Christian JUTTEN

Détection,
Estimation,
Information.

Notions de base et exercices

Univ. Grenoble Alpes - Polytech' Grenoble

Département *Informatique et Electronique des Systèmes Embarqués*

IESE5, option *Images et Signaux* et *Automatique*

Juillet 2018

Table des matières

Chapitre 1

Introduction

1.1 Problèmes de détection

Pour illustrer les objectifs des théories de la détection et de l'estimation, considérons l'exemple d'un système de communications numériques (Fig. 1.1) destiné à transmettre un message de la source vers un récepteur. La source émet des mots binaires toutes les T secondes. Le canal de transmission transmet le message. Sa nature, et la façon dont l'information sera transmise, peuvent être très variables : fil pour une transmission électrique (téléphone), air pour des transmissions électromagnétiques ou acoustiques, eau pour des transmissions en acoustique sous-marine, fibres optiques, etc.

Considérons par exemple une transmission radio avec une porteuse sinusoïdale telle que, pendant T secondes :

$$s(t) = \begin{cases} s_0(t) &= \sin(\omega_0 t) \\ s_1(t) &= \sin(\omega_1 t) \end{cases} \tag{1.1}$$

avec une pulsation *connue* ω_0 ou ω_1, selon que la source émet le mot binaire 0 ou 1 (Fig. 1.2).

1.1.1 Détection d'un signal connu

Dans le meilleur des cas, le signal, après transmission dans le canal, arrive avec atténuation mais sans distorsion au niveau du récepteur. Il suffit d'un amplificateur pour restituer l'amplitude initiale du signal. Cependant, le canal et l'amplificateur introduisent du bruit, si bien que le signal reçu, $r(t)$ s'écrit :

$$r(t) = \begin{cases} \sin(\omega_0 t) &+ n(t) \\ \sin(\omega_1 t) &+ n(t). \end{cases} \tag{1.2}$$

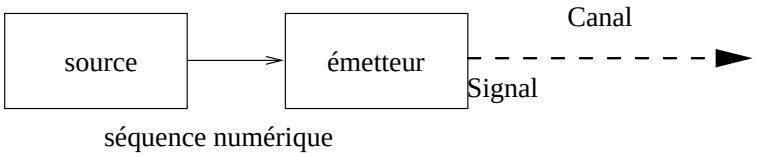


FIGURE 1.1 – Schéma d'un système de communication numérique.

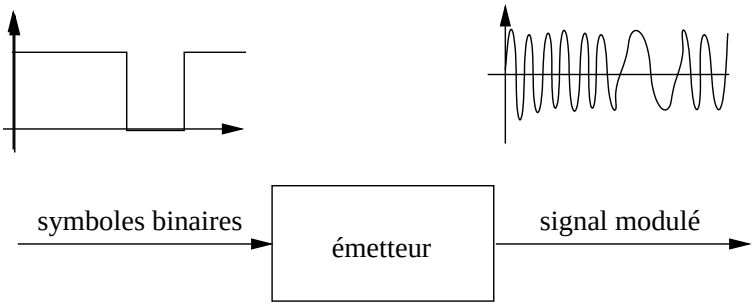


FIGURE 1.2 – Schéma d'un système avec modulation binaire.

Le problème consiste donc à décider, en observant le signal $r(t)$, lequel des deux symboles 0 ou 1 a été émis. C'est donc une problème simple de détection d'un signal connu dans du bruit.

1.1.2 Détection d'un signal inconnu

Supposons maintenant que les oscillateurs délivrant les signaux $s_0(t)$ et $s_1(t)$ aient un glissement de phase, alors pendant T secondes on observe :

$$r(t) = \begin{cases} \sin(\omega_0 t + \theta_0) & + \quad n(t) \\ \sin(\omega_1 t + \theta_1) & + \quad n(t), \end{cases} \tag{1.3}$$

où θ_0 et θ_1 sont des angles *inconnus et constants* [1]. Même en absence de bruit, le signal reçu n'est pas parfaitement connu. Le problème est alors plus complexe que dans le paragraphe précédent : il s'agit de détecter un signal inconnu (ou à paramètres inconnus) dans du bruit.

Exemple : écho radar. Le radar émet pendant T secondes, avec une certaine période de répétition T_r, un signal de pulsation ω_0, c'est-à-dire :

$$s(t) = \begin{cases} \sin(\omega_0 t), & \text{si } t \in [kT_r, kT_r + T[\\ 0, & \text{sinon.} \end{cases} \tag{1.4}$$

Si une cible immobile [2] est présente, on observe, en réponse à $s(t)$ un signal $r(t)$:

$$r(t) = \begin{cases} A_r \sin(\omega_0(t-\tau) + \theta_r) + n(t), & \text{si } t \in [kT_r + \tau, kT_r + T + \tau[\\ n(t), & \text{sinon.} \end{cases} \tag{1.5}$$

où A_r est l'atténuation qui dépend de la distance de la cible, de sa taille et de son coefficient de réflexion, τ est égal au temps aller-retour entre la cible et la source, θ_r est le glissement de phase en réception, et $n(t)$ est un bruit. On voit qu'en absence de cible, on observe toujours $n(t)$. Même s'il n'y a pas de bruit, l'écho est un signal inconnu, puisqu'il comporte trois paramètres inconnus : l'atténuation A_r, le retard τ et le glissement de phase θ_r.

1. Généralement, les glissements de phase varient lentement dans le temps
2. Il n'y a donc pas d'effet Doppler

1.1.3 Détection d'un signal aléatoire

Quittons l'exemple des communications numériques, et considérons maintenant un signal sonar dans lequel on cherche à détecter le signal émis par un navire parmi les signaux émis par d'autres bâtiments et les signaux d'origine biologique émis par les crevettes, les dauphins, les poissons, etc. Le signal émis par le navire est inconnu, il est complexe car dû à des causes multiples : signaux des moteurs et de transmission, signaux des hélices et de cavitation, bruits produits par l'équipage et transmis par la coque, etc. On le modélise donc par un signal aléatoire $s_a(t)$, et toutes les perturbations sont résumées par le bruit $n(t)$. On observe donc :

$$r(t) = \begin{cases} s_a(t) + n(t), & \text{si le navire est présent,} \\ n(t), & \text{sinon.} \end{cases} \tag{1.6}$$

Cette situation se retrouve dans le traitement de signaux sismiques ou radio-astronomiques, mais aussi pour des communications numériques lorsque les caractéristiques sont inconnues. Ce problème est encore plus compliqué, puisqu'il s'agit de détecter un *signal aléatoire dans du bruit*, c'est-à-dire dans un autre signal aléatoire.

1.1.4 Quelques exemples

La détection est une étape essentielle dans de nombreux problèmes. Nous avons évoqué des problèmes de communications numériques où il s'agit de détecter un signal parmi deux possibles : il s'agit de détection binaire. Les communications numériques peuvent utiliser des codes plus compliqués, si bien qu'à chaque instant un signal parmi 4, 16, voire 256 peut être attendu : on parle alors de détection M-aire, avec $M = 4$, $M = 16$ ou $M = 256$.

Les problèmes de classification sont également des problèmes de détection généralement M-aire. En reconnaissance optique de caractères, il s'agit de détecter un caractère alphanumétique parmi 26 lettres (en distinguant éventuellement les majuscules), plus 10 chiffres et les signes de ponctuation. Dans le cas le plus simple, on connait le type de caractères (par exemple, la fonte *Time Roman* $11pt$) : on connaît donc parfaitement les "signaux" possibles. Dans le cas où la taille de la fonte est inconnue, la détection doit être invariante ; c'est aussi le cas pour les classifieurs multi-fontes. En reconnaissance de la parole (ou de locuteur), il s'agit de détecter un mot (ou un locuteur) à partir d'un signal de parole.

Dans la plupart des cas, le signal est multi-dimensionnel. Dans le cas de signaux de communication, on observe le signal sur un temps T, ce qui permet de mesurer N échantillons. Pour un signal de parole de 100 ms, échantillonné entre 10 et 20 kHz, on mesure entre 1000 et 2000 échantillons. Il est illusoire de travailler dans un espace qui comporte autant de dimensions [3]. La première étape consiste à extraire un nombre réduit des caractéristiques pertinentes (*features* en anglais) qui seront discriminantes. De la pertinence de ces caractéristiques dépendront les performances du classifieur. La même étape de codage qui permet de réduire la dimension de l'observation de façon judicieuse est primordiale en reconnaissance de caractères, analyse de scènes, etc.

Certains détecteurs sont déjà tout à fait usuels : c'est le cas des logiciels de lecture optique associés à n'importe quel scanner. D'autres sont encore à concevoir : par exemple, l'archivage et

3. En raison de la *malédiction de la dimensionalité*, les données sont généralement trop éparses dans un espace de dimension supérieure à 5 pour que les estimations aient un sens

la recherche de données complexes (images, sons, séquences audio-vidéo, etc.) dans des bases de données de très grandes tailles (par exemple, sur Internet) suppose de savoir indexer automatiquement des données pour les ranger et les retrouver efficacement : l'indexation est le calcul de caractéristiques pertinentes de ces données. On voudrait aboutir à des systèmes capables de rechercher (et vite) des données à partir d'une requête du style : *je veux les images de coucher de soleil en montagne* ou bien une recherche par le contenu à partir d'une image de coucher de soleil en montagne, c'est-à-dire : *je veux les images qui ressemblent à celle-ci*.

Pour résumer, nous avons rassemblé ces exemples par niveau de complexité dans le tableau ci-dessous.

Résumé des problèmes de classification

Complexité	**Types de problèmes**
Niveau 1 *Détection d'un signal connu dans du bruit*	Communications numériques synchrones Reconnaissances de formes (connues)
Niveau 2 *Détection d'un signal inconnu dans du bruit*	Détection d'une cible en radar ou sonal actifs Classification invariante Communication sans référence de phase Reconnaissance de la parole pour un locuteur connu
Niveau 3 *Détection d'un signal aléatoire dans du bruit*	Communications numériques Reconnaissance multi-locuteurs Radar ou sonar passifs, signaux sismiques, Radio-astronomie

1.2 Problèmes d'estimation

On peut également considérer plusieurs niveaux de complexité, similaires, dans les problèmes d'estimation. En détection, il s'agit de décider d'une hypothèse parmi un nombre fini M d'hypothèses possibles. En estimation, il s'agit d'évaluer la valeur d'une quantité continue (au moins par intervalles) et qui peut prendre une infinité de valeurs.

1.2.1 Estimation d'un signal connu

Considérons de nouveau un système de communications, dans lequel la source émet un signal analogique, $A(t)$, échantillonné à T_e. A chaque instant kT_e, on mesure l'amplitude $A(kT_e) = A_k$ qui est transmise vers le récepteur et perturbée par les bruits du canal, de l'émetteur et du récepteur. Ainsi, avec un émetteur utilisant une modulation d'amplitude, le signal émis est de la forme :

$$s(t, A_k) = A_k \sin(\omega_c t),\ t \in [kT_e, (k+1)T_e[. \tag{1.7}$$

Dans le cas d'une modulation de fréquence, on aura :

$$s(t, A_k) = \sin(\omega_c t + A_k t),\ t \in [kT_e, (k+1)T_e[. \tag{1.8}$$

Sur le récepteur, en tenant compte des divers bruits, on a :

$$r(t) = s(t, A_k) + n(t),\ t \in [kT_e, (k+1)T_e[. \tag{1.9}$$

Le signal reçu, $r(t)$ dépend donc de l'amplitude A du signal. Si l'on connait parfaitement la modulation : type et paramètre (ω_c), et si l'application entre A et $s(t, A)$ est bijective, c'est un problème *d'estimation (des paramètres) d'un signal connu dans du bruit*.

1.2.2 Estimation d'un signal inconnu

Considérons de nouveau le signal reçu en réponse à une émission radar, pour une cible qui se déplace. Le signal reçu est similaire à l'équation (1.5), mais avec un effet Doppler qui se traduit par une variation de pulsation ω_d :

$$r(t) = \begin{cases} A_r \sin((\omega_0 + \omega_d)(t - \tau) + \theta_r) + n(t), & \text{si } t \in [kT_r + \tau, kT_r + T + \tau[\\ n(t), & \text{sinon.} \end{cases} \tag{1.10}$$

Ici, on sait qu'il y a une cible et on désire évaluer sa distance et sa vitesse (et même éventuellement sa taille et son type) à partir de l'estimation de τ et ω_d. Dans cet exemple, on remarque qu'il existe, en plus des paramètres utiles, des paramètres supplémentaires inconnus, A_r et θ_r, qui vont rendre l'estimation difficile. Le problème est donc *l'estimation d'un signal inconnu dans du bruit*.

1.2.3 Estimation d'un signal aléatoire dans du bruit

Dans ce cas, le récepteur mesure :

$$r(t) = s_a(t, A) + n(t), \tag{1.11}$$

où $s_a(t, A)$ est la réalisation d'un processus aléatoire. Ce type de situations est fréquent en radar ou sonar passif [4] (estimation de la vitesse d'un avion ou d'un navire), en radio-astronomie (estimation de la vitesse d'un object céleste), etc.

1.2.4 Quelques exemples

L'estimation est une étape essentielle, qui succède généralement à la détection (inutile d'estimer la vitesse d'une cible s'il n'y a pas de cible !). Aux problèmes de détection évoqués ci-dessus, on peut associer des problèmes d'estimation.

Comme en détection, dans la plupart des cas et pour les mêmes raisons, le signal est multidimensionnel. La relation entre l'observation et les grandeurs que l'on cherche à estimer est aussi essentielle.

Certains systèmes d'estimation sont tout à fait usuels : c'est le cas des systèmes radar qui sont capables d'estimer avec précision la position et la vitesse des avions.

4. c'est-à-dire sans émission d'un signal

Pour résumer, nous avons rassemblé ces exemples par niveau de complexité dans le tableau ci-dessous.

Résumé des problèmes d'estimation	
Complexité	**Types de problèmes**
Niveau 1 *Estimation d'un signal connu dans du bruit*	Systèmes de communications à modulation d'amplitude, de fréquence, etc. connues
Niveau 2 *Estimation d'un signal inconnu dans du bruit*	Vitesse, distance en radar ou sonal actifs Communication analogique avec modulation d'amplitude, de fréquence, etc. inconnues.
Niveau 3 *Estimation d'un signal aléatoire dans du bruit*	Estimation des paramètres d'un spectre Vitesse en radio-astronomie Paramètres d'une cible en radar ou sonar passifs ou en sismique.

1.3 Approche

Il est évident, comme nous l'avons déjà souligné, que dans les problèmes de détection comme ceux d'estimation, on trouve des aspects aléatoires, à différents niveaux de complexité. De façon naturelle, ceci conduit à une approche statistique.

Par ailleurs, nous pourrons considérer un cadre structuré ou non. Dans le cadre structuré, la solution est recherchée dans le cadre d'un modèle paramétrique. Dans la suite, on considèrera deux exemples, le premier dans lequel le modèle est un système linéaire invariant, le second dans lequel le modèle est non linéaire.

1.3.1 Système linéaire invariant

L'entrée $r(t)$ d'un système linéaire invariant, de réponse impulsionnelle $h(\tau)$, vaut :

$$r(t) = \begin{cases} s(t) + n(t), & \text{si } 0 \leq t \leq T \\ 0, & \text{sinon.} \end{cases} \tag{1.12}$$

Le signal déterministe $s(t)$ a une énergie connue :

$$E_s = \int_0^T s^2(t)dt, \tag{1.13}$$

et le bruit $n(t)$ est supposé centré, blanc et de covariance :

$$\Gamma_n(t,u) = E[n(t)n(t-u)] = N_0\delta(u). \tag{1.14}$$

En absence de bruit, la sortie du système à l'instant T serait :

$$s_o(T) = \int_0^T h(\tau)s(T-\tau)d\tau. \tag{1.15}$$

De même, la réponse au bruit seul serait :

$$n_o(T) = \int_0^T h(\tau)n(T-\tau)d\tau. \tag{1.16}$$

Comme critère de qualité du système, on peut définir le rapport signal à bruit en sortie :

$$\begin{aligned}
\frac{S}{N} &= \frac{s_o^2(T)}{E[n_o^2(T)]} \\
&= \frac{\left[\int_0^T h(\tau)s(T-\tau)d\tau\right]^2}{E\left[\int_0^T\int_0^T h(\tau)n(T-\tau)h(u)n(T-u)d\tau du\right]} \\
&= \frac{\left[\int_0^T h(\tau)s(T-\tau)d\tau\right]^2}{\int_0^T\int_0^T h(\tau)h(u)E\left[n(T-\tau)n(T-u)\right]d\tau du} \\
&= \frac{\left[\int_0^T h(\tau)s(T-\tau)d\tau\right]^2}{\int_0^T\int_0^T h(\tau)h(u)N_0\delta(u-\tau)d\tau du} \\
&= \frac{\left[\int_0^T h(\tau)s(T-\tau)d\tau\right]^2}{N_0\int_0^T h^2(\tau)d\tau}.
\end{aligned} \tag{1.17}$$

On peut alors choisir le système $h(\tau)$ qui maximise le rapport signal à bruit.

Cet exemple montre que la solution du problème repose sur trois ingrédients :
— une structure : ici, nous avons choisi un système linéaire invariant,
— un critère : ici, c'est le rapport signal à bruit,
— des informations *a priori* : sur le signal (pour calculer $s_o^2(T)$) et sur le bruit (covariance).

Des informations supplémentaires sur le signal, par exemple sa densité de probabilité, ne seraient d'aucune utilité. Réciproquement, si on avait moins d'informations, on ne pourrait plus de résoudre le problème.

De plus, si nous changions le critère, les informations requises seraient différentes, ainsi peut-être que la solution.

Ainsi, les trois ingrédients : structure, critère et information *a priori*, sont fortement liées. Mentionnons également dans la mise en œuvre pratique un quatrième ingrédient : l'algorithme d'optimisation du critère.

1.3.2 Système non linéaire

La structure n'est pas limitée à un modèle linéaire. Considérons par exemple un système non linéaire sans mémoire (la sortie à l'instant t ne dépend que de l'entrée à l'instant t) dont l'entrée vaut $r(t) = s(t) + n(t)$, où $s(t)$ est un signal aléatoire de densité de probabilité connue $p_s(u)$ et $n(t)$ est un bruit de densité connue $p_n(v)$. Le système est un dispositif quadratique dont la sortie $y(t)$ s'exprime par :

$$y(t) = a_0 + a_1 r(t) + a_2 r^2(t). \tag{1.18}$$

On détermine le système, c'est-à-dire ses coefficients a_0, a_1 et a_2, en minimisant l'erreur quadratique moyenne :

$$\begin{aligned} e &= E[(y(t) - s(t))^2], \\ &= E[(a_0 + a_1 r(t) + a_2 r^2(t) - s(t))^2]. \end{aligned} \tag{1.19}$$

Cet exemple propose une autre approche structurée et montre bien les liens entre la structure, le critère et les informations. En effet,

— un système linéaire invariant ou un modèle paramétrique non linéaire se formalisent avec des équations différentes,

— minimiser l'erreur quadratique moyenne (1.19) ne requiert pas du tout les mêmes informations que maximiser le rapport signal à bruit.

1.3.3 Approche structurée ou non

Dans le cas non structurée, on n'impose pas de structure au système recherché, seulement un critère. L'avantage de cette méthode est que, si nous savons trouver une solution, ce sera la meilleure solution vis-à-vis de ce critère. La difficulté est que, en raison de l'absence de structure, on doit disposer d'informations très complètes sur le signal et le bruit. Au contraire, les approches structurées nécessitent moins d'informations sur le signal et le bruit. En revanche, les performances sont fortement liées à la qualité du modèle.

1.4 Notations

Dans ce document, nous noterons les quantités scalaires par des caractères simples et les vecteurs en caractères gras : par exemple, r correspond à un scalaire alors que $\mathbf{r}$ est un vecteur.

Pour simplifier les notations, nous ne distinguerons pas (en général) une variable aléatoire (scalaire) R ou un vecteur aléatoire $\mathbf{R}$ de leur réalisations r ou $\mathbf{r}$, respectivement : toutes seront notées en minuscules. La densité conditionnelle de r étant donné une quantité (variable, paramètre, hypothèse) a sera notée $p(r/a)$.

Une variable aléatoire s gaussienne, de moyenne m et de variance σ^2 sera notée de façon compacte $s \sim N(m, \sigma^2)$.

Les intégrales, intervenant notamment dans le risque et dans les moyennes, qui sont souvent des intégrales multiples, seront notées avec un seul signe $\int$; l'élément différentiel sera scalaire, par exemple dr, pour une intégrale simple, ou vectoriel, par exemple $\mathbf{dr}$, pour une intégrale multiple.

1.5 Références

De très nombreux ouvrages ont été écrits sur ces théories et les outils nécessaires à leur compréhension. Quelques références disponibles à la bibliothèque universitaire et à celle de Polytech' sont proposées ci-dessous.

Rappels de probabilités et de statistiques

H. Stark, *Probability, Random Processes, and Estimation Theory for Engineers*, Prentice Hall, 1994

B. Picinbono, *Signaux aléatoires, Tomes 1 et 2*, Dunod université, 1994

A. Papoulis, *Signal Analysis*, McGraw-Hill, 1977

Théories de la détection et de l'estimation

D. Declecq, *Signaux et systèmes en questions (IV) Détection et estimation des signaux*, Hermès, 1996

M. Akay, *Detection and estimation methods for biomedical signals*, California Academic Press, 1996

P. Comon, C. Jutten, *Handbook of Blind source Separation - Independent Component Analysis and Applications*, Elsevier, 2010

R. Deutsch, *Estimation Theory*, Prentice Hall, 1965

M. Gugliemi, *Signaux aléatoires : modélisation, estimation, détection*, Hermès, 2004

J. Hérault, C. Jutten, *Réseaux neuronaux et traitement du signal*, Hermès, 1994

M. Kunt, G. Coray, Granlund G. H., J.-P. Haton, *Reconnaissance de formes et analyse de scènes*, Presses polytechniques et universitaires romandes, CNET-ENST, 2000

A. Quinquis, C.-S. Maroni, *Détection et estimations des signaux : exercices et ptoblèmes corrigés*, Hermès, 1999

L. L. Scharf, *Statistical Signal Processing - Detection, Estimation and Time Series Analysis*, Addison-Wesley, 1991

H. Van Trees, *Detection, Estimation and Modulation Theory*, John Wiley and Sons, 1968 (Tomes 1, 2 et 3)

Théorie de l'information

G. Battail, *Théorie de l'information - Application aux techniques de communication*, Masson, 1997

L. Brillouin, *La science et la théorie de l'information*, Masson, 1959

T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley, 1991

R. McEliece, *The theory of information and coding (student edition)*, Cambridge University Press, 2003

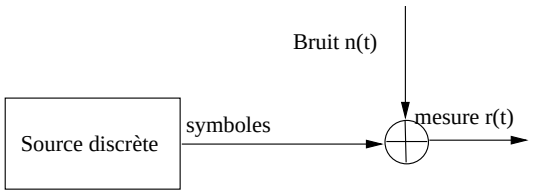
J. F. Young, *Information theory*, Butterworth, 1971

1.6 Plan du document

Outre cette introduction, ce cours est partagé en 4 parties. La première concerne la théorie de la détection, la seconde la théorie de l'estimation, la troisième la théorie de l'information. Ces trois parties de cours sont suivies d'une quatrième partie, qui est un recueil des exercices proposés en travaux dirigés.

