, we can add a prior density for x that is log-concave, and the resulting MAP estimation problem will be convex.

Linear measurements with IID noise

Suppose that $x \in \mathbf{R}^n$ and $y \in \mathbf{R}^m$ are related by

$$y_i = a_i^T x + v_i, \quad i = 1, \ldots, m,$$

where v_i are IID with density p_v on $\mathbf{R}$, and x has prior density p_x on $\mathbf{R}^n$. The joint density of x and y is then

$$p(x,y) = p_x(x) \prod_{i=1}^m p_v(y_i - a_i^T x),$$

and the MAP estimate can be found by solving the optimization problem

$$\text{maximize} \quad \log p_x(x) + \textstyle\sum_{i=1}^m \log p_v(y_i - a_i^T x). \tag{7.7}$$

If p_x and p_v are log-concave, this problem is convex. The only difference between the MAP estimation problem (7.7) and the associated ML estimation problem (7.2) is the extra term $\log p_x(x)$.

For example, if v_i are uniform on $[-a, a]$, and the prior distribution of x is Gaussian with mean $\bar{x}$ and covariance Σ, the MAP estimate is found by solving the QP

$$\begin{array}{ll} \text{minimize} & (x-\bar{x})^T \Sigma^{-1} (x-\bar{x}) \\ \text{subject to} & \|Ax - y\|_\infty \leq a, \end{array}$$

with variable x.

MAP with perfect linear measurements

Suppose $x \in \mathbf{R}^n$ is a vector of parameters to be estimated, with prior density p_x. We have m perfect (noise free, deterministic) linear measurements, given by $y = Ax$. In other words, the conditional distribution of y, given x, is a point mass with value one at the point Ax. The MAP estimate can be found by solving the problem

$$\begin{array}{ll} \text{maximize} & \log p_x(x) \\ \text{subject to} & Ax = y. \end{array}$$

If p_x is log-concave, this is a convex problem.

If under the prior distribution, the parameters x_i are IID with density p on $\mathbf{R}$, then the MAP estimation problem has the form

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n \log p(x_i) \\ \text{subject to} & Ax = y, \end{array}$$

which is a least-penalty problem ((6.6), page 304), with penalty function $\phi(u) = -\log p(u)$.

Conversely, we can interpret any least-penalty problem,

$$\begin{array}{ll} \text{minimize} & \phi(x_1) + \cdots + \phi(x_n) \\ \text{subject to} & Ax = b \end{array}$$

as a MAP estimation problem, with m perfect linear measurements (*i.e.*, $Ax = b$) and x_i IID with density

$$p(z) = \frac{e^{-\phi(z)}}{\int e^{-\phi(u)}\, du}.$$

7.2 Nonparametric distribution estimation

We consider a random variable X with values in the finite set $\{\alpha_1, \ldots, \alpha_n\} \subseteq \mathbf{R}$. (We take the values to be in $\mathbf{R}$ for simplicity; the same ideas can be applied when the values are in $\mathbf{R}^k$, for example.) The distribution of X is characterized by $p \in \mathbf{R}^n$, with $\mathbf{prob}(X = \alpha_k) = p_k$. Clearly, p satisfies $p \succeq 0$, $\mathbf{1}^T p = 1$. Conversely, if $p \in \mathbf{R}^n$ satisfies $p \succeq 0$, $\mathbf{1}^T p = 1$, then it defines a probability distribution for a random variable X, defined as $\mathbf{prob}(X = \alpha_k) = p_k$. Thus, the probability simplex

$$\{p \in \mathbf{R}^n \mid p \succeq 0,\ \mathbf{1}^T p = 1\}$$

is in one-to-one correspondence with all possible probability distributions for a random variable X taking values in $\{\alpha_1, \ldots, \alpha_n\}$.

In this section we discuss methods used to estimate the distribution p based on a combination of prior information and, possibly, observations and measurements.

Prior information

Many types of prior information about p can be expressed in terms of linear equality constraints or inequalities. If $f : \mathbf{R} \to \mathbf{R}$ is any function, then

$$\mathbf{E}\, f(X) = \sum_{i=1}^n p_i f(\alpha_i)$$

is a linear function of p. As a special case, if $C \subseteq \mathbf{R}$, then $\mathbf{prob}(X \in C)$ is a linear function of p:

$$\mathbf{prob}(X \in C) = c^T p, \qquad c_i = \left\{ \begin{array}{ll} 1 & \alpha_i \in C \\ 0 & \alpha_i \notin C. \end{array} \right.$$

It follows that known expected values of certain functions (*e.g.*, moments) or known probabilities of certain sets can be incorporated as linear equality constraints on $p \in \mathbf{R}^n$. Inequalities on expected values or probabilities can be expressed as linear inequalities on $p \in \mathbf{R}^n$.

For example, suppose we know that X has mean $\mathbf{E}\, X = \alpha$, second moment $\mathbf{E}\, X^2 = \beta$, and $\mathbf{prob}(X \geq 0) \leq 0.3$. This prior information can be expressed as

$$\mathbf{E}\, X = \sum_{i=1}^n \alpha_i p_i = \alpha, \qquad \mathbf{E}\, X^2 = \sum_{i=1}^n \alpha_i^2 p_i = \beta, \qquad \sum_{\alpha_i \geq 0} p_i \leq 0.3,$$

which are two linear equalities and one linear inequality in p.

We can also include some prior constraints that involve nonlinear functions of p. As an example, the variance of X is given by

$$\mathbf{var}(X) = \mathbf{E}\, X^2 - (\mathbf{E}\, X)^2 = \sum_{i=1}^n \alpha_i^2 p_i - \left(\sum_{i=1}^n \alpha_i p_i \right)^2.$$

The first term is a linear function of p and the second term is concave quadratic in p, so the variance of X is a concave function of p. It follows that a *lower bound* on the variance of X can be expressed as a convex quadratic inequality on p.

As another example, suppose A and B are subsets of $\mathbf{R}$, and consider the conditional probability of A given B:

$$\mathbf{prob}(X \in A | X \in B) = \frac{\mathbf{prob}(X \in A \cap B)}{\mathbf{prob}(X \in B)}.$$

This function is linear-fractional in $p \in \mathbf{R}^n$: it can be expressed as

$$\mathbf{prob}(X \in A | X \in B) = c^T p / d^T p,$$

where

$$c_i = \left\{ \begin{array}{ll} 1 & \alpha_i \in A \cap B \\ 0 & \alpha_i \notin A \cap B \end{array} \right. , \qquad d_i = \left\{ \begin{array}{ll} 1 & \alpha_i \in B \\ 0 & \alpha_i \notin B. \end{array} \right.$$

Therefore we can express the prior constraints

$$l \leq \mathbf{prob}(X \in A | X \in B) \leq u$$

as the linear inequality constraints on p

$$l d^T p \leq c^T p \leq u d^T p.$$

Several other types of prior information can be expressed in terms of nonlinear convex inequalities. For example, the entropy of X, given by

$$-\sum_{i=1}^{n} p_i \log p_i,$$

is a concave function of p, so we can impose a minimum value of entropy as a convex inequality on p. If q represents another distribution, *i.e.*, $q \succeq 0$, $\mathbf{1}^T q = 1$, then the Kullback-Leibler divergence between the distribution q and the distribution p is given by

$$\sum_{i=1}^{n} p_i \log(p_i/q_i),$$

which is convex in p (and q as well; see example 3.19, page 90). It follows that we can impose a maximum Kullback-Leibler divergence between p and a given distribution q, as a convex inequality on p.

In the next few paragraphs we express the prior information about the distribution p as $p \in \mathcal{P}$. We assume that $\mathcal{P}$ can be described by a set of linear equalities and convex inequalities. We include in the prior information $\mathcal{P}$ the basic constraints $p \succeq 0$, $\mathbf{1}^T p = 1$.

Bounding probabilities and expected values

Given prior information about the distribution, say $p \in \mathcal{P}$, we can compute upper or lower bounds on the expected value of a function, or probability of a set. For example to determine a lower bound on $\mathbf{E}\, f(X)$ over all distributions that satisfy the prior information $p \in \mathcal{P}$, we solve the convex problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} f(\alpha_i) p_i \\ \text{subject to} & p \in \mathcal{P}. \end{array}$$

Maximum likelihood estimation

We can use maximum likelihood estimation to estimate p based on observations from the distribution. Suppose we observe N independent samples $x_1, \ldots, x_N$ from the distribution. Let k_i denote the number of these samples with value α_i, so that $k_1 + \cdots + k_n = N$, the total number of observed samples. The log-likelihood function is then

$$l(p) = \sum_{i=1}^{n} k_i \log p_i,$$

which is a concave function of p. The maximum likelihood estimate of p can be found by solving the convex problem

$$\begin{array}{ll} \text{maximize} & l(p) = \sum_{i=1}^n k_i \log p_i \\ \text{subject to} & p \in \mathcal{P}, \end{array}$$

with variable p.

Maximum entropy

The maximum entropy distribution consistent with the prior assumptions can be found by solving the convex problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n p_i \log p_i \\ \text{subject to} & p \in \mathcal{P}. \end{array}$$

Enthusiasts describe the maximum entropy distribution as the most equivocal or most random, among those consistent with the prior information.

Minimum Kullback-Leibler divergence

We can find the distribution p that has minimum Kullback-Leibler divergence from a given prior distribution q, among those consistent with prior information, by solving the convex problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n p_i \log(p_i/q_i) \\ \text{subject to} & p \in \mathcal{P}, \end{array}$$

Note that when the prior distribution is the uniform distribution, *i.e.*, $q = (1/n)\mathbf{1}$, this problem reduces to the maximum entropy problem.

Example 7.2 We consider a probability distribution on 100 equidistant points α_i in the interval $[-1, 1]$. We impose the following prior assumptions:

$$\begin{array}{rcl} \mathbf{E}\, X & \in & [-0.1, 0.1] \\ \mathbf{E}\, X^2 & \in & [0.5, 0.6] \\ \mathbf{E}(3X^3 - 2X) & \in & [-0.3, -0.2] \\ \mathbf{prob}(X < 0) & \in & [0.3, 0.4]. \end{array} \tag{7.8}$$

Along with the constraints $\mathbf{1}^T p = 1$, $p \succeq 0$, these constraints describe a polyhedron of probability distributions.

Figure 7.2 shows the maximum entropy distribution that satisfies these constraints. The maximum entropy distribution satisfies

$$\begin{array}{rcl} \mathbf{E}\, X & = & 0.056 \\ \mathbf{E}\, X^2 & = & 0.5 \\ \mathbf{E}(3X^3 - 2X) & = & -0.2 \\ \mathbf{prob}(X < 0) & = & 0.4. \end{array}$$

To illustrate bounding probabilities, we compute upper and lower bounds on the cumulative distribution $\mathbf{prob}(X \le \alpha_i)$, for $i = 1, \ldots, 100$. For each value of i,

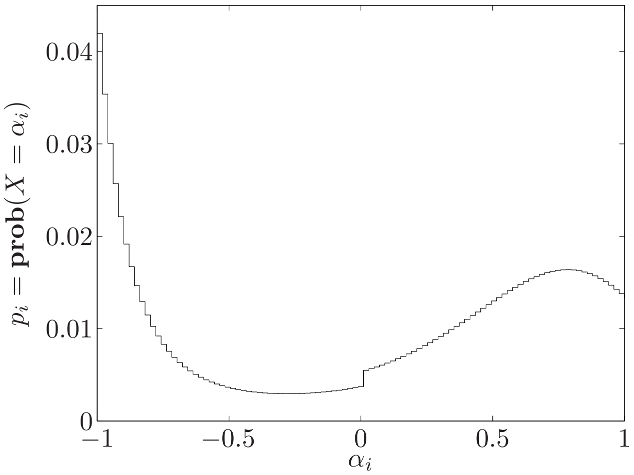

Figure 7.2 Maximum entropy distribution that satisfies the constraints (7.8).

we solve two LPs: one that maximizes $\mathbf{prob}(X \leq \alpha_i)$, and one that minimizes $\mathbf{prob}(X \leq \alpha_i)$, over all distributions consistent with the prior assumptions (7.8). The results are shown in figure 7.3. The upper and lower curves show the upper and lower bounds, respectively; the middle curve shows the cumulative distribution of the maximum entropy distribution.

Example 7.3 *Bounding risk probability with known marginal distributions.* Suppose X and Y are two random variables that give the return on two investments. We assume that X takes values in $\{\alpha_1, \ldots, \alpha_n\} \subseteq \mathbf{R}$ and Y takes values in $\{\beta_1, \ldots, \beta_m\} \subseteq \mathbf{R}$, with $p_{ij} = \mathbf{prob}(X = \alpha_i, Y = \beta_j)$. The marginal distributions of the two returns X and Y are known, *i.e.*,

$$\sum_{j=1}^{m} p_{ij} = r_i, \quad i = 1, \ldots, n, \qquad \sum_{i=1}^{n} p_{ij} = q_j, \quad j = 1, \ldots, m, \tag{7.9}$$

but otherwise nothing is known about the joint distribution p. This defines a polyhedron of joint distributions consistent with the given marginals.

