Introduction to Detection Theory

Reading:

- Ch. 3 in Kay-II.
- Notes by Prof. Don Johnson on detection theory, see http://www.ece.rice.edu/~dhj/courses/elec531/notes5.pdf.
- Ch. 10 in Wasserman.

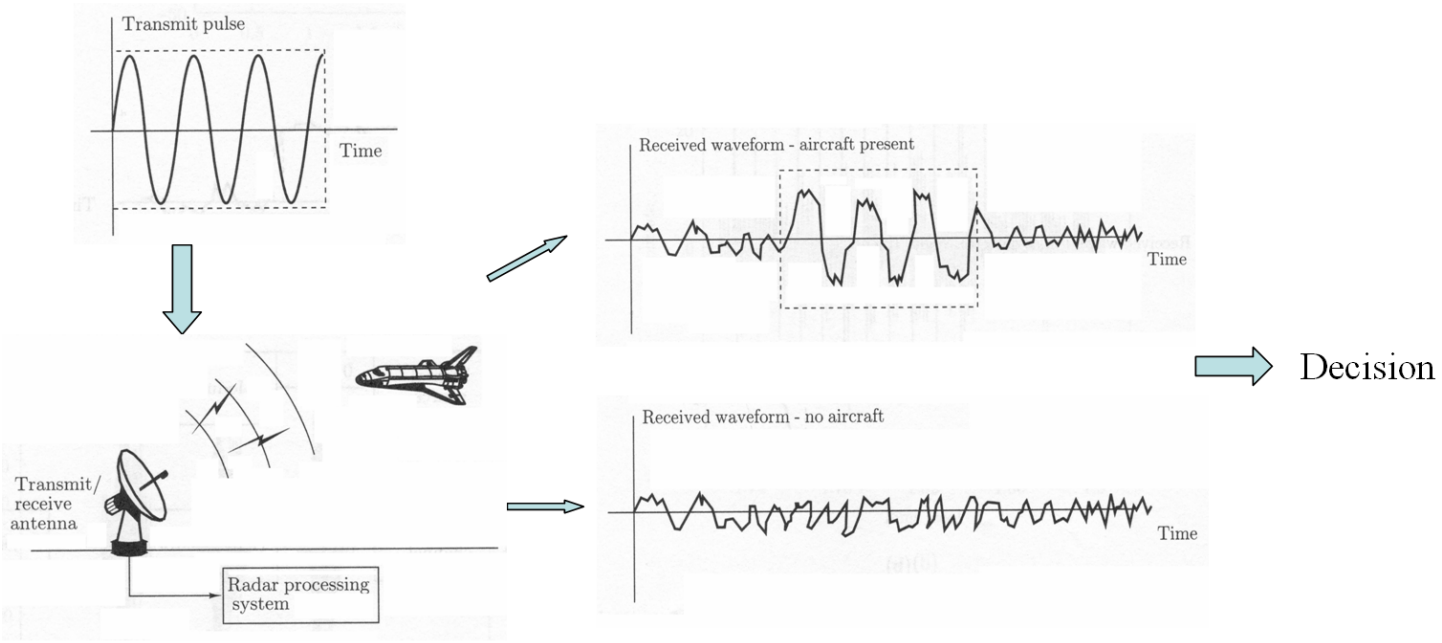
Introduction to Detection Theory (cont.)

We wish to make a decision on a signal of interest using noisy measurements. Statistical tools enable systematic solutions and optimal design.

Application areas include:

- Communications,
- Radar and sonar,
- Nondestructive evaluation (NDE) of materials,
- Biomedicine, etc.

Example: Radar Detection. We wish to decide on the presence or absence of a target.



Introduction to Detection Theory

We assume a parametric measurement model $p(x \,|\, \theta)$ [or $p(x\,;\,\theta)$, which is the notation that we sometimes use in the classical setting].

In point estimation theory, we estimated the parameter $\theta \in \Theta$ given the data x.

Suppose now that we choose Θ_0 and Θ_1 that form a *partition* of the parameter space Θ:

$$\Theta_0 \cup \Theta_1 = \Theta, \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

In detection theory, we wish to identify *which* hypothesis is true (i.e. make the *appropriate decision*):

$$\begin{aligned} \mathcal{H}_0 &: \theta \in \Theta_0, \quad \text{null hypothesis} \\ \mathcal{H}_1 &: \theta \in \Theta_1, \quad \text{alternative hypothesis}. \end{aligned}$$

Terminology: If θ can only take two values,

$$\Theta = \{\theta_0, \theta_1\}, \quad \Theta_0 = \{\theta_0\}, \quad \Theta_1 = \{\theta_1\}$$

we say that the hypotheses are *simple*. Otherwise, we say that they are *composite*.

Composite Hypothesis Example: $\mathcal{H}_0 : \theta = 0$ versus $\mathcal{H}_1 : \theta \in (0, \infty)$.

The Decision Rule

We wish to design a decision rule (function) $\phi(\boldsymbol{x}) : \mathcal{X} \to (0, 1)$:

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & \text{decide } \mathcal{H}_1, \\ 0, & \text{decide } \mathcal{H}_0. \end{cases}$$

which partitions the data space $\mathcal{X}$ [i.e. the support of $p(\boldsymbol{x} \,|\, \theta)$] into two regions:

$$\text{Rule } \phi(\boldsymbol{x}): \quad \mathcal{X}_0 = \{\boldsymbol{x} : \phi(\boldsymbol{x}) = 0\}, \quad \mathcal{X}_1 = \{\boldsymbol{x} : \phi(\boldsymbol{x}) = 1\}.$$

Let us define probabilities of false alarm and miss:

$$\begin{aligned} P_{\text{FA}} &= \mathrm{E}_{\boldsymbol{x} \,|\, \theta}[\phi(\boldsymbol{X}) \,|\, \theta] = \int_{\mathcal{X}_1} p(\boldsymbol{x} \,|\, \theta)\, dx \qquad \text{for } \theta \text{ in } \Theta_0 \\ P_{\text{M}} &= \mathrm{E}_{\boldsymbol{x} \,|\, \theta}[1 - \phi(\boldsymbol{X}) \,|\, \theta] = 1 - \int_{\mathcal{X}_1} p(\boldsymbol{x} \,|\, \theta)\, d\boldsymbol{x} \\ &= \int_{\mathcal{X}_0} p(\boldsymbol{x} \,|\, \theta)\, d\boldsymbol{x} \qquad \text{for } \theta \text{ in } \Theta_1. \end{aligned}$$

Then, the probability of detection (correctly deciding $\mathcal{H}_1$) is

$$P_{\text{D}} = 1 - P_{\text{M}} = \mathrm{E}_{\boldsymbol{x} \,|\, \theta}[\phi(\boldsymbol{X}) \,|\, \theta] = \int_{\mathcal{X}_1} p(\boldsymbol{x} \,|\, \theta)\, d\boldsymbol{x} \qquad \text{for } \theta \text{ in } \Theta_1.$$

Note: P_{FA} and $P_{\mathrm{D}}/P_{\mathrm{M}}$ are generally functions of the parameter θ (where $\theta \in \Theta_0$ when computing P_{FA} and $\theta \in \Theta_1$ when computing $P_{\mathrm{D}}/P_{\mathrm{M}}$).

More Terminology. Statisticians use the following terminology:

- False alarm $\equiv$ "Type I error"
- Miss $\equiv$ "Type II error"
- Probability of detection $\equiv$ "Power"
- Probability of false alarm $\equiv$ "Significance level."

Bayesian Decision-theoretic Detection Theory

Recall (a slightly generalized version of) the *posterior expected loss*:

$$\rho(\text{action} \,|\, \boldsymbol{x}) = \int_{\Theta} \mathrm{L}(\theta, \text{action})\, p(\theta \,|\, \boldsymbol{x})\, d\theta$$

that we introduced in handout # 4 when we discussed Bayesian decision theory. Let us now apply this theory to our easy example discussed here: *hypothesis testing*, where our action space consists of only two choices. We first assign a loss table:

decision rule $\phi \downarrow$	true state $\rightarrow$	Θ_1	Θ_0
$\boldsymbol{x} \in \mathcal{X}_1$		$\mathrm{L}(1\|1) = 0$	$\mathrm{L}(1\|0)$
$\boldsymbol{x} \in \mathcal{X}_0$		$\mathrm{L}(0\|1)$	$\mathrm{L}(0\|0) = 0$

with the loss function described by the quantities $\mathrm{L}(\text{declared} \,|\, \text{true})$:

- $\mathrm{L}(1 \,|\, 0)$ quantifies loss due to a false alarm,
- $\mathrm{L}(0 \,|\, 1)$ quantifies loss due to a miss,
- $\mathrm{L}(1 \,|\, 1)$ and $\mathrm{L}(0 \,|\, 0)$ (losses due to correct decisions) — typically set to zero in real life. Here, we adopt zero losses for correct decisions.

Now, our posterior expected loss takes two values:

$$\begin{aligned}
\rho_0(\boldsymbol{x}) &= \int_{\Theta_1} \mathrm{L}(0 \,|\, 1)\, p(\theta \,|\, \boldsymbol{x})\, d\theta \\
&\quad + \int_{\Theta_0} \underbrace{\mathrm{L}(0 \,|\, 0)}_{0}\, p(\theta \,|\, \boldsymbol{x})\, d\theta \\
&= \int_{\Theta_1} \mathrm{L}(0 \,|\, 1)\, p(\theta \,|\, \boldsymbol{x})\, d\theta \\
&\overset{\mathrm{L}(0 \,|\, 1) \text{ is constant}}{=} \mathrm{L}(0 \,|\, 1) \underbrace{\int_{\Theta_1} p(\theta \,|\, \boldsymbol{x})\, d\theta}_{P[\theta \in \Theta_1 \,|\, \boldsymbol{x}]}
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\rho_1(\boldsymbol{x}) &= \int_{\Theta_0} \mathrm{L}(1 \,|\, 0)\, p(\theta \,|\, \boldsymbol{x})\, d\theta \\
&\overset{\mathrm{L}(1 \,|\, 0) \text{ is constant}}{=} \mathrm{L}(1 \,|\, 0) \underbrace{\int_{\Theta_0} p(\theta \,|\, \boldsymbol{x})\, d\theta}_{P[\theta \in \Theta_0 \,|\, \boldsymbol{x}]} .
\end{aligned}$$

We define the *Bayes' decision rule* as the rule that *minimizes the posterior expected loss*; this rule corresponds to choosing the data-space partitioning as follows:

$$\mathcal{X}_1 = \{\boldsymbol{x} : \rho_1(\boldsymbol{x}) \leq \rho_0(\boldsymbol{x})\}$$

or

$$\mathcal{X}_1 = \left\{ \boldsymbol{x} \,:\, \frac{\overbrace{\int_{\Theta_1} p(\theta \,|\, \boldsymbol{x})\, d\theta}^{P[\theta \in \Theta_1 \,|\, \boldsymbol{x}]}}{\underbrace{\int_{\Theta_0} p(\theta \,|\, \boldsymbol{x})\, d\theta}_{P[\theta \in \Theta_0 \,|\, \boldsymbol{x}]}} \geq \frac{\mathrm{L}(1\,|\,0)}{\mathrm{L}(0\,|\,1)} \right\} \tag{1}$$

or, equivalently, upon applying the Bayes' rule:

$$\mathcal{X}_1 = \left\{ \boldsymbol{x} \,:\, \frac{\int_{\Theta_1} p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)\, d\theta}{\int_{\Theta_0} p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)\, d\theta} \geq \frac{\mathrm{L}(1\,|\,0)}{\mathrm{L}(0\,|\,1)} \right\}. \tag{2}$$

