I. SUMMARY

Here are the key ideas and results:

• The Hypothesis Testing problem is to maximize $P[Z = 1 | X = 1]$ subject to $P[Z = 1 | X = 0] \leq \beta$ where $Z$ is based on $Y$ and $f_{Y|X}$ is known.
• The solution of the HT problem is given by the Neyman-Pearson theorem 1.
• Some composite HT problems have a simple solution; most do not.
• $k(Y)$ is sufficient for $X$ if $X \rightsquigarrow k(Y) \rightsquigarrow Y$ (see Definition 2).

II. BINARY HYPOTHESIS TESTING

So far we have considered the Bayesian detection problem of minimizing $\mathbb{E}(c(X, g(Y)))$ when both $p_X$ and $f_{Y|X}$ are known. Recall that, when $X \in \{0, 1\}$ the optimal decision has the form $g(Y) = h(\Lambda(Y))$ where $\Lambda(y) = f_{Y|X}[y|1]/f_{Y|X}[y|0]$ is the likelihood ratio. In particular, if $\Lambda(y) = f(k(y))$, then $k(Y)$ is a sufficient statistic for detecting $X$ given $Y$.

In this section we explore the detection problem when $p_X$, the prior, is unknown. We start with the case when $X \in \{0, 1\}$.

A. Formulation

**Definition 1: Binary Hypothesis Testing Problem**

One is given $f_{Y|X}$. For each observed value $y$ of $Y$, one chooses $\phi(y) \in [0, 1]$ and one lets $Z = 1$ with probability $\phi(y)$ and $Z = 0$ otherwise. One is also given $\beta \in (0, 1)$. The objective is to choose $\phi$ to

$$
\text{maximize } P[Z = 1 | X = 1] \text{ subject to } P[Z = 1 | X = 0] \leq \beta.
$$

(1)

One interpretation is that $X = 1$ means that your house is on fire. In that case the problem is to design the alarm system to detect a fire with the largest probability compatible with a probability of false alarm at most equal to $\beta$.

B. Neyman-Pearson Theorem

The key result is the following.

**Theorem 1: Neyman-Pearson**

The solution to the binary hypothesis testing problem is as follows:

$$
\phi(y) = \begin{cases} 
1, & \text{if } \Lambda(y) := \frac{f_{Y|X}[y|1]}{f_{Y|X}[y|0]} > \lambda; \\
\gamma, & \text{if } \Lambda(y) = \lambda; \\
0, & \text{if } \Lambda(y) < \lambda,
\end{cases}
$$

(2)

where $\lambda > 0$ and $\gamma \in [0, 1]$ are the only values such that

$$
P[Z = 1 | X = 0] = \beta \text{ when } P[Z = 1 | Y = y] = \phi(y).
$$

(3)

The interpretation of the result is that if $\Lambda(y) > \lambda$, then one is pretty sure that $X = 1$ because the observed value would be much less likely if $X = 0$. In that case, one safely decides $Z = 1$. If $\Lambda(y) < \lambda$, one decides $Z = 0$. If $\Lambda(y) = \lambda$, one edges the bet by deciding $Z = 1$ with probability $\gamma$. The threshold $\lambda$ is an adjustment of the ‘sensitivity’ of the alarm: the lower $\lambda$, the most likely the alarm is to sound. The randomization $\gamma$ is designed to achieve exactly the probability of false alarm.

Before looking at the proof, let us examine two representative examples.
C. Two Examples

Example 1: We consider the binary symmetric channel. That is, one is given some distribution $P[Z = 1 | X = 0]$ and, for $x, y \in \{0, 1\}$, one has

$$P[Y = y | X = x] = \begin{cases} 1 - \epsilon, & \text{if } x = y; \\ \epsilon, & \text{if } x \neq y. \end{cases}$$

We find

$$\Lambda(y) = \frac{P[Y = y | X = 1]}{P[Y = y | X = 0]} = \begin{cases} \frac{1-\epsilon}{\epsilon}, & \text{if } y = 1; \\ \frac{\epsilon}{1-\epsilon}, & \text{if } y = 0. \end{cases}$$

Note that $\Lambda(1) > \Lambda(0)$. Let us fix $\gamma$ and examine $P[Z = 1 | X = 0]$ as a function of $\lambda$. For instance, if $\lambda = \Lambda(1)$, then we find that $P[Z = 1 | X = 0] = \gamma P[Y = 1 | X = 0] = \gamma \epsilon$. Similarly, if $\lambda \in (\Lambda(0), \Lambda(1))$, then $P[Z = 1 | X = 0] = P[Y = 1 | X = 0] = \epsilon$.

