Chapter 2

Bayes Decision Theory: Yuille, Coughlan, Kersten, Schrater

This chapter describes the concepts of Bayes Decision Theory and the basic tools associated with it. It is the fundamental approach to the statistical classification of data and was developed in the 1950's (see Duda and Hart for a basic introduction and see De Groot, Berger for more advanced treatment).

We will formulate Decision theory to include the visual tasks such as classification and estimation described in the previous chapter. This will give the basic theory behind the Granny example and the signal detection theory example. In this chapter, however, we restrict ourselves to simple generative models with trivial graph structures and no secondary variables. This restriction will be removed in later chapters of the book.

The general problem is to make an inference about a state s based on an observation x. The Bayesian framework assumes knowledge of how the observations are generated which is specified by conditional distributions $p(x|s)$. It also assume that there is prior knowledge about the state given by $P(s)$. The Bayesian approaches gives a posterior distribution $P(s|x)$ which combines the information from the observation with the prior knowledge. Typically the posterior distribution $P(s|x)$ does not determine the state uniquely (i.e. for any data x there are several different states s for which $P(s|x)$ is non-zero) and so inference of the state s is therefore an imperfect process. In general, any decision criterion for determining the state will make errors. Bayesian decision theory specifies the optimal procedure for estimating the state given one's relative tolerance for different types of errors.

Why do errors happen? Consider, for example, the task of discriminating between a dim and a bright light (described in the first chapter). There are many sources of stochastic fluctuations, for example from the measuring devices, so that there is always a chance that a sample of light from the dim source actually appears brighter than a sample from the bright source see figure (2.1).

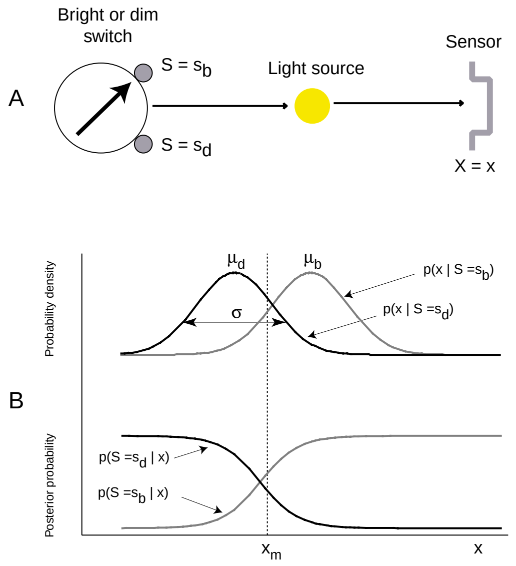


Figure 2.1 Panel A. From a measurement of the amount of light x, at the sensor, an observer's task is to infer whether the light switch was set to bright ($S = s_b$) or dim ($S = s_d$). Uncertainty due to quantal or thermal fluctuations in light emission, or sensor noise could result in the kind of distributions shown in the bottom panel. Other kinds of uncertainty, such as an unreliable "off switch" could result in bumpier distributions with more than one mode (such as in the right panel of figure 2.7). The upper graph of panel B shows the distributions of the photon counts we might measure depending on whether the switch is set to s_b or s_d. These are the likelihoods, $p(x|S = s_b)$ and $p(x|S = s_d)$, i.e. the probabilities of x conditioned on signal state s. The lower graph of panel B shows graphs of the posterior probabilities $P(S = s_b|x)$ and $P(S = s_d|x)$. For a given measurement x_m, the "dim" likelihood can be greater than the "bright" likelihood, while at the same time the "dim" posterior probability is less that the "bright" posterior. The reason is that the dim setting is twice as frequent as the bright setting, so $P(S = s_b|x) = p(x|S = s_b)(1/3)/p(x)$, and $P(S = s_b|x) = p(x|S = s_b)(2/3)/p(x)$. A decision based on the posteriors gives the smallest average error rate possible.

2.1 Bayes Decision Theory (I): Discrete state spaces.

Bayes decision theory assumes that we make an *observation* x which lies within *an observation set* X and we wish to deduce the *state* s which we know lies within a *state space* S. We assume that we have probability models $p(x|s)$, the probability of x *conditional* on the state s, which determine the probability of making the observation x when the state of the system is s. In addition, we have *a prior probability model* $P(s)$ for the set of possible states. We assume, for now, that the set of possible states is *discrete* and *finite*. The observation set X, however, may consist of either discrete or continuous variables.

Say we make an observation, x, and deduce that the state is $\hat{s}$ lying within a *decision space* D. The deduced state is presumed to be the result of a deterministic decision rule: $\hat{s} = \alpha(x)$, which may also be expressed as a probability $P(D = \hat{s}|x)$ that takes on values of exactly one or zero.

Due to uncertainty, our decision could be right or wrong–the deduced state $\hat{s}$ may not be the same as the actual state s–so a decision induces a joint event in the space $D \times S$. A performance cost or loss L is associated with this joint event. The influence relationships between the state, observation, decision and performance variables are illustrated in figure (2.2).

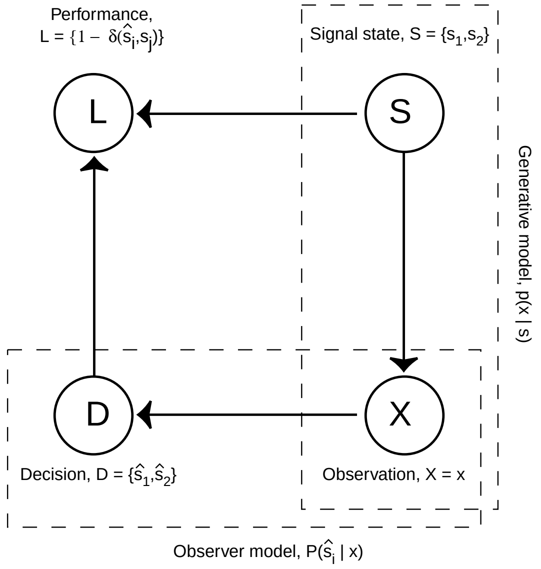

Performance,
L = {1 − δ(ŝ_i,s_j)}
Signal state, S = {s_1,s_2}
L
S
Generative model, p(x | s)
D
X
Decision, D = {ŝ_1,ŝ_2}
Observation, X = x
Observer model, P(ŝ_i | x)

Figure 2.2 Signal detection. The graphical model showing the relationship between the generative model, the observer model, and performance evaluation. The generative model specifies how observations (e.g. $x =$ photon count) are determined by signal states (e.g. $S = s_1 = s_b$ *for* "*bright*" or $S = s_2 = s_d$ *for* "*dim*"). In order to evaluate performance for a given decision rule, we need to consider two additional random variables, the decisions $\hat{s}$ and a loss variable (e.g. $l = 1 - \delta_{\hat{s},s}$) representing costs of correct and incorrect classifications. For detection a correct decision is either a hit, or correct rejection. An incorrect decision is either a false alarm or a miss. With these defined, we can then determine the distributions that characterize performance, $P(\hat{s}, s)$, $p(L = l)$, or average loss. Bayesian decision theory turns the problem around, and asks: given a desired performance cost such as average loss, what should the decision rule for the observer model be?

2.1.1 Light discrimination: An example from psychophysics

For example, consider the signal detection task described in Chapter 1 (figure 4). The state space consists of two elements s_2, s_1, which we will designate as s_d, s_b, for "dim" and "bright". The conditional distribution for generating the data assumes that the photons are generated by Gaussian distributions with mean and variance μ_b, σ^2 for bright light and μ_d, σ^2 for dim light.[1] So,

[1] Note that we are approximating physics by setting the number of photons to be a continuous value which might even be negative. This assumption can be justified based on the Central Limit theorem, see later chapter. Under ideal conditions, photon absorption is described by a Poisson distribution.

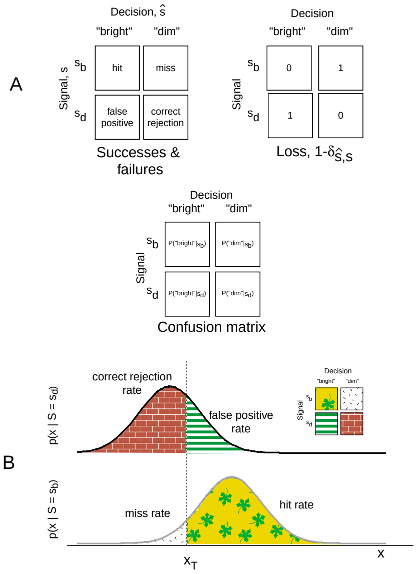


Figure 2.3 Detection performance. Panel A shows the different types of error, losses, and the confusion matrix. Panel B shows how the relationship between correct and incorrect classification rates and the areas under the densities for the dim and bright cases.

$$
\begin{aligned}
p(x|S=s_b) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu_b)^2/2\sigma^2} \\
p(x|S=s_d) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu_d)^2/2\sigma^2}.
\end{aligned}
\tag{2.1}
$$

The conditional probability densities, $p(x|S=s_b)$ and $(x|S=s_b)$, are shown in upper panel of figure 2.1 B. Let the prior probability that the state is dim be 1/3, and bright 2/3.

The task is to determine whether a bright ($S=s_b$) or dim ($S=s_d$) light was flashed. The observation x is the number of photons measured by the viewer who has to respond "dim" or "bright". We will represent the decision space $\{\hat{s}_2, \hat{s}_1\}$ as $\{\text{"}dim\text{"}, \text{"}bright\text{"}\}$. There are two ways of being right, and two ways of being wrong. If the switch is set to bright ($S=s_b$) and the observer says "bright", the observer scores a "hit". If the switch is set to dim ($S=s_d$), and the observer decides "dim", the observer scores a "correct" rejection. A false positive (or false alarm) is when the observer decides "bright", when the "dim" switch was set. A false negative (or miss) is when the observer decides "dim" when the "bright" was sent (see figure 2.3 A).

Suppose the bright light is defined to be the target (or "signal") and the dim light the non-target or "noise"[2]. One way of noting successes is to use the Kronecker delta

[2] In signal detection theory, the target is often referred to as the "signal", and the non-target as the

function $\delta_{\hat{s},s}$, to assign a one to a correct decision ($\hat{s} = s$), and a zero to an incorrect decision ($\hat{s} \neq s$).[3] Alternatively, we could view the glass as half-empty and count "losses", where a misclassification is one and a correct decision zero: $L(\hat{s}, s) = 1 - \delta_{\hat{s},s}$. This is a particular example of a *loss function*. Later we will introduce loss functions $L(d, s)$ that allow for alternative performance costs and decisions d which do not necessarily map onto the signal space.

Now suppose a series of observations x are made leading to decisions, $\hat{s}$. We can summarize average performance in terms of a "confusion matrix", $P(\hat{s}|s)$, which for the detection task is a 2×2 array representing the hit ($P(\hat{s}_1|s_1)$), false positive ($P(\hat{s}_1|s_2)$), false negative ($P(\hat{s}_2|s_1)$), and correct rejection rates ($P(\hat{s}_2|s_2)$) as shown in figure 2.3 B. These conditional probabilities are estimated by finding the proportions of "bright" and "dim" decisions under $S = s_b$ and $S = s_d$ conditions. E.g. the false alarm rate is calculated by counting the number of times the observer decides bright when the actual state was dim, divided by the number of dim state trials. Note that the hit and miss rates sum to one, and similarily for the false positive and correction rejection rates.

A useful summary measure of performance is the total error:

$$P(\hat{s} \neq s) = \sum_{\hat{s},s}(1 - \delta_{\hat{s},s})P(\hat{s}, s) = \sum_{i \neq j} P(\hat{s}_i, s_j),$$

which for our light detection example is,

$$P(\hat{s}_2, s_1) + P(\hat{s}_1, s_2) = P(D = \text{“}dim\text{”}, S = bright) + P(D = \text{“}bright\text{”}, S = dim).$$

The probability of error is, of course, one minus the probability of being correct: $P(\hat{s} \neq s) = 1 - P(\hat{s} = s)$.

How does the decision rule affect performance? Suppose we make an observation x_m, as shown in panel B of figure (2.1). One could base the decision on the size of x_m, i.e. the bigger the more likely it is to have come from the bright light. We'll see later that this intuition is right when the evidence grows monotonically with x, but could be quite wrong in other cases (e.g. see figure (2.9)). Further, it doesn't tell us at what value of x to start deciding "bright" rather than "dim". We require rules that have a well-defined relationship to performance. Another intuitively plausible observer model is

"noise". This is merely a convention, as both "signal" and "noise" cases can be viewed as two different signal states

[3] Kronecker was a nineteenth century German mathematician. An ex-banker who, believing that "the integers were made by God: all else is the work of man", took the unfashionable view that only "constructivist mathematics" was real and was an outspoken enemy of abstract mathematics such as Cantor's theory of sets. Curiously, Kronecker's views are in tune with modern computer science and the notion of proof by algorithm.

to choose the switch setting that has the biggest likelihood, i.e. either $p(x|S = s_b)$ or $p(x|S = s_d)$. This is called the maximum likelihood rule (ML). It does divide up the observation space into disjoint regions corresponding to the two decisions, and we'll see shortly how ML determines performance. However, it doesn't explicitly incorporate prior information about the probabilities of the two signal states, and thus doesn't provide sufficient information to achieve the best performance. A third possibility is to choose the switch setting that has the biggest posterior probability, i.e. either $P(S = s_b|x)$ or $P(S = s_d|x)$. This is the maximum a posteriori rule (MAP). As shown in the lower graph of panel B of figure (2.1), the MAP and ML decision rules affect performance differently. In fact, as we will prove, the MAP rule minimizes the probability of error.

There is a special condition for which using the ML rule gives the same performance as the MAP rule. That is when the prior probability is that the state has *equal probability* of being dim or bright, and hence $P(S = s_b) = P(S = s_d) = 1/2$. In this case the MAP decision is equivalent to the maximum likelihood rule. This is an important case because in the absence of information about the priors, ML is a good strategy. Further, we will see in the next section, how to use likelihoods and still take into account information about priors as well as performance goals that generalize beyond minimizing error.

2.1.2 Edge detection: An example from computer vision

Another example is to detect the boundaries of objects in real images, see figure (2.4). It is clear that the image intensity values typically change by a large amount at object boundaries. Conversely, intensity changes are typically small away from object boundaries. One may therefore hope to detect object boundaries by finding places where the image intensity changes a lot. But these are only typical properties and it is straightforward to find situations where the intensity change across an object boundary is small or, alternatively, situations where the intensity changes are large away from object boundaries. We can tackle this problem probabilistically be defining a filter $|\vec{\nabla} I(\vec{x})|$ which measures the local intensity gradient of the image. At each image location $(\vec{x})$, the gradient $\vec{\nabla} I(\vec{x})$ points in direction of maximum intensity change. By training on a set of examples images, with the ground-truth boundaries extracted by hand (see figure (2.4)), we can learn probability distributions $P(|\vec{\nabla} I(\vec{x})| = y| \ on)$ and $P(|\vec{\nabla} I(\vec{x})| = y| \ off)$. Learning is described in a later chapter, but the basic idea is to compile two sets of histograms. In the "on" histogram, we count the frequency of each filter response at those image locations where there is an edge in the ground-truth segmentation (figure (2.4)). The "off" histogram is compiled over those pixels where the ground-truth segmentation shows no edges. Such distributions are shown in figure (2.5). Just as with the light discrimination example, we can use these likelihoods to classify pixels into edges or non-edges. At each image location, and observation, $|\vec{\nabla} I(\vec{x})|$, is made. For ML, we choose $D =$ "*edge*" if $p(|\vec{\nabla} I(\vec{x})|S = edge)$ is larger than $p(|\vec{\nabla} I(\vec{x})|S = non - edge)$, whereas MAP says to choose $D =$ "*edge*" if $p(S = edge||\vec{\nabla} I(\vec{x}))$ is larger than $p(S = non - edge||\vec{\nabla} I(\vec{x}))$.

It should be realized that the observations need not be scalars (i.e. numbers). For example, a sophisticated boundary detection filter may process the image at a range of different scales (or resolutions). A standard way to achieve multiple scales is by convolving the image with a Gaussian filter $G(\vec{x};\vec{0},\sigma^2)$ to obtain a blurred image $I_\sigma(\vec{x}) = \int G((\vec{x} - \vec{z});\vec{0},\sigma^2)I(\vec{z})d\vec{z}$. (The intuition is that blurring the image will destroy the small scale structure of the image while preserving the large scale structure). We can then compute the *joint probability distributions of multiple edge filters* evaluated at four scales $\sigma = 0,1,2,4$:

$$\begin{aligned} &P(\{|\vec{\nabla} I_\sigma(\vec{x})| = y_\sigma :\ \sigma = 0,1,2,4\}|\ on) \\ &P(\{|\vec{\nabla} I_\sigma(\vec{x})| = y_\sigma :\ \sigma = 0,1,2,4\}|\ off). \end{aligned} \tag{2.2}$$

The joint probability distributions are needed because, in general, we expect the results of the filters at different scales will be dependent (e.g. correlated). In mathematical terms, we do not expect the distributions to factorize (i.e. we do expect that $P(\{|\vec{\nabla} I_\sigma(\vec{x}) = y_\sigma : \sigma = 0,1,2,4\}|\ on) = \prod_{\sigma=0,1,2,4} P(|\vec{\nabla} I_\sigma(\vec{x}) = y_\sigma|\ on)$.)

We expect that filters which combine information from multiple scales will be more effective than those which do not. This is illustrated in our results using ML estimation, see figure (2.6), and MAP estimation, see figure (2.10).

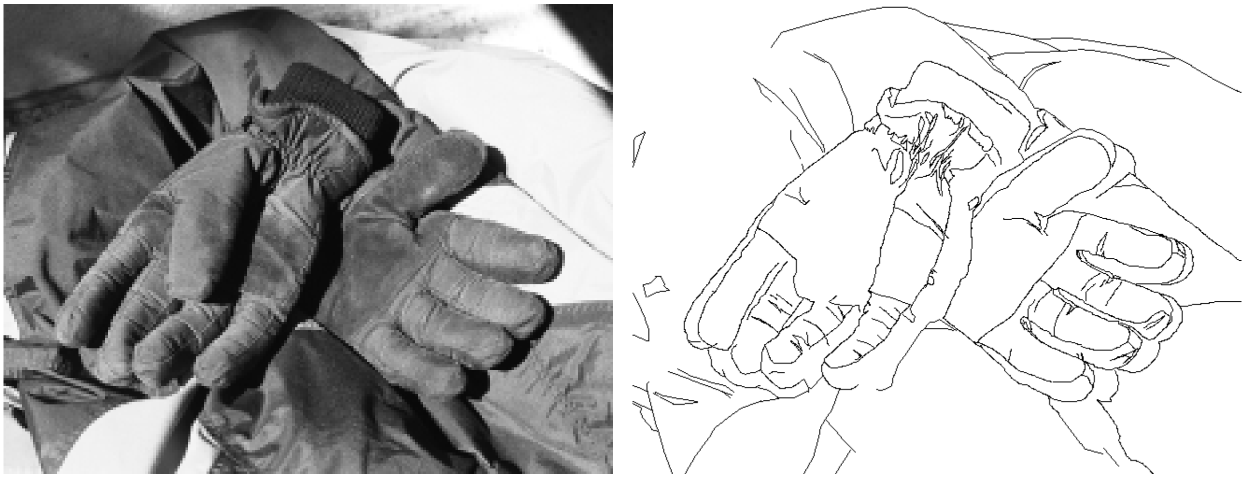


Figure 2.4 A typical image (left) and the ground truth segmentation (right). courtesy of K. Bowyer at U. South Florida.