0-1 loss: For $\mathrm{L}(1|0) = \mathrm{L}(0|1) = 1$, we have

decision rule $\phi \downarrow$	true state $\rightarrow$	Θ_1	Θ_0
$\boldsymbol{x} \in \mathcal{X}_1$		$\mathrm{L}(1\|1) = 0$	$\mathrm{L}(1\|0) = 1$
$\boldsymbol{x} \in \mathcal{X}_0$		$\mathrm{L}(0\|1) = 1$	$\mathrm{L}(0\|0) = 0$

yielding the following Bayes' decision rule, called the maximum *a posteriori* (MAP) rule:

$$\mathcal{X}_1 = \left\{ \boldsymbol{x} \,:\, \frac{P[\theta \in \Theta_1 \,|\, \boldsymbol{x}]}{P[\theta \in \Theta_0 \,|\, \boldsymbol{x}]} \geq 1 \right\} \tag{3}$$

or, equivalently, upon applying the Bayes' rule:

$$\mathcal{X}_1 = \left\{ \boldsymbol{x} \,:\, \frac{\int_{\Theta_1} p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)\, d\theta}{\int_{\Theta_0} p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)\, d\theta} \geq 1 \right\}. \tag{4}$$

Simple hypotheses. Let us specialize (1) to the case of simple hypotheses ($\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$):

$$\mathcal{X}_1 = \left\{ \boldsymbol{x} : \underbrace{\frac{p(\theta_1 \,|\, \boldsymbol{x})}{p(\theta_0 \,|\, \boldsymbol{x})}}_{\text{posterior-odds ratio}} \geq \frac{\mathrm{L}(1 \,|\, 0)}{\mathrm{L}(0 \,|\, 1)} \right\}. \tag{5}$$

We can rewrite (5) using the Bayes' rule:

$$\mathcal{X}_1 = \left\{ \boldsymbol{x} : \underbrace{\frac{p(\boldsymbol{x} \,|\, \theta_1)}{p(\boldsymbol{x} \,|\, \theta_0)}}_{\text{likelihood ratio}} \geq \frac{\pi_0 \,\mathrm{L}(1 \,|\, 0)}{\pi_1 \,\mathrm{L}(0 \,|\, 1)} \right\} \tag{6}$$

where

$$\pi_0 = \pi(\theta_0), \quad \pi_1 = \pi(\theta_1) = 1 - \pi_0$$

describe the prior probability mass function (pmf) of the binary random variable θ (recall that $\theta \in \{\theta_0, \theta_1\}$). Hence, for binary simple hypotheses, the prior pmf of θ is the Bernoulli pmf.

Preposterior (Bayes) Risk

The preposterior (Bayes) risk for rule $\phi(\boldsymbol{x})$ is

$$\begin{aligned} \mathrm{E}_{x,\theta}[\text{loss}] &= \int_{\mathcal{X}_1} \int_{\Theta_0} \mathrm{L}(1 \mid 0)\, p(\boldsymbol{x} \mid \theta) \pi(\theta)\, d\theta\, d\boldsymbol{x} \\ &\quad + \int_{\mathcal{X}_0} \int_{\Theta_1} \mathrm{L}(0 \mid 1)\, p(\boldsymbol{x} \mid \theta) \pi(\theta)\, d\theta\, d\boldsymbol{x}. \end{aligned}$$

How do we choose the rule $\phi(\boldsymbol{x})$ that minimizes the preposterior

risk?

$$\int_{\mathcal{X}_1}\int_{\Theta_0} \mathrm{L}(1\,|\,0)\,p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

$$+\int_{\mathcal{X}_0}\int_{\Theta_1} \mathrm{L}(0\,|\,1)\,p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

$$=\int_{\mathcal{X}_1}\int_{\Theta_0} \mathrm{L}(1\,|\,0)\,p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

$$-\int_{\mathcal{X}_1}\int_{\Theta_1} \mathrm{L}(0\,|\,1)\,p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

$$+\int_{\mathcal{X}_0}\int_{\Theta_1} \mathrm{L}(0\,|\,1)\,p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

$$+\int_{\mathcal{X}_1}\int_{\Theta_1} \mathrm{L}(0\,|\,1)\,p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\,d\boldsymbol{x}$$

$$=\underbrace{\text{const}}_{\text{not dependent on } \boldsymbol{\phi}(\boldsymbol{x})}$$

$$+\int_{\mathcal{X}_1}\Big\{\mathrm{L}(1\,|\,0)\cdot\int_{\Theta_0} p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta$$

$$-\mathrm{L}(0\,|\,1)\cdot\int_{\Theta_1} p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta\Big\}\,d\boldsymbol{x}$$

implying that $\mathcal{X}_1$ should be chosen as

$$\Big\{\mathcal{X}_1 : \mathrm{L}(1\,|\,0)\cdot\int_{\Theta_0} p(\boldsymbol{x}\,|\,\theta)\pi(\theta)\,d\theta - \mathrm{L}(0\,|\,1)\cdot\int_{\Theta_1} p(\boldsymbol{x}\,|\,\theta)\pi(\theta) < 0\Big\}$$

which, as expected, is the same as (2), since

minimizing the posterior expected loss

$\Longleftrightarrow$ minimizing the preposterior risk for every $\boldsymbol{x}$

as showed earlier in handout # 4.

0-1 loss: For the 0-1 loss, i.e. $\mathrm{L}(1|0) = \mathrm{L}(0|1) = 1$, the preposterior (Bayes) risk for rule $\phi(\boldsymbol{x})$ is

$$\begin{aligned} \mathrm{E}_{\boldsymbol{x},\theta}[\text{loss}] &= \int_{\mathcal{X}_1} \int_{\Theta_0} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta\, d\boldsymbol{x} \\ &\quad + \int_{\mathcal{X}_0} \int_{\Theta_1} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta\, d\boldsymbol{x} \end{aligned} \tag{7}$$

which is simply the *average error probability*, with averaging performed over the joint probability density or mass function (pdf/pmf) or the data $\boldsymbol{x}$ and parameters θ.

Bayesian Decision-theoretic Detection for Simple Hypotheses

The Bayes' decision rule for simple hypotheses is (6):

$$\underbrace{\Lambda(\boldsymbol{x})}_{\text{likelihood ratio}} = \frac{p(\boldsymbol{x}\,|\,\theta_1)}{p(\boldsymbol{x}\,|\,\theta_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\pi_0\,\mathrm{L}(1|0)}{\pi_1\,\mathrm{L}(0|1)} \equiv \tau \tag{8}$$

see also Ch. 3.7 in Kay-II. (Recall that $\Lambda(x)$ is the sufficient statistic for the detection problem, see p. 37 in handout # 1.) Equivalently,

$$\log \Lambda(\boldsymbol{x}) = \log[p(\boldsymbol{x}\,|\,\theta_1)] - \log[p(\boldsymbol{x}\,|\,\theta_0)] \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \log \tau \equiv \tau'.$$

Minimum Average Error Probability Detection: In the familiar 0-1 loss case where $\mathrm{L}(1|0) = \mathrm{L}(0|1) = 1$, we know that the preposterior (Bayes) risk is equal to the *average error probability*, see (7). This average error probability greatly

simplifies in the simple hypothesis testing case:

$$\text{av. error probability} = \int_{\mathcal{X}_1} \underbrace{\mathrm{L}(1\,|\,0)}_{1}\, p(\boldsymbol{x}\,|\,\theta_0)\pi_0\, d\boldsymbol{x}$$

$$+ \int_{\mathcal{X}_0} \underbrace{\mathrm{L}(0\,|\,1)}_{1}\, p(\boldsymbol{x}\,|\,\theta_1)\pi_1\, d\boldsymbol{x}$$

$$= \quad \pi_0 \cdot \underbrace{\int_{\mathcal{X}_1} p(\boldsymbol{x}\,|\,\theta_0)\, d\boldsymbol{x}}_{P_{\mathrm{FA}}} + \pi_1 \cdot \underbrace{\int_{\mathcal{X}_0} p(\boldsymbol{x}\,|\,\theta_1)\, d\boldsymbol{x}}_{P_{\mathrm{M}}}$$

where, as before, the averaging is performed over the joint pdf/pmf of the data $\boldsymbol{x}$ and parameters θ, and

$$\pi_0 = \pi(\theta_0), \quad \pi_1 = \pi(\theta_1) = 1 - \pi_0 \quad \text{(the Bernoulli pmf)}.$$

In this case, our Bayes' decision rule simplifies to the MAP rule (as expected, see (5) and Ch. 3.6 in Kay-II):

$$\underbrace{\frac{p(\theta_1\,|\,\boldsymbol{x})}{p(\theta_0\,|\,\boldsymbol{x})}}_{\text{posterior-odds ratio}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} 1 \tag{9}$$

or, equivalently, upon applying the Bayes' rule:

$$\underbrace{\frac{p(\boldsymbol{x}\,|\,\theta_1)}{p(\boldsymbol{x}\,|\,\theta_0)}}_{\text{likelihood ratio}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\pi_0}{\pi_1}. \tag{10}$$

which is the same as

- (4), upon substituting the Bernoulli pmf as the prior pmf for θ and

- (8), upon substituting $\mathrm{L}(1|0) = \mathrm{L}(0|1) = 1$.

Bayesian Decision-theoretic Detection Theory: Handling Nuisance Parameters

We apply the same approach as before — integrate the nuisance parameters (φ, say) out!

Therefore, (1) still holds for testing

$$\begin{aligned} \mathcal{H}_0 &: \theta \in \Theta_0 \quad \text{versus} \\ \mathcal{H}_1 &: \theta \in \Theta_1 \end{aligned}$$

but $p_{\theta \,|\, \boldsymbol{x}}(\theta \,|\, \boldsymbol{x})$ is computed as follows:

$$p_{\theta \,|\, \boldsymbol{x}}(\theta \,|\, \boldsymbol{x}) = \int p_{\theta,\varphi \,|\, \boldsymbol{x}}(\theta, \varphi \,|\, \boldsymbol{x})\, d\varphi$$

and, therefore,

$$\frac{\int_{\Theta_1} \overbrace{\int p_{\theta,\varphi \,|\, \boldsymbol{x}}(\theta, \varphi \,|\, \boldsymbol{x})\, d\varphi}^{p(\theta \,|\, \boldsymbol{x})}\, d\theta}{\int_{\Theta_0} \underbrace{\int p_{\theta,\varphi \,|\, \boldsymbol{x}}(\theta, \varphi \,|\, \boldsymbol{x})\, d\varphi}_{p(\theta \,|\, \boldsymbol{x})}\, d\theta} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\mathrm{L}(1 \,|\, 0)}{\mathrm{L}(0 \,|\, 1)} \tag{11}$$

or, equivalently, upon applying the Bayes' rule:

$$\frac{\int_{\Theta_1} \int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta,\varphi)\,\pi_{\theta,\varphi}(\theta,\varphi)\,d\varphi\,d\theta}{\int_{\Theta_0} \int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta,\varphi)\,\pi_{\theta,\varphi}(\theta,\varphi)\,d\varphi\,d\theta} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\mathrm{L}(1\,|\,0)}{\mathrm{L}(0\,|\,1)}. \tag{12}$$

Now, if θ and φ are independent *a priori*, i.e.

$$\pi_{\theta,\varphi}(\theta,\varphi) = \pi_\theta(\theta) \cdot \pi_\varphi(\varphi) \tag{13}$$

then (12) can be rewritten as

$$\frac{\int_{\Theta_1} \pi_\theta(\theta) \overbrace{\int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta,\varphi)\,\pi_\varphi(\varphi)\,d\varphi}^{p(\boldsymbol{x}\,|\,\boldsymbol{\theta})}\,d\theta}{\int_{\Theta_0} \pi_\theta(\theta) \underbrace{\int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta,\varphi)\,\pi_\varphi(\varphi)\,d\varphi}_{p(\boldsymbol{x}\,|\,\boldsymbol{\theta})}\,d\theta} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\mathrm{L}(1\,|\,0)}{\mathrm{L}(0\,|\,1)}. \tag{14}$$

Simple hypotheses and independent priors for θ and φ: Let us specialize (11) to the simple hypotheses ($\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$):

$$\frac{\int p_{\theta,\varphi\,|\,\boldsymbol{x}}(\theta_1,\varphi\,|\,\boldsymbol{x})\,d\varphi}{\int p_{\theta,\varphi\,|\,\boldsymbol{x}}(\theta_0,\varphi\,|\,\boldsymbol{x})\,d\varphi} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\mathrm{L}(1\,|\,0)}{\mathrm{L}(0\,|\,1)}. \tag{15}$$