Première partie

Théorie de la Détection



Définitions

Introduisons d'abord les définitions générales de la théorie de la détection.

Source

La source génère le signal du côté de l'émetteur. En détection, la source est discrète, c'est-à-dire que le signal émis ne peut prendre qu'un nombre fini de valeurs :

— source binaire : 2 hypothèses (valeurs) notées H_0 et H_1,

— source M-aire : M hypothèses (valeurs) notées $H_0, \ldots, H_{M-1}$.

Exemples.

— Communications numériques : 2 symboles 0 (H_0) et 1 (H_1) sont émis,

— Radar : absence (H_0) ou présence (H_1) d'une cible,

— Classification de locuteurs parmi 6 hypothèses : 3 nationalités (Allemand, Anglais, Français) $\times$ 2 genres (homme ou femme)

Loi de probabilité

La décision optimale nécessite de connaître la loi de probabilité du signal reçu. Ce signal, illustré à la figure ci-dessus, suit la relation :

$$r(t) = \begin{cases} s_0(t) + n(t), & \text{si } H_0 \text{ est vraie,} \\ \ldots & \\ s_{M-1}(t) + n(t), & \text{si } H_{M-1} \text{ est vraie.} \end{cases}$$

Si l'on connait la loi de probabilité du bruit, on peut déduire les lois de probabilité conditionnelles, sachant les différentes hypothèses H_i.

Supposons que l'hypothèse H_i génère le signal $s_i(t)$, et que le bruit $n(t)$ admette la densité de probabilité $p_n(u)$, on peut écrire :

$$p(r/H_i) = p_n(r - s_i).$$

La loi de probabilité du signal observé est l'ensemble des lois conditionnelles $p(r/H_i)$, $i = 1, \ldots, M-1$.

Espace d'observation

Pour une valeur donnée (hypothèse) de la source, on répète généralement la mesure. Chaque observation est ainsi un ensemble de k mesures que l'on peut associer à un vecteur $\mathbf{r}$ dans $\mathbb{R}^k$, l'espace d'observation. La dimension de l'espace d'observation k est donc indépendante du nombre M d'hypothèses. On supposera simplement que ce nombre k est fini.

Règle de décision

A partir des mesures dans l'espace d'observation, on doit finalement décider quelle est l'hypothèse la plus vraisemblable, au moyen d'une règle qui assigne une hypothèse à chaque point de l'espace d'observation. De façon plus globale, la règle de décision partitionne l'espace d'observation en M régions (qui peuvent être non connexes), chacune associée à une hypothèse.

Organisation

Cette partie consacrée à la théorie de la détection est organisée en trois chapitres : cette introduction, un chapitre sur la détection binaire et un chapitre qui généralise à la détection M-aire et détaille le cas ternaire ($M = 3$).

Chapitre 2

Détection binaire

Dans ce chapitre, on suppose des sources binaires, c'est-à-dire que deux hypothèses (valeurs), H_0 et H_1, sont possibles avec des probabilités *a priori*, P_0 et P_1. Chaque observation $\mathbf{r}$ est un vecteur de $\mathbb{R}^k$, dont on suppose connues les lois de probabilité conditionnelles $p(\mathbf{r}/H_0)$ et $p(\mathbf{r}/H_1)$.

Dans le cas binaire, on a 2 hypothèses possibles à l'émission et 2 décisions possibles à la réception, soit 4 situations différentes :

— 1 : H_0 est vraie et on décide H_0,
— 2 : H_0 est vraie et on décide H_1,
— 3 : H_1 est vraie et on décide H_0,
— 4 : H_1 est vraie et on décide H_1.

Les situations 1 et 4 correspondent à des bonnes décisions, les deux autres à des décisions erronnées. La régle de décision que l'on cherche à concevoir doit bien entendu donner le plus souvent possible de bonnes décisions. Pour cela, on associe un critère qui mesure la qualité de la décision.

Dans la suite de cette partie, nous étudierons trois critères : le critère de Bayes, le critère MINIMAX et le critère de Neyman-Pearson.

2.1 Critère de Bayes

On attribue à chacune des quatre situations : "on décide H_i alors que H_j est vraie", un coût C_{ij}. A la règle de décision, on associe un coût moyen, appelé risque de Bayes et noté $\mathcal{R}_{Bayes}$:

$$\begin{aligned} \mathcal{R}_{Bayes} \quad = \quad & C_{00}P_0\text{Pr(choisir } H_0/H_0 \text{ vraie)} \\ & +C_{10}P_0\text{Pr(choisir } H_1/H_0 \text{ vraie)} \\ & +C_{01}P_1\text{Pr(choisir } H_0/H_1 \text{ vraie)} \\ & +C_{11}P_1\text{Pr(choisir } H_1/H_1 \text{ vraie).} \end{aligned} \tag{2.1}$$

Le critère de décision doit permettre de choisir entre les deux hypothèses H_0 et H_1. Ceci revient à partager l'espace d'observation, noté Z (l'ensemble des points $\mathbf{r} \in \mathbb{R}^k$), en deux régions : Z_0 associée à la décision H_0 et Z_1 associée à la décision H_1 (Figure 2.1). On peut donc écrire chaque probabilité Pr(choisir H_i/H_j vraie) comme l'intégrale (multiple, dans $\mathbb{R}^k$) de la densité conditionnelle $p(\mathbf{r}/H_j)$ sur le domaine Z_i :

$$\text{Pr(choisir } H_i/H_j \text{ vraie)} = \int_{Z_i} p(\mathbf{r}/H_j)\mathbf{dr}. \tag{2.2}$$

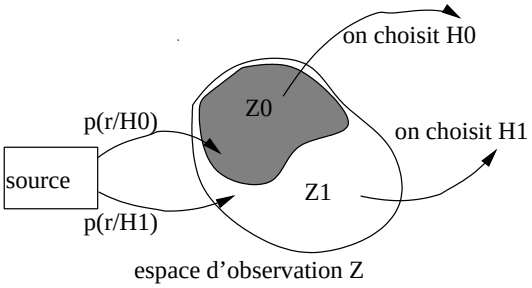


FIGURE 2.1 – La loi de décision vise à partitionner l'espace d'observation.

Le risque de Bayes s'écrit alors :

$$\begin{array}{rcl} \mathcal{R}_{Bayes} & = & C_{00}P_0 \int_{Z_0} p(\mathbf{r}/H_0)\mathbf{dr} \\ & & +C_{10}P_0 \int_{Z_1} p(\mathbf{r}/H_0)\mathbf{dr} \\ & & +C_{01}P_1 \int_{Z_0} p(\mathbf{r}/H_1)\mathbf{dr} \\ & & +C_{11}P_1 \int_{Z_1} p(\mathbf{r}/H_1)\mathbf{dr}. \end{array} \tag{2.3}$$

Les domaines Z_0 et Z_1 formant une partition, on a $Z = Z_0 \bigcup Z_1$ et $Z_0 \bigcap Z_1 = \emptyset$. On peut donc écrire les intégrales en les décomposant en deux termes, en notant $Z_0 = Z \backslash Z_1$ où $\backslash$ représente la différence de deux ensembles. Ainsi :

$$\begin{array}{rcl} \int_{Z_0} p(\mathbf{r}/H_j)\mathbf{dr} & = & \int_{Z\backslash Z_1} p(\mathbf{r}/H_j)\mathbf{dr} \\ & = & \int_{Z} p(\mathbf{r}/H_j)\mathbf{dr} - \int_{Z_1} p(\mathbf{r}/H_j)\mathbf{dr} \end{array} \tag{2.4}$$

De plus, l'intégrale d'une densité sur le domaine tout entier étant égale à 1, on a finalement :

$$\int_{Z_0} p(\mathbf{r}/H_j)\mathbf{dr} = 1 - \int_{Z_1} p(\mathbf{r}/H_j)\mathbf{dr}. \tag{2.5}$$

En utilisant (2.4), on peut écrire l'équation (2.3) sous forme d'intégrales sur les domaines Z et Z_0 uniquement :

$$\begin{array}{rcl} \mathcal{R}_{Bayes} & = & C_{00}P_0 \int_{Z_0} p(\mathbf{r}/H_0)\mathbf{dr} \\ & & +C_{10}P_0 \int_{Z\backslash Z_0} p(\mathbf{r}/H_0)\mathbf{dr} \\ & & +C_{01}P_1 \int_{Z_0} p(\mathbf{r}/H_1)\mathbf{dr} \\ & & +C_{11}P_1 \int_{Z\backslash Z_0} p(\mathbf{r}/H_1)\mathbf{dr}. \end{array} \tag{2.6}$$

En regroupant les termes constants, obtenus en utilisant la propriété (2.5), et ceux sous l'intégrale, on arrive à :

$$\begin{array}{rcl} \mathcal{R}_{Bayes} & = & [C_{11}P_1 + C_{10}P_0] \\ & & + \int_{Z_0} [P_1(C_{01} - C_{11})p(\mathbf{r}/H_1) - P_0(C_{10} - C_{00})p(\mathbf{r}/H_0)]\mathbf{dr}. \end{array} \tag{2.7}$$

2.2 Rapport de vraisemblance

Dans (2.7), le premier terme de droite entre crochets correspond à un coût fixe ; le second, sous l'intégrale, est variable selon le domaine Z_0.

Dans le cas général, les coûts C_{ij} sont quelconques, mais le coût relatif à une décision juste doit être naturellement plus faible que celui relatif à une décision erronnée :

$$\begin{array}{rcl} C_{10} & > & C_{00}, \\ C_{01} & > & C_{11}. \end{array} \tag{2.8}$$

Ainsi, dans le terme entre crochet à l'intérieur de l'intégrale (l'intégrande), les coefficients $(C_{01} - C_{11})$ et $(C_{10} - C_{00})$ sont supposés positifs.

2.2.1 Minimisation

Pour minimiser l'intégrale, on construit Z_0 de sorte que chaque point $\mathbf{r} \in Z_0$ minimise l'intégrale, c'est à dire corresponde à un intégrande négatif. Tous les points $\mathbf{r} \in Z_0$ doivent donc satisfaire :

$$[P_1(C_{01} - C_{11})p(\mathbf{r}/H_1) - P_0(C_{10} - C_{00})p(\mathbf{r}/H_0)] < 0 \tag{2.9}$$

$$P_1(C_{01} - C_{11})p(\mathbf{r}/H_1) < P_0(C_{10} - C_{00})p(\mathbf{r}/H_0) \tag{2.10}$$

soit finalement :

$$\frac{p(\mathbf{r}/H_1)}{p(\mathbf{r}/H_0)} < \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}. \tag{2.11}$$

Réciproquement, pour minimiser le risque de Bayes, les points $\mathbf{r} \in Z_1$ doivent satisfaire l'inégalité :

$$\frac{p(\mathbf{r}/H_1)}{p(\mathbf{r}/H_0)} > \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}. \tag{2.12}$$

2.2.2 Rapport de vraisemblance

On appelle *rapport de vraisemblance*, et on note $\Lambda(\mathbf{r})$ la quantité :

$$\Lambda(\mathbf{r}) = \frac{p(\mathbf{r}/H_1)}{p(\mathbf{r}/H_0)}. \tag{2.13}$$

On voit que $\Lambda(\mathbf{r})$ (grand lambda) est une variable aléatoire positive à une dimension, qui ne dépend que des deux densités de probabilité conditionnelles (c'est leur rapport). La décision optimale (qui minimise le critère de Bayes) est alors obtenue en comparant le rapport de vraisemblance, $\Lambda(\mathbf{r})$, à un seuil scalaire, noté η (eta) :

$$\eta = \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}, \tag{2.14}$$

qui ne dépend que des probabilités *a priori* et des coûts. On rassemble les deux équations (2.11) et (2.12) sous la notation compacte :

$$\Lambda(\mathbf{r}) \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \eta. \tag{2.15}$$

Ce test est appelé test du rapport de vraisemblance (en anglais *likelihood ratio test*).

2.2.3 Logarithme du rapport de vraisemblance

Le rapport de vraisemblance étant une quantité positive, et le logarithme étant une fonction croissante de $\mathbb{R}^{*+}$ dans $\mathbb{R}$, on peut aussi écrire le critère de décision sous la forme :

$$\ln \Lambda(\mathbf{r}) \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \ln \eta. \tag{2.16}$$

Cette forme est très pratique lorsque les densités conditionnelles, qui interviennent dans le rapport de vraisemblance, s'expriment sous forme de produits.

2.3 Exemples

Dans ce paragraphe, nous illustrons l'utilisation du test de rapport de vraisemblance sur trois exemples.

2.3.1 Exemple 1 : Détection d'un signal déterministe dans du bruit

Enoncé. On mesure une tension électrique qui vaut m Volts sous l'hypothèse H_1 et 0 Volt sous l'hypothèse H_0. Une observation est constituée de k mesures prélevées toutes les T_e secondes. On suppose que la valeur du signal ne change pas pendant la durée de l'observation. Chaque mesure est polluée par un bruit additif gaussien, centré, de même variance σ^2 et indépendant du bruit de la mesure précédente. Déterminer le test du rapport de vraisemblance.

Solution. D'après l'énoncé, pour une mesure i, on obtient :

$$r_i = \begin{cases} m & + & n_i, \text{ si } H_1, \\ 0 & + & n_i, \text{ si } H_0. \end{cases} \tag{2.17}$$

On connaît par ailleurs la loi de probabilité de chaque échantillon de bruit :

$$p_n(u) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{u^2}{2\sigma^2}) \tag{2.18}$$

On peut facilement calculer les deux lois conditionnelles. Sous l'hypothèse H_0, on a :

$$\begin{aligned} p(r_i/H_0) &= p_n(r_i) \\ &= \tfrac{1}{\sqrt{2\pi}\sigma} \exp(-\tfrac{r_i^2}{2\sigma^2}). \end{aligned} \tag{2.19}$$

Sous l'hypothèse H_1, on a :

$$\begin{aligned} p(r_i/H_1) &= p_n(r_i - m) \\ &= \tfrac{1}{\sqrt{2\pi}\sigma} \exp(-\tfrac{(r_i-m)^2}{2\sigma^2}). \end{aligned} \tag{2.20}$$

L'observation est constituée de k mesures. C'est donc un vecteur à k composantes : $\mathbf{r} = (r_1, r_2, \ldots, r_k)^T$, où T représente l'opération de transpostion. Puisque les échantillons de bruits,

n_i, sont indépendants, la loi conditionnelle $p(\mathbf{r}/H_j)$ est simplement le produit des densités $p(r_i/H_j)$, c'est-à-dire sous l'hypothèse H_0 :

$$\begin{aligned} p(\mathbf{r}/H_0) &= \Pi_{i=1}^{k} p(r_i/H_0) \\ &= \Pi_{i=1}^{k} p_n(r_i) \\ &= \Pi_{i=1}^{k} \frac{1}{\sqrt{2\pi}\sigma} \exp\big(-\frac{r_i^2}{2\sigma^2}\big). \end{aligned} \tag{2.21}$$

et sous l'hypothèse H_1 :

$$\begin{aligned} p(\mathbf{r}/H_1) &= \Pi_{i=1}^{k} p(r_i/H_1) \\ &= \Pi_{i=1}^{k} p_n(r_i - m) \\ &= \Pi_{i=1}^{k} \frac{1}{\sqrt{2\pi}\sigma} \exp\big(-\frac{(r_i-m)^2}{2\sigma^2}\big). \end{aligned} \tag{2.22}$$

On peut écrire le rapport de vraisemblance $\Lambda(\mathbf{r})$:

$$\begin{aligned} \Lambda(\mathbf{r}) &= \frac{p(\mathbf{r}/H_1)}{p(\mathbf{r}/H_0)} \\ &= \frac{\Pi_{i=1}^{k} \frac{1}{\sqrt{2\pi}\sigma} \exp\big(-\frac{(r_i-m)^2}{2\sigma^2}\big)}{\Pi_{i=1}^{k} \frac{1}{\sqrt{2\pi}\sigma} \exp\big(-\frac{r_i^2}{2\sigma^2}\big)} \end{aligned} \tag{2.23}$$

En simplifiant et en prenant le logarithme, on obtient :

$$\begin{aligned} \ln \Lambda(\mathbf{r}) &= \ln \frac{\Pi_{i=1}^{k} \exp\big(-\frac{(r_i-m)^2}{2\sigma^2}\big)}{\Pi_{i=1}^{k} \exp\big(-\frac{r_i^2}{2\sigma^2}\big)} \\ &= \frac{1}{2\sigma^2} \Sigma_{i=1}^{k} (-(r_i - m)^2 + r_i^2) \\ &= \frac{m}{\sigma^2} (\Sigma_{i=1}^{k} r_i - \frac{km}{2}) \end{aligned} \tag{2.24}$$

Le test du rapport de vraisemblance est obtenu en comparant $\ln \Lambda(\mathbf{r})$ à un seuil $\ln \eta$ où $\eta = \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})}$. On arrive finalement à :

$$\Sigma_{i=1}^{k} r_i \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \frac{\sigma^2}{m} \ln \eta + \frac{km}{2} = \gamma \tag{2.25}$$

On remarque que toute l'information sur les mesures qui permet de faire le test optimal est contenu dans la somme des mesures. La quantité $l(\mathbf{r}) = \Sigma_{i=1}^{k} r_i$ constitue une *statistique suffisante*. Il est donc inutile de conserver toutes les mesures r_i : seule la somme est importante. Cette remarque est très importante car elle conditionne la mise en œuvre, logicielle ou matérielle, du test.

2.3.2 Exemple 2 : Détection d'un signal aléatoire continu dans du bruit

Enoncé. Chaque observation est un ensemble de k mesures : r_1, r_2, ..., r_k. Les r_i sont des variables aléatoires gaussiennes, centrées, indépendantes et de variance σ_0^2 sous l'hypothèse H_0 et σ_1^2 sous l'hypothèse H_1. On suppose que la source ne change pas pendant la durée de l'observation. Déterminer le test du rapport de vraisemblance.

Solution. On peut calculer les densités conditionnelles de chaque mesure r_i. Sous l'hypothèse H_0, on a :

$$p(r_i/H_0) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\,(\,-\frac{r_i^2}{2\sigma_0^2}). \tag{2.26}$$

Sous l'hypothèse H_1, on a :

$$p(r_i/H_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\,(\,-\frac{r_i^2}{2\sigma_1^2}). \tag{2.27}$$

Chaque mesure étant indépendante, on peut facilement calculer les lois conditionnelles de l'observation $\mathbf{r}$. Sous l'hypohèse H_0, on a :

$$\begin{array}{rcl} p(\mathbf{r}/H_0) & = & \Pi_{i=1}^k p(r_i/H_0) \\ & = & \Pi_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma_0} \exp\,(\,-\frac{r_i^2}{2\sigma_0^2}). \end{array} \tag{2.28}$$

et sous l'hypothèse H_1 :

$$\begin{array}{rcl} p(\mathbf{r}/H_1) & = & \Pi_{i=1}^k p(r_i/H_1) \\ & = & \Pi_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma_1} \exp\,(\,-\frac{r_i^2}{2\sigma_1^2}). \end{array} \tag{2.29}$$

On peut maintenant calculer le rapport de vraisemblance $\Lambda(\mathbf{r})$:

$$\begin{array}{rcl} \Lambda(\mathbf{r}) & = & \frac{p(\mathbf{r}/H_1)}{p(\mathbf{r}/H_0)} \\ & = & \frac{\Pi_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{r_i^2}{2\sigma_1^2}\right)}{\Pi_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{r_i^2}{2\sigma_0^2}\right)}. \end{array} \tag{2.30}$$

En simplifiant et en prenant le logarithme, on obtient :

$$\begin{array}{rcl} \ln \Lambda(\mathbf{r}) & = & \ln \frac{\Pi_{i=1}^k \frac{1}{\sigma_1} \exp\left(-\frac{r_i^2}{2\sigma_1^2}\right)}{\Pi_{i=1}^k \frac{1}{\sigma_0} \exp\left(-\frac{r_i^2}{2\sigma_0^2}\right)} \\ & = & k \ln \frac{\sigma_0}{\sigma_1} - \frac{1}{2}(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2})\Sigma_{i=1}^k r_i^2. \end{array} \tag{2.31}$$

Le test du rapport de vraisemblance est obtenu en comparant $\ln \Lambda(\mathbf{r})$ à un seuil $\ln \eta$ où $\eta = \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})}$). Après simplification, si $\sigma_1^2 - \sigma_0^2 > 0$, on arrive finalement à :

$$\Sigma_{i=1}^k r_i^2 \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}(\ln \eta - k \ln \frac{\sigma_0}{\sigma_1}) = \gamma. \tag{2.32}$$

Dans le cas contraire, c'est-à-dire si $\sigma_1^2 - \sigma_0^2 < 0$, il faut inverser le sens des inégalités en divisant par ce terme, et on obtient le test contraire :

$$\Sigma_{i=1}^k r_i^2 \underset{\mathrm{H_1}}{\overset{\mathrm{H_0}}{\gtrless}} \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}(\ln \eta - k \ln \frac{\sigma_0}{\sigma_1}) = \gamma. \tag{2.33}$$

De nouveau, le test de rapport de vraisemblance met en évidence une statistique suffisante. Contrairement à l'exemple précédent, c'est ici la quantité $l(\mathbf{r}) = \Sigma_{i=1}^k r_i^2$. Cette statistique est importante pour la mise en œuvre du test.

2.3.3 Exemple 3 : Détection d'un signal aléatoire discret dans du bruit

Enoncé. Pendant un temps d'observation T, on recueille les impulsions envoyées par une source dont l'émission suit une loi de Poisson, de moyenne m_0 sous l'hypothèse H_0 et m_1 sous l'hypothèse H_1. Déterminer le test du rapport de vraisemblance.

Solution. La source émettant selon une loi de Poisson, on a directement la probabilité d'observer n impulsions :

$$\Pr(n \text{ impulsions}) = \begin{cases} \frac{m_0^n}{n!}\exp(-m_0), & \text{sous l'hypothèse } H_0, \\ \frac{m_1^n}{n!}\exp(-m_1), & \text{sous l'hypothèse } H_1. \end{cases} \tag{2.34}$$

Par rapport aux deux exemples précédents, celui-ci est caractérisé par des évènements discrets, et donc modélisé par une loi discrète car la densité n'existe pas. Le rapport de vraisemblance est le rapport des probabilités conditionnelles :

$$\begin{aligned} \Lambda(\mathbf{r}) &= \frac{\Pr(n \text{ impulsions}/H_1)}{\Pr(n \text{ impulsions}/H_0)} \\ &= \frac{\frac{m_1^n}{n!}\exp(-m_1)}{\frac{m_0^n}{n!}\exp(-m_0)} \\ &= (\tfrac{m_1}{m_0})^n \exp[-(m_1 - m_0)]. \end{aligned} \tag{2.35}$$

En prenant le logarithme, on arrive au test :

$$\ln \Lambda(\mathbf{r}) = n \ln \frac{m_1}{m_0} - (m_1 - m_0) \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \ln \eta. \tag{2.36}$$

Après quelques simplifications, on obtient la forme suivante, si $m_1 > m_0$:

$$n \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \frac{\ln \eta + m_1 - m_0}{\ln m_1 - \ln m_0}, \tag{2.37}$$

ou bien, si $m_1 < m_0$:

$$n \underset{\mathrm{H_1}}{\overset{\mathrm{H_0}}{\gtrless}} \frac{\ln \eta + m_1 - m_0}{\ln m_1 - \ln m_0}. \tag{2.38}$$

Dans cet exemple, on remarque encore une statistique suffisante : le test ne demande que la connaissance du nombre d'impulsions. Inutile de mémoriser les dates d'arrivée de ces impulsions. Cette remarque permet également de concevoir le récepteur optimal le plus simple : c'est un simple compteur, suivi d'un comparateur.

