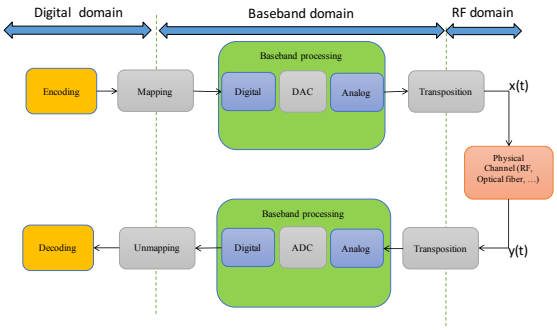
INSA de Lyon
Département Télécommunications,
Services et Usages

Communications Numériques Avancées (CNA) - Partie I



Edition 2020-2021
INSA de Lyon

Table des matières

Chapitre 1

Description générale du cours

1.1 Introduction

L'objectif d'un système de transmission est de permettre de transmettre une certaine quantité d'information entre un émetteur et un récepteur de façon fiable et sécurisée, tout en consommant un nombre minimal de ressources (énergie et bande passante). Le but de cette section est de décrire ce problème de façon très générale mais rigoureuse.

Nous pourrons alors définir les méthodes de transmission les plus robustes et démontrer leurs performances. Car si il est relativement facile et à la portée de tous, de développer un système de transmission qui fonctionne, réaliser un système qui soit efficace voire optimal, relève d'une démarche rigoureuse alliant mathématique et physique.

Le but de ce cours est d'aller plus loin qu'une description des systèmes mais bien de comprendre les principes théoriques à la base de leur conception. Nous allons montrer comment les outils mathématiques de la théorie de l'estimation de la théorie de la détection et de la théorie de l'information sont puissamment utilisés pour concevoir des systèmes extrêmement performants et même parfois proches de l'optimal. Sans être un cours de mathématiques, ce cours s'appuie sur des principes rigoureux, les met en oeuvre et permettra d'aller jusqu'à la réalisation de systèmes complets, tels qu'ils seront mis en oeuvre dans l'EC suivant, PSC.

Ce cours s'adresse à des étudiants de 2ième année du département Télécommunications, Services et Usages. Il vise à donner aux étudiants les bases nécessaires à la modélisation et l'évaluation des systèmes de communications numériques. La première partie du cours est consacrée aux outils fondamentaux de la théorie de l'estimation, de la détection et de l'information. La deuxième partie exploite ces outils pour modéliser et concevoir les systèmes modernes de communications numériques incluant les systèmes orthogonal frequency division multiplexing, accès multiple par division en fréquences orthogonales (OFDM), multiple input multiple output, entrées multiples sorties multiples (MIMO) et multi-utilisateurs.

Les étudiants qui suivent ce cours sont supposés avoir des connaissances solides en traitement du signal (filtrage, transformée de Fourier, transformée en Z, auto-corrélation, convolution), en probabilité et statistiques (densité de probabilité, probabilités conditionnelles). Ils connaissent les principaux blocs constituant les systèmes de transmission (modulation, codage, etc...). Le cours propose une approche assez théorique, en poursuivant l'objectif de donner à l'ingénieur les outils permettant de comprendre ce qu'est un système optimal. Un système de communications numériques n'est pas seulement un système qui marche, c'est

un système dont les performances sont pratiquement optimales. Cette notion d'optimalité est fondamentale, car dans un domaine extrêmement compétitif, et où les ressources (fréquences, énergie) sont précieuses, l'ingénieur se doit de concevoir un système optimal.

Nous étudions de façon détaillée les méthodes modernes de transmission : codage, égalisation, étalement de spectre, transmission multi-porteuses, diversité et systèmes MIMO.

Les éléments vus dans ce cours pourront être mis en oeuvre dans le module PSC où les étudiants mettent en oeuvre un système de transmission complet lors de projets en groupes. Les éléments de systèmes de communication vus dans ce cours seront indispensable pour le cours RAN du 2ieme semestre, qui s'intéresse aux réseaux radio multi-utilisateurs et en particuliers au réseaux radio-mobiles.

1.2 Compétences, capacités et connaissances attendues

Le but de ce cours est de vous faire acquérir les bases théoriques nécessaires à la compréhension, à la conception et à l'optimisation des systèmes de radiocommunication (communications numériques, canal radio, modélisation de performances).

1.2.1 Connaissances

Les connaissances développées sont :

— Estimation Bayésienne, estimateur optimal.

— Modélisation théorique en bande de base.

— Modélisation d'un canal radio en bande de base.

— Méthodes d'évaluation des performances d'un système de communication.

— Techniques de transmission en canal réel : codage canal, égalisation, diversité, étalement de spectre.

— Techniques de transmissions multi-porteuses : OFDM, peak to average power ratio, rapport de puissance crête à puissance moyenne (PAPR), préfixe cyclique, single carrier frequency division multiplexing, accès multiple par division en fréquences à porteuse unique (SC-FDMA).

— Techniques de transmissions MIMO : diversité, beamforming, capacité, codage spatio-temporel.

1.2.2 Capacités

A l'issu de ce cours vous devez être capables de :

— Concevoir un système de transmission suivant une démarche cohérente et validée, adapté à un canal donné.

— Analyser un standard : identifier les différentes fonction et leur rôle.

1.2.3 Auto-évaluation

Vous trouverez tout au long de ce document des exercices, dont certains seront faits en TD, mais pas tous, et qui ont vocation à vous permettre de vous auto-évaluer sur vos capacités. Certains éléments de corrections sont fournis en fin de chapitres.

Un QCM par chapitre est disponible en ligne et permet de vous évaluer sur vos connaissances élémentaires.

1.3 Notations et rappels

Nous résumons ici les principales notations utilisées tout au long du cours.

Notons que nous utilisons la notation := lorsqu'il s'agit de la définition d'une variable, et la notation = lorsqu'il s'agit d'une égalité qui découle des propriétés des opérandes.

Glossaire

ALOHA protocole d'accès aléatoire au medium. 41

AWGN bruit blanc additif gaussien –additive white gaussian noise–. 23, 26, 27, 36, 49, 51, 61, 69, 73

BER taux d'erreur binaire –bit error rate– . 19, 25, 45

BPSK modulation par déplacement de phase à deux états –binary phase-shift keying–. 50

BSC canal binaire symmétrique –binary symmetric channel–. 63

CRLB Cramer-Rao lower bound. 47

CSMA accès multiple par détection de porteuse –carrier sense multiple access–. 15

DMC discrete memoryless channel, canal discret sans mémoire. 63, 67, 68, 78

FA fausse alarme –false alarm–. 19, 26, 28, 31

GMC gaussian memoryless channel, canal gaussien sans mémoire. 62, 78

iid indépendantes et identiquement distribuées –independently and identically distributed– . 23, 24, 26, 35, 43, 45, 53–55, 58, 62

LDPC test de parité à faible densité –low density parity check–. 37

LRT test du log-vraisemblance –log-likelihood ratio test–. 23, 24, 53–55

MAC contrôle d'accès multiple –medium access control–. 15

MAE mean absolute error, erreur moyenne au sens de la valeur absolue. 39

MAP maximum a posteriori. 27, 28, 31, 37, 39–41, 43, 48, 52–54, 58, 59

MIMO multiple input multiple output, entrées multiples sorties multiples. 5, 6, 63, 75

ML maximum de vraisemblance –maximum likelihood–. 25–28, 31, 36, 37, 42–44, 46–49, 51–54, 57, 59, 69

MMAE minimum mean absolute error. 39–43, 59

MMSE erreur minimale au sens des moindres carrés –minimum mean square error–. 39, 40, 42, 43, 49, 58, 59

MMSPE erreur de prédiction minimale au sens des moindres carrés –minimum mean square prediction error–. 26, 27

MSE erreur moyenne au sens des moindres carrés –mean square error–. 39, 52

ND non detection –misdetection–. 19, 26, 28, 31

OFDM orthogonal frequency division multiplexing, accès multiple par division en fréquences orthogonales. 5, 6

PAPR peak to average power ratio, rapport de puissance crête à puissance moyenne. 6

PSK phase-shift keying. 50

QAM quadrature amplitude modulation. 37, 50, 80

QPSK quadrature phase-shift keying. 50, 80

RF radio-fréquences. 73

ROC caractéristiques opérationnelles du récepteur –receiver operating characteristics–. 19, 31, 32, 34, 36, 52–55, 57

RSSI received signal strength indicator, indicateur de puissance de signal reçu. 51

SC-FDMA single carrier frequency division multiplexing, accès multiple par division en fréquences à porteuse unique. 6

SNR signal to noise ratio, rapport signal à bruit. 80

v.a. variable aléatoire –random variable–. 15–17, 23, 24

1.3.1 Fonctions

$\lfloor a \rfloor$	:	partie entière de a.
$\text{sign}(a)$	:	fonction signe : -1 si $a < 0$, 0 si $a = 0$ et 1 si $a > 0$.
$\text{sinc}(t)$	:	fonction sinus cardinal de période 1.
$\text{rect}_T(t)$	:	fonction rectangle de période T.
$U(t)$	:	fonction échelon unité.
δ_{ij}	:	fonction delta $\delta_{ij} = 1$ si $i = j$, et $\delta_{ij} = 0$ sinon.
$\max.$	:	valeur max de l'argument de la fonction.
$\min.$	:	valeur min de l'argument de la fonction.
$\arg\max_{x \in \boldsymbol{\mathcal{X}}} f(x)$	:	argument du max de$f(x)$. C'est à dire que la fonction retourne la valeur (ou les valeurs) de x pour lequel $f(x)$ est maximal.
$\text{med}(f(x))$	:	valeur médiane de $f(x)$.
$\mathcal{H}(x(t))$	:	transformée de Hilbert de $x(t)$
$\mathbb{1}_{\{test\}}$	:	fonction indicatrice, égale à 1 si le test est vrai et à 0 sinon.
$Q(x)$	:	fonction d'erreur : $\text{Q}(x) := \frac{1}{\sqrt{2\pi}} \int_{z=x}^{\infty} \exp(-z^2/2) dz$
$\text{erfc}(x)$	:	fonction erreur complémentaire : $\text{erfc}(x) := \frac{2}{\sqrt{\pi}} \int_{z=x}^{\infty} \exp(-z^2) \cdot dz$.

remarque : la fonction $\text{Q}(x)$ est usuelle dans les documents anglophones, alors que la fonction $\text{erfc}(x)$ est utilisée dans les documents francophones. Elles sont liées par $\text{erfc}(x) = 2 \cdot \text{Q}\left(\sqrt{2} \cdot x\right)$.

1.3.2 Probabilités

Notations

— Soit X une variable aléatoire. x une réalisation de cette variable. Le domaine de définition de la variable aléatoire est notée $\boldsymbol{\mathcal{X}}$. Lorsqu'il est utile de distinguer plusieurs réalisations différentes, on pourra les indexer et les noter x_i.

— On note $\boldsymbol{X}$ les vecteurs aléatoires lorsqu'il est nécessaire de faire apparaitre explicitement que l'on travaille sur des vecteurs et non sur des scalaires.

— Tout ensemble à valeurs discrètes, ou tout domaine continu, est noté par une variable majuscule calligraphique, par exemple $\boldsymbol{\mathcal{E}}$, $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{X}}$.

— Les fonction de probabilité à valeurs discrètes sont notées $P_A(a)$, où A représente la variable aléatoire à valeurs dans $\boldsymbol{\mathcal{A}}$. $a \in \boldsymbol{\mathcal{A}}$ est une réalisation de A.

— Les fonction de densité de probabilité (à valeurs continues) sont notées $f_X(x)$, où X représente la variable aléatoire à valeurs dans $\boldsymbol{\mathcal{X}}$. $x \in \boldsymbol{\mathcal{X}}$ est une réalisation de X.

— La probabilité de X conditionnée à une réalisation y d'une autre variable aléatoire Y est notée $P_{X|Y}(x|y)$ lorsque la variable X est à valeurs discrètes, et sa densité est notée $f_{X|Y}(x|y)$ lorsque la variable X est à valeur continues.

— $\underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}}$ test d'hypothèse.

— lois de probabilité :

  — $\mathcal{N}(m, v)$: loi normale de moyenne m et de variance v.

  — $\underline{\mathcal{N}}(\underline{m}, v)$: loi normale complexe à symétrie circulaire est de moyenne $\underline{m}$ et de variance v.

Propriétés et théorèmes

Nous rappelons dans cette section les résultats de probabilité supposés connus pour ce cours. TODO : voir slides, compléter avec les principales propriétés (Bayes, etc...)

1.3.3 Systèmes linéaires

Concernant les matrices et les vecteurs, soit M une matrice et X un vecteur

— On note $X^\top$ la transposée de X et $X^\dagger$ sa transposée conjuguée.

— On note I_d la matrice identité.

— La matrice de covariance d'un signal aléatoire $x(n)$ est notée K_x, où $k_x(i,j) = \mathbb{E}[x^*(i)x(j)]$. Si le signal est ergodique alors on a $k_x(i,j) = k_x(a, a+j-i)$, et si les échantillons sont indépendants, alors $K_x = \sigma_x^2 \cdot I_d$, où σ^2 est la puissance moyenne du signal.

TODO : a compléter

1.4 Variables

e	:	évènement à retrouver (v.a.).
$\hat{e}$	:	évènement estimé (v.a.).
$f_{enc}(w)$ et $f_{dec}(y)$	:	fonctions d'encodage et de décodage associées à un système de transmiss
f_c	:	fréquence porteuse (ou modulant)
x	:	signal à support temporel fini, ou vecteur.
$\underline{x}_{BB}$	:	ou simplement $\underline{x}$ ou $\underline{x}(t)$ signal continu en bande de base (donc complexe
x_{RF}	:	signal radio-fréquence .
$\underline{x}_{RF}$	:	signal analytique associé au signal x_{RF}.
$\hat{x}_{RF}$	:	transformée de Hilbert du signal x_{RF}.
C_{ij}	:	coût associé à la décision e_i lorsque l'évènement réel est e_j.
C_{fp}	:	coût associé à un faux positif.
C_{nd}	:	coût associé à une non détection.
L_s	:	Latence maximale
P_{fp}	:	probabilité d'erreur de fausse alarme (faux positif).
P_{nd}	:	probabilité de non-détection (faux négatif).
Q_c	:	Quantité d'information par paquet de couche physique
Q_s	:	Quantité d'information insécable
R	:	rate, débit
T_c	:	
$\Lambda(y)$	:	rapport de vraisemblance.
$\boldsymbol{\mathcal{E}} = \{e_0, e_1, \ldots, e_N\}$	:	domaine de définition des évènements à estimer (estimation-détection).
$\boldsymbol{\mathcal{W}} = \{w_1, w_2, \ldots, w_N\}$	:	dictionnaire de la source (théorie de l'information).
$\boldsymbol{\mathcal{X}}$	:	domaine de définition des signaux d'entrée (communications numériques)
$\boldsymbol{\mathcal{Y}}$	:	domaine d'observation (estimation-détection, communications numérique
$\mathbb{H}_k$	:	hypothèse k.
$\mathfrak{R}$	:	règle de décision.

Chapitre 2

Estimation-détection

2.1 Introduction

Ce chapitre a pour but de poser rigoureusement les bases de la théorie de l'estimation et de la détection, utilisée pour la conception et l'évaluation des systèmes de communication. Mais il faut souligner que l'usage de cette théorie va bien au-delà des télécommunications. On peut citer les applications dans le domaine des radars, et plus généralement dans tous les problèmes d'observation de systèmes complexes : formation d'image, astronomie, médecine, etc...

Cette théorie doit beaucoup aux travaux de Thomas Bayes (mathématicien anglais, 1702-1761), qui a écrit *Essay towards solving a problem in the doctrine of chances* , publié après sa mort en 1764 (Encyclopaedia Universalis).

Ces outils théoriques sont très liés au traitement du signal ou de l'image. Plusieurs livres traitent de l'estimation et de la détection sous l'angle du traitement du signal, voir par exemple le livre du Professeur H. Vincent Poor de l'Université de Princeton [9] qui est une référence sur le sujet, sur lequel s'appuie en partie les deux premiers chapitres de cours. On peut également suggérer la lecture de [5] ou en français de [2, 4].

En ce qui concerne l'application spécifiquement aux télécommunications, vous pourrez trouver en ligne les supports de cours de Philippe Ciblat (IMT-Paris) ou encore de Christian Jutten (Polytech Grenoble).

Comme dit précédemment, ces outils théoriques ont une portée beaucoup plus large que les télécommunications : observation spatiale, localisation, poursuite de cibles, analyse statistique en particulier dans le domaine médical. Enfin, ces modèles jouent également un rôle crucial dans le développement de ce qu'on appelle aujourd'hui l'apprentissage machine ou l'intelligence artificielle.

Pour ceux d'entre vous qui portent un intérêt au traitement du signal et à la théorie de la décision, je vous invite à creuser ce domaine, au-delà du cours. Certains exercices d'approfondissement sont là pour vous aider à aller plus loin.

Bien entendu, toutes ces notions sont fortement liées aux signaux aléatoires. Vous pouvez consulter le livre de référence de Bernard Picinbono [8] pour une présentation détaillée des signaux aléatoires et de leurs propriétés, mais vous avez acquis les bases essentielles pour comprendre ce chapitre, dans les cours PBS, CMN ou TSA, que je vous invite à relire au début de ce cours. Dans ce chapitre, nous ne parlerons pas explicitement de télécommunications, et sa portée est donc générale. Les éléments développés dans ce

chapitre seront utilisés dans tous les autres chapitres du cours pour construire, optimiser et justifier les éléments des systèmes de communication numérique.

Nous commençons par poser formellement ce que sont les problèmes de détection et d'estimation, puis nous étudions plusieurs approches pour les résoudre (estimation Bayésienne, estimation de Neyman-Pearson).

2.2 Définition d'un problème de détection

La problématique centrale d'un problème de détection est de déterminer si un évènement donné a eu lieu. Pour un radar, il s'agit de déterminer si une cible est présente, pour un système de communication, il peut s'agir de déterminer le message qui a été envoyé, en médecine, on souhaite évaluer si un médicament a un effet positif sur tel virus.

L'objectif de poser un cadre théorique pour un tel problème, est de répondre à cette question avec la plus grande fiabilité possible, ou mieux, d'adapter l'expérience afin d'obtenir la fiabilité voulue. La théorie de la détection repose sur les probabilités et l'analyse statistique, qui offrent tous les outils permettant de prendre une décision, en contrôlant le risque d'erreur.

En théorie de la détection, l'espace des évènements est un espace discret (le nombre d'évènement possible est fini et comptabilisable), noté :

$$\boldsymbol{\mathcal{E}} := \{e_1, e_2, \ldots, e_n\}\,.$$

Pour en faire un espace probabilisé, il faut définir une *tribu*[1] $\boldsymbol{\mathcal{T}}$ sur $\boldsymbol{\mathcal{E}}$, et associer à chaque élément $\boldsymbol{\mathcal{T}}_k$ de $\boldsymbol{\mathcal{T}}$, une mesure noté $\mathbb{P}(\boldsymbol{\mathcal{T}}_k)$, telle que $\mathbb{P}(\boldsymbol{\mathcal{E}}) = 1$. L'espace probabilisé correspondant est noté $(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{T}}, \mathbb{P})$.

La variable aléatoire E est définie par la fonction $\mathbb{P}(e_k)$ pour tout $k \in \{1, \ldots, n\}$, qui est la probabilité de l'évènement $E = e_k$.

Exercice 2.1 : Espace probabilisé

Il s'agit d'un petit exercice de mise en jambe et de rappel sur les probabilités. Soit un jeu de lancé de dés, constitué de 3 dés à 6 faces, numérotées de 1 à 6.

1. Identifiez $\boldsymbol{\mathcal{E}}$. Donnnez son cardinal, noté $|\boldsymbol{\mathcal{E}}|$.
2. Identifiez $\boldsymbol{\mathcal{T}}$. Donnnez son cardinal, noté $|\boldsymbol{\mathcal{T}}|$.
3. Quelle est la probabilité $\mathbb{P}$ associée à chaque évènement de $\boldsymbol{\mathcal{E}}$?
4. Quelle est la probabilité $\mathbb{P}$ associée à chaque élément de $\boldsymbol{\mathcal{T}}$?

A un instant donné, l'évènement E étant aléatoire, sa réalisation n'est pas connue. Cependant il est observé au travers d'un système bruité. L'observation est effectuée dans le domaine d'observation $\mathcal{Y}$, qui lui, peut être à valeurs discrètes ou continues, suivant le problème étudié. La principe de la détection, est d'effectuer des hypothèses H_i, et choisir parmi les hypothèses celle qui est la plus probable.

Comme déjà mentionné, de très nombreux problèmes peuvent être formulés de la sorte :

1. une tribu sur un ensemble $\boldsymbol{\mathcal{E}}$ est un ensemble non vide de parties de $\boldsymbol{\mathcal{E}}$, stable par passage au complémentaire et par union dénombrable

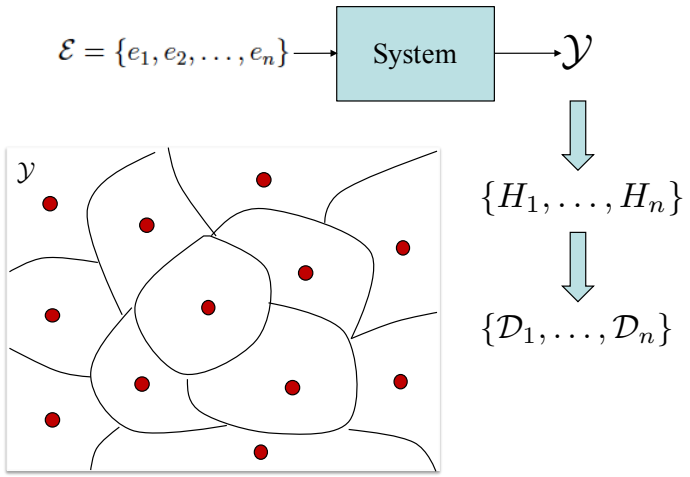


FIGURE 2.1 – Représentation du problème de détection optimale.

— Lorsqu'un radar émet des signaux de tests, il exploite les signaux reçus par réflexion pour déterminer la présence éventuelle d'une cible.

— Lorsqu'un télescope observe l'espace, il cherchera à détecter la présence ou l'absence d'une étoile ou d'un objet céleste, à une distance donnée.

— Lorsqu'un capteur de recul embarqué sur un véhicule, observe les signaux réfléchis par l'environnement, il doit déterminer si il y a un obstacle dans son champ de vision.

Le problème de détection est une composante essentielle dans les processus de décision ou dans les approches par apprentissage, ou en intelligence artificielle.

Plus généralement, le raisonnement probabiliste est un élément essentiel du processus de décision en ingénierie.

En télécommunications, si l'on considère un message reçu, l'espace $\boldsymbol{\mathcal{E}}$ représente l'ensemble des messages possibles, la variable aléatoire –random variable– (v.a.) E représente donc le mot code émis par la source. Mais ce n'est pas le seul cas d'utilisation de la théorie de la détection. La théorie de la détection est présente également dans les protocoles contrôle d'accès multiple –medium access control– (MAC) et dans les algorithmes de gestion des ressources radio, en particulier pour la détection de présence de signal, pour la détection d'un code d'étalement ou simplement pour détecter si le canal est libre, en accès multiple par détection de porteuse –carrier sense multiple access– (CSMA).

Après avoir probabilisé l'espace (discret) des évènements, la théorie de la détection repose sur la modélisation probabiliste du système d'observation lui-même. Lorsque l'espace d'observation est discret et comptabilisable ($\boldsymbol{\mathcal{Y}} = \{y_1, \ldots, y_M\}$), le système est caractérisé par un ensemble de probabilités conditionnelles, reliant l'observation à l'évènement :

$$\begin{array}{ccccc} P_{Y|E}(.|.) & : & \{1,\ldots,n\} \times \{1,\ldots,m\} & \longrightarrow & [0,1] \\ & & (i,j) & & P_{Y|E}(y_j|e_i) \end{array}.$$

Cette fonction se lit comme la probabilité d'observer y_j sachant que l'évènement e_i s'est réalisé. C'est une probabilité conditionnelle qui se nomme la vraisemblance de e_i. C'est une mesure qui ne dépend que du système d'observation et non des probabilités a priori sur $\boldsymbol{\mathcal{E}}$.

Lorsque l'espace d'observation est un espace continu, non dénombrable, la fonction de

densité de probabilité [2] est utilisée. Le système d'observation est alors décrit par la loi de densité conditionnelle (également appelée vraisemblance) :

$$\begin{array}{rccc} f_{Y|E}(.|.) & : & \{1,\ldots,n\}\times\boldsymbol{\mathcal{Y}} & \longrightarrow & [0,1] \\ & & (i,y) & & f_{Y|E}(y|e_i) \end{array}.$$

Notez que les fonctions $P_{Y|E}(y|e_i)$ et $f_{Y|E}(y|e_i)$ sont indexées par la v.a. associée, ici $Y|E$ (qui se lit Y sachant E). L'argument $y|e_i$, indique la réalisation y sachant que l'évènement est e_i. Cette notation est un peu lourde, mais permet de bien préciser la loi par rapport à laquelle sont définies les probabilités. La notation majuscule indique une v.a. dont la loi est supposée connu. La notation minuscule indique une réalisation de cette v.a..