Now, if θ and φ are independent *a priori*, i.e. (13) holds, then

we can rewrite (14) [or (15) using the Bayes' rule]:

$$\underbrace{\frac{\int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta_1,\varphi)\,\pi_\varphi(\varphi)\,d\varphi}{\int p_{\boldsymbol{x}\,|\,\theta,\varphi}(\boldsymbol{x}\,|\,\theta_0,\varphi)\,\pi_\varphi(\varphi)\,d\varphi}}_{\text{integrated likelihood ratio}} = \overbrace{\frac{p(\boldsymbol{x}\,|\,\theta_1)}{p(\boldsymbol{x}\,|\,\theta_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\pi_0\,\mathrm{L}(1\,|\,0)}{\pi_1\,\mathrm{L}(0\,|\,1)}}^{\text{same as (6)}} \tag{16}$$

where

$$\pi_0 = \pi_\theta(\theta_0), \quad \pi_1 = \pi_\theta(\theta_1) = 1 - \pi_0.$$

Chernoff Bound on Average Error Probability for Simple Hypotheses

Recall that minimizing the average error probability

$$\text{av. error probability} = \int_{\mathcal{X}_1} \int_{\Theta_0} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta\, d\boldsymbol{x} + \int_{\mathcal{X}_0} \int_{\Theta_1} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta\, d\boldsymbol{x}$$

leads to the MAP decision rule:

$$\mathcal{X}_1^\star = \left\{ \boldsymbol{x} : \int_{\Theta_0} p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)\, d\theta - \int_{\Theta_1} p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)\, d\theta < 0 \right\}.$$

In many applications, we *may not be able to obtain* a simple closed-form expression for the minimum average error

probability, but we *can bound it* as follows:

$$\text{min av. error probability} = \int_{\mathcal{X}_1^\star} \int_{\Theta_0} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta\, d\boldsymbol{x}$$

$$+ \int_{\mathcal{X}_0^\star} \int_{\Theta_1} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta\, d\boldsymbol{x}$$

$$\overset{\text{using the def. of } \mathcal{X}_1^\star}{=} \int_{\underbrace{\mathcal{X}}_{\text{data space}}} \min\left\{ \int_{\Theta_0} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta, \int_{\Theta_1} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta) \right\} d\boldsymbol{x}$$

$$\leq \int_{\mathcal{X}} \Big[\overbrace{\int_{\Theta_0} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta}^{\triangleq\, q_0(\boldsymbol{x})} \Big]^{\lambda} \Big[\overbrace{\int_{\Theta_1} p(\boldsymbol{x} \,|\, \theta)\pi(\theta)\, d\theta}^{\triangleq\, q_1(\boldsymbol{x})} \Big]^{1-\lambda} d\boldsymbol{x}$$

$$= \int_{\mathcal{X}} [q_0(\boldsymbol{x})]^{\lambda}\, [q_1(\boldsymbol{x})]^{1-\lambda}\, d\boldsymbol{x}$$

which is the *Chernoff bound on the minimum average error probability*. Here, we have used the fact that

$$\min\{a, b\} \leq a^{\lambda}\, b^{1-\lambda}, \quad \text{for} \quad 0 \leq \lambda \leq 1, \quad a, b \geq 0.$$

When

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

with

- $x_1, x_2, \ldots x_N$ conditionally independent, identically distributed (i.i.d.) given θ and
- simple hypotheses (i.e. $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$)

then

$$q_0(\boldsymbol{x}) = p(\boldsymbol{x} \,|\, \theta_0) \cdot \pi_0 = \pi_0 \prod_{n=1}^{N} p(x_n \,|\, \theta_0)$$

$$q_1(\boldsymbol{x}) = p(\boldsymbol{x} \,|\, \theta_1) \cdot \pi_1 = \pi_1 \prod_{n=1}^{N} p(x_n \,|\, \theta_1)$$

yielding

Chernoff bound for N conditionally i.i.d. measurements (given θ) and simple hyp.

$$= \int \left[\pi_0 \prod_{n=1}^{N} p(x_n \,|\, \theta_0)\right]^{\lambda} \left[\pi_1 \prod_{n=1}^{N} p(x_n \,|\, \theta_1)\right]^{1-\lambda} d\boldsymbol{x}$$

$$= \pi_0^{\lambda}\, \pi_1^{1-\lambda} \cdot \prod_{n=1}^{N} \left\{ \int [p(x_n \,|\, \theta_0)]^{\lambda} \, [p(x_n \,|\, \theta_1)]^{1-\lambda} \, dx_n \right\}$$

$$= \pi_0^{\lambda}\, \pi_1^{1-\lambda} \cdot \left\{ \int [p(x_1 \,|\, \theta_0)]^{\lambda} \, [p(x_1 \,|\, \theta_1)]^{1-\lambda} \, dx_1 \right\}^{N}$$

or, in other words,

$$\frac{1}{N} \cdot \log(\text{min av. error probability}) \leq \log(\pi_0^{\lambda}\, \pi_1^{1-\lambda})$$
$$+ \log \int [p(x_1 \,|\, \theta_0)]^{\lambda}\, [p(x_1 \,|\, \theta_1)]^{1-\lambda}\, dx_1, \quad \forall \lambda \in [0,1].$$

If $\pi_0 = \pi_1 = 1/2$ (which is almost always the case of interest when evaluating average error probabilities), we can say that, as $N \to \infty$,

$$\underset{\text{for } N \text{ cond. i.i.d. measurements (given } \theta) \text{ and simple hypotheses}}{\text{min av. error probability}} \to$$

$$f(N) \cdot \exp\left(- N \cdot \underbrace{\left\{ - \min_{\lambda \in [0,1]} \log \int [p(x_1 \,|\, \theta_0)]^{\lambda}\, [p(x_1 \,|\, \theta_1)]^{1-\lambda} \right\}}_{\text{Chernoff information for a single observation}} \right)$$

where $f(N)$ is a slowly-varying function compared with the exponential term:

$$\lim_{N \to \infty} \frac{\log f(N)}{N} = 0.$$

Note that the Chernoff information in the exponent term of the above expression *quantifies* the asymptotic behavior of the minimum average error probability.

We now give a useful result, taken from

K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., San Diego, CA: Academic Press, 1990

for evaluating a class of Chernoff bounds.

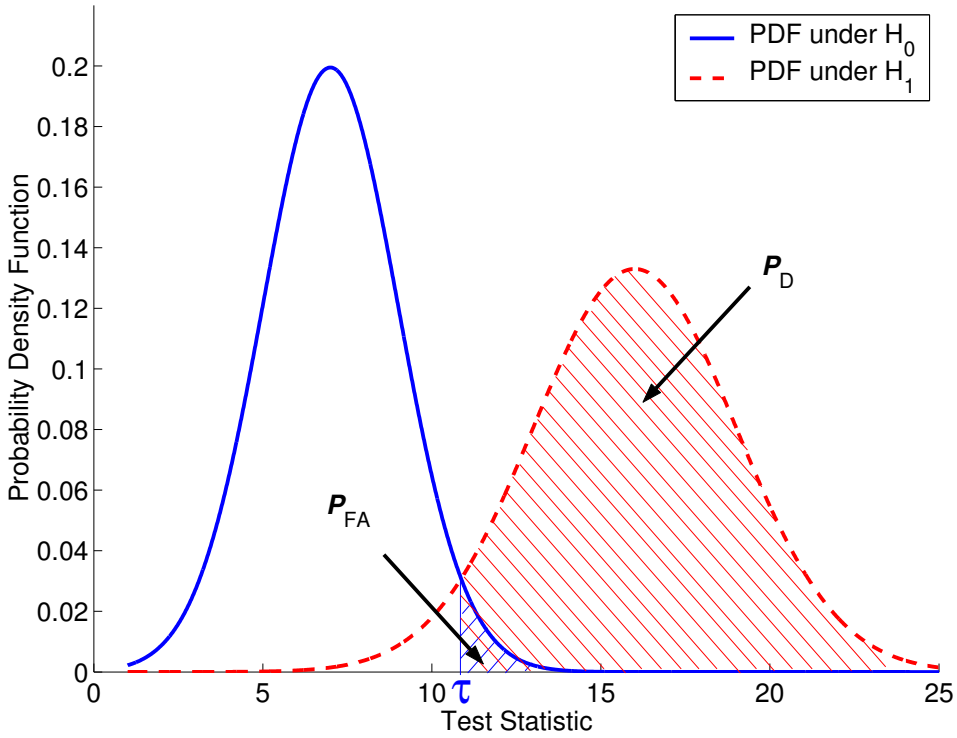
Lemma 1. *Consider* $p_1(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ *and* $p_2(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$. *Then*

$$\int [p_1(\boldsymbol{x})]^{\lambda} \cdot [p_2(\boldsymbol{x})]^{1-\lambda} \, d\boldsymbol{x} = \exp[-g(\lambda)]$$

where

$$\begin{aligned} g(\lambda) &= \frac{\lambda\,(1-\lambda)}{2} \cdot (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T [\lambda\,\Sigma_1 + (1-\lambda)\,\Sigma_2]^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ &\quad + \tfrac{1}{2} \log \left[\frac{|\lambda\,\Sigma_1 + (1-\lambda)\,\Sigma_2|}{|\Sigma_1|^{\lambda} \cdot |\Sigma_2|^{1-\lambda}} \right]. \end{aligned}$$

Probabilities of False Alarm (P_{FA}) and Detection (P_{D}) for Simple Hypotheses



$$P_{\mathrm{FA}} = P[\overbrace{\underbrace{\text{test statistic} > \tau}_{\boldsymbol{X} \in \mathcal{X}_1}}^{\Lambda(x)} \,|\, \theta = \theta_0]$$

$$P_{\mathrm{D}} = P[\underbrace{\text{test statistic} > \tau}_{\boldsymbol{X} \in \mathcal{X}_0} \,|\, \theta = \theta_1].$$

Comments:

(i) As the region $\mathcal{X}_1$ shrinks (i.e. $\tau \nearrow \infty$), both of the above probabilities shrink towards zero.

(ii) As the region $\mathcal{X}_1$ grows (i.e. $\tau \searrow 0$), both of these probabilities grow towards unity.

(iii) Observations **(i)** and **(ii)** *do not imply equality* between P_{FA} and P_{D}; in most cases, as $\mathcal{R}_1$ grows, P_{D} grows more rapidly than P_{FA} (i.e. we better be right more often than we are wrong).

(iv) However, the *perfect case* where our rule is always right and never wrong ($P_{\mathrm{D}} = 1$ and $P_{\mathrm{FA}} = 0$) cannot occur when the conditional pdfs/pmfs $p(\boldsymbol{x} \,|\, \theta_0)$ and $p(\boldsymbol{x} \,|\, \theta_1)$ overlap.

(v) Thus, to increase the detection probability P_{D}, we must also allow for the false-alarm probability P_{FA} to increase. This behavior

- represents the fundamental tradeoff in hypothesis testing and detection theory and
- motivates us to introduce a (classical) approach to testing simple hypotheses, pioneered by Neyman and Pearson (to be discussed next).

Neyman-Pearson Test for Simple Hypotheses

Setup:

- Parametric data models $p(\boldsymbol{x}\,;\,\theta_0)$, $p(\boldsymbol{x}\,;\,\theta_1)$,
- Simple hypothesis testing:

$$\begin{aligned} \mathcal{H}_0 &: \quad \theta = \theta_0 \quad \text{versus} \\ \mathcal{H}_1 &: \quad \theta = \theta_1. \end{aligned}$$

- No prior pdf/pmf on θ is available.