In both cases, the task is to infer the state from the observation x. We've seen that there are two natural procedures *Maximum Likelihood* (ML) estimation and *Maximum a Posteriori* (MAP) estimation. Later we will see that both of these are subsumed as special cases of Bayesian decision procedures and correspond to different choices of *loss functions.*

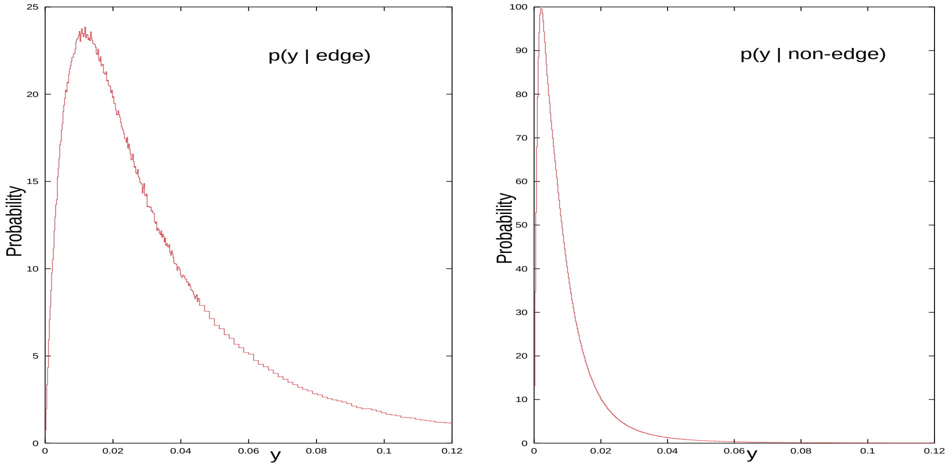


Figure 2.5 Empirical distributions for a gradient filter response on boundaries (left) and off boundaries (right).

2.1.3 ML estimation

ML estimation says that one should choose the state $s^* \in S$ which maximizes $p(x|s)$ (i.e. $s^* = \arg\max_{s \in S} p(x|s)$). Recall that $p(x|s)$ is the conditional probability distribution for x (i.e. $\sum_{x \in X} P(x|s) = 1, \ \forall \ s \in S$) but it is also known as the *likelihood function* for s (note: it is usually *not* a probability distribution for s because it is not normalized, i.e. $\sum_{s \in S} p(x|s)$ is usually not equal to 1, but it can happen for a subclass of Gaussian distributions.)

For our signal detection example, the ML estimate says that we should estimate the state to be "bright" if $P(x|S = s_b) > P(x|S = s_d)$ and to be "dim" otherwise. It is easily checked that this reduces to the condition – select "bright" if $x > (1/2)(\mu_b + \mu_d)$ and select "dim" otherwise. (Decision rules, such as ML, reduce to very simple forms if the distributions used are Gaussians – see later section!!).

Similarly, we can attempt to detect object boundaries using ML estimation based on the joint distributions $P(\{|\vec{\nabla} I_\sigma(\vec{x})| = y_\sigma \ : \ \sigma = 0, 1, 2, 4\}| \ on)$ and $P(\{|\vec{\nabla} I_\sigma(\vec{x})| = y_\sigma \ : \ \sigma = 0, 1, 2, 4\}| \ off)$ described earlier. This involves estimating that a pixel x with intensity gradients (at multiple scales) $|\vec{\nabla} I_\sigma(\vec{x})|$ is an object boundary if $P(\{|\vec{\nabla} I_\sigma(\vec{x})| = y_\sigma \ : \ \sigma = 0, 1, 2, 4\}| \ on) > P(\{|\vec{\nabla} I_\sigma(\vec{x})| = y_\sigma \ : \ \sigma = 0, 1, 2, 4\}| \ off)$ and is not a boundary otherwise. The results are shown in figure (2.6). Observe that the process produces a large number of pixels classified as edges. (Note that we are simplifying the problem greatly by assuming that we can classify boundary pixels independently of each other and ignoring the fact that they tend to be spatially correlated).

As we saw earlier, any estimation procedure will (almost always) make errors. Hence an important concept for any estimator is the *error rate* – the expected amount of data

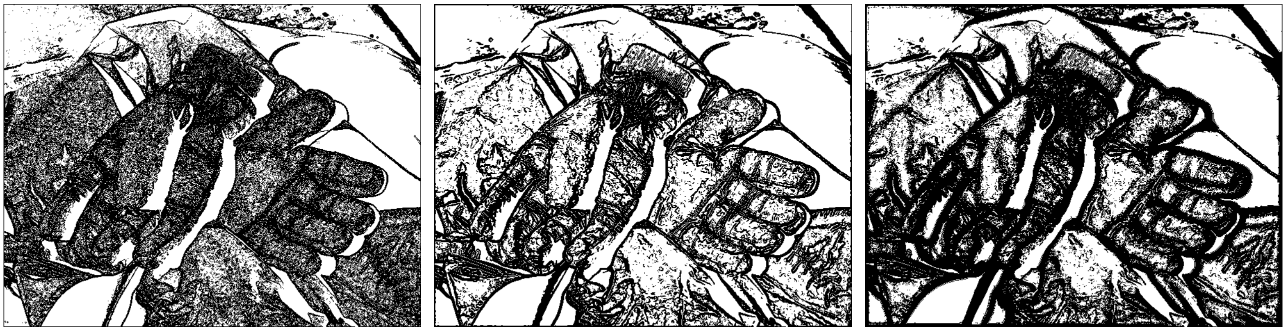


Figure 2.6 The edge estimated on the glove image by ML with the filter at scale 0 (left), filter at scale 1 (centre), and filter with scales $0, 1, 2, 4$ (right). Observe that ML significantly overestimates the number of edges in this image.

which is misclassified. For a binary decision process, where one state is labelled the target, this can be expressed as the expected number of false positives (non-target stimuli which are classified as target) and false negatives (i.e. "missses"–target stimuli which are classified as non-target). Let the target state be represented by s_1, and the non-target by s_2.

For the ML estimator, we define a target acceptance region $X_1 = \{x \in X : \log \frac{p(x|S=s_1)}{p(x|S=s_2)} \geq 0\}$ and $X_2 = \{x \in X : \log \frac{p(x|S=s_1)}{p(x|S=s_2)} < 0\}$. I.e. X_1 is the set of observations for which the ML estimate is "target", $D = \hat{s}_1$ and similarly X_2 corresponds to "non-target", $D = \hat{s}_2$. (Note that $X_1 \cup X_2 = X$ and $X_2 \cap X_2 = \emptyset$). The false positive rate is defined to be $F_+ = \int_{X_1} dx p(x|S = s_2)$ and the false negative rate is $F_- = \int_{X_2} dx p(x|S = s_1)$. I.e. F_+ is chance of misclassifying a non-target as a target and vice versa for F_-. It is clear, that the only way we can avoid errors is when the conditional distributions for the two states do not overlap – i.e. $P(x|S = s_1) \times P(x|S = s_2) = 0$ for all x (see figure (2.7)).

For our signal detection example, we can set the target to be "bright". Then the false positive and false negative rates are given by:

$$\begin{aligned} F_+ &= \int_{(1/2)(\mu_b+\mu_d)}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu_d)^2/2\sigma^2} dx. \\ F_- &= \int_{-\infty}^{(1/2)(\mu_b+\mu_d)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu_d)^2/2\sigma^2} dx. \end{aligned} \tag{2.3}$$

We get similar results for the false positives of the boundary detection based on the ML estimator. See figure.

The ML estimator is extremely powerful and has many applications. But it suffers from two weaknesses. The first is that it ignores any prior knowledge we may have about which states $s \in S$ are most likely to occur. The second is that it does not take into account the fact that certain types of estimation errors can be more serious than others

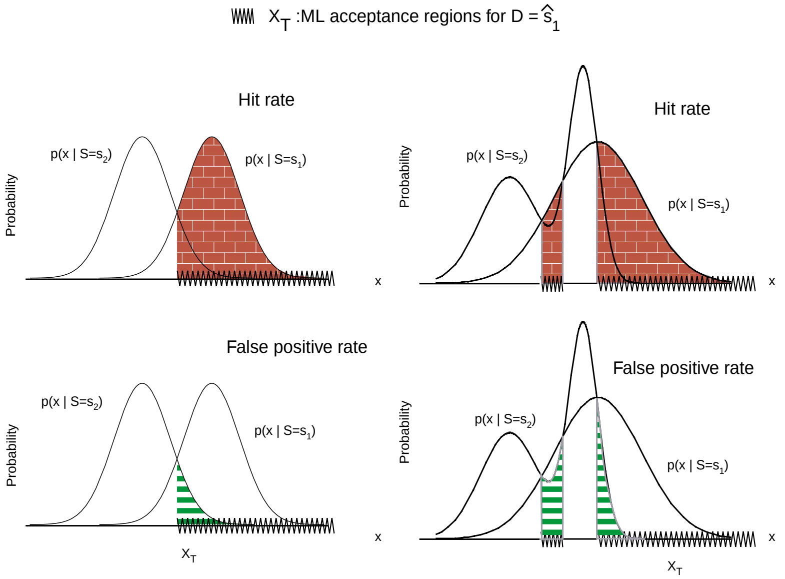


Figure 2.7 Hit and false positive rates for the maximum likelihood rule. The left panel shows the areas corresponding to the hit and false positive rates for our light detection example. The right panel shows the hit and false positive rates when one of the distributions has more than one mode. In this case, the ML acceptance region is disconnected.

– see example table.

****

EXAMPLE TABLE: THE NEED FOR PRIORS. (SOME OF THESE EXAMPLES COME FROM BERGER!!)

To motivate the importance of prior knowledge consider the following two examples: (I) A friend claims he can always tell Mozart's music from Handel and you want to test if he tells the truth or if he lies. You play him ten short pieces (randomly chosen – half from Mozart and half from Handel) and ask him to judge each piece. You represent his judgments depending on whether they are correct or not – this gives a string of true/false judgments, i.e. $T, F, T, T, F, F...$ which are the observables of the system. If he tells the truth then he will always get the right answer so $P(T| \; Truth) = 1$. But if he is lying, then he will probably just guess at random –so $P(T| \; Liar) = P(F| \; Liar) = 1/2$. Suppose his answers are all true –$T, T, T, T, T, T, T, T, T, T$ – then the likelihood function is $P(\vec{A}| \; Truth) = 1$ and $P(\vec{A}| \; Liar) = 1/2^{10}$, where $\vec{A} = (T, T, T, T, T, T, T, T, T, T)$. So the most likely interpretation is that your friend tells the truth.

(II) Now suppose you meet somebody in a bar who claims he can predict the tosses of an unbiased coin. This can be formulated in exactly the same way to estimate whether the person is a liar or not. Now suppose this person guesses right all 10 times. Do you really believe that they are telling the truth? Instead you probably suspect that somehow the person is tricking you. The difference in the two cases is the inherent probability of the claim in the two cases. Distinguishing Mozart from Handel is not hard (if you like classical music). But predicting a coin toss is generally believed to be impossible.

This problem arises also in more serious situations. Suppose you are being tested for HIV. It is known that certain population groups, such as intravenous drug takers, are particularly likely to get this disease and are classified as high risk groups. You need to take this prior knowledge into account when evaluating the results of tests for HIV which, like all tests, are not perfect. It has been estimated that somebody from a low risk group can have less than 20 % chance of being HIV positive even if the test comes back positive. We will discuss later, when we introduce loss functions, why having false positives may be a good idea.

****

2.1.4 MAP Estimation

MAP estimation differs from ML by making use of prior knowledge $p(s)$ of the state $s \in S$. This is the probability of s *before* we have made any observations. If no observations are available, or if an observer has dozed off, then his/her best strategy is to guess the state based on the prior. For our signal detection task the priors for the light being bright or dim were both set to 0.5. So the best strategy, without observation, is to guess "dim" or "bright" equally likely. If, however, the experiment is set up so that $P(s = dim) = 0.8$ then the best strategy, if no observation is available, is to guess that the light was "dim". (We note that, in some cases the use of priors can be avoided by simply redefining the state space and putting all the information into the likelihood function.)

Bayesian probability theory gives a way to combine information from measurements with prior knowledge. It yields the *posterior* distribution $p(s|x)$, the conditional probability of s given that we have made the observation x. *Bayes theorem* (Bayes 1783) expresses the posterior distribution in terms of the prior and the likelihood function:

$$P(s|x) = \frac{P(x|s)P(s)}{P(x)}, \tag{2.4}$$

where $P(x) = \sum_{s'} P(x|s')P(s')$ is a normalization factor which ensures that $p(s|x)$ is normalized (i.e. $\sum_{s \in S} p(s|x) = 1$). Bayes theorem can be proven by expressing the *joint* distribution $P(s, x)$ either as $P(s|x)P(x)$ or as $P(x|s)P(s)$, and then equating these expressions.[4] Bayes rule combines the likelihood function with the prior, by multiplication,

[4] Bayes was a church of England clergyman who dabbled in mathematics but never published anything in his lifetime. In his will, he left one hundred pounds and two scientific papers to another clergyman friend who liked the papers and got them published. They made little impact until the results were independently obtained a few years later by the great French mathematician Laplace. Laplace was initially given the credit until English scholars rediscovered Bayes' work.

to obtain a posterior distribution which is, hopefully, more sharply peaked than either. See figure (2.8) for an example.

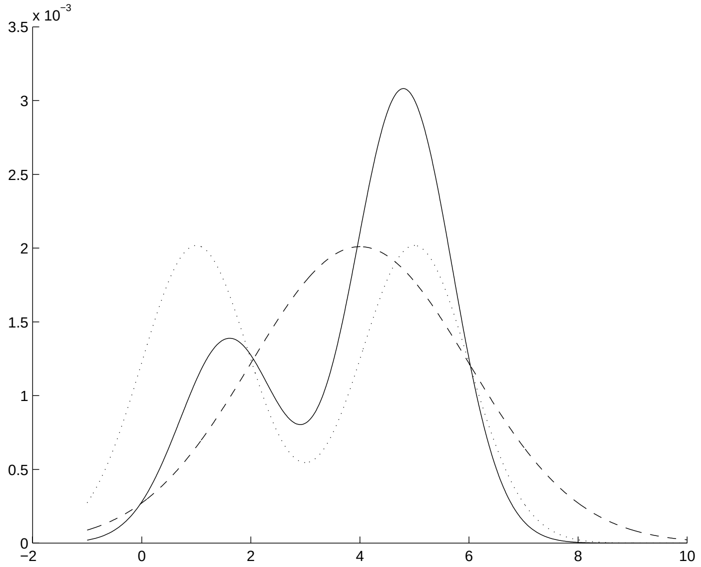


Figure 2.8 The likelihood and prior are shown as the dotted and dashed lines respectively. The posterior is shown by the regular curve. Observe that the posterior is more sharply peaked than either the prior or the likelihood function.

Given data x we want to seek the most probable, or *maximum a posteriori* (MAP) estimation of the state s. This corresponds to selecting the state s^* which maximizes the posterior distribution $P(s|x)$.

For example, if $P(S = s_d) = P(S = s_b) = 0.5$, then the MAP estimate for whether the light is "dim" or "bright" is given by the same criterion as ML – namely select "bright" if $x > (1/2)(\mu_b + \mu_d)$ and select "dim" otherwise. However, suppose we modify the experiment so that the priors can take arbitrary values. Then, applying MAP estimation says that you select "dim" if $x < x_T$, where the decision threshold, x_T, is found by solving:

$$log \frac{p(x|S = s_1)p(s_1)}{p(x|S = s_2)p(s_2)} = 0 \tag{2.5}$$

for x. Using the gaussian likelihoods for "dim" and "bright", we have

$$x_T = (1/2)(\mu_b + \mu_d) + \frac{\sigma^2}{(\mu_b - \mu_d)} \log \frac{P(s_d)}{P(s_b)}. \tag{2.6}$$

We can again calculate the error rates for the MAP estimator. The only difference from the ML case is that we now use the prior probabilities to balance the relative weighting of different types of errors (for example, false positive and false negative).

For the MAP estimator, we define $X_1 = \{x \in X : \log \frac{p(x|S=s_1)p(s_1)}{p(x|S=_1)p(s_2)} \geq 0\}$ and $X_2 = \{x \in X : \log \frac{p(x|S=s_1)p(s_1)}{p(x|S=s_2)p(s_2)} < 0\}$ (note that $X_1 \cup X_2 = X$ and $X_2 \cap X_2 = \emptyset$ as before) (But note that for non-uniform priors, these MAP regions X_1 and X_2 are *not* the same as the ML ones determined earlier!–see Panel B of figure (2.9)).

The false positive rate is defined to be $F_+ = \int_{X_1} dxp(x|S = s_2)$ and the false negative rate is $F_- = \int_{X_2} dxp(x|S = s_1)$. The overall error rate F_{MAP} is

$$\begin{aligned} F_{MAP} = p(s_2) \int_{X_1} dxp(x|S = s_2) + p(s_1) \int_{X_2} dxp(x|S = s_1) \\ = p(s_2)F_+ + p(s_1)F_- \qquad . \end{aligned} \tag{2.7}$$

One can also calculate the error rate starting from a Bayes net perspective using the joint probability structure represented in the graph of figure (2.2). Let the loss $L = 1 - \delta_{\hat{s},s}$ be a random variable, then $p(L = 1)$ is the error rate. We calculate $P(L = l)$ by marginalizing $p(l, \hat{s}, s, x)$ with respect to $\hat{s}$,s, and x:

$$p(l) = \sum_{\hat{s},s} \int_{-\infty}^{\infty} p(l, \hat{s}, s, x)dx =$$

$$\sum_{\hat{s},s} \int_{-\infty}^{\infty} p(l|\hat{s}, s)p(\hat{s}|x)p(x|s)p(s)dx$$

where in order to factor the joint probability, we have used the fact that $\hat{s}$ is conditionally independent of s (given x), and that l is conditionally independent of x, (given $\hat{s}$ and s) (see the influence arrows in the graph of figure 2.2). The probability $p(\hat{s}|x)$, represents the deterministic decision rule, or "indicator function", $\phi_{\hat{s}}(x)$ for the acceptance region for $\hat{s}$ (see panel A of figure (2.9)). The probability of l conditional on $\hat{s}$ and s also represents a deterministic rule and can be written:

$$P(L = l|\hat{s}, s) = (1 - \delta_{\hat{s},s})l + \delta_{\hat{s},s}(1 - l)$$

.

Letting $L = 1$ gives us the overall error rate, F_{MAP}:

$$\begin{aligned} P(L = 1) = \sum_{\hat{s},s} \int_{-\infty}^{\infty} (1 - \delta_{\hat{s},s})p(\hat{s}|x)p(x|s)p(s)dx = \\ p(s_2) \int_{-\infty}^{\infty} dxp(\hat{s}_1|x)p(x|s_2) + p(s_1) \int_{-\infty}^{\infty} dxp(\hat{s}_2|x)p(x|s_1) \\ = p(s_2)F_+ + p(s_1)F_- \qquad . \end{aligned} \tag{2.8}$$

where $p(\hat{s}_i|x)$ is the indicator function ($\phi_{\hat{s}_i}(x)$) for the acceptance region, X_i for $\hat{s}_i$.