Now suppose we make both investments, so our total return is the random variable $X + Y$. We are interested in computing an upper bound on the probability of some level of loss, or low return, *i.e.*, $\mathbf{prob}(X + Y < \gamma)$. We can compute a tight upper bound on this probability by solving the LP

$$\begin{array}{ll} \text{maximize} & \sum \{p_{ij} \mid \alpha_i + \beta_j < \gamma\} \\ \text{subject to} & (7.9), \quad p_{ij} \geq 0, \quad i = 1, \ldots n, \quad j = 1, \ldots, m. \end{array}$$

The optimal value of this LP is the maximum probability of loss. The optimal solution $p^\star$ is the joint distribution, consistent with the given marginal distributions, that maximizes the probability of the loss.

The same method can be applied to a derivative of the two investments. Let $R(X, Y)$ be the return of the derivative, where $R : \mathbf{R}^2 \to \mathbf{R}$. We can compute sharp lower

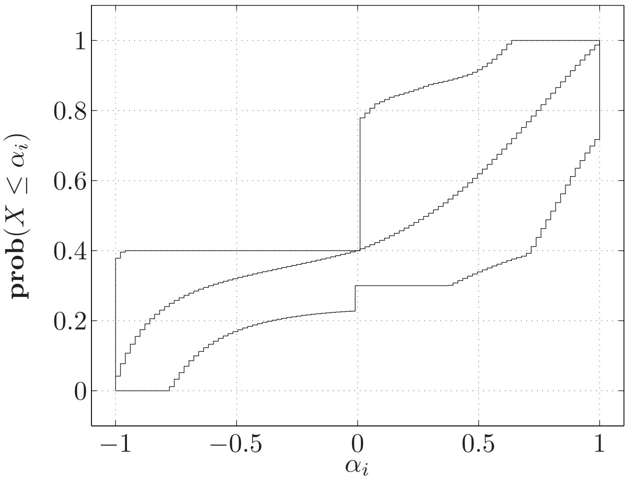

Figure 7.3 The top and bottom curves show the maximum and minimum possible values of the cumulative distribution function, $\mathbf{prob}(X \leq \alpha_i)$, over all distributions that satisfy (7.8). The middle curve is the cumulative distribution of the maximum entropy distribution that satisfies (7.8).

and upper bounds on $\mathbf{prob}(R < \gamma)$ by solving a similar LP, with objective function

$$\sum \{p_{ij} \mid R(\alpha_i, \beta_j) < \gamma\},$$

which we can minimize and maximize.

7.3 Optimal detector design and hypothesis testing

Suppose X is a random variable with values in $\{1, \ldots, n\}$, with a distribution that depends on a parameter $\theta \in \{1, \ldots, m\}$. The distributions of X, for the m possible values of θ, can be represented by a matrix $P \in \mathbf{R}^{n \times m}$, with elements

$$p_{kj} = \mathbf{prob}(X = k \mid \theta = j).$$

The jth column of P gives the probability distribution associated with the parameter value $\theta = j$.

We consider the problem of estimating θ, based on an observed sample of X. In other words, the sample X is generated from one of the m possible distributions, and we are to guess which one. The m values of θ are called *hypotheses*, and guessing which hypothesis is correct (*i.e.*, which distribution generated the observed sample X) is called *hypothesis testing*. In many cases one of the hypotheses corresponds to some normal situation, and each of the other hypotheses corresponds to some abnormal event. In this case hypothesis testing can be interpreted as observing a value of X, and then guessing whether or not an abnormal event has occurred, and if so, which one. For this reason hypothesis testing is also called *detection*.

In most cases there is no significance to the ordering of the hypotheses; they are simply m different hypotheses, arbitrarily labeled $\theta = 1, \ldots, m$. If $\hat{\theta} = \theta$, where $\hat{\theta}$ denotes the estimate of θ, then we have correctly guessed the parameter value θ. If $\hat{\theta} \neq \theta$, then we have (incorrectly) guessed the parameter value θ; we have mistaken $\hat{\theta}$ for θ. In other cases, there is significance in the ordering of the hypotheses. In this case, an event such as $\hat{\theta} > \theta$, *i.e.*, the event that we overestimate θ, is meaningful.

It is also possible to parametrize θ by values other than $\{1, \ldots, m\}$, say as $\theta \in \{\theta_1, \ldots, \theta_m\}$, where θ_i are (distinct) values. These values could be real numbers, or vectors, for example, specifying the mean and variance of the kth distribution. In this case, a quantity such as $\|\hat{\theta} - \theta\|$, which is the norm of the parameter estimation error, is meaningful.

7.3.1 Deterministic and randomized detectors

A (deterministic) *estimator* or *detector* is a function ψ from $\{1, \ldots, n\}$ (the set of possible observed values) into $\{1, \ldots, m\}$ (the set of hypotheses). If X is observed to have value k, then our guess for the value of θ is $\hat{\theta} = \psi(k)$. One obvious deterministic detector is the *maximum likelihood detector*, given by

$$\hat{\theta} = \psi_{\mathrm{ml}}(k) = \underset{j}{\operatorname{argmax}}\, p_{kj}. \qquad (7.10)$$

When we observe the value $X = k$, the maximum likelihood estimate of θ is a value that maximizes the probability of observing $X = k$, over the set of possible distributions.

We will consider a generalization of the deterministic detector, in which the estimate of θ, given an observed value of X, is random. A *randomized detector* of θ is a random variable $\hat{\theta} \in \{1, \ldots, m\}$, with a distribution that depends on the observed value of X. A randomized detector can be defined in terms of a matrix $T \in \mathbf{R}^{m \times n}$ with elements

$$t_{ik} = \mathbf{prob}(\hat{\theta} = i \mid X = k).$$

The interpretation is as follows: if we observe $X = k$, then the detector gives $\hat{\theta} = i$ with probability t_{ik}. The kth column of T, which we will denote t_k, gives the probability distribution of $\hat{\theta}$, when we observe $X = k$. If each column of T is a unit vector, then the randomized detector is a deterministic detector, *i.e.*, $\hat{\theta}$ is a (deterministic) function of the observed value of X.

At first glance, it seems that intentionally introducing additional randomization into the estimation or detection process can only make the estimator worse. But we will see below examples in which a randomized detector outperforms all deterministic estimators.

We are interested in designing the matrix T that defines the randomized detector. Obviously the columns t_k of T must satisfy the (linear equality and inequality) constraints

$$t_k \succeq 0, \qquad \mathbf{1}^T t_k = 1. \qquad (7.11)$$

7.3.2 Detection probability matrix

For the randomized detector defined by the matrix T, we define the *detection probability matrix* as $D = TP$. We have

$$D_{ij} = (TP)_{ij} = \mathbf{prob}(\hat{\theta} = i \mid \theta = j),$$

so D_{ij} is the probability of guessing $\hat{\theta} = i$, when in fact $\theta = j$. The $m \times m$ detection probability matrix D characterizes the performance of the randomized detector defined by T. The diagonal entry D_{ii} is the probability of guessing $\hat{\theta} = i$ when $\theta = i$, *i.e.*, the probability of correctly detecting that $\theta = i$. The off-diagonal entry D_{ij} (with $i \neq j$) is the probability of mistaking $\theta = i$ for $\theta = j$, *i.e.*, the probability that our guess is $\hat{\theta} = i$, when in fact $\theta = j$. If $D = I$, the detector is perfect: no matter what the parameter θ is, we correctly guess $\hat{\theta} = \theta$.

The diagonal entries of D, arranged in a vector, are called the *detection probabilities*, and denoted P^{d}:

$$P^{\mathrm{d}}_i = D_{ii} = \mathbf{prob}(\hat{\theta} = i \mid \theta = i).$$

The *error probabilities* are the complements, and are denoted P^{e}:

$$P^{\mathrm{e}}_i = 1 - D_{ii} = \mathbf{prob}(\hat{\theta} \neq i \mid \theta = i).$$

Since the columns of the detection probability matrix D add up to one, we can express the error probabilities as

$$P^{\mathrm{e}}_i = \sum_{j \neq i} D_{ji}.$$

7.3.3 Optimal detector design

In this section we show that a wide variety of objectives for detector design are linear, affine, or convex piecewise-linear functions of D, and therefore also of T (which is the optimization variable). Similarly, a variety of constraints for detector design can be expressed in terms of linear inequalities in D. It follows that a wide variety of optimal detector design problems can be expressed as LPs. We will see in §7.3.4 that some of these LPs have simple solutions; in this section we simply formulate the problem.

Limits on errors and detection probabilities

We can impose a lower bound on the probability of correctly detecting the jth hypothesis,

$$P^{\mathrm{d}}_j = D_{jj} \geq L_j,$$

which is a linear inequality in D (hence, T). Similarly, we can impose a maximum allowable probability for mistaking $\theta = i$ for $\theta = j$:

$$D_{ij} \leq U_{ij},$$

which are also linear constraints on T. We can take any of the detection probabilities as an objective to be maximized, or any of the error probabilities as an objective to be minimized.

Minimax detector design

We can take as objective (to be minimized) the *minimax error probability*, $\max_j P^{\mathrm{e}}_j$, which is a piecewise-linear convex function of D (hence, also of T). With this as the only objective, we have the problem of minimizing the maximum probability of detection error,

$$\begin{array}{ll} \text{minimize} & \max_j P^{\mathrm{e}}_j \\ \text{subject to} & t_k \succeq 0, \quad \mathbf{1}^T t_k = 1, \quad k = 1, \ldots, n, \end{array}$$

where the variables are $t_1, \ldots, t_n \in \mathbf{R}^m$. This can be reformulated as an LP. The minimax detector minimizes the worst-case (largest) probability of error over all m hypotheses.

We can, of course, add further constraints to the minimax detector design problem.

Bayes detector design

In Bayes detector design, we have a prior distribution for the hypotheses, given by $q \in \mathbf{R}^m$, where

$$q_i = \mathbf{prob}(\theta = i).$$

In this case, the probabilities p_{ij} are interpreted as conditional probabilities of X, given θ. The probability of error for the detector is then given by $q^T P^{\mathrm{e}}$, which is an affine function of T. The Bayes optimal detector is the solution of the LP

$$\begin{array}{ll} \text{minimize} & q^T P^{\mathrm{e}} \\ \text{subject to} & t_k \succeq 0, \quad \mathbf{1}^T t_k = 1, \quad k = 1, \ldots, n. \end{array}$$

We will see in §7.3.4 that this problem has a simple analytical solution.

One special case is when $q = (1/m)\mathbf{1}$. In this case the Bayes optimal detector minimizes the average probability of error, where the (unweighted) average is over the hypotheses. In §7.3.4 we will see that the maximum likelihood detector (7.10) is optimal for this problem.

Bias, mean-square error, and other quantities

In this section we assume that the ordering of the values of θ have some significance, *i.e.*, that the value $\theta = i$ can be interpreted as a larger value of the parameter than $\theta = j$, when $i > j$. This might be the case, for example, when $\theta = i$ corresponds to the hypothesis that i events have occurred. Here we may be interested in quantities such as

$$\mathbf{prob}(\hat{\theta} > \theta \mid \theta = i),$$

which is the probability that we overestimate θ when $\theta = i$. This is an affine function of D:

$$\mathbf{prob}(\hat{\theta} > \theta \mid \theta = i) = \sum_{j>i} D_{ji},$$

so a maximum allowable value for this probability can be expressed as a linear inequality on D (hence, T). As another example, the probability of misclassifying θ by more than one, when $\theta = i$,

$$\mathbf{prob}(|\hat{\theta} - \theta| > 1 \mid \theta = i) = \sum_{|j-i|>1} D_{ji},$$

is also a linear function of D.

We now suppose that the parameters have values $\{\theta_1, \ldots, \theta_m\} \subseteq \mathbf{R}$. The estimation or detection (parameter) error is then given by $\hat{\theta} - \theta$, and a number of quantities of interest are given by linear functions of D. Examples include:

- *Bias.* The bias of the detector, when $\theta = \theta_i$, is given by the linear function

  $$\mathop{\mathbf{E}}_i(\hat{\theta} - \theta) = \sum_{j=1}^{m} (\theta_j - \theta_i) D_{ji},$$

  where the subscript on $\mathbf{E}$ means the expectation is with respect to the distribution of the hypothesis $\theta = \theta_i$.