Goal: Design a test that maximizes the probability of detection

$$P_{\mathrm{D}} = P[\boldsymbol{X} \in \mathcal{X}_1\,;\,\theta = \theta_0]$$

(equivalently, minimizes the miss probability P_{M}) under the constraint

$$P_{\mathrm{FA}} = P[\boldsymbol{X} \in \mathcal{X}_1\,;\,\theta = \theta_0] = \alpha' \leq \alpha.$$

Here, we consider simple hypotheses; classical version of testing composite hypotheses is much more complicated. The Bayesian version of testing composite hypotheses is trivial (as

is everything else Bayesian, at least conceptually) and we have already seen it.

Solution. We apply the Lagrange-multiplier approach: maximize

$$\begin{aligned} L &= P_{\mathrm{D}} + \lambda \cdot (P_{\mathrm{FA}} - \alpha') \\ &= \int_{\mathcal{X}_1} p(\boldsymbol{x}\,;\,\theta_1)\,d\boldsymbol{x} + \lambda \cdot \Big[\int_{\mathcal{X}_1} p(\boldsymbol{x};\theta_0)\,d\boldsymbol{x} - \alpha' \Big] \\ &= \int_{\mathcal{X}_1} [p(\boldsymbol{x}\,;\,\theta_1) - \lambda\, p(\boldsymbol{x}\,;\,\theta_0)]\,d\boldsymbol{x} - \lambda \cdot \alpha'. \end{aligned}$$

To maximize L, set

$$\mathcal{X}_1 = \{\boldsymbol{x} : p(\boldsymbol{x}\,;\,\theta_1) - \lambda \cdot p(\boldsymbol{x}\,;\,\theta_0) > 0\} = \left\{\boldsymbol{x} \,:\, \frac{p(\boldsymbol{x}\,;\,\theta_1)}{p(\boldsymbol{x}\,;\,\theta_0)} > \lambda \right\}.$$

Again, we find the likelihood ratio:

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}\,;\,\theta_1)}{p(\boldsymbol{x}\,;\,\theta_0)}.$$

Recall our constraint:

$$\int_{\mathcal{X}_1} p(\boldsymbol{x}\,;\,\theta_0)\,d\boldsymbol{x} = P_{\mathrm{FA}} = \alpha' \leq \alpha.$$

If we increase λ, $P_{\rm FA}$ and $P_{\rm D}$ go down. Similarly, if we decrease λ, $P_{\rm FA}$ and $P_{\rm D}$ go up. Hence, to maximize $P_{\rm D}$, choose λ so that $P_{\rm FA}$ is as big as possible under the constraint.

Two useful ways for determining the threshold that achieves a specified false-alarm rate:

- Find λ that satisfies

$$\int_{\boldsymbol{x}\,:\,\Lambda(\boldsymbol{x})>\lambda} p(\boldsymbol{x}\,;\,\theta_0)\,d\boldsymbol{x} = P_{\rm FA} = \alpha$$

  or,

- expressing in terms of the pdf/pmf of $\Lambda(\boldsymbol{x})$ under $\mathcal{H}_0$:

$$\int_{\lambda}^{\infty} p_{\Lambda;\theta_0}(l\,;\,\theta_0)\,dl = \alpha.$$

  or, perhaps, in terms of a monotonic function of $\Lambda(\boldsymbol{x})$, say $T(\boldsymbol{x}) = \text{monotonic function}(\Lambda(\boldsymbol{x}))$.

Warning: We have been implicitly assuming that $P_{\rm FA}$ is a continuous function of λ. Some (not insightful) technical adjustments are needed if this is not the case.

A way of handling nuisance parameters: We can utilize the integrated (marginal) likelihood ratio (16) under the Neyman-Pearson setup as well.

Chernoff-Stein Lemma for Bounding the Miss Probability in Neyman-Pearson Tests of Simple Hypotheses

Recall the definition of the *Kullback-Leibler (K-L) distance* $D(\boldsymbol{p} \,\|\, \boldsymbol{q})$ from one pmf $(\boldsymbol{p})$ to another $(\boldsymbol{p})$:

$$D(\boldsymbol{p} \,\|\, \boldsymbol{q}) = \sum_k p_k \log \frac{p_k}{q_k}.$$

The complete proof of this lemma for the discrete (pmf) case is given in

Additional Reading: T.M. Cover and J.A. Thomas, *Elements of Information Theory*. Second ed., New York: Wiley, 2006.

Setup for the Chernoff-Stein Lemma

- Assume that $x_1, x_2, \ldots, x_N$ are conditionally i.i.d. given θ.

- We adopt the Neyman-Pearson framework, i.e. obtain a decision threshold to achieve a fixed P_{FA}. Let us study the asymptotic $P_{\mathrm{M}} = 1 - P_{\mathrm{D}}$ as the number of observations N gets large.

- To keep P_{FA} constant as N increases, we need to make our decision threshold (γ, say) vary with N, i.e.

$$\gamma = \gamma_N(P_{\mathrm{FA}})$$

  Now, the miss probability is

$$P_{\mathrm{M}} = P_{\mathrm{M}}(\gamma) = P_{\mathrm{M}}\Big(\gamma_N(P_{\mathrm{FA}})\Big).$$

Chernoff-Stein Lemma

The Chernoff-Stein lemma says:

$$\lim_{P_{\mathrm{FA}} \to 0} \lim_{N \to \infty} \frac{1}{N} \log P_{\mathrm{M}} = - \underbrace{D\Big(p(X_n \,|\, \theta_0) \,\|\, p(X_n \,|\, \theta_1)\Big)}_{\text{K-L distance for a single observation}}$$

where the K-L distance between $p(x_n \,|\, \theta_0)$ and $p(x_n \,|\, \theta_1)$

$$D\Big(p(X_n \,|\, \theta_0) \,\|\, p(X_n \,|\, \theta_1)\Big) = -\mathrm{E}_{\,p(x_n \,|\, \theta_0)} \left[\log \frac{p(X_n \,|\, \theta_1)}{p(X_n \,|\, \theta_0)} \right]$$

$$\overset{\text{discrete (pmf) case}}{=} - \sum_{x_n} p(x_n \,|\, \theta_0) \log \left[\frac{p(x_n \,|\, \theta_1)}{p(x_n \,|\, \theta_0)} \right]$$

does not depend on the observation index n, since x_n are conditionally i.i.d. given θ.

Equivalently, we can state that

$$P_{\mathrm{M}} \overset{N \to +\infty}{\longrightarrow} f(N) \cdot \exp \Big[- N \cdot D\big(p(X_n \,|\, \theta_0) \,\|\, p(X_n \,|\, \theta_1)\big) \Big]$$

as $P_{\mathrm{FA}} \to 0$ and $N \to \infty$, where $f(N)$ is a slowly-varying function compared with the exponential term (when $P_{\mathrm{FA}} \to 0$ and $N \to \infty$).

Detection for Simple Hypotheses: Example

Known positive DC level in additive white Gaussian noise (AWGN), Example 3.2 in Kay-II.

Consider

$$\begin{aligned} \mathcal{H}_0: &\quad x[n] = w[n], \quad n = 1, 2, \ldots, N \quad \text{versus} \\ \mathcal{H}_1: &\quad x[n] = A + w[n], \quad n = 1, 2, \ldots, N \end{aligned}$$

where

- $A > 0$ is a known constant,
- $w[n]$ is zero-mean white Gaussian noise with known variance σ^2, i.e.

$$w[n] \sim \mathcal{N}(0, \sigma^2).$$

The above hypothesis-testing formulation is EE-like: noise versus signal plus noise. **It is similar to the on-off keying scheme in communications**, which gives us an idea to rephrase it so that it fits our formulation on p. 4 (for which we have developed all the theory so far). Here is such an

alternative formulation: consider a family of pdfs

$$p(\boldsymbol{x}\,;\,a) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \cdot \exp\Big[-\frac{1}{2\,\sigma^2}\sum_{n=1}^{N}(x[n]-a)^2\Big] \qquad (17)$$

and the following (equivalent) hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad a = 0 \quad \text{(off)} \quad \text{versus} \\ \mathcal{H}_1: &\quad a = A \quad \text{(on)}. \end{aligned}$$

Then, the likelihood ratio is

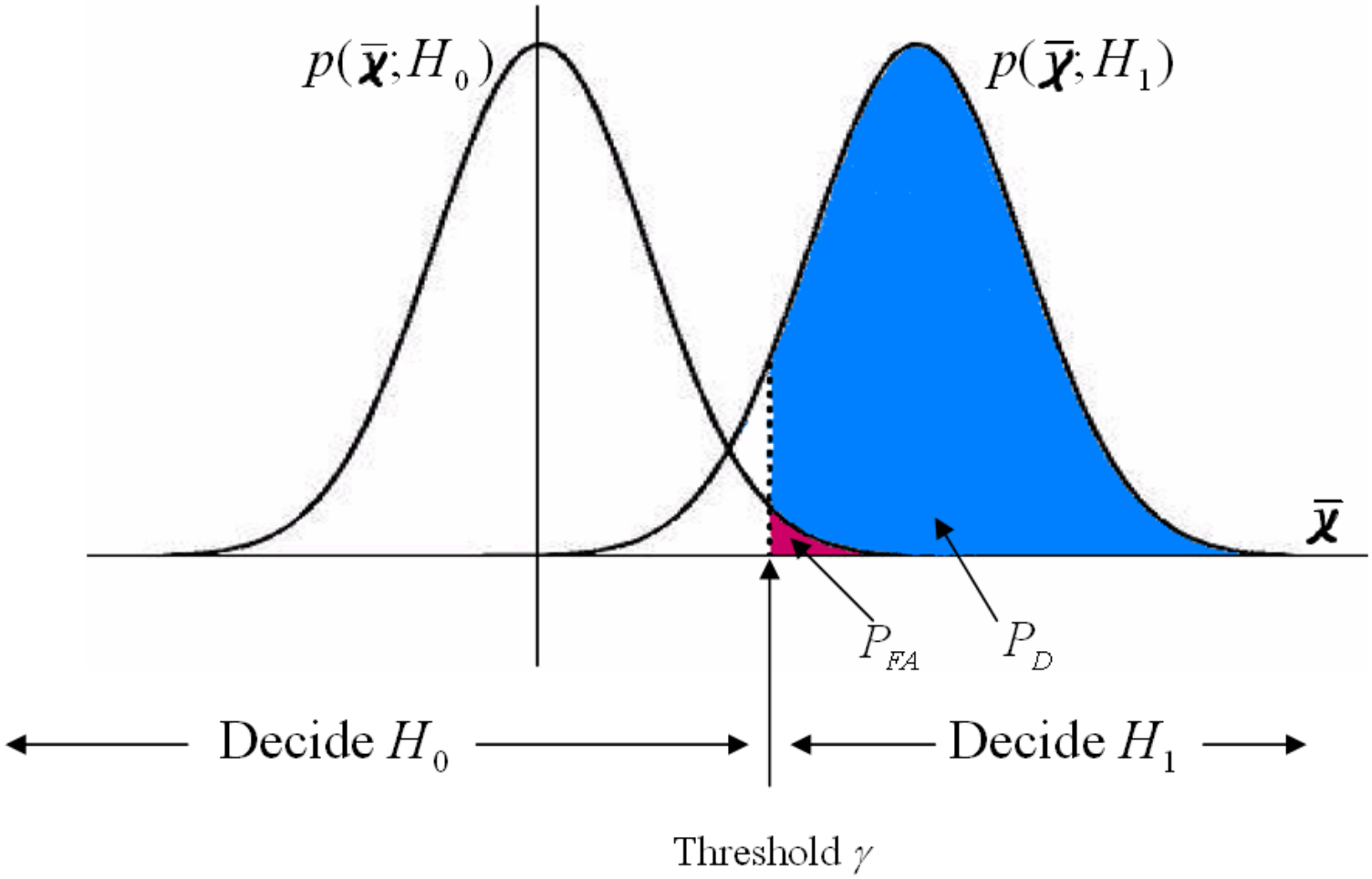
$$\begin{aligned} \Lambda(\boldsymbol{x}) &= \frac{p(\boldsymbol{x}\,;\,a=A)}{p(\boldsymbol{x}\,;\,a=0)} \\ &= \frac{1/(2\pi\sigma^2)^{N/2}\cdot\exp[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x[n]-A)^2]}{1/(2\pi\sigma^2)^{N/2}\cdot\exp(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}x[n]^2)}. \end{aligned}$$

Now, take the logarithm and, after simple manipulations, reduce our likelihood-ratio test to comparing

$$T(\boldsymbol{x}) = \overline{x} \stackrel{\triangle}{=} \frac{1}{N}\sum_{n=1}^{N}x[n]$$

with a threshold γ. [Here, $T(\boldsymbol{x})$ is a monotonic function of $\Lambda(x)$.] If $T(\boldsymbol{x}) > \gamma$ accept $\mathcal{H}_1$ (i.e. reject $\mathcal{H}_0$), otherwise accept

$\mathcal{H}_0$ (well, not exactly, we will talk more about this decision on p. 59).