Similarly, calculating $P(L = 0)$ gives us the probability of being correct.

For our signal detection example, this gives

$$F_{MAP} = P(S=s_d)\int_{x_T}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu_d)^2/2\sigma^2} dx + P(S=s_b)\int_{-\infty}^{x_T} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu_d)^2/2\sigma^2} dx, \tag{2.9}$$

where $x_T = (1/2)(\mu_b+\mu_d) + \frac{\sigma^2}{(\mu_b-\mu_d)} \log \frac{P(s_d)}{P(s_b)}$.

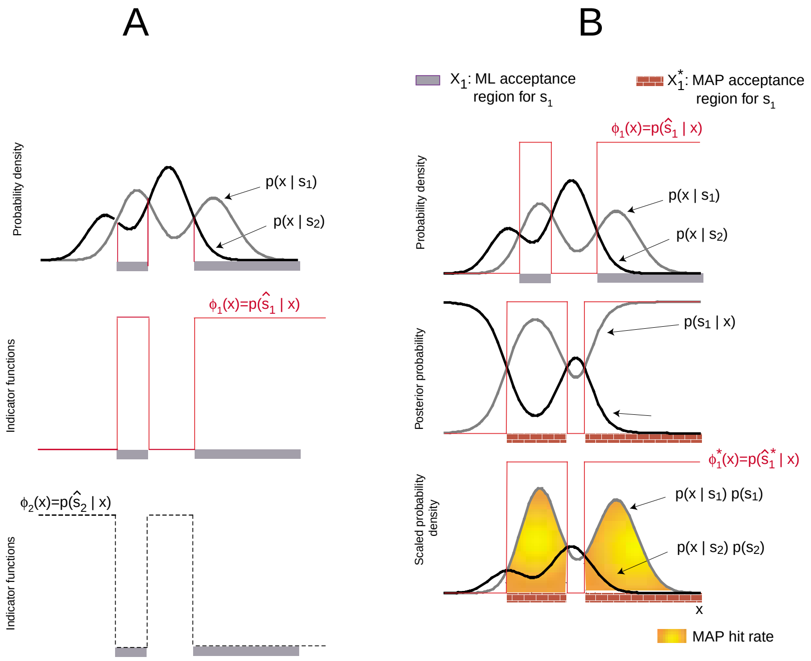


Figure 2.9 Acceptance regions for ML and MAP.

We can now modify our boundary detection method to take into account prior information. It is certainly clear that most typical images all contain far more pixels which are off-boundaries than on-boundaries. From image datasets we can estimate that about ninety percent of pixels in a typical image are off-boundaries. This gives us a prior $P(off) = 0.9$ and $P(on) = 0.1$. We apply MAP estimation in this case, see figure (2.10). Observe that number of pixels classified as boundaries has been greatly reduced compared to the ML estimator. (Note that, as before, we are simplifying the problem greatly by assuming that we can classify boundary pixels independently of each other and ignoring the fact that they tend to be spatially correlated).

So far, we have discussed how error rates occur starting from specific choices of estimators (ML and MAP). In the next section, we give a more general formulation based on loss functions whereby the estimators can be derived from the desired criteria of the error rates. This also allows us to take into account the fact that certain errors may be more serious than others.

****

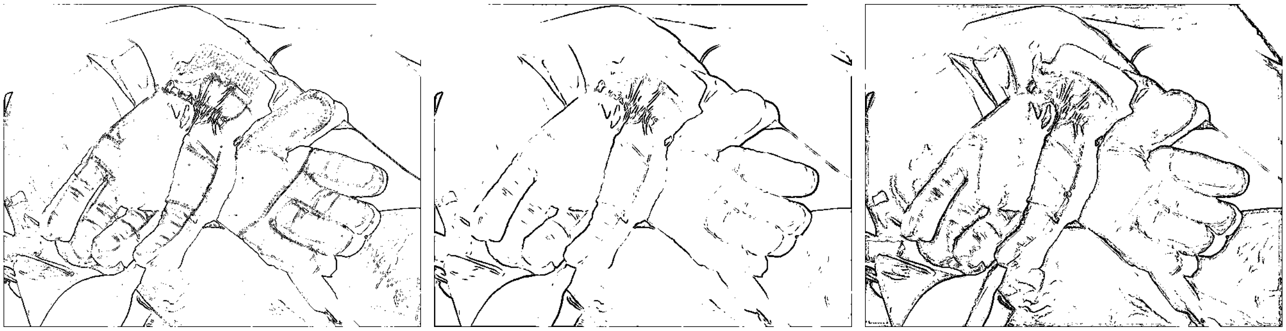


Figure 2.10 The edge estimated on the glove image by MAP with the filter at scale 0 (left), filter at scale 1 (centre), and filter with scales $0, 1, 2, 4$ (right). Observe that the MAP estimates are significantly better than the ML estimates for this image. Also combining filters at multiple scales improves the performance.

THE NEED FOR LOSS FUNCTIONS.

In certain decision tasks the consequences of making different types of error may be very different. For example, when testing for a dangerous infectious disease such as AIDS it is clearly better to make the test conservative so that it produces more false positives than false negatives. Loss functions give a way of taking into account the consequences of errors made in the decision process.

****

2.1.5 Loss Functions and Decision Spaces

To obtain the most general decision criteria, we must introduce two extra ingredients. The first is a *decision space* D. Up until this point, we've assumed that the decision space maps directly to the state space ($d = \hat{s}$). We now relax this restriction. The second is a loss function, defined on $D \times S$, which gives a loss $L(d, s)$ for making a decision d when the true state is s.[5] If we had perfect knowledge of the state s then we would simply select the decision d which minimizes $L(d, s)$. In practice, direct observation of s is impossible and we can only estimate s indirectly from observations. We therefore define a *risk function*:

$$R(d; x) = \sum_s L(d, s) P(s|x) = \sum_s L(d, s) \frac{P(x|s)P(s)}{P(x)}. \tag{2.10}$$

The Bayes decision d^*, for the input stimulus x, is chosen to minimize the risk $R(d; x)$. The Bayes risk is the value $R(d^*; x)$.

A decision procedure, however, requires specifying a decision for *all possible input stimuli* x. Moreover, its effectiveness should be evaluated over this stimulus set. In other

[5] The dominant convention is to view the glass as half-empty and calculate losses, but the complementary *gain function* is also used by more optimistic statisticians.

words, we should care about the risks for *all the possible input stimuli x.* Not all stimuli, however, are equally likely and so we should weight their importance by how frequently they occur. This motivates the following definitions: the optimal *decision rule* $\alpha(x)$ is defined by $\alpha(x) = \arg\min_{d \in D} R(d;x)$. This decision rule also minimizes the *expected risk* averaged over all observations:

$$R(\alpha) = \sum_x R(\alpha(x);x)P(x). \tag{2.11}$$

We now describe how we can get all standard estimation methods just by suitable choices of: (i) the likelihood $p(x|s)$, (ii) the prior $p(s)$, and the (iii) the loss function $L(d,s)$.

To obtain the ML estimator, we simply set the prior to be the *uniform* distribution, so that $P(s)$ is constant. This means that we need only consider the likelihood function. Then we choose a loss function which give us a loss of 1 unless we have exactly the right answer, which gives us a loss of zero. For discrete variables, we can set $L(d,s) = 1 - \delta_{d,s}$, where $\delta_{d,s}$ is the Kronecker delta function ($\delta_{d,s} = 1$ if $d = s$ and is zero otherwise.)

To get the MAP estimator we simply keep the prior term $p(s)$ and use the same loss functions as above. So the risk $R(d;x)$ can be expressed as:

$$R(d;x) = \sum_s L(d,s)P(x|s) = \sum_s \{1 - \delta_{d,s}\}P(s|x) = 1 - P(d|x), \tag{2.12}$$

and the decision which minimizes the risk is the one that maximizes the posterior probability $P(d|x)$, and hence is the MAP estimator.

Observe also that using the loss function $L(d,s) = 1 - \delta_{d,s}$ means that we pick the decision rule which minimizes the expected number of classification errors. This is because:

$$R(\alpha) = \sum_x \sum_s L(s,\alpha(x)) \frac{P(x|s)P(s)}{P(x)} P(x) = \sum_s P(s) \sum_x \{1 - \delta_{\alpha(x),s}\}P(x|s), \tag{2.13}$$

and the second term is the expected number of errors using the decision rule when the state is s.[6]

[6] Here is an alternative perspective on why MAP minimizes average error. Suppose that x is fixed at a value for which $P(S = s_b|x) > P(S = s_d|x)$. This is exactly like the problem of guessing "heads" or "tails" for a biased coin, say with a probability of heads $P(S = s_b|x)$. Imagine the light discrimination experiment repeated many times and you have to decide whether the switch was set to bright or not–but only on those trials for which you measured exactly x. The optimal strategy is to always say "bright" (see Exercise 1), and this results in a probability of error $P(error|x) = P(S = s_d|x)$. That's the best that can be done on average. On the other hand, if the observation is in a region for which $P(S = s_d|x) > P(S = s_b|x)$, the minimum error strategy is to always pick "dim" with a resulting $P(error|x) = P(S = s_b|x)$. Of course, x isn't fixed from trial to trial, so we calculate the total probability of error which is determined by the specific values where signal states and decisions don't agree:

Loss functions are often used to take into account the fact that different types of error can be more serious than others. For example, suppose in our signal detection case the observer receives different positive feedback based on which type of correct estimates he/she makes. For example, the observer may receive a Hershey bar if he/she correctly estimates the state to be "dim" and Godiva chocolate if correctly estimates it to be "bright". This will lead the observer to have an asymmetric loss function. For example, $L(s,d) = -A_s \delta_{d,s}$ where $A_{S=b}$ and $A_{S=d}$ reflect the observer's preference for Hershey versus Godiva. (We will return to the issue of observers' loss functions in the signal detection theory section later in this chapter).

For detecting boundaries, it may sometimes be best to prefer a strategy which overestimates the number of boundaries rather than underestimates them. This is because, in many practical computer vision applications, it is often easier to "prune out" false positives rather than recover from false negatives.

From our perspective, there is no ideal estimator suitable for all problems. The choice of estimators should depend on the problem formulation, whether the posterior distributions are sharp enough to allow precise inferences to be made, and on the goals of the system. In many cases, MAP and ML are excellent estimators but it is necessary to understand the, often hidden, assumptions they make before applying them and to determine form of the problem structure, and whether the assumptions are valid.

2.1.6 Two Hypotheses: the Log-Likelihood Ratio Test

We now concentrate on the special case where we have only two hypotheses s_1, s_2 and two corresponding decisions d_1, d_2. The loss function $L(d,s)$ can be represented by a two-dimensional square matrix $\mathbf{L}$ with elements $L_{ij} = L(d_i, s_j)$ for $i,j = 1,2$. We calculate the risk to be:

$$R(d_i; x) = \sum_j L_{ij} P(s_j|x), \tag{2.14}$$

which for each of the two decisions is:

$$R(d_1 : x) = \frac{1}{P(x)} \{L_{11} P(x|s_1) P(s_1) + L_{12} P(x|s_2) P(s_2)\}$$

$$P(error) = \sum_{i \neq j} P(\hat{s}_i, s_j) =$$

$$\sum_{i \neq j} \int P(\hat{s}_i, s_j|x) p(x) dx$$

Because the MAP rule ensures that $P(\hat{s}_i, s_j|x)$ is the minimum for each x, the integral over all x minimizes the total probability of error.

$$R(d_2 : x) = \frac{1}{P(x)} \{L_{21} P(x|s_1) P(s_1) + L_{22} P(x|s_2) P(s_2)\}. \tag{2.15}$$

The decision, whether $R(d_1; x) > R(d_2; x)$, is then equivalent to whether:

$$\log \frac{P(x|s_1)}{P(x|s_2)} > \log \frac{P(s_2)}{P(s_1)} + \log \frac{(L_{22} - L_{12})}{(L_{11} - L_{21})}. \tag{2.16}$$

Therefore the decision depends on the data x only in terms of the *log-likelihood ratio* $\log \frac{P(x|s_1)}{P(x|s_2)}$. The decision process can then be reformulated as a *log-likelihood ratio test* with a decision threshold T[7]:

$$\begin{aligned} &If\ \ \log \frac{P(x|s_1)}{P(x|s_2)} > T\ select\ s_1, \\ &If\ \ \log \frac{P(x|s_1)}{P(x|s_2)} < T\ select\ s_2. \end{aligned} \tag{2.17}$$

The log-likelihood ratio is a very important concept and will reoccur throughout this book. It can be derived from first principles by the Neyman-Pearson lemma, given below, without requiring the assumptions of Bayesian decision theory. (In our derivation the decision threshold T is determined in terms of the priors, $P(s_1), P(s_2)$, and the loss function $\mathbf{L}$.).

The functional forms of the log-likelihood ratios for the bright/dim and the boundary estimation tasks are given in figure (2.11). Observe that for the bright/dim task, the log-likelihood ratio becomes $x\{(\mu_1 - \mu_2)/\sigma^2\} + (\mu_2^2 - \mu_1^2)/2\sigma^2$. (We've let $\mu_b = \mu_1$ and $\mu_d = \mu_2$.) It is therefore linear in x with gradient $(\mu_1 - \mu_2)/\sigma^2$. From the forms of the log-likelihood functions we can get an idea about how sensitive an estimation result is to the choice of threshold. In some experimental situations, such as those treated in signal detection theory, the experiments conditions are modified so that the decision criterion of the subject keeps changing. The experimenter is thus able to probe the entire form of the log-likelihood function. For historical reasons, see later section, these experiments have usually been done with the likelihood functions being Gaussians with equal variance (i.e. like our bright/dim task). Determining the likelihood function is therefore equivalent to determining the *signal-to-noise ratio* $(\mu_1 - \mu_2)/\sigma$, which is called d', and which determines the difficulty of the discrimination task.

The log-likelihood ratio can also be thought of as "evidence". The point is that if we have two pieces of independent data x_1, x_2, then the log-likelihood ratios for their combination add. This is because, by assumption, $P(x_1, x_2|s) = P(x_1|s)P(x_2|s)$, so that

[7] Psychophysicists typically refer to the decision threshold as the "criterion", and reserve the term "threshold" to mean the value of an experimenter-controlled parameter, such as light intensity, required to achieve a fixed performance level (e.g. percent correct in a two-alternative forced-choice task). In this chapter, the term threshold is always taken to mean decision threshold or criterion.

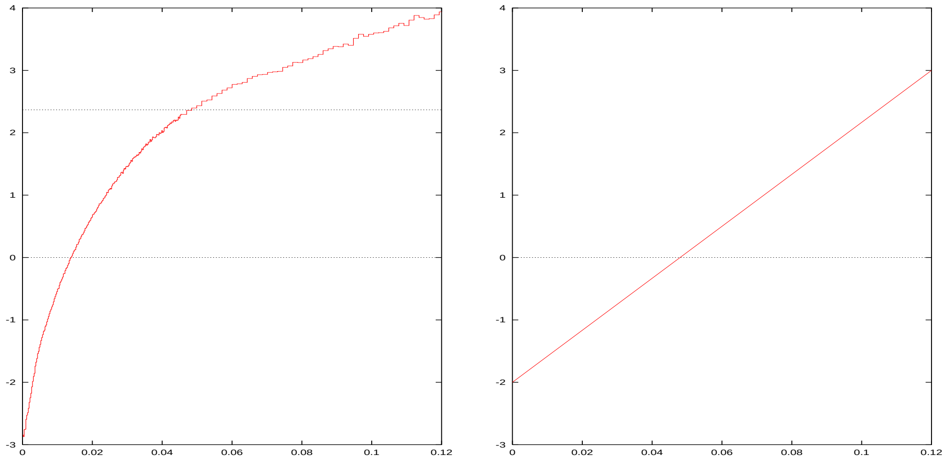


Figure 2.11 The log-likelihood ratios for the boundary detection task (left) and for Gaussian estimation of bright/dim (right). The ML estimates are made where the log-likelihood crosses the axis. The MAP estimate for boundary detection occurs at the threshold shown by the upper horizontal line (left) at approximately 2.3.

$log\frac{P(x_1,x_2|s_1)}{P(x_1,x_2|s_2)} = \log\frac{P(x_1|s_1)}{P(x_1|s_2)} + \log\frac{P(x_2|s_1)}{P(x_2|s_2)}$. It is more intuitive to think of evidences as adding.

Justification for the log-likelihood ratio test (independent of the Bayesian Decision theory setup) is given by the classic Neyman-Pearson lemma. This states:

Theorem. *Neyman-Pearson lemma). Let $x_1, ..., x_n$ be drawn i.i.d. and let the error rates be $\alpha(T)$ and β for a log-likelihood ratio test with threshold $T \geq 0$. Then for any other test with error rates α', β' then either $\alpha^{p}rime \geq \alpha$ or $\beta' \geq \beta$.*

Proof. *Let $\phi_1(x)$ and $\phi_1'(x)$ be the indicator functions associated with the likelihood ratio test and second test, respectively. That is, $\phi_1(x)$ is unity where the likelihood ratio classifies x as s_1, and zero otherwise, and similarly $\phi_1'(x)$ indicates where the second test classifies x as s_1. Formally, $\phi_1(x) = 1$ if $x \in X_1$ else $\phi_1(x) = 0$, and $\phi_1'(x) = 1$ if $x \in X_1'$ else $\phi_1'(x) = 0$, where X_1 and X_1' are the sets of x for which the likelihood ratio and second tests classify x as s_1, respectively (see Panel A of figure (2.9)). Then for all x: $(\phi_1(x) - \phi_1'(x))(P(x|s_1) - e^T P(x|s_2)) \geq 0$ (this can be verified by considering the cases $x \in X_1$ and $x \notin X_1$ separately). Integrating this inequality leads to $\int_{X_1}(P(x|s_1) - e^T P(x|s_2))dx - \int_{X_1'}(P(x|s_1) - e^T P(x|s_2))dx \geq 0$. This yields $e^T(\beta' - \beta) - (\alpha - \alpha') \geq 0$. Hence result.*

2.1.7 Classification with multiple decisions

Bayesian decision theory extends straightforwardly to situations where there are multiple decisions. For example, we may attempt to classify pixels in an image based on their colours as road, sky, vegetation, edge, or other. See figure (2.14).

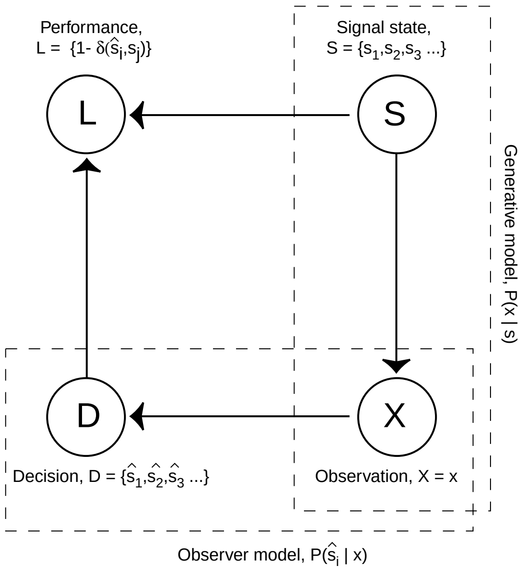


Figure 2.12 Classification with discrete multiple decisions.

