Chapter 3

Basic concepts of general hypothesis testing

In previous chapters, we introduced the main basis of the simple binary hypothesis testing problem. In this chapter, we consider the more general case of the M-ary hypothesis testing. First, we give the basic definitions for the case of simple hypothesis testing for the well-known stochastic processes. Then, we consider the case of composite hypothesis testing where the stochastic processes are parametrically known. In each case, we try to make simple classification of different decision rules, describing their optimality criterion and their performances.

3.1 A general M-ary hypothesis testing problem

To consider a general M-ary hypothesis testing problem, let consider the necessary steps that any decision making procedure has to follow:

1. Get the data: observe $x(t)$ a realization of $X(t)$ in some time interval $[0, T]$.
2. Define a library of hypothesis $\{H_i, i = 1, \cdots, M\}$ where H_i is the hypothesis that the data $x(t)$ come from a stochastic processes $X_i(t)$ with either finite or infinite membership.
3. Define a performance criterion for evaluating the decisions $\{\delta_i, i = 1, \cdots, M\}$.
4. If possible, define a probability measure determining the *a priori* probabilities of stochastic processes in the library.
5. Use all the available assets to formulate a general optimization problem whose solution is a decision.

Evidently, the nature of the optimization problem and the subsequent decisions vary significantly with the specifities of the library of the stochastic processes, with the availability of the *a priori* probability distribution on these stochastic processes, the necessary performance criterion to optimize and the possibility of controling the observation time $[0, T]$ dynamically.

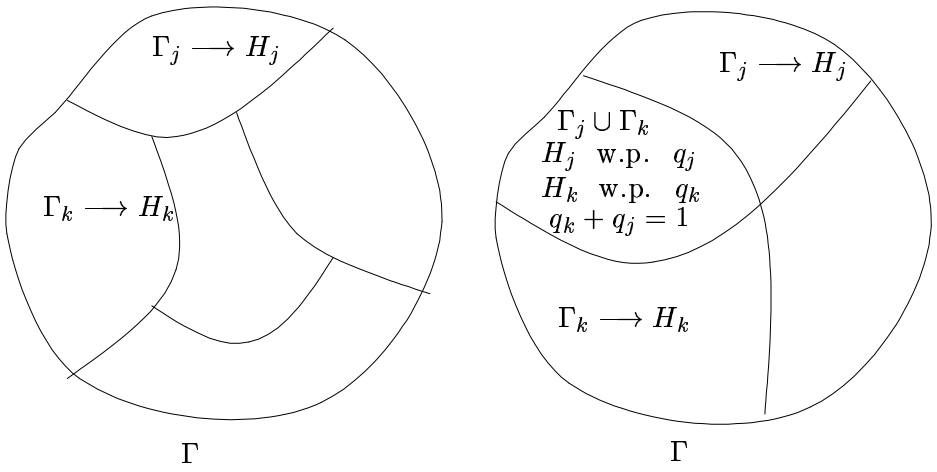

Figure 3.1: Partition of the observation space and deterministic or probabilistic decision rules.

For any fixed specifications on the above issues, the decision then depends on the data. This is what is called the *decision rule* or *test*.

We can distinguish the following special cases:

- If the library has a finite number of members, the decision process is classified as *hypothesis testing*.
- If this number is only two, then the decision process is called *detection*.
- If the stochastic processes are well-known, the hypothesis testing is called *simple*. If they are defined parametrically, then the decision process is called *parametric*, if not, it is called *nonparametric*.

3.1.1 Deterministic or probabilistic decision rules

A decision rule $\delta = \{\delta_j(\boldsymbol{x}), j = 1, \cdots, M\}$ subdivides the space of the observations Γ into M subspaces $\{\Gamma_j, j = 1, \cdots, M\}$. One can distinguish two types of decision rules:

- *Deterministic decision rule*:
  If these subspaces are all disjoints and for a given data set $\boldsymbol{x}$, the hypothesis H_j is decided with probability one, the decision rule is called deterministic.
- *Probabilistic decision rule*:
  If some of theses subspaces overlap and for a given data set $\boldsymbol{x}$, none of the hypothesis can be decided with probability one, then the decision rule is called probabilistic. That is, given $\boldsymbol{x}$, the hypothesis H_k is decided with probability q_k and H_j with probability q_j with $\sum_j q_j = 1$.

3.1.2 Conditional, A priori and Joint Probability Distributions

For a given decision rule δ, we can define the following probability distributions:

- Conditional probability distribution:
  $P_{ki}(\delta)$ is the conditional probability that H_k is chosen given that H_i is true. These probabilities can be calculated from the probability distribution of the stochastic process:

$$\begin{aligned} P_{ki}(\delta) &= \Pr\{H_k \text{ decided by rule } \delta | H_i \text{ true}\} \\ &= \int_\Gamma \mathrm{dPr}\{H_k \text{ decided and } \boldsymbol{x} \text{ observed } | H_i \text{ true}\} \\ &= \int_\Gamma \Pr\{H_k \text{ decided} | \boldsymbol{x} \text{ observed}\}\, \mathrm{dPr}\{\boldsymbol{x} \text{ observed } | H_i \text{ true}\} \\ &= \int_\Gamma \delta_k(\boldsymbol{x})\, \mathrm{d}F_i(\boldsymbol{x}) = \int_\Gamma \delta_k(\boldsymbol{x}) f_i(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \end{aligned} \tag{3.1}$$

  Note that, in this derivation, we have used the theorem of total probabilities and the Bayes rule and the fact that the decision induced by the decision rule δ is independent of the true hypothesis. Note also that the decision rule δ consists of probabilities such that

$$\sum_{j=1}^{n} \delta_j(\boldsymbol{x}) = 1 \quad \forall \boldsymbol{x} \in \Gamma \tag{3.2}$$

  Using this fact, it is easy to show that

$$\sum_{k=1}^{M} P_{ki} = 1 \quad \forall i = 1, \cdots, M \tag{3.3}$$

  This can be interpreted as: Given the true hypothesis H_i, the decision induced by the decision rule δ is restricted to one of the M hypotheses.