- *Mean square error.* The mean square error of the detector, when $\theta = \theta_i$, is given by the linear function

  $$\mathop{\mathbf{E}}_i(\hat{\theta} - \theta)^2 = \sum_{j=1}^{m} (\theta_j - \theta_i)^2 D_{ji}.$$

- *Average absolute error.* The average absolute error of the detector, when $\theta = \theta_i$, is given by the linear function

  $$\mathop{\mathbf{E}}_i |\hat{\theta} - \theta| = \sum_{j=1}^{m} |\theta_j - \theta_i| D_{ji}.$$

7.3.4 Multicriterion formulation and scalarization

The optimal detector design problem can be considered a multicriterion problem, with the constraints (7.11), and the $m(m-1)$ objectives given by the off-diagonal entries of D, which are the probabilities of the different types of detection error:

$$\begin{array}{lll} \text{minimize (w.r.t. } \mathbf{R}_+^{m(m-1)}) & D_{ij}, \quad i, \ j = 1, \ldots, m, \quad i \neq j \\ \text{subject to} & t_k \succeq 0, \quad \mathbf{1}^T t_k = 1, \quad k = 1, \ldots, n, \end{array} \qquad (7.12)$$

with variables $t_1, \ldots, t_n \in \mathbf{R}^m$. Since each objective D_{ij} is a linear function of the variables, this is a multicriterion linear program.

We can scalarize this multicriterion problem by forming the weighted sum objective

$$\sum_{i,j=1}^{m} W_{ij} D_{ij} = \mathbf{tr}(W^T D)$$

where the weight matrix $W \in \mathbf{R}^{m\times m}$ satisfies

$$W_{ii} = 0, \quad i = 1,\ldots,m, \qquad W_{ij} > 0, \quad i,\ j = 1,\ldots,m, \quad i \neq j.$$

This objective is a weighted sum of the $m(m-1)$ error probabilities, with weight W_{ij} associated with the error of guessing $\hat{\theta} = i$ when in fact $\theta = j$. The weight matrix is sometimes called the *loss matrix*.

To find a Pareto optimal point for the multicriterion problem (7.12), we form the scalar optimization problem

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(W^T D) \\ \text{subject to} & t_k \succeq 0, \quad \mathbf{1}^T t_k = 1, \quad k = 1,\ldots,n, \end{array} \tag{7.13}$$

which is an LP. This LP is separable in the variables $t_1,\ldots,t_n$. The objective can be expressed as a sum of (linear) functions of t_k:

$$\mathbf{tr}(W^T D) = \mathbf{tr}(W^T TP) = \mathbf{tr}(PW^T T) = \sum_{k=1}^n c_k^T t_k,$$

where c_k is the kth column of WP^T. The constraints are separable (*i.e.*, we have separate constraints on each t_i). Therefore we can solve the LP (7.13) by separately solving

$$\begin{array}{ll} \text{minimize} & c_k^T t_k \\ \text{subject to} & t_k \succeq 0, \quad \mathbf{1}^T t_k = 1, \end{array}$$

for $k = 1,\ldots,n$. Each of these LPs has a simple analytical solution (see exercise 4.8). We first find an index q such that $c_{kq} = \min_j c_{kj}$. Then we take $t_k^\star = e_q$. This optimal point corresponds to a deterministic detector: when $X = k$ is observed, our estimate is

$$\hat{\theta} = \underset{j}{\operatorname{argmin}}(WP^T)_{jk}. \tag{7.14}$$

Thus, for every weight matrix W with positive off-diagonal elements we can find a deterministic detector that minimizes the weighted sum objective. This seems to suggest that randomized detectors are not needed, but we will see this is not the case. The Pareto optimal trade-off surface for the multicriterion LP (7.12) is piecewise-linear; the deterministic detectors of the form (7.14) correspond to the vertices on the Pareto optimal surface.

MAP and ML detectors

Consider a Bayes detector design with prior distribution q. The mean probability of error is

$$q^T P^{\mathrm{e}} = \sum_{j=1}^m q_j \sum_{i\neq j} D_{ij} = \sum_{i,j=1}^m W_{ij} D_{ij},$$

if we define the weight matrix W as

$$W_{ij} = q_j, \quad i,\ j = 1,\ldots,m, \quad i \neq j, \qquad W_{ii} = 0, \quad i = 1,\ldots,m.$$

Thus, a Bayes optimal detector is given by the deterministic detector (7.14), with

$$(WP^T)_{jk} = \sum_{i \neq j} q_i p_{ki} = \sum_{i=1}^{m} q_i p_{ki} - q_j p_{kj}.$$

The first term is independent of j, so the optimal detector is simply

$$\hat{\theta} = \underset{j}{\operatorname{argmax}} (p_{kj} q_j),$$

when $X = k$ is observed. The solution has a simple interpretation: Since $p_{kj}q_j$ gives the probability that $\theta = j$ and $X = k$, this detector is a maximum a posteriori probability (MAP) detector.

For the special case $q = (1/m)\mathbf{1}$, *i.e.*, a uniform prior distribution on θ, this MAP detector reduces to a maximum likelihood (ML) detector:

$$\hat{\theta} = \underset{j}{\operatorname{argmax}}\, p_{kj}.$$

Thus, a maximum likelihood detector minimizes the (unweighted) average or mean probability of error.

7.3.5 Binary hypothesis testing

As an illustration, we consider the special case $m = 2$, which is called *binary hypothesis testing*. The random variable X is generated from one of two distributions, which we denote $p \in \mathbf{R}^n$ and $q \in \mathbf{R}^n$, to simplify the notation. Often the hypothesis $\theta = 1$ corresponds to some normal situation, and the hypothesis $\theta = 2$ corresponds to some abnormal event that we are trying to detect. If $\hat{\theta} = 1$, we say the test is negative (*i.e.*, we guess that the event did not occur); if $\hat{\theta} = 2$, we say the test is positive (*i.e.*, we guess that the event did occur).

The detection probability matrix $D \in \mathbf{R}^{2\times 2}$ is traditionally expressed as

$$D = \left[\begin{array}{cc} 1 - P_{\rm fp} & P_{\rm fn} \\ P_{\rm fp} & 1 - P_{\rm fn} \end{array} \right].$$

Here $P_{\rm fn}$ is the probability of a *false negative* (*i.e.*, the test is negative when in fact the event has occurred) and $P_{\rm fp}$ is the probability of a *false positive* (*i.e.*, the test is positive when in fact the event has not occurred), which is also called the *false alarm probability*. The optimal detector design problem is a bi-criterion problem, with objectives $P_{\rm fn}$ and $P_{\rm fp}$.

The optimal trade-off curve between $P_{\rm fn}$ and $P_{\rm fp}$ is called the *receiver operating characteristic* (ROC), and is determined by the distributions p and q. The ROC can be found by scalarizing the bi-criterion problem, as described in §7.3.4. For the weight matrix W, an optimal detector (7.14) is

$$\hat{\theta} = \begin{cases} 1 & W_{21} p_k > W_{12} q_k \\ 2 & W_{21} p_k \leq W_{12} q_k \end{cases}$$

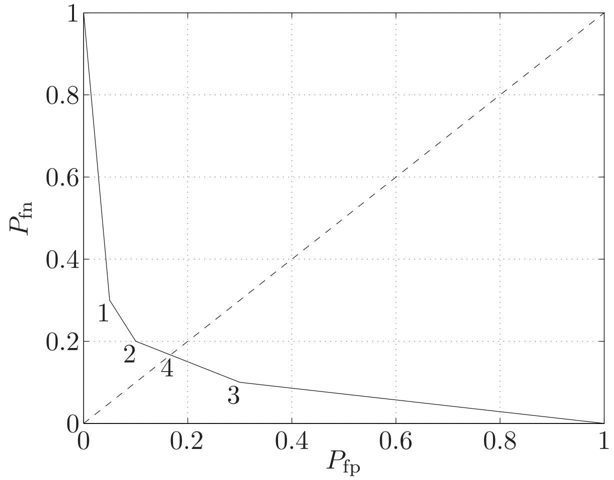

Figure 7.4 Optimal trade-off curve between probability of a false negative, and probability of a false positive test result, for the matrix P given in (7.15). The vertices of the trade-off curve, labeled 1–3, correspond to deterministic detectors; the point labeled 4, which is a randomized detector, is the minimax detector. The dashed line shows $P_{\rm fn} = P_{\rm fp}$, the points where the error probabilities are equal.

when $X = k$ is observed. This is called a *likelihood ratio threshold test*: if the ratio p_k/q_k is more than the threshold W_{12}/W_{21}, the test is negative (*i.e.*, $\hat{\theta} = 1$); otherwise the test is positive. By choosing different values of the threshold, we obtain (deterministic) Pareto optimal detectors that give different levels of false positive versus false negative error probabilities. This result is known as the Neyman-Pearson lemma.

The likelihood ratio detectors do not give all the Pareto optimal detectors; they are the vertices of the optimal trade-off curve, which is piecewise-linear.

Example 7.4 We consider a binary hypothesis testing example with $n = 4$, and

$$P = \begin{bmatrix} 0.70 & 0.10 \\ 0.20 & 0.10 \\ 0.05 & 0.70 \\ 0.05 & 0.10 \end{bmatrix}. \qquad (7.15)$$

The optimal trade-off curve between $P_{\rm fn}$ and $P_{\rm fp}$, *i.e.*, the receiver operating curve, is shown in figure 7.4. The left endpoint corresponds to the detector which is always negative, independent of the observed value of X; the right endpoint corresponds to the detector that is always positive. The vertices labeled 1, 2, and 3 correspond to the deterministic detectors

$$\begin{aligned} T^{(1)} &= \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \\ T^{(2)} &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \end{aligned}$$

$$T^{(3)} \quad = \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix},$$

respectively. The point labeled 4 corresponds to the nondeterministic detector

$$T^{(4)} = \begin{bmatrix} 1 & 2/3 & 0 & 0 \\ 0 & 1/3 & 1 & 1 \end{bmatrix},$$

which is the minimax detector. This minimax detector yields equal probability of a false positive and false negative, which in this case is 1/6. Every deterministic detector has either a false positive or false negative probability that exceeds 1/6, so this is an example where a randomized detector outperforms every deterministic detector.

7.3.6 Robust detectors

So far we have assumed that P, which gives the distribution of the observed variable X, for each value of the parameter θ, is known. In this section we consider the case where these distributions are not known, but certain prior information about them is given. We assume that $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of possible distributions. With a randomized detector characterized by T, the detection probability matrix D now depends on the particular value of P. We will judge the error probabilities by their worst-case values, over $P \in \mathcal{P}$. We define the *worst-case detection probability matrix* $D^{\rm wc}$ as

$$D^{\rm wc}_{ij} = \sup_{P \in \mathcal{P}} D_{ij}, \quad i,\ j = 1, \ldots, m, \quad i \neq j$$

and

$$D^{\rm wc}_{ii} = \inf_{P \in \mathcal{P}} D_{ii}, \quad i = 1, \ldots, m.$$

The off-diagonal entries give the largest possible probability of errors, and the diagonal entries give the smallest possible probability of detection, over $P \in \mathcal{P}$. Note that $\sum_{i=1}^n D^{\rm wc}_{ij} \neq 1$ in general, *i.e.*, the columns of a worst-case detection probability matrix do not necessarily add up to one.

We define the worst-case probability of error as

$$P^{\rm wce}_i = 1 - D^{\rm wc}_{ii}.$$

Thus, $P^{\rm wce}_i$ is the largest probability of error, when $\theta = i$, over all possible distributions in $\mathcal{P}$.

Using the worst-case detection probability matrix, or the worst-case probability of error vector, we can develop various robust versions of detector design problems. In the rest of this section we concentrate on the robust minimax detector design problem, as a generic example that illustrates the ideas.

We define the *robust minimax detector* as the detector that minimizes the worst-case probability of error, over all hypotheses, *i.e.*, minimizes the objective

$$\max_i P^{\rm wce}_i = \max_{i=1,\ldots,m} \sup_{P \in \mathcal{P}} (1 - (TP)_{ii}) = 1 - \min_{i=1,\ldots,m} \inf_{P \in \mathcal{P}} (TP)_{ii}.$$

The robust minimax detector minimizes the worst possible probability of error, over all m hypotheses, and over all $P \in \mathcal{P}$.

Robust minimax detector for finite $\mathcal{P}$

When the set of possible distributions is finite, the robust minimax detector design problem is readily formulated as an LP. With $\mathcal{P} = \{P_1, \dots, P_k\}$, we can find the robust minimax detector by solving

$$\begin{array}{ll} \text{maximize} & \min_{i=1,\dots,m} \inf_{P\in\mathcal{P}} (TP)_{ii} = \min_{i=1,\dots,m} \min_{j=1,\dots,k} (TP_j)_{ii} \\ \text{subject to} & t_i \succeq 0, \quad \mathbf{1}^T t_i = 1, \quad i = 1, \dots, n, \end{array}$$

The objective is piecewise-linear and concave, so this problem can be expressed as an LP. Note that we can just as well consider $\mathcal{P}$ to be the polyhedron $\mathbf{conv}\,\mathcal{P}$; the associated worst-case detection matrix, and robust minimax detector, are the same.

Robust minimax detector for polyhedral $\mathcal{P}$

It is also possible to efficiently formulate the robust minimax detector problem as an LP when $\mathcal{P}$ is a polyhedron described by linear equality and inequality constraints. This formulation is less obvious, and relies on a dual representation of $\mathcal{P}$.