By looking at the other possibilities $\lambda > \Lambda(1), \lambda = \Lambda(0)$, and $\lambda < \Lambda(0)$, one finds the results shown in Figure 1. Unfortunately, because of a bug in Visio, the figure shows $g$ instead of $\gamma$ and $L(0), L(1)$ instead of $\lambda \epsilon, \Lambda(0), \epsilon$, and $l$ instead of $\lambda$. Thanks again Bill!

![Fig. 1. Probability of false alarm as a function of threshold.](image)

Consequently, one finds that

$$(\gamma_0, \gamma_1) = \begin{cases} (0, \frac{\beta}{\epsilon}), & \text{if } \beta \leq \epsilon; \\ (\frac{\beta - \gamma}{1-\epsilon}, 1), & \text{if } \beta > \epsilon. \end{cases}$$

Example 2: In this example, $Y = X + V$ where $X$ and $V$ are independent and $V = N(0, 1)$. Here,

$$\Lambda(y) = \exp\{-\frac{1}{2} (y - 1)^2\} / \exp\{-\frac{1}{2} y^2\} = \exp\{y - \frac{1}{2}\}.$$

Since $\Lambda(y)$ is increasing in $y$, we see that the solution of the hypothesis testing problem has the following form:

$$\phi(y) = \begin{cases} 1, & \text{if } y > y_0; \\ \gamma, & \text{if } y = y_0; \\ 0, & \text{if } y < y_0, \end{cases}$$

where $y_0 = \Lambda^{-1}(\lambda)$. However, since $P[Y = y_0 | X = x] = 0$ for $x \in \{0, 1\}$, one can ignore the middle possibility. Accordingly, the solution is

$$\phi(y) = \begin{cases} 1, & \text{if } y > y_0; \\ 0, & \text{if } y < y_0. \end{cases}$$

The value of $y_0$ is such that $P[Z = 1 | X = 0] = \beta$, i.e.,

$$\beta = P[Y > y_0 | X = 0] = P(V > y_0).$$

We now turn to the proof of Theorem 1.

D. Proof of Neyman-Pearson Theorem

Define $Z$ as indicated by the theorem and let $V$ be some random variable that corresponds to another choice $\phi'(y)$ instead of $\phi(y)$. We assume that $V$ satisfies the bound on the false alarm probability, i.e., that $P[V = 1 | X = 0] \leq \beta$. We show that $P[V = 1 | X = 1] \leq P[Z = 1 | X = 1]$. To do this, note that $\Lambda(Y)(Z - V) \geq \lambda(Z - V)$, where $\lambda$ is defined as (2). Hence,

$$E[\Lambda(Y)(Z - V) | X = 0] \geq \lambda E[Z - V | X = 0] = \lambda (P[Z = 1 | X = 0] - P[V = 1 | X = 0]) \geq 0.$$
But
\[ E[\Lambda(Y)(Z - V)|X = 0] = E[Z - V|X = 1]. \]

Indeed,
\[
E[\Lambda(Y)(Z - V)|X = 0] = \int \Lambda(y)[\phi(y) - \phi'(y)]f_{Y|X}[y|0]dy
= \int \frac{f_{Y|X}[y|1]}{f_{Y|X}[y|0]}[\phi(y) - \phi'(y)]f_{Y|X}[y|0]dy
= \int [\phi(y) - \phi'(y)]f_{Y|X}[y|1]dy = E[Z - V|X = 1].
\]

Hence,
\[ 0 \leq E[Z - V|X = 1] = P[Z = 1|X = 1] - P[V = 1|X = 1], \]
as was to be shown.

For an intuitive discussion of this result, see [2], page 126.

E. Important Observations

One remarkable fact is that the optimal decision is again a function of the likelihood ratio. Thus, as in the Bayesian case, if \( \Lambda(y) = f(k(y)) \), the optimal decision for the hypothesis testing problem is a function of the sufficient statistic \( k(Y) \).

We used two simple but useful observations in the examples. The first one is that when \( \Lambda(y) \) is increasing in \( y \), the decision rule is a threshold on \( Y \). The second is that when \( Y \) has a density, there is no need to randomize by introducing some \( \gamma \).