- A priori probability distribution:
  $\{\pi_i, i = 1, \cdots, M\}$ is a prior probability distribution on the hypothesis $\{H_i, i = 1, \cdots, M\}$. Naturally we have

$$\sum_{k=1}^{M} \pi_k = 1 \tag{3.4}$$

- Joint probability distribution:
  Using the conditional probabilities $P_{kj}(\delta)$ and the prior probabilities π_i, we can calculate the joint probabilities $Q_{ki}(\delta)$, denoting the probabilities that H_k is decided by the decision rule δ while H_i is true. We then have:

$$Q_{ki}(\delta) = \pi_i P_{ki}(\delta) = \pi_i \int_\Gamma \delta_k(\boldsymbol{x}) f_i(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{3.5}$$

$$\sum_{k=1}^{M} Q_{ki} = \pi_i \quad \forall i = 1, \cdots, M \tag{3.6}$$

$$\sum_{k=1}^{M}\sum_{i=1}^{M} Q_{ki} = 1 \tag{3.7}$$

$$\tag{3.8}$$

3.1.3 Probabilities of false and correct detection

For a given decision rule δ, we can define the following probabilities:

- Probability of false decision:
  $P_e(\delta)$ is the probability that the decision induced by δ is erroneous, *i.e.*; H_k is decided while $H_{i\neq k}$ is true.

$$P_e(\delta) = \sum_{k\neq i} Q_{ki}(\delta) \tag{3.9}$$

- Probability of correct decision:
  $P_d(\delta)$ is the probability that the decision induced by δ is correct, *i.e.*H_k is decided while H_k is true.

$$P_d(\delta) = \sum_{k} Q_{kk}(\delta) = 1 - P_e(\delta) \tag{3.10}$$

- One can use these probabilities to define optimal decision procedures:

$$P_e(\delta^*) \leq P_e(\delta), \quad \forall \delta \in \mathcal{D} \tag{3.11}$$
$$P_d(\delta^*) \geq P_d(\delta), \quad \forall \delta \in \mathcal{D} \tag{3.12}$$

3.1.4 Penalty or costs coefficients

In addition to the M known hypotheses and their *a priori* probabilities, the analyst may be equipped with a set of real *cost* or *penalty* coefficients c_{ki} such that

$$c_{ki} \geq 0 \quad \text{and} \quad c_{ki} \geq c_{kk} \quad \forall k, i = 1, \cdots, M, \tag{3.13}$$

where c_{ki} is the penalty paied when H_k is decided while H_i is true. The implication behind the condition that each coefficient is nonnegative is that there is no gain associated with any decision, thus the term penalty and, in general, the penalties c_{ki} are chosen greater than c_{kk}.

3.1.5 Conditional and Bayes risks

For a given decision rule δ and a given set of cost functions c_{ki}, one can calculate the following quantities:

- Conditional expected penalties or conditional risks:

$$R_i(\delta) = \sum_{k=1}^{M} c_{ki} P_{ki}(\delta) \tag{3.14}$$

- Expected penalty or Bayes risk:

$$r(\delta) = \sum_{k=1}^{M} \sum_{i=1}^{M} c_{ki}\, Q_{ki}(\delta) \tag{3.15}$$

3.1.6 Bayesian and non Bayesian hypothesis testing

Different optimization problems which are classically used to define a decision process are:

- *Bayesian* hypothesis testing:
  - If a specific cost function that penalizes wrong decisions is provided, then the minimization of the *expected penalty* or the *Bayes risk* is chosen as the performance criterion.
  - If not, the *probability of making a decision error* is minimized instead. This decision process is called *ideal observation test.*
- *Non Bayesian* hypothesis testing:
  When an *a priori* probability distribution is unavailable, then
  - If a specific cost function is available, then first a least favorable *a priori* probability distribution is defined and then the expected penalty with this least favorable *a priori* probability distribution is minimized to obtain a decision rule. This decision process is what is called the *minimax* decision process.
  - If a cost function is not available then first one of the hypothesis is selected in advance as to be the most important and then the performance criterion used is the maximization of the probability of the detection of that hypothesis subject to the constraint that the probability of its false alarm does not exceed a given value α. This is what is called the *Neyman-Pearson* test procedure.

3.1.7 Admissible decision rules and stopping rule

- Admissible decision rules:
  It may happen that, for a given set of optimal criterions, there exist more than one best decision rule which satisfy these performances criterion, then these rules are called admissible.
- When a decision rule is designed, one may be intended to know how this decision rule performs with respect of the observed time interval $[0, T]$, *i.e.*; the number of data. The study of the behavior of the decision rule is called stoping rule.