To simplify the discussion, we assume that $\mathcal{P}$ has the form

$$\mathcal{P} = \left\{ P = [p_1 \ \cdots \ p_m] \;\middle|\; A_k p_k = b_k, \ \mathbf{1}^T p_k = 1, \ p_k \succeq 0 \right\}. \tag{7.16}$$

In other words, for each distribution p_k, we are given some expected values $A_k p_k = b_k$. (These might represent known moments, probabilities, etc.) The extension to the case where we are given inequalities on expected values is straightforward.

The robust minimax design problem is

$$\begin{array}{ll} \text{maximize} & \gamma \\ \text{subject to} & \inf\{\tilde{t}_i^T p \mid A_i p = b_i, \ \mathbf{1}^T p = 1, \ p \succeq 0\} \geq \gamma, \quad i = 1, \dots, m \\ & t_i \succeq 0, \quad \mathbf{1}^T t_i = 1, \quad i = 1, \dots, n, \end{array}$$

where $\tilde{t}_i^T$ denotes the ith row of T (so that $(TP)_{ii} = \tilde{t}_i^T p_i$). By LP duality,

$$\inf\{\tilde{t}_i^T p \mid A_i p = b_i, \ \mathbf{1}^T p = 1, \ p \succeq 0\} = \sup\{\nu^T b_i + \mu \mid A_i^T \nu + \mu\mathbf{1} \preceq \tilde{t}_i\}.$$

Using this, the robust minimax detector design problem can be expressed as the LP

$$\begin{array}{ll} \text{maximize} & \gamma \\ \text{subject to} & \nu_i^T b_i + \mu_i \geq \gamma, \quad i = 1, \dots, m \\ & A_i^T \nu_i + \mu_i \mathbf{1} \preceq \tilde{t}_i, \quad i = 1, \dots, m \\ & t_i \succeq 0, \quad \mathbf{1}^T t_i = 1, \quad i = 1, \dots, n, \end{array}$$

with variables $\nu_1, \dots, \nu_m$, $\mu_1, \dots, \mu_n$, and T (which has columns t_i and rows $\tilde{t}_i^T$).

Example 7.5 *Robust binary hypothesis testing.* Suppose $m = 2$ and the set $\mathcal{P}$ in (7.16) is defined by

$$A_1 = A_2 = A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ a_1^2 & a_2^2 & \cdots & a_n^2 \end{bmatrix}, \qquad b_1 = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \qquad b_2 = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Designing a robust minimax detector for this set $\mathcal{P}$ can be interpreted as a binary hypothesis testing problem: based on an observation of a random variable $X \in \{a_1, \dots, a_n\}$, choose between the following two hypotheses:

1. $\mathbf{E}\,X = \alpha_1$, $\mathbf{E}\,X^2 = \alpha_2$
2. $\mathbf{E}\,X = \beta_1$, $\mathbf{E}\,X^2 = \beta_2$.

Let $\tilde{t}^T$ denote the first row of T (and so, $(\mathbf{1}-\tilde{t})^T$ is the second row). For given $\tilde{t}$, the worst-case probabilities of correct detection are

$$\begin{array}{rcl} D_{11}^{\mathrm{wc}} & = & \inf \left\{ \tilde{t}^T p \;\middle|\; \sum_{i=1}^n a_i p_i = \alpha_1,\ \sum_{i=1}^n a_i^2 p_i = \alpha_2,\ \mathbf{1}^T p = 1,\ p \succeq 0 \right\} \\ D_{22}^{\mathrm{wc}} & = & \inf \left\{ (\mathbf{1}-\tilde{t})^T p \;\middle|\; \sum_{i=1}^n a_i p_i = \beta_1,\ \sum_{i=1}^n a_i^2 p_i = \beta_2,\ \mathbf{1}^T p = 1,\ p \succeq 0 \right\}. \end{array}$$

Using LP duality we can express D_{11}^{wc} as the optimal value of the LP

$$\begin{array}{ll} \text{maximize} & z_0 + z_1\alpha_1 + z_2\alpha_2 \\ \text{subject to} & z_0 + a_i z_1 + a_i^2 z_2 \le \tilde{t}_i, \quad i = 1,\ldots,n, \end{array}$$

with variables z_0, z_1, $z_2 \in \mathbf{R}$. Similarly D_{22}^{wc} is the optimal value of the LP

$$\begin{array}{ll} \text{maximize} & w_0 + w_1\beta_1 + w_2\beta_2 \\ \text{subject to} & w_0 + a_i w_1 + a_i^2 w_2 \le 1 - \tilde{t}_i, \quad i = 1,\ldots,n, \end{array}$$

with variables w_0, w_1, $w_2 \in \mathbf{R}$. To obtain the minimax detector, we have to maximize the minimum of D_{11}^{wc} and D_{22}^{wc}, *i.e.*, solve the LP

$$\begin{array}{ll} \text{maximize} & \gamma \\ \text{subject to} & z_0 + z_1\alpha_2 + z_2\alpha_2 \ge \gamma \\ & w_0 + \beta_1 w_1 + \beta_2 w_2 \ge \gamma \\ & z_0 + z_1 a_i + z_2 a_i^2 \le \tilde{t}_i, \quad i = 1,\ldots,n \\ & w_0 + w_1 a_i + w_2 a_i^2 \le 1 - \tilde{t}_i, \quad i = 1,\ldots,n \\ & 0 \preceq \tilde{t} \preceq \mathbf{1}. \end{array}$$

The variables are z_0, z_1, z_2, w_0, w_1, w_2 and $\tilde{t}$.

7.4 Chebyshev and Chernoff bounds

In this section we consider two types of classical bounds on the probability of a set, and show that generalizations of each can be cast as convex optimization problems. The original classical bounds correspond to simple convex optimization problems with analytical solutions; the convex optimization formulation of the general cases allow us to compute better bounds, or bounds for more complex situations.

7.4.1 Chebyshev bounds

Chebyshev bounds give an upper bound on the probability of a set based on known expected values of certain functions (*e.g.*, mean and variance). The simplest example is Markov's inequality: If X is a random variable on $\mathbf{R}_+$ with $\mathbf{E}\,X = \mu$, then we have $\mathbf{prob}(X \geq 1) \leq \mu$, no matter what the distribution of X is. Another simple example is Chebyshev's bound: If X is a random variable on $\mathbf{R}$ with $\mathbf{E}\,X = \mu$ and $\mathbf{E}(X-\mu)^2 = \sigma^2$, then we have $\mathbf{prob}(|X-\mu| \geq 1) \leq \sigma^2$, again no matter what the distribution of X is. The idea behind these simple bounds can be generalized to a setting in which convex optimization is used to compute a bound on the probability.

Let X be a random variable on $S \subseteq \mathbf{R}^m$, and $C \subseteq S$ be the set for which we want to bound $\mathbf{prob}(X \in C)$. Let 1_C denote the 0-1 indicator function of the set C, *i.e.*, $1_C(z) = 1$ if $z \in C$ and $1_C(z) = 0$ if $z \notin C$.

Our prior knowledge of the distribution consists of known expected values of some functions:

$$\mathbf{E}\,f_i(X) = a_i, \quad i = 1, \ldots, n,$$

where $f_i : \mathbf{R}^m \to \mathbf{R}$. We take f_0 to be the constant function with value one, for which we always have $\mathbf{E}\,f_0(X) = a_0 = 1$. Consider a linear combination of the functions f_i, given by

$$f(z) = \sum_{i=0}^{n} x_i f_i(z),$$

where $x_i \in \mathbf{R}$, $i = 0, \ldots, n$. From our knowledge of $\mathbf{E}\,f_i(X)$, we have $\mathbf{E}\,f(X) = a^T x$.

Now suppose that f satisfies the condition $f(z) \geq 1_C(z)$ for all $z \in S$, *i.e.*, f is pointwise greater than or equal to the indicator function of C (on S). Then we have

$$\mathbf{E}\,f(X) = a^T x \geq \mathbf{E}\,1_C(X) = \mathbf{prob}(X \in C).$$

In other words, $a^T x$ is an upper bound on $\mathbf{prob}(X \in C)$, valid for all distributions supported on S, with $\mathbf{E}\,f_i(X) = a_i$.

We can search for the best such upper bound on $\mathbf{prob}(X \in C)$, by solving the problem

$$\begin{array}{ll} \text{minimize} & x_0 + a_1 x_1 + \cdots + a_n x_n \\ \text{subject to} & f(z) = \sum_{i=0}^n x_i f_i(z) \geq 1 \text{ for } z \in C \\ & f(z) = \sum_{i=0}^n x_i f_i(z) \geq 0 \text{ for } z \in S,\ z \notin C, \end{array} \qquad (7.17)$$

with variable $x \in \mathbf{R}^{n+1}$. This problem is always convex, since the constraints can be expressed as

$$g_1(x) = 1 - \inf_{z \in C} f(z) \leq 0, \qquad g_2(x) = -\inf_{z \in S \setminus C} f(z) \leq 0$$

(g_1 and g_2 are convex). The problem (7.17) can also be thought of as a semi-infinite linear program, *i.e.*, an optimization problem with a linear objective and an infinite number of linear inequalities, one for each $z \in S$.

In simple cases we can solve the problem (7.17) analytically. As an example, we take $S = \mathbf{R}_+$, $C = [1, \infty)$, $f_0(z) = 1$, and $f_1(z) = z$, with $\mathbf{E}\,f_1(X) = \mathbf{E}\,X = \mu \leq 1$ as our prior information. The constraint $f(z) \geq 0$ for $z \in S$ reduces to $x_0 \geq 0$, $x_1 \geq 0$. The constraint $f(z) \geq 1$ for $z \in C$, *i.e.*, $x_0 + x_1 z \geq 1$ for all $z \geq 1$, reduces to $x_0 + x_1 \geq 1$. The problem (7.17) is then

$$\begin{array}{ll} \text{minimize} & x_0 + \mu x_1 \\ \text{subject to} & x_0 \geq 0, \quad x_1 \geq 0 \\ & x_0 + x_1 \geq 1. \end{array}$$

Since $0 \leq \mu \leq 1$, the optimal point for this simple LP is $x_0 = 0$, $x_1 = 1$. This gives the classical Markov bound $\mathbf{prob}(X \geq 1) \leq \mu$.

In other cases we can solve the problem (7.17) using convex optimization.

Remark 7.1 *Duality and the Chebyshev bound problem.* The Chebyshev bound problem (7.17) determines a bound on $\mathbf{prob}(X \in C)$ for all probability measures that satisfy the given expected value constraints. Thus we can think of the Chebyshev bound problem (7.17) as producing a bound on the optimal value of the infinite-dimensional problem

$$\begin{array}{ll} \text{maximize} & \int_C \pi(dz) \\ \text{subject to} & \int_S f_i(z)\pi(dz) = a_i, \quad i = 1, \ldots, n \\ & \int_S \pi(dz) = 1 \\ & \pi \geq 0, \end{array} \tag{7.18}$$

where the variable is the measure π, and $\pi \geq 0$ means that the measure is nonnegative.

Since the Chebyshev problem (7.17) produces a bound on the problem (7.18), it should not be a surprise that they are related by duality. While semi-infinite and infinite-dimensional problems are beyond the scope of this book, we can still formally construct a dual of the problem (7.17), introducing a Lagrange multiplier *function* $p : S \to \mathbf{R}$, with $p(z)$ the Lagrange multiplier associated with the inequality $f(z) \geq 1$ (for $z \in C$) or $f(z) \geq 0$ (for $z \in S \backslash C$). Using an integral over z where we would have a sum in the finite-dimensional case, we arrive at the formal dual

$$\begin{array}{ll} \text{maximize} & \int_C p(z)\, dz \\ \text{subject to} & \int_S f_i(z)p(z)\, dz = a_i, \quad i = 1, \ldots, n \\ & \int_S p(z)\, dz = 1 \\ & p(z) \geq 0 \text{ for all } z \in S, \end{array}$$

where the optimization variable is the *function* p. This is, essentially, the same as (7.18).

Probability bounds with known first and second moments

As an example, suppose that $S = \mathbf{R}^m$, and that we are given the first and second moments of the random variable X:

$$\mathbf{E}\, X = a \in \mathbf{R}^m, \qquad \mathbf{E}\, XX^T = \Sigma \in \mathbf{S}^m.$$

In other words, we are given the expected value of the m functions z_i, $i = 1, \ldots, m$, and the $m(m+1)/2$ functions $z_i z_j$, $i, j = 1, \ldots, m$, but no other information about the distribution.

In this case we can express f as the general quadratic function

$$f(z) = z^T P z + 2q^T z + r,$$

where the variables (*i.e.*, the vector x in the discussion above) are $P \in \mathbf{S}^m$, $q \in \mathbf{R}^m$, and $r \in \mathbf{R}$. From our knowledge of the first and second moments, we find that

$$\begin{array}{rcl} \mathbf{E}\, f(X) & = & \mathbf{E}(X^T P X + 2q^T X + r) \\ & = & \mathbf{E}\, \mathbf{tr}(PXX^T) + 2\, \mathbf{E}\, q^T X + r \\ & = & \mathbf{tr}(\Sigma P) + 2q^T a + r. \end{array}$$

The constraint that $f(z) \geq 0$ for all z can be expressed as the linear matrix inequality

$$\left[\begin{array}{cc} P & q \\ q^T & r \end{array}\right] \succeq 0.$$

In particular, we have $P \succeq 0$.