2.4 Choix de coûts

Dans tous les exemples précédents, le test du rapport de vraisemblance, optimal au sens du risque de Bayes, fait intervenir le seuil η. Ce seuil s'exprime en fonction des probabilités *a priori*, P_0 et P_1, et des coûts, C_{ij}. Si les probabilités choisies sont fausses, le seuil n'est plus optimal. Si on change les coûts, le seuil de décision change également.

2.4.1 Coûts uniformes

On suppose que les erreurs ont un coût identique : $C_{01} = C_{10} = 1$ et que les bonnes décisions ont un coût nul : $C_{00} = C_{11} = 0$.

Le risque de Bayes (2.1) se simplifie alors :

$$\begin{array}{rcl} \mathcal{R}_{Bayes} & = & P_0\text{Pr(choisir } H_1/H_0 \text{ vraie)} + P_1\text{Pr(choisir } H_0/H_1 \text{ vraie)} \\ & = & P_0 \int_{Z_1} p(\mathbf{r}/H_0)\mathbf{dr} + P_1 \int_{Z_0} p(\mathbf{r}/H_1)\mathbf{dr}. \end{array} \tag{2.39}$$

Ce risque s'interprète facilement comme la probabilité moyenne d'erreur. Dans le test du rapport de vraisemblance associé, seul le seuil η est modifié. Avec les hypothèses ci-dessus, on a :

$$\eta = \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} = \frac{P_0}{P_1}, \tag{2.40}$$

d'où le test :

$$\Lambda(\mathbf{r}) \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \frac{P_0}{P_1}. \tag{2.41}$$

ou bien, sous forme logarithmique :

$$\ln(\Lambda(\mathbf{r})) \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \ln \frac{P_0}{P_1}. \tag{2.42}$$

2.4.2 Communications numériques

Dans le cas de communications numériques, les erreurs sont pénalisées de façon identique. De plus, on peut supposer que les probabilités *a priori* des symboles 0 et 1 sont identiques, autrement dit : $P_0 = P_1 = 1/2$. Le seuil du test vaut donc :

$$\eta = \frac{P_0}{P_1} = 1. \tag{2.43}$$

Le test du rapport de vraisemblance prend alors la forme très simple :

$$\Lambda(\mathbf{r}) \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} 1, \tag{2.44}$$

ou bien, sous forme logarithmique :

$$\ln(\Lambda(\mathbf{r})) \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} 0. \tag{2.45}$$

2.4.3 Probabilités *a priori* inconnues.

La décision optimale dépend du seuil η, et revient à déterminer les régions Z_0 et Z_1. A coûts fixés, le seuil ne dépend que des probabilités *a priori*. Dans le cas binaire, on peut l'exprimer en

fonction de P_1 (car $P_0 = 1 - P_1$). On introduit la probabilité de fausse alarme, P_F, la probabilité de détection, P_D et la probabilité d'oubli, P_M (M pour Missing) :

$$\begin{array}{rcl} P_F & = & \text{Pr(choisir } H_1/H_0 \text{ vraie)} = \int_{Z_1} p(\mathbf{r}/H_0)\mathbf{dr} \\ P_D & = & \text{Pr(choisir } H_1/H_1 \text{ vraie)} = \int_{Z_1} p(\mathbf{r}/H_1)\mathbf{dr} \\ P_M & = & \text{Pr(choisir } H_0/H_1 \text{ vraie)} = \int_{Z_0} p(\mathbf{r}/H_1)\mathbf{dr}. \end{array} \tag{2.46}$$

Les relations sous forme d'intégrales supposent l'existence des densités $p(\mathbf{r}/H_i)$ (ce ne sera pas le cas si la variable aléatoire est discrète - voir l'exemple 3 au paragraphe 2.3.3 ci-dessus). Dans ce cas, les trois probabilités qui minimisent le risque de Bayes sont fonctions du seuil η, lui-même fonction continue de P_1. Si P_1 change, le seuil η change de façon continue ainsi que les régions Z_0 et Z_1 et par conséquent les trois probabilités, P_F P_D et P_M. Ces probabilités, P_F P_D et P_M, sont donc des fonctions continues de P_1. Pour simplifier, dans ce paragraphe, on omettra cette dépendance et on écrira simplement : $P_F(P_1) = P_F$, etc. On peut alors ré-écrire le risque de Bayes (2.6) :

$$\mathcal{R}_{Bayes} = C_{00}P_0(1 - P_F) + C_{10}P_0P_F + C_{01}P_1P_M + C_{11}P_1P_D. \tag{2.47}$$

En remarquant que $P_M = 1 - P_D$, et que $P_0 = 1 - P_1$, on peut exprimer le risque de Bayes en fonction uniquement de P_F, P_M et P_1 :

$$\begin{array}{rcl} \mathcal{R}_{Bayes}(P_1) & = & C_{00}(1 - P_F) + C_{10}P_F \\ & & +P_1[(C_{11} - C_{00}) + (C_{01} - C_{11})P_M - (C_{10} - C_{00})P_F]. \end{array} \tag{2.48}$$

2.5 Courbes de risque

Les courbes de risque sont les courbes $\mathcal{R}_{Bayes}(P_1)$, fonction de P_1. Si les coûts et les probabilités *a priori* sont connus, on connait η et on peut en déduire le test optimal de Bayes, en calculant les valeurs exactes de P_F et de P_M. Si les probabilités *a priori* sont inconnues, le seuil η n'est pas connu avec exactitude et les valeurs de P_F et P_M ne sont pas les valeurs optimales.

Supposons $P_1 = P_1^*$, et notons $P_F^* = P_F(P_1^*)$ et $P_M^* = P_M(P_1^*)$ les probabilités de fausse alarme et d'oubli calculées pour P_1^*. La relation (2.48) ne dépend plus que de la variable P_1. On notera $\mathcal{R}(P_1^*, P_1)$ le risque calculé avec les probabilités P_F^* et P_M^* :

$$\begin{array}{rcl} \mathcal{R}(P_1^*, P_1) & = & C_{00}(1 - P_F^*) + C_{10}P_F^* \\ & & +P_1[(C_{11} - C_{00}) + (C_{01} - C_{11})P_M^* - (C_{10} - C_{00})P_F^*]. \end{array} \tag{2.49}$$

On remarque que ce risque est une fonction affine de P_1. Ce risque $\mathcal{R}(P_1^*, P_1)$ ne coïncide avec le risque optimal de Bayes que pour $P_1 = P_1^*$. $\mathcal{R}(P_1^*, P_1)$ est tangent au risque de Bayes $\mathcal{R}_{Bayes}(P_1)$ en $P_1 = P_1^*$, et au-dessus partout ailleurs (Figure 2.2) :

$$\mathcal{R}(P_1^*, P_1) \geq \mathcal{R}_{Bayes}(P_1). \tag{2.50}$$

Pour mieux caractériser les courbes de risque de Bayes, on peut calculer quelques valeurs particulières, par exemple pour $P_1 = 0$ et $P_1 = 1$. Pour cela, calculons P_M et P_F pour ces valeurs de P_1.

Si $P_1 = 0$, on sait que le symbole 1 n'est jamais émis. Par conséquent, on ne décidera jamais H_1, et :

$$\begin{array}{rcll} P_F(0) & = & \text{Pr(décider } H_1/H_0 \text{ vraie)} & = 0 \\ P_M(0) & = & \text{Pr(décider } H_0/H_1 \text{ vraie)} & = 0. \end{array} \tag{2.51}$$

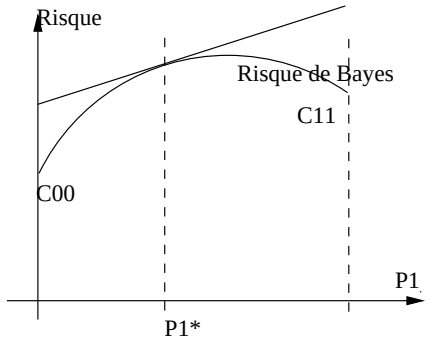


FIGURE 2.2 – Risque de Bayes et risque à P_1^* fixé.

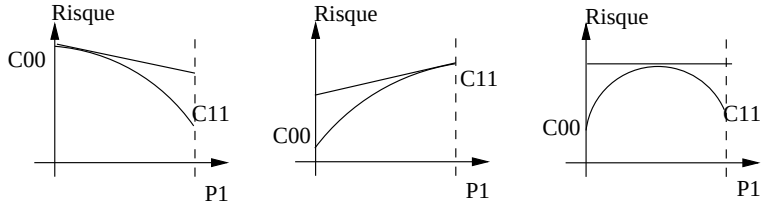


FIGURE 2.3 – Exemples de courbes de risques

Le risque de Bayes devient alors :

$$\begin{aligned} \mathcal{R}_{Bayes}(0) &= C_{00}(1 - P_F(0)) + C_{10}P_F(0) \\ &= C_{00}. \end{aligned} \tag{2.52}$$

Si $P_1 = 1$, on sait que le symbole 1 est toujours émis (H_0 n'est jamais vraie). Par conséquent, on décide toujours H_1, et :

$$\begin{aligned} P_F(1) &= \text{Pr(décider } H_1/H_0 \text{ vraie)} = 0 \\ P_M(1) &= \text{Pr(décider } H_0/H_1 \text{ vraie)} = 0. \end{aligned} \tag{2.53}$$

Le risque de Bayes devient alors :

$$\begin{aligned} \mathcal{R}_{Bayes}(1) &= C_{00}(1 - P_F(1)) + C_{10}P_F(1) \\ &\quad +[(C_{11} - C_{00}) + (C_{01} - C_{11})P_M(1) - (C_{10} - C_{00})P_F(1)], \\ &= C_{11}. \end{aligned} \tag{2.54}$$

Les courbes de risque ont les allures typiques suivantes (Figure 2.3).

2.6 Critère MINIMAX

Dans le cas où les probabilités *a priori* sont inconnues, on pourrait minimiser le risque calculé pour une valeur fixée P_1^* de P_1. Mais, en regardant les différentes courbes de risques (Fig. 2.3), on remarque que cette stratégie est risquée. En effet, pour certaines valeurs de P_1, le critère minimisé serait très différent du risque de Bayes.

Pour éviter cette situation, dans le cas où la courbe de risque présente un maximum pour $0 < P_1 < 1$, une stratégie consiste à minimiser le risque maximum. Soit P_1^* la valeur où le risque

est maximal, la droite $\mathcal{R}_{Bayes}(P_1^*, P_1)$ est donc tangente au maximum de la courbe de risque $\mathcal{R}_{Bayes}(P_1)$. Elle est caractérisée par sa pente nulle, c'est-à-dire d'après (2.48) par la relation :

$$(C_{11} - C_{00}) + (C_{01} - C_{11})P_M - (C_{10} - C_{00})P_F = 0. \tag{2.55}$$

Ce test qui minimise le risque maximal s'appelle test *MINIMAX*. Il est caractérisé par l'équation (2.55) et le risque vaut :

$$\mathcal{R}_{Minimax} = C_{00}(1 - P_F) + C_{10}P_F. \tag{2.56}$$

Dans le cas fréquent où l'on choisit $C_{00} = C_{11} = 0$, le test *MINIMAX* est caractérisé par l'équation :

$$C_{01}P_M - C_{10}P_F = 0, \tag{2.57}$$

et le risque vaut :

$$\mathcal{R}_{Minimax} = C_{10}P_F. \tag{2.58}$$

2.7 Test de Neyman-Pearson

En pratique, il est souvent difficile d'attribuer des coûts réalistes et des probabilités *a priori*. Pour contourner cette difficulté, on peut utiliser une autre stratégie à partir de probabilités P_F et P_D. En effet, on peut chercher le test qui produit la probabilité de fausse alarme, P_F, aussi petite que possible et la probabilité de détection, P_D aussi grande que possible.

Fixons $P_F \leq \alpha$ et cherchons un test qui maximise P_D (ou minimise $P_M = 1 - P_D$). Pour cela, on construit la fonction de coût F :

$$\begin{array}{rcl} F & = & P_M + \lambda[P_F - \alpha], \\ & = & \int_{Z_0} p(\mathbf{r}/H_1)\mathbf{dr} + \lambda[\int_{Z\setminus Z_0} p(\mathbf{r}/H_0)\mathbf{dr} - \alpha] \\ & = & \lambda(1-\alpha) + \int_{Z_0}[p(\mathbf{r}/H_1) - \lambda p(\mathbf{r}/H_0)]\mathbf{dr}, \end{array} \tag{2.59}$$

où λ est un multiplieur de Lagrange. Dans la dernière équation, on remarque que le premier terme de droite est un coût fixe. Pour minimiser F, il faut donc choisir $\mathbf{r} \in Z_0$ si l'intégrande $p(\mathbf{r}/H_1) - \lambda p(\mathbf{r}/H_0)$ est négatif, c'est-à-dire si :

$$\frac{p(\mathbf{r}/H_1)}{p(\mathbf{r}/H_0)} < \lambda. \tag{2.60}$$

On obtient donc le test :

$$\text{Si } \Lambda(\mathbf{r}) < \lambda, \text{ alors on choisit } H_0, \tag{2.61}$$

où le seuil λ est calculé par l'équation :

$$P_F = \Pr(\Lambda(\mathbf{r}) > \lambda/H_0) = \int_{\lambda}^{+\infty} p_\Lambda(u/H_0)du = \alpha. \tag{2.62}$$

Dans cette équation intégrale, l'inconnue λ est une borne de l'intégrale. Puisque la variable aléatoire $\Lambda(\mathbf{r})$ ne prend que des valeurs positives (c'est le rapport de deux densités), le seuil λ doit aussi être positif. Faire décroître λ revient à augmenter la région Z_1 où l'on décide H_1 : la probilité de fausse alarme P_F ainsi que la probabilité de détection P_D augmentent si λ diminue. La résolution de cette équation (2.62) est en général impossible analytiquement. Si P_F est une fonction continue de λ, le test de Neyman-Pearson, comme le montre l'expression (2.60), est un test du rapport de vraisemblance, ce qui rend ce test très intéressant.

2.8 Statistique suffisante

Dans les paragraphes précédents, deux idées essentielles ont été développées :

— Relativement à un critère de Bayes ou de Neyman-Pearson, le meilleur test revient à comparer le rapport de vraisemblance $\Lambda(\mathbf{r}) = \frac{p(\mathbf{r}/H_1)}{p(\mathbf{r}/H_0)}$ à un seuil scalaire η. Quelle que soit la dimension k de l'espace d'observation ($\mathbf{r} \in \mathbb{R}^k$) l'espace de décision est mono-dimentionnel.

— En calculant le rapport de vraisemblance, on met en évidence une statistique suffisante, qui renseigne sur la structure (logicielle et/ou matérielle) du détecteur optimal.

Ce dernier point n'apparaît pas directement dans le calcul théorique, mais généralement dans le calcul explicite (voir en particulier les trois exemples). L'explication théorique est assez simple, et peut parfois s'illustrer géométriquement. D'un point de vue théorique, on peut décomposer l'espace d'observation $\mathbf{R} = \{\mathbf{r} \in \mathbb{R}^k\}$ de dimension k en un espace de dimension 1 correspondant à la statistique suffisante l et un espace de dimension $(k-1)$ dont on notera les éléments $\mathbf{y}$. On peut donc écrire le rapport de vraisemblance :

$$\Lambda(\mathbf{r}) = \Lambda(l, \mathbf{y}) = \frac{p_{l,\mathbf{y}/H_1}(l, \mathbf{y}/H_1)}{p_{l,\mathbf{y}/H_0}(l, \mathbf{y}/H_0)} \tag{2.63}$$

En utilisant le théorème de Bayes, on a :

$$\Lambda(l, \mathbf{y}) = \frac{p_{l/H_1}(l/H_1) p_{\mathbf{y}/l,H_1}(\mathbf{y}/l, H_1)}{p_{l/H_0}(l/H_0) p_{\mathbf{y}/l,H_0}(\mathbf{y}/l, H_0)} \tag{2.64}$$

Puisque l est une statistique suffisante, la décision ne dépend que de l et pas de $\mathbf{y}$, et $\Lambda(l, \mathbf{y})$ doit donc se réduire à $\Lambda(l)$. On doit donc avoir :

$$p_{\mathbf{y}/l,H_1}(\mathbf{y}/l, H_1) = p_{\mathbf{y}/l,H_0}(\mathbf{y}/l, H_0), \tag{2.65}$$

car la densité de $\mathbf{y}$ ne dépend pas des hypothèses H_0 ou H_1.

Dans l'exemple 1 (paragraphe 2.3.1), en nous restreignant à $k = 2$ (2 mesures = 2 dimensions), la statistique suffisante est la somme des observations : $l \propto r_1 + r_2$. On peut donc transformer l'espace initial (r_1, r_2) par une simple rotation en $l = (r_1 + r_2)/\sqrt{2}$ et $y = (r_1 - r_2)/\sqrt{2}$. Sur la seconde coordonnée y, on mesure :

$$y = \begin{cases} (m + n_1 - m - n_2)/\sqrt{2} \;=\; (n_1 - n_2)/\sqrt{2}, & \text{si } H_1 \text{ est vraie,} \\ (n_1 - n_2)/\sqrt{2}, & \text{si } H_0 \text{ est vraie.} \end{cases} \tag{2.66}$$

On remarque que la variable y est identique pour les deux hypothèses : elle n'est donc d'aucune utilité dans la décision.

2.9 Performance du test

On caractérise un détecteur par ses performances, c'est-à-dire le couple (P_D, P_F). De façon exhaustive (et théorique), on peut tracer les courbes P_D en fonction de P_F pour différentes valeurs du seuil η (test de Bayes). Dans le cas d'un test de Neyman-Pearson, le seuil λ permet directement de calculer P_F selon la relation (2.62) et P_D selon :

$$P_D = \Pr(\Lambda(\mathbf{r}) > \lambda / H_1) = \int_{\lambda}^{+\infty} p_{\Lambda}(u/H_1) du. \tag{2.67}$$

Dans ce paragraphe, nous allons calculer ces courbes pour les trois exemples du paragraphe 2.3, puis en déterminer les propriétés essentielles.

2.9.1 Performance de l'exemple 1

A partir du test (2.25), en divisant les deux termes par $\sqrt{k}\sigma$ (normalisation), on a :

$$l = \frac{1}{\sqrt{k}\sigma}\Sigma_{i=1}^{k} r_i \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \frac{\sigma}{m\sqrt{k}} \ln \eta + \frac{\sqrt{k}m}{2\sigma} \tag{2.68}$$

Cette division réalise une normalisation de la variable aléatoire l, c'est-à-dire de la statistique suffisante. En effet, chaque mesure r_i est la réalisation d'une variable aléatoire (VA) R_i qui suit une loi gaussienne, de moyenne m (si H_1) ou 0 (si H_0) et de variance σ^2, ce que nous noterons :

$$\left\{ \begin{array}{lcll} R_i & \sim & N(0,\sigma^2), & \text{si } H_0 \text{ est vraie,} \\ R_i & \sim & N(m,\sigma^2), & \text{si } H_1 \text{ est vraie.} \end{array} \right. \tag{2.69}$$

La somme $S = \Sigma_{i=1}^{k} R_i$ est la somme de k VA gaussiennes, de même moyenne (m ou 0) et de variance σ^2. En utilisant les résultats classiques sur la somme de VA gaussiennes identiques (la moyenne de la somme est égale à la somme des moyennes et les variances s'ajoutent) on peut donc écrire :

$$\left\{ \begin{array}{lcll} S & \sim & N(0,k\sigma^2), & \text{si } H_0 \text{ est vraie,} \\ S & \sim & N(km,k\sigma^2), & \text{si } H_1 \text{ est vraie.} \end{array} \right. \tag{2.70}$$

En divisant S par $\sqrt{k}\sigma$, on obtient une variable normalisée L, c'est-à-dire de variance unité :

$$\left\{ \begin{array}{lcll} L & \sim & N(0,1), & \text{si } H_0 \text{ est vraie,} \\ L & \sim & N(\sqrt{k}m/\sigma,1), & \text{si } H_1 \text{ est vraie.} \end{array} \right. \tag{2.71}$$

Les lois conditionnelles $p_{L/H_0}(l/H_0)$ et $p_{L/H_1}(l/H_1)$ suivent des densités gaussiennes normées de moyenne 0 et $\sqrt{k}m/\sigma$, respectivement. Ces deux gaussiennes sont tracées à la figure 2.4. En posant $d = \sqrt{k}m/\sigma$, on remarque que la décision optimale (2.68) utilise le seuil scalaire $\theta = (\ln \eta)/d + d/2$. Selon les valeurs de η, on a (pour $d > 0$) :

$$\left\{ \begin{array}{lcll} \theta & = & d/2, & \text{si } \eta = 1, \\ \theta & > & d/2, & \text{si } \eta > 1, \\ \theta & < & d/2, & \text{si } \eta < 1. \end{array} \right. \tag{2.72}$$

En reprenant les définitions des probabilités de fausse alarme et de détection, on peut écrire :

$$\begin{array}{rcl} P_F & = & \text{Pr(choisir } H_1 \text{ /} H_0 \text{ vraie)} \\ & = & \text{Pr}(L > \theta / H_0) \\ & = & \int_{\theta}^{+\infty} p_{L/H_0}(u/H_0)du, \end{array} \tag{2.73}$$

et

$$\begin{array}{rcl} P_D & = & \text{Pr(choisir } H_1 \text{ /} H_1 \text{ vraie)} \\ & = & \text{Pr}(L > \theta / H_1) \\ & = & \int_{\theta}^{+\infty} p_{L/H_1}(u/H_1)du. \end{array} \tag{2.74}$$

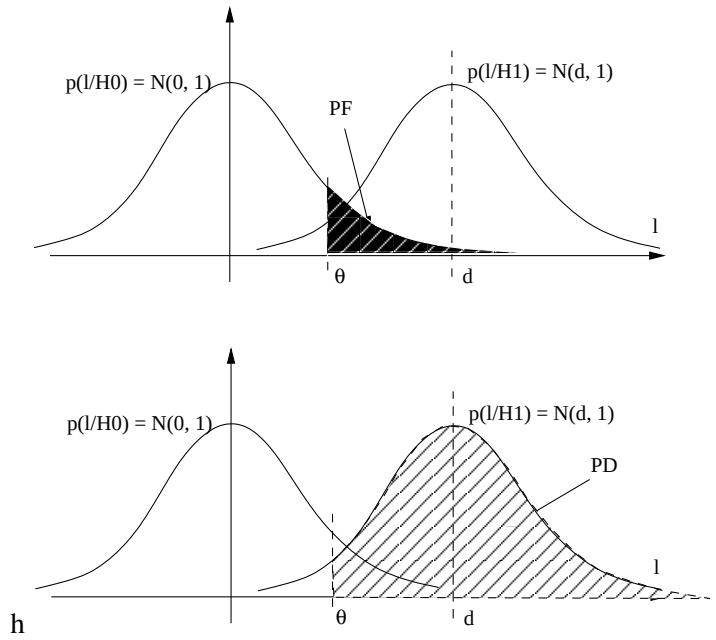


FIGURE 2.4 – Lois conditionnelles de la variable L, avec P_F (en haut) et P_D en bas.