All the above test procedures take a dynamic form if the observation time interval $[0, T]$ can be controlled dynamically.

3.1.8 Classification of simple hypothesis testing schemes

To summarize, let again list the richest possible set of assets available to the analyst:

i. A library of M distinct hypothesis $\{H_i, i = 1, \cdots, M\}$;

ii. A set of data $\boldsymbol{x}$ which is assumed to be a realization of a well known stochastic process under only one of these hypotheses;

iii. A prior probability distribution $\{\pi_i, i = 1, \cdots, M\}$ for the M hypotheses;

iv. A set of penalty coefficients $\{c_{ki}, k, i = 1, \cdots, M\}$;

The minimum set of assets that is (or must be) always available consists of those in i and ii and the performance criterion will suffer limitations as the number of remaining available assets decrease.

Now, to continue, first we assume that all assets in i to iv are available. Then an optimal rule δ^* is such that the expected penalty $r(\delta)$ is lower that any others, *i.e.*

$$\delta^* : r(\delta^*) \leq r(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.16}$$

This rule then guarantees a minimum average cost due to the wrong decisions. Note that this rule may not be unique. When the uniqueness is not satisfied, this means that there exist a number of admissible rules, among them, we can choose the one which is the simplest to implement.

If assets in i to iii are available, then an optimal rule δ^* can be defined by using the induced probability of error $P_e(\delta)$, or the probability of the detection,*i.e.*

$$\delta^* : P_e(\delta^*) \leq P_e(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.17}$$

or

$$\delta^* : P_d(\delta^*) \geq P_d(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.18}$$

Again, note that there may not exist a unique decision rule, but a set of admissible rules, among them, we can choose the one which is the simplest to implement.

The hypothesis testing rules based on the above criterions are called *Bayesian* due to the basic ingredient which is the availability of the prior probabilities π_i on the hypothesis space. When the asset iii is not available, then the decision rules are called *non Bayesian.*

Now assume all assets i, ii and iv are available, then the analyst can choose an arbitrary prior probability distribution $\pi = \{\pi_i\}$ and calculate the induced conditional expected penalty $R(\delta, \pi)$. Then, the decision rule

$$\delta^* : \sup_{\pi} R(\delta^*, \pi) \leq \sup_{\pi} R(\delta, \pi) \quad \forall \delta \in \mathcal{D} \tag{3.19}$$

defines admissible ones. The analyst then can choose between these admissible rules the one with the lowest complexity. This procedure, when successful, isolates the decision rule that induces the minimum maximum value of the conditional expected penalty and protects the analyst against the most costly case. This formalism and procedure is called *minimax.*

Finally, assume that only the assets in i and ii are available, Then, the main idea is to select one of the hypothesis as to be the principal and use the notion of the *power function* $P(\delta)$.

General Hypothesis Testing Schemes			
Scheme	A priori	Cost	Decision rule
Bayesian	Yes	Yes	Minimization of expected penalty $r(\delta)$
	Yes	No	Minimization of error probability $P_e(\delta)$
Non Bayesian	No	Yes	Minimax test rule using conditional risks $R_j(\delta)$
	No	No	Neyman-Pearson test rule using $P_e(\delta)$ and $P_d(\delta)$

Classes of Hypothesis Testing Schemes for Well Known Stochastic Processes.				
Scheme	Assets used	Optimization function	Optimal Decision rule δ^*	Specific Name
Bayesian	i, ii, iii, iv, v	$r(\delta)$	$r(\delta^*) \leq r(\delta)$	Bayesian
	i, ii, iii, v	$P_e(\delta)$	$P_e(\delta^*) \leq P_e(\delta)$	Bayesian
Non Bayesian	i, ii, iii	$\sup_p r(\delta, \pi)$	$\sup_p r(\delta^*, \pi) \leq \sup_p r(\delta, \pi)$	Minimax
	ii, iii	$P_d(\delta)$ subject to $P_e(\delta) \leq \alpha$	$P_d(\delta^*) \geq P_d(\delta)$ and $P_e(\delta) \leq \alpha$	Neyman-Pearson

3.2 Composite hypothesis

Now assume that, the hypothesis H_i means that $\boldsymbol{x}$ is a realization of the process $\boldsymbol{X}_i$ but the process $\boldsymbol{X}_i$ is only parametrically known, i.e.; its probability distribution is known within a set of unknown parameters $\boldsymbol{\theta}$ so that, the prior probabilities $\{\pi_i\}$ depend on the parameter $\boldsymbol{\theta}$. Now, assume that the partitioning of the decision rule is due to the partition of the parameter space $\mathcal{T}$ of possible values of $\boldsymbol{\theta}$, i.e.;

$$\mathcal{T} = \{\mathcal{T}_1, \cdots, \mathcal{T}_M\}, \quad \cup_i \mathcal{T}_i = \mathcal{T} \tag{3.20}$$

Assume also that, for each value of $\boldsymbol{\theta}$, the stochastic process is defined through its probability distribution $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and assume that we can define a probability density function $\pi(\boldsymbol{\theta})$ over the space $\mathcal{T}$. Then we have

$$\pi_i = \int_{\mathcal{T}_i} \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{3.21}$$

and

$$P_{k,\boldsymbol{\theta}}(\delta) = \int_{\Gamma} \delta_k(\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{3.22}$$

where

$$\sum_{k=1}^{M} P_{k,\boldsymbol{\theta}}(\delta) = 1 \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.23}$$