Now suppose that the set C is the complement of an open polyhedron,

$$C = \mathbf{R}^m \setminus \mathcal{P}, \qquad \mathcal{P} = \{z \mid a_i^T z < b_i, \ i = 1, \ldots, k\}.$$

The condition that $f(z) \geq 1$ for all $z \in C$ is the same as requiring that

$$a_i^T z \geq b_i \implies z^T P z + 2q^T z + r \geq 1$$

for $i = 1, \ldots, k$. This, in turn, can be expressed as: there exist $\tau_1, \ldots, \tau_k \geq 0$ such that

$$\left[\begin{array}{cc} P & q \\ q^T & r-1 \end{array}\right] \succeq \tau_i \left[\begin{array}{cc} 0 & a_i/2 \\ a_i^T/2 & -b_i \end{array}\right], \quad i = 1, \ldots, k.$$

(See §B.2.)

Putting it all together, the Chebyshev bound problem (7.17) can be expressed as

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(\Sigma P) + 2q^T a + r \\ \text{subject to} & \left[\begin{array}{cc} P & q \\ q^T & r-1 \end{array}\right] \succeq \tau_i \left[\begin{array}{cc} 0 & a_i/2 \\ a_i^T/2 & -b_i \end{array}\right], \quad i = 1, \ldots, k \\ & \tau_i \geq 0, \quad i = 1, \ldots, k \\ & \left[\begin{array}{cc} P & q \\ q^T & r \end{array}\right] \succeq 0, \end{array} \tag{7.19}$$

which is a semidefinite program in the variables P, q, r, and $\tau_1, \ldots, \tau_k$. The optimal value, say α, is an upper bound on $\mathbf{prob}(X \in C)$ over all distributions with mean a and second moment Σ. Or, turning it around, $1 - \alpha$ is a lower bound on $\mathbf{prob}(X \in \mathcal{P})$.

Remark 7.2 *Duality and the Chebyshev bound problem.* The dual SDP associated with (7.19) can be expressed as

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^k \lambda_i \\ \text{subject to} & a_i^T z_i \geq b\lambda_i, \quad i = 1, \ldots, k \\ & \sum_{i=1}^k \left[\begin{array}{cc} Z_i & z_i \\ z_i^T & \lambda_i \end{array}\right] \preceq \left[\begin{array}{cc} \Sigma & a \\ a^T & 1 \end{array}\right] \\ & \left[\begin{array}{cc} Z_i & z_i \\ z_i^T & \lambda_i \end{array}\right] \succeq 0, \quad i = 1, \ldots, k. \end{array}$$

The variables are $Z_i \in \mathbf{S}^m$, $z_i \in \mathbf{R}^m$, and $\lambda_i \in \mathbf{R}$, for $i = 1, \ldots, k$. Since the SDP (7.19) is strictly feasible, strong duality holds and the dual optimum is attained.

We can give an interesting probability interpretation to the dual problem. Suppose Z_i, z_i, λ_i are dual feasible and that the first r components of λ are positive, and the rest are zero. For simplicity we also assume that $\sum_{i=1}^k \lambda_i < 1$. We define

$$\begin{aligned} x_i &= (1/\lambda_i)z_i, \quad i=1,\ldots,r, \\ w_0 &= \frac{1}{\mu}\left(a - \sum_{i=1}^r \lambda_i x_i\right), \\ W &= \frac{1}{\mu}\left(\Sigma - \sum_{i=1}^r \lambda_i x_i x_i^T\right), \end{aligned}$$

where $\mu = 1 - \sum_{i=1}^k \lambda_i$. With these definitions the dual feasibility constraints can be expressed as

$$a_i^T x_i \geq b_i, \quad i=1,\ldots,r$$

and

$$\sum_{i=1}^r \lambda_i \left[\begin{array}{cc} x_i x_i^T & x_i \\ x_i^T & 1 \end{array}\right] + \mu \left[\begin{array}{cc} W & w_0 \\ w_0^T & 1 \end{array}\right] = \left[\begin{array}{cc} \Sigma & a \\ a^T & 1 \end{array}\right].$$

Moreover, from dual feasibility,

$$\begin{aligned} \mu \left[\begin{array}{cc} W & w_0 \\ w_0^T & 1 \end{array}\right] &= \left[\begin{array}{cc} \Sigma & a \\ a^T & 1 \end{array}\right] - \sum_{i=1}^r \lambda_i \left[\begin{array}{cc} x_i x_i^T & x_i \\ x_i^T & 1 \end{array}\right] \\ &= \left[\begin{array}{cc} \Sigma & a \\ a^T & 1 \end{array}\right] - \sum_{i=1}^r \left[\begin{array}{cc} (1/\lambda_i) z_i z_i^T & z_i \\ z_i^T & \lambda_i \end{array}\right] \\ &\succeq \left[\begin{array}{cc} \Sigma & a \\ a^T & 1 \end{array}\right] - \sum_{i=1}^r \left[\begin{array}{cc} Z_i & z_i \\ z_i^T & \lambda_i \end{array}\right] \\ &\succeq 0. \end{aligned}$$

Therefore, $W \succeq w_0 w_0^T$, so it can be factored as $W - w_0 w_0^T = \sum_{i=1}^s w_i w_i^T$. Now consider a discrete random variable X with the following distribution. If $s \geq 1$, we take

$$\begin{array}{ll} X = x_i & \text{with probability } \lambda_i, \quad i=1,\ldots,r \\ X = w_0 + \sqrt{s}\, w_i & \text{with probability } \mu/(2s), \quad i=1,\ldots,s \\ X = w_0 - \sqrt{s}\, w_i & \text{with probability } \mu/(2s), \quad i=1,\ldots,s. \end{array}$$

If $s = 0$, we take

$$\begin{array}{ll} X = x_i & \text{with probability } \lambda_i, \quad i=1,\ldots,r \\ X = w_0 & \text{with probability } \mu. \end{array}$$

It is easily verified that $\mathbf{E}\, X = a$ and $\mathbf{E}\, XX^T = \Sigma$, *i.e.*, the distribution matches the given moments. Furthermore, since $x_i \in C$,

$$\mathbf{prob}(X \in C) \geq \sum_{i=1}^r \lambda_i.$$

In particular, by applying this interpretation to the dual optimal solution, we can construct a distribution that satisfies the Chebyshev bound from (7.19) with equality, which shows that the Chebyshev bound is sharp for this case.

7.4.2 Chernoff bounds

Let X be a random variable on $\mathbf{R}$. The Chernoff bound states that

$$\mathbf{prob}(X \geq u) \leq \inf_{\lambda \geq 0} \mathbf{E}\, e^{\lambda(X-u)},$$

which can be expressed as

$$\log \mathbf{prob}(X \geq u) \leq \inf_{\lambda \geq 0} \{-\lambda u + \log \mathbf{E}\, e^{\lambda X}\}. \qquad (7.20)$$

Recall (from example 3.41, page 106) that the righthand term, $\log \mathbf{E}\, e^{\lambda X}$, is called the cumulant generating function of the distribution, and is always convex, so the function to be minimized is convex. The bound (7.20) is most useful in cases when the cumulant generating function has an analytical expression, and the minimization over λ can be carried out analytically.

For example, if X is Gaussian with zero mean and unit variance, the cumulant generating function is

$$\log \mathbf{E}\, e^{\lambda X} = \lambda^2/2,$$

and the infimum over $\lambda \geq 0$ of $-\lambda u + \lambda^2/2$ occurs with $\lambda = u$ (if $u \geq 0$), so the Chernoff bound is (for $u \geq 0$)

$$\mathbf{prob}(X \geq u) \leq e^{-u^2/2}.$$

The idea behind the Chernoff bound can be extended to a more general setting, in which convex optimization is used to compute a bound on the probability of a set in $\mathbf{R}^m$. Let $C \subseteq \mathbf{R}^m$, and as in the description of Chebyshev bounds above, let 1_C denote the 0-1 indicator function of C. We will derive an upper bound on $\mathbf{prob}(X \in C)$. (In principle we can compute $\mathbf{prob}(X \in C)$, for example by Monte Carlo simulation, or numerical integration, but either of these can be a daunting computational task, and neither method produces guaranteed bounds.)

Let $\lambda \in \mathbf{R}^m$ and $\mu \in \mathbf{R}$, and consider the function $f : \mathbf{R}^m \to \mathbf{R}$ given by

$$f(z) = e^{\lambda^T z + \mu}.$$

As in the development of Chebyshev bounds, if f satisfies $f(z) \geq 1_C(z)$ for all z, then we can conclude that

$$\mathbf{prob}(X \in C) = \mathbf{E}\, 1_C(X) \leq \mathbf{E}\, f(X).$$

Clearly we have $f(z) \geq 0$ for all z; to have $f(z) \geq 1$ for $z \in C$ is the same as $\lambda^T z + \mu \geq 0$ for all $z \in C$, *i.e.*, $-\lambda^T z \leq \mu$ for all $z \in C$. Thus, if $-\lambda^T z \leq \mu$ for all $z \in C$, we have the bound

$$\mathbf{prob}(X \in C) \leq \mathbf{E} \exp(\lambda^T X + \mu),$$

or, taking logarithms,

$$\log \mathbf{prob}(X \in C) \leq \mu + \log \mathbf{E} \exp(\lambda^T X).$$

From this we obtain a general form of Chernoff's bound:

$$\begin{array}{rcl}\log \mathbf{prob}(X \in C) & \leq & \inf\{\mu + \log \mathbf{E}\exp(\lambda^T X) \mid -\lambda^T z \leq \mu \text{ for all } z \in C\} \\ & = & \inf_{\lambda} \left(\sup_{z\in C}(-\lambda^T z) + \log \mathbf{E}\exp(\lambda^T X) \right) \\ & = & \inf \left(S_C(-\lambda) + \log \mathbf{E}\exp(\lambda^T X) \right),\end{array}$$

where S_C is the support function of C. Note that the second term, $\log \mathbf{E}\exp(\lambda^T X)$, is the cumulant generating function of the distribution, and is always convex (see example 3.41, page 106). Evaluating this bound is, in general, a convex optimization problem.

Chernoff bound for a Gaussian variable on a polyhedron

As a specific example, suppose that X is a Gaussian random vector on $\mathbf{R}^m$ with zero mean and covariance I, so its cumulant generating function is

$$\log \mathbf{E}\exp(\lambda^T X) = \lambda^T \lambda/2.$$

We take C to be a polyhedron described by inequalities:

$$C = \{x \mid Ax \preceq b\},$$

which we assume is nonempty.

For use in the Chernoff bound, we use a dual characterization of the support function S_C:

$$\begin{array}{rcl}S_C(y) & = & \sup\{y^T x \mid Ax \preceq b\} \\ & = & -\inf\{-y^T x \mid Ax \preceq b\} \\ & = & -\sup\{-b^T u \mid A^T u = y,\ u \succeq 0\} \\ & = & \inf\{b^T u \mid A^T u = y,\ u \succeq 0\}\end{array}$$

where in the third line we use LP duality:

$$\inf\{c^T x \mid Ax \preceq b\} = \sup\{-b^T u \mid A^T u + c = 0,\ u \succeq 0\}$$

with $c = -y$. Using this expression for S_C in the Chernoff bound we obtain

$$\begin{array}{rcl}\log \mathbf{prob}(X \in C) & \leq & \inf_{\lambda} \left(S_C(-\lambda) + \log \mathbf{E}\exp(\lambda^T X) \right) \\ & = & \inf_{\lambda} \inf_{u} \{b^T u + \lambda^T \lambda/2 \mid u \succeq 0,\ A^T u + \lambda = 0\}.\end{array}$$

Thus, the Chernoff bound on $\mathbf{prob}(X \in C)$ is the exponential of the optimal value of the QP

$$\begin{array}{ll}\text{minimize} & b^T u + \lambda^T \lambda/2 \\ \text{subject to} & u \succeq 0, \quad A^T u + \lambda = 0,\end{array} \qquad (7.21)$$

where the variables are u and λ.

This problem has an interesting geometric interpretation. It is equivalent to

$$\begin{array}{ll} \text{minimize} & b^T u + (1/2)\|A^T u\|_2^2 \\ \text{subject to} & u \succeq 0, \end{array}$$

which is the dual of

$$\begin{array}{ll} \text{maximize} & -(1/2)\|x\|_2^2 \\ \text{subject to} & Ax \preceq b. \end{array}$$

In other words, the Chernoff bound is

$$\mathbf{prob}(X \in C) \leq \exp(-\,\mathbf{dist}(0, C)^2/2), \tag{7.22}$$

where $\mathbf{dist}(0, C)$ is the Euclidean distance of the origin to C.