More generally, let there be c classes $s_1, ..., s_c$. Patterns belong to class s_i with prior $P(s_i)$. Pattern x (some feature vector) modelled by $P(x|s_i)$ if x belongs to i^{th} class. We compute the posterior probability:

$$P(s_j|x) = \frac{P(x|s_j)P(s_j)}{\sum_k P(x|s_k)P(s_k)}. \tag{2.18}$$

The MAP choice is the j_{th} class with the largest posterior:

$$argmax_j\{P(s_j|x)\}. \tag{2.19}$$

This a straight forward extension of the detection case to more than one signal state (see figure (2.12)). As before we can use loss functions to take into account other performance costs.

Let $\hat{s}(x)$ be the decision rule. Let $L(\hat{s}_i; s_j)$ be the loss if decision $\hat{s}_i$ is made when the true state is s_j. The risk is the expected loss given data x and decision $\hat{s}_i$ (averaged over the states):

$$R(\hat{s}_i; x) = \sum_{j=1}^{c} L(\hat{s}_i; s_j)P(s_j|x) \tag{2.20}$$

Bayes decision rule: $\hat{s}^*(x) = s_i, \;\; if \; R(\hat{s}_i; x) \leq R(\hat{s}_j; x), \; \jmath = 1, ...c$

Various possible loss functions and associated acceptance regions. E.g. set $L(\hat{s}_i; s_i) = 0$ $\forall i$, $L(\hat{s}_i; s_j) = 1$ for $i \neq j$ and $L(\hat{s}_0|s_j) = L_r$ for states that we reject completely (i.e.

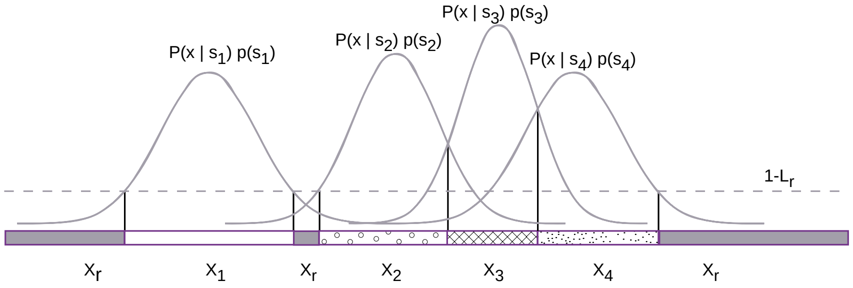


Figure 2.13 Classification with multiple decisions.

don't classify as one of the classes). Then Ω_i, the set of all observations that we classify as belonging to class i, is:

$$\Omega_i = \{x|P(s_i|x) = max_j P(s_j|x) \geq 1 - L_r\} \tag{2.21}$$

and

$$\Omega_0 = \{x|1 - L_r > max_j P(s_j|x)\}, \tag{2.22}$$

is the set of observations that we reject (see figure (2.13)).

So the acceptance probability and reject rates for decision $\hat{s}^*(x) = s_i$ are:

$$A^{*i} = \int_{\Omega_i} P(x)dx \quad R^* = \int_{\Omega_0} P(x)dx. \tag{2.23}$$

We can also compute the confusion matrix $\mathbf{C}$ with elements $C_{ij} = P(s_i|\hat{s}_j)$.

In figure (2.14), we give show the classification error rates for this procedure taken over a dataset set of images taken in the English countryside. The likelihood functions and the priors are learnt from previously classified data. Observe, that some classes like sky, road, and vegetation can be classified with good accuracy rates using this approach.

2.2 Decision Theory for Continuous Variables

We now consider what happens if the state space is continuous. This occurs, for example, when we seek to estimate the depth to an object or estimate the brightness of a stimulus. The basic concepts of decision theory transfer over directly(see figure (2.15). But there are some new features that we must pay attention to and some dangers that we must avoid.

First, recall from Appendix!!, that continuous probability distributions require specifying a probability density function $p(x)$ so that the probability of an outcome being in the infinitesimal interval $x, x + \delta x$ is given by $p(x)\delta x$.

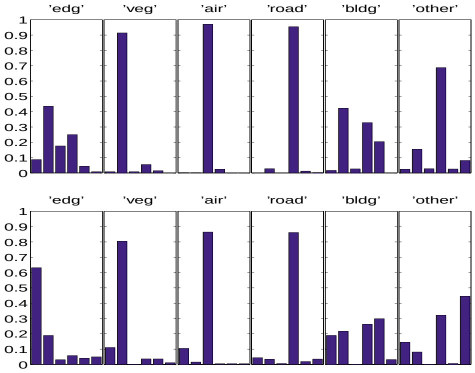


Figure 2.14 Colour for classification of image pixels: Confusion plot with true prior. The top row shows the probabilities, $P(s^*|s)$, of the classification s^* when the true class is s. I.e. the top leftmost panel shows $P(s^*|edge)$ for s^* being $\{edge, /vegetation, /air, /road, /building, /other\}$. Observe that the classifications of *veg, air, road* are over 90%. The bottom row shows $P(s|s^*)$ with similar conventions. Observe that if a pixel is classified as *vegetation* then it has an over 80% chance of really being vegetation.

2.2.1 Basic Decision Theory for Continuous Variables

Again, we have observations x, state variables s lying in a state space W, conditional density function $P(x|s)$, priors $P(s)$, decisions d in a decision space D, and the loss functions $L(s, d)$. The estimator aims to minimize the risk:

$$R(d; x) = \int L(s, d) \frac{P(x|s)P(s)}{P(x)} ds. \tag{2.24}$$

The optimal *decision rule* $\alpha(x)$ is defined by $\alpha(x) = \arg\min_{d \in D} R(d; x)$. This will also minimize the expected risk averaged over all possible distributions $P(x)$ of the observations. (This density $P(x) = \int P(x|s)P(s)ds$ so it assumes that the prior is correct.)

To obtain the MAP estimator we set $L(s, d) = -\delta(s - d)$ where $\delta(s - d)$ is the Dirac delta function.[8]

Observe that this loss function only rewards us if we estimate the answer correctly, i.e. if $d = s$. It does not give partial credit for a near miss. This is a serious problem if the task is to estimate a continuous variable such as the angle of rotation of wheel. At best one could hope to estimate this angle to within a small error range, perhaps of a few degrees. What sense does it make to use a loss function which rewards you only if your estimate is perfectly correct? In such cases, it make be better to use alternative loss functions – such as quantizing the angles and rewarding the decision if you are within a

[8] Dirac was a twentieth century English physicist who received a Nobel prize for his fundamental contributions to the theory of quantum mechanics. He invented the delta function as a non-rigourous heuristic tool which mathematicians later made honest by developing distribution theory.

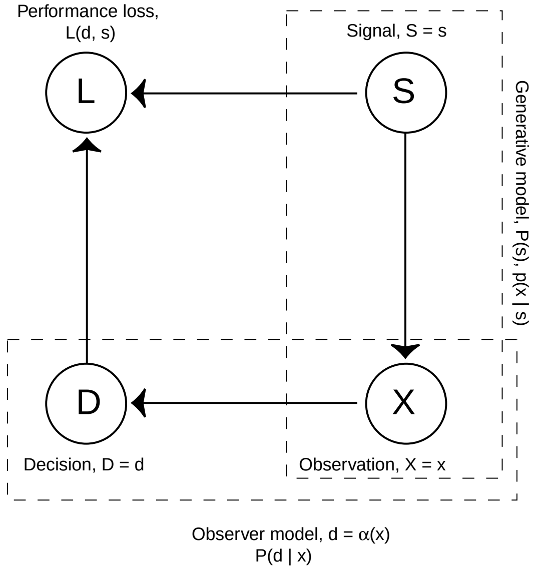


Figure 2.15 The general case. Signal states, s and estimations d are discrete or continuous. Observations x may be vectors. Performance is evaluated in terms of a loss function $L(d, s)$. d .

fixed number of degrees.

For example, suppose we only want to estimate the angle to within a tolerance of $2\pi\Delta$ radians. We can quantize the decision space to take $N = 1/\Delta$ values $\{2n\pi/N : n = 1, ..., N\}$. We choose a loss function $L(\theta, n) = -1$ if $|\theta - 2n\pi/N| < 2\pi\Delta$, and $L(\theta, n) = 0$ otherwise. This will enable us to estimate the angle to the closest element of the decision space.

Another possibility is to try to minimize the expected variance of the estimate. This can be done by setting $L(d, s) = (d - s)^2$. It is a simple exercise to deduce that the best estimate is then the *mean* $\int dssP(s|x)$ of the posterior distribution. This, however, is a bad estimator if the posterior distribution $P(s|x)$ has multiple peaks.

One way to get insight into how the use of loss functions affects the estimation process is by considering the special case where the loss functions are of form $L(d, s) = f(d - s)$ for some function $f(.)$. (Observe that the error functions $L(s, d) = -\delta(s - d)$ and $L(d, s) = (d - s)^2$ discussed above are both of this form). The risk function, $R(d; x) = \int dsf(d-s)p(s|x)$, is then the *convolution of the posterior density function* $p(s|x)$ *with the function* $f(.)$. We than therefore think of the loss function as a way of "smoothing" the posterior distribution. One possibility (Freeman and Brainard, Yuille and Bulthoff) is to use a Gaussian loss function, $L(d, s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(s-d)^2/2\sigma^2}$ because of the nice smoothing properties of such a function, see figure (2.17).

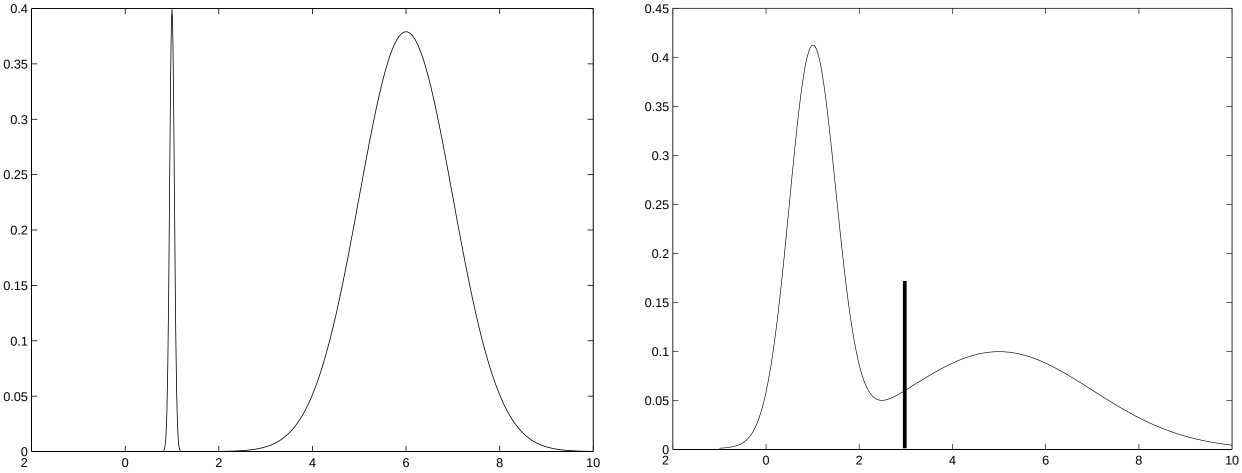


Figure 2.16 For some distributions (left) the minimal variance estimator (the mean) give results which seem better result than those of the MAP estimator. But for other distributions (right) the results from the minimal variance estimator look bad. In general, the choice of estimator should depend on the distributions and the importance of different types of errors.

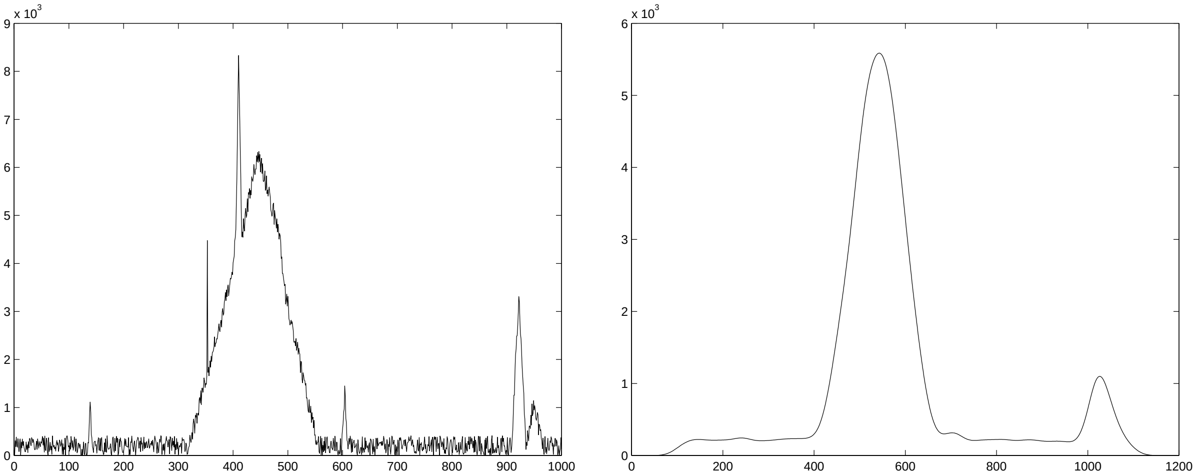


Figure 2.17 We have a jagged posterior distribution (left) which gets "smoothed" using a Gaussian loss function (right). Observe that the estimate will shift from the narrow spike peak (on left) to a wider peak with more support.

2.2.2 The dangers of coordinate system transforms

Another problem with MAP, or ML, estimation for continuous variables is that the state estimate depends on the coordinate system that the problem is formulated in. Suppose, for example, that we estimate the binocular *disparity* d of a stimulus using MAP, or ML. We may then want to compute the *depth* D using the assumption that depth from the fixation plane is inversely proportional to the disparity, $D = F/d$, where F is a constant. If the theory measures the disparity of a binocular stimulus to be μ then a simple application of the formula above gives the depth from fixation to be F/μ. This seems obvious, but is it correct? The (perhaps) surprising answer that F/μ is the best maximum likelihood estimate of the depth *only if we have measured the disparity to perfect precision.* In practice, any system will only be able to estimate the disparity to within a certain error. For illustration, we assume that we can model this error by a Gaussian distribution with standard deviation σ. This means that, after our disparity measurement process, we can say that:

$$P(d) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(d-\mu)^2}{2\sigma^2}}, \tag{2.25}$$

.

where the *best disparity estimate* is indeed μ.

To determine the best maximum likelihood estimate for the depth from fixation we must derive a density function $P(D)$, where $D = F/d$, see figure (2.18). This involves changing the coordinate system. Recall, see appendix, that this involves equating *the probabilities* $P(D)\delta D$ with $P(d)\delta d$. More formally, the rule for transforming between density functions must take into account the *change in area* $\delta D/\delta d$ caused by the transformation. This gives $P(D) = P(d)\frac{\delta d}{\delta D}$. For our specific transformation we have $\frac{\delta d}{\delta D} = F/D^2$. Therefore, we obtain:

$$P(D) = \frac{F}{D^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(F/D-\mu)^2}{2\sigma^2}}. \tag{2.26}$$

.

Applying ML to $P(D)$ is equivalent to minimizing $-(F/D - \mu)^2(1/\sigma^2) - 2\log D$. By elementary calculus and algebra we obtain the equation $2D^{*2}\sigma^2/F = F - \mu D^*$ for the maximum likelihood estimator D^*. This shows that the naive estimate $D^* = F/\mu$ is, at best, approximately correct provided σ^2/F is small. The real estimate is given by the solutions $D = -\frac{\mu F}{4\sigma^2} \pm \frac{1}{4}\sqrt{\frac{\mu^2 F^2}{\sigma^4} + \frac{8F^2}{\sigma^2}}$.

This example illustrates the dangers of naively transforming between two different coordinate systems (the disparity and the depth from fixation). Such naive transformations are only correct in the unattainable limit of infinite precision. Otherwise, it is necessary to quantify the error processes to determine whether they are either small enough to be neglected or large enough to require precise modelling. In addition, this example also

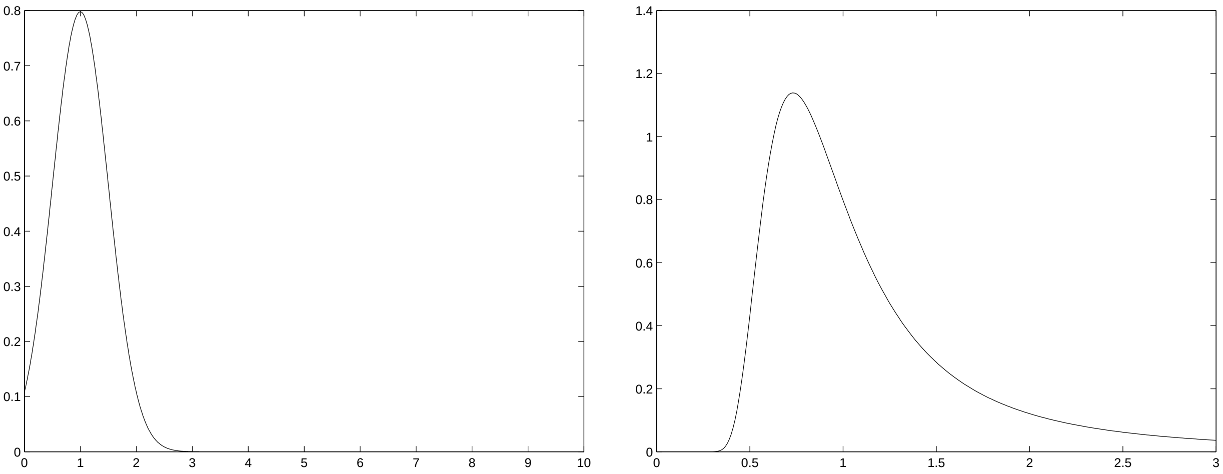


Figure 2.18 The distribution for disparity (left) and for depth (right). Observe that the peaks do not correspond.

helps illustrate the sensitivity of the maximum likelihood estimator. Other estimators with different loss functions would be more stable to coordinate transformations of this type.

Another, more dramatic example, is when we start with a Gaussian probability distribution $P(x) = 1/(\sqrt{2\pi\sigma})e^{-(x)^2/(2\sigma^2)}$ and then take the coordinate transformation $y = 1/x$. This transformation, see figure (2.19), changes the number of peaks in the distribution!

Our final example, see figure (2.20), illustrates the dangers of trying to avoid the use of priors. Suppose we try to estimate disparity by ML estimation. This is equivalent to MAP estimation but with a uniform prior on disparity. However, a uniform prior on disparity induces, by coordinate transform, a highly non-uniform prior on depth, see figure (2.20). So estimating disparity by ML (uniform prior on disparity) is equivalent to a MAP estimate of depth with a rather unusual prior on depth. In general, ML estimation in one coordinate system is not equivalent to ML estimation in another.

2.2.3 A Brief Comment on Discrete and Continuous Probability Distribution

Throughout this book we will deal with both continuous and discrete probability distributions defined over finite sets[9]. The same intuitions typically apply to both but there are some important differences we must point out.