We can now calculate the conditional probabilities $P_{ki}(\delta)$

$$\begin{aligned} P_{ki}(\delta) &= \pi_i^{-1} \int_{\mathcal{T}_i} P_{k,\boldsymbol{\theta}}(\delta)\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ &= \left[\int_{\mathcal{T}_i} \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}\right]^{-1} \int_{\Gamma} \mathrm{d}\boldsymbol{x}\, \delta_k(\boldsymbol{x}) \int_{\mathcal{T}_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \end{aligned} \tag{3.24}$$

and the joint probabilities Q_{ki} as follows:

$$\begin{aligned} Q_{ki}(\delta) &= \pi_i P_{ki}(\delta) = \int_{\mathcal{T}_i} P_{k,\boldsymbol{\theta}}(\delta)\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\Gamma} \mathrm{d}\boldsymbol{x}\, \delta_k(\boldsymbol{x}) \int_{\mathcal{T}_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \end{aligned} \tag{3.25}$$

3.2.1 Penalty or cost functions

In addition to the M parametrically known hypotheses, we may be equiped with a set of real *penalty* or *cost* functions $c_k(\boldsymbol{\theta})$.

We can then calculate:

- Conditional expected penalty or conditional risk function:

$$R(\delta, \boldsymbol{\theta}) = \sum_{k=1}^{M} c_k(\boldsymbol{\theta}) P_{k,\boldsymbol{\theta}}(\delta) \tag{3.26}$$

$$= \int \mathrm{d}\boldsymbol{x} \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) c_k(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{3.27}$$

- Expected penalty or Bayes risk:
  If $\pi(\boldsymbol{\theta})$ is available, then the expected penaly can be calculated by

$$r(\delta) = \int_{\mathcal{T}} R(\delta, \boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = \sum_{k=1}^{M} \int c_k(\boldsymbol{\theta}) P_{k,\boldsymbol{\theta}}(\delta)\,\mathrm{d}\boldsymbol{\theta} \tag{3.28}$$

$$= \int \mathrm{d}\boldsymbol{x} \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) \int c_k(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{x})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{3.29}$$

3.2.2 Case of binary hypothesis testing

Now, consider the particular case of *binary hypothesis testing*, where there are only hypotheses, and assume that they are determined through a single parametrically known stochastic process and two disjoint subdivisions of the parameter space $\mathcal{T}$. Let note these two hypotheses H_0 and H_1 and the decision rules $\delta_0(\boldsymbol{x})$ and $\delta_1(\boldsymbol{x})$ with $\delta_0(\boldsymbol{x}) + \delta_1(\boldsymbol{x}) = 1 \quad \forall \boldsymbol{x} \in \Gamma$. Now, if we emphasize the hypothesis H_1 (detection), then $\delta_0(\boldsymbol{x}) = 1 - \delta_1(\boldsymbol{x})$, so that we can drop the indices in the decision rule and denote by $\delta(\boldsymbol{x}) = \delta_1(\boldsymbol{x})$. Now, we have

$$P_{\boldsymbol{\theta}}(\delta) = \int_{\Gamma} \delta(\boldsymbol{x})\, f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{3.30}$$

This expression represents the probability that the emphasized hypothesis H_1 is decided, conditioned on the value $\boldsymbol{\theta}$ of the vector parameter. This function is called *the power function of the decision rule*. This is due to the fact that it provides the probability with which the emphatic hypothesis is decided for each fixed parameter value $\boldsymbol{\theta}$.

3.2.3 Classification of hypothesis testing schemes for a parametrically known stochastic process

As before, we now summarize the richest possible set of assets available for a composite hypothesis testing problem:

1. A library of M distinct hypotheses $\{H_i, i = 1, \cdots, M\}$
2. A set of data $\boldsymbol{x}$ which is assumed to be a realization of a parametrically known stochastic process, the hypotheses $\{H_i, i = 1, \cdots, M\}$ corresponding to the M disjoint subdivisions of the parameter space $\mathcal{T}$.
3. A prior probability distribution $\pi(\boldsymbol{\theta})$ on the parameter space $\mathcal{T}$
4. A set of penalty functions $\{c_k(\boldsymbol{\theta}), k = 1, \cdots, M\}$ defined on $\mathcal{T}$.

First we assume that all assets in i to iv are available. Then an optimal rule δ^* is such that the expected penalty $r(\delta)$ is lower that any others, *i.e.*

$$\delta^* : r(\delta^*) \leq r(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.31}$$

If assets in i to iii are available, then an optimal rule δ^* can be defined by using the induced probability of error $P_e(\delta)$, or the probability of the detection, *i.e.*;

$$\delta^* : P_e(\delta^*) \leq P_e(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.32}$$

or

$$\delta^* : P_d(\delta^*) \geq P_d(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.33}$$

Again, note that there may not exist a unique decision rule, but a set of admissible rules, among which we can choose the one which is the simplest to implement.