Remark 7.3 The bound (7.22) can also be derived without using Chernoff's inequality. If the distance between 0 and C is d, then there is a halfspace $\mathcal{H} = \{z \mid a^T z \geq d\}$, with $\|a\|_2 = 1$, that contains C. The random variable $a^T X$ is $\mathcal{N}(0, 1)$, so

$$\mathbf{prob}(X \in C) \leq \mathbf{prob}(X \in \mathcal{H}) = \Phi(-d),$$

where Φ is the cumulative distribution function of a zero mean, unit variance Gaussian. Since $\Phi(-d) \leq e^{-d^2/2}$ for $d \geq 0$, this bound is at least as sharp as the Chernoff bound (7.22).

7.4.3 Example

In this section we illustrate the Chebyshev and Chernoff probability bounding methods with a detection example. We have a set of m possible symbols or signals $s \in \{s_1, s_2, \ldots, s_m\} \subseteq \mathbf{R}^n$, which is called the *signal constellation*. One of these signals is transmitted over a noisy channel. The received signal is $x = s + v$, where v is a noise, modeled as a random variable. We assume that $\mathbf{E}\, v = 0$ and $\mathbf{E}\, vv^T = \sigma^2 I$, *i.e.*, the noise components $v_1, \ldots, v_n$ are zero mean, uncorrelated, and have variance σ^2. The receiver must estimate which signal was sent on the basis of the received signal $x = s + v$. The *minimum distance detector* chooses as estimate the symbol s_k closest (in Euclidean norm) to x. (If the noise v is Gaussian, then minimum distance decoding is the same as maximum likelihood decoding.)

If the signal s_k is transmitted, correct detection occurs if s_k is the estimate, given x. This occurs when the signal s_k is closer to x than the other signals, *i.e.*,

$$\|x - s_k\|_2 < \|x - s_j\|_2, \qquad j \neq k.$$

Thus, correct detection of symbol s_k occurs if the random variable v satisfies the linear inequalities

$$2(s_j - s_k)^T(s_k + v) < \|s_j\|_2^2 - \|s_k\|_2^2, \quad j \neq k.$$

These inequalities define the *Voronoi region* V_k of s_k in the signal constellation, *i.e.*, the set of points closer to s_k than any other signal in the constellation. The probability of correct detection of s_k is $\mathbf{prob}(s_k + v \in V_k)$.

Figure 7.5 shows a simple example with $m = 7$ signals, with dimension $n = 2$.

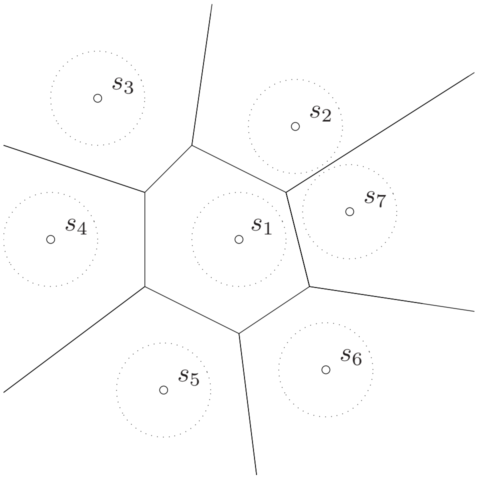

Figure 7.5 A constellation of 7 signals $s_1, \ldots, s_7 \in \mathbf{R}^2$, shown as small circles. The line segments show the boundaries of the corresponding Voronoi regions. The minimum distance detector selects symbol s_k when the received signal lies closer to s_k than to any of the other points, *i.e.*, if the received signal is in the interior of the Voronoi region around symbol s_k. The circles around each point have radius one, to show the scale.

Chebyshev bounds

The SDP bound (7.19) provides a lower bound on the probability of correct detection, and is plotted in figure 7.6, as a function of the noise standard deviation σ, for the three symbols s_1, s_2, and s_3. These bounds hold for *any* noise distribution with zero mean and covariance $\sigma^2 I$. They are tight in the sense that there exists a noise distribution with zero mean and covariance $\Sigma = \sigma^2 I$, for which the probability of error is equal to the lower bound. This is illustrated in figure 7.7, for the first Voronoi set, and $\sigma = 1$.

Chernoff bounds

We use the same example to illustrate the Chernoff bound. Here we assume that the noise is Gaussian, *i.e.*, $v \sim \mathcal{N}(0, \sigma^2 I)$. If symbol s_k is transmitted, the probability of correct detection is the probability that $s_k + v \in V_k$. To find a lower bound for this probability, we use the QP (7.21) to compute upper bounds on the probability that the ML detector selects symbol i, $i = 1, \ldots, m$, $i \neq k$. (Each of these upper bounds is related to the distance of s_k to the Voronoi set V_i.) Adding these upper bounds on the probabilities of mistaking s_k for s_i, we obtain an upper bound on the probability of error, and therefore, a lower bound on the probability of correct detection of symbol s_k. The resulting lower bound, for s_1, is shown in figure 7.8, along with an estimate of the probability of correct detection obtained using Monte Carlo analysis.

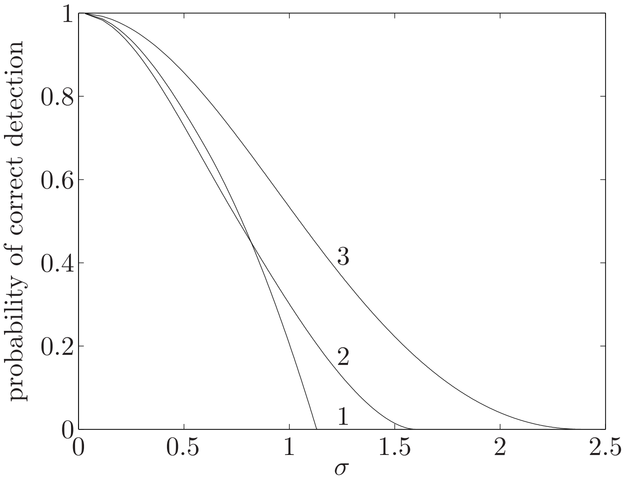

Figure 7.6 Chebyshev lower bounds on the probability of correct detection for symbols s_1, s_2, and s_3. These bounds are valid for any noise distribution that has zero mean and covariance $\sigma^2 I$.

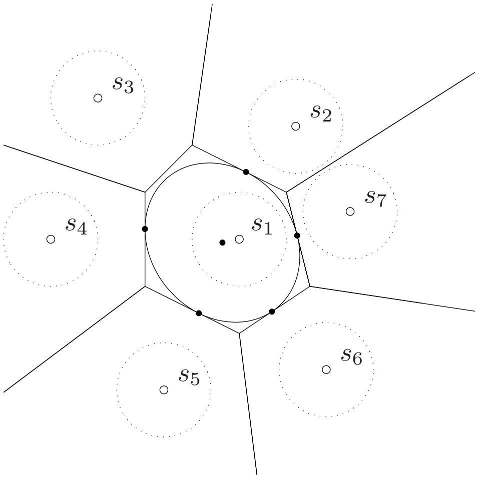

Figure 7.7 The Chebyshev lower bound on the probability of correct detection of symbol 1 is equal to 0.2048 when $\sigma = 1$. This bound is achieved by the discrete distribution illustrated in the figure. The solid circles are the possible values of the received signal $s_1 + v$. The point in the center of the ellipse has probability 0.2048. The five points on the boundary have a total probability 0.7952. The ellipse is defined by $x^T P x + 2q^T x + r = 1$, where P, q, and r are the optimal solution of the SDP (7.19).

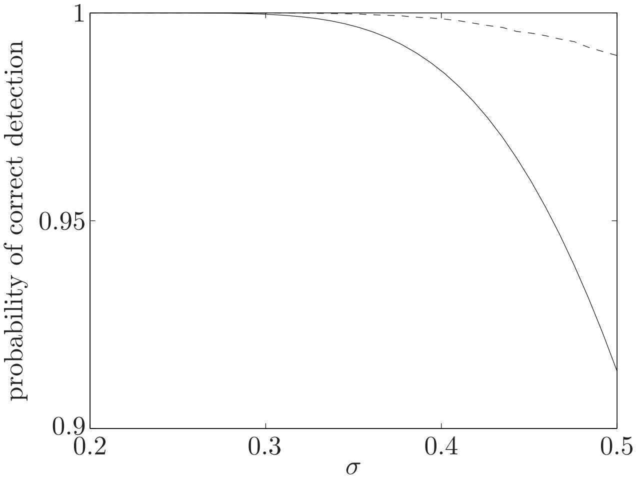

Figure 7.8 The Chernoff lower bound (solid line) and a Monte Carlo estimate (dashed line) of the probability of correct detection of symbol s_1, as a function of σ. In this example the noise is Gaussian with zero mean and covariance $\sigma^2 I$.

7.5 Experiment design

We consider the problem of estimating a vector $x \in \mathbf{R}^n$ from measurements or experiments

$$y_i = a_i^T x + w_i, \quad i = 1, \ldots, m,$$

where w_i is measurement noise. We assume that w_i are independent Gaussian random variables with zero mean and unit variance, and that the measurement vectors $a_1, \ldots, a_m$ span $\mathbf{R}^n$. The maximum likelihood estimate of x, which is the same as the minimum variance estimate, is given by the least-squares solution

$$\hat{x} = \left(\sum_{i=1}^m a_i a_i^T \right)^{-1} \sum_{i=1}^m y_i a_i.$$

The associated estimation error $e = \hat{x} - x$ has zero mean and covariance matrix

$$E = \mathbf{E}\, ee^T = \left(\sum_{i=1}^m a_i a_i^T \right)^{-1}.$$

The matrix E characterizes the accuracy of the estimation, or the informativeness of the experiments. For example the α-confidence level ellipsoid for x is given by

$$\mathcal{E} = \{ z \mid (z - \hat{x})^T E^{-1} (z - \hat{x}) \leq \beta \},$$

where β is a constant that depends on n and α.

We suppose that the vectors $a_1, \ldots, a_m$, which characterize the measurements, can be chosen among p possible test vectors $v_1, \ldots, v_p \in \mathbf{R}^n$, *i.e.*, each a_i is one of the v_j. The goal of *experiment design* is to choose the vectors a_i, from among the possible choices, so that the error covariance E is small (in some sense). In other words, each of m experiments or measurements can be chosen from a fixed menu of p possible experiments; our job is to find a set of measurements that (together) are maximally informative.

Let m_j denote the number of experiments for which a_i is chosen to have the value v_j, so we have

$$m_1 + \cdots + m_p = m.$$

We can express the error covariance matrix as

$$E = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1} = \left(\sum_{j=1}^{p} m_j v_j v_j^T \right)^{-1}.$$

This shows that the error covariance depends only on the numbers of each type of experiment chosen (*i.e.*, $m_1, \ldots, m_p$).

The basic experiment design problem is as follows. Given the menu of possible choices for experiments, *i.e.*, $v_1, \ldots, v_p$, and the total number m of experiments to be carried out, choose the numbers of each type of experiment, *i.e.*, $m_1, \ldots, m_p$, to make the error covariance E small (in some sense). The variables $m_1, \ldots, m_p$ must, of course, be integers and sum to m, the given total number of experiments. This leads to the optimization problem

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = \left(\sum_{j=1}^p m_j v_j v_j^T \right)^{-1} \\ \text{subject to} & m_i \geq 0, \quad m_1 + \cdots + m_p = m \\ & m_i \in \mathbf{Z}, \end{array} \tag{7.23}$$

where the variables are the integers $m_1, \ldots, m_p$.

The basic experiment design problem (7.23) is a vector optimization problem over the positive semidefinite cone. If one experiment design results in E, and another in $\tilde{E}$, with $E \preceq \tilde{E}$, then certainly the first experiment design is as good as or better than the second. For example, the confidence ellipsoid for the first experiment design (translated to the origin for comparison) is contained in the confidence ellipsoid of the second. We can also say that the first experiment design allows us to estimate $q^T x$ better (*i.e.*, with lower variance) than the second experiment design, for any vector q, since the variance of our estimate of $q^T x$ is given by $q^T E q$ for the first experiment design and $q^T \tilde{E} q$ for the second. We will see below several common scalarizations for the problem.

7.5.1 The relaxed experiment design problem

The basic experiment design problem (7.23) can be a hard combinatorial problem when m, the total number of experiments, is comparable to p, since in this case the m_i are all small integers. In the case when m is large compared to p, however, a good approximate solution of (7.23) can be found by ignoring, or relaxing, the constraint that the m_i are integers. Let $\lambda_i = m_i/m$, which is the fraction of the total number of experiments for which $a_j = v_i$, or the relative frequency of experiment i. We can express the error covariance in terms of λ_i as

$$E = \frac{1}{m}\left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)^{-1}. \tag{7.24}$$

The vector $\lambda \in \mathbf{R}^p$ satisfies $\lambda \succeq 0$, $\mathbf{1}^T\lambda = 1$, and also, each λ_i is an integer multiple of $1/m$. By ignoring this last constraint, we arrive at the problem

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = (1/m)\left(\sum_{i=1}^p \lambda_i v_i v_i^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1, \end{array} \tag{7.25}$$

with variable $\lambda \in \mathbf{R}^p$. To distinguish this from the original combinatorial experiment design problem (7.23), we refer to it as the *relaxed experiment design problem.* The relaxed experiment design problem (7.25) is a convex optimization problem, since the objective E is an $\mathbf{S}_+^n$-convex function of λ.