Exercice 2.2 : Vraisemblance de variables discrètes

Continuons avec le jeu de lancé de dés, constitué cette fois de 2 dés à 6 faces, numérotées de 1 à 6. Après avoir lancé les 2 dés, un partenaire vous donne uniquement la somme des valeurs obtenues.

1. Quel est le domaine d'observation $\boldsymbol{\mathcal{Y}}$?
2. Déterminez la vraisemblance $P_{Y|E}(y|e)$ pour tous les évènements possibles.
3. Déterminez la loi $P_Y(y)$.
4. Dans quels cas peut-on déterminer avec exactitude la valeur de chaque dé ?
5. Dans quels cas peut-on déterminer avec exactitude le couple de valeurs ?
6. Quel est le pire cas en termes de prise de décision ?

Votre partenaire ajoute un aléa. Il utilise une pièce de monnaie pour décider si il ajoute un aléa : soit il vous donne la vraie valeur (pile), soit pour chaque dé, il prend le complément à 7 et vous donne la somme (face). Par exemple, considérons que les dés ont donné (6,5). Votre partenaire vous indique alors 11 si il a fait pile, et 3 si il a fait face. Bien entendu, vous ne connaissez pas le résultat de la pièce.

7. Reprennez les questions ci-dessus.

La notion de vraisemblance vraiment essentielle pour la suite peut se référer tant à une densité qu'à une probabilité, en fonction des caractéristiques de l'espace d'observation.

Dans le cas de la détection bayésienne, ces probabilités conditionnelles sont supposées connues (en communications numériques, on dira que le canal est connu).

Nous pouvons maintenant formuler le problème de la détection optimale. L'observateur effectue un ensemble d'hypothèses possibles, notées $\mathbb{H}_1,\ldots\mathbb{H}_n$, et veut déterminer la plus vraisemblable : il réalise un *test d'hypothèses*. L'observateur retient l'hypothèse qui minimise un certain risque qu'il nous reste à préciser.

Notons que les hypothèses faites peuvent être associées à des sous-ensembles d'évènements (donc des éléments de $\mathcal{A}$) . A partir de l'exercice 2.2, on peut s'intéresser à la probabilité que les deux dés aient la même valeur. Il s'agit dans ce cas de calculer la vraisemblance de l'ensemble $\mathcal{A}_k = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$, sous condition d'une observation donnée.

2. Nous ne discutons pas ici des conditions d'existence de cette densité et supposons qu'elle existe sur les espaces d'observation étudiés dans ce cours

L'objectif de ce chapitre est d'exploiter cette notion de vraisemblance pour prendre la meilleure décision possible, et la quantifier. Nous allons voir que mathématiquement, la solution peut être exprimée assez facilement, bien que parfois très difficile à calculer explicitement. Une approche exhaustive (tester toutes les entrées possibles) est rarement efficace, car le nombre de possibilités est très grand. Prenons le cas de la transmission d'un paquet de 1000 bits d'information, le nombre d'évènements possibles est égal à 2^{1000}. Si l'on détermine chaque bit indépendamment des autres, alors il y a seulement 1000 tests binaires à faire, ce qui est beaucoup plus simple. Malheureusement, en communications numériques, les observations sont liées entre elles par le codage, la modulation et l'effet du canal, et une telle approche n'est pas optimale.

La formulation du problème est la suivante :

Définition 2.1 : Détection-décision

Soit un espace probabilisé $(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{T}}, \mathbb{P})$ et une v.a. sur cet espace notée E. Soit un système d'observation de E, à valeurs dans $\boldsymbol{\mathcal{Y}}$, défini par sa vraisemblance $f_{Y|E}$ (cas continu) ou $P_{Y|E}$ (cas discret).

Soit un ensemble d'hypothèses formulées par l'observateur $\{\mathbb{H}_n; n \in [1, \ldots, N]\}$. Dans le cas général, l'ensemble des hypothèses constitue une partition de $\boldsymbol{\mathcal{E}}$, (par exemple dan le cas du jet de dés, on peut faire deux hypothèses : pair ou impair). Mais dans ce cours,par défaut, chaque hypothèse correspondra à un évènement de $\boldsymbol{\mathcal{E}}$.

Une fonction de décision déterministe (qui implémente une règle de décision $\mathfrak{R}$) fait correspondre à toute observation $y \in \boldsymbol{\mathcal{Y}}$, un choix unique parmi les hypothèses.

Définition 2.2 : Sous-espaces de décision

Dans un problème de décision, l'ensemble des points qui conduisent à la même décision $\hat{E} = e_i$ est appelé le sous-espace de décision associé à l'évènement e_i, noté $\mathcal{D}_i$. L'ensemble des $\mathcal{D}_i$ constitue une partition de $\boldsymbol{\mathcal{Y}}$ (voir figure 2.1).

Exercice 2.3 : Canal binaire Gaussien

Considérons le cas d'un évènement binaire (ON/OFF) probabiliste E sur $\boldsymbol{\mathcal{E}} = \{e_0, e_1\}$. Il s'agit par exemple d'un capteur automobile de recul, qui sur la base d'une observation y doit déterminer la présence éventuelle d'un obstacle.

L'observation est effectuée au travers d'une fonction déterministe $h(E)$ (la fonction de transfert du capteur et de l'ensemble de la chaîne de mesure), perturbée par un bruit additif noté N, qui est également une v.a. :

$$Y = h(E) + N.$$

Considérons que N est une variable continue, de densité de probabilité $f_N(n)$.

1. Calculez la fonction de vraisemblance $f_{Y|E}(y|e_i)$, pour $i \in \{1, 2\}$.

La fonction $h(E)$ étant déterministe, la vraisemblance $f_{Y|E}(y|e_i)$ est donnée par :

$$f_{Y|E}(y|e_i) = f_N(y - h(e_i)).$$

En effet, du point de vue de l'observation et pour un évènement e_i donné, $h(e_i)$ est une constante. La distribution de Y et égale à celle de N, décalée de $h(e_i)$.

Les deux fonctions de vraisemblance, en fonction de y, sont illustrées à la figure 2.2 pour une observation $y = x + n$, où $\boldsymbol{\mathcal{X}} = \{-2, 2\}$ et où n suit une loi normale $\mathcal{N}(0, 1)$.

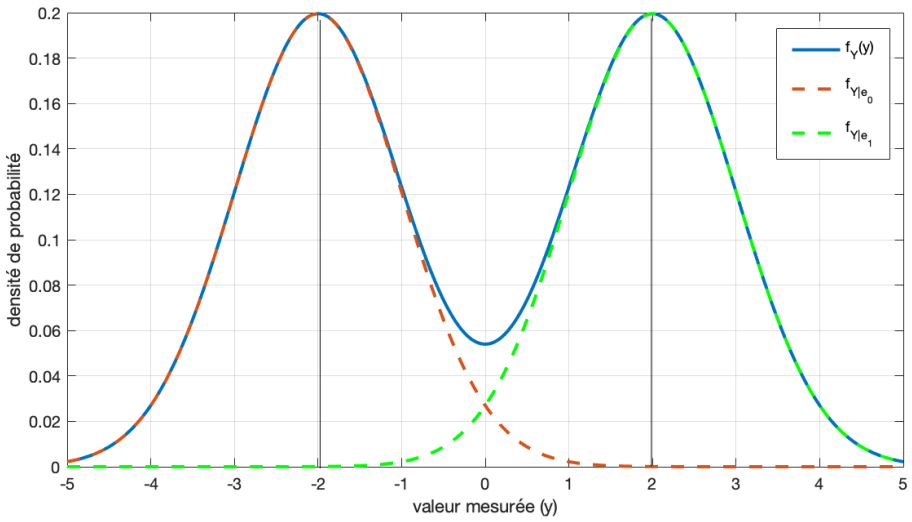


FIGURE 2.2 – densités a posteriori $f_{Y|E}(y|e_0)$ et $f_{Y|E}(y|e_1)$ où l'évènement e_0 correspond à un signal d'entrée $x = -2$ et l'évènement e_1 à $x = 2$. La densité $f_Y(y)$ est également représentée (avec un facteur 2).

On observe dans cet exemple que les lois des vraisemblance sont régies par les propriétés du bruit additif.

Une fois le problème posé, on souhaite développer un processus de décision, qui à partir d'une observation y, retourne la valeur de E la plus vraisemblable. Un question additionnelle, est de savoir, lorsque la mesure peut être répétée K fois, comment prendre une décision à partir de ce vecteur observé et comment déterminer le nombre de mesures à réaliser. La suite du cours va permettre de répondre à ces questions.

2.3 Détection binaire

Nous allons construire les règles de décision optimales, dans le cas relativement simple où l'espace d'évènements est binaire, $\boldsymbol{\mathcal{E}} := \{e_0, e_1\}$. C'est typiquement le cas d'un capteur de fumée qui surveille une certaine zone et qui doit déclencher une alerte en cas de fumée (e_1), à partir d'un signal bruité. Il fait deux hypothèses. On note $\mathbb{H}_0$ l'hypothèse qu'il n'y a pas de fumée, et $\mathbb{H}_1$ l'hypothèse contraire. Ces hypothèses sont formulées à partir d'une observation Y (scalaire ou vectorielle), mesurée dans un certain domaine $\boldsymbol{\mathcal{Y}}$. A titre d'exemple, considérons le cas où Y est une tension, mesurée aux bornes d'un capteur de fumée, et prenant ses valeurs sur $\boldsymbol{\mathcal{Y}} := [0, 5]$ Volts.

Le capteur doit prendre une décision à partir des observations et construit une décision sur E, notée $\hat{E}$. 4 cas sont possibles :

— l'évènement est $E = e_0$, et l'hypothèse choisie est $\mathbb{H}_0$, i.e. $\hat{E} = e_0$: pas d'erreur ; vrai négatif.

— l'évènement est $E = e_0$, et l'hypothèse choisie est $\mathbb{H}_1$, i.e. $\hat{E} = e_1$: erreur dite de faux positif, souvent appelée fausse alarme –false alarm– (FA).

— l'évènement est $E = e_1$,l'hypothèse choisie est$\mathbb{H}_0$, i.e. $\hat{E} = e_0$: erreur dite de faux négatif, ou encore non detection –misdetection– (ND).

— l'évènement est $E = e_1$, et l'hypothèse choisie est $\mathbb{H}_1$, i.e. $\hat{E} = e_1$: pas d'erreur ; bonne détection (vrai positif).

Chacun des états conjoints est notée (e_i, e_j), avec $i, j \in \{0, 1\}$. Nous avons donc deux cas de mauvaise décision et deux cas de bonnes décision.

Le sens des erreur de FA et de ND est particulièrement important dans le cas d'un problème de détection d'alerte, où l'état e_1 revêt un caractère particulièrement critique.

Au contraire, pour une transmission d'un bit d'information, où e_0 correspond au bit 0 et e_1 au bit 1, les deux types d'erreurs ont le même impact sur la qualité de transmission, et les deux types d'erreur ont donc le même poids et se mêlent dans le calcul du taux d'erreur binaire –bit error rate– (BER) :

$$Pr(err) = P(e_0) \cdot Pr(fa) + P(e_1) \cdot Pr(md).$$

Nous allons voir que la définition d'une bonne règle de décision repose sur un compromis entre probabilité de bonne détection et probabilité de faux positif, qui est représentée au travers de la courbe ROC. L'approche de Neyman-Pearson ou l'approche Minimax permettent alors de prendre la décision finale (voir section 2.4).

Lorsque les probabilités a priori des évènements $P(e_j)$ sont connues, ainsi que les coûts associés aux erreurs, le détecteur optimal est le détecteur bayésien, par lequel nous allons commencer.

2.3.1 Détecteur bayésien

L'approche bayésienne repose sur la définition du risque de Bayes.

Pour chacune des 4 situations définies ci-dessus ($E = e_j \rightarrow \hat{E} = e_i$), nous supposons connu un coût C_{ij}, qui a pour simple contrainte d'être plus élevé en cas d'erreur : $C_{ij} > C_{jj}$. C_{00} et C_{11} correspondent aux coûts associés aux bonnes décisions et C_{01} et C_{10} aux coûts associés aux mauvaises décisions.

Par exemple, pour un canal de communication binaire, le coût représente les erreurs, et on aura $C_{00} = C_{11} = 0$ et $C_{01} = C_{10} = 1$.

Définition 2.3 : Risque de Bayes

Soit une règle de décision $\mathfrak{R}$ qui à toute observation y fait correspondre une décision $\hat{e} : \mathfrak{R} : y \in \boldsymbol{\mathcal{Y}} \rightarrow \hat{e} \in \boldsymbol{\mathcal{E}}$. Il s'agit d'une règle de décision déterministe.

Dans certains cas, il peut être utile d'introduire la notion de règle randomisée, c'est à dire que la décision est probabiliste : $\mathfrak{R}^{(r)} : y \in \boldsymbol{\mathcal{Y}} \rightarrow Pr(e) \forall e \in \boldsymbol{\mathcal{E}}$.

Pour une règle déterministe, le coût global appelé *risque de Bayes* est défini comme suit :

$$Q(\mathfrak{R}) := \sum_{i \in \{0,1\}} \sum_{j \in \{0,1\}} C_{ij} P_{E\hat{E}}(e_i, e_j),$$

où $P_{E\hat{E}}(e_i, e_j)$ est la loi de probabilité conjointe des évènements E et $\hat{E}$.

La généralisation de cette définition du risque de Bayes au cas M-aire [3] est triviale : il suffit de sommer sur l'ensemble des éléments de $\boldsymbol{\mathcal{E}}$.

Définition 2.4 : Règle de Bayes

La règle de Bayes $\mathfrak{R}_{Bayes}$ est la règle de décision qui minimise le risque de Bayes.

Pour obtenir cette règle de décision, on va réécrire le risque de Bayes en y faisant apparaitre la vraisemblance.

La loi de Bayes [4] permet de réécrire le coût global sous la forme :

$$Q(\mathfrak{R}) = \sum_{i\in\{0,1\}} \sum_{j\in\{0,1\}} C_{ji} P_{\hat{E}|E}(e_j|e_i) \cdot P_E(e_i). \tag{2.1}$$

De plus, les lois de probabilités vérifient

$$\begin{aligned} P_{\hat{E}|E}(e_0|e_0) &= 1 - P_{\hat{E}|E}(e_1|e_0) \\ P_{\hat{E}|E}(e_1|e_1) &= 1 - P_{\hat{E}|E}(e_0|e_1), \end{aligned}$$

qui, injectées dans (2.1) donnent :

$$\begin{aligned} Q(\mathfrak{R}) &= C_{00} \cdot P_{\hat{E}|E}(e_0|e_0) \cdot P_E(e_0) + C_{01} \cdot P_{\hat{E}|E}(e_0|e_1) \cdot P_E(e_1), \\ &\quad + C_{11} \cdot P_{\hat{E}|E}(e_1|e_1) \cdot P_E(e_1) + C_{10} \cdot P_{\hat{E}|E}(e_1|e_0) \cdot P_E(e_0), \\ &= C_{00} \cdot P_E(e_0) + C_{11} \cdot P_E(e_1) \\ &\quad + (C_{10} - C_{00}) \cdot P_E(e_0) \cdot P_{\hat{E}|E}(e_1|e_0) + (C_{01} - C_{11}) \cdot P_E(e_1) \cdot P_{\hat{E}|E}(e_0|e_1). \end{aligned} \tag{2.2}$$

On peut réduire les degrés de liberté dans la formulation du problème. En effet, les deux premiers termes $C_{00} \cdot P_E(e_0)$ et $C_{11} \cdot P_E(e_1)$ ne dépendent pas de la règle de décision. Ce sont donc des constantes du problème.

(2.2) s'écrit :

$$Q(\mathfrak{R}) \propto C_{fp} \cdot P_0 \cdot P_{\hat{E}|E}(e_1|e_0) + C_{nd} \cdot P_1 \cdot P_{\hat{E}|E}(e_0|e_1), \tag{2.3}$$

où l'on a noté $P_0 := P_E(e_0)$, $P_1 := P_E(e_1)$, avec $P_0 + P_1 = 1$. $C_{fp} := (C_{10} - C_{00})$ et $C_{nd} := (C_{01} - C_{11})$ sont les différentiels de coût. Le premier correspond au surcoût associé à un faux positif, et le deuxième au surcoût d'une non-détection. Ce sont des termes positifs.

Pour faire apparaitre plus clairement le rôle de la règle de décision dans cette équation, il faut développer les probabilités conditionnelles et faire apparaitre la vraisemblance.

Dans le cas où le domaine d'observation est continu, on obtient :

$$\begin{aligned} P_{\hat{E}|E}(e_1|e_0) &= \int_{\mathcal{D}_1} f_{Y|E}(y|e_0) \cdot dy \\ P_{\hat{E}|E}(e_0|e_1) &= \int_{\mathcal{D}_0} f_{Y|E}(y|e_1) \cdot dy. \end{aligned}$$

3. On qualifie un problème de détection de M-aire lorsque l'espace d'évènement est de cardinalité M.
4. On rappelle la loi de Bayes :

$$P_{AB}(a, b) = P_{A|B}(a|b) \cdot P_B(b),$$

Et en insérant ces intégrales dans (2.3) :

$$Q(\mathfrak{R}) \propto C_{fp} \cdot P_0 \cdot \int_{\mathcal{D}_1} f_{Y|E}(y|e_0) \cdot dy + C_{nd} \cdot P_1 \cdot \int_{\mathcal{D}_0} f_{Y|E}(y|e_1) \cdot dy.$$

La même démarche peut être suivie lorsque l'espace d'observation $\boldsymbol{\mathcal{Y}}$ est discret. Vous trouverez en fin de chapitre l'exercice2.16 qui traite le cas du canal binaire, dont l'espace d'observation est discret. Ces développements nous amène à une réécriture importante du risque de Bayes :

Propriété 2.1 : Risque de Bayes

Le risque de Bayes associé à une règle de décision dans un problème de détection binaire, est donné, dans le cas où le domaine d'observation $\mathcal{Y}$ est continu, par :

$$Q(\mathfrak{R}) = \int_{\boldsymbol{\mathcal{Y}}} \left(\mathbb{1}_{\{y \in \mathcal{D}_1\}} \cdot C_{fp} \cdot P_0 \cdot f_{Y|E}(y|e_0) + \mathbb{1}_{\{y \in \mathcal{D}_0\}} C_{nd} \cdot P_1 \cdot f_{Y|E}(y|e_1) \right) \cdot dy,$$

Dans le cas discret, le risque de Bayes est donné par :

$$Q(\mathfrak{R}) = \sum_{y_i \in \boldsymbol{\mathcal{Y}}} \left(\mathbb{1}_{\{y \in \mathcal{D}_1\}} C_{fp} \cdot P_0 \cdot P_{Y|E}(y_i|e_0) + \mathbb{1}_{\{y \in \mathcal{D}_0\}} C_{nd} \cdot P_1 \cdot P_{Y|E}(y_i|e_1) \right).$$

Notons que pour tout y, soit $\mathbb{1}_{\{y \in \mathcal{D}_1\}} = 1$, soit $\mathbb{1}_{\{y \in \mathcal{D}_0\}} = 1$. Ainsi, pour toute observation y, et quelle que soit la règle de décision, il y a un coût à payer : soit $C_{fp} \cdot P_0 \cdot f_{Y|E}(y|e_0)$, soit $C_{nd} \cdot P_1 \cdot f_{Y|E}(y|e_1)$. La règle optimale consiste donc à affecter y au domaine qui minimsera le coût. Et ce n'est pas forcément l'évènement qui a la vraisemblance maximale, car les coûts et les probabilités a priori interviennent dans le calcul.

Trouver une bonne stratégie revient donc à partitionner l'espace d'observation $\boldsymbol{\mathcal{Y}}$ associé à Y en deux régions $\mathcal{D}_0$ et $\mathcal{D}_1$, telles que :

$$\boldsymbol{\mathcal{Y}} = \mathcal{D}_0 \cup \mathcal{D}_1.$$

avec $\mathcal{D}_0 \cap \mathcal{D}_1 = \emptyset$.

La bonne décision dépend donc des coûts et des probabilités a priori P_0 et P_1 associées à $\boldsymbol{\mathcal{E}}$.

Exercice 2.4 : Détecteur de fumée

Reprenons l'exemple du capteur de fumée. On suppose que le coût d'une non détection est estimé à $C_{nd} = 100$ et le coût d'un faux positif à $C_{fp} = 1$ (On estime dans cet exemple la non détection plus dommageable que le faux positif). On suppose d'autre part que la probabilité qu'une alerte se produise est égale à $P_1 = 1e^{-6}$. On mesure en sortie une tension comprise entre $[0, 5]$Volts. Le détecteur a un bruit additif Gaussien de variance $\sigma^2 = 0.04\text{Volts}^2$.

Quelle est la bonne décision à prendre en fonction de la tension mesurée ? Pour

chaque valeur de Y, il faut donc calculer les 2 coûts :

$$C_{fp} \cdot P_0 \cdot f_{Y|E}(y|e_0) = 1 \cdot (1 - 1e^{-6}) \cdot \frac{1}{\sqrt{2\pi}0,2} e \frac{-y^2}{0,08}$$
$$C_{nd} \cdot P_1 \cdot f_{Y|E}(y|e_1) = 100 \cdot 1e^{-6} \cdot \frac{1}{\sqrt{2\pi}0,2} e \frac{-(y-5)^2}{0,08}$$

Pour chaque valeur de y, il reste alors à choisir la meilleure hypothèse du point de vu du critère de Bayes. Cette règle minimise le coût, à condition de connaître parfaitement : les lois a priori des évènements (P_0 et P_1), les coûts associés aux différentes erreurs, et le modèle de bruit du système qui conduit aux lois de probabilités conditionnelles $f_{Y|E}$.

Rapport de vraisemblance

Nous venons de voir dans l'expression du coût global de la propriété 2.1, la bonne règle de décision dépend de la comparaison entre deux termes :

$$C_{nd} \cdot P_1 \cdot f_{Y|E}(y|e_1) \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} C_{fp} \cdot P_0 \cdot f_{Y|E}(y|e_0) \tag{2.4}$$

Effectuer cette comparaison s'appelle faire un test d'hypothèse (très utilisé dans toutes les sciences expérimentales : médecine, biologie, etc...).

Ce test d'hypothèse s'exprime à partir du rapport de vraisemblance :

Définition 2.5 : Rapport de vraisemblance

Le rapport de vraisemblance associé à un problème de détection binaire s'écrit :

$$\Lambda(y) := \frac{f_{Y|E}(y|e_1)}{f_{Y|E}(y|e_0)}$$

Le test d'hypothèse consiste à comparer ce rapport à une valeur seuil η. D'après (2.4), on a le théorème suivant :

Théorème 2.1 : Test du rapport de vraisemblance

Le test du rapport de vraisemblance est donné par :

$$\Lambda(y) \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \eta,$$

où le seuil optimal (qui minimise le risque de Bayes) est égal à :

$$\eta = \frac{C_{fp} \cdot P_0}{C_{nd} \cdot P_1}.$$

On écrit également ce test de vraisemblance sous la forme de la log-vraisemblance, l'intérêt étant principalement calculatoire, en particulier pour les signaux gaussiens :

Théorème 2.2 : Test de log-vraisemblance

Soit la log-vraisemblance donnée par $\log\left(f_{Y|E}\left(y|e_i\right)\right)$.
Le test du log-vraisemblance –log-likelihood ratio test– (LRT) s'écrit :

$$\log\left(\Lambda(y)\right) = \log f_{Y|E}\left(y|e_1\right) - \log f_{Y|E}\left(y|e_0\right) \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \log(\eta).$$

2.3.2 Cas d'une observation multi-dimensionnelle

L'observation est rarement limitée à une seule mesure. L'espace d'observation est souvent multidimensionnel et est associé à un vecteur d'observation

$$\boldsymbol{Y} = [Y[1], \ldots, Y[N]]^t.$$

Par exemple, un capteur de fumée peut effectuer plusieurs mesures successives avant de décider la présence ou non d'un incendie. Nous allons traiter l'exercice 2.13, très classique, qui s'appelle le test de position en canal bruit blanc additif gaussien –additive white gaussian noise– (AWGN).

Soit une v.a. X à valeurs dans $\boldsymbol{\mathcal{X}} = \{m_0, m_1\}$, mesurée au travers d'un système bruité, tel que

$$\boldsymbol{Y} = X + \boldsymbol{V},$$

où $V[k]$, avec $k \in [1; N]$, est une séquence de variables indépendantes et identiquement distribuées –independently and identically distributed– (iid) de loi normale $\mathcal{N}\left(0, \sigma^2\right)$.
Notez bien que Xest constant durant l'expérience, c'est donc un scalaire dans cette expression.

D'un point de vue applicatif, il peut s'agir d'un terminal qui mesure l'occupation d'un canal radio, où $m_0 = 0$ traduit l'absence de signal et m_1 représente le niveau de puissance attendu sur ce canal.

Le test d'hypothèse associé est le suivant :

$$\begin{array}{rcl} \mathbb{H}_0 & : & \boldsymbol{Y} = m_0 + \boldsymbol{V} \\ \mathbb{H}_1 & : & \boldsymbol{Y} = m_1 + \boldsymbol{V}, \end{array}$$

où l'on supposera sans perte de généralité que $m_1 > m_0$.