Now assume that the assets i, ii and iv are available, then the analyst can use the induced conditional expected penalty $R(\delta, \boldsymbol{\theta})$. An optimal rule would induce relatively low $R(\delta, \boldsymbol{\theta})$ values for all values of $\boldsymbol{\theta} \in \mathcal{T}$. So, if there exist two rules $\delta^{(1)}$ and $\delta^{(2)}$ such that

$$R(\delta^{(1)}, \boldsymbol{\theta}) \leq R(\delta^{(2)}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.34}$$

then $\delta^{(2)}$ should be rejected in the presence of $\delta^{(1)}$. The rule $\delta^{(1)}$ is said to be *uniformly superior* than the rule $\delta^{(2)}$. But, it may happen that $R(\delta^{(1)}, \boldsymbol{\theta}) \leq R(\delta^{(2)}, \boldsymbol{\theta})$ for some values of $\boldsymbol{\theta}$ and $R(\delta^{(1)}, \boldsymbol{\theta}) > R(\delta^{(2)}, \boldsymbol{\theta})$ for other values of $\boldsymbol{\theta}$. In this case, we may ask to prefer $\delta^{(1)}$ to $\delta^{(2)}$ if

$$\sup_{\boldsymbol{\theta} \in \mathcal{T}} R(\delta^{(1)}, \boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \mathcal{T}} R(\delta^{(2)}, \boldsymbol{\theta}) \tag{3.35}$$

Thus, the selection procedure has, in general, two steps: first reject all the *uniformly inadmissible* rules, and then between the remaining ones, define the optimal rules:

$$\delta^* : \sup_{\boldsymbol{\theta} \in \mathcal{T}} R(\delta^*, \boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \mathcal{T}} R(\delta, \boldsymbol{\theta}) \quad \forall \delta \in \mathcal{D} \tag{3.36}$$

which are admissible. Finally, the analyst then can choose between these admissible rules the one with the lowest complexity. This procedure, when successful, isolates the decision rule that induces the minimum maximum value of the conditional expected penalty and protects the analyst against the most costly case. This formalism and procedure is called *minimax*.

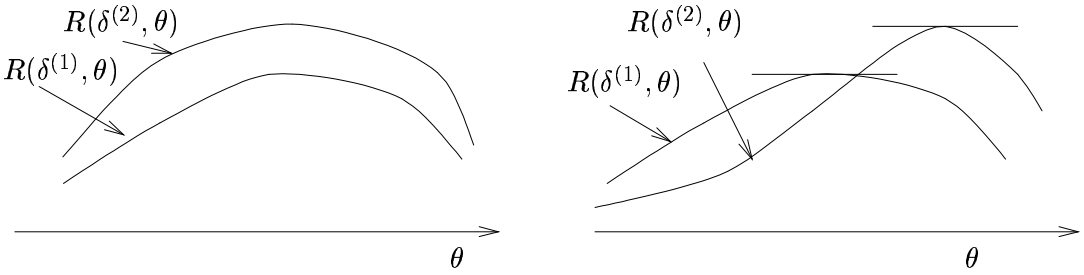

Figure 3.2: Two decision rules $\delta^{(1)}$ and $\delta^{(2)}$ and thier respectives risk functions. In both cases $\delta^{(2)}$ is rejected in presence of $\delta^{(1)}$.

Finally, assume that only the assets in i and ii are available. Then, the main idea is to select one of the hypotheses as to be the principal and use the notion of the *power function* $P_{\boldsymbol{\theta}}(\delta)$. Let note H_1 the emphasized hypothesis, $\mathcal{T}_1$ its associated region in $\mathcal{T}$ and $P_{\boldsymbol{\theta}}(\delta)$ the power function associated to it. It is then desirable that $P_{\boldsymbol{\theta}}(\delta)$ for any $\boldsymbol{\theta} \in \mathcal{T}_1$ has a value higher that its value for other hypotheses, *i.e.*;

$$P_{\boldsymbol{\theta} \in \mathcal{T}_1}(\delta) \geq P_{\boldsymbol{\theta} \in \mathcal{T}_j}(\delta) \tag{3.37}$$

The quantity $\sup_{\boldsymbol{\theta}\in\mathcal{T}_0} P_{\boldsymbol{\theta}}(\delta)$ is the false alarm induced by δ. The value of $P_{\boldsymbol{\theta}}(\delta)$ for a given value of $\boldsymbol{\theta} \in \mathcal{T}_1$ is the power induced by the decision rule δ. If the subspaces $\mathcal{T}_0$ and $\mathcal{T}_1$ are fixed, then the best decision rule δ^* is the one that induces the highest power subject to a false alarm constraint, *i.e.*;

$$\delta^* : P_{\boldsymbol{\theta}}(\delta^*) \leq P_{\boldsymbol{\theta}}(\delta) \quad \forall \boldsymbol{\theta} \in \mathcal{T}_1 \quad \text{subject to} \quad \sup_{\boldsymbol{\theta}\in\mathcal{T}_0} P_{\boldsymbol{\theta}}(\delta^*) \leq \alpha, \quad \forall \delta \in \mathcal{D} \tag{3.38}$$

The procedure, as in the minimax scheme, may have more than one solution.