Discrete probability distributions are specified by a finite set of numbers corresponding to the probabilities of the (finite) possible outcomes. By contrast, continuous probability

[9] Mathematicians have given probability theory a solid foundation as a branch of Analysis by introducing terms such as Borel sets and measure theory. Such advanced material is not needed in the applications we consider in this book. Moreover, it has been argued (Mumford) that such a formulation is unintuitive and that random variables should be the basic building blocks.

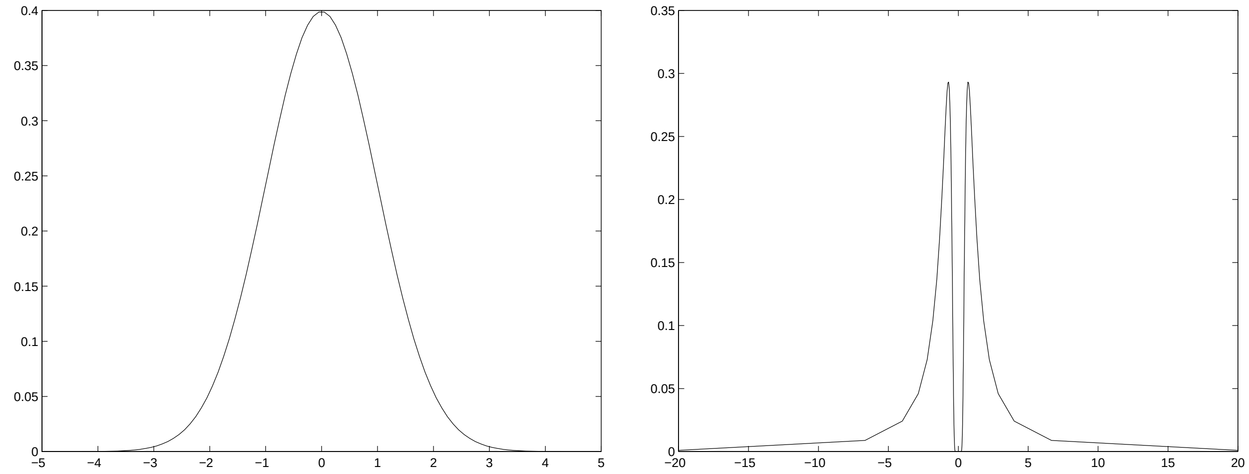


Figure 2.19 The distribution for disparity (left) and for depth (right) related by (depth)= 1/(disparity). Observe that the peaks do not correspond, so the most probable depth is *not* the inverse of the most probable disparity (if these distributions are used).

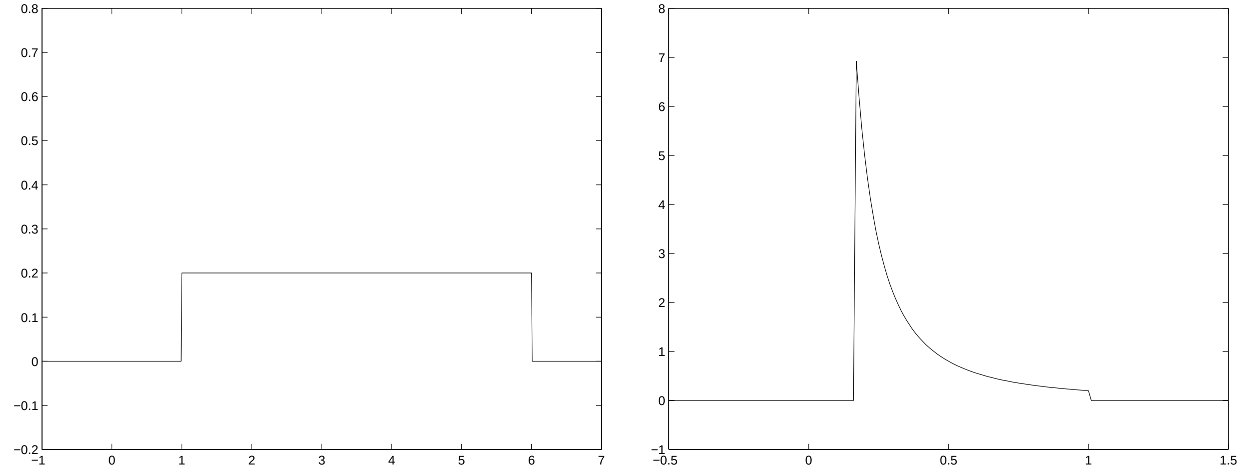


Figure 2.20 The uniform prior for a variable x (left) and the corresponding prior for $y = 1/x$ right. So assuming that all disparities are equally likely does not imply that all depths are equally likely.

distributions require specifying a probability density function $p(x)$ so that the probability of an outcome being in the infinitesimal interval $x, x+\delta x$ is given by $p(x)\delta x$.

A simple link between probability density functions and discrete probabilities is by letting the density function be a discrete sum of Dirac delta functions: $p(x) = \sum_{i=1}^{N} a_i \delta(x - x_i)$, where the $\{a_i : i = 1, ..., N\}$ are positive numbers such that $\sum_{i=1}^{N} a_i = 1$. This is equivalent to a discrete probability distribution with $P(X = x_i) = a_i, \; i = 1, ..., N$.

One natural way in which discrete distributions give rise to continuous density functions can be illustrated by a classical result by De Moivre who showed that a binomial distribution can be approximated by a Gaussian[10]. More precisely, suppose we have a sample x from a *binomial distribution* of form:

$$P(x|n,p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \tag{2.27}$$

where x is a discrete variable taking integer values in the range $0, ..., n$. The mean of the distribution is $\mu = np$ and the variance is $\sigma^2 = np(1-p)$.

De Moivre's result states that the binomial distribution can be well approximated by a Gaussian distribution for large n. Because the binomial is discrete and the Gaussian is continuous we make the approximation:

$$P(x|n,p) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \int_{x-1/2}^{x+1/2} e^{-(y-\mu)^2/2\sigma^2} dy. \tag{2.28}$$

This result is a special case of the Central Limit Theorem which will be discussed, and proved, in a later chapter!! This theorem states that *the sum of many independent random samples* is specified by a Gaussian distribution even though the individual samples are *not*. De Moivre's result follows from the observation that the binomial distribution arises by taking n independent samples from the *Bernoulli distribution* $P(x=1) = p$ over the binary variable $x = 0, 1$. Hence the Central Limit theorem can be applied.

Another example, is the approximation of the binomial distribution $P(x|n,p)$ by a *Poisson distribution* $P(x|\lambda) = e^{-\lambda}\lambda^x / x!$. This occurs in the limit as $n \mapsto \infty$ and $p \mapsto 0$ in such a way that $np = \lambda$ is finite. The result can be deduced using the approximation $\lim_{n \mapsto \infty} n!/x!(n-x)! = n^x/!x$ and $\lim_{n \mapsto \infty} \lim_{p \mapsto 0} (1-p)^{n-x} = e^{-\lambda}$. (This approximation follows from *Sterlings approximation* that $!n \approx n^n$ for large n.)

By comparing to De Moivre's result, we see that the Poisson distribution itself can be approximated by a Gaussian with mean $\mu = \lambda$ and variance $\sigma^2 = \lambda$ (the mean of a Poisson distribution always equals the variance). This is because both are limits of the binomial distribution $P(x|n,p)$ for large n. The Gaussian distribution has mean $\mu = np$ and variance $\sigma^2 = np(1-p)$. To obtain the Gaussian which approximates the Poisson we

[10] De Moivre was a French Huguenot who fled to England to escape religious persecution. Curiously he proved his result relating the Binomial distribution to the Gaussian long before Gauss was born.

must take the limit $n \mapsto \infty$ and $p \mapsto 0$ in such a way that $np = \lambda$ is finite. This gives $\mu = \lambda$ and $\sigma^2 = \lambda$ as $n \mapsto \infty$.

In practice, it is usually necessary, or convenient, to approximate continuous distributions by discrete ones. This is particularly important in information theory and so techniques, such as rate distortion theory [?], were invented to deal with it. Often, it is straightforward to *quantize* a continuous distribution to approximate it by a discrete one. But there are certain distributions for which this is difficult (we will warn the reader if any of these appear in the book).

The main differences between continuous and discrete distributions are the following: (i) probability density functions are sensitive to change of coordinate systems but discrete distributions are not, (ii) the entropies (which we have not defined yet!!) of discrete distributions are finite and positive but continuous distributions can have negative entropies. Moreover, certain mathematical results can only be proven for discrete distributions. For example, in a later chapter!! we will introduce the theory of types which is only derived for discrete distributions.

2.3 Multi-dimensional inputs: the Geometry of Decision Surfaces

So far in this chapter, the examples we have given have assumed that the input data is a scalar variable (i.e. a single number). Bayesian decision theory, however, applies without any changes when the input data are vectors. There is, however, a nice geometrical way of visualizing how decisions are made in this case. This is the geometry of decision surfaces.

If the state space is a set of discrete hypotheses $s \in S$ then the decision rules, $\hat{s} = \alpha(x)$, divide the observation space up into regions R_s so that all $x \in R_s$ are classified as being state s.

In the Bayesian framework that we propose these decision regions are determined by the underlying probability distributions and loss functions. In figure (2.21) we show an example of the decision boundaries in cases where the input variables are two dimensional. For example the chrominance.

In the next subsection, we analyze the special case where the likelihood functions are Gaussian distributions and demonstrate that the boundaries of these decision regions are specified by quadratic equations and are reduced to hyperplanes if the covariances of the Gaussians are identical.

There is, however, another approach. One can attempt to estimate the decision boundaries directly *without having any explicit underlying probability distributions*. Instead one assumes specific forms for the decision boundaries either by elementary analytic expressions, such as hyperplanes, or by parameterized functions such as multilevel perceptrons. It is then attempted to learn these decision boundaries from training data with, or without, a teacher. Much work in the statistical learning, neural networks, and statistics

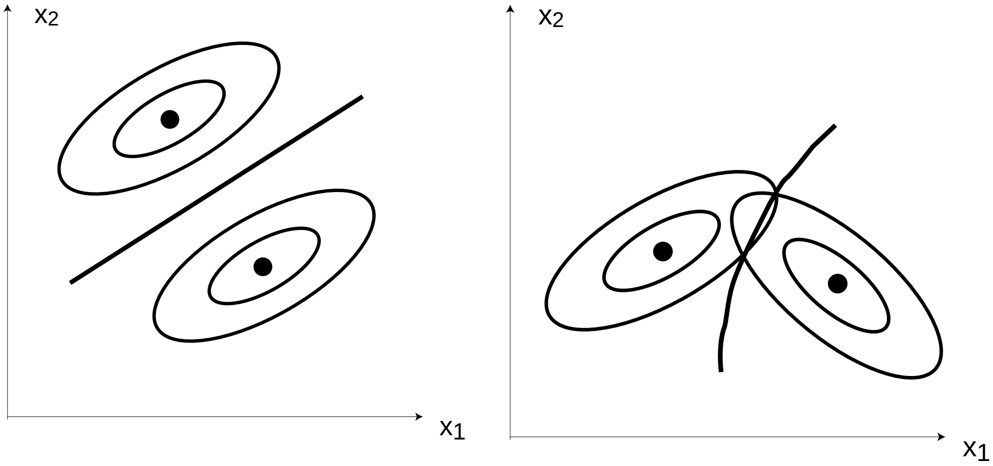


Figure 2.21 The decision regions for two-dimensional Gaussians. On the left, both Gaussians have the same covariances and the decision boundary is given by a straight line. If the covariances are not equal, see right, the decision boundary is a quadratic polynomial.

literature can be framed in this manner. We discuss this approach in a later chapter!!

2.3.1 The Gaussian Case

First recall the decision rule for surfaces when the distributions are univariate Gaussians with means μ_1, μ_2 and variances σ_1^2, σ_2^2. They are of form $P(x|s_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x-\mu_1)^2/2\sigma_1^2}$, $P(x|s_2) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(x-\mu_2)^2/2\sigma_2^2}$.

The log-likelihood ratio is simply:

$$\log\{\frac{P(x|s_1)}{P(x|s_2)}\} = \frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \log\frac{\sigma_2}{\sigma_1}, \tag{2.29}$$

which means a data point x is classified as "s_1" or "s_2" if it lies closest to μ_1 or μ_2 weighted by σ_1, σ_2.

These results can be generalized to multivariate distributions in n-dimensional space with means $\vec{\mu}_1, \vec{\mu}_2$ and covariances Σ_1, Σ_2. The distributions can be written as:

$$\begin{aligned} P(\vec{x}|s_1) &= \frac{1}{\sqrt{(2\pi)^n \det \Sigma_1}} e^{-(1/2)(\vec{x}-\vec{\mu}_1)^T \Sigma_1^{-1}(\vec{x}-\vec{\mu}_1)}, \\ P(\vec{x}|s_2) &= \frac{1}{\sqrt{(2\pi)^n \det \Sigma_2}} e^{-(1/2)(\vec{x}-\vec{\mu}_2)^T \Sigma_2^{-1}(\vec{x}-\vec{\mu}_2)}. \end{aligned} \tag{2.30}$$

The log-likelihood ratios can be computed. The decision boundary occurs at:

$$\frac{1}{2}(\vec{x}-\vec{\mu}_2)^T\Sigma_2^{-1}(\vec{x}-\vec{\mu}_2)-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T\Sigma_1^{-1}(\vec{x}-\vec{\mu}_1)-\frac{1}{2}\log\frac{\det\Sigma_1}{\det\Sigma_2}=0. \tag{2.31}$$

This has a nice geometrical interpretation. Let the data $\vec{x}$ lie in a parameter space D. Then the decision rule is given by a quadratic curve (a conic) separating the data $\vec{x}$ into two regions. If $\vec{x}$ lies in the region which contains $\vec{\mu}_1$, then it is classified as being generated by 1. Otherwise it is classified as being 2. See figure (2.21).

Again, an important special case is when the covariances are identical. The quadratic terms in x cancel and we get a linear rule:

$$\vec{x}^T\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_2) < (1/2)\{\vec{\mu}_1^T\Sigma^{-1}\vec{\mu}_1-\vec{\mu}_2^T\Sigma^{-1}\vec{\mu}_2.\}=0 \tag{2.32}$$

The decision criterion depends only on which side of the plane, see equation (2.32) and figure (2.21), the data falls on.

These results can be generalized to many classes. In parameter space, each class will be distinguished from each other class by which side of a decision boundary it lies on. In the special case of equal covariances, we get a parameter space which is divided by hyperplanes into the appropriate decision regions. For Gaussians with unequal covariances we get a parameter space divided into regions by quadratic surfaces.

2.4 An Information Theoretic Perspective

We can get another perspective on the estimation task by considering information theory. In this section, we introduce some of the basic concepts, such as entropy, which will be used throughout the book.

The *entropy* $H(P)$ of a discrete probability distribution $P(x)$ is given by:

$$H(P)=-\sum_{x=1}^{N}P(x)\log P(x). \tag{2.33}$$

The values of the entropy vary from $H=\log N$ for the distribution $P(x)=1/N,\ \forall\ x$ to $H=0$ if $P(x)=1\ if\ x=x_0$ and $P(x)=0$ otherwise (where x_o is an integer in range $1,...,N$).

The entropy is a measure of how much information is gained when we obtain a sample. If the entropy is zero then no information is gained because this means the distribution is deterministic and so we know what the sample will be beforehand. The maximum gain of information $\log N$ occurs in the opposite extreme where are samples are equally likely (i.e. $P(x)=1/N,\ \forall\ x$.)

Intuitively the entropy is a measure of the variability of the distribution. Thus it is analogous to the variance of a Gaussian distribution (but this analogy must not be pressed too far!).

For a decision task, the entropy of the posterior distribution $p(s|x)$ is a way of characterizing the difficulty of the estimation task. For example, suppose we have N possible states s. Then the ideal situation would be if $H(p(s|x))$ is zero, because then we would know that $p(s^*|x) = 1$ for one state s^* and the other states have zero probability – hence the decision s^* must be the correct choice. Conversely, that worst situation would be if $H(p(s|x)) = \log N$ because then all states s would be equally likely and our estimator would be picking the state effectively at random. In general, the lower the entropy of the conditional distribution then the better the estimate and the lower the error rate.

Another useful concept is the *Kullback-Leibler divergence* $D(P_A||P_B)$ between two distributions $P_A(x), P_B(x)$. It is defined by:

$$D(P_A||P_B) = \sum_x P_A(x) \log \frac{P_A(x)}{P_B(x)}, \quad D(P_B||P_A) = \sum_x P_B(x) \log \frac{P_B(x)}{P_A(x)}. \tag{2.34}$$

Observe that, in general, $D(P_A||P_B) \neq D(P_B||P_A)$ and so Kullback-Leibler is not symmetric as a function of the two probability distributions. However, it is straightforward to prove that $D(P_A||P_B) \geq 0$ with equality only if $P_A(x) = P_B(x) \ \forall \ x$. (this proof uses *Jensen's inequality*: if $f(\bullet)$ is a convex function and X is a random variable with distribution $P(x)$ then $\sum_x P(x)f(x) \geq f(\sum_x P(x)x)$.)

The Kullback-Leibler divergence has a natural role in the log-likelihood ratio test. In fact, we see that the expected value of the test $\log P_A(x)/P_B(x)$ when the data x is generated by $P_A(x)$ is exactly $D(P_A||P_B)$. Conversely, the expected value of the test when the data is generated by $P_B(x)$ is $-D(P_B||P_A)$. Hence, we can measure the effectiveness of the test on average by the expected *difference between the expected values of the test for the two different situation* by $D(P_A||P_B) + D(P_B||P_A)$.

An alternative way of thinking of the entropy of $P(x)$ is in terms of the Kullback-Leibler divergence between $P(x)$ and the uniform distribution $U(x) = 1/N, \ \forall \ x$. It is clear that

$$D(P||U) = \sum_x P(x) \log \frac{P(x)}{U(x)} = \sum_x P(x) \log P(x) - \sum_x P(x) \log U(x) = -H(P) + \log N. \tag{2.35}$$

The definitions of entropy and Kullback-Leibler can be automatically adapted to continuous probability distributions. One difference is that the entropy of a distribution can now be negative. However, the Kullback-Leibler divergences stay positive as before. What does a negative entropy mean? Recall that for a uniform distribution defined over N states the entropy is $H(U) = \log N$. In other words the number of allowed states is $2^{H(U)}$. For a continuous distribution $P(x)$, we can interpret $2^{H(P)}$ as the volume of state space. It therefore does not matter if the entropy is negative (it simply means a small volume).

It is a useful exercise to compute the Kullback-Leibler divergence between two univariate Gaussian distributions with the identical variances. Show that Kullback-Leibler is proportional to the signal to noise ratio $(\mu_1 - \mu_2)/\sigma$.

Another very important concept in information theory is the *mutual information* between two distributions. Consider the task of estimating the state s from observations x. The mutual information $I(S, X)$ is given by:

$$I(S, X) = \sum_{s,x} P(s, x) \log \frac{P(s, x)}{P(s)P(x)}. \tag{2.36}$$

The greater the mutual information $I(S, X)$ then knowledge of an observation x is more successful at predicting the correct state s. The worst case, with lowest mutual information, is when $P(s, x) = P(s)P(x)$ and then $I(S, X) = 0$. In this case knowledge of the observation x gives no help at all in estimating the state s.