The choice of γ depends on the approach that we take. For the Bayes' decision rule, γ is a function of π_0 and π_1. For the Neyman-Pearson test, γ is chosen to achieve (control) a desired P_{FA}.

Bayesian decision-theoretic detection for 0-1 loss (corresponding to minimizing the average error

probability):

$$\log \Lambda(\boldsymbol{x}) = -\frac{1}{2\,\sigma^2}\sum_{n=1}^{N}(x[n]-A)^2 + \frac{1}{2\,\sigma^2}\sum_{n=1}^{N}(x[n])^2 \overset{\mathcal{H}_1}{\gtrless} \log\left(\frac{\pi_0}{\pi_1}\right)$$

$$\Longleftrightarrow \frac{1}{2\,\sigma^2}\sum_{n=1}^{N}(\underbrace{x[n]-x[n]}_{0}+A)\,(x[n]+x[n]-A) \overset{\mathcal{H}_1}{\gtrless} \log\left(\frac{\pi_0}{\pi_1}\right)$$

$$\Longleftrightarrow 2\,A\cdot\left(\sum_{n=1}^{N}x[n]\right) - A^2\,N \overset{\mathcal{H}_1}{\gtrless} 2\,\sigma^2\,\log\left(\frac{\pi_0}{\pi_1}\right)$$

$$\Longleftrightarrow \left(\sum_{n=1}^{N}x[n]\right) - \frac{AN}{2} \overset{\mathcal{H}_1}{\gtrless} \frac{\sigma^2}{A}\log\left(\frac{\pi_0}{\pi_1}\right) \quad \text{since } A>0$$

and, finally,

$$\overline{x} \overset{\mathcal{H}_1}{\gtrless} \underbrace{\frac{\sigma^2}{N\,A}\cdot\log(\pi_0/\pi_1)+\frac{A}{2}}_{\gamma}$$

which, for the practically most interesting case of equiprobable hypotheses

$$\pi_0 = \pi_1 = \tfrac{1}{2} \tag{18}$$

simplifies to

$$\overline{x} \overset{\mathcal{H}_1}{\gtrless} \underbrace{\frac{A}{2}}_{\gamma}$$

known as the *maximum-likelihood test* (i.e. the Bayes' decision rule for 0-1 loss and *a priori* equiprobable hypotheses is defined as the maximum-likelihood test). This maximum-likelihood detector *does not* require the knowledge of the noise variance

σ^2 to declare its decision. However, the knowledge of σ^2 is *key* to assessing the detection performance. Interestingly, these observations will carry over to a few maximum-likelihood tests that we will derive in the future.

Assuming (18), we now derive the minimum average error probability. First, note that $\overline{X}\,|\,a=0 \sim \mathcal{N}(0,\sigma^2/N)$ and $\overline{X}\,|\,a=A \sim \mathcal{N}(A,\sigma^2/N)$. Then

$$\text{min av. error prob.} = \tfrac{1}{2}\underbrace{P[\overline{X} > \frac{A}{2}\,|\,a=0]}_{P_{\rm FA}} + \tfrac{1}{2}\underbrace{P[\overline{X} < \frac{A}{2}\,|\,a=A]}_{P_{\rm M}}$$

$$= \tfrac{1}{2}P\Big[\underbrace{\frac{\overline{X}}{\sqrt{\sigma^2/N}}}_{\substack{\text{standard}\\ \text{normal}\\ \text{random var.}}} > \frac{A/2}{\sqrt{\sigma^2/N}}\,;\, a=0\Big]$$

$$+\tfrac{1}{2}P\Big[\underbrace{\frac{\overline{X}-A}{\sqrt{\sigma^2/N}}}_{\substack{\text{standard}\\ \text{normal}\\ \text{random var.}}} < \frac{A/2-A}{\sqrt{\sigma^2/N}}\,;\, a=A\Big]$$

$$= \tfrac{1}{2}\underbrace{Q\Big(\sqrt{\frac{N\,A^2}{4\,\sigma^2}}\Big)}_{\substack{\text{standard}\\ \text{normal complementary cdf}}} + \tfrac{1}{2}\underbrace{\Phi\Big(-\sqrt{\frac{N\,A^2}{4\,\sigma^2}}\Big)}_{\substack{\text{standard}\\ \text{normal cdf}}} = Q\Big(\tfrac{1}{2}\sqrt{\frac{N\,A^2}{\sigma^2}}\Big)$$

since

$$\Phi(-x) = Q(x).$$

The minimum probability of error decreases monotonically with $N\,A^2/\sigma^2$, which is known as the *deflection coefficient*. In this case, the Chernoff information is equal to $\frac{A^2}{8\,\sigma^2}$ (see Lemma 1):

$$\min_{\lambda\in[0,1]} \log\left\{ \int \underbrace{[\mathcal{N}(0,\sigma^2/N)]^{\lambda}}_{p_0(x)}\, \underbrace{[\mathcal{N}(A,\sigma^2/N)]^{1-\lambda}}_{p_1(x)}\, dx \right\}$$

$$= \min_{\lambda\in[0,1]} \left[\frac{\lambda\,(1-\lambda)}{2}\cdot\frac{A^2}{\sigma^2}\right] = \frac{A^2}{8\,\sigma^2}$$

and, therefore, we expect the following asymptotic behavior:

$$\text{min av. error probability} \overset{N\to+\infty}{\longrightarrow} f(N)\cdot\exp\left(-N\cdot\frac{A^2}{8\,\sigma^2}\right) \quad (19)$$

where $f(N)$ varies slowly with N when N is large, compared with the exponential term:

$$\lim_{N\to\infty}\frac{\log f(N)}{N} = 0.$$

Indeed, in our example,

$$Q\left(\sqrt{\frac{N\,A^2}{4\,\sigma^2}}\right) \overset{N\to+\infty}{\longrightarrow} \underbrace{\frac{1}{\sqrt{\frac{N\,A^2}{4\,\sigma^2}\cdot\sqrt{2\pi}}}}_{f(N)}\cdot\exp\left(-\frac{N\,A^2}{8\,\sigma^2}\right)$$

where we have used the asymptotic formula

$$Q(x) \overset{x\to+\infty}{\longrightarrow} \frac{1}{x\sqrt{2\pi}} \cdot \exp(-x^2/2) \tag{20}$$

given e.g. in equation (2.4) of Kay-II.

Known DC Level in AWGN: Neyman-Pearson Approach

We now derive the Neyman-Pearson test for detecting a known DC level in AWGN. Recall that our likelihood-ratio test compares

$$T(\boldsymbol{x}) = \overline{x} \stackrel{\triangle}{=} \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

with a threshold γ.

- If $T(\boldsymbol{x}) > \gamma$, decide $\mathcal{H}_1$ (i.e. reject $\mathcal{H}_0$),
- otherwise decide $\mathcal{H}_0$ (see also the discussion on p. 59).

Performance evaluation: Assuming (17), we have

$$T(\boldsymbol{x}) \,|\, a \sim \mathcal{N}(a, \sigma^2/N).$$

Therefore, $T(\boldsymbol{X}) \,|\, a = 0 \sim \mathcal{N}(0, \sigma^2/N)$, implying

$$P_{\text{FA}} \quad = \quad P[T(\boldsymbol{X}) > \gamma \,;\, a = 0] = Q\Big(\frac{\gamma}{\sqrt{\sigma^2/N}}\Big)$$

and we obtain the decision threshold as follows:

$$\gamma = \sqrt{\frac{\sigma^2}{N}} \cdot Q^{-1}(P_{\mathrm{FA}}).$$

Now, $T(\boldsymbol{X}) \,|\, a = A \sim \mathcal{N}(A, \sigma^2/N)$, implying

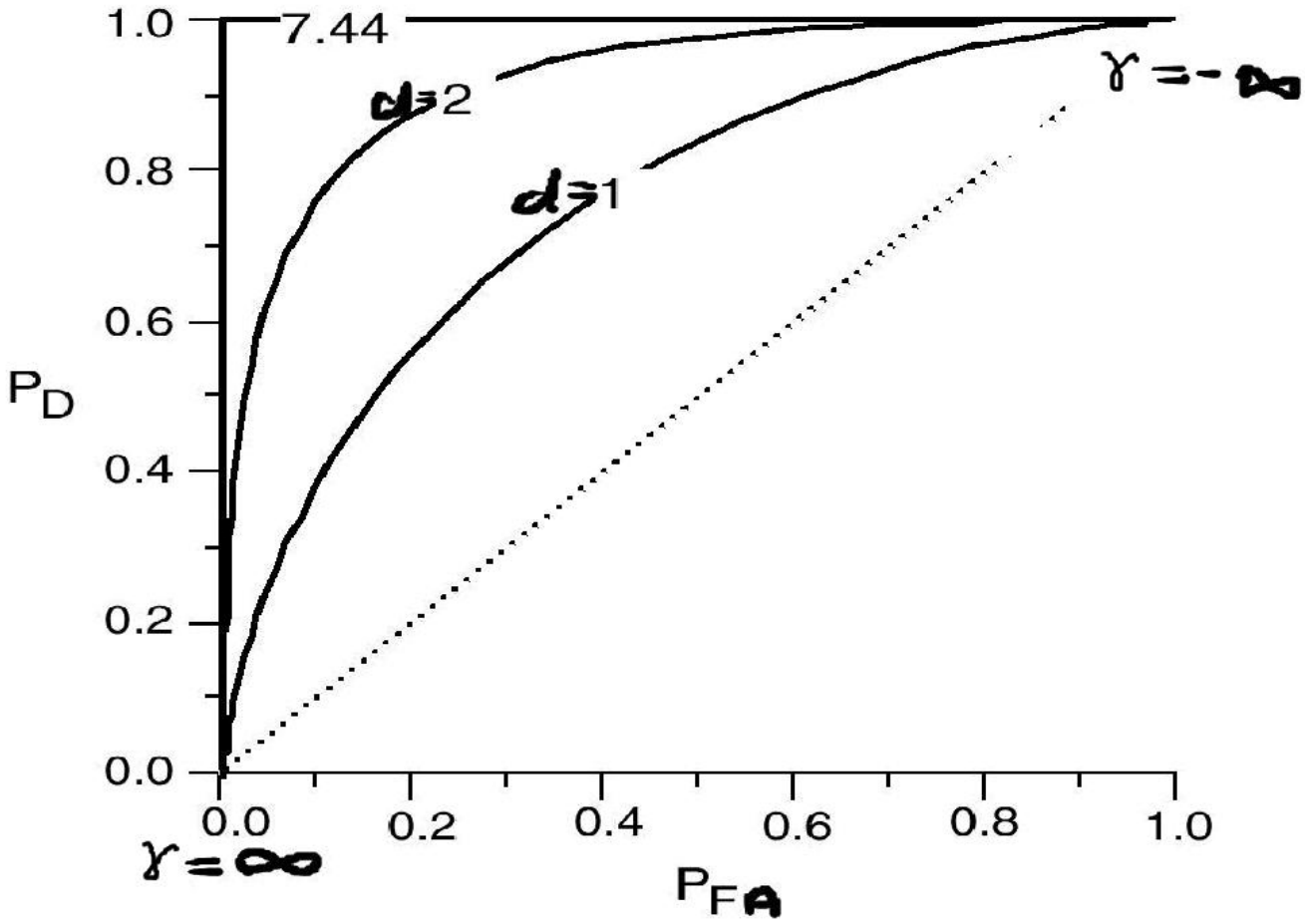
$$\begin{aligned} P_{\mathrm{D}} &= P(T(\boldsymbol{X}) > \gamma \,|\, a = A) = Q\Big(\frac{\gamma - A}{\sqrt{\sigma^2/N}}\Big) \\ &= Q\left(Q^{-1}(P_{\mathrm{FA}}) - \sqrt{\frac{A^2}{\sigma^2/N}}\right) \\ &= Q\left(Q^{-1}(P_{\mathrm{FA}}) - \sqrt{\underbrace{NA^2/\sigma^2}_{\triangleq\ \mathrm{SNR} = d^2}}\right). \end{aligned}$$

Given the false-alarm probability P_{FA}, the detection probability P_{D} depends only on the *deflection coefficient*:

$$d^2 = \frac{NA^2}{\sigma^2} = \frac{\{\mathrm{E}\,[T(\boldsymbol{X}) \,|\, a = A] - \mathrm{E}\,[T(\boldsymbol{X}) \,|\, a = 0]\}^2}{\mathrm{var}[T(\boldsymbol{X} \,|\, a = 0)]}$$

which is also (a reasonable definition for) the signal-to-noise ratio (SNR).