Pour un seuil θ fixé, les probabilités de fausse alarme P_F et de détection P_D sont représentées par les aires hachurées sur la figure 2.4. Dans cet exemple, les lois $p_{L/H_0}(u/H_0)$ et $p_{L/H_1}(u/H_1)$ étant connues, on peut expliciter le calcul :

$$P_F = \int_{(\ln \eta)/d+d/2}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}) du, \tag{2.75}$$

Il n'est pas possible de continuer le calcul analytique. Pour la distribution gaussienne, il existe des tables (indirectes) calculées (voir Travaux Dirigés) pour la fonction $\text{erf}(x)$ définie par :

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-u^2) du. \tag{2.76}$$

On peut alors exprimer les probabilités P_F et P_D à partir de cette fonction :

$$P_F = \frac{1 - \text{erf}(\theta/\sqrt{2})}{2}, \tag{2.77}$$

$$P_D = \frac{1 - \text{erf}((\theta - d)/\sqrt{2})}{2}. \tag{2.78}$$

.

On remarque que l'on a $P_D \geq P_F$, avec égalité si $d = 0$ (c'est-à-dire si $m = 0$!) et pour $P_F = 0$ et pour $P_F = 1$. La fonction $\text{erf}(x)$ étant croissante, on remarque que P_D comme P_F sont des fonctions décroissantes du seuil θ. Puisque $\theta = \ln \eta/d + d/2$, on déduit que :

— à d fixé, P_D et P_F sont fonctions décroissantes de η,

— à P_F fixée, P_D croît avec d.

On obtient des courbes typiques (Fig. 2.5) appelées courbes opérationnelles du récepteur (COR) dont l'acronyme anglais est ROC pour Receiver Operating Characteristic.

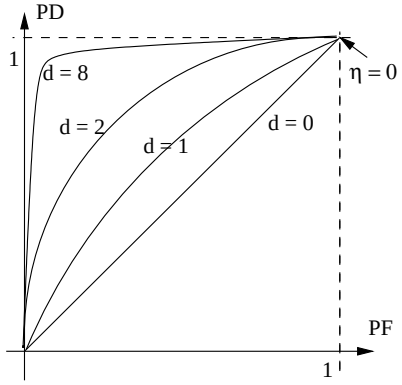


FIGURE 2.5 – Allure typique des courbes ROC. Pour $d = 0$, les deux densités conditionnelles sont identiques, par conséquent $P_D = P_F$. Si d augmente, les courbes de performance s'écartent de la diagonale et on a toujours $P_D \geq P_F$. A $d = cte$, chaque courbe passe par $P_D = P_F = 1$ pour $\eta = 0$ et tend vers $P_D = P_F = 0$ pour $\eta \to +\infty$.

2.9.2 Performances pour la minimisation de l'erreur totale.

Dans ce cas particulier, associé aux coûts $C_{ij} = 1 - \delta_{ij}$ où δ_{ij} est le symbole de Kronecker, le risque vaut $\mathcal{R}_{Bayes} = P_0 P_F + P_1 P_M$. Si l'on suppose aussi $P_0 = P_1 = 0.5$, on a $\mathcal{R}_{Bayes} = (P_F + P_M)/2$, et on a $\eta = 1$, d'où le seuil $\theta = d/2$. Les probabilités de fausse alarme et de détection s'écrivent simplement :

$$\begin{aligned} P_F &= \int_{d/2}^{+\infty} p_{L/H_0}(u/H_0)du \\ &= \int_{d/2}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{u^2}{2}\Big)du, \end{aligned} \tag{2.79}$$

et

$$\begin{aligned} P_D &= \int_{d/2}^{+\infty} p_{L/H_1}(u/H_1)du \\ &= \int_{d/2}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(u-d)^2}{2}\Big)du \\ &= \int_{-d/2}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{u^2}{2}\Big)du. \end{aligned} \tag{2.80}$$

En comparant les deux équations (2.79) et (2.80), et en utilisant la symétrie de la gaussienne, on remarque que les valeurs sont situées sur une droite :

$$P_D = 1 - P_F = \frac{1 + \text{erf}(\frac{d}{2\sqrt{2}})}{2} \tag{2.81}$$

2.9.3 Performance de l'exemple 3

Reprenons maintenant l'exemple 3, du paragraphe 2.3.3. Pour $m_1 > m_0$, on avait trouvé le test (2.37) :

$$n \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \frac{\ln\eta + m_1 - m_0}{\ln m_1 - \ln m_0} = \gamma. \tag{2.82}$$

Bien sûr, le seuil γ est généralement un réel, alors que le nombre d'événements n est un entier. En notant $\gamma_I = \text{int}(\gamma + 1)$ où $\text{int}(u)$ représente la partie entière de u, on peut alors remplacer le test

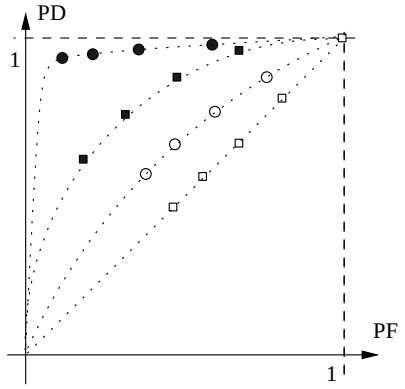


FIGURE 2.6 – Caractéristiques COR dans le cas d'une variable discrète. Seules certaines paires (P_D, P_F) existent. Elles sont représentées ici par des points, carrés ou ronds, vides ou plein, selon les valeurs de m_0 et m_1.

précédent par :

$$\begin{array}{rcll} n & \geq & \gamma_I, & \text{on choisit } H_1, \\ n & < & \gamma_I, & \text{on choisit } H_0. \end{array} \tag{2.83}$$

On peut alors écrire les probabilités de détection et de fausse alarme. Détaillons d'abord le calcul de P_D :

$$\begin{array}{rcl} P_D & = & \text{Pr(choisir } H_1 \text{ /} H_1 \text{ vraie)} \\ & = & \text{Pr}(n \geq \gamma_I / H_1 \text{ vraie}) \\ & = & \sum_{n=\gamma_I}^{+\infty} \text{Pr}(n/H_1) \\ & = & 1 - \sum_{n=0}^{\gamma_I - 1} \text{Pr}(n/H_1) \\ & = & 1 - \sum_{n=0}^{\gamma_I - 1} \frac{m_1^n}{n!} \exp(-m_1). \end{array} \tag{2.84}$$

De façon similaire, on obtient :

$$\begin{array}{rcl} P_F & = & \text{Pr(choisir } H_1 \text{ /} H_0 \text{ vraie)} \\ & = & \text{Pr}(n \geq \gamma_I / H_0 \text{ vraie}) \\ & = & \sum_{n=\gamma_I}^{+\infty} \text{Pr}(n/H_0) \\ & = & 1 - \sum_{n=0}^{\gamma_I - 1} \text{Pr}(n/H_0) \\ & = & 1 - \sum_{n=0}^{\gamma_I - 1} \frac{m_0^n}{n!} \exp(-m_0). \end{array} \tag{2.85}$$

En conséquence, P_D et P_F sont fonctions de la variable entière γ_I, et ne prennent qu'un nombre infini dénombrable de valeurs. On ne peut plus parler de courbes COR, puisque dans le plan (P_D, P_F), les performances sont caractérisées par des points isolés $(P_D(\gamma_I), P_F(\gamma_I))$ (Figure 2.6).

A priori, il est en général impossible de fixer une valeur de P_F^*, puisqu'il n'existe en général pas d'entier k tel que $P_F(k) = P_F^*$. Les test de Neyman-Pearson sont donc *a priori* difficiles.

Notion de test aléatoire

Pour surmonter ce problème, on peut introduire la notion de test aléatoire, qui aura un sens en moyenne si l'on réalise un grand nombre de tests.

Supposons que l'on veuille obtenir une probabilité P_F comprise entre $P_F(i)$ et $P_F(i+1)$. Il existe un nombre $0 < p < 1$ tel que :

$$P_F = pP_F(i+1) + (1-p)P_F(i). \tag{2.86}$$

Pour atteindre en moyenne P_F, on propose de faire le test avec $\gamma_I = i+1$ avec la probabilité p et et avec $\gamma_I = i$ avec la probabilité $1-p$. Ce test *aléatoire* permet d'atteindre en moyenne n'importe quelle valeur. La probabilité P_D vaut alors :

$$P_D = pP_D(i+1) + (1-p)P_D(i). \tag{2.87}$$

Ce test aléatoire réalise une interpolation linéaire par morceaux du test discret initial. Il n'a un sens qu'en moyenne, si on a les moyens de prendre une décision en réalisant un grand nombre de tests élémentaires.

2.9.4 Propriétés des courbes COR

Les probabilités de fausse alarme et de détection, à d fixé, sont des fonction du seuil (positif ou nul) de Bayes η. Pour des variables aléatoires continues (exemples 1 et 2), on peut les écrire sous la forme d'intégrales :

$$\begin{array}{rcl} P_F(\eta) & = & \int_\eta^{+\infty} p_{\Lambda/H_0}(l/H_0)dl, \\ P_D(\eta) & = & \int_\eta^{+\infty} p_{\Lambda/H_1}(l/H_1)dl, \end{array} \tag{2.88}$$

où $p_{\Lambda/H_i}(l/H_i$ représente la densité du rapport de vraisemblance $\Lambda(\mathbf{r})$ sachant l'hypothèse H_i. A partir de ces équations, on peut déduire les propriétés suivantes des courbes COR.

Propriété 1. Ce sont donc des fonctions continues de η, monotones et décroissantes, avec les limites suivantes, en zéro :

$$\begin{array}{rcl} \lim_{\eta\to 0} P_F(\eta) & = & 1, \\ \lim_{\eta\to 0} P_D(\eta) & = & 1, \end{array} \tag{2.89}$$

et à l'infini :

$$\begin{array}{rcl} \lim_{\eta\to \infty} P_F(\eta) & = & 0, \\ \lim_{\eta\to \infty} P_D(\eta) & = & 0. \end{array} \tag{2.90}$$

Propriété 2. Dans le plan (P_F, P_D), les courbes COR sont convexes, car $P_D \geq P_F$.

Propriété 3. Tous les tests du rapport de vraisemblance ont des courbes COR situées au dessus de la droite $P_D = P_F$.

Propriété 4. Dans le plan (P_F, P_D), en tout point d'une courbe COR, la pente de la tangente est égale à la valeur du seuil η correspondant à ce point.

En effet, en dérivant les équations (2.88) par rapport à η, on a :

$$\begin{array}{rcl} dP_F(\eta)/d\eta & = & -p_{\Lambda/H_0}(\eta/H_0), \\ dP_D(\eta)/d\eta & = & -p_{\Lambda/H_1}(\eta/H_1). \end{array} \tag{2.91}$$

La tangente à la courbe (P_F, P_D) s'écrit alors :

$$\frac{dP_D(\eta)/d\eta}{dP_F(\eta)/d\eta} = \frac{p_{\Lambda/H_1}(\eta/H_0)}{p_{\Lambda/H_0}(\eta/H_1)}. \tag{2.92}$$

Notons maintenant $\Omega(\eta)$ l'ensemble :

$$\Omega(\eta) = \Big\{\mathbf{r}/\Lambda(\mathbf{r}) > \eta\Big\} = \Big\{\mathbf{r}/\frac{p_{\mathbf{r}/H_1}(\mathbf{r}/H_1)}{p_{\mathbf{r}/H_0}(\mathbf{r}/H_0)} > \eta\Big\}. \tag{2.93}$$

On peut alors exprimer $P_D(\eta)$ sous la forme :

$$\begin{array}{rcl} P_D(\eta) & = & \int_{\Omega(\eta)} p_{\mathbf{r}/H_1}(\mathbf{r}/H_1)\mathbf{d}r, \\ & = & \int_{\Omega(\eta)} \Lambda(\mathbf{r}) p_{\mathbf{r}/H_0}(\mathbf{r}/H_0)\mathbf{d}r. \end{array} \tag{2.94}$$

En utilisant la définition de $\Omega(\eta)$, on peut écrire :

$$P_D(\eta) = \int_{\Omega(\eta)} \Lambda(\mathbf{r}) p_{\mathbf{r}/H_0}(\mathbf{r}/H_0)\mathbf{d}r = \int_{\eta}^{+\infty} l p_{\Lambda/H_0}(l/H_0)dl. \tag{2.95}$$

En dérivant la dernière équation de (2.95) par rapport à η, on trouve :

$$dP_D(\eta)/d\eta = -\eta p_{\Lambda/H_0}(\eta/H_0). \tag{2.96}$$

En reportant dans (2.92), on arrive au résultat prévu.

Propriété 5. Lorsque la valeur maximale du risque de Bayes est atteinte pour une valeur de P_1 intérieure à l'intervalle $]0, 1[$, le point de fonctionnement MINIMAX est l'intersection de la droite d'équation (condition MINIMAX) :

$$(C_{11} - C_{00}) + (C_{01} - C_{11})(1 - P_D) - (C_{10} - C_{00})P_F = 0, \tag{2.97}$$

avec la courbe COR appropriée (valeur de d).

Dans le cas où $C_{00} = C_{11} = 0$, en notant $C_M = C_{01}$ et $C_F = C_{10}$, l'équation MINIMAX se réduit à :

$$P_D = 1 - \frac{C_F}{C_M} P_F. \tag{2.98}$$

Si on a en plus $C_M = C_F = 1$ (c'est-à-dire globalement $C_{ij} = 1 - \delta_{ij}$), on a simplement :

$$P_D = 1 - P_F. \tag{2.99}$$

La figure 2.7 montre les points des courbes COR associés aux conditions MINIMAX, pour quelques valeurs de C_F et C_M.

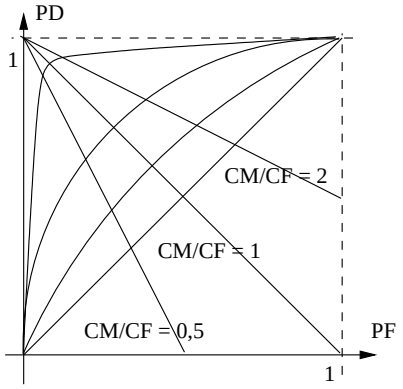


FIGURE 2.7 – Performances dans le cas de tests MINIMAX. Pour d fixé, les performances (P_D, P_F) sont les points d'intersection entre la courbe COR correspondante et la droite associée à la condition MINIMAX.

2.10 Résumé sur la détection binaire

Trois résultats importants doivent être compris concernant la détection binaire.

— En utilisant un critère de Bayes ou de Neyman-Pearson, on trouve que le test optimal est un test du rapport de vraisemblance. Quelle que soit la dimension de l'espace d'observation, ce test consiste à comparer une variable aléatoire à une dimension à un seuil scalaire.

— La mise en œuvre d'un test du rapport de vraisemblance est simplifiée par la détermination de la statistique suffisante. D'un point de vue géométrique, on peut interpréter cette statistique comme la coordonnée (de l'espace d'observation) qui contient toutes les informations nécessaires à la décision.

— On peut décrire les performances d'un test du rapport de vraisemblance en traçant les courbes COR (courbes $P_D(P_F)$ pour différentes valeurs du seuil η. En pratique, il est inutile (et coûteux en calcul) de tracer la courbe complète : la valeur correspondant au test suffit.

Peu de formules sont à savoir par cœur, hormis le rapport de vraisemblance $\Lambda(\mathbf{r}) = p(\mathbf{r}/H_1)/p(\mathbf{r}/H_0)$, les définitions des probabilités de détection et de fausse alarme. Le seuil de Bayes et les autres quantités sont disponibles dans le cours.

Chapitre 3

Détection non binaire

Il s'agit maintenant de concevoir un récepteur capable de décider d'une hypothèse parmi M : c'est le cas M-aire. Puisqu'il y a M hypothèses possibles et M décisions possibles, on a au total M^2 situations de la forme : *choisir H_i alors que H_j est vraie*.

On peut facilement étendre au cas M-aire le risque de Bayes introduite en décision binaire. En revanche, la mise en œuvre d'un critère de Neyman-Pearson n'est pas utilisée en pratique. C'est pourquoi nous ne développerons dans ce chapitre que le critère de Bayes, d'abord succintement dans le cas M-aire, puis de façon plus détaillée dans le cas ternaire ($M = 3$).

3.1 Critère de Bayes dans le cas M-aire

On associe à chacune des M^2 situations "*choisir H_i alors que H_j est vraie*" un coût C_{ij}. En notant les probabilités *a priori* P_i et les densités conditionnelles $p(\mathbf{r}/H_i)$, le risque de Bayes est le coût moyen :

$$\mathcal{R}_{Bayes} = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} P_j C_{ij} \text{Pr(décider } H_i/H_j \text{ vraie)} \tag{3.1}$$

En notant Z l'espace d'observation, la décision consiste à déterminer la partition $(Z_0, Z_1, \ldots, Z_{M-1})$ de Z qui minimise le risque de Bayes, que l'on peut aussi écrire :

$$\mathcal{R}_{Bayes} = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} P_j C_{ij} \int_{Z_i} p(\mathbf{r}/H_j)\mathbf{dr} \tag{3.2}$$

On peut transformer l'équation précédente :

$$\mathcal{R}_{Bayes} = \sum_{i=0}^{M-1} \int_{Z_i} \left[\sum_{j=0}^{M-1} P_j C_{ij}.p(\mathbf{r}/H_j) \right] \mathbf{dr}. \tag{3.3}$$

En notant I_j les intégrandes des intégrales sur Z_j, on minimise le risque en suivant la procédure :

- Calculer les I_j, $j = 1, \ldots, M-1$,
- Choisir l'hypothèse H_{i_0}, d'indice i_0 tel que I_{i_0} est la plus petite des intégrandes I_j, $j = 1, \ldots, M-1$, autrement dit tel que $i_0 = \text{Argmin}_j I_j$.

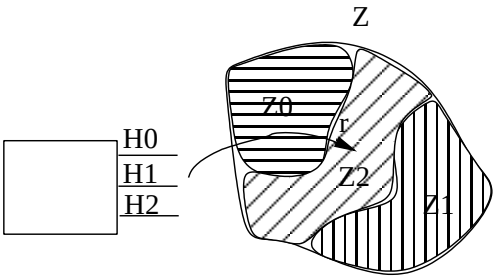


FIGURE 3.1 – Détection ternaire. La règle de décision consiste à partitionner l'espace d'observation Z en trois régions de façon à minimiser le risque.

3.2 Critère de Bayes dans le cas ternaire

Dans le cas ternaire ($M = 3$), on a :

$$\mathcal{R}_{Bayes} = \sum_{i=0}^{2}\sum_{j=0}^{2} P_j C_{ij} \int_{Z_i} p(\mathbf{r}/H_j)\mathbf{dr}, \tag{3.4}$$

En utilisant le fait que les Z_i forment une partition de Z, on peut écrire :

$$\begin{array}{l}\mathcal{R}_{Bayes} = \\ P_0[C_{00}\int_{Z\setminus(Z_1\bigcup Z_2)} p(\mathbf{r}/H_0)\mathbf{dr} + C_{10}\int_{Z_1} p(\mathbf{r}/H_0)\mathbf{dr} + C_{20}\int_{Z_2} p(\mathbf{r}/H_0)\mathbf{dr}] \\ +P_1[C_{01}\int_{Z_0} p(\mathbf{r}/H_1)\mathbf{dr} + C_{11}\int_{Z\setminus(Z_0\bigcup Z_2)} p(\mathbf{r}/H_1)\mathbf{dr} + C_{21}\int_{Z_2} p(\mathbf{r}/H_1)\mathbf{dr}] \\ +P_2[C_{02}\int_{Z_0} p(\mathbf{r}/H_2)\mathbf{dr} + C_{12}\int_{Z_1} p(\mathbf{r}/H_1)\mathbf{dr} + C_{22}\int_{Z\setminus(Z_0\bigcup Z_1)} p(\mathbf{r}/H_2)\mathbf{dr}].\end{array} \tag{3.5}$$

En isolant les termes constants (obtenus en intégrant sur Z) et en regroupant les termes dans les intégrales sur Z_i, on a :

$$\begin{array}{rcl}\mathcal{R}_{Bayes} & = & C_{00}P_0 + C_{11}P_1 + C_{22}P_2 \\ & + & \int_{Z_0}[P_2(C_{02}-C_{22})p(\mathbf{r}/H_2) + P_1(C_{01}-C_{11})p(\mathbf{r}/H_1)]\mathbf{dr} \\ & + & \int_{Z_1}[P_0(C_{10}-C_{00})p(\mathbf{r}/H_0) + P_2(C_{12}-C_{22})p(\mathbf{r}/H_2)]\mathbf{dr} \\ & + & \int_{Z_2}[P_0(C_{20}-C_{00})p(\mathbf{r}/H_0) + P_1(C_{21}-C_{11})p(\mathbf{r}/H_1)]\mathbf{dr}.\end{array} \tag{3.6}$$

Dans cette expression, la première ligne correspond à un coût fixe, les trois suivantes à des coûts variables, selon le choix des domaines Z_j. Notons I_j les trois intégrandes (termes entre crochets) dans les intégrales sur Z_j. Pour minimiser le risque de Bayes, on compare simplement les intégrandes et on choisit la décision H_j associée à l'intégrande I_j la plus petite. Autrement dit :

$$\begin{array}{llll}\text{si } I_0(\mathbf{r}) < I_1(\mathbf{r}) & \text{et} & I_0(\mathbf{r}) < I_2(\mathbf{r}), & \text{on choisit } H_0, \\ \text{si } I_1(\mathbf{r}) < I_0(\mathbf{r}) & \text{et} & I_1(\mathbf{r}) < I_2(\mathbf{r}), & \text{on choisit } H_1, \\ \text{si } I_2(\mathbf{r}) < I_0(\mathbf{r}) & \text{et} & I_2(\mathbf{r}) < I_1(\mathbf{r}), & \text{on choisit } H_2.\end{array} \tag{3.7}$$

3.3 Test dans le cas ternaire

A partir des expressions précédentes, on peut introduire deux rapports de vraisemblance :

$$\begin{array}{rcl}\Lambda_1(\mathbf{r}) & = & \frac{p(\mathbf{r}/H_1)}{p(\mathbf{r}/H_0)}, \\ \Lambda_2(\mathbf{r}) & = & \frac{p(\mathbf{r}/H_2)}{p(\mathbf{r}/H_0)}.\end{array} \tag{3.8}$$

En reportant dans (3.6) et en divisant par $p(\mathbf{r}/H_0)$, les inégalités (3.7) peuvent alors s'exprimer dans le plan (Λ_1, Λ_2).