2.5 ROC curve and two-alternative forced choice.

We can ask how well does an observer do in signal detection. We describe basic properties of signal detection theory from our Bayesian decision theory perspective. In particular, we describe the classic forced choice experiment and the resulting ROC (receiver operating characteristic) curve. We relate it to the two alternative forced choice experiments. In both cases, following psychophysics tradition, we pay special attention to the cases of Gaussian distributions with identical variances (univariate distributions of course).

The forced choice, or yes/no, experiment is as follows. There is a signal source S and a noise source N. The observer's task is to determine whether a specific input is signal or noise.

It is assumed that each source generates a response x on which the observer has to make his/her decision. There are distributions $P(x|S)$ and $P(x|N)$ which determines this response. Why introduce probability distributions here? There are at least three possible forms of uncertainty which may require probabilistic modelling: (I) the output from the signal and noise sources may be uncertain (e.g. a light source emits a quantum flux of photons not a deterministic signal), (II) the observer (human or camera) will induce uncertainty when the light is detected (e.g. the photoreceptors in humans are governed by Poisson's law), and (III) the signal must be transmitted from the measurement device to a place where a decision is made (e.g. information must be transferred from the retina to the cortex). Uncertainty can be introduced at all three stages.

Signal detection theory assumes that any observer's performance depends on the two distributions $P(x|S)$ and $P(x|N)$. Different observers may have different distributions (e.g. some observers may be better at this task). But we know, from decision theory, that the best decision is based on the log-likelihood ratio test which involves comparing the value of $\log P(x|S)/P(x|N)$ to a decision threshold T. This threshold T is determined by the loss function and the prior probability of the input being signal or noise.

The problem is that the loss function of a specific observer is unknown. A conservative observer may have a loss function which penalizes false positives. Alternatively, a nervous observer may not like having false negatives. Therefore the choice of threshold may be influenced by subjective aspects of the observer.

Signal detection theory gives a way round this difficultly. The suggestion is to plot the ROC curve of the proportion of correct responses to the proportion of false positives. The point is *that each point of the curve corresponds to a particular choice of threshold*. The experimenter can manipulate the observer, by offering rewards or changing the prior, or that the observer's decision threshold changes. The resulting ROC curve *therefore represents information which is independent of the observer's loss function and prior*. It therefore depends only on the probability distributions $P(x|S)$ and $P(x|N)$. If two observers have the same distributions then their ROC curves will be identical (i.e. independent of their loss functions or priors).

More precisely, for a given threshold T, we can write the proportion of correct responses $C_+(T)$ and false positives $F_+(T)$ as:

$$\begin{aligned} C_+(T) &= \int_{\{x:\log P(x|S)/P(x|N)>T\}} P(x|S)dx, \\ F_+(T) &= \int_{\{x:\log P(x|S)/P(x|N)>T\}} P(x|N)dx, \end{aligned} \tag{2.37}$$

As we vary the threshold T (by the experimenter manipulating the experiment) then we obtain the ROC curve. The gradient of the curve is dC_+/dF_+. If we assume that $\log P(x|S)/P(x|N)$ is a monotonic function of x (we will verify later than this assumption is true for Gaussian distributions) then it follows that

$$\frac{dC_+}{dF_+}(T) = \frac{P(x(T)|S)}{P(x(T)|N)}, \tag{2.38}$$

where $x(T)$ is defined by the equation $\log P(x(T)|S)/P(x(T)|N) = T$. Therefore the gradient of the ROC curve is related to the log-likelihood ratio.

Another standard experimental technique is two-alternative forced choice. This occurs when two signals x_1, x_2 are presented, either simultaneously or in sequence, and the observer is told that one is signal and the other is noise. The observer's task is to determine which is signal. (We assume that the order in which the signal and noise are presented are equally likely).

This experimental design is simple because the best decision procedure is independent of the observer's loss function (and priors). The best strategy is to decide that the signal S is x_1 if $\log P(x_1|S)/P(x_1|N) > \log P(x_2|S)/P(x_2|N)$ and otherwise the signal is x_2.

Note that this involves comparing the likelihood ratios only (the loss function and prior drop out of the Bayesian a posteriori estimate).

There is a simple relationship between the proportion of correct results for the two-alternative forced choice and the ROC curve for the forced choice experiment. We will derive this result with the assumption that the log-likelihood ratio $\log P(y|S)/P(y|N)$ is a monotonic function of y (for example, if the distributions are Gaussians with the same variance).. Hence, the equation $\log P(y|S)/P(y|N) = T$ defines a unique invertible function $y = y(T)$.

The error rate for two-alternative forced choice is given by

$$\int\int_{\{(x_1,x_2):\log P(x_1|S)/P(x_1|N)>\log P(x_2|S)/P(x_2|N)\}} P(x_1|N)P(x_2|S)dx_1dx_2. \tag{2.39}$$

This can be expressed as

$$\int_{-\infty}^{\infty} dx_1 P(x_1|N)\int_{-\infty}^{x_1} P(x_2|S)dx_2. \tag{2.40}$$

Because of the monotonicity we can write:

$$F_+(T) = \int_{y(T)}^{\infty} P(x|N)dx \quad C_+(T) = \int_{y(T)}^{\infty} P(x|S)dx. \tag{2.41}$$

Now the area under the ROC curve is given by

$$\begin{aligned}\int_{\infty}^{-\infty} C_+(T)\frac{dF_+}{dT}dT &= -\int_{-\infty}^{\infty} dT\int_{y(T)}^{\infty} P(x|S)dx\{-P(y(T)|N)\}\frac{dy}{dT}\\ = \int_{-\infty}^{\infty} dyP(y|N)\int_{y}^{\infty} P(x|S)dx &= 1-\int_{-\infty}^{\infty} dyP(y|N)\int_{-\infty}^{y} dxP(x|S),\end{aligned} \tag{2.42}$$

which is the one minus the error rate, and thus the area under the ROC curve is the frequency of correct responses. (Where we have used $\int_y^\infty dxP(x|S) = 1 - \int_{-\infty}^y dxP(x|S)$.) Note this result only assumes monotonicity of the log-likelihood ratios and so does not require Gaussian distributions. (Exercise for the reader: is this result true if we don't assume monotonicity of the loss functions?)

2.5.1 Gaussian case

We now treat the historically very important case that both distributions $P(x|S)$ and $P(x|N)$ are Gaussians with identical variances. This case leads to very simple expression for the form of the ROC curve and to a "natural" *measure of sensitivity called d'* (note that this is the signal-to-noise ratio which arose earlier in our bright/dim example!!). In our view, the best argument for the Gaussian distribution is the pragmatic – it is the easiest thing to try and if it fits the data one should use it. We are, however, sceptical of arguments for the "naturalness" of the Gaussian based on the Central Limit Theorem (see later chapter for a statement and proof of this theorem). We stress that the choice of distribution should, if possible, be determined empirically.

For Gaussian distributions we have:

$$\begin{aligned} P(x|S) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_S)^2/2\sigma^2}, \\ P(x|N) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_N)^2/2\sigma^2}. \end{aligned} \tag{2.43}$$

The success rate and the false positive rate are given (as a function of threshold) by:

$$\begin{aligned} F_+(T) &= \int_T^\infty dx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu_N)^2/2\sigma^2}, \\ C_+(T) &= \int_T^\infty dx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu_S)^2/2\sigma^2}. \end{aligned} \tag{2.44}$$

We now define the error function by:

$$Z(y) = \int_y^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy. \tag{2.45}$$

We can therefore write $F_+(T) = Z\{(T-\mu_N)/\sigma\}$ and $C_+(T) = Z\{(T-\mu_S)/\sigma\}$. By inverting the error function, we have:

$$\frac{T-\mu_N}{\sigma} = Z^{-1}\{F_+(T)\}, \quad \frac{T-\mu_S}{\sigma} = Z^{-1}\{C_+(T)\} \tag{2.46}$$

so we can write:

$$\frac{\mu_S-\mu_N}{\sigma} = Z^{-1}\{F_+(T)\} - Z^{-1}\{C_+(T)\}, \tag{2.47}$$

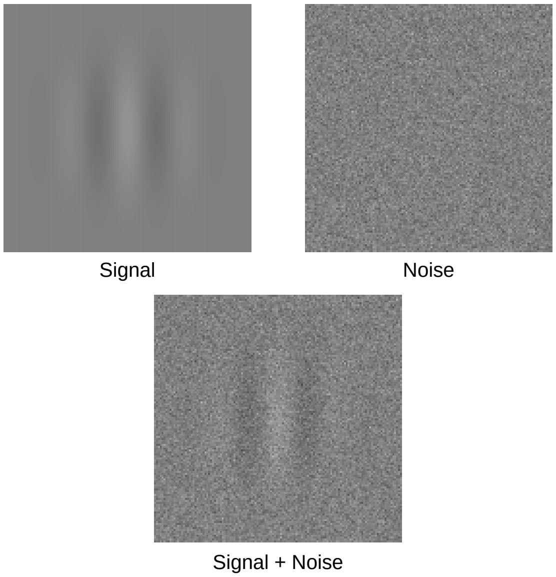
and so, if the equal variance Gaussian assumption is correct, we see that the ROC curve will become a straight line at forty five degrees in this transformed space (i.e. the space defined by $Z^{-1}(.)$). The position of the curve in this space is determined by the parameter $d' = \frac{\mu_S-\mu_N}{\sigma}$.

Observe, we can also extract a quantity that depends on the *observer's bias* by summing $Z^{-1}(C_+(T)) + Z^{-1}(F_+(T)) = (2T-\mu_N-\mu_S)/\sigma$. The optimal *unbiased* strategy for this task would be to set the threshold to be $T^* = (\mu_N+\mu_S)/2$, so we can re-express the bias as $Z^{-1}(C_+(T^*)) + Z^{-1}(F_+(T^*)) = 2(T-T^*)/\sigma$, in other words, it is the difference between the threshold T chosen by the observer and the unbiased threshold T^* multiplied by 2 and normalized by the standard deviation σ.

From the perspective of decision theory, we see that the bias depends on the observer's choice of threshold and therefore gives a measure of the prior probabilities assumed and the loss function used. (Details are left as an exercise for the reader).

2.6 Signal Known Exactly Gaussian Model

A much studied model is the *Signal Known Exactly* model. In this case, the task is to detect whether a signal is present or not. The input signal is a spatial pattern which we can characterize by a set of values $\{I_i\}$ (where i labels the spatial position). The signal is specified by values a set $\{S_i\}$ and we assume that the measurement process is corrupted by additive zero mean Gaussian noise with variance σ^2. There is also a background intensity level B at each pixel (see figure (2.22)).



Signal

Noise

Signal + Noise

Figure 2.22 Signal plus noise. The upper left and right panels show the signal and a sample of independently sampled Gaussian noise, respectively. The bottom panel shows the observation: "signal + noise". Human observers can detect this pattern in noise (over a large range of sizes) with an efficiency of 10% or more.

The image generated by the signal is specified probabilistically by:

$$\begin{aligned} P(\{I_i\}|S) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(I_i - S_i - B)^2/2\sigma^2}, \\ P(\{I_i\}|N) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(I_i - B)^2/2\sigma^2}. \end{aligned} \tag{2.48}$$

The values I_i/B are called the contrast of the intensity at this pixel point, i.e. the amount by which the intensity differs from the background. This gives the contrast as a specific pixel position i. There are several definitions of contrast for the entire signal (i.e. taking into account the values at all the pixels). For example, for sinusoid gratings with $I(x) = A\cos\omega x + B$, a historically important class of signals, the contrast is defined by dividing the amplitude A of the grating divided by B to get A/B.

To discriminate between the signal and noise case we compute the log-likelihood ratio. This gives:

$$\log \frac{P(\{I_i\}|\{S_i\})}{P(\{I_i\}|N)} = -\frac{1}{\sigma^2}\{\sum_i s_i^2 + 2B\sum_i s_i - 2\sum_i s_i I_i\}. \qquad (2.49)$$

We see that the log-likelihood test only depends on the data by the value of the scalar variable $y = \vec{s}\cdot\vec{I} = \sum_{i=1}^N s_i I_i$. Therefore we *need only analyze* how this scalar is estimated (despite the fact that the original data is a vector).

We can therefore calculate the induced distributions on this variable $\vec{s}\cdot\vec{I}$ assuming that the it is generated either by the signal or by the noise. It is clear that both distributions $P(\vec{s}\cdot\vec{I}|S)$ and $P(\vec{s}\cdot\vec{I}|N)$ are Gaussian distributions. The mean of the first is $\vec{s}\cdot(\vec{s}+\vec{B})$ with variance $\sigma^2\vec{s}\cdot\vec{s}$ and the second has mean $\vec{s}\cdot\vec{I}$ and the same variance. We then calculate

$$d'^2 = \frac{\vec{s}\cdot\vec{s}}{\sigma^2}. \qquad (2.50)$$

This allows us to compare the experimental d'_h found for human observers, calculated from their ROC curves, with the ideal d' computed from the theory.

Even under the best of conditions, human sensitivity to contrast patterns added to white (i.i.d. pixel intensities) Gaussian noise is generally less than ideal, and can be described by:

$$d'^2_h = k\frac{\vec{s}\cdot\vec{s}}{\sigma^2 + {\sigma_{int}}^2}$$

where d'_h is calculated from the hit and false positive rates in a yes/no experiment, or from the proportion correct in a two-alternative forced-choice experiment. Particularly, in cases near threshold where S_i/B is small, human observers behave as if there is internal noise, which can be modeled as an "equivalent" added external noise variance σ^2_{int}. This internal noise might partially be due to photoreceptor processes but may also reflect higher up processing of the data. One argument in favour of this hypothesis is that it predicts that humans should have closer performance to the ideal when the external noise σ^2 is so high that it dwarfs the equivalent noise σ^2_{int}. In fact, when $\sigma^2 \gg {\sigma_{int}}^2$, human sensitivity can approach that of the ideal, but generally it still falls short by a fraction k,

$$d'^2_h \approx k\frac{\vec{s}\cdot\vec{s}}{\sigma^2} = kd'^2,\ k \leq 1.$$

One benchmark for measuring how observers fail to perform compared to the ideal is *statistical efficiency.* This idea, originally due to Fisher, asks how many additional samples are needed to obtain a reliable estimate compared to the number given by the ideal.

For example, when estimating the mean of a set of samples from a Gaussian distribution it is known that the variance of the optimal estimator falls off as $1/N$. If the variance of the estimator falls off as $1/2N$ then it has an efficiency of $1/2$. For discriminating between two Gaussian distributions it can also be shown that the difficulty, and d' scales by $1/N$. This is explained more in a later chapter. It can be shown that efficiency is equal to ${d'}_h^2/{d'}^2$. Contrast detection under ordinary daylight conditions is very inefficient (e.g. efficiency is less than a tenth of a percent [3, 4]) where the dominant external noise, due to photon fluctuations, has a negligible contribution to human performance. However, if artificial noise is added then as $\sigma^2 \gg {\sigma_{int}}^2$, efficiency ($\approx k$) can be quite high. The exact value of k depends on the type of pattern, and has been reported between 30% and 70% for contrast discrimination of wavelet-like patterns [1, 2]. (See figure (2.22)).

2.6.1 Predictions of the model for edge detection

What would the Signal Known Exactly Gaussian model predict for edge detection? Suppose we model a step edge by a step function $\{S_i\}$ with additive Gaussian noise. We can compute the probability distributions for an edge detection filter both at and away from an edge.

Away from an edge, we calculate that:

$$P(I_{i+1} - I_i = y) = \int dy dx_i dx_{i+1} \delta(y - x_{i+1} + x_i) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_{i+1}-B)^2/2\sigma^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-B)^2/2\sigma^2}$$
$$= \frac{1}{\sqrt{4\pi\sigma^2}} e^{-y^2/4\sigma^2}. \qquad (2.51)$$

This predicts $P_{off}(y)$ to be a Gaussian distribution with zero mean and variance $2\sigma^2$.

Similarly, at an edge, we would obtain $P(I_{i+1} - I_i = y)$ to be a Gaussian distribution with mean the size h of the edge (i.e. the intensity change across the edge) and variance again equal to $2\sigma^2$. This would give $P_{on}(y|h)$ to be a Gaussian distribution also. This model is limited because it assumes that edges all have the same size h. This assumption could be relaxed by assuming that there is a probability distribution $P(h)$ on the size of edges.

Such edge models have been studied in the computer vision literature and "ideal edge detector filters" based on them have been designed (Canny). Such filters have been used as the basis for edge detectors to detect object boundaries from real images. (It should be emphasized that such edge detectors have additional ingredients such as local assumptions about edge smoothness).

By contrast, the empirical studies of P_{on}, P_{off} evaluated on real image datasets, with the object boundaries extracted by hand, give rather different distributions as described earlier, see figure (2.5). In our opinion, the distribution of intensities changes both at object boundaries (and away from them) are properties of the world and must be determined by the statistics of real world images. Idealizations, such as step edge models and the use of Gaussian distributions to model noise, are assumptions that should be checked against reality.

2.6.2 ROC curves on the log-likelihood ratios

In this subsection we address an important issue: how much information does the ROC curve supply about the underlying probability distributions $P(x|S)$ and $P(x|N)$? To put the question another way: which probability distributions generate identical ROC curves? For example, there is a "standard" ROC curve shape that results if the distributions $P(x|S)$ and $P(x|N)$ are univariate Gaussians with equal variances. In this case the precise shape of the ROC curve depends on the parameter d'. But how many non-Gaussian distributions $P(x|S)$ and $P(x|N)$ give rise to ROC curves of this shape?

To understand this issue we observe that the ROC curve is based on *decisions* made by the theory, or by the observer. These decisions are, in turn, based on the the log-likelihood ratios $\log\{P(x|S)/P(x|N)\}$. This suggests that the ROC curve can only yield information about $P(x|S)$ and $P(x|N)$ which is available in the log-likelihood ratio. We must therefore study the probability distributions of the log-likelihood ratio.

More precisely, we define two *induced* probability distributions, $\hat{P}(r|S)$ and $\hat{P}(r|N)$, on the log-likelihood ratio r which depend on whether the data is generated by $P(x|S)$ or $P(x|N)$. Formally:

$$\hat{P}(r|S) = \int dx \delta(r - \log \frac{P(x|S)}{P(x|N)}) P(x|S),$$

$$\hat{P}(r|N) = \int dx \delta(r - \log \frac{P(x|S)}{P(x|N)}) P(x|N). \tag{2.52}$$

Dividing these equations by each other and using the delta function identity $\delta(r - \log \frac{P(x|S)}{P(x|N)}) P(x|S) = \delta(r - \log \frac{P(x|S)}{P(x|N)}) P(x|N) e^r$ (since $P(x|S) = P(x|N) e^{\log\{P(x|S)/P(x|N)\}}$), we obtain that:

$$\frac{\hat{P}(r|S)}{\hat{P}(r|N)} = e^r. \tag{2.53}$$

We can now answer the question posed at the start of this subsection. The answer is that *the shape of the ROC curve is determined only by the induced distributions $\hat{P}(r|S)$ and $\hat{P}(r|N)$ on the log-likelihood ratio r. Moreover, the ROC curve will determine these distributions uniquely.*

To prove these claims, we recall that the ROC curve is parameterized by the threshold T and plots the number of false positive against the number of correct positives. At threshold T the number of false positives is $\int_T^\infty \hat{P}(r|N)dr$ and the number of correct positives is $\int_T^\infty \hat{P}(r|S)dr$. So if we know how the points on the ROC curve correspond to the threshold T then we can read off the *x probability distributions* of $\hat{P}(r|S)$ and $\hat{P}(r|N)$ as the horizontal and vertical coordinates of the curve, see figure (2.23). But the gradient of the ROC curve at threshold T is just given by $\frac{\hat{P}(r=T|S)}{\hat{P}(r=T|N)}$ which is e^T. Thus by measuring the gradient at any point on the ROC curve we can determine the threshold T which it corresponds to. Therefore we can determine the cumulative probability distributions of $\hat{P}(r|S)$ and $\hat{P}(r|N)$, and hence the distributions themselves. Conversely, it is clear that knowing these distributions determines the ROC curve uniquely.