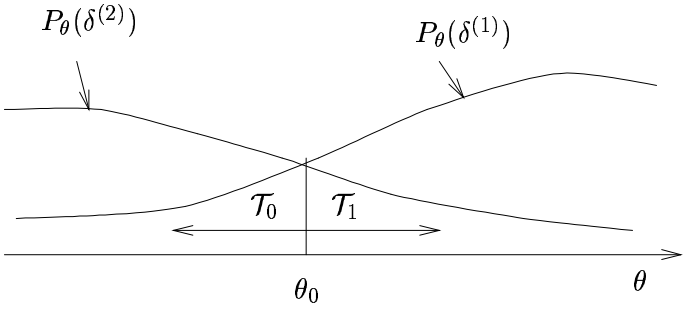

Figure 3.3: Two decision rules $\delta^{(1)}$ and $\delta^{(2)}$ and thier respectives power functions. In this case $\delta^{(1)}$ is prefered to $\delta^{(2)}$.

Classes of Hypothesis Testing Schemes for Parametrically Known Stochastic Processes.				
Scheme	Assets used	Optimization function	Optimal Estimate δ^*	Specific Name
Bayesian	i, ii, iii, iv, v	$r(\delta)$	$r(\delta^*) \leq r(\delta)$	Bayesian
	i, ii, iii, v	$P_e(\delta)$	$P_e(\delta^*) \leq P_e(\delta)$	Bayesian
Non Bayesian	i, ii, iii	$\sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\delta, \boldsymbol{\theta})$	$\sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\delta^*, \boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\delta, \boldsymbol{\theta})$	Minimax
	ii, iii	$P_{\theta\in\Theta_1}(\delta)$ subject to $\sup_{\theta\in\Theta_0} P_\theta(\delta) \leq \alpha$	$P_{\theta\in\Theta_1}(\delta^*) \geq P_{\theta\in\Theta_1}(\delta)$ and $\sup_{\theta\in\Theta_0} P_\theta(\delta) \leq \alpha$	Neyman-Pearson

3.3 Classification of parameter estimation schemes

The basic ingredient that distinguishes the parameter estimation from the hypothesis testing is the dimension of the hypothesis space and the nature of the stochastic process corresponding to each alternative. In hypothesis testing the dimension of the hypothesis space is finite and any of the M alternatives are represented by one stochastic process. In parameter estimation, we are face to an infinite number of alternatives represented by some m dimentional vector parameter $\boldsymbol{\theta}$ that takes its values in $\mathcal{T}$.

The basic elements of parameter estimation are then the vector parameter $\boldsymbol{\theta}$ and a stochastic process $X(t)$ which is parameterized by $\boldsymbol{\theta}$ and we still can distinguish two cases:

- If for a fixed $\boldsymbol{\theta}$ the stochastic process is well-known, then we have a *parametric* parameter estimation scheme.
- If for a fixed $\boldsymbol{\theta}$ the stochastic process is a member of some class $\mathcal{F}_{\boldsymbol{\theta}}$ of processes, then we have a *non parametric* or *robust parameter estimation scheme.*

In both cases, the main assumption is that the value of the parameter and so the nature of the stochastic process remains unchanged during the observation time $[0, T]$. The main objective is then to determine the active value of the parameter $\boldsymbol{\theta}$. Given a set of data $\boldsymbol{x}$, the solution is noted $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ and is called *parameter estimate.*

Between the different criteria to measure the performances of an estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$, one can mention the following:

- *Bias* :
  For a real valued parameter vector $\boldsymbol{\theta}$, the Euclidean norm
  $$\left\| \boldsymbol{\theta} - \mathrm{E}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{X})\right] \right\|^{1/2}$$
  is called the *bias* of the estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ at the process. If the bias is zero for all $\boldsymbol{\theta} \in \mathcal{T}$, then the estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ is called *unbiased* at the process.
- *Conditional variance* :
  The quantity
  $$\mathrm{E}\left[\left\| \widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \mathrm{E}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{X})\right] \right\| \mid \boldsymbol{\theta} \right]$$
  is called the *Conditional variance* of the estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$.

In general, the bias and the conditional variance present a tradeoff. Indeed, an unbiased estimate may induce a relatively large variance, and very often, admitting a small bias may result in a significant reduction of the conditional variance.

A parameter estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ is called *efficient* at the process, if the conditional variance equals a lower bound known as the *Cramer-Rao bound.*

A more general criterion is here also the expected penalty, if we define a penalty function $c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}]$– a scalar, non negative function whose values vary as $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ vary in $\mathcal{T}$. We can then define:

- Conditional expected penalty or conditional risk function:

$$R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}] \mid \boldsymbol{\theta}\right] = \int_{\Gamma} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}]\, f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{3.39}$$

  where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the probability density function of the stochastic process defined by $\boldsymbol{\theta}$ at the point $\boldsymbol{x}$.

- Expected penalty or Bayes risk function:
  When an *a priori* probability density function $\pi(\boldsymbol{\theta})$ is available, we can calculate the expected value of $R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathcal{T}$, and thus definec the total expected penalty or the Bayes risk function by

$$r(\widehat{\boldsymbol{\theta}}) = r(\widehat{\boldsymbol{\theta}}, \pi) = \int_{\mathcal{T}} R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{3.40}$$

Now, let try to make a classification of parameter estimation schemes. For this, we list the richest possible set of assets:

i. A parametric or nonparametric description of a stochastic process depending on a finite dimensional parameter vector $\boldsymbol{\theta}$.

ii. A set of data $\boldsymbol{x}$ which is assumed to be a realization of one of the active stochastic processes with the implicite assumption that this process remains unchanged during the observation time.

iii. A parameter space $\mathcal{T}$ where $\boldsymbol{\theta}$ takes its values.

iv. An *a priori* probability distribution $\pi(\boldsymbol{\theta})$ defined on the parameter space $\mathcal{T}$.

v. A penalty function $c[\widehat{\theta}(\boldsymbol{x}), \boldsymbol{\theta}]$ defined for each data sequence $\boldsymbol{x}$, parameter vector $\boldsymbol{\theta}$ and the estimated parameter vector $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$.