Commençons par l'inégalité $I_0(\mathbf{r}) < I_1(\mathbf{r})$:

$$\begin{aligned} P_2(C_{02} - C_{22})\Lambda_2(\mathbf{r}) + P_1(C_{01} - C_{11})\Lambda_1(\mathbf{r}) &< P_0(C_{10} - C_{00}) + P_2(C_{12} - C_{22})\Lambda_2(\mathbf{r}) \\ \text{d'où :} & \\ P_1(C_{01} - C_{11})\Lambda_1(\mathbf{r}) &< P_0(C_{10} - C_{00}) + P_2(C_{12} - C_{02})\Lambda_2(\mathbf{r}). \end{aligned} \tag{3.9}$$

De la même façon, on obtient pour $I_0(\mathbf{r}) < I_2(\mathbf{r})$:

$$P_2(C_{02} - C_{22})\Lambda_2(\mathbf{r}) < P_0(C_{20} - C_{00}) + P_1(C_{21} - C_{01})\Lambda_1(\mathbf{r}), \tag{3.10}$$

et pour $I_1(\mathbf{r}) < I_2(\mathbf{r})$:

$$P_2(C_{12} - C_{22})\Lambda_2(\mathbf{r}) < P_0(C_{20} - C_{00}) + P_1(C_{21} - C_{11})\Lambda_1(\mathbf{r}). \tag{3.11}$$

Les trois autres inégalités : $I_0(\mathbf{r}) > I_1(\mathbf{r})$, $I_0(\mathbf{r}) > I_2(\mathbf{r})$ et $I_1(\mathbf{r}) > I_2(\mathbf{r})$, sont obtenus en changeant $<$ par $>$ dans les trois relations précédentes. Chacune des inégalités ci-dessus permet donc de décider d'une hypothèse parmi deux, la troisième ne jouant aucun rôle, c'est-à-dire qu'on ne peut rien dire. Autrement dit, le test peut être résumé par les trois inégalités :

$$\begin{cases} P_1(C_{01} - C_{11})\Lambda_1(\mathbf{r}) & \begin{matrix} \mathrm{H}_1 \text{ (ou } \mathrm{H}_2) \\ \gtrless \\ \mathrm{H}_0 \text{ (ou } \mathrm{H}_2) \end{matrix} & P_0(C_{10} - C_{00}) + P_2(C_{12} - C_{02})\Lambda_2(\mathbf{r}), \\ P_2(C_{02} - C_{22})\Lambda_2(\mathbf{r}) & \begin{matrix} \mathrm{H}_2 \text{ (ou } \mathrm{H}_1) \\ \gtrless \\ \mathrm{H}_0 \text{ (ou } \mathrm{H}_1) \end{matrix} & P_0(C_{20} - C_{00}) + P_1(C_{21} - C_{01})\Lambda_1(\mathbf{r}), \\ P_2(C_{12} - C_{22})\Lambda_2(\mathbf{r}) & \begin{matrix} \mathrm{H}_2 \text{ (ou } \mathrm{H}_0) \\ \gtrless \\ \mathrm{H}_1 \text{ (ou } \mathrm{H}_0) \end{matrix} & P_0(C_{20} - C_{10}) + P_1(C_{21} - C_{11})\Lambda_1(\mathbf{r}). \end{cases} \tag{3.12}$$

3.4 Représentation graphique dans le plan (Λ_2, Λ_1)

Dans le plan (Λ_2, Λ_1), il est facile de voir que les frontières des régions sont délimitées par les droites d'équation :

$$\begin{cases} P_1(C_{01} - C_{11})\Lambda_1(\mathbf{r}) &= P_0(C_{10} - C_{00}) + P_2(C_{12} - C_{02})\Lambda_2(\mathbf{r}), \\ P_2(C_{02} - C_{22})\Lambda_2(\mathbf{r}) &= P_0(C_{20} - C_{00}) + P_1(C_{21} - C_{01})\Lambda_1(\mathbf{r}), \\ P_2(C_{12} - C_{22})\Lambda_2(\mathbf{r}) &= P_0(C_{20} - C_{10}) + P_1(C_{21} - C_{11})\Lambda_1(\mathbf{r}). \end{cases} \tag{3.13}$$

On obtient alors les régions de décision de la figure (3.2).

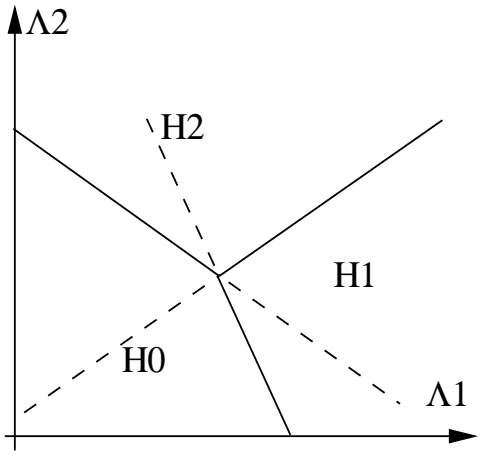


FIGURE 3.2 – Régions de décision dans le plan $(\Lambda_1(\mathbf{r}), \Lambda_2(\mathbf{r}))$.

3.4.1 Représentation graphique dans le cas particulier $C_{ij} = 1 - \delta_{ij}$

Dans ce cas, les équations (3.12) se simplifient :

$$\begin{cases} P_1\Lambda_1(\mathbf{r}) & \underset{\mathrm{H}_0 \text{ (ou } \mathrm{H}_2)}{\overset{\mathrm{H}_1 \text{ (ou } \mathrm{H}_2)}{\gtrless}} & P_0, \\ P_2\Lambda_2(\mathbf{r}) & \underset{\mathrm{H}_0 \text{ (ou } \mathrm{H}_1)}{\overset{\mathrm{H}_2 \text{ (ou } \mathrm{H}_1)}{\gtrless}} & P_0, \\ P_2\Lambda_2(\mathbf{r}) & \underset{\mathrm{H}_1 \text{ (ou } \mathrm{H}_0)}{\overset{\mathrm{H}_2 \text{ (ou } \mathrm{H}_0)}{\gtrless}} & P_1\Lambda_1(\mathbf{r}). \end{cases} \tag{3.14}$$

ou encore, en prenant les logarithmes :

$$\begin{cases} \ln\Lambda_1(\mathbf{r}) & \underset{\mathrm{H}_0 \text{ (ou } \mathrm{H}_2)}{\overset{\mathrm{H}_1 \text{ (ou } \mathrm{H}_2)}{\gtrless}} & \ln\frac{P_0}{P_1}, \\ \ln\Lambda_2(\mathbf{r}) & \underset{\mathrm{H}_0 \text{ (ou } \mathrm{H}_1)}{\overset{\mathrm{H}_2 \text{ (ou } \mathrm{H}_1)}{\gtrless}} & \ln\frac{P_0}{P_2}, \\ \ln\Lambda_2(\mathbf{r}) & \underset{\mathrm{H}_1 \text{ (ou } \mathrm{H}_0)}{\overset{\mathrm{H}_2 \text{ (ou } \mathrm{H}_0)}{\gtrless}} & \ln\Lambda_1(\mathbf{r}) + \ln\frac{P_1}{P_2}. \end{cases} \tag{3.15}$$

Dans le plan (Λ_2, Λ_1), comme dans le plan $(\ln\Lambda_2, \ln\Lambda_1)$, les régions de décision définies par ces équation sont limitées par des droites, très simples. Les deux représentations sont données aux figures 3.3 et 3.4.

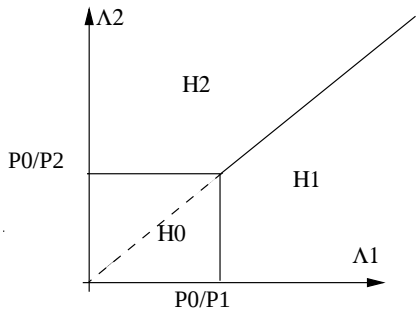


FIGURE 3.3 – Régions de décision dans le plan $(\Lambda_1(\mathbf{r}), \Lambda_2(\mathbf{r}))$, pour le cas particulier $C_{ij} = 1 - \delta_{ij}$.

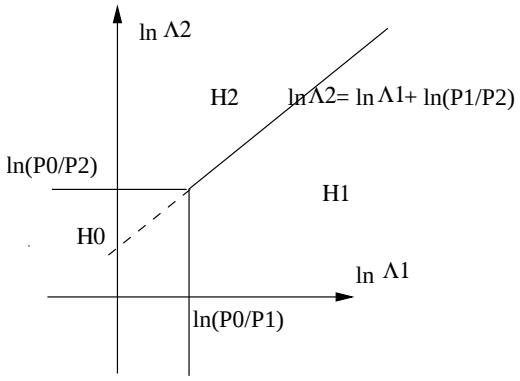


FIGURE 3.4 – Régions de décision dans le plan $(\ln \Lambda_1(\mathbf{r}), \ln \Lambda_2(\mathbf{r}))$, pour le cas particulier $C_{ij} = 1 - \delta_{ij}$.

3.4.2 Interprétation des équations dans le cas $C_{ij} = 1 - \delta_{ij}$

Les équations (3.14), quoique fort simples, sont encore difficiles à interpréter. En remplaçant Λ_1 et Λ_2 par leurs définitions et en multipliant par $p(\mathbf{r}/H_0)$, on arrive à :

$$\begin{cases} P_1 p(\mathbf{r}/H_1) & \underset{\mathrm{H_0\ (ou\ H_2)}}{\overset{\mathrm{H_1\ (ou\ H_2)}}{\gtrless}} & P_0 p(\mathbf{r}/H_0), \\ P_2 p(\mathbf{r}/H_2) & \underset{\mathrm{H_0\ (ou\ H_1)}}{\overset{\mathrm{H_2\ (ou\ H_1)}}{\gtrless}} & P_0 p(\mathbf{r}/H_0), \\ P_2 p(\mathbf{r}/H_2) & \underset{\mathrm{H_1\ (ou\ H_0)}}{\overset{\mathrm{H_2\ (ou\ H_0)}}{\gtrless}} & P_1 p(\mathbf{r}/H_1). \end{cases} \tag{3.16}$$

En appliquant le théorème de Bayes à chaque probabilité conditionnelle : $P_i p(\mathbf{r}/H_i) = \mathrm{Pr}(H_i/\mathbf{r}) p(\mathbf{r})$, on obtient finalement :

$$\begin{cases} \mathrm{Pr}(H_1/\mathbf{r}) & \underset{\mathrm{H_0\ (ou\ H_2)}}{\overset{\mathrm{H_1\ (ou\ H_2)}}{\gtrless}} & \mathrm{Pr}(H_0/\mathbf{r}), \\ \mathrm{Pr}(H_2/\mathbf{r}) & \underset{\mathrm{H_0\ (ou\ H_1)}}{\overset{\mathrm{H_2\ (ou\ H_1)}}{\gtrless}} & \mathrm{Pr}(H_0/\mathbf{r})), \\ \mathrm{Pr}(H_2/\mathbf{r}) & \underset{\mathrm{H_1\ (ou\ H_0)}}{\overset{\mathrm{H_2\ (ou\ H_0)}}{\gtrless}} & \mathrm{Pr}(H_1/\mathbf{r}). \end{cases} \tag{3.17}$$

Ces équations permettent une interprétation simple et logique du test : en effet, on choisit l'hypothèse H_i dont la probabilité sachant les observations, $\mathrm{Pr}(H_i/\mathbf{r})$ (c'est la probabilité *a posteriori*), est la plus grande.

3.5 Résumé sur l'estimation ternaire

La décision optimale au sens du risque de Bayes met en évidence les deux points suivants :

— La dimension de l'espace de décision est inférieure ou égale à 2. Les frontières des régions de décisions sont des droites dans le plan (Λ_2, Λ_1).

— Le test optimal est généralement facile à trouver. Dans le cas où l'on cherche à minimiser la probabilité d'erreur totale (cas où $C_{ij} = 1 - \delta_{ij}$), le test consiste à choisir l'hypothèse H_i dont la probabilité *a posteriori* $\mathrm{Pr}(H_i/\mathbf{r})$ est la plus grande.

3.6 Extension au cas M-aire

Les résultats se généralisent facilement au cas M-aire. D'une façon générale, le décision, quelle que soit la dimension k de l'espace d'observation, s'effectue dans un espace de décision de dimension inférieure ou égale à $M-1$. Dans cet espace, les frontières entre les régions de décision sont des hyperplans.

Deuxième partie

Théorie de l'estimation

Définitions et position du problème

Dans la partie précédente, le problème consistait à choisir une hypothèse parmi 2 (cas binaire) ou M (dans le cas général), de façon la plus vraisemblable en minimisant une fonction de coût. Dans cette partie, nous voulons aller plus loin et prédire la valeur de paramètres.

Par exemple, on veut mesurer une tension v. D'après la physique du système, on peut par exemple supposer que cette tension est comprise entre $[-V, +V]$. La mesure est entâchée d'un bruit n que l'on peut modéliser par un bruit additif, gaussien, de moyenne nulle et de variance σ_n^2. On observe donc :

$$r = v + n. \tag{3.18}$$

La densité de l'observation, étant donné le paramètre inconnu v, notée $p(r/v)$ peut alors s'écrire :

$$p(r/v) = p_N(r - v) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\Big(- \frac{(r-v)^2}{2\sigma_n^2}\Big). \tag{3.19}$$

Le problème est de prédire la valeur de v à partir de l'observation r.

Le principe général de l'estimation peut se résumer par les 4 points essentiels :

— Espace des paramètres : la sortie de la source est une variable que l'on peut considérer comme un point dans l'espace des paramètres. Dans le cas d'un seul paramètre, cet espace est *a priori* l'axe réel $-\infty < v < +\infty$, ou une partie de l'axe réel $-V \leq v \leq +V$ si des informations permettent de réduire la gamme des valeurs possibles.

— Espace d'observation : c'est en général un espace de dimension k finie. Chaque mesure ou observation est un point $\mathbf{r}$ de cet espace.

— Loi de probabilité : elle décrit la relation probabiliste $p(\mathbf{r}/v)$ entre l'observation et le paramètre v.

— Règle d'estimation : c'est la relation qui permet de prédire v, sous la forme d'un estimateur $\hat{v}(\mathbf{r})$ dépendant de l'observation $\mathbf{r}$.

Par rapport à la théorie de la détection, la nouveauté réside dans l'espace des paramètres et la règle d'estimation.

L'objectif de ce chapitre est de proposer quelques méthodes générales pour l'estimation de paramètres. Les observations étant perturbées par du bruit, la formalisation du problème repose encore sur un modèle statistique. On distinguera deux approches selon que le paramètre à estimer est aléatoire ou déterministe.

Organisation

Cette seconde partie est organisée en trois chapitres : un chapitre consacré à l'estimation d'un paramètre aléatoire, un autre pour l'estimation d'un paramètre déterministe et un dernier sur l'estimation de paramètres multiples.

Chapitre 4

Estimation d'un paramètre aléatoire

Dans ce chapitre, le paramètre à estimer est une variable aléatoire notée a. L'observation est un vecteur de dimension k, noté $\mathbf{r}$.

4.1 Principe et fonctions de coût

Essayons d'abord d'étendre l'idée du risque de Bayes au problème d'estimation.

Dans le problème de détection M-aire, le risque de Bayes était construit en associant un coût C_{ij} à chaque situation : *décider H_i alors que H_i est vraie*. Aux M^2 situations possibles, on pouvait donc associer la matrice de coût $\mathbf{C} = (C_{ij})$.

Dans le problème d'estimation, le paramètre a et son estimée $\hat{a}(\mathbf{r})$ sont des variables continues. Pour chaque paire $(a, \hat{a}(\mathbf{r}))$, on peut donc associer un coût $C(a, \hat{a}(\mathbf{r}))$, qui est une fonction de deux variables.

Généralement, on considère, et c'est réaliste, que le coût dépend uniquement de l'erreur d'estimation $e(\mathbf{r}) = \hat{a}(\mathbf{r}) - a$. La fonction de coût se réduit alors à une fonction d'une seule variable, l'erreur $e(\mathbf{r})$. Dans la suite, on supposera que cette fonction est à valeurs positives.

Quelques fonctions de coût classiques (Figure 4.1) sont :

1. La fonction de coût quadratique : $C_{ls}(e(\mathbf{r})) = (\hat{a}(\mathbf{r}) - a)^2$,
2. La fonction de coût *valeur absolue* : $C_{abs}(e(\mathbf{r})) = |\hat{a}(\mathbf{r}) - a|$,
3. La fonction de coût uniforme :

$$C_{unif}(e(\mathbf{r})) = \begin{cases} 0 & \text{si } |e(\mathbf{r})| \leq \frac{\Delta}{2}, \\ 1 & \text{si } |e(\mathbf{r})| > \frac{\Delta}{2}. \end{cases} \tag{4.1}$$

Le rôle de chaque fonction de coût est :
— de mesurer la qualité de l'estimation,
— d'aboutir à une solution en minimisant la fonction de coût.

Dans le problème de détection, le risque de Bayes permettait de mesurer un coût moyen, à partir des coût C_{ij}, des probabilités *a priori* P_i et des densités conditionnelles. Dans le problème d'estimation, on étend cette idée pour mesurer l'erreur moyenne : le coût est remplacé par la fonction

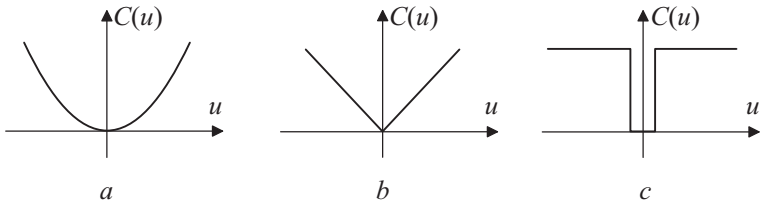


FIGURE 4.1 – Exemples de quelques fonctions de coût : (a) coût quadratique, (b) coût valeur absolue, (c) coût uniforme.

de coût $C(e(\mathbf{r}))$, la probabilité *a priori* P_i est remplacé par la densité de probabilité du paramètre (variable aléatoire) a. On a alors :

$$\mathcal{R} = E[C(e(\mathbf{r}))] = \int_{\mathcal{A}} da \int_{\mathbb{R}^k} C(\hat{a}(\mathbf{r}) - a) p(a, \mathbf{r}) \mathbf{dr}. \tag{4.2}$$

Cette équation montre que la moyenne (opérateur $E[.]$) porte à la fois sur toutes les valeurs de a (première intégrale, simple) et sur toutes les valeurs $\mathbf{r}$ (seconde intégrale, multiple - k) de l'espace d'observation.

La densité de probabilité conjointe, $p(a, \mathbf{r})$, peut être factorisée en appliquant le théorème de Bayes : $p(a, \mathbf{r}) = p(\mathbf{r})p(a/\mathbf{r})$. En supposant la convergence uniforme des intégrales, on peut intervertir les deux intégrales, ce qui conduit à la relation :

$$\mathcal{R} = \int_{\mathbb{R}^k} p(\mathbf{r}) \Big[\int_{\mathcal{A}} C(\hat{a}(\mathbf{r}) - a) p(a/\mathbf{r}) da \Big] \mathbf{dr}. \tag{4.3}$$

Dans cette expression, les termes $p(\mathbf{r})$ et $C(\hat{a}(\mathbf{r}) - a)p(a/\mathbf{r})$ étant positifs, la minimisation du risque $\mathcal{R}$ sera obtenue en minimisant simplement l'intégrale intérieure. Le meilleur estimateur $\hat{a}(\mathbf{r})$, au sens de la fonction de coût $C(.)$, est donc obtenu en calculant la valeur de $\hat{a}(\mathbf{r})$ pour laquelle l'intégrale intérieure est minimale.

Dans les paragraphes suivants, nous allons détailler le calcul pour les trois fonctions de coût introduites plus haut.

4.2 Calcul pour le coût *quadratique*

Le risque, noté $\mathcal{R}_{LS}$ (LS pour *least square*), s'écrit :

$$\mathcal{R}_{ls} = \int_{\mathbb{R}^k} p(\mathbf{r}) \Big[\int_{\mathcal{A}} (\hat{a}(\mathbf{r}) - a)^2 p(a/\mathbf{r}) da \Big] \mathbf{dr}. \tag{4.4}$$

Pour minimiser l'intégrale intérieure, $I_{ls}(\mathbf{r})$, par rapport $\hat{a}(\mathbf{r})$, nous calculons la dérivée de cette intégrale par rapport à $\hat{a}(\mathbf{r})$ [1] :

$$\begin{array}{rcl} \frac{d}{d\hat{a}} I_{ls}(\mathbf{r}) & = & \frac{d}{d\hat{a}} \int_{\mathcal{A}} (\hat{a}(\mathbf{r}) - a)^2 p(a/\mathbf{r}) da \\ & = & 2 \int_{\mathcal{A}} (\hat{a}(\mathbf{r}) - a) p(a/\mathbf{r}) da \\ & = & 2 \int_{\mathcal{A}} \hat{a}(\mathbf{r}) p(a/\mathbf{r}) da - 2 \int_{\mathcal{A}} a p(a/\mathbf{r}) da \\ & = & 2\hat{a}(\mathbf{r}) \int_{\mathcal{A}} p(a/\mathbf{r}) da - 2 \int_{\mathcal{A}} a p(a/\mathbf{r}) da \\ & = & 2\hat{a}(\mathbf{r}) - 2 \int_{\mathcal{A}} a p(a/\mathbf{r}) da. \end{array} \tag{4.5}$$

1. Dans la suite, pour simplifier les notations, $d\hat{a}(\mathbf{r})$ sera simplement noté $d\hat{a}$ dans les dérivations

En utilisant la dernière relation, on déduit que la valeur $\hat{a}$ qui annule la dérivée vérifie :

$$\hat{a}_{ls}(\mathbf{r}) = \int_{\mathcal{A}} a p(a/\mathbf{r}) da. \tag{4.6}$$

En dérivant une seconde fois I_{ls}, on trouve $\frac{d^2 I_{ls}}{d\hat{a}^2} = 2 > 0$, ce qui montre que la valeur trouvée est un minimum. Cet estimateur, appelé estimateur des moindres carrés (en abrégé : LS pour least square, ou parfois MS pour mean square) est noté $\hat{a}_{ls}(\mathbf{r})$ pour bien montrer qu'il est relatif à la fonction de coût quadratique. On remarque que l'estimateur des moindres carrés est simplement la moyenne conditionnelle ou moyenne *a posteriori*. En revenant à la relation (4.4), on en déduit que $\mathcal{R}_{ls}$ est alors la variance conditionnelle ou variance a posteriori sur l'espace d'observation.