Receiver Operating Characteristics (ROC)



$$P_{\mathrm{D}} = Q\big(Q^{-1}(P_{\mathrm{FA}}) - d\big).$$

Comments:

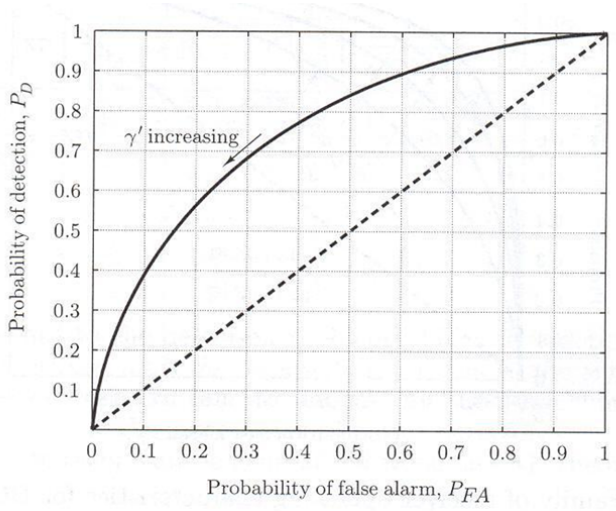
- As we raise the threshold γ, P_{FA} goes down but so does P_{D}.
- ROC should be above the 45° line — otherwise we can do better by flipping a coin.
- Performance improves with d^2.

Typical Ways of Depicting the Detection Performance Under the Neyman-Pearson Setup

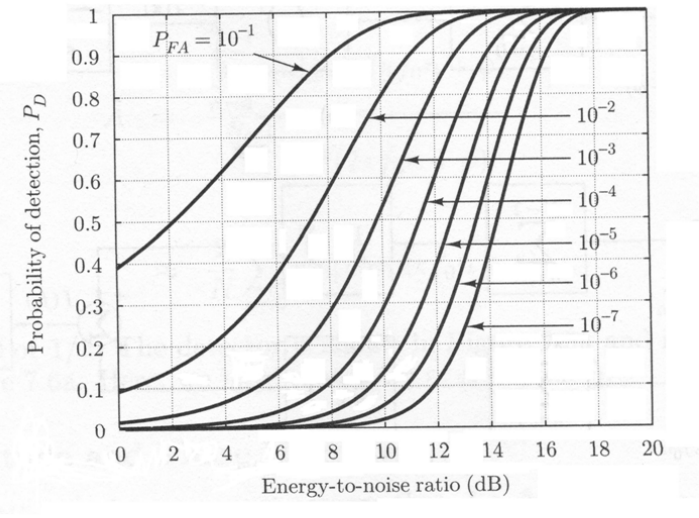
To analyze the performance of a Neyman-Pearson detector, we examine two relationships:

- Between P_{D} and P_{FA}, for a given SNR, called the *receiver operating characteristics* (ROC).
- Between P_{D} and SNR, for a given P_{FA}.

Here are examples of the two:



(a) ROC : P_D vs P_{FA}



(b) P_D vs SNR

see Figs. 3.8 and 3.5 in Kay-II, respectively.

Asymptotic (as $N \to \infty$ and $P_{\mathrm{FA}} \searrow 0$) P_{D} and P_{M} for a Known DC Level in AWGN

We apply the Chernoff-Stein lemma, for which we need to compute the following K-L distance:

$$\begin{aligned} & D\Big(p(X_n \,|\, a=0) \,\|\, p(X_n \,|\, a=A)\Big) \\ = \quad & -\mathrm{E}_{\,p(x_n \,|\, a=0)}\left\{ \log\left[\frac{p(X_n \,|\, a=A)}{p(X_n \,|\, a=0)}\right]\right\} \end{aligned}$$

where

$$\begin{aligned} p(x_n \,|\, a=0) &= \mathcal{N}(0, \sigma^2) \\ \log\left[\frac{p(x_n \,|\, a=A)}{p(x_n \,|\, a=0)}\right] &= -\frac{(x_n - A)^2}{2\,\sigma^2} + \frac{x_n^2}{2\,\sigma^2} = \frac{1}{2\sigma^2} \cdot (2\,A\,x_n - A^2). \end{aligned}$$

Therefore,

$$D\Big(p(X_n \,|\, \theta_0) \,\|\, p(X_n \,|\, \theta_1)\Big) = \frac{A^2}{2\,\sigma^2}$$

and the Chernoff-Stein lemma predicts the following behavior of the detection probability as $N \to \infty$ and $P_{\mathrm{FA}} \to 0$:

$$P_{\mathrm{D}} \approx 1 - \underbrace{f(N) \cdot \exp\left(-N \cdot \frac{A^2}{2\,\sigma^2}\right)}_{\approx P_{\mathrm{M}}}$$

where $f(N)$ is a slowly-varying function of N compared with the exponential term. In this case, the exact expression for $P_{\rm M}$ ($P_{\rm D}$) is available and consistent with the Chernoff-Stein lemma:

$$
\begin{aligned}
P_{\rm M} &= 1 - Q\Big(Q^{-1}(P_{\rm FA}) - \sqrt{NA^2/\sigma^2}\Big) \\
&= Q\Big(\sqrt{NA^2/\sigma^2} - Q^{-1}(P_{\rm FA})\Big) \\
&\overset{N\to+\infty}{\longrightarrow} \frac{1}{[\sqrt{NA^2/\sigma^2} - Q^{-1}(P_{\rm FA})]\cdot\sqrt{2\pi}} \\
&\quad \cdot \exp\Big\{ -[\sqrt{NA^2/\sigma^2} - Q^{-1}(P_{\rm FA})]^2/2 \Big\} \\
&= \frac{1}{[\sqrt{NA^2/\sigma^2} - Q^{-1}(P_{\rm FA})]\cdot\sqrt{2\pi}} \\
&\quad \cdot \exp\Big\{ -NA^2/(2\sigma^2) - [Q^{-1}(P_{\rm FA})]^2/2 \\
&\quad + Q^{-1}(P_{\rm FA})\sqrt{NA^2/\sigma^2} \Big\} \\
&= \frac{1}{[\sqrt{NA^2/\sigma^2} - Q^{-1}(P_{\rm FA})]\cdot\sqrt{2\pi}} \\
&\quad \cdot \exp\Big\{ -[Q^{-1}(P_{\rm FA})]^2/2 + Q^{-1}(P_{\rm FA})\sqrt{NA^2/\sigma^2} \Big\} \\
&\quad \cdot \underbrace{\exp[-NA^2/(2\sigma^2)]}_{\text{as predicted by the Chernoff-Stein lemma}} \qquad (21)
\end{aligned}
$$

where we have used the asymptotic formula (20). When $P_{\rm FA}$

is small and N is large, the first two (green) terms in the above expression make a slowly-varying function $f(N)$ of N. Note that the exponential term in (21) *does not* depend on the false-alarm probability $P_{\rm FA}$ (or, equivalently, on the choice of the decision threshold). The Chernoff-Stein lemma asserts that *this is not a coincidence*.

Comment. For detecting a known DC level in AWGN:

- The slope of the exponential-decay term of $P_{\rm M}$ is

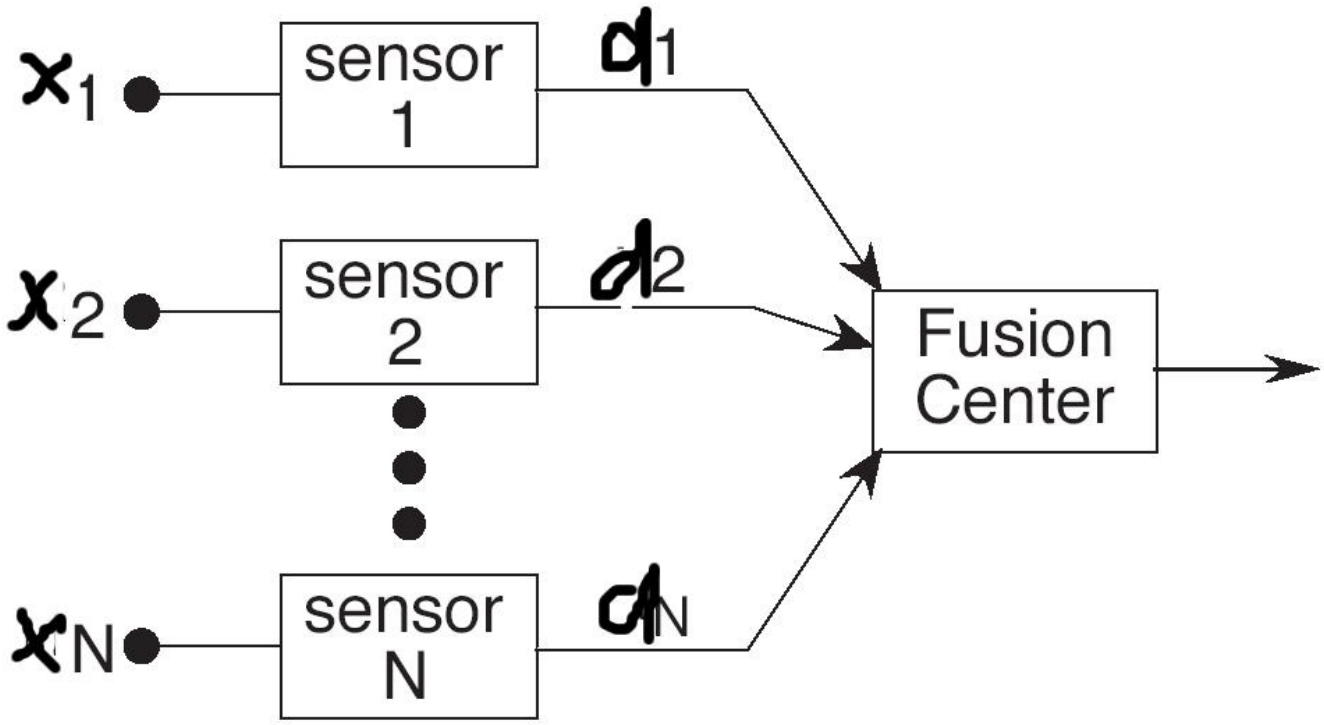
$$A^2/(2\sigma^2)$$

  which is *different from* (larger than, in this case) the slope of the exponential decay of the minimum average error probability:

$$A^2/(8\sigma^2).$$

Decentralized Neyman-Pearson Detection for Simple Hypotheses

Consider a sensor-network scenario depicted by



Assumption: Observations made at N spatially distributed sensors (nodes) follow the same marginal probabilistic model:

$$\mathcal{H}_i : x_n \sim p(x_n \,|\, \theta_i)$$

where $n = 1, 2, \ldots, N$ and $i \in \{0, 1\}$ for binary hypotheses.