Du fait de l'indépendance entre échantillons de bruit, on obtient :

$$f_{\boldsymbol{Y}|E}\left(\boldsymbol{y}|e_i\right) = \prod_{k=1}^{N} f_{Y|E}\left(y[k]|e_i\right).$$

Comme le bruit est AWGN, la vraisemblance de $Y[k]$ est :

$$f_{Y|E}\left(y[k]|e_i\right) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{|y_k - m_i|^2}{2\sigma^2}\right).$$

La vraisemblance du vecteur observé est alors égale à :

$$f_{\boldsymbol{Y}|E}\left(\boldsymbol{y}|e_i\right) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \cdot \exp\left(-\frac{\sum_{k=1}^{N} |y_k - m_i|^2}{2\sigma^2}\right).$$

Finalement, le rapport de vraisemblance est donné par :

$$\Lambda(\boldsymbol{y}) = \exp\left(\frac{m_1 - m_0}{\sigma^2} \cdot \sum_{k=1}^{N} y_k - \frac{N(m_1^2 - m_0^2)}{2\sigma^2}\right).$$

Le rapport de log-vraisemblance associé est

$$\log\left(\Lambda(\boldsymbol{y})\right) = \frac{m_1 - m_0}{\sigma^2} \cdot \sum_{k=1}^{N} y_k - \frac{N(m_1^2 - m_0^2)}{2\sigma^2}.$$

Définissons la v.a. $S := \frac{1}{N}\sum_{k=1}^{N} Y[k]$. On montre alors aisément que le LRT se réécrit sous la forme :

$$S \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \gamma.$$

avec $\gamma = \frac{m_1+m_0}{2} + \frac{\sigma^2 \log(\eta)}{N(m_1-m_0)}$.

Remarque : vous vérifierez ce calcul.

Notons que si $P_0 = P_1$ et $C_{fa} = C_{nd}$, alors le seuil γ n'est autre que la valeur moyenne des deux valeurs possibles de X, $\gamma = \frac{m_1+m_0}{2}$. Il peut être intéressant d'étudier les lois conditionnelles de $S_{|X=m_0}$ et de $S_{|X=m_1}$. Dans le premier cas, on a

$$\begin{aligned} S_{|X=m_0} &= \frac{1}{N}\sum_{k=1}^{N} Y[k] \\ &= m_0 + \frac{1}{N}\sum_{k=1}^{N} V[k]. \end{aligned}$$

La somme dans cette expression est la somme de variables aléatoires iid normales. $S|X = m_i$ est donc une v.a. de loi $\mathcal{N}\left(m_i, \sigma^2/N\right)$.

Pour établir ce résultat, rappelons la propriété suivante qui sera utile dans beaucoup d'exercices utilisant les lois normales :

Propriété 2.2 : sommes de v.a. normales iid

Soient $X_1, X_2, \ldots, X_N$, N variables aléatoires iid, normales de loi $\mathcal{N}\left(0, \sigma^2\right)$. Alors la somme $U = X_1 + X_2 + \cdots + X_N$ est une variable aléatoire normale de loi $\mathcal{N}\left(0, \sigma^2/N\right)$.

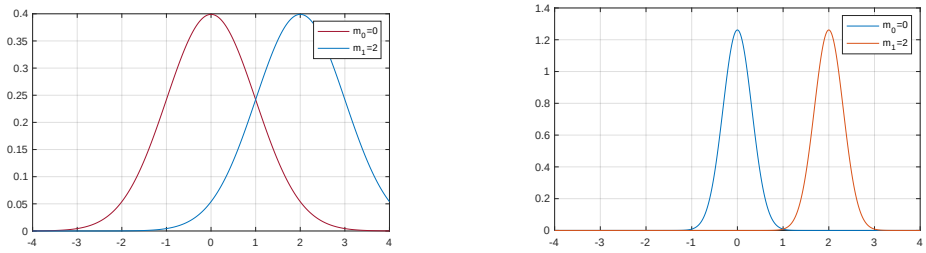
La variance décroit donc avec N, ce qui permet de garantir que la détection s'améliore en $1/N$ avec le nombre de mesures. Ces distributions sont illustrées à la figure 2.3.

Les exercices 2.14 et 2.15 mettent en oeuvre la même démarche sur d'autres exemples, avec d'autres lois.

Ce développement est un des éléments clés du savoir-faire présenté dans ce chapitre.

Exercice 2.5 : Log-vraisemblance

Dans l'exercice traité ci-dessus, analysez l'intérêt du test de log-vraisemblance par rapport au test de vraisemblance.



(a) N=1 (b) N=10

FIGURE 2.3 – Distributions des observations conditionnées de la variable S pour 1 seule observation (a) et pour 10 observations (b), avec $m_0 = 0$, $m_1 = 2$ et $\sigma^2 = 1$.

2.3.3 Maximum de vraisemblance

En télécommunications, lorsqu'on utilise ces techniques pour la détection des bits transmis, les coûts C_{fp} et C_{nd} sont identiques. Avec $C_{fp} = C_{nd} = 1$, la fonction de coût global mesure le BER. Enfin, si la source est équiprobable ($P_0 = P_1$), alors $\eta = 1$. Le rapport de vraisemblance doit simplement être comparé à 1 :

$$\Lambda(y) \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} 1.$$

et le test sur la log-vraisemblance à

$$\log\left(\Lambda(y)\right) \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} 0.$$

Cette règle de décision s'appelle le détecteur du ML.

Définition 2.6 : Détecteur ML

Le détecteur du ML est défini par :

$$\hat{e}_{ML}(y) = \arg \max_{e_i; i \in \{0,1\}} f_{Y|E}\left(y|e_i\right).$$

Ce détecteur est optimal au sens du risque de Bayes, seulement si les deux hypothèses sont équiprobables si les coûts sont symétriques. Cette remarque est également généralisable au cas M-aire.

Exercice 2.6 : Détecteur ML

Reprenez l'exercice 2.5 et déterminez les régions de décision correspondant au détecteur du maximum de vraisemblance.

Réponse : sur la figure 2.3, le seuil de détection du ML est localisé en $S = 1$ qui correspond au croisement des courbes. En effet le test ML sélectionne simplement la vraisemblance la plus grande. L'axe des abscisses est partagé en deux : si $S < 1$, la décision ML est $\mathbb{H}_0$ et sinon $\mathbb{H}_1$. Pour $S = 1$, les deux hypothèses ont une probabilité de $0,5$.

Enfin, pour calculer les probabilités d'erreur de ND et de FA, il faut intégrer sur les queues des gaussiennees à partir de $S = 1$.

Voici un exercice d'application en télécommunications.

Exercice 2.7 : Signal aléatoire continu dans du bruit

Un récepteur radio observe un signal aléatoire échantillonné $Y[k]$. Il veut tester la présence ou l'absence d'un signal radio, à partir de N mesures indépendantes. Le test repose sur 2 hypothèses :

— $\mathbb{H}_0$: pas de signal présent, $Y[k] = N[k]$. Le signal mesuré est alors simplement un bruit gaussien additif de loi $\mathcal{N}\left(0, \sigma_n^2\right)$.

— $\mathbb{H}_1$: un signal est présent $Y[k] = X[k] + N[k]$. Comme on ne connait pas la source, on modélise la source par une autre loi normale $\mathcal{N}\left(0, \sigma_x^2\right)$, où les échantillons $X[k]$ sont iid.

1. Quelle est la différence principale avec l'exercice traité ci-dessus ?
2. Déterminez les log-vraisemblances associées à chacune des hypothèses.
3. Déterminez le test du log-vraisemblance et la valeur de décision.
4. Explicitez le calcul des probabilités d'erreur de FA et de ND ?

2.3.4 Détecteur aux moindres carrés

Un détecteur alternatif parfois utilisé consiste à minimiser l'erreur de prédiction. On l'appelle erreur de prédiction minimale au sens des moindres carrés –minimum mean square prediction error– (MMSPE).

Il est construit à partir des évènements possibles. Pour chaque évènement e_i on construire une prédiction $\tilde{y}_i$ qui est l'observation la plus vraisemblable.

Par exemple, dans le cas du canal AWGN, $Y = X + N$, la valeur la plus probable du bruit est $n = 0$. La prédiction est donc $\tilde{y}_i = x_i$. Dans le cas binaire, on construit donc deux prédictions $\tilde{y}_0$ ou $\tilde{y}_1$.

Le détecteur aux moindres carrés cherche à minimiser la somme du carré des erreurs de détection (quand on a plusieurs observations), en choisissant entre $\tilde{\boldsymbol{y}}_\mathbf{0}$ ou $\tilde{\boldsymbol{y}}_\mathbf{1}$ celui qui minimise la distance avec $\boldsymbol{y}$.

Définition 2.7 : Détecteur MMSPE

Le détecteur MMSPE choisi l'hypothèse qui minimise l'erreur de prédiction :

$$\hat{e}_{MMSPE}(\boldsymbol{y}) = \arg \min_{e_i; i \in \{0,1\}} ||\boldsymbol{y} - \tilde{\boldsymbol{y}}_\boldsymbol{i}||^2 .$$

Ce détecteur est en général assez facile à calculer, surtout lorsque l'observation est

une fonction linéaire de l'entrée. Malheureusement, ce détecteur peut s'avérer très sous-optimal, car minimiser l'erreur de prédiction moyenne ($\mathbb{E}[y-\tilde{y}]$) n'est pas équivalent, dans le cas général, à minimiser le risque de Bayes. Rappelez-vous que le risque de Bayes se calcule sur les valeurs d'entrées non observables, alors que l'erreur de prédiction se calcule sur les observations.

Exercice 2.8 : Comparaison MMSPE et ML

1. Vous montrerez que ces deux détecteurs sont équivalents dans le cas d'un signal binaire (par exemple à valeurs -1 ou 1), dans un canal AWGN.
2. Trouvez un autre exemple dans lequel les deux détecteurs ne sont pas équivalents.

2.3.5 Détecteur MAP

Le dernier détecteur que nous présentons est appelé détecteur du maximum a posteriori (MAP). Pour le relier au détecteur de Bayes, conservons l'hypothèse que les coûts de faux positif et de non détection sont identiques. Par contre, on suppose connues les probabilités marginales des évènements, $P_0 = P_E(e_0)$ et $P_1 = P_E(e_1)$, qui ne sont pas équiprobables.

Dans ce cas le test du rapport de vraisemblance Th.2.1 devient :

$$\Lambda(y) \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \frac{P_0}{P_1}.$$

En passant les probabilités P_0 et P_1 à gauche du test, on obtient :

$$\Lambda(y) \cdot \frac{P_1}{P_0} \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} 1.$$

Effectuer ce test revient donc à choisir l'hypothèse qui donne le minimum entre $f_{Y|E}(y|e_1) \cdot P_E(e_1)$ et $f_{Y|E}(y|e_0) \cdot P_E(e_0)$. ce qui conduit, par l'égalité de Bayes, à la définition du MAP.

Définition 2.8 : Détecteur MAP

Le détecteur MAP choisi l'hypothèse qui maximise la probabilité a posteriori :

$$\hat{e}_{MAP}(y) = \arg \max_{e_i; i \in \{0,1\}} P_{E|Y}(e_i|y).$$

D'après Bayes on peut écrire en effet : $P_{E|Y}(e_i|y) = f_{Y|E}(y|e_i) \cdot P_E(e_i) / f_Y(y)$. comme $f_Y(y)$ est la densité d'observation qui est une constante pour y donné, la maximisation de $P_{E|Y}(e_i|y)$ est bien équivalent à maximiser $f_{Y|E}(y|e_i) \cdot P_E(e_i)$.

Bien entendu, si $P_E(e_1) = P_E(e_2)$, alors le détecteur MAP est équivalent au détecteur ML.

On retiendra, notamment pour l'extension $M-aire$, les relations suivantes entre les log-probabilités :

$$\log \underbrace{P_{E|Y}(e_i|y)}_{aposteriori} \propto \log \underbrace{f_{Y|E}(y|e_i)}_{vraisemblance} + \log \underbrace{P_E(e)}_{apriori}.$$

Cette expression met en évidence la différence entre l'estimateur ML et l'estimateur MAP. Ce dernier permet d'ajouter dans le modèle une connaissance a priori sur les évènements à détecter. Il s'agit d'une connaissance sur la loi de probabilité $\mathbb{P}$ associée à l'espace probabilisé des évènements $(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{T}}, \mathbb{P})$.

Pour résumer, nous pouvons dire que

— le détecteur ML est optimal si les deux types d'erreurs (ND et FA) ont le même coût et si les évènements sont équiprobables.

— Le détecteur MAP permet d'introduire un a priori sous forme de probabilité marginale (a priori) des évènements. Il est plus performant que le ML si l'a priori introduit est juste.

— Le détecteur optimal de Bayes est nécessaire si l'on a des coûts d'erreur asymétriques et permet, par rapport au détecteur du MAP d'adapter les seuils de décision.

2.3.6 Statistique suffisante

La notion de statistique suffisante est importante pour l'étude de problèmes complexes à grandes dimensions. Prenons un vecteur d'observation $\boldsymbol{Y}$de dimension N. La recherche d'une statistique suffisante consiste à rechercher un espace de représentation des données de dimension $D < N$. Cet espace caractérise alors une statistique suffisante du problème si la résolution du problème de détection est aussi efficace dans ce sous-espace que dans l'espace original.

Propriété 2.3 : Statistique suffisante

Soit un problème de détection défini par un espace d'évènements $\boldsymbol{\mathcal{E}}$, et une observation $\boldsymbol{Y} \in \boldsymbol{\mathcal{Y}}$. Soit $\boldsymbol{S} \in \boldsymbol{\mathcal{S}}$, un vecteur aléatoire construit à partir de $\boldsymbol{Y}$: $\boldsymbol{S} = g(\boldsymbol{Y})$ et tel que la dimension de $\boldsymbol{\mathcal{S}}$ est inférieure à celle de $\boldsymbol{\mathcal{Y}}$.

On dit que $\boldsymbol{S}$ est une statistique suffisante de $\boldsymbol{Y}$ pour le la détection de E, si la connaissance de $\boldsymbol{S}$ est strictement suffisante pour déterminer la solution optimale du problème de détection.

Prenons l'exemple d'un réseau de capteurs, qui remonte des informations vers une entité centrale qui doit prendre une décision (par exemple déclencer une alerte si la valeur moyenne remontée est supérieure à une certaine valeur). L'approche basique consiste à faire remonter toutes les mesures à l'entité centrale. Celle ci pouvant alors calculer la température moyenne. Mais est-il vraiment nécessaire de faire remonter toutes les informations ?

Supposons qu'il existe une transformation bijective permettant de décomposer $\boldsymbol{Y}$ en deux composantes :

$$\boldsymbol{Y} \xrightarrow{g} \begin{bmatrix} \boldsymbol{S} \\ \boldsymbol{A} \end{bmatrix}, \tag{2.5}$$

de telle sorte que la v.a. $\boldsymbol{A}|\boldsymbol{S}$ (lire A sachant S) soit indépendante de E. Autrement dit $\boldsymbol{A}$ n'apporte pas de connaissance supplémentaire sur E par rapport à $\boldsymbol{S}$.

Cela s'écrit :

$$f_{\boldsymbol{A}|\boldsymbol{S},E}\left(\boldsymbol{a}|\boldsymbol{s}, e_0\right) = f_{\boldsymbol{A}|\boldsymbol{S},E}\left(\boldsymbol{a}|\boldsymbol{s}, e_1\right) = f_{\boldsymbol{A}|\boldsymbol{S}}\left(\boldsymbol{a}|\boldsymbol{s}\right). \tag{2.6}$$

L'application $g(\cdot)$ peut correspondre à un changement de base, par exemple une transformée de Fourier. On notera la fonction inverse g^{-1}.

Si cette application est linéaire, elle peut alors être représentée par sa matrice M_g, telle que

$$\begin{bmatrix} \boldsymbol{S} \\ \boldsymbol{A} \end{bmatrix} = \boldsymbol{M_g} \cdot \boldsymbol{Y}.$$

Montrons que si la décomposition (2.5) vérifie (2.6), alors la connaissance de $\boldsymbol{S}$ est suffisante pour effectuer correctement le test d'hypothèse.

Le test d'hypothèse étant basé sur l'étude du rapport de vraisemblance (Def.2.5), décomposons la vraisemblance (ici dans le cas d'un espace d'observation continu). En utilisant le théorème des densités composées (voir cours de PBS), on obtient :

$$f_{\boldsymbol{Y}|E}(\boldsymbol{y}|e_i) = J(g^{-1}) \cdot f_{\boldsymbol{A},\boldsymbol{S}|E}(\boldsymbol{a},\boldsymbol{s}|e_i),$$

où $J(g^{-1})$ est le Jacobien [5] de la fonction g^{-1}. C'est pour pouvoir écrire cette relation, qu'on a du introduire la variable $\boldsymbol{A}$ ci-dessus.

Enfin, la formule des probabilités conditionnelles conduit à :

$$f_{\boldsymbol{Y}|E}(\boldsymbol{y}|e_i) = J(g^{-1}) \cdot f_{\boldsymbol{A}|\boldsymbol{S},E}(\boldsymbol{a}|\boldsymbol{s},e_i) \cdot f_{\boldsymbol{S}|E}(\boldsymbol{s}|e_i)$$

En injectant ce résultat dans la formule du rapport de vraisemblance, on obtient :

$$\Lambda(\boldsymbol{y}) = \frac{J(g^{-1}) \cdot f_{\boldsymbol{A}|\boldsymbol{S},E}(\boldsymbol{a}|\boldsymbol{s},e_1) \cdot f_{\boldsymbol{S}|E}(\boldsymbol{s}|e_1)}{J(g^{-1}) \cdot f_{\boldsymbol{A}|\boldsymbol{S},E}(\boldsymbol{a}|\boldsymbol{s},e_0) \cdot f_{\boldsymbol{S}|E}(\boldsymbol{s}|e_0)}$$

Les jacobiens s'annulent mutuellement car ils sont indépendants de e_i. En injectant (2.6), on obtient alors : $\Lambda(\boldsymbol{y}) = \Lambda(\boldsymbol{s})$.

Le test de vraisemblance peut donc être conduit uniquement à partir de $\boldsymbol{S}$.

L'utilisation d'une statistique suffisante permet de réduire la dimensionalité du problème. Ceci est particulièrement utile lorsqu'on manipule des données massives.

Etudions la construction d'une statistique suffisante sur un problème relativement simple.

Exercice 2.9 : Construction d'une statistique suffisante

Dans l'exemple étudié à la section 2.3.2 qui résoud l'exercice 2.13, la valeur moyenne du vecteur observé notée S constitue une statistique suffisante. En effet, nous avons montré que le test de vraisemblance s'exprime uniquement à partir de la valeur de S.

Nous prenons ici un exercice un petit peu différent.

$$\begin{array}{rcl} \mathbb{H}_0 & : & Y[k] \sim \mathcal{N}\left(m_0, \sigma_0^2\right) \\ \mathbb{H}_1 & : & Y[k] \sim \mathcal{N}\left(m_1, \sigma_1^2\right) \end{array}$$

pour $1 \leq k \leq N$ avec $\sigma_1^2 > \sigma_0^2$ et $m_1 > m_0$.

1. Construire une statistique suffisante de $\boldsymbol{Y}$.

5. Le Jacobien d'une fonction multivariée est une matrice dont l'élément (i,j) contient la dérivée de la i^{eme} composante par rapport à la j^{ieme} variable

La seule différence avec l'exercice 2.13 est qu'ici la variance est également différente pour les deux hypothèses. Mais comme précédemment, la vraisemblance s'écrit à partir des propriétés des lois normales :

$$\begin{aligned} f_{\boldsymbol{Y}|E}\left(\boldsymbol{y}|e_i\right) &= \prod_{k=1}^{N} f_{Y|E}\left(y[k]|e_i\right) \\ &= \frac{1}{(2\pi\sigma_i^2)^{N/2}} \cdot \exp\left(\frac{-\sum_{k=1}^{N}(y[k]-m_i)^2}{2\sigma_i^2}\right) \end{aligned}$$

La première égalité découle de l'indépendance entre les échantillons. La deuxième égalité dérive de la définition de la loi normale.

En développant l'exposant de l'exponentielle, on peut écrire :

$$f_{\boldsymbol{Y}|E}\left(\boldsymbol{y}|e_i\right) = b_i \cdot \exp\left(-\boldsymbol{\theta_i^t} \cdot \boldsymbol{s}\right).$$

avec comme paramètres :

$$\begin{aligned} \boldsymbol{\theta_i} &= \begin{bmatrix} N \cdot m_i/\sigma_i^2 & -N/2\sigma_i^2 \end{bmatrix}^t \\ b_i &= \frac{1}{(2\pi\sigma_i^2)^{N/2}} \cdot \exp\left(-\frac{N m_i^2}{2\sigma_i^2}\right) \end{aligned}$$

Et la statistique suffisante :

$$\boldsymbol{s} = g(\boldsymbol{y}) = \frac{1}{N}\begin{bmatrix} \sum_{k=1}^{N} y[k] & \sum_{k=1}^{N} y[k]^2 \end{bmatrix}^t$$

Le rapport de vraisemblance est alors :

$$\Lambda(\boldsymbol{y}) = \Lambda(\boldsymbol{s}) = \frac{b_1}{b_0} \cdot \exp\left((\boldsymbol{\theta_0} - \boldsymbol{\theta_1})^t \cdot \boldsymbol{s}\right)$$

Définition 2.9 : Statistique suffisante minimale

Une statistique suffisante est dite minimale si elle a la plus petite dimension parmi toutes les statistiques suffisantes.

La statistique suffisante (de dimension 2) obtenue à l'exercice précédent est-elle minimale ? Et bien non. En fait, pour un problème binaire tel que celui étudié ici, la vraisemblance $\Lambda(y)$ est en elle-même une statistique suffisante ! Et c'est une variable scalaire d'ordre 1.

Propriété 2.4 : Propriété du rapport de vraisemblance binaire

Pour tout problème de test d'hypothèse binaire, le rapport de vraisemblance constitue une statistique suffisante minimale de dimension 1.

Attention, ce n'est pas le cas pour un problème de décision M-aire. En effet, Si $|\boldsymbol{\mathcal{E}}| > 2$, il faut alors effectuer plusieurs tests, deux à deux, entre les hypothèses. L'étude des rapports de vraisemblance (entre hypothèses, 2 à 2) nécessite un espace multivarié qui constitue une statistique suffisante mais non nécessairement minimale.

2.4 Caractéristiques opérationnelles d'un récepteur

2.4.1 courbe ROC

Dans les section précédentes, nous nous sommes intéressés à la détection bayésienne, conduisant au test de vraisemblance et aux détecteurs MAP ou ML. L'inconvénient de l'approche bayésienne est qu'elle nécessite de connaitre les probabilités a priori des évènements et les coûts associés aux erreurs. Dans le cas d'une transmission numérique binaire (modèle du canal binaire par exemple), l'hypothèse que les erreurs de ND ou de FA sont équivalentes est légitime. De plus les embrouilleurs utilisés du côté des encodeurs garantissent en général l'équiprobabilité des symboles, ce qui justifie l'utilisation du ML.

Mais dans de nombreux cas applicatifs (notamment en radar, en localisation, en détection), les erreurs de ND ou de FA n'ont pas le même impact.

En partant de (2.2) et en y intégrant la définition des probabilités de fausse alarme $p_{fa} := P_{\hat{E}|E}(e_1|e_0)$ et de détection [6] $p_d := P_{\hat{E}|E}(e_1|e_1)$. L'étude des caractéristiques opérationnelles (avec les courbes ROC) consiste à étudier l'ensemble des paires (p_d, p_{fa}) simultanément atteignables.

Reprenons l'exemple du détecteur de fumée.

Si le détecteur retourne systématiquement une détection de fumée, alors :

$$P_{\hat{E}|E}(e_1|e_0) = P_{\hat{E}|E}(e_1|e_1) = P_{\hat{E}}(e_1) = 1,$$

conduisant à $p_d = 1, p_{fa} = 1$. Au contraire, si le détecteur retourne systématiquement une absence de fumée, alors :

$$P_{\hat{E}|E}(e_1|e_0) = P_{\hat{E}|E}(e_1|e_1) = P_{\hat{E}}(e_1) = 0.$$

Ce qui donne cette fois $p_d = 0, p_{fa} = 0$.

Aucun de ces deux tests n'est satisfaisant !

L'espace des tests possibles, en termes de performance, est représenté par un carré $(0 \leq p_d \leq 1) \times (0 \leq p_{fa} \leq 1)$ (voir figure 2.4). Les deux détecteurs naifs défini ci-dessus se retrouvent aux extrémités $(1, 1)$ et $(0, 0)$.

Le point opérationnel idéal a pour coordonnées $(1, 0)$, en haut à gauche.

Plus généralement, pour un problème de détection, l'ensemble de points réalisables est délimité par 2 courbes, telles que représentées à la figure 2.4. La courbe supérieure de la zone des tests réalisables contient l'ensemble des points optimaux au sens de Pareto [7]. Le choix d'un détecteur, sur cette courbe dépend de l'importance relative des erreurs de fausse alarme et de non détection. On appelle cette courbe, la courbe ROC. Plus elle s'approche du point optimal, plus le test est performant.

La courbe inférieure est en réalité la symétrique de la borne supérieure par rapport au point $(0,5; 0,5)$. Ceci se démontre de la façon suivante. Soit un test réalisable de probabilités $\mathfrak{R}_1 : (p_d, p_{fa})$. Alors on peut définir un test qui prend systématique la décision

6. Nous avons défini préalablement la probabilité de non détection p_{nd}. La probabilité de détection est son complément : $p_d = 1 - p_{nd}$.

