Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

5 Minimax Hypothesis Testing

Thus far, we have seen two approaches to hypothesis testing. In the Bayesian approach, we model the hypothesis as a random variable H with prior $p = \mathbb{P}(\mathsf{H} = H_1)$, and choose the function $\tilde{C}(H_j, H_i) = C_{ij}$ as the cost of deciding that the hypothesis is $\hat{H} = H_i$ when the correct hypothesis is $\mathsf{H} = H_j$. We then find the decision rule $\hat{H}_{\mathrm{B}}(\cdot)$ that minimizes the Bayes risk $\mathbb{E}[\tilde{C}(\mathsf{H}, f(\mathbf{y}))]$, where the expectation is over both H and $\mathbf{y}$. Equivalently, we can write

$$\hat{H}_{\mathrm{B}}(\cdot) = \arg\min_{f(\cdot)} \varphi_{\mathrm{B}}(f), \quad \text{where } \varphi_{\mathrm{B}}(f) = \mathbb{E}[\tilde{C}(\mathsf{H}, f(\mathbf{y}))], \tag{1}$$

with $\varphi_{\mathrm{B}}(f)$ as our notation for the Bayes risk.

In the Neyman-Pearson approach, we avoid assuming a prior and explicitly assigning costs by fixing some $\alpha > 0$ and obtaining a decision rule as the solution to

$$\hat{H}_{\mathrm{NP}}(\cdot) = \mathop{\arg\max}_{\{f(\cdot)\,:\,P_{\mathrm{F}}(f)\leq\alpha\}} P_{\mathrm{D}}(f), \tag{2}$$

where $P_{\mathrm{D}}(f)$ and $P_{\mathrm{F}}(f)$ are the probabilities of detection and false-alarm, respectively, when the decision rule $f(\cdot)$ is used. Equivalently, we can write

$$\hat{H}_{\mathrm{NP}}(\cdot) = \arg\min_{f(\cdot)} \varphi_{\mathrm{NP}}^{\alpha}(f), \quad \text{where } \varphi_{\mathrm{NP}}^{\alpha}(f) = \begin{cases} -P_{\mathrm{D}}(f) & \text{if } P_{\mathrm{F}}(f) \leq \alpha \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

with $\varphi_{\mathrm{NP}}^{\alpha}(f)$ as our notation for the Neyman-Pearson risk with parameter α.

In this section, we consider a third alternative that can be used when we are comfortable assigning costs $\tilde{C}(H_j, H_i)$, but not priors. In such a scenario, we can adopt an adversarial model. In particular, we can model the problem as a (zero-sum) game in which nature will pick the most detrimental prior for whatever decision rule we choose, and we must choose the best possible decision rule given that we know nature will behave in this manner. Such worst-case analysis is referred to as the *minimax* approach to hypothesis testing, and leads to robust (though conservative) decision rules.

In minimax hypothesis testing, we define $\varphi_{\mathrm{M}}(f, p)$ as the Bayes risk if the prior for H_1 were p, i.e.,

$$\varphi_{\mathrm{M}}(f, p) = (1 - p)\, \mathbb{E}[\tilde{C}(\mathsf{H}, f(\mathbf{y})) \mid \mathsf{H} = H_0] + p\, \mathbb{E}[\tilde{C}(\mathsf{H}, f(\mathbf{y})) \mid \mathsf{H} = H_1], \tag{4}$$

and express the optimum decision rule $\hat{H}_{\mathrm{M}}(\cdot)$ as (with a slight temporary abuse of notation)

$$\hat{H}_{\mathrm{M}}(\cdot) = \arg\min_{f(\cdot)} \varphi_{\mathrm{M}}(f), \quad \text{where } \varphi_{\mathrm{M}}(f) = \max_{p\in[0,1]} \varphi_{\mathrm{M}}(f, p), \tag{5}$$

with $\varphi_{\mathrm{M}}(f)$ as our notation for the minimax risk. Note that we are implicitly restricting our attention to deterministic decision rules by our formulation (5); we discuss randomized decision rules later. In what follows, since there will be no risk of confusion we drop the subscript, using φ instead of φ_{M}.