Several statements can be made about the relation between the (combinatorial) experiment design problem (7.23) and the relaxed problem (7.25). Clearly the optimal value of the relaxed problem provides a lower bound on the optimal value of the combinatorial one, since the combinatorial problem has an additional constraint. From a solution of the relaxed problem (7.25) we can construct a suboptimal solution of the combinatorial problem (7.23) as follows. First, we apply simple rounding to get

$$m_i = \mathbf{round}(m\lambda_i), \quad i = 1, \ldots, p.$$

Corresponding to this choice of $m_1, \ldots, m_p$ is the vector $\tilde{\lambda}$,

$$\tilde{\lambda}_i = (1/m)\mathbf{round}(m\lambda_i), \quad i = 1, \ldots, p.$$

The vector $\tilde{\lambda}$ satisfies the constraint that each entry is an integer multiple of $1/m$. Clearly we have $|\lambda_i - \tilde{\lambda}_i| \le 1/(2m)$, so for m large, we have $\lambda \approx \tilde{\lambda}$. This implies that the constraint $\mathbf{1}^T\tilde{\lambda} = 1$ is nearly satisfied, for large m, and also that the error covariance matrices associated with $\tilde{\lambda}$ and λ are close.

We can also give an alternative interpretation of the relaxed experiment design problem (7.25). We can interpret the vector $\lambda \in \mathbf{R}^p$ as defining a probability distribution on the experiments $v_1, \ldots, v_p$. Our choice of λ corresponds to a *random experiment*: each experiment a_i takes the form v_j with probability λ_j.

In the rest of this section, we consider only the relaxed experiment design problem, so we drop the qualifier 'relaxed' in our discussion.

7.5.2 Scalarizations

Several scalarizations have been proposed for the experiment design problem (7.25), which is a vector optimization problem over the positive semidefinite cone.

D-optimal design

The most widely used scalarization is called *D-optimal design*, in which we minimize the determinant of the error covariance matrix E. This corresponds to designing the experiment to minimize the volume of the resulting confidence ellipsoid (for a fixed confidence level). Ignoring the constant factor $1/m$ in E, and taking the logarithm of the objective, we can pose this problem as

$$\begin{array}{ll} \text{minimize} & \log\det\left(\sum_{i=1}^{p}\lambda_i v_i v_i^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1, \end{array} \tag{7.26}$$

which is a convex optimization problem.

E-optimal design

In *E-optimal design*, we minimize the norm of the error covariance matrix, *i.e.*, the maximum eigenvalue of E. Since the diameter (twice the longest semi-axis) of the confidence ellipsoid $\mathcal{E}$ is proportional to $\|E\|_2^{1/2}$, minimizing $\|E\|_2$ can be interpreted geometrically as minimizing the diameter of the confidence ellipsoid. E-optimal design can also be interpreted as minimizing the maximum variance of q^Te, over all q with $\|q\|_2 = 1$.

The E-optimal experiment design problem is

$$\begin{array}{ll} \text{minimize} & \left\|\left(\sum_{i=1}^{p}\lambda_i v_i v_i^T\right)^{-1}\right\|_2 \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1. \end{array}$$

The objective is a convex function of λ, so this is a convex problem.

The E-optimal experiment design problem can be cast as an SDP

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & \sum_{i=1}^{p}\lambda_i v_i v_i^T \succeq tI \\ & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1, \end{array} \tag{7.27}$$

with variables $\lambda \in \mathbf{R}^p$ and $t \in \mathbf{R}$.

A-optimal design

In *A-optimal experiment design*, we minimize $\mathbf{tr}\, E$, the trace of the covariance matrix. This objective is simply the mean of the norm of the error squared:

$$\mathbf{E}\,\|e\|_2^2 = \mathbf{E}\,\mathbf{tr}(ee^T) = \mathbf{tr}\, E.$$

The A-optimal experiment design problem is

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\left(\sum_{i=1}^{p}\lambda_i v_i v_i^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1. \end{array} \tag{7.28}$$

This, too, is a convex problem. Like the E-optimal experiment design problem, it can be cast as an SDP:

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T u \\ \text{subject to} & \begin{bmatrix} \sum_{i=1}^{p}\lambda_i v_i v_i^T & e_k \\ e_k^T & u_k \end{bmatrix} \succeq 0, \quad k = 1, \ldots, n \\ & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1, \end{array}$$

where the variables are $u \in \mathbf{R}^n$ and $\lambda \in \mathbf{R}^p$, and here, e_k is the kth unit vector.

Optimal experiment design and duality

The Lagrange duals of the three scalarizations have an interesting geometric meaning.

The dual of the D-optimal experiment design problem (7.26) can be expressed as

$$\begin{array}{ll} \text{maximize} & \log\det W + n \log n \\ \text{subject to} & v_i^T W v_i \leq 1, \quad i = 1, \ldots, p, \end{array}$$

with variable $W \in \mathbf{S}^n$ and domain $\mathbf{S}^n_{++}$ (see exercise 5.10). This dual problem has a simple interpretation: The optimal solution $W^\star$ determines the minimum volume ellipsoid, centered at the origin, given by $\{x \mid x^T W^\star x \leq 1\}$, that contains the points $v_1, \ldots, v_p$. (See also the discussion of problem (5.14) on page 222.) By complementary slackness,

$$\lambda_i^\star (1 - v_i^T W^\star v_i) = 0, \quad i = 1, \ldots, p, \tag{7.29}$$

i.e., the optimal experiment design only uses the experiments v_i which lie on the surface of the minimum volume ellipsoid.

The duals of the E-optimal and A-optimal design problems can be given a similar interpretation. The duals of problems (7.27) and (7.28) can be expressed as

$$\begin{array}{ll} \text{maximize} & \mathbf{tr}\, W \\ \text{subject to} & v_i^T W v_i \leq 1, \quad i = 1, \ldots, p \\ & W \succeq 0, \end{array} \tag{7.30}$$

and

$$\begin{array}{ll} \text{maximize} & (\mathbf{tr}\, W^{1/2})^2 \\ \text{subject to} & v_i^T W v_i \leq 1, \quad i = 1, \ldots, p, \end{array} \tag{7.31}$$

respectively. The variable in both problems is $W \in \mathbf{S}^n$. In the second problem there is an implicit constraint $W \in \mathbf{S}^n_+$. (See exercises 5.40 and 5.10.)

As for the D-optimal design, the optimal solution $W^\star$ determines a minimal ellipsoid $\{x \mid x^T W^\star x \leq 1\}$ that contains the points $v_1, \ldots, v_p$. Moreover $W^\star$ and $\lambda^\star$ satisfy the complementary slackness conditions (7.29), *i.e.*, the optimal design only uses experiments v_i that lie on the surface of the ellipsoid defined by $W^\star$.

Experiment design example

We consider a problem with $x \in \mathbf{R}^2$, and $p = 20$. The 20 candidate measurement vectors a_i are shown as circles in figure 7.9. The origin is indicated with a cross. The D-optimal experiment has only two nonzero λ_i, indicated as solid circles in figure 7.9. The E-optimal experiment has two nonzero λ_i, indicated as solid circles in figure 7.10. The A-optimal experiment has three nonzero λ_i, indicated as solid circles in figure 7.11. We also show the three ellipsoids $\{x \mid x^T W^\star x \leq 1\}$ associated with the dual optimal solutions $W^\star$. The resulting 90% confidence ellipsoids are shown in figure 7.12, along with the confidence ellipsoid for the 'uniform' design, with equal weight $\lambda_i = 1/p$ on all experiments.

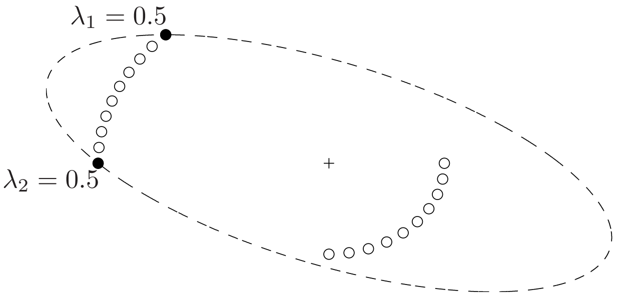

Figure 7.9 Experiment design example. The 20 candidate measurement vectors are indicated with circles. The D-optimal design uses the two measurement vectors indicated with solid circles, and puts an equal weight $\lambda_i = 0.5$ on each of them. The ellipsoid is the minimum volume ellipsoid centered at the origin, that contains the points v_i.

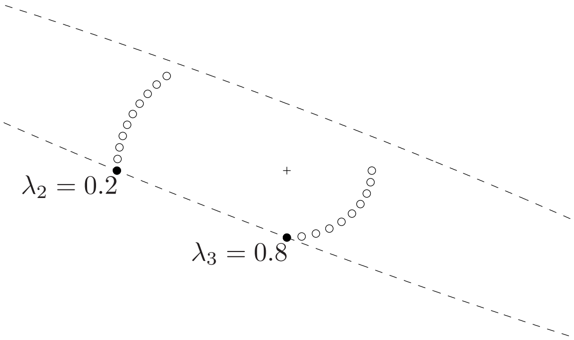

Figure 7.10 The E-optimal design uses two measurement vectors. The dashed lines are (part of) the boundary of the ellipsoid $\{x \mid x^T W^\star x \leq 1\}$ where $W^\star$ is the solution of the dual problem (7.30).

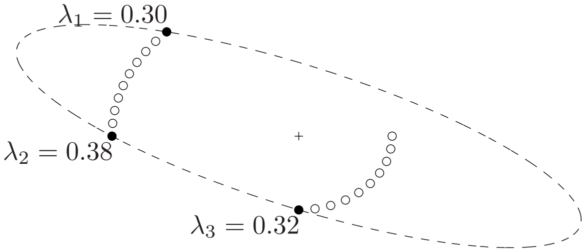

Figure 7.11 The A-optimal design uses three measurement vectors. The dashed line shows the ellipsoid $\{x \mid x^T W^\star x \leq 1\}$ associated with the solution of the dual problem (7.31).

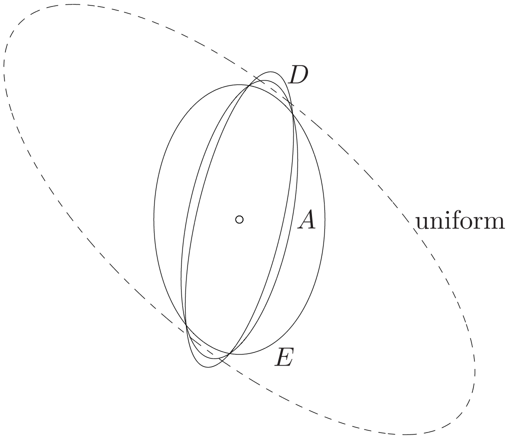

Figure 7.12 Shape of the 90% confidence ellipsoids for D-optimal, A-optimal, E-optimal, and uniform designs.

7.5.3 Extensions

Resource limits

Suppose that associated with each experiment is a cost c_i, which could represent the economic cost, or time required, to carry out an experiment with v_i. The total cost, or time required (if the experiments are carried out sequentially) is then

$$m_1 c_1 + \cdots + m_p c_p = mc^T \lambda.$$

We can add a limit on total cost by adding the linear inequality $mc^T \lambda \leq B$, where B is a budget, to the basic experiment design problem. We can add multiple linear inequalities, representing limits on multiple resources.

Multiple measurements per experiment

We can also consider a generalization in which each experiment yields multiple measurements. In other words, when we carry out an experiment using one of the possible choices, we obtain several measurements. To model this situation we can use the same notation as before, with v_i as matrices in $\mathbf{R}^{n \times k_i}$:

$$v_i = \left[\begin{array}{ccc} u_{i1} & \cdots & u_{ik_i} \end{array} \right],$$

where k_i is the number of (scalar) measurements obtained when the experiment v_i is carried out. The error covariance matrix, in this more complicated setup, has the exact same form.

In conjunction with additional linear inequalities representing limits on cost or time, we can model discounts or time savings associated with performing groups of measurements simultaneously. Suppose, for example, that the cost of simultaneously making (scalar) measurements v_1 and v_2 is less than the sum of the costs of making them separately. We can take v_3 to be the matrix

$$v_3 = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$$

and assign costs c_1, c_2, and c_3 associated with making the first measurement alone, the second measurement alone, and the two simultaneously, respectively.

When we solve the experiment design problem, λ_1 will give us the fraction of times we should carry out the first experiment alone, λ_2 will give us the fraction of times we should carry out the second experiment alone, and λ_3 will give us the fraction of times we should carry out the two experiments simultaneously. (Normally we would expect a choice to be made here; we would not expect to have $\lambda_1 > 0$, $\lambda_2 > 0$, and $\lambda_3 > 0$.)