7. L'optimalité au sens de Pareto est définie pour des problèmes multi-objectifs. Une solution est dite Pareto optimale si il n'existe pas de solution qui permette d'améliorer le résultat relativement à un des objectifs sans dégrader le résultat sur les autres objectifs.

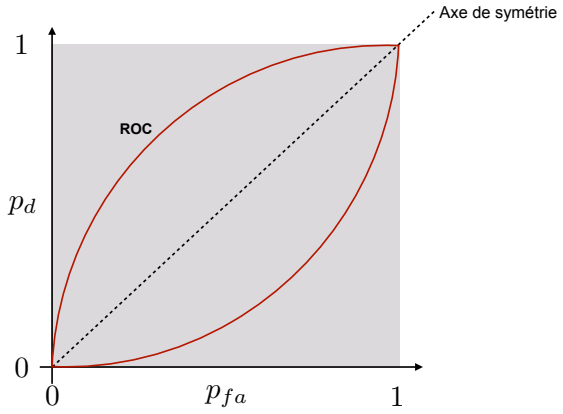


FIGURE 2.4 – Domaine de définition et propriétés des courbes ROC.

contraire : là où le premier test décide qu'il y a une fumée, le deuxième teste décide qu'il n'y en a pas, et vice-versa. Ce test aura les performances suivantes $\mathfrak{R}_2 : (1 - p_d, 1 - p_{fa})$.

Ainsi, à partir de tout test dont les performances sont situées sous la diagonale, il est possible de construire un test situé au-dessus de cette diagonale. La zone inférieure à la diagonale n'a donc pas d'intérêt, er les pires tests sont ceux situés sur la diagonale. Si le lieu des points réalisables est réduit à la diagonale, cela indique qu'il n'est pas possible de faire mieux que de prendre une décision complètement arbitraire.

2.4.2 Test de Neyman-Pearson

Pour éviter d'avoir à définir des coûts associés aux différents évènements, comme dans l'approche bayésienne, Jerzy Neyman et Egon Sharpe Pearson ont proposé en 1931 [6], le raisonnement suivant, plus facile à exploiter d'un point de vue opérationnel.

Fixons une probabilité de fausse alarme cible α, qui ne doit pas être dépassée, i.e. $p_{fa} \leq \alpha$. Sa valeur peut être librement choisie par l'expérimentateur.

A toute règle $\mathfrak{R}$, correspondent deux probabilités $p_d(\mathfrak{R})$ et $p_{fa}(\mathfrak{R})$. On note

$$\boldsymbol{\mathcal{D}}_\alpha = \{\mathfrak{R}; p_{fa}(\mathfrak{R}) \leq \alpha\},$$

l'ensemble des règles qui vérifient la condition sur la probabilité de fausse alarme.

L'objectif du test de Neyman-Pearson est de trouver une règle $\mathfrak{R}$, garantissant la contrainte de faussse alarme, tout en cherchant à maximiser la probabilité de détection p_d.

Cette formulation priorise le taux de fausse alarme. On peut aussi dans certains cas, poser le problème dans l'autre sens, i.e. fixer p_d et chercher à minimiser p_{fa}.

Définition 2.10 : Test de Neyman-Pearson

Le test de Neyman-Pearson est défini par la règle suivante

$$\mathfrak{R}_{NP} := \arg \max_{\mathfrak{R} \in \mathcal{D}_\alpha} p_d(\mathfrak{R}).$$

où α exprime la contrainte sur la probabilité de fausse alarme, et $\mathcal{D}_\alpha$ l'ensemble des

règles qui vérifient cette contrainte.

Cette définition se traduit littéralement par : *la règle associée au test de Neyman-Pearson est celle qui maximise la probabilité de détection sous contrainte que la probabilité de fausse alarme soit inférieure ou égale à α.*

Pour trouver ce test, il faut envisager tous les tests possibles, puis sélectionner celui qui maximise la probabilité de détection, parmi ceux qui vérifient le taux de fausse alarme.

Dans l'approche bayésienne, nous avons démontré qu'un choix déterministe était suffisant pour trouver une règle de décision optimale au sens bayésien, c'est à dire qu'à tout y correspond une décision unique.

Mais pour développe l'approche de Neymann-Pearson, il faut considérer tous les tests possibles, y compris des tests dont la décision est randomisée, c'est à dire que pour chaque observation y, la règle fait correspondre une décision probabiliste : la sortie du test est la probabilité de sélectionner l'évènement e_1 lorsque y est observé :

Définition 2.11 : Test randomisé

Un test binaire randomisé est défini par une règle qui a toute observation y associe une détection positive avec une certaine probabilité.

$$\begin{array}{rccl} \mathfrak{R}: & \boldsymbol{\mathcal{Y}} & \rightarrow & [0;1] \\ & y & & \tilde{\delta}(y) = \mathbb{P}(\hat{e}=1|y) \end{array} .$$

Avant de donner les résultats obtenus par Neymann et Pearson, définissons un sous-ensemble de tests, dits tests de vraisemblance randomisés :

Définition 2.12 : Test de vraisemblance randomisé

$$\tilde{\delta}(y) \begin{cases} = 1 & si \quad \Lambda(y) > \eta \\ = q & si \quad \Lambda(y) = \eta \\ = 0 & si \quad \Lambda(y) < \eta \end{cases} ,$$

où $q \in [0,1]$ et $\eta \in [0;\infty)$ sont des paramètres du test.

Par rapport au test de vraisemblance déterministe, la différence est localisée à l'interface entre les deu régions de décision $\boldsymbol{\mathcal{D}}_0$ et $\boldsymbol{\mathcal{D}}_1$. A cette interface, e_1 est choisi avec la probabilité q, alors que dans le test de vraisemblance déterministe il aurait fallu choisir lorsque $\Lambda(y) = \eta$, l'une ou l'autre des décisions.

Notons qu'ici la probabilité q est identique pour tous les y pour lesquels le rapport de vraisemblance est égal à η.

On a alors tous les éléments pour exprimer le théorème de Neyman-Pearson.

Théorème 2.3 : Neyman Pearson

Soit un test de vraisemblance randomisé tel que défini dans Def.2.12, de règle de décision $\tilde{\delta}(y)$ et qui vérifie $p_{fa}(\tilde{\delta}) = \alpha$.

1. Optimalité : ce test est optimal au sens de Neymann-Pearson, c'est à dire que

tout autre test (de tout type possible, du moment qu'il n'utilise pas d'information a priori complémentaire) de règle $\tilde{\delta}'(y)$ et qui vérifie aussi $p_{fa}(\tilde{\delta}') \leq \alpha$, conduit à une probabilité de détection inférieure ou égale : $p_d(\tilde{\delta}') \leq p_d(\tilde{\delta})$.

2. Existence : Pour tout $\alpha \in [0,1]$ il existe un test de vraisemblance randomisé, de règle de décision $\tilde{\delta}_{NP}$ pour laquelle $p_{fa}(\tilde{\delta}) = \alpha$.
3. Unicité : Soit une règle $\tilde{\delta}''$ qui est un test optimal au sens de Neyman-Pearson de paramètre α, alors $\tilde{\delta}''$ est un test de vraisemblance randomisé, tel que défini en Def.2.12.

La preuve de ce théroème nécessite de manipuler la résolution Lagrangienne, que nous n'avons pas détaillée dans ce cours. Le lecteur intéressé pourra trouver la preuve dans [9]. TODO : Il serait intéressant cependant d'étudier ce type d'algorithme, utile pour beaucoup de problèmes en CNA.

Nous retiendrons surtout l'interprétation de ce théorème, qui dit littéralement :

1. Tout test de vraisemblance randomisé est optimal au sens de Neyman-Pearson (situé sur la courbe ROC).
2. Pour toute contrainte α, il existe un test de vraisemblance qui permet de la vérifier.
3. Tout test qui est optimal au sens de Neyman-Pearson est un test de vraisemblance randomisé.

Ainsi en choisissant les paramètres η et q d'un test de vraisemblance, l'ensemble des solutions Pareto-optimale du problème de détection binaire peut être parcouru.

Enfin, comme nous le verrons en exercice, la randomisation n'est nécessaire que dans le cas où l'espace d'observation est discret ou disjoint (comme dans l'exercice2.16). Lorsque le domaine d'observation $\mathcal{Y}$ est continu, alors la probabilité que le rapport de vraisemblance soit exactement égal à η tend vers 0 et la randomisation n'est pas nécessaire. On a alors une stricte équivalence entre les solutions optimales au sens de Bayes et les solutions de Neymann-Pearson, qui se trouvent sur le front de Pareto.

Notons que le choix de du seuil η conditionne le poids relatif des deux types d'erreur. Pour tenir compte de l'état intermédiaire probabiliste, on peut écrire les probabilités de détection et de fausse alarme :

$$\begin{aligned} p_d(\mathfrak{R}) &= \mathbb{E}_{Y|E=e_1}\left[\tilde{\delta}(y)\right] = \int_{\mathcal{Y}} \tilde{\delta}(y) \cdot f_{Y|E}\left(y|e_1\right) \cdot dy \\ p_{fa}(\mathfrak{R}) &= \mathbb{E}_{Y|E=e_0}\left[\tilde{\delta}(y)\right] = \int_{\mathcal{Y}} \tilde{\delta}(y) \cdot f_{Y|E}\left(y|e_0\right) \cdot dy \end{aligned} \tag{2.7}$$

2.4.3 Application à l'exercice 2.13, partie B

Pour simplifier l'analyse, nous prenons comme indiqué dans l'énoncé $m_0 = 0$ et$m_1 = \rho$. Pour rappel, nous avons démontré que le test de vraisemblance optimal pour le risque de Bayes s'écrit :

$$S \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \gamma, \tag{2.8}$$

avec $\gamma = \frac{\rho}{2} + \frac{\sigma^2 \log(\eta)}{N\rho}$. Ce test fait intervenir S et non Y. En effet, nous avons ramené le problème à l'étude d'une statistique suffisante S. D'après le théorème de Neymann Pearson

et l'équivalence entre la formulation du test de vraisemblance et du test de Neymann-Pearson dans le cas où $\boldsymbol{\mathcal{Y}}$ est continu, la décision sur Y se ramène à une décision relative à S :

$$\tilde{\delta}(y) \equiv \tilde{\delta}(s = f(y)).$$

Ce qui est confortable, car dès lors les intégrales multiples peuvent être remplacées par des intégrales mono-variables, et (2.7) devient :

$$p_d(\mathfrak{R}) = \int_{s=-\infty}^{\infty} \tilde{\delta}'(s) \cdot f_{S|E}\left(s|e_1\right) \cdot ds$$
$$p_{fa}(\mathfrak{R}) = \int_{s=-\infty}^{\infty} \tilde{\delta}'(s) \cdot f_{S|E}\left(s|e_0\right) \cdot ds.$$

Nous venons de voir que tout test de Neyman-Pearson peut s'exprimer comme un test de vraisemblance, en faisant varier la valeur du seuil dans (2.8). C'est à dire que la fonction $\tilde{\delta}'(s)$ s'identifie à :

$$\tilde{\delta}'(s) \begin{cases} = 1 & si \quad s \geq \gamma \\ = 0 & sinon \end{cases}.$$

qui n'est autre que la fonction échelon unité $U\left(s - \gamma\right)$. Les probabilités de détection et de fausse alarme sont dans cet exemple données par :

$$p_d(\mathfrak{R}) = \int_{s=\gamma}^{\infty} f_{S|E}\left(s|e_1\right) \cdot ds$$
$$p_{fa}(\mathfrak{R}) = \int_{s=\gamma}^{\infty} f_{S|E}\left(s|e_0\right) \cdot ds,$$

La variable S est dans cet exemple la moyenne de variables aléatoires iid de loi normale. Elle suit donc une loi normale d'après la propriété 2.2. $\mathcal{N}\left(0, \sigma^2/N\right)$. On obtient alors :

$$p_d(\mathfrak{R}) = \int_{s=\gamma}^{\infty} \frac{1}{\left(2\pi\sigma^2/N\right)} \cdot \exp\left(-\frac{(s-\rho)^2}{2\sigma^2/N}\right) \cdot ds$$
$$p_{fa}(\mathfrak{R}) = \int_{s=\gamma}^{\infty} \frac{1}{\left(2\pi\sigma^2/N\right)} \cdot \exp\left(-\frac{s^2}{2\sigma^2/N}\right) \cdot ds,$$

Par changement de variable, on peut faire apparaitre la fonction $Q(x)$ (on utilise également la propriété $Q(-x) = 1 - Q(x)$. On obtient :

$$p_d(\mathfrak{R}) = Q\left(\frac{\sqrt{N}}{\sigma}(\gamma - \rho)\right)$$
$$p_{fa}(\mathfrak{R}) = Q\left(\frac{\sqrt{N}}{\sigma}\gamma\right)$$

Ces expressions caractérisent bien l'intégration de la queue des gaussiennes jusqu'à γ. Quand $\gamma = -\infty$, on obtient $p_d = p_{fa} = 1$, et pour $\gamma = \infty$, $p_d = p_{fa} = 0$. Ces deux points correspondent aux positions extrêmes de la courbe.

On peut introduire une distance normalisée entre les 2 positions, notée $d = \rho \cdot \frac{\sqrt{N}}{\sigma}$ et un paramètre normalisé de seuil $\tau = \frac{\sqrt{N}}{\sigma}\gamma$. On obtient alors :

$$p_d(\mathfrak{R}) = Q\left(\tau - d\right))$$
$$p_{fa}(\mathfrak{R}) = Q\left(\tau\right)$$

Enfin, on peut exprimer la probabilité de détection en fonction de la probabilité de fausse alarme, caractérisant ainsi les courbes ROC associées à ce problème :

$$p_d = 1 - Q\left(d - Q^{-1}(p_{fa})\right), \tag{2.9}$$

représentées à la figure 2.5.

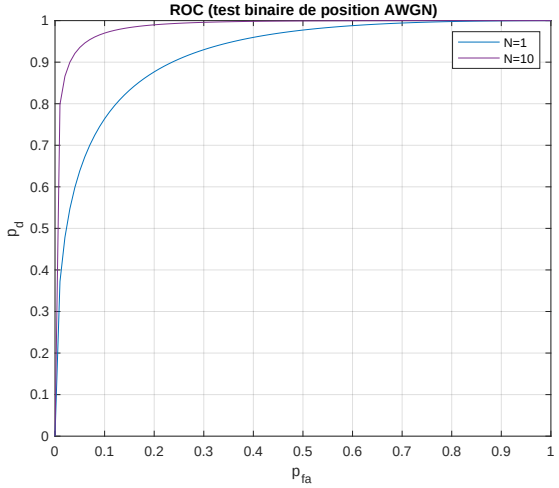


FIGURE 2.5 – Courbes ROC associées au problème du test binaire de position en canal AWGN, d'après (2.9).

On peut observer deux propriétés des courbes ROC, qui sont toujours vraies :

— La courbe $p_d = f(p_{fa})$ est concave.

— La courbe démarre et termine aux points $(0, 1)$ et $(1, 0)$.

D'autre part, le test ML obtenu pour $\eta = 1$, peut se retrouver à partir de la courbe ROC, au point où sa dérivée est égale à -1, car on cherche dans ce cas à maximiser $p_{fa} + 1 - p_d$.

Pour mettre en évidence l'apport de la randomisation, il faut étudier un cas où l'espace d'observation est discret. C'est le cas de la deuxième partie de l'exercice 2.16, qui permet de calculer la courbe ROC dans le cas du canal binaire. Si cette approche peut paraître un peu artificielle dans le cas du canal binaire car les erreurs de fausse alarme et de non détection ont le même poids. Cependant, dans d'autres situations, typiquement celle de notre détecteur de fumée, il peut être très intéressant de caractériser la courbe ROC.

Dans des problèmes complexes, où vous pouvez avoir à comparer plusieurs méthodes de détection, comparer les courbes ROC associées à chacune des méthodes est très intéressant. TODO : prévoir un exo à ce sujet

2.5 Détection M-aire

La généralisation du cas binaire à un cas de dimension supérieure (appelé M-aire) n'est pas difficile du point de vue conceptuel, mais conduit à des difficultés calculatoires.

Soit un évènement à valeurs dans $\boldsymbol{\mathcal{E}} = \{e_0, e_1, \ldots, e_{M-1}\}$. Pour développer un cadre baysésien, il faut définir les coûts associés à chaque couple $(E, \hat{E})$. Il faut donc définir N^2 coûts : C_{ij} et calculer le risque de Bayes global associé à chaque règle de décision.

En extrapolant le théorème 2.1, le risque de Bayes s'exprime maintenant par :

$$Q(\mathfrak{R}) = \sum_{i=0}^{M-1} \int_{\mathcal{Y}_i} C_i(y) f_Y(y)\, dy, \tag{2.10}$$

où

$$C_i(y) := \sum_{j=0}^{M-1} C_{ij} \cdot P_{E|Y}(e_j|y).$$

Clairement, la règle sera optimale au sens de Bayes, si pour tout y observé, on choisi l'évènement e_k tel quel $k = \arg\min_{i\in[0,M-1]} C_i(y)$.

Si les coûts sont donnés par $C_{ij} = 1 - \delta_{ij}$, alors minimiser $C_i(y)$ revient à sélectionner l'évènement dont la probabilité a posteriori est la plus élevée. On obtient donc le détecteur du MAP.

$$\hat{e}_{MAP} = \arg \max_{e_i; i\in\{0,1,\ldots,M-1\}} P_{E|Y}(e_i|y).$$

Enfin, si les évènements sont équiprobables, alors maximiser $P_{E|Y}(e_j|y)$ revient à maximiser la vraisemblance $f_{Y|E}(y|e_i)$ et on retrouve le détecteur ML.

Définition 2.13 : Détecteurs M-aire

Dans un problème M-aire à coût uniforme on défini les détecteurs suivants :

- ML : $\hat{e}_{ML} = \arg\max_{e_i; i\in\{0,1,\ldots,M-1\}} f_{Y|E}(y|e_i)$.
- MAP : $\hat{e}_{MAP} = \arg\max_{e_i; i\in\{0,1,\ldots,M-1\}} P_{E|Y}(e_j|y)$.

Ces détecteurs sont utiles par exemple pour étudier les propriétés des modulations d'ordre élevé, par exemple une quadrature amplitude modulation (QAM) à 16 états, où chaque symbole représente une des hypothèses. Egalement, lorsque les transmissions sont codées, le décodage doit se faire sur le mot code complet. Considérons un mot de N_s symboles M-aire, transmis conjointement dans un paquet, dans un canal perturbé qui engendre des échos. Le récepteur ML correspondant doit chercher parmi tous les mots codes celui qui a la vraisemblance la plus forte. Mais il y a M^{N_s} mots codes possible. Dès lors implémenter un ML, bien que théoriquement optimal, s'avère impossible. C'est pour obtenir des décodeurs/démodulateurs moins complexes que les techniques de modulations et codages numériques continuent de se développer (turbo-codes, test de parité à faible densité –low density parity check– (LDPC),...).

2.6 Estimation

La théorie de l'estimation s'intéresse à un problème similaire à celui de la détection, mais conduit malgré tout à la mise en oeuvre de méthodes assez différentes. L'unique différence dans la formulation du problème est que l'ensemble d'évènements discrets $\boldsymbol{\mathcal{E}}$, est remplacé par un espace de paramètres à estimer qui prend ses valeurs dans un espace continu $\boldsymbol{\mathcal{A}}$. On note a l'élément de $\boldsymbol{\mathcal{A}}$ à estimer.

Si l'on reprend l'exemple du capteur de fumée, il s'agit maintenant d'estimer le taux de CO2 et non seulement la présence de fumée. Le capteur mesure ce niveau d'émission et la converti en tension, par exemple entre 0 et 5 volts. Imaginons que la tension (notée y) est une fonction du taux de particules (noté x), bruitée par un bruit additif noté n. On peut écrire :

$$y = g(x) + n \tag{2.11}$$

Mettre en place une technique d'estimation a deux objectifs :

— Trouver quelle est la valeur de x la plus probable.

— Définir la fiabilité de cette mesure, en estimant par exemple un intervalle de confiance sur cette estimation.

Dans de nombreux problèmes d'estimation, l'observation est constituée de plusieurs observations, c'est à dire que y est un vecteur. Une question très importante est d'analyser le comportement de la prédiction en fonction du nombre de mesures.

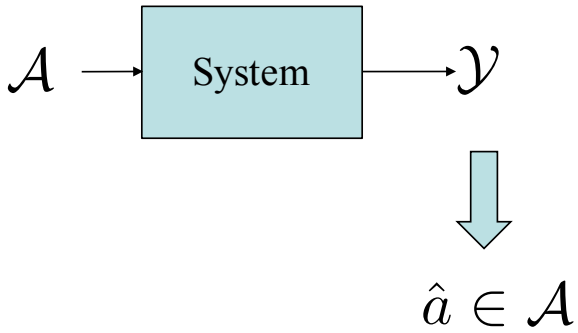


FIGURE 2.6 – Représentation du problème d'estimation optimale.

Définition 2.14 : Problème d'estimation

Un problème d'estimation est défini par les éléments suivants :

1. Un espace de paramètres noté $\boldsymbol{\mathcal{A}}$, qui représente l'ensemble des valeurs que peut prendre le (ou les) paramètre(s) que l'on cherche à estimer.
2. Un espace d'observation de dimension finie, et noté $\boldsymbol{\mathcal{Y}}$. Une observation y est un élément de cet espace.
3. Une loi de probabilités conditionnelles, connue, qui relie les observations aux paramètres recherchés : $f_{Y|A}(y|a)$ ou $P_{Y|A}(y_i|a)$ si l'espace d'observation est discret.
4. Une règle d'estimation qui prédit a sous la forme d'un estimateur $\hat{a}(y)$.

Ce type de problème trouve de nombreuses applications en Télécoms : estimation de la fréquence d'une porteuse, du niveau de bruit d'un canal, du niveau de puissance utile, des coefficients du canal, etc...

Définir un bon estimateur poursuit plusieurs objectifs :

— Minimiser l'erreur (justesse et précision).

— Minimiser le nombre d'observations nécessaires

— Limiter la complexité mathématique ou algorithmique de cet estimateur.

Nous allons nous appuyer sur ce que nous avons vu en théorie de la détection, mais la notion de test d'hypothèse ne s'applique plus directement, car nous travaillons sur un espace continu. Nous cherchons la valeur de a la plus probable dans un espace continu $\boldsymbol{\mathcal{A}}$. Ceci constitue la différence fondamentale entre estimation et détection.

2.6.1 Minimisation d'une fonction de coût

Nous commençons par définir l'erreur d'estimation comme la différence entre la valeur du paramètre et la valeur estimée :

$$\epsilon(y|a) := \tilde{a}(y) - a.$$

La fonction de coût se défini alors en fonction de l'erreur : $C(\epsilon)$. On peut imaginer une infinité de fonctions de coût. En voici 3 exemples très importants.

1. La fonction de coût quadratique : $C_{MSE}(\epsilon) := \epsilon^2$, appelée erreur moyenne au sens des moindres carrés –mean square error– (MSE).
2. La fonction de coût valeur absolue : $C_{MAE}(\epsilon) := |\epsilon|$, appelée mean absolute error, erreur moyenne au sens de la valeur absolue (MAE).
3. La fonction de coût uniforme : $C_{\Delta}(\epsilon) := 1$ si $|\epsilon| > \Delta/2$, et 0 sinon.

Ces fonctions de coût proposent trois façons différentes d'évaluer le coût de l'erreur d'estimation à adapter en fonction du problème étudié.

Définition 2.15 : Rsque de Bayes pour l'estimation

Le critère qui est minimisé dans un problème d'estimation est le risque de Bayes, défini comme le coût moyen :

$$Q(\mathfrak{R}) := \mathbb{E}_{A,Y}\left[C(\epsilon)\right],$$

pour une certaine fonction $C(\epsilon)$, positive, croissante, à valeurs dans $\mathbb{R}$.

Chaque fonction de coût conduit donc à un estimateur optimal différent. L'estimateur erreur minimale au sens des moindres carrés –minimum mean square error– (MMSE) minimise le coût quadratique moyen, l'estimateur minimum mean absolute error (MMAE) minimise le coût valeur absolu et l'estimateur MAP minimise le coût uniforme pour $\Delta \to 0$.

Théorème 2.4 : Estimateurs optimaux

Les estimateurs optimaux associés respectivement aux fonctions de coût quadratique, absolue et uniforme, se calculent à partir de la fonction de densité a posteriori :

— MMSE : $\hat{a}_{MMSE}(y) = \mathbb{E}_{A|Y}[a]$.
— MMAE : $\hat{a}_{MMAE}(y) = \text{med}\left(f_{A|Y}(a|y)\right)$.
— MAP : $\hat{a}_{MAP}(y) = \hat{a}_{\Delta\to 0}(y) = \arg\max_{a\in\boldsymbol{\mathcal{A}}}\left[f_{A|Y}(a|y)\right]$.