4.3 Calcul pour le coût *erreur absolue*

Le risque, noté $\mathcal{R}_{abs}$, s'écrit :

$$\mathcal{R}_{abs} = \int_{\mathbb{R}^k} p(\mathbf{r}) \Big[\int_{\mathcal{A}} |\hat{a}(\mathbf{r}) - a| p(a/\mathbf{r}) da \Big] \mathbf{dr}. \tag{4.7}$$

Pour traiter la valeur absolue, on décompose l'intégrale intérieure, $I_{abs}(\mathbf{r})$, en deux intégrales :

$$I_{abs}(\mathbf{r}) = \int_{-\infty}^{\hat{a}(\mathbf{r})} (\hat{a}(\mathbf{r}) - a) p(a/\mathbf{r}) da - \int_{\hat{a}(\mathbf{r})}^{+\infty} (\hat{a}(\mathbf{r}) - a) p(a/\mathbf{r}) da. \tag{4.8}$$

En différentiant par rapport à $\hat{a}$, variable qui apparaît dans les bornes des deux intégrales et dans les deux intégrales elles-mêmes, on a :

$$\frac{dI_{abs}}{d\hat{a}}(\mathbf{r}) = \int_{-\infty}^{\hat{a}(\mathbf{r})} p(a/\mathbf{r}) da - \int_{\hat{a}(\mathbf{r})}^{+\infty} p(a/\mathbf{r}) da. \tag{4.9}$$

La valeur $\hat{a}_{abs}(\mathbf{r})$, qui annule cette expression, vérifie :

$$\int_{-\infty}^{\hat{a}_{abs}(\mathbf{r})} p(a/\mathbf{r}) da = \int_{\hat{a}_{abs}(\mathbf{r})}^{+\infty} p(a/\mathbf{r}) da. \tag{4.10}$$

C'est la médiane de la densité *a posteriori*.

4.4 Calcul pour le coût *uniforme*

Le risque, noté $\mathcal{R}_{unif}$, s'écrit :

$$\mathcal{R}_{unif} = \int_{\mathbb{R}^k} p(\mathbf{r}) \Big[\int_{\mathcal{A}} C_{unif}(\hat{a}(\mathbf{r}) - a) p(a/\mathbf{r}) da \Big] \mathbf{dr}, \tag{4.11}$$

que l'on peut écrire, en tenant compte de la définition (4.1) :

$$\mathcal{R}_{unif} = \int_{\mathbb{R}^k} p(\mathbf{r}) \Big[1 - \int_{\hat{a}_{unif}(\mathbf{r})-\Delta/2}^{\hat{a}_{unif}(\mathbf{r})+\Delta/2} p(a/\mathbf{r}) da \Big] \mathbf{dr}. \tag{4.12}$$

Pour minimiser le risque, on minimise le terme entre crochets, ce qui revient à maximiser l'intégrale $I_{unif}(\mathbf{r})$:

$$I_{unif}(\mathbf{r}) = \int_{\hat{a}_{unif}(\mathbf{r})-\Delta/2}^{\hat{a}_{unif}(\mathbf{r})+\Delta/2} p(a/\mathbf{r})da. \tag{4.13}$$

Lorsque $\Delta \to 0$, $I_{unif}(\mathbf{r})$ est maximale pour la valeur de $a = \hat{a}_{unif}(\mathbf{r})$ telle que $p(a/\mathbf{r})$ est maximale. L'estimateur $\hat{a}_{unif}(\mathbf{r})$ est donc le maximum de la densité *a posteriori*. On le note généralement $\hat{a}_{map}(\mathbf{r})$, avec *map* pour "maximum *a posteriori*", notation que nous utiliserons dans la suite.

4.5 Equation du maximum a posteriori (MAP)

L'estimateur $\hat{a}_{map}(\mathbf{r})$ est donc la solution de l'équation :

$$\frac{\partial p(a/\mathbf{r})}{\partial a} = 0. \tag{4.14}$$

La fonction logarithme étant strictement monotone, on utilise fréquemment :

$$\frac{\partial \ln p(a/\mathbf{r})}{\partial a} = 0. \tag{4.15}$$

Cette dernière équation est souvent la plus pratique, notamment en raison des formes produits fréquentes de $p(a/\mathbf{r})$, obtenues avec le théorème de Bayes ou à partir de mesures indépendantes.

Les solutions de cette équation donnent bien entendu tous les extréma de $p(a/\mathbf{r})$, maxima et minima. S'il n'y a qu'une seule solution, il faut vérifier qu'il s'agit bien d'un maximum. S'il y en a plusieurs, il faut rechercher le *maximum maximorum*.

Ainsi, en utilisant le théorème de Bayes :

$$p(a/\mathbf{r}) = \frac{p(\mathbf{r}/a)p(a)}{p(\mathbf{r})}, \tag{4.16}$$

on a :

$$\ln p(a/\mathbf{r}) = \ln p(\mathbf{r}/a) + \ln p(a) - \ln p(\mathbf{r}). \tag{4.17}$$

Puisque l'on cherche le maximum sur a, le dernier terme ne joue aucun rôle. L'estimateur MAP est donc le maximum de la quantité :

$$l(a) = \ln p(\mathbf{r}/a) + \ln p(a). \tag{4.18}$$

Cette expression montre le rôle des données (premier terme de droite) et le rôle de l'information *a priori* sur le paramètre a (second terme à droite).

L'équation MAP est alors :

$$\frac{\partial l(a)}{\partial a}\Big|_{a=\hat{a}_{map}(\mathbf{r})} = \frac{\partial \ln p(\mathbf{r}/a)}{\partial a}\Big|_{a=\hat{a}_{map}(\mathbf{r})} + \frac{\partial \ln p(a)}{\partial a}\Big|_{a=\hat{a}_{map}(\mathbf{r})} = 0. \tag{4.19}$$

4.6 Exemple

4.6.1 Enoncé

On réalise un ensemble de k mesures regroupées dans le vecteur $\mathbf{r}$. Chaque mesure est de la forme $r_i = a + n_i$, où a est un paramètre inconnu et n_i sont des échantillons indépendants et identiquement distribués (iid) selon une loi gaussienne : $n_i \sim N(0, \sigma_n^2)$. La connaissance *a priori* sur le paramètre a est résumée par sa densité de probabilité : $a \sim N(0, \sigma_a^2)$. On peut donc écrire :

$$p(a) = \frac{1}{\sqrt{2\pi}\sigma_a} \exp\,(\,-\,\frac{a^2}{2\sigma_a^2}), \tag{4.20}$$

et pour une observation r_i :

$$p(r_i/a) = p_n(r_i - a) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\Big(\,-\,\frac{(r_i-a)^2}{2\sigma_n^2}\Big). \tag{4.21}$$

Puisque les bruits n_i sont indépendants, on peut écrire :

$$\begin{array}{rcl} p(\mathbf{r}/a) & = & \prod_{i=1}^k p(r_i/a) \\ & = & \prod_{i=1}^k p_n(r_i - a) \\ & = & \prod_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma_n} \exp\Big(\,-\,\frac{(r_i-a)^2}{2\sigma_n^2}\Big). \end{array} \tag{4.22}$$

On se propose de calculer les trois estimateurs étudiés ci-dessus : l'estimateur LS, l'estimateur ABS et l'estimateur MAP.

4.6.2 Calcul de $\hat{a}_{ls}(\mathbf{r})$

On rappelle que $\hat{a}_{ls}(\mathbf{r})$ est la moyenne *a posteriori* (de la densité *a posteriori*) :

$$\hat{a}_{ls}(\mathbf{r}) = \int_{-\infty}^{+\infty} a p(a/\mathbf{r}) da. \tag{4.23}$$

Pour calculer cet estimateur, il faut donc calculer la densité *a posteriori* $p(a/\mathbf{r})$. En utilisant les hypothèses (4.20) et (4.22), et en appliquant le théorème de Bayes, on peut écrire :

$$\begin{array}{rcl} p(a/\mathbf{r}) & = & \frac{p(\mathbf{r}/a)p(a)}{p(\mathbf{r})} \\ & = & \frac{1}{p(\mathbf{r})} \frac{1}{\sqrt{2\pi}\sigma_a} \exp\,(\,-\,\frac{a^2}{2\sigma_a^2}) \prod_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma_n} \exp\,(\,-\,\frac{(r_i-a)^2}{2\sigma_n^2}). \\ & = & \frac{1}{p(\mathbf{r})} \frac{1}{\sqrt{2\pi}\sigma_a} (\prod_{i=1}^k \sqrt{2\pi}\sigma_n) \exp\Big[\,-\,\frac{1}{2}\Big(\frac{a^2}{\sigma_a^2} + \sum_{i=1}^k \frac{(r_i-a)^2}{\sigma_n^2}\Big)\Big] \\ & = & f(\mathbf{r}) \exp\Big[\,-\,\frac{1}{2}\Big(\frac{a^2}{\sigma_a^2} + \sum_{i=1}^k \frac{(r_i-a)^2}{\sigma_n^2}\Big)\Big]. \end{array} \tag{4.24}$$

Dans la dernière expression, nous avons regroupé tous les termes ne dépendants pas de a (donc fonction de $\mathbf{r}$ ou constants) dans un terme noté $f(\mathbf{r})$. Dans la suite, nous allons traiter uniquement le terme exponentiel que l'on note T, en développant les termes quadratiques et en factorisant ceux qui ne dépendent pas de a :

$$\begin{array}{rcl} T & = & \exp\Big[\,-\,\frac{1}{2}\Big(\frac{a^2}{\sigma_a^2} + \sum_{i=1}^k \frac{(r_i^2 - 2ar_i + a^2}{\sigma_n^2}\Big)\Big] \\ & = & \exp\Big[\,-\,\frac{1}{2}\Big(\frac{a^2}{\sigma_a^2} + \frac{ka^2}{\sigma_n^2} + \sum_{i=1}^k \frac{r_i^2}{\sigma_n^2} - 2a \sum_{i=1}^k \frac{r_i}{\sigma_n^2}\Big)\Big]. \end{array} \tag{4.25}$$

En posant $k/\sigma_n^2 + 1/\sigma_a^2 = 1/\sigma_p^2$ et en factorisant, T s'écrit alors :

$$T = \exp\Big[-\frac{1}{2\sigma_p^2}\Big(a^2 - \frac{2a\sigma_p^2}{\sigma_n^2}\sum_{i=1}^{k} r_i + \frac{\sigma_p^2}{\sigma_n^2}\sum_{i=1}^{k} r_i^2\Big)\Big]. \tag{4.26}$$

On remarque que le terme entre parenthèses est le début du développement d'un terme au carré de la forme $(a-u)^2$, que l'on explicite :

$$T = \exp\Big[-\frac{1}{2\sigma_p^2}\Big\{\Big(a - \frac{\sigma_p^2}{\sigma_n^2}\sum_{i=1}^{k} r_i\Big)^2 - \Big(\frac{\sigma_p^2}{\sigma_n^2}\sum_{i=1}^{k} r_i\Big)^2 + \frac{\sigma_p^2}{\sigma_n^2}\sum_{i=1}^{k} r_i^2\Big\}\Big]. \tag{4.27}$$

Les deux derniers termes ne dépendant pas de a, on peut les sortir de l'exponentielle et les regrouper dans le terme $f(\mathbf{r})$. Finalement, la densité *a posteriori* s'écrit :

$$p(a/\mathbf{r}) = f(\mathbf{r}) \exp\Big[-\frac{1}{2\sigma_p^2}\Big(a - \frac{\sigma_p^2}{\sigma_n^2}\sum_{i=1}^{k} r_i\Big)^2\Big]. \tag{4.28}$$

Le terme $f(\mathbf{r})$ est un simple terme de normalisation, tel que $\int p(a/\mathbf{r})da = 1$. Son expression exacte n'a pas ici d'importance, dans la mesure où l'on remarque que $p(a/\mathbf{r})$ a la forme typique d'une densité gaussienne :

$$p(a/\mathbf{r}) \sim N\Big(\frac{\sigma_p^2}{\sigma_n^2}\sum_{i=1}^{k} r_i, \sigma_p^2\Big). \tag{4.29}$$

Puisque l'on sait que l'estimateur des moindres carrés, $\hat{a}_{ls}(\mathbf{r})$ est la moyenne *a posteriori*, on a directement :

$$\hat{a}_{ls}(\mathbf{r}) = \frac{\sigma_p^2}{\sigma_n^2}\sum_{i=1}^{k} r_i, \tag{4.30}$$

que l'on mettra sous la forme finale :

$$\hat{a}_{ls}(\mathbf{r}) = \frac{\sigma_a^2}{\sigma_a^2 + \frac{\sigma_n^2}{k}}\Big(\frac{1}{k}\sum_{i=1}^{k} r_i\Big). \tag{4.31}$$

Remarques.

— Cette expression met en évidence l'existence d'une statistique suffisante, comme dans les problèmes de détection. En effet, toute l'information sur les mesures est résumée dans leur somme. Cette remarque est très importante pour la mise en œuvre pratique, logicielle ou matérielle, de l'estimateur.

— La règle d'estimation (4.31) utilise de façon pertinente l'information disponible.
Si $\sigma_a^2 \ll \frac{\sigma_n^2}{k}$, c'est-à-dire lorsque la variance de a est très petite par rapport à celle du bruit, l'information *a priori* donnée par $p(a)$ est bien plus précise que les mesures. Puisque $a \sim N(0, \sigma_a^2)$, a doit tendre vers 0, qui est la limite de (4.31) pour $\sigma_a^2 \ll \frac{\sigma_n^2}{k}$. L'estimateur ne tient pas du tout compte des mesures r_i.
Au contraire, si $\sigma_a^2 \gg \frac{\sigma_n^2}{k}$, l'information *a priori* est très floue (a possède une très grande variance) par rapport aux mesures r_i. La limite de (4.31) sous cette condition est alors :

$$\hat{a}_{ls}(\mathbf{r}) \longrightarrow \frac{1}{k}\sum_{i=1}^{k} r_i, \tag{4.32}$$

et ne tient aucun compte de l'information *a priori* sur a.

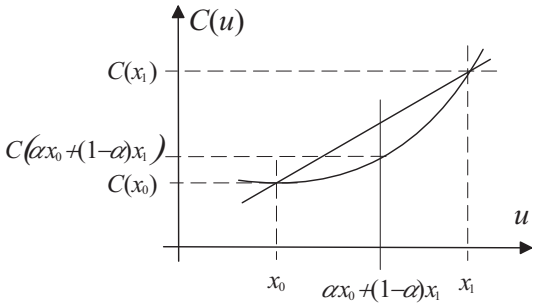


FIGURE 4.2 – Fonction convexe.

4.6.3 Calcul de $\hat{a}_{abs}(\mathbf{r})$ et de $\hat{a}_{map}(\mathbf{r})$

A partir de la densité *a posteriori* (4.28), on peut aussi déduire facilement ces deux estimateurs.

D'après (4.10), l'estimateur $\hat{a}_{abs}(\mathbf{r})$ n'est autre que la médiane de la densité *a posteriori*. Pour une densité gaussienne, la médiane est égale à la moyenne, donc :

$$\hat{a}_{abs}(\mathbf{r}) = \hat{a}_{ls}(\mathbf{r}) = \frac{\sigma_a^2}{\sigma_a^2 + \frac{\sigma_n^2}{k}} \Big(\frac{1}{k} \sum_{i=1}^{k} r_i \Big). \tag{4.33}$$

D'après (4.10), l'estimateur $\hat{a}_{unif}(\mathbf{r})$ n'est autre que la valeur qui correspond au maximum de la densité *a posteriori*. Pour une densité gaussienne, le maximum est obtenu pour la moyenne, donc :

$$\hat{a}_{map}(\mathbf{r}) = \hat{a}_{ls}(\mathbf{r}) = \frac{\sigma_a^2}{\sigma_a^2 + \frac{\sigma_n^2}{k}} \Big(\frac{1}{k} \sum_{i=1}^{k} r_i \Big). \tag{4.34}$$

Dans cet exemple, les trois estimateurs coïncident. En général, ceci n'est pas vrai, et l'estimateur peut dépendre du critère de coût utilisé. Remarquons enfin que l'estimateur $\hat{a}_{map}(\mathbf{r})$ peut être obtenu directement de $p(a/\mathbf{r})$ (4.24), par simple dérivation par rapport à a.

4.7 Invariance de l'estimateur

Dans l'exemple prédédent, les estimateurs $\hat{a}_{ls}(\mathbf{r})$, $\hat{a}_{abs}(\mathbf{r})$ et $\hat{a}_{map}(\mathbf{r})$) étaient identiques. On peut se demander à quelles conditions, on a invariance des estimateurs.

On donnera dans ce paragraphe deux propositions qui conduisent à l'invariance, mais sans démonstration.

Proposition 4.7.1 *Si la fonction de coût, $C(x)$, est paire (C1) et convexe (Figure 4.2 c'est-à-dire à concavité tournée vers le haut (C2), et si la densité* a posteriori $p(a/\mathbf{r})$ *est symétrique par rapport à sa moyenne (la moyenne conditionnelle) (C3), l'estimateur $\hat{a}(\mathbf{r})$ minimisant la fonction de coût vérifie $\hat{a}(\mathbf{r}) = \hat{a}_{ls}(\mathbf{r})$.*

D'un point de vue mathématique, ces conditions s'écrivent :
— (C1) : $C(x) = C(-x)$,
— (C2) : $\forall (x_0, x_1) \in \mathbb{R}^2$, $C(\alpha x_0 + (1-\alpha)x_1) \leq \alpha C(x_0) + (1-\alpha)C(x_1)$,
— (C3) : $\forall u \in \mathbb{R}, p((m-u)/\mathbf{r}) = p((m+u)/\mathbf{r})$.

Essayons d'appliquer cette proposition à l'exemple précédent. Prenons d'abord la fonction de coût $C(x) = |x|$. On vérifie facilement les conditions (C1) et (C2). De plus, la densité *a posteriori*, qui est gaussienne, vérifie la condition (C3). On peut donc conclure que $\hat{a}_{abs}(\mathbf{r}) = \hat{a}_{ls}(\mathbf{r})$.

En revanche, on ne peut pas utiliser cette proposition pour conclure en ce qui concerne l'estimateur $\hat{a}_{map}(\mathbf{r})$, car (C2) n'est pas vérifiée.

Proposition 4.7.2 *Si la fonction de coût, $C(x)$, est paire (C1) et non décroissante pour $x > 0$ (C2), et si la densité* a posteriori $p(a/\mathbf{r})$ *est unimodale, symétrique par rapport à sa moyenne (la moyenne conditionnelle) (C3) et vérifie (C4) :*

$$\lim_{x \longrightarrow +\infty} C(x)p(x/\mathbf{r}) = 0, \tag{4.35}$$

alors l'estimateur $\hat{a}(\mathbf{r})$ minimisant la fonction de coût vérifie $\hat{a}(\mathbf{r}) = \hat{a}_{ls}(\mathbf{r})$.

On peut appliquer cette proposition dans l'exemple prédédent :

— $C(x) = |x|$: (C1), (C2) et C(3) sont vérifiées. De plus, $\lim_{x\to+\infty} xp(x/\mathbf{r}) = 0$, donc (C4) est aussi vérifiée. On peut donc appliquer la proposition et conclure $\hat{a}_{abs}(\mathbf{r}) = \hat{a}_{ls}(\mathbf{r})$.

— Estimateur MAP, associé au coût uniforme : (C1), C(2) et (C3) sont vérifiées. De plus, (C4) est aussi vérifiée, car $\lim_{x\to+\infty} p(x/\mathbf{r}) = 0$. On peut donc appliquer la proposition et conclure : $\hat{a}_{map}(\mathbf{r}) = \hat{a}_{ls}(\mathbf{r})$.

Ces deux propositions sont importantes, en particulier parce qu'elles s'appliquent pour une large gamme de fonctions de coût et de densités *a posteriori*. Ces propositions justifient l'intérêt de l'estimateur des moindres carrés, puisqu'asymptotiquement, en raison de la loi des grands nombres, dans de nombreux problèmes, la densité *a posteriori* tendra vers une loi gaussienne, et par conséquent de nombreux critères conduiront à un estimateur égal à celui des moindres carrés.

4.8 Exemple d'une observation non linéaire

4.8.1 Enoncé

La variable inconnue à estimer a est mesurée au travers d'un système non linéaire qui fournit :

$$r_i = g(a) + n_i, \quad i = 1, \dots k. \tag{4.36}$$

Comme dans l'exemple précédent, les n_i sont de échantillons iid, gaussiens (de loi $N(0, \sigma_n^2)$, et la densité *a priori* de a suit une loi $N(0, \sigma_a^2)$.

4.8.2 Solution

En utilisant les calculs (4.24), la densité *a posteriori* est alors :

$$p(a/\mathbf{r}) = f(\mathbf{r}) \exp\Big[-\frac{1}{2}\Big(\frac{a^2}{\sigma_a^2} + \sum_{i=1}^{k} \frac{(r_i - g(a))^2}{\sigma_n^2}\Big)\Big]. \tag{4.37}$$

Si la fonction $g(a)$ n'est pas connue, on ne peut pas aller plus loin.

L'équation MAP peut être écrite facilement en maximisant $l(a) = \ln p(\mathbf{r}/a) + \ln p(a)$ (4.18). On a ici :

$$l(a) = -\frac{1}{2}\Big(\frac{a^2}{\sigma_a^2} + \sum_{i=1}^{k} \frac{(r_i - g(a))^2}{\sigma_n^2}\Big). \tag{4.38}$$

L'estimateur MAP est donc solution de :

$$\frac{\partial l(a)}{\partial a} = -\frac{1}{2}\Big[\frac{2a}{\sigma_a^2} + \frac{1}{\sigma_n^2}\sum_{i=1}^{k} 2(r_i - g(a))\Big(-\frac{\partial g(a)}{\partial a}\Big)\Big] = 0, \tag{4.39}$$

d'où :

$$\hat{a}_{map}(\mathbf{r}) = \frac{\sigma_a^2}{\sigma_n^2}\sum_{i=1}^{k}(r_i - g(a))\frac{\partial g(a)}{\partial a}\Big|_{a=\hat{a}_{map}(\mathbf{r})}. \tag{4.40}$$

En général, cette équation ne peut pas être résolue analytiquement, même lorsque g est connue.

4.9 Estimation d'une loi de Poisson

4.9.1 Enoncé

Le nombre d'événements n d'une expérience suit une loi de Poisson de paramètre a inconnu. La probabilité d'observer n événements est donc :

$$\text{Pr}(n \text{ événements}/a) = \frac{a^n}{n!}\exp(-a). \tag{4.41}$$

On suppose que le paramètre a est distribué selon une loi exponentielle unilatérale :

$$p(a) = \begin{cases} \lambda\exp(-\lambda a), & \text{si } a > 0, \\ 0, & \text{sinon.} \end{cases} \tag{4.42}$$

On veut estimer le paramètre a à partir du nombre n.