5.1 Mismatched Bayes Decision Rules

Our development will require the analysis of mismatched decision rules, i.e., decision rules not properly matched to the actual scenario in which they are used. For this purpose, we adopt some additional notation.

To begin, we use $\hat{H}_{\mathrm{B}}(\cdot, p)$ to denote the Bayes decision rule one would use if the prior were p, i.e., the likelihood ratio test

$$\hat{H}_{\mathrm{B}}(\mathbf{y}, p) = \begin{cases} H_1, & L(\mathbf{y}) \geq \dfrac{(1-p)}{p} \dfrac{C_{10} - C_{00}}{C_{01} - C_{11}}, \\ H_0, & \text{otherwise}. \end{cases} \tag{6}$$

As related notation, we use $P_{\mathrm{D}}(p)$ and $P_{\mathrm{F}}(p)$ to denote the detection and false-alarm probabilities, respectively, of the Bayes decision rule $\hat{H}_{\mathrm{B}}(\cdot, p)$.

Next, we define the *mismatch Bayes risk* as

$$\varphi_{\mathrm{B}}(q, p) \triangleq \varphi(\hat{H}_{\mathrm{B}}(\cdot, q), p), \tag{7}$$

which is the Bayes risk when a Bayes decision rule is designed assuming the prior is q, but the actual prior in effect is p.

Some properties of the mismatch Bayes risk will be useful. First, $\varphi_{\mathrm{B}}(q, p)$ is linear in p (for any fixed q):

$$\begin{aligned} \varphi_{\mathrm{B}}(q, p) &= (1-p)\ \mathbb{E}[\tilde{C}(\mathsf{H}, \hat{H}_{\mathrm{B}}(\mathbf{y}, q)) \mid \mathsf{H} = H_0] + p\ \mathbb{E}[\tilde{C}(\mathsf{H}, \hat{H}_{\mathrm{B}}(\mathbf{y}, q)) \mid \mathsf{H} = H_1] && (8) \\ &= (1-p)\big[C_{00}(1 - P_{\mathrm{F}}(q)) + C_{10} P_{\mathrm{F}}(q)\big] + p\big[C_{01}(1 - P_{\mathrm{D}}(q)) + C_{11} P_{\mathrm{D}}(q)\big] && (9) \\ &= \varphi_{\mathrm{B}}(q, 0) + p\, \frac{\partial \varphi_{\mathrm{B}}(q, p)}{\partial p}, && (10) \end{aligned}$$

where

$$\varphi_{\mathrm{B}}(q, 0) = C_{00} + (C_{10} - C_{00}) P_{\mathrm{F}}(q) \tag{11}$$

and

$$\frac{\partial \varphi_{\mathrm{B}}(q, p)}{\partial p} = (C_{01} - C_{00}) - (C_{01} - C_{11}) P_{\mathrm{D}}(q) - (C_{10} - C_{00}) P_{\mathrm{F}}(q). \tag{12}$$

Second, the mismatch and matched Bayes risks are related according to

$$\varphi_{\mathrm{B}}(q, p) \geq \varphi_{\mathrm{B}}(p, p), \tag{13}$$

with equality if $q = p$. These relationships are depicted graphically in Fig. 1. In particular, when plotted as functions of p for a particular q, the mismatch Bayes risk