Here also, some of the assets listed above may not be available and we will see how different schemes come out from partial availability of these assets.

The minimum set of assets that is (or must be) always available consists of those in i, ii and iii and the performance criterion of the estimation will suffer limitations as the number of remaining available assets decrease.

First we assume that a parametric description of the stochastic process is available. When all the assets are available, we will have the *Bayesian parameter estimation scheme* where the performance criterion is the expected penalty or the Bayes risk

$$r(\widehat{\boldsymbol{\theta}}) = \mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}]\right] = \int_{\mathcal{T}} \int_{\Gamma} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}]\, f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{\theta} \tag{3.41}$$

with respect to the estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ for all $\boldsymbol{x}$ in the observation space Γ.

The Bayesian optimal estimate is then defined as the estimate $\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x})$ which minimizes the expected penalty function, *i.e.*

$$\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}): \quad r(\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x})) \leq r(\boldsymbol{\theta}(\boldsymbol{x})) \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.42}$$

If assets in i to iv are available, then we can calculate the posterior probability density function of $\boldsymbol{\theta}$, using the Bayes rule:

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})} = \frac{f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})} \tag{3.43}$$

where

$$m(\boldsymbol{x}) = \int_{\mathcal{T}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{3.44}$$

and define an estimate, called *maximum a posteriori (MAP)* estimate by

$$\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}): \quad \pi(\widehat{\boldsymbol{\theta}}^*|\boldsymbol{x}) \geq \pi(\boldsymbol{\theta}|\boldsymbol{x}) \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.45}$$

or written differently

$$\widehat{\boldsymbol{\theta}}^* = \arg\max_{\boldsymbol{\theta}\in\mathcal{T}} \{\pi(\boldsymbol{\theta}|\boldsymbol{x})\} \tag{3.46}$$

If assets in i to iii and v are available, then an optimal estimate $\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x})$ can be defined by using the expected conditional penalty

$$R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}]|\boldsymbol{\theta}]\right] = \int_{\Gamma} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}] f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{3.47}$$

where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the probability density function of the stochastic process defined by $\boldsymbol{\theta}$ at the point $\boldsymbol{x}$. We are then in the *minimax parameter estimation* scheme which is based on the saddle-point game formalization, with payoff function the expected penalty $r(\widehat{\boldsymbol{\theta}}, \pi)$ and with variables the parameters estimate $\widehat{\boldsymbol{\theta}}$ and the *a priori* probability density function π.

In summary, we can say that, if a minimax estimate $\widehat{\boldsymbol{\theta}}^*$ exists, it is an optimal Bayesian estimate at some least favorable *a priori* probability distribution p_0, *i.e.*

$$\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}): \exists \pi_0: \quad r[\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}), \pi] \leq r[\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}), \pi_0] \leq r[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \pi_0] \quad \forall \boldsymbol{\theta} \in \mathcal{T} \text{ and } \forall \pi \tag{3.48}$$

When only the assets i to iii are available, then the analyst can use the induced conditional probability density function $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. The scheme is called *maximum likelihood* and the main idea is to use the induced conditional probability density function $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ as a function $l(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{x})$, called *likelihood* of the vector parameter $\boldsymbol{\theta}$ and define the maximum likelihood (ML) estimate $\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x})$ as the one who maximizes the likelihood $l(\boldsymbol{\theta})$, *i.e.*

$$\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}): \quad l(\widehat{\boldsymbol{\theta}}^*) \geq l(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.49}$$

All the above schemes comprise the class of parametric parameter estimation procedures with the common characteristic of the assumtion that the stochastic process that generates the data is parametrically well-known.

When, for a given vector parameter $\boldsymbol{\theta}$, the stochastic process is nonparametrically described, then the parameter estimation is called *nonparametric* or sometimes *robust*. As in minimax scheme, the robust estimation scheme uses a saddle-point game procedure, but here the payoff function originates from the likelihood. So, in this scheme, in addition to the nonparametric description of the stochastic process, the only assets ii and iii are used to define a performance criterion using the likelihood function. We will be back more in details on this scheme in future chapters.

Classes of Parameter Estimation Schemes for Parametrically Known Stochastic Processes.			
Scheme	A priori	Cost	Decision rule
Bayesian	Yes	Yes	Minimization of expected penalty $r(\widehat{\boldsymbol{\theta}})$
	Yes	No	Maximization of the posterior probability $\pi(\boldsymbol{\theta}\|\boldsymbol{x})$
Non Bayesian	No	Yes	Minimax estimation using $r(\widehat{\boldsymbol{\theta}}, \pi)$
	No	No	Maximum likelihood tests using $l(\boldsymbol{\theta})$