Bibliography

ML and MAP estimation, hypothesis testing, and detection are covered in books on statistics, pattern recognition, statistical signal processing, or communications; see, for example, Bickel and Doksum [BD77], Duda, Hart, and Stork [DHS99], Scharf [Sch91], or Proakis [Pro01].

Logistic regression is discussed in Hastie, Tibshirani, and Friedman [HTF01, §4.4]. For the covariance estimation problem of page 355, see Anderson [And70].

Generalizations of Chebyshev's inequality were studied extensively in the sixties, by Isii [Isi64], Marshall and Olkin [MO60], Karlin and Studden [KS66, chapter 12], and others. The connection with semidefinite programming was made more recently by Bertsimas and Sethuraman [BS00] and Lasserre [Las02].

The terminology in §7.5 (A-, D-, and E-optimality) is standard in the literature on optimal experiment design (see, for example, Pukelsheim [Puk93]). The geometric interpretation of the dual D-optimal design problem is discussed by Titterington [Tit75].

Exercises

Estimation

7.1 *Linear measurements with exponentially distributed noise.* Show how to solve the ML estimation problem (7.2) when the noise is exponentially distributed, with density

$$p(z) = \begin{cases} (1/a)e^{-z/a} & z \geq 0 \\ 0 & z < 0, \end{cases}$$

where $a > 0$.

7.2 *ML estimation and ℓ_∞-norm approximation.* We consider the linear measurement model $y = Ax + v$ of page 352, with a uniform noise distribution of the form

$$p(z) = \begin{cases} 1/(2\alpha) & |z| \leq \alpha \\ 0 & |z| > \alpha. \end{cases}$$

As mentioned in example 7.1, page 352, any x that satisfies $\|Ax - y\|_\infty \leq \alpha$ is a ML estimate.

Now assume that the parameter α is not known, and we wish to estimate α, along with the parameters x. Show that the ML estimates of x and α are found by solving the ℓ_∞-norm approximation problem

$$\text{minimize} \quad \|Ax - y\|_\infty,$$

where a_i^T are the rows of A.

7.3 *Probit model.* Suppose $y \in \{0, 1\}$ is random variable given by

$$y = \begin{cases} 1 & a^T u + b + v \leq 0 \\ 0 & a^T u + b + v > 0, \end{cases}$$

where the vector $u \in \mathbf{R}^n$ is a vector of explanatory variables (as in the logistic model described on page 354), and v is a zero mean unit variance Gaussian variable.

Formulate the ML estimation problem of estimating a and b, given data consisting of pairs (u_i, y_i), $i = 1, \ldots, N$, as a convex optimization problem.

7.4 *Estimation of covariance and mean of a multivariate normal distribution.* We consider the problem of estimating the covariance matrix R and the mean a of a Gaussian probability density function

$$p_{R,a}(y) = (2\pi)^{-n/2} \det(R)^{-1/2} \exp(-(y - a)^T R^{-1} (y - a)/2),$$

based on N independent samples $y_1, y_2, \ldots, y_N \in \mathbf{R}^n$.

(a) We first consider the estimation problem when there are no additional constraints on R and a. Let μ and Y be the sample mean and covariance, defined as

$$\mu = \frac{1}{N} \sum_{k=1}^N y_k, \qquad Y = \frac{1}{N} \sum_{k=1}^N (y_k - \mu)(y_k - \mu)^T.$$

Show that the log-likelihood function

$$l(R, a) = -(Nn/2)\log(2\pi) - (N/2)\log\det R - (1/2)\sum_{k=1}^N (y_k - a)^T R^{-1} (y_k - a)$$

can be expressed as

$$l(R,a) = \frac{N}{2}\left(-n\log(2\pi) - \log\det R - \mathbf{tr}(R^{-1}Y) - (a-\mu)^T R^{-1}(a-\mu)\right).$$

Use this expression to show that if $Y \succ 0$, the ML estimates of R and a are unique, and given by

$$a_{\rm ml} = \mu, \qquad R_{\rm ml} = Y.$$

(b) The log-likelihood function includes a convex term ($-\log\det R$), so it is not obviously concave. Show that l is concave, jointly in R and a, in the region defined by

$$R \preceq 2Y.$$

This means we can use convex optimization to compute simultaneous ML estimates of R and a, subject to convex constraints, as long as the constraints include $R \preceq 2Y$, *i.e.*, the estimate R must not exceed twice the unconstrained ML estimate.

7.5 *Markov chain estimation.* Consider a Markov chain with n states, and transition probability matrix $P \in \mathbf{R}^{n\times n}$ defined as

$$P_{ij} = \mathbf{prob}(y(t+1) = i \mid y(t) = j).$$

The transition probabilities must satisfy $P_{ij} \geq 0$ and $\sum_{i=1}^n P_{ij} = 1$, $j = 1,\ldots,n$. We consider the problem of estimating the transition probabilities, given an observed sample sequence $y(1) = k_1$, $y(2) = k_2$, ..., $y(N) = k_n$.

(a) Show that if there are no other prior constraints on P_{ij}, then the ML estimates are the empirical transition frequencies: $\hat{P}_{ij}$ is the ratio of the number of times the state transitioned from j into i, divided by the number of times it was j, in the observed sample.

(b) Suppose that an equilibrium distribution p of the Markov chain is known, *i.e.*, a vector $q \in \mathbf{R}^n_+$ satisfying $\mathbf{1}^T q = 1$ and $Pq = q$. Show that the problem of computing the ML estimate of P, given the observed sequence and knowledge of q, can be expressed as a convex optimization problem.

7.6 *Estimation of mean and variance.* Consider a random variable $x \in \mathbf{R}$ with density p, which is normalized, *i.e.*, has zero mean and unit variance. Consider a random variable $y = (x+b)/a$ obtained by an affine transformation of x, where $a > 0$. The random variable y has mean b/a and variance $1/a^2$. As a and b vary over $\mathbf{R}_+$ and $\mathbf{R}$, respectively, we generate a family of densities obtained from p by scaling and shifting, uniquely parametrized by mean and variance.

Show that if p is log-concave, then finding the ML estimate of a and b, given samples $y_1,\ldots,y_n$ of y, is a convex problem.

As an example, work out an analytical solution for the ML estimates of a and b, assuming p is a normalized Laplacian density, $p(x) = e^{-2|x|}$.

7.7 *ML estimation of Poisson distributions.* Suppose x_i, $i = 1,\ldots,n$, are independent random variables with Poisson distributions

$$\mathbf{prob}(x_i = k) = \frac{e^{-\mu_i}\mu_i^k}{k!},$$

with unknown means μ_i. The variables x_i represent the number of times that one of n possible independent events occurs during a certain period. In emission tomography, for example, they might represent the number of photons emitted by n sources.

We consider an experiment designed to determine the means μ_i. The experiment involves m detectors. If event i occurs, it is detected by detector j with probability p_{ji}. We assume the probabilities p_{ji} are given (with $p_{ji} \geq 0$, $\sum_{j=1}^{m} p_{ji} \leq 1$). The total number of events recorded by detector j is denoted y_j,

$$y_j = \sum_{i=1}^{n} y_{ji}, \quad j = 1, \ldots, m.$$

Formulate the ML estimation problem of estimating the means μ_i, based on observed values of y_j, $j = 1, \ldots, m$, as a convex optimization problem.

Hint. The variables y_{ji} have Poisson distributions with means $p_{ji}\mu_i$, *i.e.*,

$$\mathbf{prob}(y_{ji} = k) = \frac{e^{-p_{ji}\mu_i}(p_{ji}\mu_i)^k}{k!}.$$

The sum of n independent Poisson variables with means $\lambda_1, \ldots, \lambda_n$ has a Poisson distribution with mean $\lambda_1 + \cdots + \lambda_n$.

7.8 *Estimation using sign measurements.* We consider the measurement setup

$$y_i = \mathbf{sign}(a_i^T x + b_i + v_i), \quad i = 1, \ldots, m,$$

where $x \in \mathbf{R}^n$ is the vector to be estimated, and $y_i \in \{-1, 1\}$ are the measurements. The vectors $a_i \in \mathbf{R}^n$ and scalars $b_i \in \mathbf{R}$ are known, and v_i are IID noises with a log-concave probability density. (You can assume that $a_i^T x + b_i + v_i = 0$ does not occur.) Show that maximum likelihood estimation of x is a convex optimization problem.

7.9 *Estimation with unknown sensor nonlinearity.* We consider the measurement setup

$$y_i = f(a_i^T x + b_i + v_i), \quad i = 1, \ldots, m,$$

where $x \in \mathbf{R}^n$ is the vector to be estimated, $y_i \in \mathbf{R}$ are the measurements, $a_i \in \mathbf{R}^n$, $b_i \in \mathbf{R}$ are known, and v_i are IID noises with log-concave probability density. The function $f : \mathbf{R} \to \mathbf{R}$, which represents a measurement nonlinearity, is *not* known. However, it is known that $f'(t) \in [l, u]$ for all t, where $0 < l < u$ are given.

Explain how to use convex optimization to find a maximum likelihood estimate of x, as well as the function f. (This is an infinite-dimensional ML estimation problem, but you can be informal in your approach and explanation.)

7.10 *Nonparametric distributions on* $\mathbf{R}^k$. We consider a random variable $x \in \mathbf{R}^k$ with values in a finite set $\{\alpha_1, \ldots, \alpha_n\}$, and with distribution

$$p_i = \mathbf{prob}(x = \alpha_i), \quad i = 1, \ldots, n.$$

Show that a lower bound on the covariance of X,

$$S \preceq \mathbf{E}(X - \mathbf{E}\,X)(X - \mathbf{E}\,X)^T,$$

is a convex constraint in p.

Optimal detector design

7.11 *Randomized detectors.* Show that every randomized detector can be expressed as a convex combination of a set of deterministic detectors: If

$$T = \left[\begin{array}{cccc} t_1 & t_2 & \cdots & t_n \end{array}\right] \in \mathbf{R}^{m \times n}$$

satisfies $t_k \succeq 0$ and $\mathbf{1}^T t_k = 1$, then T can be expressed as

$$T = \theta_1 T_1 + \cdots + \theta_N T_N,$$

where T_i is a zero-one matrix with exactly one element equal to one per column, and $\theta_i \geq 0$, $\sum_{i=1}^N \theta_i = 1$. What is the maximum number of deterministic detectors N we may need?

We can interpret this convex decomposition as follows. The randomized detector can be realized as a bank of N deterministic detectors. When we observe $X = k$, the estimator chooses a random index from the set $\{1, \ldots, N\}$, with probability $\mathbf{prob}(j = i) = \theta_i$, and then uses deterministic detector T_j.

7.12 *Optimal action.* In detector design, we are given a matrix $P \in \mathbf{R}^{n \times m}$ (whose columns are probability distributions), and then design a matrix $T \in \mathbf{R}^{m \times n}$ (whose columns are probability distributions), so that $D = TP$ has large diagonal elements (and small off-diagonal elements). In this problem we study the dual problem: Given P, find a matrix $S \in \mathbf{R}^{m \times n}$ (whose columns are probability distributions), so that $\tilde{D} = PS \in \mathbf{R}^{n \times n}$ has large diagonal elements (and small off-diagonal elements). To make the problem specific, we take the objective to be maximizing the minimum element of $\tilde{D}$ on the diagonal.

We can interpret this problem as follows. There are n *outcomes*, which depend (stochastically) on which of m inputs or *actions* we take: P_{ij} is the probability that outcome i occurs, given action j. Our goal is find a (randomized) strategy that, to the extent possible, causes any specified outcome to occur. The strategy is given by the matrix S: S_{ji} is the probability that we take action j, when we want outcome i to occur. The matrix $\tilde{D}$ gives the action error probability matrix: $\tilde{D}_{ij}$ is the probability that outcome i occurs, when we want outcome j to occur. In particular, $\tilde{D}_{ii}$ is the probability that outcome i occurs, when we want it to occur.

Show that this problem has a simple analytical solution. Show that (unlike the corresponding detector problem) there is always an optimal solution that is deterministic.

Hint. Show that the problem is separable in the columns of S.

Chebyshev and Chernoff bounds

7.13 *Chebyshev-type inequalities on a finite set.* Assume X is a random variable taking values in the set $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$, and let S be a subset of $\{\alpha_1, \ldots, \alpha_m\}$. The distribution of X is unknown, but we are given the expected values of n functions f_i:

$$\mathbf{E}\, f_i(X) = b_i, \quad i = 1, \ldots, n. \tag{7.32}$$

Show that the optimal value of the LP

$$\begin{array}{ll} \text{minimize} & x_0 + \sum_{i=1}^n b_i x_i \\ \text{subject to} & x_0 + \sum_{i=1}^n f_i(\alpha) x_i \geq 1, \quad \alpha \in S \\ & x_0 + \sum_{i=1}^n f_i(\alpha) x_i \geq 0, \quad \alpha \notin S, \end{array}$$

with variables $x_0, \ldots, x_n$, is an upper bound on $\mathbf{prob}(X \in S)$, valid for all distributions that satisfy (7.32). Show that there always exists a distribution that achieves the upper bound.