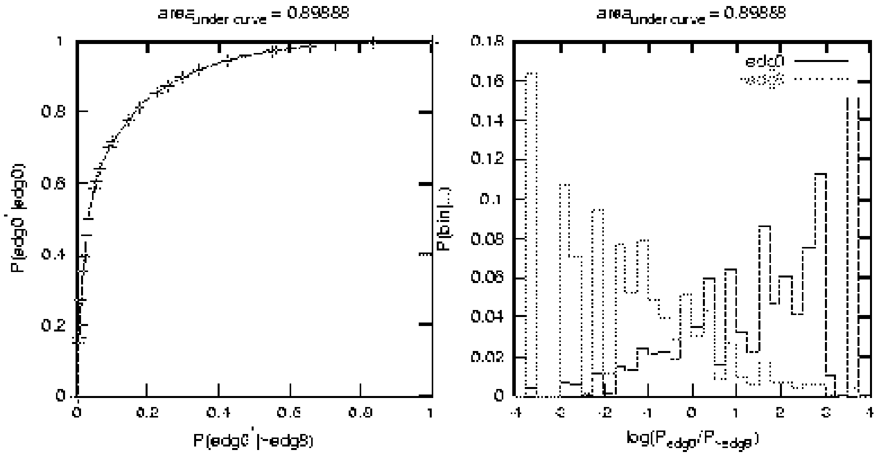


Figure 2.23 The ROC curve determines a unique $\hat{P}(r|S)$ and $\hat{P}(r|N)$. This is evaluated at the point on the curve with gradient e^r.

This result is important. In particular, it says that there are many probability distributions $P(x|S)$ and $P(x|N)$ which generate identical ROC curves. For example, suppose we have two univariate Gaussian distributions $\hat{P}(r|S)$ and $\hat{P}(r|N)$ which have equal variance. Now consider the distributions $P(x|S)$ and $P(x|N)$ obtained by the coordinate transformation $x = 1/r$. These two distributions will appear highly non-Gaussian (in fact, they are bimodal) but nevertheless the ROC curves associated with them as the same as if the data had been generated by Gaussians with equal variances. See figure (2.24).

We stress that the same result will apply even if the data is multi-dimensional. So, for example, we could have $P(\vec{y}|S)$ and $P(\vec{y}|N)$ where $\vec{y}$ is an M-dimensional filter measurement (such as a boundary detector working at multiple scales and/or coupling different segmentation cues). Then the ROC curve still depends only on the induced distributions on the log-likelihood ratio. It is very possible that the ROC curve may appear similar to that obtained in the data came from univariate Gaussians with equal variance even

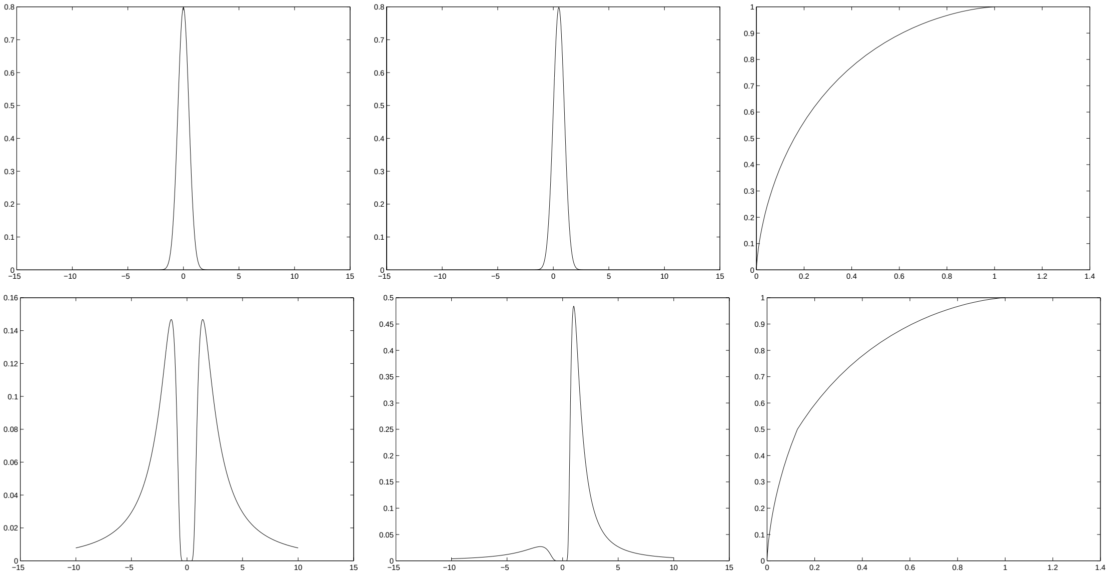


Figure 2.24 The ROC curve between two univariate Gaussians (top left and top centre) takes a standard form (top right). But we get exactly the same ROC curve for the two distributions obtained by the coordinate transformation $x = 1/r$ (bottom left and bottom centre). Thus many, apparently quite different distributions can have the same ROC curve.

though the underlying distributions are multi-dimensional and highly non-Gaussian.

From a more abstract perspective, consider a coordinate mapping $z = f(x)$ of the distributions $P(x|S)$ and $P(x|N)$ with the restriction that $f(x_i) \neq f(x_j)$ for any two points x_i, x_j with *different* loglikelihood ratios (i.e $\log\{P(x_i|S)/P(x_i|N)\} \neq \log\{P(x_i|S)/P(x_i|N)\}$). In other words, the restriction means that points in x-space with different log-likelihood ratios cannot be mapped to the same point in z-space. Such a mapping *will not affect the ability to discriminate between samples from S and N*. The ROC curves for $\hat{P}(z|S)$ and $\hat{P}(z|N)$ will be identical to those for $P(x|S)$ and $P(x|N)$. Points which can be discriminated in x-space by the log-likelihood ratio can be discriminated in z-space (exercise for the reader).

In summary, the ROC curve tells us everything about the induced probability distributions on the log-likelihood ratios. But it tells us little about the form of the underlying distributions $P(x|S)$ and $P(x|N)$.

2.7 Fisher Information and the Cramer-Rao lower bound

How well can we estimate a continuous variable? One measure is by determining the variance of the estimator. For example, suppose we are estimating the angle of a line in space. Our input are noisy measurements. Our estimate of the angle is therefore a random variable whose mean and variance are respectively measures of the bias and the accuracy

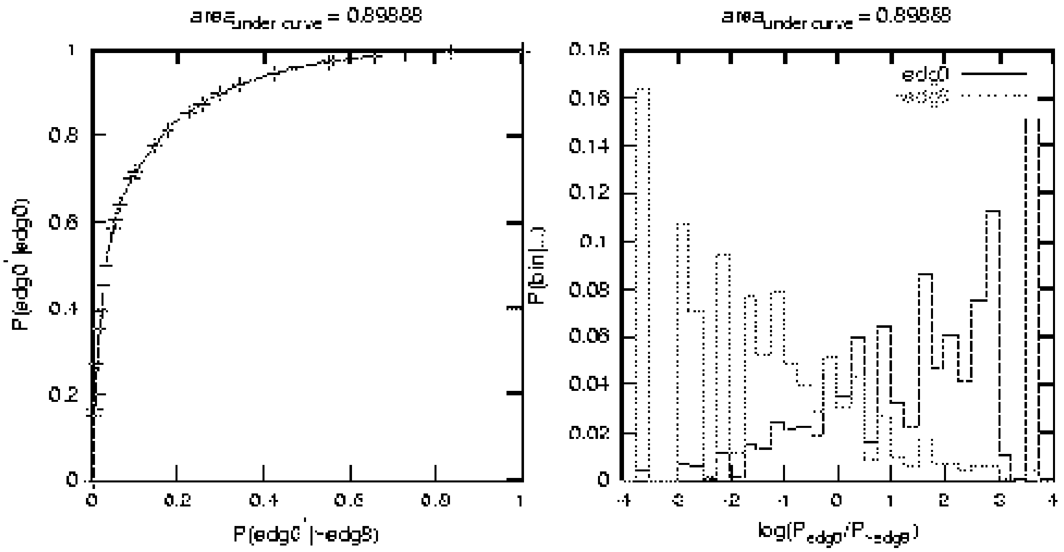


Figure 2.25 The ROC curve (left) for boundary discrimation using edge filters at multiple scales. The induced probability distributions on the log-likelihood ratios (right) are roughly of univariate Gaussian form. So it is not surprising that the ROC curve is simialr to those determined by univariate Gaussians with equal variance even though the distributions on the edge filter data are three-dimensional and non-Gaussian.

of the estimator. In experimental situations, it is possible (though time consuming) to determine such means and estimates.

If the probability distributions are known and a choice of estimator is specified then it is, in principle, possible to derive a probability distribution on the estimates and from this calculate the means and variances. But these results are determined by the choice of estimator.

A classic result of statistics, *the Cramer-Rao lower bound*, shows that the variance of *any* unbiased estimator is bounded below by the Fisher information. Moreover, when this lower bound is achieved the estimator must be the ML estimator.

Below we will state these results precisely in mathematics and give a proof of the Cramer-Rao lower bound. We should emphasize that *this result does not prove that ML is the best estimation procedure.* The criterion used attempts to minimize the variance for *each state s*. By contrast, criteria derived from decision theory take into account the prior distribution over states so that not all errors are weighted equally. Moreover, the Cramer-Rao result only applies to unbiased estimators and, as we will show later by example, there can be biased estimators which have lower variances. The concepts of biased, and unbiased, estimators were developed by Fisher[11].

[11] Fisher was one of the leading statisticians of the twentieth century. He believed that Bayesian methods,

More precisely, an estimator ϕ is a function from the space of observations X to the space of states S. Thus for any $x \in X$, we have $\phi(x) \in S$. To determine the *bias* of an estimator we evaluate $\int dx \phi(x) p(x|s)$. An estimator $\phi(.)$ is said to be *unbiased* if:

$$\int dx \phi(x) p(x|s) = s, \ \forall \ s \in S. \tag{2.54}$$

Observe that this definition of bias depends only on the likelihood function $p(x|s)$. It therefore ignores any prior knowledge about the state variable s.

An unbiased estimator, by definition, makes the correct estimate on average. But how much does this estimate vary? One measure is the variance of the estimate as a function of s,

$$Var(\phi) = \int dx p(x|s) \{\phi(x) - s\}^2. \tag{2.55}$$

The Fisher information $J(s)$ is a measure of how much information about the state s that is present in the data. It is given by:

$$J(s) = \int dx p(x|s) \{\frac{\partial}{\partial s} log p(x|s)\}^2, \tag{2.56}$$

and can also be expressed as:

$$J(s) = -\int p(x|s) \frac{\partial^2}{\partial s^2} \log p(x|s) dx. \tag{2.57}$$

The proof of the equivalence of these expressions is left as an exercise for the reader. (Hint: make use of the identity $\partial^2 / \partial s^2 \int p(x|s) dx = 0$.

We will now state and prove the Cramer-Rao lower bound.

Theorem: the Cramer-Rao lower bound. *The variance $Var(\phi)$ of an unbiased estimator $\phi(.)$ is always bounded below by the inverse of a function, the Fisher information $J(s)$, of the distribution $p(x|s)$. Moreover, when this lower bound is attained the estimator must be the ML estimator.*

Proof. *Why does Fisher information have anything to do with $Var(\phi)$? The connection arises from differentiating the expression for an unbiased estimator with respect to s. This can be expressed as:*

$$\int dx \phi(x) \{\frac{\partial}{\partial s} \log p(x|s)\} p(x|s) = 1, \tag{2.58}$$

which can be interpreted geometrically as saying that the dot product, with respect to $p(x|s)$, of $\phi(x)$ with $\frac{\partial}{\partial s} \log p(x|s)$ is equal to 1.

which he called “inverse probabilities”, were flawed and argued that Bayes did not publish his papers while he was alive because he was aware of their weaknesses.

Now we apply the Cauchy-Schwartz inequality which states:

$$\{\int dx p(x|s)\{\phi(x)-s\}^2\}\{\int dx p(x|s)\{\frac{\partial}{\partial s}\log p(x|s)\}^2\} \geq \{\int dx p(x|s)\{\phi(x)-s\}\frac{\partial}{\partial s}\log p(x|s)\}\}^2. \tag{2.59}$$

Again this equation can be interpreted geometrically. It says that the dot product, with respect to $p(x|s)$*, of* $\{\phi(x)-s\}$ *with* $\{\frac{\partial}{\partial s}\log p(x|s)\}$ *must have smaller, or equal, magnitude to the product of the magnitudes of both vectors. We can simplify the right hand side by the observation that:*

$$\int dx p(x|s) s \frac{\partial}{\partial s}\log p(x|s) = s\int dx \frac{\partial}{\partial s} p(x|s) = 0. \tag{2.60}$$

Applying equation ?? the right hand side simplifies to be one. This implies that $Var(\phi) \geq 1/J(s)$ *as required.*

When is the lower bound attained? By Cauchy-Schwartz, we know that this happens when:

$$\phi(x) - s = \lambda(s)\frac{\partial}{\partial s}\log p(x|s), \tag{2.61}$$

for some scale function $\lambda(s)$*. The geometric intuition is that Cauchy-Schwartz becomes an equality only if the two vectors, of which you take the dot product, are parallel and so can be related by a scale factor. In our case, we apply Cauchy-Schwartz for each value of* s *and so the scale factor is a function of* s*.*

Equation (2.61) implies that $\phi(x) = s$ *if, and only if,* $\frac{\partial}{\partial s}\log p(x|s) = 0$*. But this is the exact specification of an ML estimator – recall that the ML estimator specifies that* $\phi(x) = s^*$ *where* $s^* = \arg\max_s P(x|s)$ *and hence satisfies* $\frac{\partial}{\partial s}\log P(x|s) = 0$ *– and so, any unbiased estimator which attains the lower bound must be an ML estimator.*

We add that not all ML estimators attain the lower-bound and not all are unbiased. For example, it is easy to check that the ML estimator for $p(x|s) = se^{-sx}, \;\; x \geq 0$ is biased. On the other hand, it is straightforward to verify that ML estimators for Gaussians are unbiased and attain the lower bound.

Moreover, biased estimators may have lower variance than unbiased estimators. For example, consider estimating the mean and variance of n samples $x_1, ..., x_n$ from a Gaussian distribution. Let the estimator for the mean be $\phi_{\mu,n} = (1/n)\sum_{i=1}^n x_i$. It is an exercise to check that it is unbiased. Now consider two possible estimators for the variance $\phi_{\sigma^2,n} = (1/n)\sum_{i=1}^n (x_i - \phi_{\mu,n})^2$ and $\phi_{\sigma^2,n-1} = (1/N)\sum_{i=1}^n (x_i - \phi_{\mu,n})^2$. Show that $\phi_{\sigma^2,n}$ is biased and $\phi_{\sigma^2,n-1}$ is unbiased. Then show that the variance of $\phi_{\sigma^2,n}$ is lower than the variance of $\phi_{\sigma^2,n-1}$.

There is a close relationship between these results and the class of *exponential distributions*. Exponential distributions, which include the Gaussian, Poisson, and Binomial as special cases, will be introduced in a later chapter. It turns out, that *if an estimator is unbiased and the Cramer-Rao lower bound is attained then the distribution lies in the exponential class.* Conversely, there is a subclass of exponential distributions of form $P(x|s) = e^{\phi(x)a(s)+b(s)}$ so that the estimator $\phi(x)$ is unbiased and the Cramer-Rao bound is attained.

APPENDIX: Paul Schrater's Class notes: Fisher Info and Cramer-Rao Bound

In this section we look at one approach to the question: How good is our estimator? An important measure of the goodness of an estimator is the variance. Given an estimator $\hat{s} = \phi(x)$, recall the variance of the estimator is:

$$
\begin{aligned}
var[\hat{s}] &= \mathrm{E}[(\phi(x) - s)^2] && (2.62)\\
&= \int_x (\phi(x) - s)^2 p(x|s)dx && (2.63)
\end{aligned}
$$

where s is the true value.

See figure 2.26 for an illustration of the problem of computing the variance of an estimator.

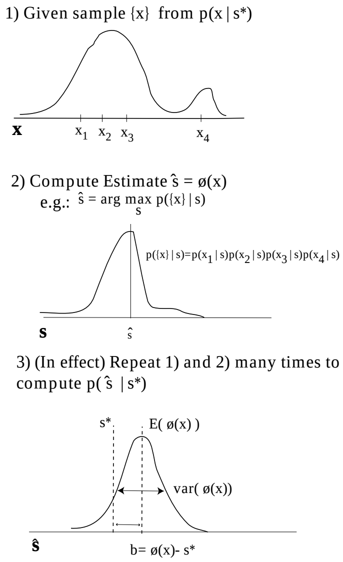

1) Given sample {x} from p(x | s*)
x
x1
x2
x3
x4
2) Compute Estimate ŝ = ø(x)
e.g.: ŝ = arg max p({x} | s)
s
p({x} | s)=p(x1 | s)p(x2 | s)p(x3 | s)p(x4 | s)
s
ŝ
3) (In effect) Repeat 1) and 2) many times to
compute p(ŝ | s*)
s*
E(ø(x))
var(ø(x))
ŝ
b= ø(x)- s*

Figure 2.26 Illustration of the problem of computing the variance of an estimator over repeated estimations.

2.7.1 Fisher Information

In order to study the variance of an estimator, we will introduce a fundamental quantity called the Fisher Information. We will treat the case of a 1-D estimator for simplicity and because the extension to N-D is straightfoward. The Fisher Information $I(s)$ is defined as:

$$I(s) = E[(\frac{d}{ds}\log p(x|s))^2] = \int_x (\frac{d}{ds}\log p(x|s))^2 p(x|s)dx$$

An equivalent definition is given by:

$$I(s) = -E[\frac{d^2}{ds^2}\log p(x|s)] = \int_x \frac{d^2}{ds^2}\log p(x|s)p(x|s)dx$$

This equivalence can be directly demonstrated:

$$\frac{\partial^2}{\partial s^2}log(p(x|s)) = \frac{\partial}{\partial s}\frac{1}{p(x|s)}\frac{\partial p(x_i|s)}{\partial s} = \frac{1}{p(x|s)}\frac{\partial^2 p(x|s)}{\partial s^2} - \frac{1}{p(x|s)^2}\frac{\partial p(x|s)}{\partial s}^2 \tag{2.64}$$

Taking expections and using $\int_x \frac{\partial^2 p(x|s)}{\partial s^2}dx = 0$, we have:

$$E[\frac{\partial^2}{\partial s^2}log(p(x|s))] = -\int_x \frac{1}{p(x|s)}\left(\frac{\partial p(x|s)}{\partial s}\right)^2 dx \tag{2.65}$$

$$= -\int_x p(x_i|s)\left(\frac{\partial \log(p(x|s))}{\partial s}\right)^2 \tag{2.66}$$

2.7.2 Cramer-Rao Bound

Although these properties of the Fisher Information are useful, what makes $I(s)$ fundamental is the fact that $1/I(s)$ forms an absolute lower bound on the variance of an unbiased estimator. An unbiased estimator is defined as an estimator that converges to the true value in expectation: $E[\hat{s}] = E[\phi(x)] = s$. The bias in an estimator $b_\phi(s)$ is defined as $b_\phi(s) = E[\phi(x)] - s$. In addition, when the lower bound is achievable, the estimator is a maximum likelihood estimator. We will now prove both of these statements.