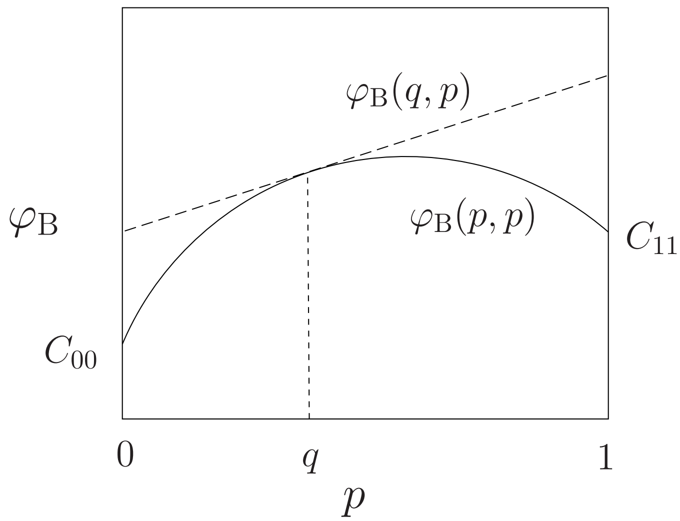


Figure 1: Mismatch Bayes risk. The solid curve indicates the Bayes risk function as a function of the actual prior probability for a Bayes decision rule that is correctly matched. The dashed line indicates the mismatch Bayes risk when the Bayes decision rule is designed for prior q but the actual prior is p.

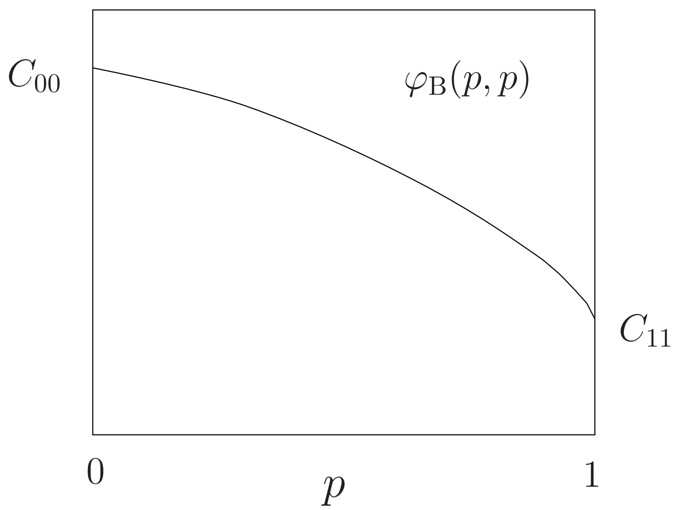
$\varphi_{\mathrm{B}}(q,p)$ is a line tangent to $\varphi_{\mathrm{B}}(p,p)$ at $p = q$. Furthermore, since this must be true for all choices of q, we further conclude that the minimum Bayes risk $\varphi_{\mathrm{B}}(p,p)$ must be concave in p, as Fig. 1 also reflects.

The example $\varphi_{\mathrm{B}}(p,p)$ depicted in Fig. 1 has the property that it exhibits a point of zero slope (corresponding to a stationary point). Whether this occurs or not depends on the details of the cost assignments. For example, when $C_{00} = C_{11} = 0$, there is always such a point of zero slope. However, more generally $\varphi_{\mathrm{B}}(p,p)$ need not have such a point. Thus, in our analysis, we will need to also consider two other possible forms of $\varphi_{\mathrm{B}}(p,p)$, corresponding to monotonically decreasing and monotonically increasing functions, as depicted in Fig. 2.