Each node n makes a *hard local decision* d_n based on its local observation x_n and sends it to the *headquarters* (fusion center), which collects all the local decisions and makes the final *global*

decision about $\mathcal{H}_0$ or $\mathcal{H}_1$. This structure is clearly suboptimal — it is easy to construct a better decision strategy in which each node sends its (quantized, in practice) likelihood ratio to the fusion center, rather than the decision only. However, such a strategy would have a higher communication (energy) cost.

We now go back to the decentralized detection problem. Suppose that each node n makes a local decision $d_n \in \{0, 1\}$, $n = 1, 2, \ldots, N$ and transmits it to the fusion center. Then, the fusion center makes the global decision based on the likelihood ratio formed from the d_ns. The simplest fusion scheme is based on the assumption that the d_ns are conditionally independent given θ (which may not always be reasonable, but leads to an easy solution). We can now write

$$p(d_n \,|\, \theta_1) = \underbrace{P_{\mathrm{D},n}^{d_n} \, (1 - P_{\mathrm{D},n})^{1-d_n}}_{\text{Bernoulli pmf}}$$

where $P_{\mathrm{D},n}$ is the nth sensor's local detection probability. Similarly,

$$p(d_n \,|\, \theta_0) = \underbrace{P_{\mathrm{FA},n}^{d_n} \, (1 - P_{\mathrm{FA},n})^{1-d_n}}_{\text{Bernoulli pmf}}$$

where $P_{\mathrm{FA},n}$ is the nth sensor's local detection false-alarm

probability. Now,

$$\begin{aligned}\log \Lambda(\boldsymbol{d}) &= \sum_{n=1}^{N} \log \left[\frac{p(d_n \,|\, \theta_1)}{p(d_n \,|\, \theta_0}\right] \\ &= \sum_{n=1}^{N} \log \left[\frac{P_{\mathrm{D},n}^{d_n} \left(1 - P_{\mathrm{D},n}\right)^{1-d_n}}{P_{\mathrm{FA},n}^{d_n} \left(1 - P_{\mathrm{FA},n}\right)^{1-d_n}}\right] \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \log \tau.\end{aligned}$$

To further simplify the exposition, we assume that all sensors have identical performance:

$$P_{\mathrm{D},n} = P_{\mathrm{D},} \quad P_{\mathrm{FA},n} = P_{\mathrm{FA}}.$$

Define the number of sensors having $d_n = 1$:

$$u_1 = \sum_{n=1}^{N} d_n$$

Then, the log-likelihood ratio becomes

$$\log \Lambda(\boldsymbol{d}) = u_1 \log \left(\frac{P_{\mathrm{D}}}{P_{\mathrm{FA}}}\right) + (N - u_1) \log \left(\frac{1 - P_{\mathrm{D}}}{1 - P_{\mathrm{FA}}}\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \log \tau$$

or

$$u_1 \log \left[\frac{P_{\mathrm{D}} \cdot (1 - P_{\mathrm{FA}})}{P_{\mathrm{FA}} \cdot (1 - P_{\mathrm{D}})}\right] \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \log \tau + N \log \left(\frac{1 - P_{\mathrm{FA}}}{1 - P_{\mathrm{D}}}\right). \quad (22)$$

Clearly, each node's local decision d_n is meaningful only if $P_{\mathrm{D}} > P_{\mathrm{FA}}$, which implies

$$\frac{P_{\mathrm{D}} \cdot (1 - P_{\mathrm{FA}})}{P_{\mathrm{FA}} \cdot (1 - P_{\mathrm{D}})} > 1$$

the logarithm of which is therefore positive, and the decision rule (22) further simplifies to

$$u_1 \underset{<}{\overset{\mathcal{H}_1}{\gtrless}} \tau'.$$

The Neyman-Person performance analysis of this detector is easy: the random variable U_1 is binomial given θ (i.e. conditional on the hypothesis) and, therefore,

$$P[U_1 = u_1] \quad = \quad \binom{N}{u_1} p^{u_1} (1-p)^{N-u_1}$$

where $p = P_{\mathrm{FA}}$ under $\mathcal{H}_0$ and $p = P_{\mathrm{D}}$ under $\mathcal{H}_1$. Hence, the "global" false-alarm probability is

$$P_{\mathrm{FA,global}} = P[U_1 > \tau' \,|\, \theta_0] = \sum_{u_1 = \lceil \tau' \rceil}^{N} \binom{N}{u_1} \cdot P_{\mathrm{FA}}^{u_1} \cdot (1 - P_{\mathrm{FA}})^{N-u_1}.$$

Clearly, $P_{\mathrm{FA,global}}$ will be a discontinuous function of τ' and therefore, we should choose our $P_{\mathrm{FA,global}}$ specification from

the available discrete choices. But, if none of the candidate choices is acceptable, this means that our current system does not satisfy the requirements and a remedial action is needed, e.g. increasing the quantity (N) or improving the quality of sensors (through changing P_{D} and P_{FA}), or both.

Testing Multiple Hypotheses

Suppose now that we choose $\Theta_0, \Theta_1, \ldots, \Theta_{M-1}$ that form a *partition* of the parameter space Θ:

$$\Theta_0 \cup \Theta_1 \cup \cdots \cup \Theta_{M-1} = \Theta, \quad \Theta_i \cap \Theta_j = \emptyset \ \forall i \neq j.$$

We wish to distinguish among $M > 2$ hypotheses, i.e. identify *which* hypothesis is true:

$$\begin{aligned} \mathcal{H}_0 &: \theta \in \Theta_0 \quad \text{versus} \\ \mathcal{H}_1 &: \theta \in \Theta_1 \quad \text{versus} \\ &\vdots \qquad\qquad \text{versus} \\ \mathcal{H}_{M-1} &: \theta \in \Theta_{M-1} \end{aligned}$$

and, consequently, our action space consists of M choices. We design a decision rule $\phi : \mathcal{X} \to (0, 1, \ldots, M-1)$:

$$\phi(\boldsymbol{x}) = \begin{cases} 0, & \text{decide } \mathcal{H}_0, \\ 1, & \text{decide } \mathcal{H}_1, \\ \vdots & \\ M-1, & \text{decide } \mathcal{H}_{M-1} \end{cases}$$

where ϕ partitions the data space $\mathcal{X}$ [i.e. the support of $p(\boldsymbol{x} \,|\, \theta)$] into M regions:

Rule ϕ: $\mathcal{X}_0 = \{\boldsymbol{x} : \phi(\boldsymbol{x}) = 0\}, \ldots, \mathcal{X}_{M-1} = \{\boldsymbol{x} : \phi(\boldsymbol{x}) = M-1\}.$

We specify the loss function using $\mathrm{L}(i \,|\, m)$, where, typically, the losses due to correct decisions are set to zero:

$$\mathrm{L}(i \,|\, i) = 0, \quad i = 0, 1, \ldots, M-1.$$

Here, we adopt zero losses for correct decisions. Now, our posterior expected loss takes M values:

$$\begin{aligned} \rho_m(\boldsymbol{x}) &= \sum_{i=0}^{M-1} \int_{\Theta_i} \mathrm{L}(m \,|\, i)\, p(\theta \,|\, \boldsymbol{x})\, d\theta \\ &= \sum_{i=0}^{M-1} \mathrm{L}(m \,|\, i) \int_{\Theta_i} p(\theta \,|\, \boldsymbol{x})\, d\theta, \quad m = 0, 1, \ldots, M-1. \end{aligned}$$

Then, the Bayes' decision rule $\phi^\star$ is defined via the following data-space partitioning:

$$\mathcal{X}_m^\star = \{\boldsymbol{x} : \rho_m(\boldsymbol{x}) = \min_{0 \le l \le M-1} \rho_l(\boldsymbol{x})\}, \quad m = 0, 1, \ldots, M-1$$

or, equivalently, upon applying the Bayes' rule

$$\begin{aligned} \mathcal{X}_m^\star &= \left\{\boldsymbol{x} : m = \arg \min_{0 \le l \le M-1} \sum_{i=0}^{M-1} \int_{\Theta_i} \mathrm{L}(l \,|\, i)\, p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)\, d\theta \right\} \\ &= \left\{\boldsymbol{x} : m = \arg \min_{0 \le l \le M-1} \sum_{i=0}^{M-1} \mathrm{L}(l \,|\, i) \int_{\Theta_i} p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)\, d\theta \right\}. \end{aligned}$$

The preposterior (Bayes) risk for rule $\phi(\boldsymbol{x})$ is

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{x},\theta}[\text{loss}] &= \sum_{m=0}^{M-1} \sum_{i=0}^{M-1} \int_{\mathcal{X}_m} \int_{\Theta_i} \mathrm{L}(m \,|\, i)\, p(\boldsymbol{x} \,|\, \theta) \pi(\theta)\, d\theta\, d\boldsymbol{x} \\
&= \sum_{m=0}^{M-1} \int_{\mathcal{X}_m} \underbrace{\sum_{i=0}^{M-1} \mathrm{L}(m \,|\, i) \int_{\Theta_i} p(\boldsymbol{x} \,|\, \theta) \pi(\theta)\, d\theta}_{h_m(\theta)}\, d\boldsymbol{x}.
\end{aligned}$$

Then, for an arbitrary $h_m(\boldsymbol{x})$,

$$\Big[\sum_{m=0}^{M-1} \int_{\mathcal{X}_m} h_m(x)\, d\boldsymbol{x} \Big] - \Big[\sum_{m=0}^{M-1} \int_{\mathcal{X}_m{}^\star} h_m(\boldsymbol{x})\, d\boldsymbol{x} \Big] \geq 0$$

which verifies that the Bayes' decision rule $\phi^\star$ minimizes the preposterior (Bayes) risk.

Special Case: $\mathrm{L}(i \,|\, i) = 0$ and $\mathrm{L}(m \,|\, i) = 1$ for $i \neq m$ (0-1 loss), implying that $\rho_m(\boldsymbol{x})$ can be written as

$$\begin{aligned} \rho_m(\boldsymbol{x}) &= \sum_{i=0,\, i\neq m}^{M-1} \int_{\Theta_i} p(\theta \,|\, \boldsymbol{x})\, d\theta \\ &= \underbrace{\text{const}}_{\text{not a function of } m} - \int_{\Theta_m} p(\theta \,|\, \boldsymbol{x})\, d\theta \end{aligned}$$

and

$$\begin{aligned} \mathcal{X}_m^\star &= \left\{ \boldsymbol{x} : m = \arg \max_{0\leq l\leq M-1} \int_{\Theta_l} p(\theta \,|\, \boldsymbol{x})\, d\theta \right\} \\ &= \left\{ \boldsymbol{x} : m = \arg \max_{0\leq l\leq M-1} P[\theta \in \Theta_l \,|\, \boldsymbol{x}] \right\} \end{aligned} \tag{23}$$

which is the MAP rule, as expected.

Simple hypotheses: Let us specialize (23) to simple hypotheses ($\Theta_m = \{\theta_m\}$, $m = 0, 1, \ldots, M-1$):

$$\mathcal{X}_m^\star = \left\{ \boldsymbol{x} : m = \arg \max_{0\leq l\leq M-1} p(\theta_l \,|\, \boldsymbol{x}) \right\}$$

or, equivalently,

$$\mathcal{X}_m^\star = \left\{ \boldsymbol{x} : m = \arg \max_{0\leq l\leq M-1} [\pi_l\, p(\boldsymbol{x} \,|\, \theta_l)] \right\},\ m = 0, 1, \ldots, M-1$$

where

$$\pi_0 = \pi(\theta_0), \quad \ldots, \pi_{M-1} = \pi(\theta_{M-1})$$

define the prior pmf of the M-ary discrete random variable θ (recall that $\theta \in \{\theta_0, \theta_1, \ldots, \theta_{M-1}\}$). If π_i, $i = 0, 1, \ldots, M-1$ are all equal:

$$\pi_0 = \pi_1 = \cdots = \pi_{M-1} = \frac{1}{M}$$

the resulting test

$$\begin{aligned} \mathcal{X}_m^\star &= \left\{ \boldsymbol{x} : m = \arg \max_{0 \le l \le M-1} \left[\frac{1}{M} p(\boldsymbol{x} \,|\, \theta_l) \right] \right\} \\ &= \left\{ \boldsymbol{x} : m = \arg \max_{0 \le l \le M-1} \underbrace{p(\boldsymbol{x} \,|\, \theta_l)}_{\text{likelihood}} \right\} \end{aligned} \tag{24}$$

is the *maximum-likelihood test*; this name is easy to justify after inspecting (24) and noting that the computation of the optimal decision region $\mathcal{X}_m^\star$ requires *the maximization of the likelihood* $p(\boldsymbol{x} \,|\, \theta)$ *with respect to the parameter* $\theta \in \{\theta_0, \theta_1, \ldots, \theta_{M-1}\}$.