Classes of Parameter Estimation Schemes			
Assets used	Optimization function	Optimal estimate $\widehat{\boldsymbol{\theta}}^*$	Scheme
i, ii, iii, iv, v	$r(\widehat{\boldsymbol{\theta}})$	$\widehat{\boldsymbol{\theta}}^* : \quad r(\widehat{\boldsymbol{\theta}}^*) \leq r(\widehat{\boldsymbol{\theta}}) \quad \forall \boldsymbol{\theta} \in \mathcal{T}$	Bayesian
i, ii, iii, iv	$\pi(\boldsymbol{\theta}\|\boldsymbol{x})$	$\pi(\widehat{\boldsymbol{\theta}}^*\|\boldsymbol{x}) \leq \pi(\widehat{\boldsymbol{\theta}}\|\boldsymbol{x}) \quad \forall \boldsymbol{\theta} \in \mathcal{T}$	MAP
i, ii, iv	$R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$	$\sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$	Minimax
i, ii	$l(\boldsymbol{\theta})$	$\widehat{\boldsymbol{\theta}}^* : \quad l(\widehat{\boldsymbol{\theta}}^*) \geq l(\widehat{\boldsymbol{\theta}}) \quad \forall \boldsymbol{\theta} \in \mathcal{T}$	Maximum likelihood
i, ii nonparametric description of the stochastic process	Based on $l(\boldsymbol{\theta})$	Appropriate saddle point optimization	Robust estimation

3.4 Summary of notations and abbreviations

- $\delta = \{\delta_k, k = 1, \cdots, M\}$
  A decision rule (or a set of possible actions)

- $\Gamma = \{\Gamma_k, k = 1, \cdots, M\}$
  The partitions of the observation space Γ corresponding to the hypotheses $\{H_k\}$ and the decision rule δ

- $\mathcal{T} = \{\mathcal{T}_k, k = 1, \cdots, M\}$
  The partitions of the parameter space $\mathcal{T}$ corresponding to the hypotheses $\{H_k\}$ and the decision rule δ

- $\{\pi_i\}$
  A prior probability distribution for the hypotheses $\{H_i\}$

- $\pi(\theta)$
  A prior probability density function for a scalar parameter θ

- $\pi(\boldsymbol{\theta})$
  A prior probability density function for a vector parameter $\boldsymbol{\theta}$

- $\{\pi_i(\boldsymbol{\theta})\}$
  Conditional prior probability density functions for the vector parameter $\boldsymbol{\theta}$ under the hypothesis H_i

- $\{r_i(\boldsymbol{\theta})\} = \{\pi_i\, \pi_i(\boldsymbol{\theta})\}$
  Unconditional prior probability density functions for the vector parameter $\boldsymbol{\theta}$ under the hypothesis H_i

- $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})$
  Conditional probability density function of the observations for a given $\boldsymbol{\theta}$

- $l(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})$
  Likelihood function of $\boldsymbol{\theta}$ for a given data $\boldsymbol{x}$

- $\pi_(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})}$
  Posterior probability density function of $\boldsymbol{\theta}$ given the observations $\boldsymbol{x}$

- $m(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$
  Marginal distribution of the observations $\boldsymbol{x}$

- $\{c_{ki}\}$
  Penalty coefficients

- $\{c_{ki}(\boldsymbol{\theta})\}$ or $\{c_k(\boldsymbol{\theta})\}$
  Penalty functions

- $\{P_{ki}(\delta)\}$
  Conditional probabilities of the decision rule δ for a well known stochastic process

- $\{P_{ki,\boldsymbol{\theta}}(\delta)\}$ or $\{P_{k,\boldsymbol{\theta}}((\delta)\}$
  Conditional probabilities of the decision rule δ for a parametrically known stochastic process

- $\{Q_{ki}(\delta) = \pi_i P_{ki}(\delta)\}$
  Probabilities of the decisions in the decision rule δ

- $P_e(\delta) = \sum_{k \neq i} Q_{ki}(\delta)$
  Probability of the error due to the decision rule δ

- $P_d(\delta) = \sum_{k} Q_{kk}(\delta) = 1 - P_e(\delta)$
  Probability of the correct detection due to the decision rule δ

- $P_{fa}(\delta) = Q_{10}(\delta)$
  Probability of false alarm in a binary hypothesis testing

- $P_{fd}(\delta) = Q_{01}(\delta)$
  Probability of false detection in a binary hypothesis testing

- $P_e(\delta) = Q_{01}(\delta) + Q_{10}(\delta)$
  Probability of the error due to the decision rule δ in a binary hypothesis testing

- $P_d(\delta) = Q_{00}(\delta) + Q_{11}(\delta)$
  Probability of the correct detection due to the decision rule δ in a binary hypothesis testing

- Conditional expected penalty or Risk function

  $$R_i(\delta) = \sum_{k=1}^{M} c_{ki} P_{ki}(\delta)$$

  for a well known stochastic process

  $$R_{\boldsymbol{\theta}}(\delta) = \sum_{k=1}^{M} c_k(\boldsymbol{\theta}) P_{k,\boldsymbol{\theta}}(\delta)$$

  for a parametrically known stochastic process

- Expected penalty or Bayes risk function

  $$r_i(\delta) = \sum_{k=1}^{M} c_{ki} Q_{ki}(\delta)$$

  for a well known stochastic process

  $$r_{\boldsymbol{\theta}}(\delta) = \sum_{k=1}^{M} c_k(\boldsymbol{\theta}) Q_{k,\boldsymbol{\theta}}(\delta)$$

  for a parametrically known stochastic process