F. Composite Hypotheses

We have considered only the case \( X \in \{0, 1\} \). The general case would be that \( X \) takes values in some set \( X \) and one wishes to determine if \( X \in A \subset X \) or \( X \notin A \). Say that one selects \( Z = 1 \) to mean that one thinks that \( X \in A \). One would then attempt to maximize \( P[Z = 1|X \in A] \) subject to the constraint that \( P[Z = 1|X \notin A] \leq \beta \). This general problem does not admit a simple answer. For instance, the likelihood ratio \( P[Y = y|X \in A]/P[Y = y|X \notin A] \) is not defined since we do not have a prior distribution of \( X \). However, some problems have a simple answer. We give one example.

Example 3: One knows that, given \( X, Y = N(X, 1) \) and we want to determine whether \( X = \mu_0 \) or \( X > \mu_0 \). First note that if the alternatives are \( X = \mu_0 \) or \( X = \mu_1 > \mu_0 \), then the optimal decision would be \( Z = 1\{Y > y_0\} \) with \( P[Y > y_0|X = \mu_0] = \beta \). Thus, the value of \( y_0 \) does not depend on \( \mu_1 \). It follows that this decision rule is optimal for the composite problem.

III. SUFFICIENT STATISTICS

Here is a discussion that may help clarify the notion of sufficient statistic. Imagine that one observes \( Z = \phi(X, V) \) where \( V \) has a given distribution and \( \phi \) is a known function. Clearly, \( Z \) contains some information about the value of \( X \). Now imagine that \( W \) is some random variable independent of \( X \) and \( V \). It is then obvious that knowing \( W \) does not help figuring out the value of \( X \). Thus, if one is given both \( Z \) and \( W \), one can safely discard \( W \) if the goal is to guess \( X \). In that sense, \( Z \) contains all the information in \( (Z, W) \) about \( X \). Now imagine that instead of being given \( (Z, W) \), one is given \( (Z, h(Z, W)) \). It should then be obvious that one can disregard \( h(Z, W) \) that only adds noise to the relevant information \( Z \) about \( X \).

As an analogy, imagine that \( Z \) is a photograph of a person one tries to identify and that \( h(Z, W) \) is a blurred version of the photograph \( Z \). Clearly, the blurred picture does not add any new information and \( Z \) is sufficient.

Now imagine that \( h(Z, W) = Y \) and that \( Z = f(Y) \). In such a situation, we say that \( Z \) is a sufficient statistic of \( Y \) for estimating \( X \).

Let us look at three examples.

A. Example 1: I.i.d. \( B(p) \)

We flip a coin \( N \) times and the outcomes are i.i.d. and equal to \( H \) with probability \( p \) and to \( T \) with probability \( 1 - p \). We claim that the number of \( H \)'s is a sufficient statistic for estimating \( p \).

To say this, assume that one is given a set of \( N \) flipped coin. Designate by \( W \) a permutation chosen uniformly among the \( N! \) possible permutations of the coins. If \( Z \) is the number of \( H \)'s, we can consider that \( h(Z, W) = Y \) determines which of the coins are \( H \) and which are \( T \). Now, \( Z \) has a distribution that depends on \( p \). However, it should be clear that \( Y \) only adds irrelevant information to \( Z \) for estimating \( p \), since \( Y \) is determined by the outcome \( W \) of an experiment that does not depend on \( p \) (the random permutation).
B. Example 2: i.i.d. $N(\mu, 1)$

Assume that $Y_1, Y_2$ are i.i.d. $N(\mu, 1)$. We want to estimate $\mu$ from $Y = (Y_1, Y_2)$. We claim that $Z = Y_1 + Y_2$ is sufficient. To see this, let $W = Y_1 - Y_2$ and note that $W = N(0, 2)$ is independent of $Z$. Thus, $W$ is independent of $Z$ and $X$. Also, $Y = ((Z + W)/2, (Z - W)/2) = h(Z, W)$ and it is clear that $Z$ is sufficient.

C. Example 3: $N(\mu(x), I)$

We extend the previous example to $Y$ being $N(a, I)$ if $X = 0$ and $N(b, I)$ if $X = 1$. Let’s look at Figure 2. The figure suggests that the component of $Y$ that is orthogonal to $b - a$ contains no information about whether $X = 0$ or $X = 1$. That component, designated by $W$ in the figure, has a distribution that does not depend on $X$. Thus, we expect that given the projection of $Y$ on $b - a$, any other information about $Y$ is irrelevant for estimating $X$. That is, $(b - a)^T Y$ is sufficient.