Summary: Bayesian Decision Approach versus Neyman-Pearson Approach

- The Neyman-Pearson approach appears particularly suitable for applications where the null hypothesis can be formulated as absence of signal or perhaps, absence of statistical difference between two data sets (treatment versus placebo, say).

- In the Neyman-Pearson approach, the null hypothesis is treated very differently from the alternative. (If the null hypothesis is true, we wish to control the false-alarm rate, which is different from our desire to maximize the probability of detection when the alternative is true). Consequently, our decisions should also be treated differently. If the likelihood ratio is large enough, we decide to *accept $\mathcal{H}_1$ (or reject $\mathcal{H}_0$)*. However, if the likelihood ratio is not large enough, we decide *not to reject $\mathcal{H}_0$* because, in this case, it may be that either

  (i) $\mathcal{H}_0$ is true or
  (ii) $\mathcal{H}_0$ is false but the test has low detection probability (power) (e.g. because the signal level is small compared with noise or we collected too small number of observations).

- The Bayesian decision framework is suitable for communications applications as it can easily handle multiple hypotheses (unlike the Neyman-Pearson framework).
  - **0-1 loss:** In communications applications, we typically select a 0-1 loss, implying that all hypotheses are treated equally (i.e. we could change the roles of null and alternative hypotheses without any problems). Therefore, interpretations of our decisions are also straightforward. Furthermore, in this case, the Bayes' decision rule is also optimal in terms of minimizing the average error probability, which is one of the most popular performance criteria in communications.

P Values

Reporting "accept $\mathcal{H}_0$" or "accept $\mathcal{H}_1$" is not very informative. Instead, we could vary $P_{\rm FA}$ and examine how our report would change.

Generally, if $\mathcal{H}_1$ is accepted for a certain specified $P_{\rm FA}$, it will be accepted for $P'_{\rm FA} > P_{\rm FA}$. Therefore, there exists a smallest $P_{\rm FA}$ at which $\mathcal{H}_1$ is accepted. This motivates the introduction of the p value.

To be more precise (and be able to handle composite hypotheses), here is a definition of a *size of a hypothesis test*.

Definition 1. *The size of a hypothesis test described by*

$$\textit{Rule } \phi: \quad \mathcal{X}_0 = \{x : \phi(x) = 0\}, \quad \mathcal{X}_1 = \{x : \phi(x) = 1\}.$$

is defined as follows:

$$\alpha = \max_{\theta \in \Theta_0} P[\boldsymbol{x} \in \mathcal{X}_1 \,|\, \theta] = \textit{max possible } P_{\rm FA}.$$

A hypothesis test is said to have level α if its size is less than or equal to α. Therefore, a level-α test is guaranteed to have a false-alarm probability less than or equal to α.

Definition 2. *Consider a Neyman-Pearson-type setup where our test*

$$\textit{Rule } \phi_\alpha\textit{:} \quad \mathcal{X}_{0,\alpha} = \{\boldsymbol{x} : \phi(\boldsymbol{x}) = 0\}, \quad \mathcal{X}_{1,\alpha} = \{\boldsymbol{x} : \phi(\boldsymbol{x}) = 1\}. \tag{25}$$

achieves a specified size α, meaning that,

$$\begin{aligned} \alpha &= \textit{max possible } P_{\text{FA}} \\ &= \max_{\theta \in \Theta_0} P[\boldsymbol{x} \in \mathcal{X}_{1,\alpha} \,|\, \theta] \quad \textit{(composite hypotheses)} \end{aligned}$$

or, in the simple-hypothesis case ($\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$):

$$\alpha = \overbrace{P[\boldsymbol{x} \in \mathcal{X}_{1,\alpha} \,|\, \theta = \theta_0]}^{P_{\text{FA}}} \quad \textit{(simple hypotheses)}.$$

We suppose that, for every $\alpha \in (0, 1)$, we have a size-α test with decision regions (25). Then, the p value for this test is the smallest level α at which we can declare $\mathcal{H}_1$:

$$p\textit{ value} = \inf\{\alpha \,:\, \boldsymbol{x} \in \mathcal{X}_{1,\alpha}\}.$$

Informally, the p value is a measure of evidence for supporting $\mathcal{H}_1$. For example, p values less than 0.01 are considered *very strong evidence* supporting $\mathcal{H}_1$.

There are a lot of warnings (and misconceptions) regarding p values. Here are the most important ones.

Warning: A large p value is *not* strong evidence in favor of $\mathcal{H}_0$; a large p value can occur for two reasons:

(i) $\mathcal{H}_0$ is true or

(ii) $\mathcal{H}_0$ is false but the test has low detection probability (power).

Warning: *Do not* confuse the p value with

$$P[\mathcal{H}_0 \,|\, \text{data}] = P[\theta \in \Theta_0 \,|\, \boldsymbol{x}]$$

which is used in Bayesian inference. **The p value is *not* the probability that $\mathcal{H}_0$ is true.**

Theorem 1. *Suppose that we have a size-α test of the form*

$$\textit{declare } \mathcal{H}_1 \textit{ if and only if } \quad T(\boldsymbol{x}) \geq c_\alpha.$$

Then, the p value for this test is

$$p \textit{ value} = \max_{\theta \in \Theta_0} P[T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \,|\, \theta]$$

where $\boldsymbol{x}$ is the observed value of $\boldsymbol{X}$. For $\Theta_0 = \{\theta_0\}$:

$$p \textit{ value} = P[T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \,|\, \theta = \theta_0].$$

In words, Theorem 1 states that

> The p value is the probability that, under $\mathcal{H}_0$, a random data realization $\boldsymbol{X}$ is observed yielding a value of the test statistic $T(\boldsymbol{X})$ that is greater than or equal to what has actually been observed (i.e. $T(\boldsymbol{x})$).

Note: This interpretation requires that we allow the experiment to be repeated many times. This is what Bayesians criticize by saying that "data that have never been observed are used for inference."

Theorem 2. *If the test statistics has a continuous distribution, then, under $\mathcal{H}_0 : \theta = \theta_0$, the p value has a uniform$(0, 1)$ distribution. Therefore, if we declare $\mathcal{H}_1$ (reject $\mathcal{H}_0$) when the p value is less than or equal to α, the probability of false alarm is α.*

In other words, if $\mathcal{H}_0$ is true, the p value is like a random draw from an uniform$(0, 1)$ distribution. If $\mathcal{H}_1$ is true and if we repeat the experiment many times, the random p values will concentrate closer to zero.

Multiple Testing

We may conduct many hypothesis tests in some applications, e.g.

- bioinformatics and
- sensor networks.

Here, we perform *many* (typically binary) *tests* (one test per node in a sensor network, say). This is different from *testing multiple hypotheses* that we considered on pp. 54–58, where we performed *a single test of multiple hypotheses*. For a sensor-network related discussion on multiple testing, see

E.B. Ermis, M. Alanyali, and V. Saligrama, "Search and discovery in an uncertain networked world," *IEEE Signal Processing Magazine*, vol. 23, pp. 107–118, Jul. 2006.

Suppose that each test is conducted with false-alarm probability $P_{\mathrm{FA}} = \alpha$. For example, in a sensor-network setup, each node conducts a test based on its local data.

Although the chance of false alarm at each node is only α, the chance of at least one falsely alarmed node is much higher, since there are many nodes. Here, we discuss two ways to deal with this problem.

Consider M hypothesis tests:

$$H_{0i} \text{ versus } H_{1i},\ i = 1, 2, \ldots, M$$

and denote by $p_1, p_2, \ldots, p_M$ the p values for these tests. Then, the *Bonferroni method* does the following:

> Given the p values $p_1, p_2, \ldots, p_m$, accept $\mathcal{H}_{1i}$ if
>
> $$p_i < \frac{\alpha}{M}.$$

Recall the *union-of-events bound*:

$$P\Big[\bigcup_{i=1}^{n} A_i\Big] \leq \sum_{i=1}^{n} P[A_i].$$

which holds for arbitrary events $A_i,\ i = 1, 2, \ldots, n$, see handout # 2 in EE 420x notes.

Theorem 3. *If we apply the Bonferroni method, the probability of any individual false alarm is less than or equal to* α.

Proof. Denote by A the event that there is at least one false alarm and by A_i the event that the ith node is falsely alarmed.

Then

$$P\{A\} = P\Big\{\bigcup_{i=1}^{M} A_i\Big\} \leq \sum_{i=1}^{n} P\{A_i\} = \sum_{i=1}^{M} \frac{\alpha}{M} = \alpha.$$

□

Comments: Suppose now that the tests are *statistically independent* and consider the smallest of the $M\,p$ values. Under $\mathcal{H}_{0i}$, p_i are $\text{uniform}(0,1)$. Then

$$P\{\min\{P_1, P_2, \ldots, P_M\} > x\} \Big|_{\substack{\text{assuming } \mathcal{H}_{0i} \\ i=1,2,\ldots,M}} = (1-x)^M$$

yielding the "proper" p value to be attached to $\min\{p_1, p_2, \ldots, p_m\}$ as

$$1 - (1 - \min\{p_1, p_2, \ldots, p_M\})^M \tag{26}$$

and if $\min\{p_1, p_2, \ldots, p_M\}$ is small and M not too large, (26) will be close to $m \min\{p_1, p_2, \ldots, p_M\}$.

False Discovery Rate: Sometimes it is reasonable to control *false discovery rate (FDR)*, which we introduce below.

Suppose that we accept $\mathcal{H}_{1i}$ for all i for which

$$p_i < \text{threshold } \tau$$

and let $m_0 + m_1 = m$ be
number of true $\mathcal{H}_{0i}$ hypotheses + number of true $\mathcal{H}_{1i}$ hypotheses = total number of hypotheses (nodes, say).

# of different outcomes	$\mathcal{H}_0$ not rejected	$\mathcal{H}_1$ declared	total
$\mathcal{H}_0$ true	U	V	m_0
$\mathcal{H}_1$ true	T	S	m_1
total	$m - R$	R	m

Define the *false discovery proportion* (FDP) as

$$\text{FDP} = \begin{cases} V/R, & R > 0, \\ 0, & R = 0 \end{cases}$$

which is simply the proportion of *incorrect* $\mathcal{H}_1$ decisions. Now, define

$$\text{FDR} = \text{E}\left[\text{FDP}\right].$$

The Benjamini-Hochberg (BH) Method

(i) Denote the ordered p values by $p_{(1)} < p_{(2)} < \cdots p_{(m)}$.

(ii) Define

$$l_i = \frac{i\,\alpha}{C_m\,m} \quad \text{and} \quad R = \max\{i \,:\, p_{(i)} < l_i\}$$

where C_m is defined to be 1 if the p values are independent and $C_m = \sum_{i=1}^{m}(1/i)$ otherwise.

(iii) Define the *BH rejection threshold* $\tau = p_{(R)}$.

(iv) Accept all $\mathcal{H}_{1i}$ for which $p_{(i)} \leq \tau$.

Theorem 4. *(formulated and proved by Benjamini and Hochberg) If the above BH method is applied, then, regardless of how many null hypotheses are true and regardless of the distribution of the p values when the null hypothesis is false,*

$$\mathrm{FDR} = \mathrm{E}\left[\mathrm{FDP}\right] \leq \frac{m_0}{m}\,\alpha \leq \alpha.$$