4.9.2 Solution

En utilisant le théorème de Bayes, la densité *a posteriori* s'écrit :

$$\begin{aligned} p(a/n) &= \frac{\text{Pr}(n/a)p(a)}{\text{Pr}(n)} \\ &= \frac{a^n}{n!}\exp(-a)\lambda\exp(-\lambda a)\frac{1}{\text{Pr}(n)} \\ &= f(n)a^n\exp[-(\lambda+1)a] \end{aligned} \tag{4.43}$$

où $f(n)$ est un terme de normalisation tel que :

$$\begin{aligned} \int_0^{+\infty} p(a/n)da &= 1, \\ f(n)\int_0^{+\infty} a^n\exp[-(\lambda+1)a]da &= 1, \\ f(n) &= \Big[\int_0^{+\infty} a^n\exp[-(\lambda+1)a]da\Big]^{-1} \end{aligned} \tag{4.44}$$

Pour calculer $f(n)$, nous devons calculer l'intégrale :

$$I_n = \int_0^{+\infty} a^n\exp[-(\lambda+1)a]da. \tag{4.45}$$

Par intégration par partie, on établit facilement une relation de récurrence :

$$\begin{array}{rcl} I_n & = & \Big[-\frac{1}{\lambda+1} a^n \exp\Big(-(\lambda+1)a \Big) \Big]_0^{+\infty} + \frac{n}{\lambda+1} \int_0^{+\infty} a^{n-1} \exp[-(\lambda+1)a] da \\ & = & \frac{n}{\lambda+1} I_{n-1} \\ & = & \frac{n}{\lambda+1} \frac{n-1}{\lambda+1} I_{n-2} \\ & = & \dots \\ & = & \frac{n!}{(\lambda+1)^n} I_0. \end{array} \tag{4.46}$$

Pour finir, calculons I_0 :

$$\begin{array}{rcl} I_0 & = & \int_0^{+\infty} a^0 \exp[-(\lambda+1)a] da, \\ & = & \int_0^{+\infty} \exp[-(\lambda+1)a] da, \\ & = & \Big[-\frac{\exp\Big(-(\lambda+1)a\Big)}{\lambda+1} \Big]_0^{+\infty}, \\ & = & \frac{1}{\lambda+1}. \end{array} \tag{4.47}$$

En combinant (4.47) et (4.46), on trouve :

$$I_n = \frac{n!}{(\lambda+1)^{n+1}}, \tag{4.48}$$

d'où :

$$f(n) = \frac{(\lambda+1)^{n+1}}{n!}, \tag{4.49}$$

et la densité

$$p(a/n) = \frac{(\lambda+1)^{n+1}}{n!} a^n \exp[-(\lambda+1)a]. \tag{4.50}$$

Estimateur LS. On peut maintenant calculer l'estimateur des moindres carrés. Puisque c'est la moyenne de la densité *a posteriori*, on a :

$$\begin{array}{rcl} \hat{a}_{ls}(n) & = & E[a/n], \\ & = & \int_0^{+\infty} a\, p(a/n) da, \\ & = & \frac{(\lambda+1)^{n+1}}{n!} \int_0^{+\infty} a\, a^n \exp[-(\lambda+1)a] da, \\ & = & \frac{(\lambda+1)^{n+1}}{n!} I_{n+1} \end{array} \tag{4.51}$$

En utilisant la relation de récurrence (4.48), on trouve immédiatement :

$$\begin{array}{rcl} \hat{a}_{ls}(n) & = & \frac{(\lambda+1)^{n+1}}{n!} \frac{(n+1)!}{(\lambda+1)^{n+2}} \\ & = & \frac{n+1}{\lambda+1}. \end{array} \tag{4.52}$$

Estimateur MAP. On peut aussi calculer l'estimateur MAP. Pour cela, nous calculons le maximum de la densité *a posteriori*. Prenons d'abord le logarithme :

$$\begin{array}{rcl} \ln p(a/n) & = & \ln \Big[f(n) a^n \exp[-(\lambda+1)a] \Big] \\ & = & \ln f(n) + n \ln a - (\lambda+1)a, \end{array} \tag{4.53}$$

puis calculons la dérivée par rapport à a :

$$\begin{array}{rcl} \frac{\partial \ln p(a/n)}{\partial a} & = & \frac{\partial}{\partial a} \Big[n \ln a - (\lambda+1)a \Big], \\ & = & \frac{n}{a} - (\lambda+1). \end{array} \tag{4.54}$$

L'estimateur MAP correspond à la solution de l'équation :

$$\begin{array}{rcl} \frac{n}{a} - (\lambda + 1) & = & 0\Big|_{a=\hat{a}_{map}(n)} \\ \text{soit : } \hat{a}_{map}(n) & = & \frac{n}{\lambda+1} \end{array} \tag{4.55}$$

4.9.3 Remarques

Dans cet exemple, les estimateurs des moindres carrés et MAP sont formellement différents. La différence est importante pour n petit. De plus, l'estimateur MAP est plus simple à obtenir, car il ne requiert pas le calcul explicite de $f(n)$. De façon générale, on remarque que l'estimateur MAP n'exige pas un calcul complet de la densité $p(a/n)$.

4.10 Résumé de l'estimation de paramètres aléatoires

On retiendra que l'estimation repose sur la minimisation d'une fonction de coût. Deux estimateurs très usuels ont été présentés : l'estimateur des moindres carrés (LS) et l'estimateur du maximum *a posteriori* (MAP).

— L'estimateur des moindres carrés, c'est-à-dire minimisant l'erreur quadratique moyenne et noté $\hat{a}_{ls}(\mathbf{r})$, est toujours la moyenne de la densité de probabilité *a posteriori* $p(a/\mathbf{r})$ parfois appelé moyenne condionnelle ou moyenne *a posteriori*.

— L'estimateur du maximum *a posteriori*, noté $\hat{a}_{map}(\mathbf{r})$, est la valeur de a pour laquelle la densité de probabilité *a posteriori* $p(a/\mathbf{r})$ est maximale. En pratique, l'estimateur MAP est obtenu en cherchant le maximum de $p(a/\mathbf{r})$ ou de $\ln p(a/\mathbf{r})$.

— Pour une large classe de fonctions de coût, l'estimateur optimal est égal à l'estimateur des moindres carrés $\hat{a}_{ls}(\mathbf{r})$, pourvu que la densité *a posteriori* $p(a/\mathbf{r})$ satisfasse quelques conditions simples (parité, unimodale, etc.). La distribution gaussienne satisfait en particulier ces conditions.

Chapitre 5

Estimation de paramètres déterministes

Il n'est pas possible de traiter l'estimation de paramètres déterministes comme celle de paramètres aléatoires. En effet, avec des paramètres déterministes, $p(a)$ et $p(a/\mathbf{r})$ n'ont plus de sens. Il faut en particulier trouver une formulation de la mesure de performance des estimateurs.

5.1 Principe et qualité de l'estimation

Puisque a est une variable déterministe, la première idée consisterait à calculer un critère de Bayes sous la forme :

$$\mathcal{R}_{Bayes}(a) = \int_{\mathbb{R}^k} (\hat{a}(\mathbf{r}) - a)^2 p(\mathbf{r}/a)\mathbf{dr}. \tag{5.1}$$

En minimisant le risque par rapport à $\hat{a}(\mathbf{r})$, on obtient $\hat{a}(\mathbf{r}) = a$. Mais a étant inconnu, ceci n'est pas applicable. Il faut donc trouver une autre mesure de la qualité de l'estimateur.

Chaque estimateur est réalisé à partir d'un jeu de mesures bruitées, l'observation $\mathbf{r}$. Chaque estimation est donc la réalisation d'une variable aléatoire, l'estimateur. Si on considère un ensemble (virtuel et arbitrairement grand) d'estimations, on peut mesurer :

— La moyenne :

$$E[\hat{a}(\mathbf{r})] = \int_{\mathbb{R}^k} \hat{a}(\mathbf{r}) p(\mathbf{r}/a)\mathbf{dr}, \tag{5.2}$$

pour évaluer l'écart (appelé biais) à la valeur vraie a. Si $E[\hat{a}(\mathbf{r})] = a$, le biais est nul et l'estimateur est dit non biaisé. Si $E[\hat{a}(\mathbf{r})] = a + b$ où b est une constante, l'estimation a un biais fixe. En estimant ce biais, il est facile de corriger l'estimateur. Si $E[\hat{a}(\mathbf{r})] = a + b(a)$, où $b(a)$ est une fonction de a, le biais devient plus complexe à compenser.

— La variance

$$\begin{aligned} \text{Var}(\hat{a}(\mathbf{r})) &= E\Big[\Big(\hat{a}(\mathbf{r}) - E[\hat{a}(\mathbf{r})]\Big)^2\Big], \\ &= \int_{\mathbb{R}^k} \Big(\hat{a}(\mathbf{r}) - E[\hat{a}(\mathbf{r})]\Big)^2 p(\mathbf{r}/a)\mathbf{dr}, \\ &= E[\hat{a}(\mathbf{r})^2] - E^2[\hat{a}(\mathbf{r})], \end{aligned} \tag{5.3}$$

qui mesure la dispersion des estimations autour de leur moyenne. La variance dépend généralement du nombre N de mesures utilisé pour réaliser une estimation, et décroît fréquemment en $1/N$.

La situation idéale serait d'avoir un biais et une variance nuls.

5.2 Maximum de vraisemblance

Reprenons l'exemple du paragraphe 4.6, où l'on observe une mesure (scalaire) $r = a + n$, dans laquelle $n \sim N(0, \sigma_n)$ mais a est maintenant déterministe. On peut donc écrire la densité conditionnelle de l'observation r sachant le paramètre a :

$$p(r/a) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\Big(- \frac{(r-a)^2}{2\sigma_n^2}\Big). \tag{5.4}$$

Choisissons comme estimation la valeur de a qui entraîne le plus vraisemblablement l'observation r, c'est-à-dire pour laquelle la densité est maximale. Dans (5.4), le terme exponentiel est maximal pour $r - a = 0$, c'est-à-dire pour $a = r$. Autrement dit, dans le cas d'une mesure avec un bruit additif comme c'est le cas ici, on retranche à r la valeur la plus probable du bruit n. C'est bien sûr 0, valeur pour laquelle $p_n(u)$ est maximale.

Cet estimateur est appelé estimateur du maximum de vraisemblance (en anglais *maximum likelihood*). Il est noté $\hat{a}_{ml}(r) = r$. La densité $p(r/a)$ est appelée fonction de vraisemblance. L'estimateur du maximum de vraisemblance, notée estimateur ML dans la suite, est donc la valeur de a qui maximise $p(r/a)$. Si le maximum est intérieur à la gamme de variation de a et si $\ln p(r/a)$ a une dérivée première continue, l'estimateur ML satisfait l'équation :

$$\frac{\partial \ln p(r/a)}{\partial a}\Big|_{a=\hat{a}_{ml}(r)} = 0. \tag{5.5}$$

On remarque que cet estimateur est similaire à l'estimateur MAP, dans le cas où le terme en $p(a)$ disparait, c'est-à-dire si la connaissance *a priori* sur a est nulle : $p(a)$ est constante.

5.3 Inégalités de Cramer-Rao

Pour mesurer les performances de l'estimateur, on calcule sa moyenne (pour déterminer le biais) et sa variance. La moyenne est généralement facile à calculer, mais la variance requiert souvent des calculs complexes. Les inégalités de Cramer-Rao, que nous allons énoncer et démontrer ci-dessous, fournissent plus facilement une borne inférieure de la variance.

5.3.1 Théorème

Nous énonçons et démontrons ci-dessous les inégalités de Cramer-Rao pour des estimateurs non biaisés. Ce résultat peut être étendu à des estimateurs biaisés (voir travaux dirigés). Ces inégalités ont d'abord été proposées par Fisher (1922) et par Dugué (1937). Elles ont été obtenues sous leur forme actuelle par Cramer (1946) et Rao (1945). Tous les estimateurs qui atteignent la borne (la variance est égale à la borne) sont appelés estimateurs efficaces.

Théorème 5.3.1 *Soit $\hat{a}(\mathbf{r})$ un estimateur quelconque non biaisé de a, tel que $p(\mathbf{r}/a)$ possède des dérivées partielles première et seconde par rapport à a absolument intégrables, on a alors*

$$\mathit{Var}[\hat{a}(\mathbf{r}) - a] \geq \Big\{ E\Big[\Big(\frac{\partial \ln p(\mathbf{r}/a)}{\partial a} \Big)^2 \Big] \Big\}^{-1}, \tag{5.6}$$

ou bien, ce qui est équivalent :

$$\mathit{Var}[\hat{a}(\mathbf{r}) - a] \geq -\Big\{ E\Big[\frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2} \Big] \Big\}^{-1}. \tag{5.7}$$

Remarque. Si l'estimateur $\hat{a}(\mathbf{r})$ est non biaisé, alors $E[\hat{a}(\mathbf{r})] = a$, et la variable aléatoire $\hat{a}(\mathbf{r}) - a$ est centrée. On a alors simplement : $\text{Var}[\hat{a}(\mathbf{r}) - a] = E[(\hat{a}(\mathbf{r}) - a)^2]$.

5.3.2 Démonstration de la première inégalité

Puisque l'estimateur $\hat{a}(\mathbf{r})$ est non biaisé, on a :

$$E[\hat{a}(\mathbf{r}) - a] = \int_{\mathbb{R}^k} (\hat{a}(\mathbf{r}) - a) p(\mathbf{r}/a) \mathbf{dr} = 0. \tag{5.8}$$

En dérivant par rapport à a, on obtient :

$$\begin{array}{rcl} \frac{\partial}{\partial a} \int_{\mathbb{R}^k} (\hat{a}(\mathbf{r}) - a) p(\mathbf{r}/a) \mathbf{dr} & = & 0 \\ \int_{\mathbb{R}^k} \frac{\partial}{\partial a} \Big\{ \hat{a}(\mathbf{r}) - a) p(\mathbf{r}/a) \Big\} \mathbf{dr} & = & 0 \\ - \int_{\mathbb{R}^k} p(\mathbf{r}/a) \mathbf{dr} + \int_{\mathbb{R}^k} \frac{\partial p(\mathbf{r}/a)}{\partial a} (\hat{a}(\mathbf{r}) - a) \mathbf{dr} & = & 0. \end{array} \tag{5.9}$$

La première intégrale de la dernière expression, qui intègre une densité sur tout son domaine, vaut évidemment 1. En remarquant par ailleurs que :

$$\frac{\partial p(\mathbf{r}/a)}{\partial a} = \frac{\partial \ln p(\mathbf{r}/a)}{\partial a} p(\mathbf{r}/a), \tag{5.10}$$

on peut modifier la seconde intégrale et écrire :

$$\begin{array}{rcl} \int_{\mathbb{R}^k} \frac{\partial \ln p(\mathbf{r}/a)}{\partial a} p(\mathbf{r}/a) (\hat{a}(\mathbf{r}) - a) \mathbf{dr} & = & 1 \\ \int_{\mathbb{R}^k} \Big[\frac{\partial \ln p(\mathbf{r}/a)}{\partial a} \sqrt{p(\mathbf{r}/a)} \Big] \Big[\sqrt{p(\mathbf{r}/a)} (\hat{a}(\mathbf{r}) - a) \Big] \mathbf{dr} & = & 1. \end{array} \tag{5.11}$$

On élève au carré la dernière expression, et en appliquant l'inégalité de Schwartz sur les intégrales :

$$\int A^2 dx \int B^2 dx \geq \Big[\int AB dx \Big]^2 \tag{5.12}$$

on arrive à :

$$\begin{array}{rcl} \Big\{ \int_{\mathbb{R}^k} \Big[\frac{\partial \ln p(\mathbf{r}/a)}{\partial a} \sqrt{p(\mathbf{r}/a)} \Big] \Big[\sqrt{p(\mathbf{r}/a)} (\hat{a}(\mathbf{r}) - a) \Big] \mathbf{dr} \Big\}^2 & = & 1 \\ \int_{\mathbb{R}^k} \Big[\frac{\partial \ln p(\mathbf{r}/a)}{\partial a} \sqrt{p(\mathbf{r}/a)} \Big]^2 \mathbf{dr} \int_{\mathbb{R}^k} \Big[\sqrt{p(\mathbf{r}/a)} (\hat{a}(\mathbf{r}) - a) \Big]^2 \mathbf{dr} & \geq & 1 \\ \Big[\int_{\mathbb{R}^k} \Big(\frac{\partial \ln p(\mathbf{r}/a)}{\partial a} \Big)^2 p(\mathbf{r}/a) \mathbf{dr} \Big] \Big[\int_{\mathbb{R}^k} (\hat{a}(\mathbf{r}) - a)^2 p(\mathbf{r}/a) \mathbf{dr} \Big] & \geq & 1 \end{array} \tag{5.13}$$

La seconde intégrale de cette équation n'est autre que la variance de $\hat{a}(\mathbf{r}) - a$. On arrive finalement à la première inégalité :

$$\int_{\mathbb{R}^k} (\hat{a}(\mathbf{r}) - a)^2 p(\mathbf{r}/a) \mathbf{dr} \geq \Big[\int_{\mathbb{R}^k} \Big(\frac{\partial \ln p(\mathbf{r}/a)}{\partial a} \Big)^2 p(\mathbf{r}/a) \mathbf{dr} \Big]^{-1} \tag{5.14}$$

On montre que l'égalité, c'est-à-dire la borne, est atteinte si et seulement si la densité vérifie :

$$\frac{\partial \ln p(\mathbf{r}/a)}{\partial a} = (\hat{a}(\mathbf{r}) - a) f(a), \tag{5.15}$$

où $f(a)$ est une fonction qui ne dépend pas des observations $\mathbf{r}$. Un estimateur qui vérifie cette condition est appelé *estimateur efficace* (en anglais *efficient*).

5.3.3 Démonstration de la seconde inégalité

Pour montrer la seconde inégalité, on remarque que $\int p(\mathbf{r}/a)\mathbf{dr} = 1$, que l'on dérive par rapport à a en tenant compte de (5.10) :

$$\begin{array}{rcl} \int_{\mathbb{R}^k} \frac{\partial p(\mathbf{r}/a)}{\partial a}\mathbf{dr} & = & 0 \\ \int_{\mathbb{R}^k} \frac{\partial \ln p(\mathbf{r}/a)}{\partial a} p(\mathbf{r}/a)\mathbf{dr} & = & 0. \end{array} \tag{5.16}$$

En dérivant de nouveau par rapport à a :

$$\begin{array}{rcl} \int_{\mathbb{R}^k} \frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2} p(\mathbf{r}/a)\mathbf{dr} + \int_{\mathbb{R}^k} \frac{\partial \ln p(\mathbf{r}/a)}{\partial a}\frac{\partial p(\mathbf{r}/a)}{\partial a}\mathbf{dr} & = & 0 \\ \int_{\mathbb{R}^k} \frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2} p(\mathbf{r}/a)\mathbf{dr} + \int_{\mathbb{R}^k} \left(\frac{\partial \ln p(\mathbf{r}/a)}{\partial a}\right)^2 p(\mathbf{r}/a)\mathbf{dr} & = & 0, \end{array} \tag{5.17}$$

et on obtient la relation :

$$E\Big[\Big(\frac{\partial \ln p(\mathbf{r}/a)}{\partial a}\Big)^2\Big] = -E\Big[\frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2}\Big]. \tag{5.18}$$

L'expression (5.18) montre l'égalité des bornes des deux inégalités et de (5.6) on déduit la seconde inégalité (5.7).

5.4 Remarques

Ce théorème mérite quelques commentaires.

— Toute estimation non biaisée a une variance plus *grande* qu'une certaine valeur. Malheureusement, dans le cas général, on ne sait pas si la variance est proche ou non de cette borne.

— Si (5.15) est vérifiée, l'estimateur $\hat{a}_{ml}(\mathbf{r})$ atteint la borne, c'est-à-dire :

$$\text{Var}[\hat{a}_{ml}(\mathbf{r}) - a] = \Big\{E\Big[\Big(\frac{\partial \ln p(\mathbf{r}/a)}{\partial a}\Big)^2\Big]\Big\}^{-1} = -\Big\{E\Big[\frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2}\Big]\Big\}^{-1}. \tag{5.19}$$

En effet, l'estimateur ML est solution de l'équation de vraisemblance :

$$\frac{\partial \ln p(\mathbf{r}/a)}{\partial a}\Big|_{a=\hat{a}_{ml}(\mathbf{r})} = f(a)(\hat{a}(\mathbf{r}) - a)\Big|_{a=\hat{a}_{ml}(\mathbf{r})} = 0. \tag{5.20}$$

Les solutions sont donc :

$$\hat{a}(\mathbf{r}) = \hat{a}_{ml}(\mathbf{r}), \tag{5.21}$$

ou

$$f(\hat{a}_{ml}) = 0. \tag{5.22}$$

Or, la solution doit bien sûr dépendre des observations $\mathbf{r}$. La solution (5.22) doit donc être éliminée. Ainsi, s'il existe un estimateur efficace, c'est l'estimateur ML, solution de l'équation de vraisemblance.

— S'il n'y a pas d'estimateur efficace, c'est-à-dire si $\partial \ln p(\mathbf{r}/a)/\partial a$ ne peut pas se mettre sous la forme $f(a)(\hat{a}(\mathbf{r}) - a)$, on ne peut pas conclure sur la qualité de l'estimateur ML. De plus, on ne sait pas comment sa variance s'approche de la borne.

— Avant d'utiliser la borne, il faut d'abord vérifier que l'estimateur est non biaisé. Ces bornes ne sont pas aplicables pour des estimateurs biaisés, pour lesquels des bornes similaires peuvent être calculées.

5.5 Variance d'un estimateur non biaisé et efficace

Le calcul de ce paragraphe est valable pour un estimateur non biaisé et efficace, c'est-à-dire tel que $\partial \ln p(\mathbf{r}/a)/\partial a = f(a)(\hat{a}(\mathbf{r}) - a)$. Calculons l'opposé de la dérivée seconde de $\ln p(\mathbf{r}/a)$:

$$-\frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2} = f(a) - \frac{df}{da}(a)(\hat{a}(\mathbf{r}) - a). \tag{5.23}$$

En prenant la moyenne :

$$E_{\mathbf{r}}\Big[-\frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2}\Big] = E_{\mathbf{r}}[f(a)] - \frac{df}{da}(a)E_{\mathbf{r}}\Big[(\hat{a}(\mathbf{r}) - a)\Big]. \tag{5.24}$$

Puisque l'estimateur est non biaisé, le dernier terme est nul. De plus, $f(a)$ ne dépend pas de $\mathbf{r}$ donc l'espérance de $f(a)$ sur $\mathbf{r}$ est égale à $f(a)$. Finalement :

$$E_{\mathbf{r}}\Big[-\frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2}\Big] = f(a). \tag{5.25}$$

En utilisant la première borne du théorème de Cramer-Rao, on a simplement :

$$\text{Var}(\hat{a}(r) - a) = f(a)^{-1}. \tag{5.26}$$

On peut donc énoncer le théorème suivant.