La relation entre la définition de ces estimateurs et ces expressions n'est pas triviale. Nous allons les démontrer. Pour cela, il faut partir du coût de Bayes, et intégrer par rapport à toutes les réalisations possibles sur $\boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{Y}}$:

$$\begin{aligned} Q(\mathfrak{R}) &= \mathbb{E}_{A,Y}\left[C(\epsilon(y,a))\right] \\ &= \int_{a\in\boldsymbol{\mathcal{A}}}\int_{y\in\boldsymbol{\mathcal{Y}}} C(\hat{a}(y)-a)\cdot f_{AY}(a,y)\cdot dy\cdot da \end{aligned}$$

En développant, avec la formule de Bayes, $f_{AY}(a,y) = f_Y(y)\cdot f_{A|Y}(a|y)$, et en supposant les intégrales uniformément convergentes, le théorème de Fubini permet d'intervertir les intégrales et d'écrire :

$$Q(\mathfrak{R}) = \int_{y\in\boldsymbol{\mathcal{Y}}} f_Y(y)\underbrace{\left[\int_{a\in\boldsymbol{\mathcal{A}}} C(\hat{a}(y)-a)\cdot f_{A|Y}(a|y)\cdot da\right]}_{I(y)}\cdot dy$$

où $I(y)$ représente l'expression à minimiser indépendamment pour chaque observation y, exactement comme nous l'avons fait pour les problèmes de détection.

Etudions les trois fonctions de coût proposées précédemment.

1. MMSE : la fonction de coût dépend de l'intégrale $I(y)$ notée ici :

$$I_{MSE}(y) = \int_{a\in\boldsymbol{\mathcal{A}}} (\hat{a}(y)-a)^2\cdot f_{A|Y}(a|y)\cdot da.$$

La minimisation de cette fonction peut se faire en cherchant pour quelle valeur de $\hat{a}$ la dérivée est nulle :

$$\begin{aligned} \frac{d}{d\hat{a}} I_{MSE}(y) &= \frac{d}{d\hat{a}}\int_{a\in\boldsymbol{\mathcal{A}}} (\hat{a}(y)-a)^2\cdot f_{A|Y}(a|y)\cdot da \\ &= 2\int_{a\in\boldsymbol{\mathcal{A}}} (\hat{a}(y)-a)\cdot f_{A|Y}(a|y)\cdot da \\ &= 2\hat{a}(y)\cdot\int_{a\in\boldsymbol{\mathcal{A}}} f_{A|Y}(a|y)\cdot da - 2\int_{a\in\boldsymbol{\mathcal{A}}} a f_{A|Y}(a|y)\cdot da \\ &= 2\hat{a}(y) - 2\int_{a\in\boldsymbol{\mathcal{A}}} a f_{A|Y}(a|y)\cdot da \end{aligned}$$

qui s'annule uniquement pour $\hat{a}_{MMSE}(y) = \int_{a\in\boldsymbol{\mathcal{A}}} a f_{A|Y}(a|y)\cdot da$, que l'on peut interpréter en disant que l'estimateur optimal au sens des moindres carrés est égal à la moyenne statistique a posteriori du paramètre, conditionnée par l'observation y. On la note $\hat{a}_{MMSE}(y) = \mathbb{E}_{A|Y}[a]$, c'est à dire la moyenne de la v.a. A, selon la loi $f_{A|Y}(.)$.
Une fois connu l'estimateur, on peut exprimer la valeur du risque de Bayes :

$$\begin{aligned} Q_{MMSE} &= \mathbb{E}_{AY}\left[(\hat{a}_{MMSE}(y)-a)^2\right] \\ &= \mathbb{E}_Y\left[\mathbb{E}_{A|Y}\left[(\hat{a}_{MMSE}(y)-a)^2\right]\right] \\ &= \mathbb{E}_Y\left[\mathbb{E}_{A|Y}\left[(\mathbb{E}_{A|Y}[a]-a)^2\right]\right] \\ &= \mathbb{E}_Y\left[\mathbb{E}_{A|Y}\left[a^2\right] - \mathbb{E}_{A|Y}[a]^2\right] \\ &= \mathbb{E}_Y\left[\mathbb{E}_{A|Y}\left[a^2\right] - \hat{a}_{MMSE}(y)^2\right] \end{aligned} \tag{2.12}$$

2. MMAE : la fonction de coût dépend de l'intégrale $I(y)$ notée ici :

$$I_{abs}(y) = \int_{a\in\boldsymbol{\mathcal{A}}} |\hat{a}(y) - a| \cdot f_{A|Y}(a|y) \cdot da.$$

Pour évaluer la fonction valeur absolue, il faut décomposer l'intégration sur deux domaines complémentaires. En supposant que $\boldsymbol{\mathcal{A}} = \mathbb{R}$, on peut écrire :

$$I_{abs}(y) = \int_{a=-\infty}^{\hat{a}(y)} (\hat{a}(y) - a) \cdot f_{A|Y}(a|y) \cdot da + \int_{a=\hat{a}(y)}^{+\infty} (a - \hat{a}(y)) \cdot f_{A|Y}(a|y) \cdot da.$$

La différentiation de cette fonction par rapport à $\hat{a}(y)$ conduit à :

$$\frac{d}{d\hat{a}} I_{abs}(y) = \hat{a}(y) \cdot f_{A|Y}(a|y) - \hat{a}(y) \cdot f_{A|Y}(a|y).$$

Il faut donc choisir $\hat{a}_{MMAE}(y)$ qui vérifie :

$$\int_{a=-\infty}^{\hat{a}_{MMAE}(y)} f_{A|Y}(a|y) \cdot da = \int_{a=\hat{a}_{MMAE}(y)}^{+\infty} f_{A|Y}(a|y) \cdot da,$$

qui n'est autre que la médiane de la densité a posteriori.

3. Coût uniforme sur $[-\Delta/2; \Delta/2]$. Il faut alors trouver $\hat{a}_\Delta$ qui maximise :

$$I_\Delta(y) = \int_{a=\hat{a}_\Delta(y)-\Delta/2}^{\hat{a}_\Delta(y)+\Delta/2} f_{A|Y}(a|y) \cdot da.$$

De fait, si l'on fait tendre $\Delta \to 0$, on obtient alors :

$$\hat{a}_{\Delta\to 0} = \arg\max_{a\in\boldsymbol{\mathcal{A}}} f_{A|Y}(a|y),$$

qui n'est autre que le MAP.

Nous avons donc trois estimateurs qui s'appuient sur les caractéristiques de la loi a posteriori : sa moyenne, sa médiane ou son maximum. Du fait que ces estimateurs soient basés sur la loi a posteriori, cela veut dire que leur mise en oeuvre nécessite de connaitre une loi a priori pour le paramètre concerné. On retrouve alors, comme pour la théorie de la détection, la notion de connaissance a priori.

Nous allons étudier un exemple d'application.

Exercice 2.10 : Estimation d'une loi exponentielle

On observe un scalaire réel, $Y \in \boldsymbol{\mathcal{Y}} = \mathbb{R}^+$, issu d'un processus aléatoire donné par une loi exponentielle de paramètre a :

$$f_{Y|A}(y|a) := \begin{cases} a \cdot e^{-ay} & si \quad y \geq 0 \\ 0 & si \quad y < 0 \end{cases}$$

La loi exponentielle est classiquement utilisée pour simuler des processus d'arrivée aléatoire. Elle est à la base de l'étude des files d'attente. Imaginons une cellule radio de type Wifi où les noeuds mobiles utilisent un protocole d'accès au medium de type protocole d'accès aléatoire au medium (ALOHA). La variable Y représente alors la distribution du temps entre arrivées de paquets, et le paramètre a caractérise la charge de la cellule. Plus a est petit, plus la cellule est chargée. La station de base peut donc estimer la charge de la cellule à partir de l'observation des intervalles entre l'arrivée des paquets.

On pose comme connaissance a priori, que le paramètre a suit lui-même une loi exponentielle :

$$f_A(a) := \begin{cases} \alpha \cdot e^{-\alpha a} & si \quad a \geq 0 \\ 0 & si \quad a < 0 \end{cases}$$

1. Calculez la loi de probabilité a posteriori correspondant au problème.
2. Calculez l'expression des estimateurs de a pour les 3 estimateurs définis dans le cours.
3. Evaluez le risque de Bayes associé au ML.

La station de base effectue cette estimation à partir de N échantillons.

4. Calculez la loi de probabilité a posteriori en fonction de N.
5. Calculez l'expression des estimateurs de a pour les 3 estimateurs en fonction de N.

La loi a posteriori est donnée par

$$f_{A|Y}(a|y) = \frac{f_{Y|A}(y|a) \cdot f_A(a)}{f_Y(y)}. \tag{2.13}$$

Les deux distributions du numérateur sont données dans l'énnoncé. La loi marginale de y, utilisée au dénominateur se calcule par la loi des probabilités totales :

$$f_Y(y) = \int_a f_{Y|A}(y|a) \cdot f_A(a) \cdot da,$$

ce qui donne :

$$\begin{aligned} f_{A|Y}(a|y) &= \frac{a\alpha e^{-(\alpha+y)a}}{\int_0^\infty a\alpha e^{-(\alpha+y)a} \cdot da} \\ &= (\alpha+y)^2 a e^{-a(\alpha+y)} \end{aligned}$$

L'estimateur MMSE est obtenu en calculant la moyenne, soit :

$$\begin{aligned} \hat{a}_{MMSE} &= \int_0^\infty a f_{A|Y}(a|y) \cdot da = (\alpha+y)^2 \int_0^\infty a^2 e^{-(\alpha+y)a} \cdot da \\ &= \frac{2}{\alpha+y}. \end{aligned} \tag{2.14}$$

En appliquant la formule (2.12), vous pourrez vérifier que le risque de Bayes associé est

$$Q_{MMSE} = \frac{2}{3\alpha^2}$$

.

L'estimateur MMAE $\hat{a}_{MMAE}(y)$ est obtenu au point médian de la distribution. Il est solution de :

$$\int_{\hat{a}_{MMAE}(y)}^\infty f_{A|Y}(a|y) \cdot da = 1/2.$$

En intégrant, on obtient :

$$\left[1 + (\alpha+y)\hat{a}_{MMAE}(y)\right] e^{-(\alpha+y)\hat{a}_{MMAE}(y)} = \frac{1}{2}.$$

ce qui donne $\hat{a}_{MMAE}(y) = \frac{T_0}{\alpha+y}$; avec $T_0 \approx 1,68$, solution de $[1+T_0]\, e^{-T_0} = \frac{1}{2}$.

Enfin, l'estimateur MAP est obtenu en cherchant le point pour lequel la dérivée de $f_{A|Y}(a|y)$ s'annule. On obtient

$$\hat{a}_{MAP}(y) = \frac{1}{\alpha + y}.$$

Notons que le calcul complet de la densité a posterori n'est nécessaire que pour les estimateurs MMSE et MMAE. L'estimateur du MAP n'ayant besoin que de calculer le max de cette densité, seul le numérateur de cette fraction est nécessaire puisque le dénominateur ne dépend pas de a.

On a donc 3 estimateurs qui retournent trois valeurs différentes. Chacun est optimal par rapport à une fonction de coût initiale.

Passons au cas où l'on observe N échantillons $Y[k]$ iid. On a alors

$$\begin{aligned} f_{\boldsymbol{Y}|A}(\boldsymbol{y}|a) &= \prod_{k=1}^{N} f_{Y|A}(y[k]|a) \\ &= a^N \cdot e^{-N \cdot a\bar{y}}, \end{aligned}$$

où $\bar{y} := \frac{1}{N}\sum_k y[k]$.

La loi a posteriori est alors donnée par :

$$\begin{aligned} f_{A|\boldsymbol{Y}}(a|\boldsymbol{y}) &= \frac{a^N \alpha e^{-(\alpha+N\bar{y})a}}{\int_0^\infty a^N \alpha e^{-(\alpha+N\bar{y})a} \cdot da} \\ &= \frac{(\alpha+N\bar{y})^{N+1}}{N!} a^N e^{-a(\alpha+N\bar{y})} \end{aligned} \tag{2.15}$$

Dont on peut tirer les 3 estimateurs. Vous terminerez l'exercice en exprimant les 3 estimateurs à partir de (2.15).

2.6.2 Estimateur ML

Nous avons défini ci-dessus l'estimateur du MAP pour l'estimation d'une variable aléatoire, en faisant tendre $\Delta \to 0$ à partir de la fonction de coût uniforme.

Revenons sur la construction de cet estimateur. Rechercher le maximum de $f_{A|Y}(a|y)$ consiste à trouver le point pour lequel sa dérivée s'annule $\frac{\partial f_{A|Y}(a|y)}{\partial a} = 0$.

Parce que la fonction log est strictement monotone et croissante, cela est également équivalent à chercher

$$\frac{\partial \log(f_{A|Y}(a|y))}{\partial a} = 0.$$

En injectant le théorème de Bayes dans la fonction log, on obtient :

$$\log(f_{A|Y}(a|y)) = \log(f_{Y|A}(y|a)) + \log(f_A(a)) - \log(f_Y(y)). \tag{2.16}$$

Le dernier terme est indépendant de a, et l'estimateur MAP est donc donné par :

$$\hat{a}_{map}(y) = \arg\max_{a \in \boldsymbol{\mathcal{A}}} \left[\log(f_{Y|A}(y|a)) + \log(f_A(a))\right]. \tag{2.17}$$

Le premier terme est la log-vraisemblance qui dépend du modèle du système et qui décrit l'observation en fonction du paramètre à estimer. Le deuxième terme représente la connaissance a priori sur la loi du paramètre recherché .

Si nous n'avons aucune connaissance a priori sur ce paramètre, on peut seulement considérer que la loi sur $\boldsymbol{\mathcal{A}}$ est uniforme (i.e. $f_A(a) = cste$). Le problème se résume à la maximisation de la log-vraisemblance.

On obtient ainsi un nouvel estimateur qui ne nécessite pas de formuler une connaissance a priori sur le paramètre recherché :

Théorème 2.5 : Estimateur ML

L'estimateur ML associé un problème d'estimation est donné par :

— $\hat{a}_{ML}(y) = \arg\max_a \left[f_{Y|A}(y|a)\right]$.

On notera qu'une condition nécessaire d'existence pour qu'un estimateur ML est qu'il existe un point pour lequel la dérivée de la vraisemblance s'annule :

$$\frac{\partial}{\partial a} \log f_{Y|A}(y|a) \Big]_{a=\hat{a}_{ML}(y)} = 0.$$

Notons qu'il existe d'autres classes d'estimateurs de paramètres, en l'absence de connaissance a priori. En particulier les estimateurs non biaisés à variance minimale (voir par exemple [9], chap.4, les estimateurs MVUE), que nous n'étudierons pas dans ce cours.

Pour l'exemple étudié dans l'exercice 2.10 on obtient :

$$\hat{a}_{ML}(\boldsymbol{y}) = \frac{1}{\bar{y}}.$$

Cet estimateur est-il performant ? Nous allons traiter cet aspect dans la section suivante, mais commençons par une rapide analyse sur cet exemple.

Prenons $A = a$ fixé. La question que l'on se pose est de savoir si lorsque le nombre de mesures tend vers l'infini, l'estimation $\hat{a}_{ML}(\boldsymbol{y})$ tend vers a ?

Nous savons que pour a fixé

$$\lim_{N\to\infty} \bar{y} = \mathbb{E}_{Y|A=a}[y],$$

et par propriété de la loi exponentielle, on a

$$\mathbb{E}_{Y|A=a}[y] = a^{-1}.$$

L'estimateur est donc consistant, c'est-à-dire que la valeur estimée tend vers la vraie valeur lorsque le nombre de mesures tend vers l'infini. Vous pourrez vérifier si cette propriété est vérifiée pour les autres estimateurs.

2.6.3 Performance des estimateurs

L'estimateur ML a de très bonnes performances en général. On s'intéressera à mesurer son biais et la variance de son erreur. En particulier, la convergence des estimateurs

avec le nombre de mesures est très importante, notamment lorsque l'on veut vérifier des résultats théoriques par simulation. Typiquement, l'évaluation du BER d'une transmission numérique, ou du taux d'erreur paquet dans un réseau. Il est d'usage d'effectuer des simulations extensives et de calculer le comportement moyen. On parle alors de simulation Monte-Carlo. Ce problème peut être assimilé à un problème d'estimation (il s'agit ici d'estimer la probabilité d'erreur). Le calcul théorique que nous développons ci-dessous permet d'une part de démontrer si la simulation a un biais (justesse du résultat), et d'autre part, de calculer la variance en fonction du nombre de mesures, ce qui permet d'en déduire le nombre d'expériences à réaliser pour garantir la fiabilité des résultats.

En résumé, on cherchera à répondre aux questions suivantes :

— L'estimateur est-il biaisé ? Autrement dit la valeur retournée est-elle en moyenne juste ?

— A quelle vitesse l'estimateur converge-t-il vers la bonne valeur lorsque l'on fait croître le nombre de mesures ?

La réponse est liée à l'étude de la variance de l'estimateur. Idéalement, il faut démontrer que la variance tend vers 0 quand le nombre de mesures tend vers l'infini, et on s'intéresse alors à la vitesse à laquelle cette variance tend vers 0.

Définition 2.16 : Performance des estimateurs

Soit un estimateur $\hat{a}(\boldsymbol{y})$ d'un paramètre a, fixé, mesuré au travers d'un vecteur d'observations iid : $\boldsymbol{Y} \in \boldsymbol{\mathcal{Y}}^N$.

1. Le biais de l'estimateur est défini comme l'erreur moyenne d'estimation pour une valeur donnée a :

$$\epsilon(a, N) := \mathbb{E}_{\boldsymbol{Y}|A=a}\left[\hat{a}(\boldsymbol{y})\right] - a.$$

2. L'estimateur est consistant (convergence en probabilité) si :

$$\lim_{N\to\infty} \epsilon(a, N) = Pr(|\hat{a}(\boldsymbol{y}) - a| > \varepsilon) = 0; \forall \varepsilon > 0.$$

3. La variance d'estimation est :

$$\mathrm{var}(a, N) := \mathbb{E}_{\boldsymbol{Y}|A=a}\left[\hat{a}(\boldsymbol{y})^2\right] - \mathbb{E}_{\boldsymbol{Y}|A=a}\left[\hat{a}(\boldsymbol{y})\right]^2.$$

Reprenons l'exemple 2.10 (pour $N > 1$). L'application direct de la formule du biais implique de calculer :

$$\mathbb{E}_{\boldsymbol{Y}|A=a}\left[\hat{a}(\boldsymbol{y})\right] = \int_{\boldsymbol{y}\in\boldsymbol{\mathcal{Y}}^N} \frac{1}{\bar{y}} \cdot f_{\boldsymbol{Y}|A}\left(\boldsymbol{y}|a\right), \tag{2.18}$$

où $f_{\boldsymbol{Y}|A}\left(\boldsymbol{y}|a\right)$ a été définie dans (2.6.1).

Le calcul direct de cette expression n'est pas triviale car elle nécesite une intégration multiple. Si $N = 1$, on obtient :

$$\mathbb{E}_{\boldsymbol{Y}|A=a}\left[\hat{a}(\boldsymbol{y})\right] = \int_{y\in\boldsymbol{\mathcal{Y}}} \frac{a}{y} e^{-a\cdot y} \cdot dy.$$

Cette intégrale est divergente, voir par exemple [3]. En effet, la fonction $x^{-1}e^{-x}$ tends vers l'infini en 0.

Pour $N > 1$, on peut exploiter le résultat trouvé dans (2.6.1), qui montre que la vraisemblance ne dépend que de $\bar{y}$. Comme pour la détection, cela indique que $\bar{y}$ est une statistique suffisante pour ce problème d'estimation. On s'intéressera alors à l'expression suivante :

$$\mathbb{E}_{\bar{Y}|A=a}\left[\hat{a}(\bar{y})\right] = \int_{\bar{y}\in\boldsymbol{\mathcal{Y}}} \frac{1}{\bar{y}} \cdot f_{\bar{Y}|A}\left(\bar{y}|a\right) \cdot d\bar{y}.$$

Pour calculer $f_{\bar{Y}|A}\left(\bar{y}|a\right)$, il faut utiliser la loi de composition des variables aléatoires. En effet, la fonction caractéristique de $\bar{y}$ est le produit des fonctions caractéristiques des y_k. On obtient :

$$f_{\bar{Y}|A}\left(\bar{y}|a\right) := \begin{cases} \frac{(Na)^N}{N!} \cdot \bar{y}^{N-1} \cdot e^{-Na\bar{y}} & si \quad \bar{y} \geq 0 \\ 0 & si \quad y < 0 \end{cases} \tag{2.19}$$

Grâce à cette expression, l'intégration [8] donne :

$$\mathbb{E}_{\boldsymbol{Y}|A=a}\left[\hat{a}_{ML}(\boldsymbol{y})\right] = \frac{N \cdot a}{N-1}.$$

Le biais est donc, pour $N > 2$:

$$\epsilon(a, N) = \frac{a}{N-1}.$$

L'estimateur ML dans ce cas est donc biaisé, avec un biais égal à $\frac{a}{N-1}$, pour $N > 1$. Cependant il est asymptotiquement non biaisé : le biais tend vers 0 quand $N \to \infty$.

Notons qu'il est facile de construire un estimateur qui compense ce biais :

$$\hat{a}_{NB}(y) := \frac{N-1}{N} \cdot \hat{a}_{ML}(y).$$

Ainsi, pour $N = 2$, on obtient $\hat{a}_{NB}(y) = 0.5 \cdot \hat{a}_{ML}(y)$, ce qui montre l'intérêt de cette modélisation pour améliorer l'estimation de paramètres statistiques.

Toujours pour notre exemple, vous vérifierez que la variance de nos 2 estimateurs est (pour $N > 2$) :

$$\begin{aligned} \mathrm{var}(\hat{a}_{ML}(y)) &= \frac{a^2 N^2}{(N-1)^2(N-2)} \\ \mathrm{var}(\hat{a}_{NB}(y)) &= \frac{a^2}{N-2}. \end{aligned}$$

On notera que la variance de l'estimateur NB est légèrement plus faible que celle du ML.

Plutôt qu'un calcul direct, il est possible d'utiliser les inégalités de Cramer-Rao (établies par Cramer et Rao respectivement en 1942 et 1945, à partir des travaux de Fisher (1922) et Dugué (1945), que nous ne démontrerons pas dans ce cours.

Théorème 2.6 : Inégalités de Cramer-Rao

Soit un estimateur $\hat{a}(y)$ quelconque non biaisé de a, tel que $f_{Y|A}\left(y|a\right)$ possède des dérivées partielles premières et secondes par rapport à a absolument intégrables. La

8. L'inrégrale à calculer est connue : $\int_0^\infty x^{\nu-1}\exp(-\mu x)dx = \frac{1}{\mu^\nu}\Gamma(\nu)$ d'après [3], 3.381 (4), p.346

variance de l'erreur vérifie alors les propriétés suivantes :

$$\mathrm{Var}[\hat{a} - a] \geq -\left\{ \mathbb{E}_{Y|A=a} \left[\frac{\partial^2 \log f_{Y|A}(y|a)}{\partial a^2} \right] \right\}^{-1}$$

Un estimateur qui atteint l'égalité est dit efficace. Il vérifie alors :

$$\frac{\partial \log f_{Y|A}(y|a)}{\partial a} = (\hat{a}(y) - a) \cdot g(a), \tag{2.20}$$

où $g(a)$ est une fonction indépendante des observations y. Si un estimateur non biaisé et efficace existe, on peut montrer que :

$$\hat{a}(y) = \hat{a}_{ML}(y)$$
$$Var(\hat{a}(y) - a) = g(a)^{-1}.$$

Autrement dit, il est équivalent à l'estimateur ML et sa variance est définie par $g(a)^{-1}$.

Revenons sur notre exemple, et calculons la borne Cramer-Rao lower bound (CRLB) à partir de la dérivée seconde de la densité a posteriori. On obtient

$$CRLB = \frac{a^2}{N}.$$

On peut observer qu'aucun de nos deux estimateurs n'est efficace, puisque leur variance est supérieure à la CRLB. Cependant, les deux estimateurs sont asymptotiquement efficaces, car la variance de ces estimateurs tend vers la CRLB. Ce résultat était prévisible, car si un estimateur non biaisé et efficace existait, ce serait le ML. De plus la borne de Cramer-Rao est une borne et n'est pas nécessairement atteignable.

Exercice 2.11 : Performance d'un estimateur

On observe une grandeur physique (toujours notre détecteur de fumée par exemple), à travers un système bruité, avec un ensemble de N mesures notées

$$y[k] = a + n[k]$$

On suppose encore une fois que le bruit est gaussien centré $N[k] \sim \mathcal{N}(0, \sigma^2)$. Les différentes mesures sont supposées indépendantes.

1. Calculez l'estimateur ML.
2. Déterminez si il est biaisé.
3. Déterminez les bornes de Cramer-Rao.
4. Déterminez si il est efficace.

2.7 Sujets avancés liés à l'estimation

Nous avons abordé dans ce cours les notions fondamentales de la théorie de l'estimation et de la détection. De nombreux sujets de recherche restent ouverts dans ce domaine,

en particulier lorsque le problème étudié est non-linéaire, à large dimension ou encore distribué.

Des bases dans ce domaine sont indispensable pour comprendre les bases des algorithmes d'apprentissage qui mettent en oeuvre de nombreuses étapes de détection ou d'estimation. Dans les approches proposées ici, la première difficulté est de modéliser la fonction de vraisemblance et l'a priori. Ensuite, il faut calculer l'estimateur, comme le ML ou le MAP, tâche qui s'avère particulièrement diffile. C'est pourquoi une approche par apprentissage profond contourne ces difficultés. Il s'agit d'aquérir une grande base de données d'observation (les $\boldsymbol{y}$ dans les notations précédentes), pour lesquelles l'entrée est connue (l'évènement recherché). Le réseau de neurone se substitue alors au modèle mathématique, et apprend la relation entre E et Y. Plus exactement il apprend d'estimation qui donne $\hat{E}$ en fonction de Y. Quelque part, la complexité de la modélisation mathématique est remplacée par l'exploitation de données massives. Ces approches sont apparues récemment dans les télécoms, voir par exemple [7]. Toutefois, cet aspect ne sera pas abordé dans ce cours.