D. Example 4: $N(\mu(x), \Sigma)$

We modify the previous example to the case of correlated random variables. Thus, $Y$ is $N(a, \Sigma)$ if $X = 0$ and $N(b, \Sigma)$ if $X = 1$. We assume $|\Sigma| \neq 0$. Assume that $\Sigma = A A^T$. Instead of looking at $Y$, let us look at $Z = A^{-1} Y$. Note that $\text{cov}(Z|X) = A^{-1} \Sigma (A^T)^{-1} = A^{-1} A A^T (A^T)^{-1} = I$. Hence, $Z = N(A^{-1} a, I)$ if $X = 0$ and $Z = N(A^{-1} b, I)$ if $X = 1$. Accordingly, $(A^{-1} b - A^{-1} a)^T Z = (b - a) (A^{-1})^T A^{-1} Y = (b - a)^T \Sigma^{-1} Y$
is sufficient. Figure 3 illustrates the result.

All this is pretty cute, but in practice, how do we tell what is sufficient?

Definition 2: Sufficient Statistic

$k(Y)$ is a sufficient statistic for $X$ if

\[ k(Y) = \varphi(X, V) \text{ and } Y = \psi(k(Y), W) \text{ where } V \text{ and } W \text{ are independent.} \]
Here is a picture:

Here is a picture:

Intuitively, this definition means that $Y$ is a noisy version of $k(Y)$ as far as $X$ is concerned. Thus, $Y$ is obtained from the sufficient statistic $k(Y)$ by adding some noise to it, that noise does not depend on $X$.

In Example 1, the additional noise is the permutation of the coins. In Example 2, the noise is the difference between the random variables. In Example 3, the noise is the component of $Y$ orthogonal to $b-a$. In example 4, the noise is the component of $A^{-1}Y$ orthogonal to $A^{-1}(b-a)$.

Mathematically, how do we check that $k(Y)$ is sufficient? We need to make sure that the distribution of $Y$ given $k(Y)$ does not depend on $X$. That is,

$$f_{Y|k(Y),X}[y|z,x] = f_{Y|k(Y)}[y|z] =: h(y,z).$$

(4)

Note that

$$f_{Y|X}[y|x] = f_{Y;k(Y)|X}[y,k(y)|x] = f_{Y|k(Y),X}[y|k(y),x]f_{k(Y)|X}[k(y)|x].$$

Thus, (4) holds if and only if

$$f_{Y|X}[y|x] = h(y,k(y))g(k(y),x)$$

where $g(k(y),x) = f_{k(Y)|X}[k(y)|x]$. We conclude that

**Fact 1:** $k(Y)$ is a sufficient statistic for $X$ iff

$$f_{Y|X}[y|x] = h(y,k(y))g(k(y),x).$$

(5)

Equivalently, if we define

$$\Lambda(y,x) = \frac{f_{Y|X}[y|x]}{f_{Y|X}[y|x_0]}$$

we find that $k(Y)$ is a sufficient statistic for $X$ iff

$$\Lambda(y,x) = \phi(k(y),x).$$

(6)

Let us revisit the detection problem in the light of this definition. Recall that

$$\text{MAP}[X|Y = y] = \arg\max_{x} P[X = x|Y = y].$$

Now, from $f_Y = f_{Y;k(Y)}$, we have

$$P[X = x|Y = y] = P[X = x|Y = y, k(Y) = k(y)] = \frac{f_{X,Y;k(Y)}[x,y,k(y)]}{f_{Y;k(Y)}[y,k(y)]} = \frac{f_{X,Y;k(Y)}[x,y,k(y)]}{f_{k(Y)}[y,k(y)]} = \frac{f_{X,Y}[x,y|k(y)]}{f_{Y}[y|k(y)]} = f_{X|k(Y)}[x|k(y)].$$

It is then clear that the maximizer is a function of $k(Y)$.

The connection with likelihood ratios is

$$\Lambda(x;y) := \frac{f_{Y|X}[y|x]}{f_{Y|X}[y|x_0]} = \frac{f_{k(Y)|X}[k(y)|x]}{f_{k(Y)|X}[k(y)|x_0]} := \Lambda(x;k(y)).$$

It follows that the solution of the HT problem, the MLE, and the MAP are all functions of the sufficient statistic.

**REFERENCES**