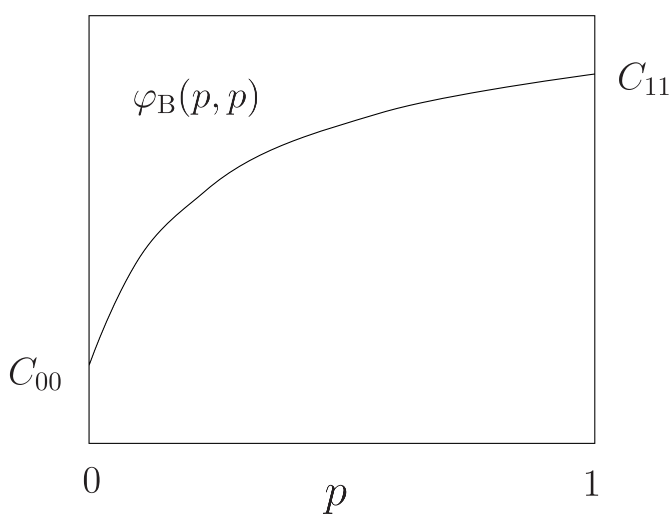
In general $\varphi_{\mathrm{B}}(p,p)$ is a continuous function of p. However, it may not be differentiable everywhere—there may be values of p where the slope is discontinuous. To simplify our subsequent development, we assume that $P_{\mathrm{D}}(q)$ and $P_{\mathrm{F}}(q)$ are continuous functions of q, as occurs when $\mathbf{y}$ is a continuous-valued random variable whose probability density function is bounded. In this case, $\varphi_{\mathrm{B}}(p,p)$ is differentiable everywhere. When $\mathbf{y}$ is discrete, we will require a slight refinement of our analysis, which we will discuss later.

5.2 The Minimax Decision Rule

Our main result is the following theorem, which establishes that the risk function $\varphi(f,p)$ exhibits a saddle point that defines the minimax-optimal decision rule.



(a)



(b)

Figure 2: Examples of Bayes risk functions $\varphi_{\mathrm{B}}(p,p)$ that do not have points of zero slope.

Theorem 1. *In terms of the notation* (7) *and* (12), *we have*

$$\varphi(f_*, p_*) \triangleq \min_f \max_{p\in[0,1]} \varphi(f,p) = \max_{p\in[0,1]} \min_f \varphi(f,p), \tag{14}$$

where, evaluating the right-hand side of (14) *we see*

$$p_* = \arg\max_{p\in[0,1]} \varphi_{\mathrm{B}}(p,p). \tag{15}$$

Note that (15) is equivalent to the statement that p_* is the solution to

$$\left.\frac{\partial \varphi_{\mathrm{B}}(q,p)}{\partial p}\right|_{q=p_*} = 0 \tag{16a}$$

when one exists; otherwise

$$p_* = \begin{cases} 0 & \text{if } \left.\frac{\partial \varphi_{\mathrm{B}}(q,p)}{\partial p}\right|_{q=p} < 0 \text{ for all } p \in [0,1], \\ 1 & \text{if } \left.\frac{\partial \varphi_{\mathrm{B}}(q,p)}{\partial p}\right|_{q=p} > 0 \text{ for all } p \in [0,1]. \end{cases} \tag{16b}$$

Theorem 1 gives us the solution to the game between the system designer and nature. First, it tells us that to make the system perform as well as possible the system designer should always use as the decision rule a likelihood ratio test corresponding to a Bayesian scenario in which the prior is p_*, i.e., $f_*(\cdot) = \hat{H}_{\mathrm{B}}(\cdot, p_*)$. Second, it tells us that nature should choose prior p_* to make the system perform as poorly as possible. For this reason, p_* is referred to as the *least favorable prior.*

There is also a useful graphical interpretation of the solution p_* to (16a). In particular, using (12) we see that (16a) is equivalent to imposing the following constraint between P_{D} and P_{F}:

$$P_{\mathrm{D}} = \frac{(C_{01} - C_{00})}{(C_{01} - C_{11})} - \frac{(C_{10} - C_{00})}{(C_{01} - C_{11})} P_{\mathrm{F}} \tag{17}$$

which when $C_{ij} > C_{ii}, C_{jj}$ for all i,j is a line with negative slope in the $(P_{\mathrm{D}}, P_{\mathrm{F}})$ unit square that intersects the operating characteristic of the likelihood ratio test at a single point. The point of intersection therefore defines $(P_{\mathrm{D}}(p_*), P_{\mathrm{F}}(p_*))$, and, implicitly, p_*.

It is also worth emphasizing that the equivalent maximin expression in (14) is much easier to evaluate than the minimax one, which is typical of such saddle-point theorems. We will see such phenomena again later in a related but more general framework for data modeling.