2.7.1 Note sur l'estimation multi-paramètres

L'estimation multi-paramètre consiste à estimer un vecteur de paramètres $a = [a_1, \ldots, a_m]$.

Le problème n'est pas fondamentalement différent, mais l'espace de recherche étant multi-dimensionnel, les calculs de performance de ces estimateurs nécessitent de manipuler les opérateurs de dérivée partielle et le calcul matriciel.

TODO : A compléter seulement si nécessaire pour la suite du cours

2.7.2 Estimation itérative

Dans de nombreux systèmes, en particulier en télécommunications, nous devons estimer un paramètre qui évolue lentement au cours du temps. Il s'agit par exemple de l'estimation de la fréquence porteuse d'un signal modulé. Du fait des propriétés des composants électroniques utilisés à l'émetteur et au récepteur, la fréquence porteuse n'est pas constante lors de la transmission d'un flux d'information. Si l'on veut démoduler un signal, il faut détecter la fréquence porteuse dans le signal reçu, et la compenser. Une première approche consiste à numériser le signal reçu, le stocker, puis corriger à posteriori la fréquence. Cependant, une telle approche à deux inconvénients : elle ne permet que de compenser une erreur statique et elle introduit un délai de traitement puisqu'elle est réalisée en post-traitement.

Une autre approche consiste à estimer la fréquence porteuse courante, en cours de réception (typiquement à partir des N derniers échantillons reçus) et à compenser la fréquence en cours de réception. Cette approche est beaucoup plus efficace, mais nécessite un algorithme complexe et assez sophistiqué, qui met en général en oeuvre un estimateur itératif de maximum de vraisemblance. Typiquement, en télécommunications, le récepteur devra se synchroniser en fréquence, en phase, en amplitude (gain) et en rythme (période d'échantillonnage). Cette étape de synchronisation est essentielle pour les communications numériques et très complexe à implémenter, car idéalement il faut estimer simultanément tous ces paramètres de façon itérative. Ce sujet sort du cadre de ce cours.

TODO : il pourrait être intéressant de développer un exemple de type estimation itérative de phase, mais également une section sur les algorithmes de type message passing ou factor graphs.

2.8 Application aux communications numériques

Pour conclure ce chapitre, nous illustrons les éléments théoriques vus dans ce chapitre, à partir de petits problèmes classiques de communications numériques, que nous détaillerons dans les chapitres suivant. Cette section est en quelque sorte une mise en bouche des chapitres à venir, et justifie les efforts faits jusque là pour comprendre ces outils théoriques.

2.8.1 Application de la détection

La théorie de la détection est utilisée en transmission dès que l'on veut prendre une décision dans un ensemble discret, à partir d'une ou plusieurs observations. Voici quelques applications

Démodulation numérique

Lorsque qu'une transmission est effectuée selon un schéma traditionnel en bande de base (voir cours CMN), la démodulation s'effectue en général symbole par symbole. On rappelle que la démodulation numérique se réfère à l'étape de démapping, c'est à dire la conversion d'un nombre complexe (symbole I/Q) en un symbole de la constellation, qui code un mot binaire.

Considérons le cas d'un canal sans mémoire. On s'intéresse donc à l'estimation des symboles, indépendamment les uns des autres, et on a la relation suivante $\underline{y}_e[k] = \underline{x}_e[k] + \underline{n}_e[k]$. Ainsi, l'espace des évènements $\boldsymbol{\mathcal{E}}$ n'est autre que la constellation, de taille 2^{N_b} avec N_b le nombre de bits transmis par symbole.

Le démodulateur numérique implémente donc bien un processus de décision.

Si les symboles sont équiprobables, l'estimateur ML est optimal, ce qui permet de définir les sous-espaces de décision associés à chaque symbole, de façon unique. Chaque nombre complexe reçu est assimilé au symbole le plus proche, comme illustré à la figure 2.7. C'est ce qui est mis en oeuvre dans l'exercice 2.12.

En effet, le calcul du ML, dans le cas d'un bruit gaussien même en complexe, revient à recherche le symbole le plus proche. Autrement dit l'estimateur ML est équivalent à l'estimateur MMSE. Cette équivalence valable uniquement ou presque, dans le cas du bruit gaussien, simplifie bien les choses.

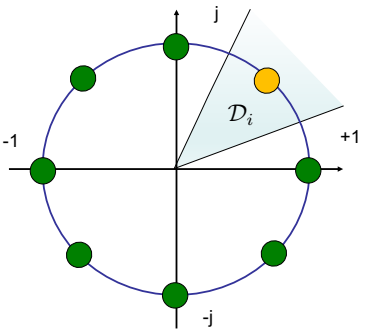


FIGURE 2.7 – Région de décision $\mathcal{D}_i$ associée au symbole s_i, lorsque le canal est AWGN.

Vous pourrez facilement résoudre l'exercice suivant en vous appuyant sur ces éléments.

Exercice 2.12 : Démodulation optimale

1. Déterminez pour différentes modulations linéaires : modulation par déplacement de phase à deux états –binary phase-shift keying– (BPSK), quadrature phase-shift keying (QPSK), 16−QAM, 8−phase-shift keying (PSK)), quels sont les sous-espaces de décision associés à chaque symbole.
2. Imaginez qu'un constellation 4−QAM possède un symbole de probabilité supérieure au trois autres, et que cette propriété soit connue du récepteur. Comment peut-il améliorer ses performances ?

Décodage

Intéressons nous maintenant au décodage d'un mot codé par un code systématique, passé dans un canal binaire symétrique, de probabilité d'erreur p. L'espace d'évènements est l'ensemble des mots codes valides. Le bruit associé ici n'est plus un bruit gaussien, puisque nous travaillons alors dans un corps F_2^n. On note ce vecteur aléatoire W^n, à réalisation dans $\boldsymbol{\mathcal{W}}^n$. C'est l'entrée du canal. On observe en sortie de canal une variable aléatoire C^n.

Soit un mot code w^n, choisi aléatoirement et c^n le mot code reçu. On note la probabilité conditionnelle de cette réalisation $P_{C^n|W^n}(c^n|w^n)$. Le canal étant sans mémoire, on peut écrire :

$$P_{C^n|W^n}(c^n|w^n) = \prod_k P_{C|W}(c[k]|w[k])$$

Pour un canal binaire symétrique, on a $P_{C|W}(c[k]|w[k]) = 1 - p_e$ si $c(k) = w(k)$ et $P_{C|W}(c[k]|w[k]) = p_e$ si $c(k) \neq w(k)$. Ainsi, la probabilité s'écrit :

$$P_{C^n|W^n}(c^n|w^n) = (1-p_e)^m \cdot p_e^{n-m}$$

où m est égal au nombre de bits pour lesquels $c(k) = w(k)$.

SI les séquences binaires sont équiprobables (d'où l'importance du codage de source), alors le détecteur optimale est celui qui minimise la distance de Hamming. La théorie de la détection permet donc de justifier le choix d'un décodeur basé sur la distance de Hamming.

Toutefois dans un canal plus complexe avec des retards et des échos multiples, le décodeur optimal n'est plus celui qui minimise la distance minimal. Nous ferons appel à des techniques plus sophistiquées, comme l'algorithme de Viterbi, ou la propagation de croyance.

Egalisation ML

TODO : revoir les notations une fois réécrit le chapitre 4

Nous verrons que l'égalisation consiste à annuler l'effet du canal pour retrouver le signal émis. Soit un signal échantillonné de durée T_x, représenté par un vecteur $\underline{\boldsymbol{x}}$. Après passage dans un canal on peut écrire :

$$\underline{\boldsymbol{y}} = \underline{H} \cdot \underline{\boldsymbol{x}} + \underline{\boldsymbol{n}}. \tag{2.21}$$

Effectivement, nous savons que $\underline{\boldsymbol{x}} \in \boldsymbol{\mathcal{X}}$ qui est l'ensemble des séquences que peut émettre le transmetteur. Si le canal est AWGN, alors chaque élément de $\underline{\boldsymbol{n}}$ suit une loi normale centrée et indépendante des autres éléments.

Prenons une séquence possible $\underline{\boldsymbol{x}}_i$. On calcule alors une prédiction du signal de sortie $\underline{\tilde{\boldsymbol{y}}}_i = \underline{H} \cdot \underline{\boldsymbol{x}}_i$. La vraisemblance de $\underline{y}$ est égale au produit des vraisemblances de chaque échantillon :

$$f_{Y|X}\left(\underline{y}|\underline{x}\right) = \prod_k f_{Y[k]|X}\left(\underline{y}[k]|\underline{x}\right) \tag{2.22}$$

Nous verrons dans le chapitre correspondant comment on peut résoudre ce problème d'estimation, qui est loin d'être trivial. Encore aujourd'hui, la recherche d'algorithmes faisant un compromis entre complexité de calcul et efficacité, occupe les ingénieurs ou les chercheurs.

2.8.2 Applications de l'estimation

La théorie de l'estimation est utilisée dès lors que l'on cherche à estimer un paramètre : estimation de la fréquence porteuse, de l'amplitude du signal, du niveau de bruit, etc...

Estimation du bruit de récepteur

Dans de nombreuses systèmes radio, le récepteur est en mesure d'estimer le niveau de bruit (amplitude ou puissance). Le récepteur écoute le canal radio, vérifie qu'aucun signal utile n'est présent (détection) et effectue une estimation de puissance de bruit. Ce bruit peut intégrer le bruit du récepteur mais également les interférences provenant d'autres systèmes. On fait classiquement l'hypothèse que ce bruit est gaussien. Il suffit alors de dérouler les techniques d'estimation vues précédemment pour obtenir un estimateur de bruit. On peut montrer que dans le cas idéal gaussien, l'estimateur ML est optimal et retourne simplement $P_{bruit} = \mathbb{E}\left[|Y[k]|^2\right]$.

Estimation du niveau de signal

Le received signal strength indicator, indicateur de puissance de signal reçu (RSSI) est une mesure faite par certains équipements radio, en particulier le Wifi, pour déterminer la puissance du bruit du signal utile. Il s'effectue en général à partir de l'étude du préambule, qui est une séquence connue. L'estimation de puissance est basée sur la corrélation avec le signal connu, voir l'exercice 2.19.

Estimation des coefficients de canal

Nous avons donné ci-dessus l'exemple de l'égalisation comme application de la détection. Pour pouvoir mettre en oeuvre cet égalisation, le récepteur doit connaitre le canal, c'est à dire les coefficients de la matrice $\underline{H}$. Ces coefficients sont estimés à partir de l'analyse de signaux pilotes, dont nous détaillerons le fonctionnement. La théorie de l'estimation est essentielle pour choisir la longueur de ces signaux pilotes.

Prendre des signaux pilotes longs permet d'améliorer les performances, mais occupe plus de bande passante, réduisant d'autant la bande passante utile. Au contraire, prendre des signaux pilotes courts optimise la bande passante, mais dégrade les performances. La théorie de l'estimation est l'outil idéal pour optimiser ces choix.

Estimation des propriétés statistiques de canal

Prenons un cas plus simple décrit par :

$$\underline{y}[k] = \underline{h} \cdot \underline{x}[k] + \underline{n}[k], \tag{2.23}$$

où $\underline{h}$ est un simple coefficient scalaire. Malheureusement, ce coefficient de canal varie lentement au cours du temps, et nous souhaitons estimer les propriétés statistique de cette variable. A-t-on un canal gaussien (coefficient statique) ? ou un canal de Rice (voir chapitre 4) ? La connaissance des paramètres statistiques du canal est très utile pour optimiser les émetteurs et récepteurs.

Pour cela, nous pouvons nous baser encore une fois sur l'estimation de paramètres statistiques tels que vus dans ce chapitre.

2.9 Synthèse du chapitre

Nous avons introduit dans ce chapitre les notions fondamentales de la théorie de la détection et de la théorie de l'estimation. Ce chapitre ne donne qu'un petit aperçu de la richesse de ces outils théoriques, abondamment utilisés dans tous les domaines de l'ingénierie. De nombreux problèmes sont encore ouverts, en particulier pour l'étude de systèmes à grande échelle ou de systèmes dynamiques. Les performances de ces outils dépendent beaucoup de la qualité de la modélisation du problème, et de nombreux travaux scientifiques consistent à améliorer, pour différentes applications, les modèles. Il ne faut pas oublier que ces problèmes sont résolument multi-objectifs et que l'amélioration d'un algorithme peut consister à réduire sa complexité, améliorer sa robustesse, augmenter ses performances en termes d'erreur.

Nous exploiterons ce formalisme dans les chapitres suivants de ce cours.

On peut résumer les éléments essentiels de ce cours

1. Principes de la théorie de la détection :
   - Risque de Bayes.
   - Rapport de vraisemblance et test d'hypothèse.
   - Log vraisemblance.
   - Courbes ROC.
   - Estimateurs ML, MSE, MAP.
2. Principes de la théorie de l'estimation :
   - Fonctions de coût.
   - Estimateurs optimaux.
   - Inégalités de Cramer-Rao.

Enfin, au travers des développements proposés et des exemples discutés nous avons développé les savoir-faire suivants :

1. Formuler un problème de détection ou d'estimation dans le domaine des communications numériques.
2. Evaluer les performances en termes de probabilité d'erreur.

2.10 Exercices

2.10.1 Exercices portant sur la théorie de la détection

Exercice 2.13 : Test binaire de position en canal gaussien

Soit $V[,]$ avec $k \in [1; N]$, une séquence iid de loi normale $\mathcal{N}\left(0, \sigma^2\right)$.
On considère le test d'hypothèse suivant :

$$\begin{array}{rcl} \mathbb{H}_0 & : & Y[k] = m_0 + V[k] \\ \mathbb{H}_1 & : & Y[k] = m_1 + V[k] \end{array}$$

pour $1 \leq k \leq N$, où on supposera que $m_1 > m_0$. L'observation est un vecteur aléatoire noté $\boldsymbol{Y} = [Y[1], Y[2], \ldots, Y[N]]^t$.

Les coût d'erreur sont donnés par c_{01} et c_{10} et les probabilités a priori des deux évènements sont p_0 et p_1.

A- Détecteur Bayésien :
On veut déterminer pour un vecteur $\mathbf{y}$ observé, quel est l'évènement qui minimise le risque de Bayes.

1. Calculez l'expression des densités de probabilité $f_{\boldsymbol{Y}|E}\left(\mathbf{y}|e_i\right)$.
2. Calculez la fonction du rapport de vraisemblance.
3. Démontrez en utilisant le LRT que la décision optimale au sens de Bayes se ramène au test :
$$S \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \gamma,$$
où $S = \frac{1}{N}\sum_{k=1}^{N} Y[k]$, et où vous déterminerez la valeur de γ.
4. Déterminez les tests à effectuer pour choisir l'estimateur ML puis l'estimateur MAP.

B- Courbes ROC et tests de Neyman-Pearson :
On prendra pour simplifier $m_0 = 0$, et $m_1 = \rho$; ρ caractérise donc la distance entre les 2 évènements. Ceci n'est pas restrictif car dans le cas général, il suffirait de modifier le signal reçu en : $\boldsymbol{Y} - m_0$.

1. Reformulez le test d'hypothèse avec ces notations.
2. Calculez les probabilités de détection et de fausse alarme. Vous utiliserez la fonction $Q(x)$ et la distance $d = \frac{N^{1/2}\rho}{\sigma}$.
3. Exprimez p_d en fonction de p_{fa}.
4. Tracez sous Matlab les courbes ROC pour différentes valeur de d.
5. Vérifiez leurs propriétés : continuité, concavité.

Exercice 2.14 : Test de variance

On observe un signal iid de loi normale sur $1 \leq k \leq N$, avec 2 hypothèses concernant sa variance :

$$\begin{array}{rcl} \mathbb{H}_0 & : & Y[k] \sim \mathcal{N}\left(0, \sigma_0^2\right) \\ \mathbb{H}_1 & : & Y[k] \sim \mathcal{N}\left(0, \sigma_1^2\right) \end{array}$$

pour $1 \leq k \leq N$, où on supposera que $\sigma_1^2 > \sigma_0^2$. L'observation est un vecteur aléatoire noté $\boldsymbol{Y} = [Y[1], Y[2], \ldots, Y[N]]^t$.

A- Détecteur Bayésien :
On veut déterminer pour un vecteur $\mathbf{y}$ observé, quel est l'évènement qui minimise le risque de Bayes.

1. Calculez l'expression des densités de probabilité $f_{\boldsymbol{Y}|E}\left(\mathbf{y}|e_i\right)$.
2. Calculez la fonction du rapport de vraisemblance.
3. Démontrez en utilisant le LRT, que la décision optimale au sens de Bayes se ramène au test :
$$S \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \gamma,$$
où $S = \frac{1}{N}\sum_{k=1}^{N} Y[k]^2$, et où vous déterminerez la valeur de γ.
4. Déterminez les tests à effectuer pour choisir l'estimateur ML puis l'estimateur MAP.

B- Courbes ROC et tests de Neyman-Pearson :
On se place dans le cas où le nombre d'observation est $N = 2$.

1. Montrez que le LRT peut s'écrire
$$T := Y[1]^2 + Y[2]^2 \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \kappa,$$
où vous expliciterez κ.
2. Calculez les probabilités de détection et de fausse alarme.
3. Exprimez p_d en fonction de p_{fa}.
4. Tracez sous Matlab les courbes ROC pour différentes valeur de $r = \sigma_0^2/\sigma_1^2$.
5. Bonus : Comment ces résultats se généralisent-ils au cas où $N > 2$. Tracez l'évolution de la courbe ROC en fonction du nombre de mesures.

Exercice 2.15 : Test Poisson

On observe un signal iid sur $\mathbb{N}^+$, distribué selon une loi de Poisson, avec 2 hypothèses :

$$\begin{array}{rcl} \mathbb{H}_0 & : & P_{Y|E}\left(y=n|e_0\right) = \frac{\lambda_0^n}{n!}\exp(-\lambda_0) \\ \mathbb{H}_1 & : & P_{Y|E}\left(y=n|e_1\right) = \frac{\lambda_1^n}{n!}\exp(-\lambda_1) \end{array}$$

pour $1 \leq k \leq N$, où on supposera que $\lambda_1 > \lambda_0$. Le vecteur observé est noté $\boldsymbol{Y} = [Y[1], Y[2], \ldots, Y[N]]^t$.

A- Détecteur Bayésien :
On veut déterminer pour un vecteur $\boldsymbol{Y}$ observé, quel est l'évènement qui minimise le risque de Bayes.

1. Calculez l'expression des densités de probabilité $P_{\boldsymbol{Y}|E}\left(\mathbf{n}|e_i\right)$.
2. Calculez la fonction du rapport de vraisemblance.
3. Démontrez en utilisant le LRT, que la décision optimale selon le critère de vraisemblance se ramène au test :

$$S \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \gamma, \tag{2.24}$$

où $S = \sum_{k=1}^{N} Y[k]$, et où vous déterminerez la valeur de γ.

B- Courbes ROC et tests de Neyman-Pearson :
On se place dans le cas où le nombre d'observation est $N = 1$, par soucis de simplification.

1. Montrez que le LRT peut s'écrire

$$Y \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \gamma,$$

où vous expliciterez γ.

2. Calculez les probabilités de détection et de fausse alarme, en fonction d'un seuil entier $\kappa = \lfloor \gamma \rfloor$.
3. Placez sous Matlab les points $(p_d(\kappa), p_{fa}(\kappa))$ pour différentes valeurs de n et pour différents couples (λ_1, λ_2). Par exemple, $\lambda_1 = 0.5$, $\lambda_2 = 5$.
4. Comment peut-on tracer entièrement la courbe ROC ? Mettez en place un test avec randomisation. Et tracez le résultat.
5. Bonus : comment se généralisent ces résultats si $N > 1$?

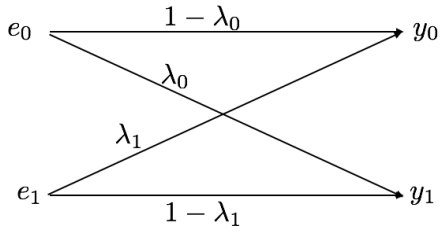
Exercice 2.16 : Canal binaire

Le canal binaire modélise la situation où un digit (0 ou 1) doit être transmis à travers un canal de communication bruité. On note $e_0 = 0$, et $e_1 = 1$ respectivement les états d'entrée du canal, soit $\boldsymbol{\mathcal{E}} = \{0, 1\}$. L'observation est une variable aléatoire $Y \in \boldsymbol{\mathcal{Y}} = \{0, 1\}$. Les erreurs de transmission sont représentées par les probabilités

conditionnelles suivantes :

$$P_{Y|E}(y=0|0) = 1-\lambda_0$$
$$P_{Y|E}(y=1|0) = \lambda_0$$
$$P_{Y|E}(y=0|1) = 1-\lambda_1$$
$$P_{Y|E}(y=1|1) = 1-\lambda_1$$

Ce modèle est représenté ci-dessous :



On cherche quelle est la meilleure décision à prendre en réception. Ce problème est typiquement un problème de détection.

A- Détecteur Bayésien :
On veut déterminer pour une observation y, l'évènement qui minimise le risque de Bayes. On choisira bien entendu $c_{01} = c_{10} = 1$ et $c_{00} = c_{11} = 0$.

1. Calculez l'expression du rapport de vraisemblance $\Lambda(y)$ pour chaque valeur possible de y.
2. Déterminez la règle de Bayes, dans le cas où les probabilités à priori sont identiques.
3. Calculez le risque de Bayes correspondant à l'estimateur optimal. Est-ce équivalent au ML ?
4. Déterminez la règle de Bayes lorsque $\lambda_0 = \lambda_1$.

B- Courbes ROC et tests de Neyman-Pearson :
Dans ce cas, discret, la randomisation est nécessaire pour déterminer la courbe ROC. On rappelle que pour trouver un test de Neyman-Pearson de contrainte $P_{fa} \leq \alpha$, il faut rechercher un test d'hypothèse $\Lambda(y) \underset{\mathbb{H}_0}{\overset{\mathbb{H}_1}{\gtrless}} \eta$, en adaptant le seuil de façon appropriée.

Nous nous plaçons dans le cas où

$$\lambda_0 + \lambda_1 < 1,$$

ce qui permet de vérifier

$$\frac{\lambda_1}{1-\lambda_0} < \frac{1-\lambda_1}{\lambda_0}.$$

1. A partir des valeurs possibles de $\Lambda(y)$, calculez en fonction du choix du seuil η, les probabilités de fausse alarme et de détection.

2. Dans le plan de représentation des courbes ROC, positionnez les points obtenus en fonctions de η.
3. Calculez et tracer la courbe ROC en prolongeant ces points par randomisation.
4. Déterminez alors le test de Neyman-Pearson pour une certaine valeur de α et exprimez en fonction de α les valeurs de seuil à choisir $\eta(\alpha)$ et de randomisation γ_0.

2.10.2 Exercices portant sur la théorie de l'estimation

Exercice 2.17 : Estimateurs optimaux pour lois gaussiennes

On réalise un ensemble de N mesures regroupées dans le vecteur Y. Chaque mesure est du type $Y[k] = a + N[k]$ où a est un paramètre inconnu, et les $N[k]$ sont des échantillons iidselon une loi Normale $\mathcal{N}\left(0, \sigma_n^2\right)$. Le paramètre a est modélisé par une variable aléatoire $A \sim \mathcal{N}\left(0, \sigma_a^2\right)$.

1. Ecrivez la loi de vraisemblance $f_{Y|A}\left(y|a\right)$.
2. Calculez la loi a posteriori $f_{A|Y}\left(a|y\right)$.
3. Calculez les 3 estimateurs $\hat{a}_{MMSE}$, $\hat{a}_{MMAE}$ et $\hat{a}_{MAP}$.
4. Vérifiez puis justifiez à partir de la forme de $f_{A|Y}\left(a|y\right)$ pourquoi les 3 estimateurs sont identiques.

Exercice 2.18 : Estimation d'une loi de Poisson

Une station de base pour l'internet des objets est déployée. Afin de dimensionner la cellule, on observe la fréquence à laquelle des paquets sont émis par les objets communicants. On détermine que le nombre de paquets n émis durant un temps T_s est une variable aléatoire notée N, qui suit une loi de Poisson en fonction d'un paramètre a qui est proportionnel au nombre d'objets dans la cellule et à la probabilité d'appel de chaque noeud durant la période T_s.

$$P_{N|A}\left(n|a\right) = \frac{a^n}{n!} exp(-a)$$

La station de base veut estimer a à partir de l'observation n.
On suppose que a suit une loi exponentielle unilatérale :

$$p(a) = \begin{array}{ll} \lambda \exp(-\lambda a) & \text{si} \quad a > 0 \\ 0 & \text{sinon} \end{array} \tag{2.25}$$

où λ est un paramètre connu.

1. Déterminez la densité de probabilité a postériori.
2. Calculez l'estimateur MMSE.
3. Calculez l'estimateur MAP.
4. Sont-ils identiques ? Discutez.