Let $\phi(x)$ be an estimator of s (assumed to be 1-D) based on a sample x from $p(x|s)$. Then $var(\phi(x)) \geq 1/I(s)$.

In order to prove the result, we will make use of the fact that the squared correlation coefficient between two random variables is always less than or equal to 1. So that the exposition is self-contained, we will sketch out the The key idea in the proof is the Cauchy-Schwarz inequality, which is just a statement that the magnitude of the inner product of two vectors is always less than the product of their magnitudes (e.g. $\vec{a}\cdot\vec{b} = ||\vec{a}||\,||\vec{b}||\cos(\theta) \leq ||\vec{a}||\,||\vec{b}|| = \sqrt{(\vec{a}\cdot\vec{a})\,(\vec{b}\cdot\vec{b})}$). Cauchy-Schwarz extends this to the case of a weighted inner product, $\vec{a}^T M\vec{b} == \sum_i a_i b_i m_i$, where M is a diagonal matrix with the weighting vector m on the diagonal. Inner products on continuous functions are defined similarly: $< a(x), b(x) >_{m(x)} = \int a(x)b(x)m(x)dx$. This latter expression is just an expectation when $m(x)$ is equal to a probability distribution. Thus we can state the Cauchy-Scharz inequality as:

$$E[a(x)b(x)] \leq \sqrt{E[a(x)^2]E[b(x)^2]}$$

Rearranging and squaring, we find:

$$\frac{E[a(x)b(x)]^2}{E[a(x)^2]E[b(x)^2]} \leq 1$$

However, we can recognize the latter quantity as the squared correlation coefficient:

$$\rho_{ab}^2 = \frac{E[a(x)b(x)]^2}{E[a(x)^2]E[b(x)^2]} \leq 1$$

In other words, we have just shown that the correlation coefficient between any two random variables is within $[-1, 1]$, which should be a familiar result.

To prove the Cramer-Rao lower bound, we will compute the squared correlation between our estimator $\phi(x)$ and $\frac{\partial \log p(x|s)}{\partial s}$. We know that the squared correlation is bounded by 1. We will show that the covariance between any *unbiased* estimator and the log likelihood must be 1, and the result will follow.

$$\begin{aligned} 1 \geq \rho^2 &= \frac{\text{cov}(\frac{\partial \log p(x|s)}{\partial s}, \phi(x))^2}{var(\frac{\partial \log p(x|s)}{\partial s})var(\phi(x))} && (2.67) \\ &\geq \frac{\left(E\left[\left(\frac{\partial \log p(x|s)}{\partial s}\right)(\phi(x)\right] - E[\frac{\partial \log p(x|s)}{\partial s}]E[\phi(x)]\right)^2}{E\left[\left(\frac{\partial \log p(x|s)}{\partial s} - E[\frac{\partial \log p(x|s)}{\partial s}]\right)^2\right]var(\phi(x))} && (2.68) \end{aligned}$$

However,

$$E[\frac{\partial \log p(x|s)}{\partial s}] = \int_x \frac{\frac{\partial p(x|s)}{\partial s}}{p(x|s)} p(x|s)dx = \frac{\partial}{\partial s}\int_x p(x|s)dx = \frac{\partial}{\partial s}1 = 0 \qquad (2.69)$$

Thus:

$$var(\phi(x)) \geq \frac{E\left[\left(\frac{\partial \log p(x|s)}{\partial s}\right)\phi(x)\right]^2}{E\left[\left(\frac{\partial \log p(x|s)}{\partial s}\right)^2\right]} \qquad (2.70)$$

$$var(\phi(x)) \geq \frac{E\left[\left(\frac{\partial \log p(x|s)}{\partial s}\right)\phi(x)\right]^2}{I(s)} \qquad (2.71)$$

Now

$$\begin{aligned} E\left[\left(\frac{\partial \log p(x|s)}{\partial s}\right)\phi(x)\right] &= \int_x \frac{\frac{\partial p(x|s)}{\partial s}}{p(x|s)} p(x|s)\phi(x)dx \\ &= \frac{\partial}{\partial s}\int_x p(x|s)\phi(x)dx = \frac{\partial}{\partial s}E[\phi(x)] = \frac{\partial}{\partial s}s = 1 \end{aligned}$$

Hence, we have proved:

$$var(\phi(x)) \geq \frac{1}{I(s)} \qquad (2.72)$$

What happens when there is a bias? Following the same argument above, but using $E[\phi(x)] = b(s)$, we let it as an exercise to show that:

$$var(\phi(x)) \geq \frac{\left(1 + \frac{db(s)}{ds}\right)^2}{I(s)} \qquad (2.73)$$

2.7.2.1 When is the lower bound achieved?

The lower bound is clearly achieved when the variance of our estimator is $1/I(s)$. From our proof, we know that equality occurs when the correlation ρ between $\frac{\partial}{\partial s}\log p(x|s)$ and $\phi(x)$ is equal to 1. However, it is easy to show that this only occurs if $\phi(x)-E[\phi(x)]$ differs from $\frac{\partial}{\partial s}\log p(x|s)$ by no more than a linear transform. Because our analysis assumed fixed s, the scale and shift parameters can be different for different values of s. Thus

$$(\phi(x) - E[\phi(x)]) = (\phi(x) - s) = \alpha(s)\frac{\partial}{\partial s}\log p(x|s)$$

. If we integrate both sides with respect to s, and solve for $\log p(x|s)$ we find:

$$\log p(x|s) = (s\phi(x) - s^2)/\alpha(s) + C(x)$$

(Question: Why is the $C(x)$ necessary?) Which leads to the constraint:

$$p(x|s) = z(x)\exp\Big((s\phi(x) - s^2)/\alpha(s)\Big) \tag{2.74}$$

where $z(x) = \exp(C(x))$. Thus we have shown that any unbiased estimator that achieves the Cramer-Rao lower bound has the general form of an exponential distribution. In addition, by differentiating the log of equation 2.74 and setting to zero, we find that the an unbiased $\phi(x)$ that achieves the Cramer-Rao lower bound is always a maximum-likelihood estimator.

In our proof of the Cramer-Rao inequality, we showed that the covariance between $\frac{\partial \log p(x|s)}{\partial s}$ and $\phi(x) - s$ is 1. This has a simple geometrical interpretation. If we remember that the covariance is just a weighted inner product of two vectors (treating functions as infinite-dimensional vectors), then we can interpret $\mathrm{cov}(\frac{\partial \log p(x|s)}{\partial s}, \phi(x) - s) = 1$ as saying that any estimator $\phi(x)$ achieving the lower bound lives in a linear subspace perpendicular to $\frac{\partial \log p(x|s)}{\partial s}$ (because the inner product equal to a constant is the equation of a hyperplane). This is illustrated in figure 2.27.

2.7.2.2 Properties and Uses of the Fisher Information

Because of the lower bound property, a convenient way to summarize an estimator's performance is by computing an efficiency ν defined as:

$$\nu = \frac{1/var\phi(x)}{I(s)}$$

and it follows directly from the Cramer-Rao bound that this quantity is in $[0, 1]$. This measure is commonly called the Fisher efficiency.

The Fisher Information has several important properties that make it a natural measure of information. In the presence of N independent data samples x,

$$I_N(s) = N\, I_1(s)$$

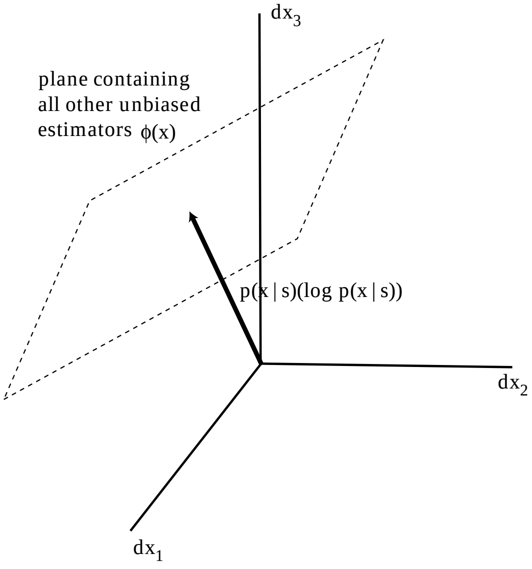


Figure 2.27 A graphical illustration of how any estimator $\phi(x)$ achieving the lower bound lives in a linear subspace of infinite dimensional function space.

where $I_1(s)$ is the Fisher Information for a single sample, defined above. This means that $I_1(s)$ is an amplification factor that determines how much the estimator's variance can decrease with each data sample.

In the presence of two independent sources of data, x and y, the Fisher Information of the combined likelihood estimator $\hat{s} = \arg\max_s (\log p(x|s)p(y|s))$ is given by:

$$I(s) = I(s_x) + I(s_y)$$

where $I(s_x)$ is the Fisher information for $\log p(x|s)$ alone. In addition, if we make separate estimates of s, $\hat{s_x} = argmax_s \log p(x|s)$ and $\hat{s_y} = argmax_s \log p(y|s)$, then the minimum variance estimate of s that combines both kinds of data is:

$$\hat{s} = (I(s_x)\hat{s_x} + I(s_y)\hat{s_y})\,/(I(s_x) + I(s_y))$$

The usefulness of this result will be seen later in the context of cue integration.

Exercise: Show that the properties of $I(s)$ above are replicated by gaussian distributed variables, if we identify $I(s)$ with $1/sigma^2$.

Finally we show how the Fisher Information can be thought of as the reciprocal of the expected variance in a Gaussian approximation to the likelihood.

Gaussian/Saddle-point/Laplace approximation

The second definition of the Fisher information provides a useful intuition about why the Fisher Information is a natural measure of the variance of an estimator.

Let us perform a Taylor series expansion to second order on the log likelihood around the estimated value $\hat{s}$:

$$\log p(x|s) \sim \log p(x|\hat{s}) + (s-\hat{s})\frac{\partial \log p(x|s)}{\partial s}|_{s=\hat{s}} + \frac{1}{2}(s-\hat{s})^2\frac{\partial^2 \log p(x|s)}{\partial s^2}|_{s=\hat{s}} + \text{h.o.t.}$$

When the estimator is a likelihood estimator, then $\frac{\partial \log p(x|s)}{\partial s}|_{s=\hat{s}} = 0$ by definition. Thus,

$$\log p(x|s) \sim \log p(x|\hat{s}) + \frac{1}{2}(s-\hat{s})^2\frac{\partial^2 \log p(x|s)}{\partial s^2}|_{s=\hat{s}}$$

which implies:

$$p(x|s) \sim p(x|\hat{s})\exp(-\frac{1}{2}(s-\hat{s})^2(-\frac{\partial^2 \log p(x|s)}{\partial s^2}|_{s=\hat{s}}))$$

This is a gaussian approximation to the likelihood around the point $\hat{s}$, and we see that $-\frac{\partial^2 \log p(x|s)}{\partial s^2}|_{s=\hat{s}}$ plays the role of $1/\sigma^2$.

An example of approximating the likelihood by a gaussian is shown in figure 2.28, for two different s points. This approximation is common and widely used to compute difficult probability integrals.

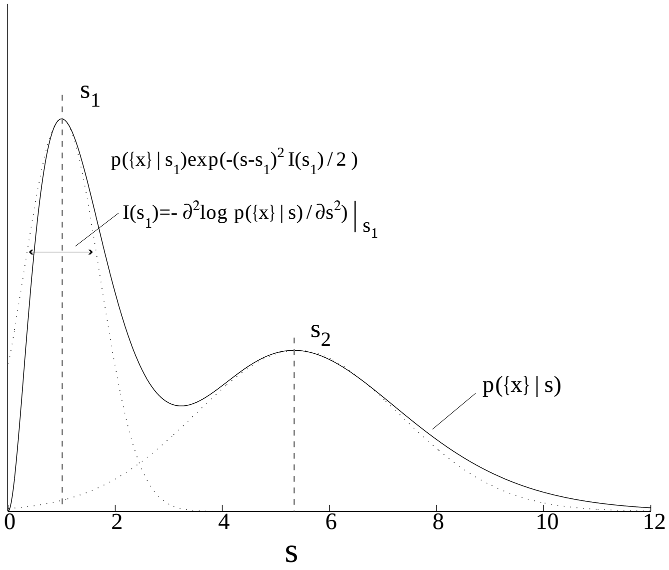


Figure 2.28 A likelihood plotted as a function of s. The gaussian approximation to the likelihood around two points are shown as dashed line curves.

If we take the expectation of the right hand side with respect to $p(x|s)$, then we find:

$$[\log p(x|s)] \sim logp(x|\hat{s}) + (s-\hat{s})^2 E\left[\frac{\partial^2 \log p(x|s)}{\partial s^2}|_{s=\hat{s}}\right] \tag{2.75}$$

$$= logp(x|\hat{s}) - (s-\hat{s})^2 I(s) \tag{2.76}$$

Thus the expected gaussian approximation to the likelihood can be written:

$$p(x|s) \sim p(x|\hat{s}) \exp(-\frac{1}{2}(s-\hat{s})^2 I(s))$$

- With data, we can estimate a lower bound on the variance from the likelihood function.
- This estimate converges to the expected lower bound variance with more data
- To find the expected variance, we average over all possible data sets of a given size.

Key points:

- Basic Bayes Decision Theory for Discrete Variables.
- Basic Bayes Decision Theory for Continuous Variables.
- Multidimensional Inputs and Decision Surfaces.
- ROC curves
- Signal Known Exactly Model
- Fisher information and the Cramer-Rao lower bound.

Exercises

1. Prove that for a fixed observation x for which $P(S = s_2|x) > P(S = s_1|x)$, one should always choose $S = s_2$ to minimize the proportion of errors. With this strategy, it is clear that the probability of error is $P(S = s_1|x)$. But can you prove that this is the smallest average error rate? To do this, let's recast this problem in terms of guessing the flip of a biased coin with a probability p of heads ($p = P(S = s_2|x)$), and q of tails ($q = (1-p) = P(S = s_1|x)$), with $p > q$. Guessing "heads" seems like a good idea, but let's prove that any departure from this strategy can only increase the number of errors.

   Let $s_2 = heads$, and $s_1 = tails$. Let the guesses be $\hat{s}_2 = \text{“}heads''$, and $\hat{s}_1 = \text{“}tails''$. Let $t = p(\hat{s}_1)$ be the proportion of times that the guesser decides to be "creative" and say "tails" instead of heads.

   a) Show that the probability of error is: p(error) $= p(s_2, \hat{s}_1) + p(s_1, \hat{s}_2) = q(1-t) + pt$.

   b) Show that $q(1-t) + pt$ is smallest when $t = 0$.

   People actually tend to perform non-optimally for the biased coin task. We tend to make guesses that match the probabilities of heads and tails, i.e. $p(\hat{s}_2) = p(s_2)$, and $p(\hat{s}_1) = p(s_1)$. We resist boredom even at the expense of error–perhaps sensing a direct connection to the goddess of fortune. Why?

   It is well-known that humans assume that even with an unbiased coin, longish strings of heads are " just too unlikely to be independent"–an assumption that could lead to changes in decisions. Or maybe humans are behaving optimally, but for another task.

   c) For what task is probability matching optimal?

2. Show that for the equal variance Gaussian signal detection problem, adding Gaussian noise to the decision threshold is equivalent to increasing the variance of the observation distributions for the signal and noise states.

3. Prove that no other decision rule can do beter than the MAP rule to minimizes error. Peformance is determined by: $p(\hat{s}, s)$, which can be obtained by marginalizing the joint, $p(\hat{s}, s, x)$, with respect to x:

$$p(\hat{s}, s) = \int p(\hat{s}, s, x)dx =$$

$$\int p(\hat{s}|s, x)p(s|x)p(x)dx$$

   Given x, $\hat{s}$ is conditionally independent of s (see arrows in graph in figure (2.2)), so this is equal to:

$$\int p(\hat{s}|x)p(s|x)p(x)dx$$

,

where $p(\hat{s}|x)$ can be seen to be the indicator function $(\phi_{\hat{s}}(x))$ for the acceptance region for $\hat{s}$.

Now the error rate is determined by the specific values where signal states and decisions don't agree:

$$p(error) = \sum_{i \neq j} p(\hat{s}_i, s_j) =$$

$$\sum_{i \neq j} \int p(\hat{s}_i|x)p(s_j|x)p(x)dx$$

Let $\hat{s}^*$ be the MAP choice. Show that for any other choice, $\hat{s}$:

$$\sum_{i \neq j} p(\hat{s}_i, s_j) - \sum_{i \neq j} p(\hat{s}^*_i, s_j) \geq 0.$$

Hint: show that

$$\int (p(\hat{s}_1|x) - p(\hat{s}^*_1|x))(p(s_2|x) - p(s_1|x))p(x)dx \geq 0.$$

4. Set the loss function to $L(d,s) = (d-s)^2$. Minimizing risk then is equivalent to minimizing the expected variance. Show that the best estimate is then the *mean* $\int sP(s|x)ds$ of the posterior distribution.

5. Show that Kullback-Leibler measure is proportional to d', the signal to noise ratio $(\mu_1 - \mu_2)/\sigma$.

6. Show that minimizing risk with the loss function $L(d,s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(s-d)^2/2\sigma^2}$ is equivalent to doing MAP on $s+n$, where n is Gaussian noise with density $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(n)^2/2\sigma^2}$.

7. The denominator $p(x)$ in Bayes expansion contains information about all the hypotheses. Calculate the decision threshold,

$$x_T = (1/2)(\mu_b + \mu_d) + \frac{\sigma^2}{(\mu_b - \mu_d)} \log \frac{P(s_d)}{P(s_b)}$$

for the light detection example by setting $\frac{(1-P(s_d|x))}{P(s_d|x)} = 1$, and expand $P(s_d|x)$ using Bayes rule.

8. Show that $I(s)$ for a gaussian distribution with mean μ and variance

9. Let $E[\phi(x)] = b(s)$. Following the argument in the chapter, but using $E[\phi(x)] = b(s)$, show that:

$$var(\phi(x)) \geq \frac{\left(1 + \frac{db(s)}{ds}\right)^2}{I(s)} \tag{2.77}$$

Bibliography

[1] Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of human visual signal discrimination. *Science*, **214**, 93-94.

[2] Kersten, D. (1984). Spatial summation in visual noise. *Vision Research*, **24**,1977-1990.

[3] Pelli, D. G. (1990). The quantum efficiency of vision. In Blakemore, C. (Ed.), *Vision:Coding and Efficiency*(pp. Cambridge: Cambridge University Press.

[4] Watson, A. B., Barlow, H. B., & Robson, J. G. (1983). What does the eye see best? *Nature*, **31**, 419-422.