In our proof of Theorem 1, we'll need the following fact.

Fact 1 (Minimax Inequality)**.** *For any function g,*

$$\min_a \max_b g(a,b) \geq \max_b \min_a g(a,b) \tag{18}$$

whenever the associated minima and maxima are well-defined.

This fact has an intuitive interpretation from the perspective of a (zero-sum) game where player A wants to minimize g and is able to choose a to this end, while player B wants to maximize g and is able to choose b to this end. In particular, the left-hand side of (18) is result of the game when player B gets to choose last, while the right-hand side is the result when player A gets to choose last. Thus (18) expresses that it is preferable to choose last in any such game, since the choice can be made having observed the opponent's move. More formally, a proof is as follows.

Proof. First, note that for any a' and b'

$$g_+(a') \triangleq \max_b g(a', b) \geq g(a', b') \geq \min_a g(a, b') \triangleq g_-(b') \tag{19}$$

Since the left- and right-hand sides of (19) depend only on a' and b', respectively, and since a' and b' can be freely chosen, we simply make the choices

$$a'_* = \arg\min_{a'} g_+(a') \quad \text{and} \quad b'_* = \arg\max_{b'} g_-(b').$$

Substituting the resulting

$$g_+(a'_*) = \min_{a'} \max_b g(a', b) \quad \text{and} \quad g_-(b'_*) = \max_{b'} \min_a g(a, b')$$

into (19) yields (18). □

Proof of Theorem 1. To establish our result, we upper and lower bound the minimax expression, and show these bounds to be equal.

We begin with our upper bound. For any $f'(\cdot)$ we have

$$\min_{f(\cdot)} \max_{p\in[0,1]} \varphi(f, p) \leq \max_{p\in[0,1]} \varphi(f', p)$$

so in particular for $f'(\cdot) = \hat{H}_{\mathrm{B}}(\cdot, q)$ with any $q \in [0, 1]$ we have

$$\min_{f(\cdot)} \max_{p\in[0,1]} \varphi(f, p) \leq \max_{p\in[0,1]} \varphi_{\mathrm{B}}(q, p), \tag{20}$$

Turning next to our lower bound, from Fact 1 we have

$$\min_{f(\cdot)} \max_{p\in[0,1]} \varphi(f, p) \geq \max_{p\in[0,1]} \min_{f(\cdot)} \varphi(f, p) \geq \min_{f(\cdot)} \varphi(f, q) \tag{21}$$

for all $q \in [0, 1]$. But the solution to the minimization on the right-hand side of (21) is the Bayes decision rule for prior q, i.e., $f(\cdot) = \hat{H}_{\mathrm{B}}(\cdot, q)$, so

$$\min_{f(\cdot)} \varphi(f, q) = \varphi_{\mathrm{B}}(q, q),$$

which when substituted into (21) yields

$$\min_{f(\cdot)} \max_{p\in[0,1]} \varphi(f, p) \geq \varphi_{\mathrm{B}}(q, q). \tag{22}$$

We now consider the three separate cases for $\varphi_{\mathrm{B}}(p, p)$, as Figs. 1 and 2 depict.

Case I: $\varphi_{\mathrm{B}}(p, p)$ Nonmonotonic

In this case,[1] there exists a q_* such that [cf. (16)]

$$\left.\frac{\partial \varphi_{\mathrm{B}}(q, p)}{\partial p}\right|_{q=q_*} = 0, \tag{23}$$

corresponding to $\mathbb{E}[\tilde{C}(\mathsf{H}, \hat{H}_{\mathrm{B}}(\mathbf{y}, q_*)) \mid \mathsf{H} = H_0] = \mathbb{E}[\tilde{C}(\mathsf{H}, \hat{H}_{\mathrm{B}}(\mathbf{y}, q_*)) \mid \mathsf{H} = H_1]$, in which case $\varphi_{\mathrm{B}}(q_*, p)$ in (10) does not depend on p. Evidently, via (9) this q_* is given implicitly via

$$(C_{01} - C_{11}) P_{\mathrm{D}}(q_*) + (C_{10} - C_{00}) P_{\mathrm{F}}(q_*) = (C_{01} - C_{00}). \tag{24}$$

Visually, this is the q_* that makes $\varphi_{\mathrm{B}}(q_*, p)$ a horizontal line when plotted as a function of p, and thus touches $\varphi_{\mathrm{B}}(p, p)$ at its maximum.