Exercice 2.19 : Estimation de l'amplitude d'un signal

Considérons le cas d'un signal observé $Y[k]$.

$$Y[k] = A \cdot s_k + N[k],$$

où s_k est un signal connu, d'amplitude unitaire, A est l'amplitude du signal à estimer, et $N[k]$ est un bruit décrit par sa loi multi-variée $\mathcal{N}(0, K_{nn})$, où K_{nn} est la matrice de covariance du bruit. On suppose que l'amplitude suit une loi normale $A \sim \mathcal{N}(\mu, \sigma_a^2)$.

1. Exprimez la vraisemblance, les les lois marginales de A et de Y.
2. Montrez que la loi a posteriori est une loi normale, et donnez l'expression de ses paramètres : moyenne et variance.
3. Calculez l'estimateur MMSE, et ainsi que le risque de Bayes associé.
4. Les 3 estimateurs MAP, MMAE et MMSE sont identiques. Pourquoi ?
5. Calculez l'estimateur ML et comparez-le au MAP.

Exercice 2.20 : Propriétés des estimateurs

Revenez, au choix, sur l'un des exercices (2.17, 2.18, 2.19). Pour un ou plusieurs estimateurs :

1. Posez le calcul du biais et de la variance de ces estimateurs.
2. Effectuez le calcul si il est faisable.
3. Calculez la borne de Cramer-Rao.

Chapitre 3

Théorie de l'information

3.1 Introduction

Dans le chapitre précédent, nous avons introduit les éléments clés de la théorie de la détection et de l'estimation, exploités abondamment dans la conception des protocoles de réseau (en couche PHY, mais pas seulement).

Avant d'attaquer l'étude des systèmes réels, il nous a semblé important de faire le lien entre la théorie de la détection, et la théorie de l'information, en particulier avec le fameux deuxième théorème de Shannon [10] sur la capacité des systèmes de transmission.

Vous connaissez l'expression de cette capacité sous la forme :

$$C = W \cdot \log_2(1 + SNR),$$

pour un canal AWGN. L'objectif de ce chapitre et de prouver ce théorème et d'en percevoir l'importance et le caractère fondamental. Cette capacité de Shannon peut être vue comme la limite fondamentale des systèmes de transmission, au même titre que la vitesse de la lumière est une limite fondamentale pour la vitesse de déplacement des particules.

C'est pourquoi la théorie de l'information joue un rôle essentiel dans la conception des systèmes de transmission.

Nous commencerons par définir précisément le canal de transmission étudié (section3.2), puis nous poursuivrons en rappelant quelques notions fondamentales de théorie de l'information, vues en 3^{ieme} année (section 3.3).

Ce qui est important dans ce chapitre, c'est de comprendre la démarche mathématique suivie par Claude Shannon et qui a permis de poser les bases d'un nouveau champ scientifique à l'interface entre les mathématiques et la physique. Dans cette partie, nous ne cherchons pas à concevoir un système mais à établir la preuve qu'il existe une limite fondamentale à tout système de communication, qui s'exprime sous la forme d'un débit maximal atteignable. Nous montrerons également que cette limite peut être approchée.

3.2 Modèle de référence d'un canal de transmission

Par anticipation sur la modélisation des systèmes qui sera détaillée au chapitre 4, la figure 3.1 représente un système de communication avec ses blocs de traitement principaux . Le canal représenté à droite de la figure par une densité de probabilité conditionnelle,

fait référence au modèle de canal analytique en bande de base. Nous en étudierons une version simplifiée qui est référencée sous le nom de canal gaussien. Nous traiterons de cas plus sophistiqués dans les chapitres suivants.

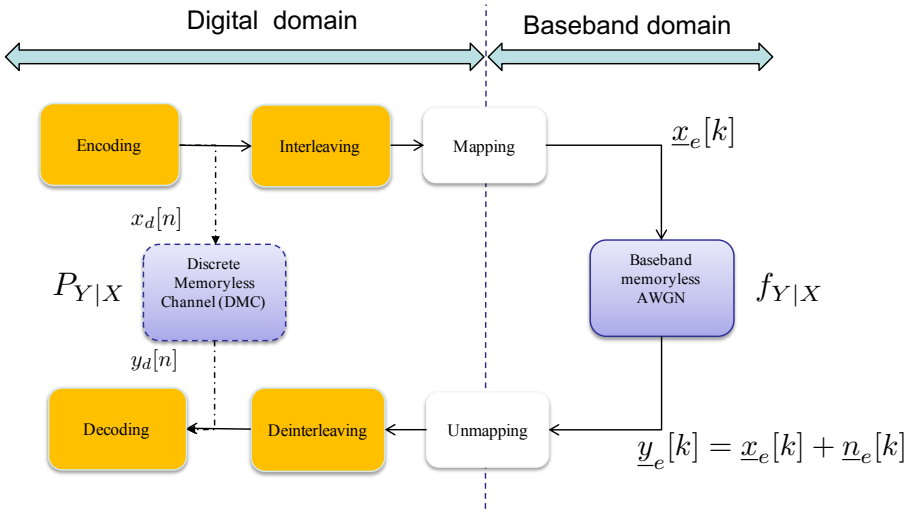


FIGURE 3.1 – Modèles de canaux sans mémoire étudiés dans cette section.

Ce canal relie le signal échantillonné, à valeurs continues, transmis en bande de base, noté $\underline{x}_e[k]$ et le signal échantillonné, à valeurs continues, reçu en bande de base, noté $\underline{y}_e[k]$. Les variables soulignées indiquent que l'on travaille ici en complexe.

Le signal reçu est synchronisé, compensé en phase et en amplitude, et le canal est supposé idéal à bruit additif gaussien, de telle sorte que l'on peut écrire :

$$\underline{y}_e[k] = \underline{x}_e[k] + \underline{n}_e[k], \tag{3.1}$$

où $\underline{n}_e[k]$ sont des échantillons de bruit gaussien, iid. Ainsi chaque échantillon est distribué selon une loi normale complexe $\underline{\mathcal{N}}\left(0, \sigma^2\right)$.

Il faut bien noter quel la sortie $\underline{y}_e[k]$ ne dépend que de $\underline{x}_e[k]$, et non des autres entrées. On peut également le décrire par ses probabilités : $f_{Y|X}\left(\underline{y}_e[k]|\underline{x}_e[k]\right)$, autrement dit par la probabilité d'observer un symbole $\underline{y}_e[k]$ lorsque l'on émet le symbole $\underline{x}_e[k]$. Cette densité de probabilité dépend directement et uniquement de la densité de probabilité du bruit. Nous retrouvons les notations vues dans le chapitre précédent et l'équivalence avec un problème de détection saute aux yeux.

Considérons un signal de n symboles émis. Le signal d'entrée est un vecteur aléatoire, noté $X^n = [X[1], X[2], \ldots, X[n]]$, et le signal de sortie est noté $Y^n = [Y[1], Y[2], \ldots, Y[n]]$. Ainsi pour ce modèle, on a

$$f_{Y^n|X^n}\left(y^n|x^n\right) = \prod_{k=1}^{n} f_{Y|X}\left(y[k]|x[k]\right). \tag{3.2}$$

Cette équivalence découle du modèle gaussien iid, où les réalisations sont indépendantes entre échantillons. En théorie de l'information, ce modèle est appelé gaussian memoryless channel, canal gaussien sans mémoire (GMC).

Bien que ce modèle soit particulièrement simple, le calcul de sa capacité, c'est à dire du débit maximal d'information, n'est pas trivial. Bien entendu, depuis les travaux de

Shannon, de nombreux cas plus complexes ont été étudiés (canaux avec évanouissements, non stationnaires, MIMO, etc...), dont nous retrouverons certains au cours des chapitres suivants.

Mais avant d'étudier ce canal, plus précisément, nous allons étudier le cas du canal discrete memoryless channel, canal discret sans mémoire (DMC), c'est à dire un modèle où les entrées et sorties prennent leurs valeurs dans des espaces discrets. Typiquement, sur la figure 3.1, nous allons nous placer en amont de la modulation numérique et de l'entrelaceur (interleaving). Rappelons que la modulation numérique transforme un symbole discret en un symbole complexe codé dans le plan $I - Q$ (in phase, quadrature). Le but de l'entrelaceur est de mieux répartir les erreurs et de réduire les corrélations possibles entre symboles successifs. Il n'introduit pas de codage, mais il permet de faire mieux coller le canal physique au modèle théorique du canal DMC.

Le signal d'entrée est un mot binaire (ou M-aire si on travaille directement dans un alphabet de plus grande dimension) de longueur n, et le signal de sortie est un mot binaire (ou M-aire) de longueur n également. Le canal correspondant, discret et sans mémoire, est caractérisé par l'ensemble des probabilités conditionnelles, d'observer y_j sachant que l'on a émis x_i et notée $P_{Y|X}\left(y_j|x_i\right)$. Sous les hypothèses d'indépendance, nous pouvons écrire

$$P_{Y^n|X^n}\left(y_i^n|x_j^n\right)) = \prod_{k=1}^{n} P_{Y|X}\left(y_i[k]|x_j[k]\right). \tag{3.3}$$

Si les symboles sont binaires, le canal correspondant est simplement le canal canal binaire symmétrique –binary symmetric channel– (BSC), que vous avez déjà étudié en cours de CMN. Ce canal est étudié à la section 3.4, puis nous passerons au canal gaussien dans la section 3.5.

3.3 Rappels de théorie de l'information

Avant d'étudier ce canaux, nous donnons quelques rappels de théorie de l'information (asymptotique). Nous ne reviendrons pas sur toutes les définitions et propriétés vues en 3^{ieme} année, mais nous rappelons les quelques éléments nécessaires à la compréhension du chapitre.

Pour une étude approfondie de la théorie de l'information, nous suggérons le livre de Thomas et Cover [1], mais qui va bien au-delà du cours.

3.3.1 Entropie

Définition 3.1 : Entropie d'une source

L'entropie d'une variable aléatoire est définie par la relation suivante :

$$H(A) := -\sum_{k} P_A\left(a_k\right) \log\left(P_A\left(a_k\right)\right).$$

Si la fonction log utilisée est en base 2, le résultat est en bits. Si on prend le logarithme néperien, on obtient un résultat en nats.

Cette entropie représente directement le nombre de bits minimal nécessaire pour coder l'information d'une telle source. Notez bien le caractère aléatoire de la source.

On rappellera que l'entropie est maximale, pour un domaine de définition donnée $\boldsymbol{\mathcal{A}}$, si la distribution est uniforme, autrement dit, si tous les symboles sont équiprobables.

Cette fonction d'entropie possède quelques propriétés intéressantes, que nous énonçons ici. L'entropie conjointe de deux variables aléatoires peut s'exprimer à partir de l'entropie conditionnelle :

$$\begin{aligned} H(A,B) &= H(A) + H(B|A), \\ &= H(B) + H(A|B). \end{aligned} \tag{3.4}$$

Si ces variables aléatoires sont conditionnées par une 3^{ieme} variable, on obtient :

$$H(A,B|C) = H(A|C) + H(B|A,C). \tag{3.5}$$

où $(A,B|C)$ s'interprète comme A et B conditionnés à C, alors que $(B|A,C)$ s'interprète comme B conditionné à (A et C).

Rappelons également que le conditionnement d'une variable aléatoire réduit toujours l'entropie :

$$H(A|B) \leq H(A). \tag{3.6}$$

Exercice 3.1 : Entropie

Quelques exercices sur l'entropie

1. Quelle est l'entropie au sens théorie de l'information d'un jet de dés non pipé ?
2. Quelle est l'entropie au sens théorie de l'information d'un double jet de dés non pipé ?
3. On effectue un jeu de pile ou face, à plusieurs tours, et on arrête lorsque l'on obtient face. On note X la variable aléatoire représentant le nombre de tirages réalisés. Quelle est l'entropie de X ?
4. soit X une variable aléatoire à valeurs discrètes (nombre fini). Pour chaque cas ci-dessous déterminez la relation entre $H(X)$ et $H(Y)$:
   - $Y = \exp(X)$
   - $Y = cos(X)$

3.3.2 Information mutuelle

Définition 3.2 : Information mutuelle

L'information mutuelle entre deux variables aléatoires discrètes A et B est définie par

$$I(A;B) = \sum_{a \in \boldsymbol{\mathcal{A}}, b \in \boldsymbol{\mathcal{B}}} P_{A,B}(a,b) \cdot \log \left[\frac{P_{A,B}(a,b)}{P_A(a) \cdot P_B(b)} \right].$$

L'information mutuelle est une mesure de l'information commune entre les 2 variables Aet B, et qui est calculée à partir d'une mesure de la différence entre la distribution conjointe et les distributions marginales de ces variables aléatoires.

On peut montrer la relation suivante, qui est illustrée à la figure 3.2 :

Propriété 3.1 : Information mutuelle et entropie conditionnelle

$$I(A;B) = H(A) - H(A|B).$$

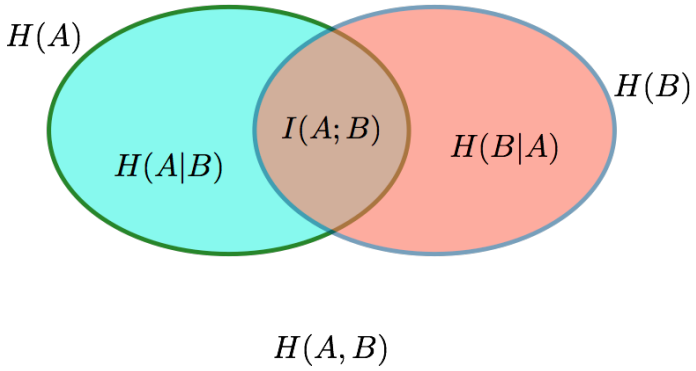


FIGURE 3.2 – Information mutuelle, entropie et équivoque.

Exercice 3.2 : Exercices sur l'information mutuelle

1. Donnez des exemples concrets mettant en jeu 3 variables aléatoires X, Y et Z, et pour lesquels, soit on a $I(X;Y|Z) < I(X;Y)$, soit on a $I(X;Y|Z) > I(X;Y)$.
2. Soient X_1 et X_2 deux variables distribuées identiquement, mais non nécessairement indépendantes. On note
$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}.$$
   - Montrez que $\rho = \frac{I(X_1;X_2)}{H(X_1)}$.
   - Montrez que $0 \leq \rho \leq 1$.
   - Quand a-t-on $\rho = 0$?
   - Quand a-t-on $\rho = 1$?

3.3.3 Règle de chaînage (Chain rule)

La règle de chaînage est utilisée dans de nombreux résultats de théorie de l'information et rentre en jeu lorsque l'on manipule plusieurs variables simultanément.

Propriété 3.2 : Chain Rule

Soit $X_1, X_2, \ldots, X_n$ un ensemble de n variables aléatoires et soit Y une variable aléatoire. On note $X^n = [X_1, \ldots, X_n]^t$. Alors :

$$I(X^n;Y) = \sum_{i=1}^{n} I(X_i;Y|X^{i-1}),$$

où $X^i = [X_1, \ldots, X_i]^t$.

Ce résultat est utilisé pour établir les preuves de capacité dans un canal sans mémoire.

3.3.4 Théorème de l'inégalité du traitement de l'information, ou Data Processing Inequality

Définissons tout d'abord ce qu'est une chaîne de Markov pour des v.a.

Définition 3.3 : chaîne de Markov

On dit que les variables X, Y et Z forment une chaîne de Markov si leur probabilité conjointe vérifie

$$P_{XYZ}(x, y, z) = P_X(x) \cdot P_{Y|X}(y|x) \cdot P_{Z|Y}(z|y).$$

On représente une chaîne de Markov sous la forme

$$X \to Y \to Z.$$

Cette définition est bien sûr extensible à N variables. L'élément important est que les variables (X, Y, Z) forment une chaine de Markov si la dépendance statistique entre X et Z n'existe qu'au travers de Y.

On peut alors établir le théorème suivant :

Théorème 3.1 : Data Processing Inequality

Si $X \to Y \to Z.$ forme une chaine de Markov, alors

$$I(X; Z) \leq I(X; Y).$$

Ce théorème donne un résultat fondamental pour la théorie de l'information appliquée aux communications numériques. Supposons que l'on observe Y à la sortie d'un canal de transmission, et que l'on veut estimer X. Et bien tout traitement des données reçues en Y, qu'il soit déterministe ou stochastique ne pourra que réduire l'information mutuelle, autrement dit l'information connue sur X. Ainsi, tout traitement a posteriori des données reçues ne peut que maintenir ou réduire l'information reçue (sauf à exploiter d'autres informations, que l'on appelle side information).

Beaucoup d'autres résultats importants relatifs à l'information mutuelle ont été établis, voir par exemple [1].

Exercice 3.3 : Exercice de révision

Tous ces théorèmes seront utilisés dans les 2 sections suivantes pour démontrer le théorème de la capacité de Shannon.

1. A titre d'exercice retrouvez dans les preuves l'usage de ces théorèmes et vérifiez qu'ils sont bien applicables.

3.4 Etude du canal DMC

3.4.1 Définition du canal discret sans mémoire (DMC)

Nous nous intéressons ici au modèle de canal qui transforme les échantillons $x_d[k]$ en échantillons $y_d[k]$, i.e.

$$x_d[k] \overset{P_{Y_d|X_d}}{\rightarrow} y_d[k]$$

, canal qui est représenté à gauche à la figure 3.1. Nous considérons de plus un canal binaire symétrique, ce qui veut dire que chaque utilisation de canal est codée sur $\mathcal{X} = \{0, 1\}$.

L'étude de la capacité de canal consiste à étudier un code long, utilisant n *channel uses* (on évitera d'utiliser le terme symbole qui peut prêter à confusion, et on préfèrera garder le terme de channel uses, pour ce qui relève de la théorie de l'information). On définit tout d'abord le canal DMC.

Définition 3.4 : Canal discret sans mémoire (DMC)

Un canal DMC, noté $\left(\mathcal{X}, \mathcal{Y}, P_{Y|X}\right)$, est défini par :

- Un alphabet $\mathcal{X} = \{x_1, \ldots, x_N\}$ des valeurs d'entrée.
- Un alphabet $\mathcal{Y} = \{y_1, \ldots, y_{N'}\}$ des valeurs de sortie.
- Une probabilité conditionnelle $P_{Y|X}$ où Y et X sont les variables aléatoires associées à un échantillon.

La transmission d'une certaine quantité d'information sur un canal discret sans mémoire se fait sur n channel uses, avec une méthode de transmission définie par :

Définition 3.5 : Transmission sur un canal discret sans mémoire

Un code (M, n) pour un canal DMC $\left(\mathcal{X}, \mathcal{Y}, P_{Y|X}\right)$ est défini par les éléments suivants :

- un ensemble d'indices $\mathcal{W} = \{1, 2, \ldots, M\}$,
- une fonction de codage $f_{enc} : \mathcal{W} \rightarrow \mathcal{X}^n$, qui constitue un dictionnaire (codeebook) de signaux d'émission $x_1^n, \ldots, x_M^n$,
- une fonction de décodage $f_{dec} : \mathcal{Y}^n \rightarrow \mathcal{W}$ basée sur la définition d'une partition de $\mathcal{Y}$: $\{\mathcal{D}_i; i \in \mathcal{W}\}$, tels que :
  - $\mathcal{D}_i = \{y^n; f_{dec}(y^n) = i\} \subset \mathcal{Y}^n$,
  - $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset; \quad \forall i \neq j$,
  - $\mathcal{D}_i^c = \mathcal{Y}^n - \mathcal{D}_i$.

Le débit de de transmission de ce code est $R = \frac{\log(M)}{n}$.

Comme nous l'avons déjà dit, l'objectif de cette section n'est pas de proposer une technique de codage efficace mais de définir les limites théoriques d'une telle technique de transmission.

Pour cela nous avons déjà défini le débit (noté R pour *rate*) qui est une métrique importante. L'autre métrique importante est la probabilité d'erreur.

On note $\lambda_i^{(n)}$ la probabilité d'erreur de décodage du mot code w_i et qui est donnée par :

$$\lambda_i^{(n)} = \sum_{y^n \in \mathcal{Y}^n} P_{Y^n|X^n}(y^n|x_i^n) \cdot \mathbb{1}_{\{\hat{w}_i = f_{dec}(y^n) \neq w_i\}} \tag{3.7}$$

L'erreur maximale est notée $\lambda^{(n)} = \max_i \lambda_i^{(n)}$.

La probabilité d'erreur est égale à la moyenne des erreurs. Si W suit une loi uniforme, alors :

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i^{(n)}. \tag{3.8}$$

Définition 3.6 : Code réalisable,

Un code (M, n, ϵ) est réalisable, si il existe un technique de transmission qui permet de transmettre $\log(M)$ bits en n *channel uses* et telle que $P_e^{(n)} \leq \epsilon$. Rappelons que le débit associé est égal à $R = \log(M)/n$. Nous avons donc clairement à étudier un compromis entre 3 critères : latence (au travers de n), fiabilité (avec P_e) et débit (avec R) .

La figure 3.3 représente la modélisation d'un système de transmission dans un canal sans mémoire, avec les différents critères à optimiser.

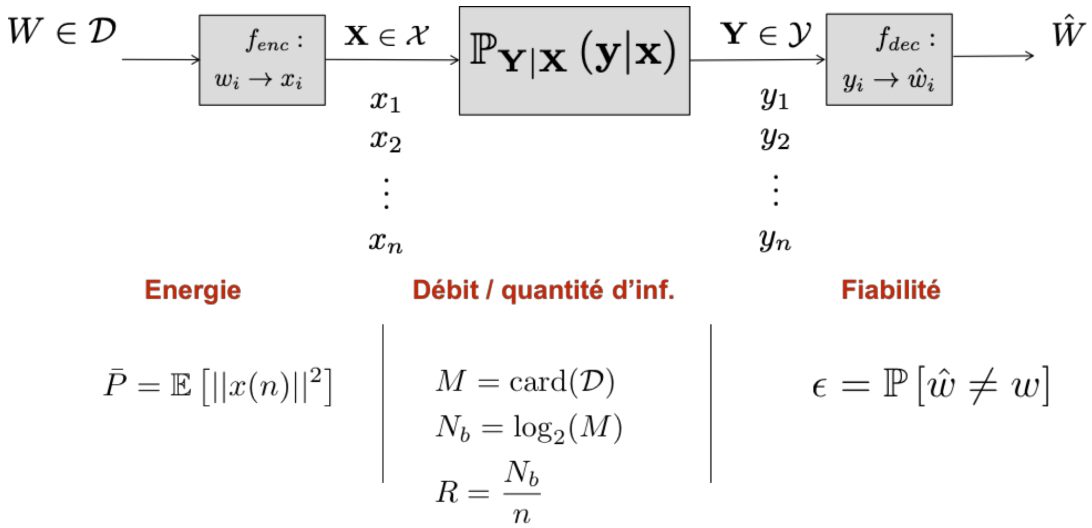


FIGURE 3.3 – Modélisation du canal DMC et des techniques de codage associées.

3.4.2 Détecteur optimal

Maintenant que nous avons décrit la formulation mathématique proposée par Claude Shannon [10] pour analyser les performances des systèmes de communication, nous allons l'analyser en partie avec les outils d'estimation/détection vus dans le chapitre précédent.

Fixons arbitrairement les fonctions d'encodage $w_i \to x_i^n$, nous avons donc défini l'espace des évènements (au sens utilisé en théorie de la détection), c'est à dire que l'on peut identifier $\boldsymbol{\mathcal{E}} \equiv \boldsymbol{\mathcal{W}} \equiv \boldsymbol{\mathcal{X}}^n$.

Pour maximiser les performances de notre système, il nous faut déterminer le détecteur optimal, c'est à dire définir la bonne décision en réception pour toute observation appartenant à $\boldsymbol{\mathcal{Y}}^n$.

Si :

— toutes les erreurs sont supposées avoir le même coût (i.e. estimer $C_{ij} = cste, \forall(i,j); i \neq j$) et,

— la source est de distribution uniforme,

alors nous savons d'après les résultats de théorie de la détection que le choix optimal pour minimiser l'erreur moyenne est l'estimateur ML, autrement dit :

$$\hat{w} = \arg \max_{i \in \{1,2,\ldots,M\}} P_{Y|X}\left(y|x_i\right). \tag{3.9}$$

Bien que très simple à écrire, cet estimateur pose quelques difficultés dans le cas général : calculer analytiquement la probabilité d'erreur associée est mathématiquement très complexe, et mettre en oeuvre ce détecteur directement nécessiterait un nombre d'opérations exponentiellement élevé (il faudrait tester la vraisemblance des 2^n mot-codes).

Le cas du canal AWGN, $y = x_i + n$, est un peu plus gérable. car l'on peut écrire $w_i = \arg\min_{i \in \{1,2,\ldots,M\}} ||y - x_i||^2$. Il faut alors rechercher le mot code le plus proche de l'observation.

Cependant, optimiser un système de transmission nécessite également de choisir les fonctions d'encodage permettant de maximiser la capacité, et de minimiser les erreurs de transmission. C'est à cette question que Claude Shannon tenta de répondre en jetant les bases de la théorie de l'information.

Et nous allons esquisser dans la suite les bases de la preuve qui a conduit à l'établissement de cette fameuse borne.

3.4.3 Enoncé du théorème de capacité

Maintenant que le problème est posé nous pouvons énoncé le fameux théorème de Shannon pour un canal sans mémoire. Nous n'aborderons pas ici le problème de façon plus général, en particulier pour des canaux à mémoire. Vous pouvez consulter les travaux de S. Verdù dans [11], qui étudie cette même capacité pour un canal quelconque avec ou sans mémoire, et non stationnaire. Ceci sort très largement du cadre de ce cours.

La première définition que C. Shannon a appelé capacité d'information est basée sur l'information mutuelle.

Définition 3.7 : Capacité d'information

La capacité d'information d'un canal discret sans mémoire est définie par :

$$C := \max_{P_X} I(X;Y).$$

Autrement dit, la capacité d'information est égale à la plus grande valeur de l'information mutuelle que l'on peut obtenir en choisissant P_X. Attention cependant, l'obtention de ce maximum peut s'avérer parfois difficile pour des canaux un peu exotiques !

Cette définition est essentiellement mathématique et ne revêt pas à ce stade de sens physique exploitable pour la transmission.

La deuxième définition s'appuie sur les (M, n, ϵ)-codes définis ci-dessus.

Définition 3.8 : Débit réalisable

Si il existe une série de codes $(M(n), n, \epsilon(n))$ telle que $M(n) \geq 2^{nR}$, et tel que $\epsilon(n) \to \infty$ quand $n \to \infty$, alors le débit R est dit réalisable.

Pour bien comprendre cette définition, il faut analyser tous les termes dans le détail. L'idée est que comme il est difficile de trouver le meilleur code, pour une valeur de n finie, nous allons regarder le comportement du codage quand $n \to \infty$. Lorsque n croit, il faut que la taille du dictionnaire, $M(n)$, croisse exponentiellement pour préserver le débit. Le but est alors de montrer que si l'on effectue notre codage sur des très grands codes, l'erreur doit tendre vers 0 et la transmission devient sans erreur.

Il faut bien comprendre que dans cette approche, nous abandonnons toute contrainte de latence et nous privilégions l'objectif de débit et la contrainte de fiabilité.

Définition 3.9 : Capacité opérationnelle

La capacité opérationnelle d'un canal est définie comme le maximum de tous les débits réalisables.

Alors que la première définition 3.7 est purement mathématique, cette deuxième définition est purement physique. Il s'agit du plus grand débit que l'on peut obtenir dans un canal, avec une erreur qui tend vers 0 et une latence infinie.

Il ne nous reste plus qu'à établir que la capacité opérationnelle et la capacité d'information sont égales, ce qui n'est n'est pas trivial.

L'approche de la théorie de l'information consiste à aborder ce problème en deux étapes. On va d'abord démontrer que pour un canal donné $P_{Y|X}$, un débit réalisable ne peut être supérieur à la capacité d'information $I(X;Y)$. Ce qui nous fournira une borne supérieure. Il restera alors à démontrer qu'il existe une méthode de codage (même un peu artificielle) qui permet d'atteindre $I(X;Y)$. Dès lors, la démonstration est complète.

3.4.4 Converse - Théorème de Fano

Pour établir cette borne supérieure, nous allons uniquement nous appuyer sur les propriétés de l'entropie et de l'information mutuelle, et sur une inégalité importante que l'on appelle l'inégalité de Fano du nom de son auteur (il a fait cette démonstration en 1942, qui lui a valu plus tard le Claude Shannon award).

Voilà les grandes lignes de ce résultat qui démontre que la capacité d'information joue le rôle d'une limite stricte au débit d'information dans un canal.

Commençons par établir une première relation entre le débit R et l'information mutuelle.

$$\begin{aligned} nR &\overset{(i)}{=} H(W) \\ &\overset{(ii)}{=} H(W|\hat{W}) + I(W;\hat{W}) \\ &\overset{(iii)}{\leq} H(W|\hat{W}) + I(X^n;Y^n) \end{aligned} \tag{3.10}$$

où (i) est établi par définition de l'entropie de la source (équiprobable), qui est égale au nombre de bits codés, soit nR. (ii) est établi par propriété de l'information mutuelle (prop.3.1). Enfin, (iii) découle de la propriété de la chaîne de Markov en considérant $W \to X^n \to Y^n \to \hat{W}$.

Nous allons étudier successivement les deux termes restant.

1. Entropie de l'erreur. L'inégalité de Fano stipule que

$$H(W|\hat{W}) \leq 1 + P_e^{(n)} \cdot nR \tag{3.11}$$

Pour ne rien laisser au hasard, prouvons cette inégalité. Nous introduisons tout d'abord une variable aléatoire E, binaire, égale à 1 si il y a erreur de transmission et à 0 sinon.

Nous pouvons écrire que :

$$\begin{aligned} H(W|\hat{W}) &\overset{(i)}{=} H(E,W|\hat{W}) - H(E|W,\hat{W}) \\ &\overset{(ii)}{=} H(E,W|\hat{W}) - 0 \\ &\overset{(iii)}{=} \underbrace{H(E|\hat{W})}_{\leq 1} + \underbrace{H(W|E,\hat{W})}_{\leq P_e \log|\mathcal{W}|} \end{aligned} \tag{3.12}$$

Nous obtenons (i) en utilisant (3.5).
(ii) est obtenu car $H(E|W,\hat{W}) = 0$. En effet, si l'on connait W et $\hat{W}$, on connait le mot émis et le mot reçu. On sait donc si il y a erreur ou pas et E est alors certain.
Enfin, (iii) est obtenu en utilisant (3.5) encore une fois, mais en permutant les variables.

Dans la dernière expression, pour le premier terme, on a $H(E|\hat{W}) \leq H(E)$ car le conditionnement réduit toujours l'entropie. On peut noter, par définition que $H(E) = h(P_e)$ qui est l'entropie d'une variable de Bernouilli. Elle est au maximum égale à 1. Le deuxième terme peut être développé en le conditionnant par les différents états possible de l'erreur :

$$H(W|E,\hat{W}) = H(W|E=0,\hat{W}) \cdot P_E(0) + H(W|E=1,\hat{W}) \cdot P_E(1). \tag{3.13}$$

Or, $H(W|E=0,\hat{W})$ est nul, car si $E=0$, il n'y a pas d'erreur, et donc l'équivoque est nulle. D'autre part, $P_E(1) = P_e$, par définition, et $H(W|E=1,\hat{W}) = \log|\mathcal{W}| = nR$, ce qui termine la preuve de l'inégalité de Fano. Notons que ce théorème est très général. Il permet d'établir une relation explicite entre la probabilité d'erreur et l'équivoque, qui est d'ordre mathématique.

2. Information mutuelle. Nous utilisons ici aussi les propriétés classiques de l'information

mutuelle :

$$\begin{aligned} I(X^n;Y^n) &\overset{i}{=} H(Y^n) - H(Y^n|X^n) \\ &\overset{ii}{=} H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\dots Y_{i-1}X^n) \\ &\overset{iii}{=} H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\overset{iv}{\leq} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \end{aligned} \tag{3.14}$$

où (i) provient encore une fois de Prop.3.1. (ii) découle de la règle de chainage (prop.3.2), et (iii) du fait que le canal est sans mémoire et du fait que $(Y_1,\dots,Y_{i-1}) \to X^n \to Y_i$ forme une chaine de Markov. La dernière inégalité provient du fait que l'entropie conjointe est inférieure ou égale à la somme des entropies individuelles.

Pour finir, on peut regrouper les deux sommes, et on obtient :

$$I(X^n;Y^n) \leq nI(X;Y) \leq nC. \tag{3.15}$$

ce qui conclue la preuve.

En effet, il découle alors de (3.10) que :

$$nR \leq 1 + P_e^{(n)} nR + nC. \tag{3.16}$$

Soit, en divisant par n et en prenant $n \to \infty$,

$$R \leq C, \tag{3.17}$$

car $P_e^{(n)} \to 0$ quand $n \to \infty$, par définition du débit atteignable (def.3.8). Ce qui prouve que pour être atteignable, un débit doit être inférieur à la capacité d'information C. La capacité opérationnelle est donc inférieure ou égale à la capacité d'information.

On a donc prouvé que la capacité d'information est une borne supérieure des débits atteignables.

3.4.5 Achievability - Typicalité

Le travail d'étude de la borne n'est pas entièrement achevé, car on ne sait pas si cette borne est atteignable ou si elle surestime très largement tout débit réalisable. Pour pouvoir dire que la capacité opérationnelle est égale à la capacité d'information, il faut maintenant prouver que C est atteignable. C'est à dire qu'il existe un code $(2^{nC}, n, \epsilon)$ quelque soit $\epsilon > 0$ et pour n suffisamment grand .

La preuve utilisée par C. Shannon repose sur plusieurs éléments :

— Les propriétés des séquences typiques.

— L'entropie encore une fois.

— La notion de codage aléatoire (random codes).

Sur la base de ces éléments, il démontre qu'un codage aléatoire permet d'atteindre les bornes. Pour plus de détails je vous invite à lire le chapitre 7 de [1], qui permet d'appréhender complètement cette preuve qui a joué un rôle essentiel dans le design des communications numériques modernes.

3.5 Capacité du canal Gaussien

Dans la section précédente nous avons donné les grandes lignes de la démonstration de C. Shannon pour la capacité des canaux discrets sans mémoire.

Nous souhaitons maintenant élargir ce résultat au cas du canal AWGN, au plus près du canal physique c'est à dire à partir des symboles émis et reçus en bande de base, tel que représenté à droite de la figure 3.1.

Le signal transmis échantillonné est construit à partir d'une séquences de symboles $\underline{x}_e(1)\ldots,\underline{x}_e(N)$, à valeurs dans $\mathbb{C}$. Cependant, chaque symbole complexe peut être vu comme la composition de deux symboles réels $x_I(k)$ et $x_Q(k)$. Ainsi, la transmission de N échantillons complexes, est équivalent à la transmission $2N$ échantillons réels. Si la démodulation radio-fréquences (RF) est parfaite, et que la phase est corrigée, alors un canal complexe échantillonné sans mémoire est équivalent à un canal réel échantillonné sans mémoire, avec 2 *channel uses* par échantillon. On parlera par la suite de n utilisations de canal, pour garder les notations usuelles en théorie de l'information.

Nous considérons donc à partir de là que le canal AWGN est un canal à valeurs réelles, représenté par : $y_e[k] = x_e[k] + n_e[k]$, ou encore décrit par la densité de probabilité conditionnelle :

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(y-x)^2}{2\sigma_n^2}}. \tag{3.18}$$

La question qui nous préoccupe est de savoir quelle est la capacité de ce canal ?

Pour répondre à cette question, nous allons suivre le même raisonnement que pour le canal discret, mais nous avons besoin pour cela de définir ce qu'est l'entropie différentielle et énoncer quelques propriétés. L'entropie $H(X)$ a été définie pour une variable X à valeurs discrètes. Nous étendons cette définion à une variable X à valeurs dans un espace continu.

3.5.1 Entropie différentielle

Définition 3.10 : Entropie différentielle

L'entropie différentielle notée $h(X)$ pour une variable aléatoire continue de densité $f_X(x)$ est définie par :

$$h(X) = -\int_{u\in\mathcal{X}} f_X(u)\cdot \log f_X(u)\cdot du$$

Exercice 3.4 : exercices sur l'entropie différentielle

— A titre d'exemple, nous vous laissons vérifier les propriétés suivantes :

1. Si $f_X(x)$ est une loi uniforme, alors $h(X) = \log(a)$.
2. Si $f_X(x)$ est une loi normale centrée $\mathcal{N}(0, \sigma^2)$, alors $h(X) = \frac{1}{2}\log(2\pi e\sigma^2)$ bits.

— Calculez l'entropie différentielle pour les cas suivants :

1. une variable aléatoire X de densité exponentielle $f_X(x) = \lambda \cdot e^{-\lambda x}, x \geq 0$.
2. une variable aléatoire X de densité de Laplace $f_X(x) = \frac{1}{2}\lambda \cdot e^{-\lambda|x|}, x \geq 0$.

3. une variable aléatoire $Z = X + Y$, où X et Y sont des variables normales, indépendantes suivant $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$, $\mathcal{N}\left(\mu_2, \sigma_2^2\right)$.

Nous retiendrons quelques propriétés importantes de l'entropie différentielle. Tout d'abord, notons que la loi normale joue un rôle essentiel. En effet, nous pouvons énoncer le théorème suivant :

Théorème 3.2 : Entropie différentielle maximale

Soit une variable aléatoire continue X de pdf $f_X : \mathbb{R} \to [0, \infty)$ de variance σ^2. Alors l'entropie différentielle $h(X)$ est maximale si et seulement si f_x suit une loi normale centrée. Dès lors :

$$h(X) \leq \frac{1}{2} \log(2\pi e \sigma^2).$$

Ce théorème nous dit qu' à puissance moyenne fixée (σ^2), la loi qui maximise l'entropie est la loi normale centrée. Nous ne le démontrerons pas ici mais vous pourrez en trouver la démonstration dans le chapitre 8 de [1].

L'entropie différentielle possède les mêmes propriétés que l'entropie. Nous pouvons donc définir l'entropie conditionnelle :

$$h(X|Y) = -\int_{\mathcal{X},\mathcal{Y}} f_{XY}(x,y) \log\left(f_{X|Y}(x|y)\right) dxdy. \tag{3.19}$$

Elle vérifie :

$$h(X|Y) = h(X,Y) - h(Y). \tag{3.20}$$

Enfin, la règle de la chaine (chain rule) s'écrit également :

$$h(X_1, \ldots, X_N) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}). \tag{3.21}$$

Un autre résultat important que nous utiliserons pour les systèmes MIMO est le suivant :

Théorème 3.3 : Entropie d'une loi Normale multi-variables

Soit $X_1 \ldots, X_N$ un ensemble de variables aléatoires normales de moyenne μ et de matrice de covariance K. Alors

$$h(X_1, \ldots, X_N) = \frac{1}{2} \log(2\pi e)^n \det(K).$$

Pour finir, l'information mutuelle se définie comme pour les variables discrètes, par :

Définition 3.11 : Information mutuelle pour les variables continues

L'information mutuelle entre deux variables aléatoires continues$X$ et Y est définie par

$$I(X;Y) = \int_{x\in\mathbb{R}} \int_{y\in\mathbb{R}} f_{X,Y}(x,y) \cdot \log\left[\frac{f_{X,Y}(x,y)}{f_X(x)\cdot P_Y(y)}\right] \cdot dx \cdot dy.$$

Et on peut démontrer également la relation suivante entre information mutuelle et entropie différentielle :

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X). \tag{3.22}$$

3.5.2 Converse

Nous avons maintenant tous les éléments pour étudier la capacité du canal Gaussien sans mémoire.

Il faut tout d'abord considérer une contrainte d'énergie ou de puissance sur nos symboles. Sans contrainte d'énergie, la capacité serait infinie !

Nous imposons donc que chaque mot code $(x_1, \dots x_n)$ vérifie une contrainte de puissance moyenne :

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P \tag{3.23}$$

Dès lors la capacité d'information est donnée par :

Définition 3.12 : Capacité d'information du canal Gaussien

La capacité d'information d'un canal Gaussien sans mémoire est définie par :

$$\begin{aligned} C &:= \max_{f_X;\mathbb{E}[X^2]\leq P} I(X;Y) \\ &= \frac{1}{2}\log\left(1+\frac{P}{N}\right) \end{aligned}$$

La première ligne est la définition de la capacité d'information. La principale différence avec le cas discret réside dans le fait que l'on contraint le domaine de X à respecter la contrainte de puissance en moyenne. La deuxième ligne est une conséquence de cette définition que nous allons prouver.

D'après (3.22), nous pouvons écrire que

$$I(X;Y) = h(Y) - h(Y|X).$$

Et comme $Y = X + Z$, où Z représente la variable aléatoire de bruit, on a également :

$$I(X;Y) = h(Y) - h(Z).$$

D'après les propriétés de l'entropie différentielle, $h(Z) = \frac{1}{2}\log(2\pi e\sigma_n^2)$. On note la puissance du bruit $N := \sigma_n^2$. Enfin, d'après le théorème de l'entropie différentielle maximale, $h(Y)$ est maximal si et seulement si Y suit une loi normale. Sa puissance est égale à la somme de la puissance de X et de celle de Z. Ce qui donne :

$$I(X;Y) \leq \frac{1}{2}\log(2\pi e(P+N)) - \frac{1}{2}\log(2\pi eN)$$
$$\frac{1}{2}\log\left(1+\frac{P}{N}\right)$$

Il reste à montrer que la capacité opérationnelle est bien bornée par cette capacité d'information. Nous suivons les étapes développées dans (3.14) :

$$\begin{aligned} I(X^n;Y^n) &\leq \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Z_i) \\ &\leq \sum_{i=1}^{n} \frac{1}{2}\log\left(1+\frac{P_i}{N}\right) \end{aligned} \tag{3.24}$$

où P_i est la puissance moyenne de l'échantillon i. Pour arriver au résultat souhaité il faut utiliser le théorème de Jensen :

Théorème 3.4 : Inégalité de Jensen

Soit $f(x)$ une fonction concave. Alors, la fonction vérifie :

$$\frac{\sum_{k=1}^{n} a_k f(x_k)}{\sum_{k=1}^{n} a_k} \leq f\left(\frac{\sum_{k=1}^{n} a_k x_k}{\sum_{k=1}^{n} a_k}\right).$$

Comme la fonction $\frac{1}{2}\log\left(1+\frac{P_i}{N}\right)$ est une fonction strictement concave, on peut poursuivre le développement (en appliquant l'inégalité de Jensen avec $a_k = 1/n$) :

$$I(X^n;Y^n) \leq \frac{n}{2}\log\left(1+\frac{P}{N}\right), \tag{3.25}$$

avec $P = \frac{1}{n}\sum_i P_i$.

Ce qui permet de terminer la preuve du converse. De même que pour le canal discret, nous ne développerons pas ici la preuve d'atteignabilité de cette capacité, qui repose sur les mêmes arguments que dans le cas discret, et en particulier sur la notion de séquences typiques.

3.6 Ouverture : théorie de l'information en réseaux

Ces résultats de capacité ont été établi dès la fin des années 50 et ont donc guidé les travaux de développement des techniques de codage.

Grâce au travail de Shannon, il est possible de confronter un code particulier à ces bornes théoriques. Cependant, l'extension de cette approche au contexte des réesaux multi-sauts est particulièrement complexe. Il s'agit d'arriver à exprimer de façon mathématique la capacité théorique d'un ensemble de paires émetteurs-récepteurs partageant un environnement radio commun.

Le problème est relativement bien résolu pour les réseaux filaires, mais le contexte des réseaux radio offre une complexité extrêmement grande. On trouve dans la littérature des résultats de capacité exacts pour des modèles de petite taille (canal à relais constitué d'un émetteur, d'un récepteur et d'un relais), ou des bornes approchées dans des grands réseaux denses. Mais à ce jour la capacité théorique de réseaux radio n'est pas établie exactement.

3.7 Ouverture : régime non asymptotique

Les résultats de capacité de Shannon sont dits asymptotiques, car il s'intéresse au débit atteignable quand $n \to \infty$. Ces résultats se sont révélés suffisants dans l'épopée récente des communications, car l'objectif poursuivi par l'industrie était d'atteindre des très haut débits pour des flux constant de type voix ou vidéo.

Cependant, le contexte de l'Internet des objets apporte de nouvelles perspectives avec l'objectif de transmettre des petits paquets (petites quantités d'information) de façon dispersée dans le réseau. Etablir les bornes théoriques de performance dans ce contexte se révèle être une tâche très ardue à laquelle une partie de la communauté scientifique se consacre actuellement, dans l'espoir de découvrir de nouveaux protocoles ou codage capables de dépasser les performances des protocoles actuels.

3.8 Synthèse du chapitre

Nous avons introduit dans ce chapitre les notions fondamentales à la base de la théorie de l'information et en particulier du théorème de Shannon sur la capacité de canal.

Ces outils mathématiques solides sont la base de nombreux travaux essentiels en ingénierie. Ils permettent de poser de façon rigoureuse de nombreux problèmes auquel peut être confronté un ingénieur. Nous n'avons fait qu'aborder les bases de ces théories, qui restent très largement exploitées dans de nombreux travaux de recherche des plus actuels.

Les connaissances développées dans ce chapitre sont les suivantes :

1. Principes de la théorie de l'information :
   - Rappels sur l'entropie et l'information mutuelle.
   - Capacité des canaux DMC.
   - Capacité du canal GMC.

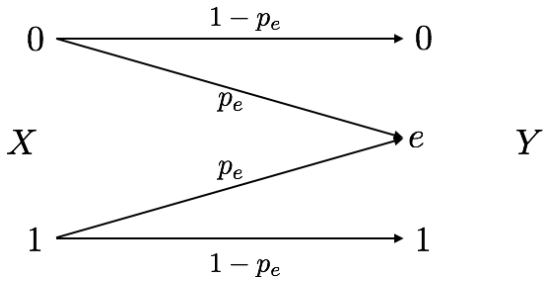
Enfin, au travers des développements proposés et des exemples discutés nous avons développé les savoir-faire suivants :

1. Evaluer les performances d'un système en termes de capacité ou de probabilité d'erreur.
2. Utiliser la théorie de la détection pour définir un récepteur optimal.

3.9 Exercices

Exercice 3.5 : Canal à effacement

Un canal binaire à effacement sans mémoire, est une extension du canal binaire sans mémoire, où nous avons un troisième état en sortie, tel que représenté à la figure suivante



Capacité du canal à effacement

1. Déterminez une borne supérieure de la capacité de ce canal dans le cas où il est symétrique. Vous pourrez utiliser l'hypothèse idéale que la source sait quand le canal produit un effacement.
2. Etendez l'étude au cas asymétrique.

Canal gaussien avec BPSK On considère maintenant un canal gaussien, utilisé avec une modulation BPSK.

1. Développez le détecteur optimal en réception lorsque l'on travaille indépendamment sur chaque symbole.
2. Calculez la probabilité d'erreur du canal binaire symétrique équivalent.
3. Calculez la capacité de ce canal et comparez à la capacité du canal gaussien.
4. Tracez les courbes en fonction du SNR.

Canal gaussien vs canal à effacement Le démodulateur retourne maintenant 3 états : 0, 1, ou incertain.

1. Montrez que ce système peut être modélisé pars la succession d'un canal binaire symétrique et d'un canal binaire à effacement.
2. Calculez la capacité du système. Vous utiliserez les propriétés de l'information mutuelle.
3. Tracer la courbe de la capacité en fonction du SNR.
4. Comparez aux courbes précédentes.

Exercice 3.6 : Perte de capacité du canal Gaussien

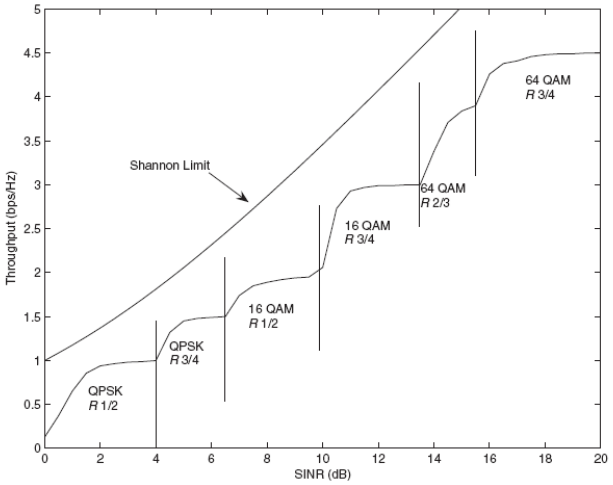
Soit un système de transmission dans un canal Gaussien, réel, décrit par :

$$y_e[k] = x_e[k] + n_e[k]. \tag{3.26}$$

1. Quelle est la capacité de ce canal ?
2. Tracez cette capacité en fonction du signal to noise ratio, rapport signal à bruit (SNR) P/N.

On souhaite utiliser pour ce canal une modulation QPSK, 16-QAM ou 32-QAM.

3. Tracez les capacités théoriques de ces 3 canaux, en fonction du SNR sur la même figure que celle de la capacité du canal gaussien.
4. Concluez.
5. Analysez et justifiez les courbes de la figure ci-jointe.



Bibliographie

[1] Thomas M Cover et Joy A Thomas : *Elements of information theory*. John Wiley & Sons, 2012.

[2] David Declercq et André Quinquis : *Détection et estimation des signaux*. Hermès, 1996.

[3] Izrail Solomonovich Gradshteyn et Iosif Moiseevich Ryzhik : *Table of integrals, series, and products*. Academic press, 2014.

[4] Michel Guglielmi *et al.* : Signaux aléatoires : modélisation, estimation, détection. *Traitement du signal et de limage. Hermes*, 2004.

[5] Bernard C Levy : *Principles of signal detection and parameter estimation*. Springer Science & Business Media, 2008.

[6] Jerzy Neyman et Egon Sharpe Pearson : Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706) :289–337, 1933.

[7] Timothy OShea et Jakob Hoydis : An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, 3(4) :563–575, 2017.

[8] Bernard Picinbono : Signaux aléatoires. *Techniques de l'ingénieur. Informatique industrielle*, 1(R7030) :R7030–1, 1993.

[9] H Vincent Poor : *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013.

[10] Claude E. Shannon : A mathematical theory of communication, part i, part ii. *Bell Syst. Tech. J.*, 27 :623–656, 1948.

[11] Sergio Verdú *et al.* : A general formula for channel capacity. *IEEE Transactions on Information Theory*, 40(4) :1147–1157, 1994.