With this choice, since $\varphi_{\mathrm{B}}(q_*, p)$ does not depend on p, we can drop the maximization on the right-hand side of (20) and just write

$$\min_{f(\cdot)} \max_{p \in [0,1]} \varphi(f, p) \leq \varphi_{\mathrm{B}}(q_*, q_*), \tag{25}$$

where we have freely chosen $p = q_*$. However, specializing the lower bound (22) to the choice $q = q_*$ yields

$$\min_{f(\cdot)} \max_{p \in [0,1]} \varphi(f, p) \geq \varphi_{\mathrm{B}}(q_*, q_*). \tag{26}$$

which coincides with the upper bound (25), and thus (14) holds for this case.

Case II: $\varphi_{\mathrm{B}}(p, p)$ Monotonic

When $\varphi_{\mathrm{B}}(p, p)$ is monotonic increasing, we choose $q = 1$ in our upper bound (20), in which case the bound becomes

$$\min_{f(\cdot)} \max_{p \in [0,1]} \varphi(f, p) \leq \max_{p \in [0,1]} \varphi_{\mathrm{B}}(1, p) = \varphi_{\mathrm{B}}(1, 1), \tag{27}$$

where that the maximizing p is 1 can be visualized from Fig. 2(b). For the lower bound, we simply choose $q = 1$ in (22), so the bounds match, proving (14) in this case of the theorem since $q_* = 1$ for this case.

When $\varphi_{\mathrm{B}}(p, p)$ is monotonic decreasing, we analogously choose $q = 0$ in (20) and (22), yielding (14) in this case as well, since $q_* = 0$ for this case.

$\square$

As a final comment, for such continuous cases, randomized minimax decision rules cannot achieve higher performance than the optimum deterministic one. We omit a proof. However, there are cases when randomized minimax decision rules offer better performance, as we discuss next.

[1] This is the only case that arises, for example, when the associated distributions for $\mathbf{y}$ are continuous and $C_{00} = C_{11} = 0$.

5.3 Discrete Observations

When $P_{\mathrm{D}}(\cdot)$ and $P_{\mathrm{F}}(\cdot)$ are not continuous functions of their arguments, such as occurs in the case the observations are discrete, the results are similar, except that a randomized test (i.e., a mixed strategy for the game) is generally required.

In particular, recall that the operating characteristic of the likelihood ratio test in the case of discrete-valued observations is a set of isolated $(P_{\mathrm{D}}, P_{\mathrm{F}})$ pairs. However, the efficient frontier of operating points is the convex hull of these points, i.e., curve formed by connecting the pairs with lines. Recall that this frontier is achieved by using a simple randomized test.

It is such randomized tests that are generally required with the minimax criterion. Recall that for continuous-valued data the p_* in (16a) is implicitly given by the point of intersection $(P_{\mathrm{D}}(p_*), P_{\mathrm{F}}(p_*))$ of the line (17) with the operating characteristic of the likelihood ratio test. When the data are discrete, it is the intersection of the line with the efficient frontier, which is the convex hull of the operating characteristic. In general, this point of intersection will correspond to a point on a line connecting two likelihood ratio test point. The minimax decision rule is then a randomization between the two associated likelihood ratio tests, where the location of the intersection along the line determines the required bias in the randomization.

While minimax hypothesis testing is of more conceptual than practical value at this point, the minimax framework will play a central role in our development of modeling later in the notes, where we develop a related by more general saddle-point theorem.