Théorème 5.5.1 *Soit $\hat{a}(\mathbf{r})$ un estimateur non biaisé et efficace tel que*

$$\frac{\partial \ln p(\mathbf{r}/a)}{\partial a} = f(a)(\hat{a}(\mathbf{r}) - a), \tag{5.27}$$

alors $\hat{a}(\mathbf{r}) = \hat{a}_{ml}(\mathbf{r})$ et Var$(\hat{a}(r) - a) = f(a)^{-1}$.

Ce théorème est très pratique mais il doit être utilisé rigoureusement. En effet, le terme $\partial \ln p(\mathbf{r}/a)/\partial a$ intervient d'abord dans l'équation de vraisemblance : $\partial \ln p(\mathbf{r}/a)/\partial a = 0$ pour laquelle le facteur $f(a)$ est sans importance. En revanche, pour l'utilisation du théorème, la forme $f(a)(\hat{a}(\mathbf{r}) - a)$ ne supporte aucune erreur de signe ou de facteur puisque la variance dépend directement du terme $f(a)$.

5.6 Applications des inégalités de Cramer-Rao

Dans ce paragraphe, nous montrons sur deux exemples comment calculer l'estimateur du maximum de vraisemblance et utiliser les inégalités de Cramer-Rao pour en déduire les performances de l'estimateur.

5.6.1 Paramètre avec bruit additif gaussien

Enoncé. On réalise un ensemble de k mesures regroupées dans le vecteur $\mathbf{r}$. Chaque mesure est de la forme $r_i = a + n_i$, où a est un paramètre déterministe inconnu et n_i sont des échantillons indépendants et identiquement distribués (iid) selon une loi gaussienne : $n_i \sim N(0, \sigma_n^2)$.

Estimateur du maximum de vraisemblance. Pour calculer l'estimateur ML, on doit calculer la densité $p(\mathbf{r}/a)$. Calculons d'abord la densité pour une observation r_i :

$$p(r_i/a) = p_n(r_i - a) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\Big(- \frac{(r_i - a)^2}{2\sigma_n^2}\Big). \tag{5.28}$$

Puisque les bruits n_i sont indépendants, on peut écrire :

$$\begin{array}{rcl} p(\mathbf{r}/a) & = & \prod_{i=1}^k p(r_i/a] \\ & = & \prod_{i=1}^k p_n(r_i - a] \\ & = & \prod_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma_n} \exp\Big(- \frac{(r_i-a)^2}{2\sigma_n^2}\Big). \end{array} \tag{5.29}$$

On calcule $\ln p(\mathbf{r}/a)$:

$$\ln p(\mathbf{r}/a) = -k \ln\Big(\sqrt{2\pi}\sigma_n\Big) + \sum_{i=1}^k \Big(- \frac{(r_i - a)^2}{2\sigma_n^2}\Big). \tag{5.30}$$

En dérivant par rapport à a :

$$\begin{array}{rcl} \frac{\partial \ln p(\mathbf{r}/a)}{\partial a} & = & \frac{1}{\sigma_n^2}(\sum_{i=1}^k (r_i - a)) \\ & = & \frac{k}{\sigma_n^2}\Big[\Big(\frac{1}{k}\sum_{i=1}^k r_i\Big) - a\Big] \end{array} \tag{5.31}$$

L'équation de vraisemblance :

$$\frac{k}{\sigma_n^2}\Big[\Big(\frac{1}{k}\sum_{i=1}^k r_i\Big) - a\Big]_{a=\hat{a}_{ml}(\mathbf{r})} = 0, \tag{5.32}$$

a pour solution :

$$\hat{a}_{ml}(\mathbf{r}) = \frac{1}{k}\sum_{i=1}^k r_i. \tag{5.33}$$

Biais de l'estimateur. Calculons d'abord le biais de (5.33). Pour cela, on calcule l'espérance de $\hat{a}_{ml}(\mathbf{r})$:

$$\begin{array}{rcl} E\Big[\frac{1}{k}\sum_{i=1}^k r_i\Big] & = & \frac{1}{k}\sum_{i=1}^k E[r_i] \\ & = & \frac{1}{k}\sum_{i=1}^k E[a + n_i] \\ & = & \frac{1}{k}\sum_{i=1}^k a \\ & = & a. \end{array} \tag{5.34}$$

L'estimateur est donc non biaisé.

L'estimateur est-il efficace ? Avant de calculer sa variance à l'aide du théorèe de Cramer-Rao, vérifions si l'estimateur est efficace. Pour cela, reprenons la dernière ligne de l'équation (5.31). On remarque que le terme de droite $(k/\sigma_n^2)[(1/k)\sum_{i=1}^k r_i - a]$ est de la forme $f(a)[\hat{a}_{ml}(\mathbf{r}) - a]$, avec $f(a) = k/\sigma_n^2$. On peut donc en déduire que l'estimateur est efficace.

De plus, en utilisant le théorème (5.5.1), on en déduit immédiatement la variance :

$$\text{Var}[\hat{a}_{ml}(\mathbf{r}) - a] = f(a)^{-1} = \frac{\sigma_n^2}{k}. \tag{5.35}$$

Calcul avec le théorème de Cramer-Rao. On peut aussi utiliser le théorème de Cramer-Rao. A titre d'exercice, nous allons ici utiliser les deux inégalités.

Pour la première, dérivons une seconde fois l'expression (5.31) :

$$\begin{array}{rcl} \frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2} & = & \frac{\partial}{\partial a}\Big[\frac{1}{\sigma_n^2}(\sum_{i=1}^k (r_i - a))\Big] \\ & = & -\frac{k}{\sigma_n^2}. \end{array} \tag{5.36}$$

En prenant l'inverse de l'opposé de l'espérance, on obtient directement la variance.

Pour la seconde inégalité, on calcule l'espérance du carré de $\frac{\partial \ln p(\mathbf{r}/a)}{\partial a}$:

$$\begin{array}{rcl} E\Big[\Big(\frac{\partial \ln p(\mathbf{r}/a)}{\partial a}\Big)^2\Big] & = & E\Big[(\frac{1}{\sigma_n^2}\sum_{i=1}^k (r_i - a))^2\Big] \\ & = & \frac{1}{\sigma_n^4} E\Big[(\sum_{i=1}^k n_i)^2\Big] \\ & = & \frac{1}{\sigma_n^4} k\sigma_n^2 \\ & = & \frac{k}{\sigma_n^2}. \end{array} \tag{5.37}$$

Pour le passage de la seconde à la troisième ligne (5.37), on utilise le fait que $n_i \sim N(0, \sigma_n^2)$, donc $\sum_{i=1}^k n_i$ est une variable aléatoire gaussienne $N(0, k\sigma_n^2)$. Donc $E\Big[(\sum_{i=1}^k n_i)^2\Big]$ n'est autre que la variance de $\sum_{i=1}^k n_i$, c'est-à-dire $k\sigma_n^2$. En prenant l'inverse du dernier terme, on obtient directement la variance.

On peut aussi faire un calcul direct de (5.37) sans utiliser le théorème de Cramer-Rao :

$$\begin{array}{rcl} E\Big[(\sum_i n_i)^2\Big] & = & E\Big[\sum_i n_i^2 + 2\sum_i \sum_{j\neq i} n_i n_j\Big], \\ & = & E\Big[\sum_i n_i^2\Big], \quad \text{car } n_i \text{ et } n_j \text{ sont indépendants et centrés} \\ & = & k\sigma_n^2. \end{array} \tag{5.38}$$

On peut noter que la mise en œuvre pratique de la seconde inégalité conduit souvent à des calculs plus complexes que la première.

5.6.2 Loi de Poisson

Enoncé. Le nombre d'événements n d'une expérience suit une loi de Poisson de paramètre déterministe a inconnu. On veut estimer le paramètre déterministe a à partir du nombre n.

Calcul de la fonction de vraisemblance. La fonction de vraisemblance est la probabilité d'observer n événements, le paramètre a étant donné :

$$\text{Pr}(n \text{ événements}/a) = \frac{a^n}{n!}\exp(-a). \tag{5.39}$$

Calculons le logarithme de la vraisemblance :

$$\ln \text{Pr}(n \text{ événements}/a) = n \ln a - \ln(n!) - a. \tag{5.40}$$

On en déduit l'équation de vraisemblance :

$$\begin{array}{rcl} \frac{\partial \ln \text{Pr}(n \text{ événements}/a)}{\partial a} & = & 0 \\ \frac{\partial}{\partial a}[n \ln a - \ln(n!) - a] & = & 0 \\ \frac{n}{a} - 1 & = & 0 \\ \frac{1}{a}[n - a] & = & 0. \end{array} \tag{5.41}$$

La dernière équation met en évidence que la dérivée de la fonction de vraisemblance se met sous la forme $f(a)[\hat{a}_{ml}(n) - a]$, avec $f(a) = 1/a$. S'il est non biaisé, l'estimateur obtenu sera efficace.

L'estimateur ML est obtenu en résolvant l'équation (5.41) :

$$\hat{a}_{ml}(n) = n. \tag{5.42}$$

Biais de l'estimateur. On calcule l'espérance de l'estimateur :

$$E[\hat{a}_{ml}(n)] = E[n] = a, \tag{5.43}$$

en utilisant les propriétés de la loi de Poisson [1].

Variance de l'estimateur. D'après les calculs précédents, on sait que l'estimateur est non biaisé et efficace. On peut utiliser le théorème des bornes d'un estimateur efficace, et on trouve directement la variance :

$$\text{Var}[\hat{a}_{ml}(n) - a] = f(a)^{-1} = a. \tag{5.44}$$

On peut également utiliser l'une ou l'autre des (in)égalités de Cramer-Rao, tâche que nous laissons au lecteur.

5.6.3 Observation non linéaire

Enoncé. La variable déterministe inconnue à estimer, a, est mesurée au travers d'un système non linéaire qui fournit :

$$r_i = g(a) + n_i, \quad i = 1, \ldots k. \tag{5.45}$$

Les n_i sont des échantillons iid, gaussiens de loi $N(0, \sigma_n^2)$.

Calcul de la fonction de vraisemblance. En raison de l'indépendance des n_i, on peut écrire :

$$\begin{array}{rcl} p(\mathbf{r}/a) & = & \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\Big[- \frac{(r_i - g(a))^2}{2\sigma_n^2}) \Big], \\ \ln p(\mathbf{r}/a) & = & k \ln (\frac{1}{\sqrt{2\pi}\sigma_n}) - \frac{1}{2\sigma_n^2} \sum_{i=1}^{k} (r_i - g(a))^2. \end{array} \tag{5.46}$$

On en déduit l'équation de vraisemblance :

$$\frac{\partial \ln p(\mathbf{r}/a)}{\partial a} = \frac{1}{\sigma_n^2} \sum_{i=1}^{k} (r_i - g(a)) \frac{\partial g(a)}{\partial a} = 0. \tag{5.47}$$

1. ce calcul a été fait de façon détaillée en TD dans les rappels de probabilités et de statistiques

En général, le second terme ne peut pas se mettre sous la forme $f(a)(\hat{a}(\mathbf{r}) - a)$, par conséquent l'estimateur n'est pas efficace. L'équation de vraisemblance peut encore s'écrire :

$$\frac{\partial g(a)}{\partial a}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k} r_i\Big) - g(a))\Big]_{a=\hat{a}_{ml}(\mathbf{r})} = 0. \tag{5.48}$$

S'il existe a tel que $g(a) = \frac{1}{k}\sum_{i=1}^{k} r_i$, on a la solution :

$$g(\hat{a}_{ml}(\mathbf{r})) = \frac{1}{k}\sum_{i=1}^{k} r_i, \tag{5.49}$$

et si la fonction g est inversible, on a finalement :

$$\hat{a}_{ml}(\mathbf{r}) = g^{-1}\Big(\frac{1}{k}\sum_{i=1}^{k} r_i\Big). \tag{5.50}$$

Dans le cas où g^{-1} n'existe pas, il n'y aucune méthode générale pour estimer a, même en absence de bruit. Par exemple, si $g(u) = u^2$, on déduit $u = \pm\sqrt{g(u)}$, mais il n'y a aucun moyen (sans information supplémentaire) de décider du signe !

Calcul de la variance On suppose que l'estimateur est non biaisé. On peut alors utiliser les bornes de Cramer-Rao, par exemple :

$$\text{Var}[\hat{a}(\mathbf{r}) - a] \geq -\Big\{E\Big[\frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2}\Big]\Big\}^{-1}. \tag{5.51}$$

A partir de (5.47), calculons donc la dérivée seconde de la fonction de vraisemblance :

$$\frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2} = \frac{1}{\sigma_n^2}\sum_{i=1}^{N}(r_i - g(a))\frac{\partial^2 g(a)}{\partial a^2} - \frac{k}{\sigma_n^2}\Big(\frac{\partial g(a)}{\partial a}\Big)^2, \tag{5.52}$$

puis son espérance :

$$E\Big[\frac{\partial^2 \ln p(\mathbf{r}/a)}{\partial a^2}\Big] = -\frac{k}{\sigma_n^2}\Big(\frac{\partial g(a)}{\partial a}\Big)^2, \tag{5.53}$$

d'où l'on déduit la borne :

$$\text{Var}[\hat{a}(\mathbf{r}) - a] \geq \frac{\sigma_n^2}{k(\partial g(a)/\partial a)^2}. \tag{5.54}$$

On remarque que la borne ressemble à celle obtenue dans le cas d'une observation linéaire, au terme correctif $(\partial g(a)/\partial a)^2$ près. Ce terme s'explique facilement en considérant la figure 5.1 : en effet, à une erreur δa sur a est associée une erreur $\delta g \approx (dg/da)\delta a$. D'un point de vue variance, on a donc $E[(\delta g)^2] \approx (dg/da)^2 E[(\delta a)^2]$.

5.7 Liens entre estimateurs ML et MAP

Les estimateurs MAP et ML sont assez voisins. En effet, en comparant l'équation de vraisemblance :

$$\partial \ln p(\mathbf{r}/a)/\partial a = 0, \tag{5.55}$$

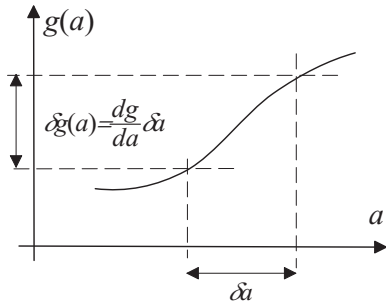


FIGURE 5.1 – Erreur dans le cas d'une estimation non linéaire $g(a)$.

et l'équation MAP :

$$\partial \ln p(\mathbf{r}/a)/\partial a + \partial \ln p(a)/\partial a = 0, \tag{5.56}$$

on voit que la différence réside dans le fait que l'estimateur MAP utilise en plus la connaissance *a priori* sur a (second terme). Par conséquent, si la connaissance sur a devient nulle, les deux estimateurs coïncident.

Dans le cas de l'observation avec bruit gaussien additif, l'estimateur MAP tend vers l'estimateur ML si la connaissance sur a devient nulle, c'est-à-dire si la variance de a tend vers l'infini :

$$\lim_{\sigma_a \to +\infty} \hat{a}_{map}(\mathbf{r}) = \lim_{\sigma_a \to +\infty} \frac{\sigma_a^2}{\sigma_a^2 + \frac{\sigma_n^2}{k}} \Big(\frac{1}{k} \sum_{i=1}^{k} r_i \Big) = \frac{1}{k} \sum_{i=1}^{k} r_i = \hat{a}_{ml}(\mathbf{r}) \tag{5.57}$$

Dans l'exemple avec la loi de Poisson, l'estimateur MAP utilise la connaissance *a priori* contenue dans le paramètre λ de la loi exponentielle de a. Si la connaissance devient nulle, c'est-à-dire lorsque λ tend vers 0, l'estimateur MAP tend vers l'estimateur ML. En effet :

$$\lim_{\lambda \to 0} \hat{a}_{map}(n) = \lim_{\lambda \to 0} \frac{n}{\lambda + 1} = n = \hat{a}_{ml}(n). \tag{5.58}$$

5.8 Propriétés de l'estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance, même lorsqu'il n'est pas efficace, possède des propriétés asymptotiques très intéressantes. De plus, l'approche ML est une approche systématique que l'on peut toujours mettre en œuvre, et souvent de façon simple. Ces remarques fournissent donc des motivations fortes pour l'utiliser.

— On montre que la solution de l'équation ML : $\partial p(\mathbf{r}/a)/\partial a = 0$, converge en probabilité vers la valeur idéale a si le nombre de mesures k tend vers l'infini, autrement dit :

$$\forall \epsilon > 0, \lim_{k \to +\infty} P\Big(|\hat{a}_{ml}(\mathbf{r}) - a| < \epsilon \Big) = 1. \tag{5.59}$$

Tous les estimateurs possédant cette propriété sont dits consistants.

— L'estimateur ML est assymptotiquement efficace, c'est-à-dire :

$$\lim_{k \to +\infty} \frac{\text{Var}[\hat{a}_{ml}(\mathbf{r}) - a]}{\Big(- E\Big[\frac{\partial^2 \ln p(\mathbf{r}/a]}{\partial a^2} \Big] \Big)^{-1}} = 1. \tag{5.60}$$

— L'estimateur ML est asymptotiquement gaussien, c'est-à-dire que pour $k \to +\infty$, $\hat{a}_{ml} \sim N(a, \sigma^2_{\hat{a}})$.

Nous terminerons par deux questions et leurs réponses.

— Existe-t-il un meilleur estimateur que l'estimateur ML ?
S'il n'y a pas d'estimateur efficace, il peut exister des estimateurs non biaisés avec des variances plus faibles que l'estimateur ML. Le problème est qu'il n'y a pas de méthodes générales pour concevoir ces estimateurs hypothétiques, contrairement à l'estimateur ML.

— Y-a-t-il des bornes plus petites que celles proposées par le théorème de Cramer-Rao ?
S'il n'y a pas d'estimateur efficace, il existe effectivement des bornes plus faibles que celles proposées par ce théorème, mais les calculs pour y parvenir sont très complexes. C'est pourquoi les bornes de Cramer-Rao restent très utilisées.

Chapitre 6

Estimation de paramètres multiples

Dans de nombreux problèmes, il est nécessaire d'estimer plusieurs paramètres. C'est par exemple le cas dans l'estimation des paramètres d'une cible radar : position, vitesse, etc. La plupart des idées introduites dans les chapitres précédents peuvent être étendues au cas multivariable. Dans ce cadre, on considèrera l'ensemble des paramètres sous forme d'un vecteur dans un espace de dimension p : $\mathbf{a} = (a_1, a_2, \ldots, a_p)^T$.

Dans ce chapitre, on étudiera les deux situations, de vecteurs paramètres aléatoires et déterministes, et on développera trois points :

— les procédures d'estimation,
— la mesure des erreurs,
— les performances.

6.1 Estimation

6.1.1 Estimation de vecteurs aléatoires

Dans le cas de l'estimation de vecteurs aléatoires, on étend le critère de Bayes afin de généraliser les estimateurs des moindres carrés (LS) et du maximum *a posteriori* (MAP).

On mesure l'erreur à l'aide d'une fonction de coût vectorielle : $C(\mathbf{a}, \hat{\mathbf{a}}(\mathbf{r}))$. Comme dans le cas scalaire, on considère généralement une fonction de l'erreur C de $\mathbb{R}^p \to \mathbb{R}$ telle que $(\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}) \in \mathbb{R}^p \to C(\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}) \in \mathbb{R}$. Ici $\mathbf{a}$ et $\hat{\mathbf{a}}(\mathbf{r})$ sont des vecteurs de dimension p, par exemple :

$$\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a} = \begin{pmatrix} \hat{a}_1(\mathbf{r}) - a_1 \\ \hat{a}_2(\mathbf{r}) - a_2 \\ \vdots \\ \hat{a}_p(\mathbf{r}) - a_p \end{pmatrix}. \tag{6.1}$$

Estimateur des moindres carrés

Dans le cas du critère d'erreur quadratique, la fonction de coût est :

$$C_{ls}(\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}) = \sum_{i=1}^{p} (\hat{a}_i(\mathbf{r}) - a_i)^2 = \hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a})^T \cdot (\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}). \tag{6.2}$$

La dernière expression de droite est la forme vectorielle : un simple produit scalaire. Le risque de Bayes associé à ce critère est simplement l'erreur moyenne, intégrée sur tout l'espace des observations $\mathbf{r} \in \mathbb{R}^k$ et sur tout l'espace des vecteurs paramètres $\mathbf{a} \in \mathbb{R}^p$:

$$\begin{aligned} \mathcal{R}_{ls} &= \int_{\mathbb{R}^k} \int_{\mathbb{R}^p} C_{ls}(\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}) p(\mathbf{r}, \mathbf{a}) \mathbf{dr} \mathbf{da}, \\ &= \int_{\mathbb{R}^k} p(\mathbf{r}) \Big[\int_{\mathbb{R}^p} (\hat{\mathbf{a}} - \mathbf{a})^T \cdot (\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}) p(\mathbf{a}/\mathbf{r}) \mathbf{da} \Big] \mathbf{dr}, \\ &= \int_{\mathbb{R}^k} p(\mathbf{r}) \Big[\int_{\mathbb{R}^p} (\sum_{i=1}^p (\hat{a}_i(\mathbf{r}) - a_i)^2) p(\mathbf{a}/\mathbf{r}) \mathbf{da} \Big] \mathbf{dr}. \end{aligned} \tag{6.3}$$

Comme dans le cas scalaire, on minimise le risque en minimisant l'intégrale intérieure. Puisque chaque terme de la somme est positif (somme de carrés), minimiser l'intégrale de la somme revient à minimiser chaque terme :

$$\begin{aligned} \min_{\hat{a}_1(\mathbf{r}), \ldots, \hat{a}_p(\mathbf{r})} \Big[\int_{\mathbb{R}^p} (\sum_{i=1}^p (\hat{a}_i(\mathbf{r}) - a_i)^2) p(\mathbf{a}/\mathbf{r}) \mathbf{da} \Big] &= \\ \min_{\hat{a}_1(\mathbf{r})} \Big[\int_{\mathbb{R}^p} (\hat{a}_1(\mathbf{r}) - a_1)^2 p(\mathbf{a}/\mathbf{r}) \mathbf{da} \Big] + \ldots & \\ + \min_{\hat{a}_p(\mathbf{r})} \Big[\int_{\mathbb{R}^p} (\hat{a}_p(\mathbf{r}) - a_p)^2 p(\mathbf{a}/\mathbf{r}) \mathbf{da} \Big]. & \end{aligned} \tag{6.4}$$

Pour le terme d'indice i, le minimum est solution de :

$$\frac{\partial}{\partial \hat{a}_i} \Big[\int_{\mathbb{R}^p} (\hat{a}_i(\mathbf{r}) - a_i)^2 p(\mathbf{a}/\mathbf{r}) \mathbf{da} \Big] = 2 \Big[\int_{\mathbb{R}^p} (\hat{a}_i(\mathbf{r}) - a_i) p(\mathbf{a}/\mathbf{r}) \mathbf{da} \Big] = 0, \tag{6.5}$$

c'est-à-dire pour :

$$(\hat{a}_i)_{ls}(\mathbf{r}) = \int_{\mathbb{R}^p} a_i p(\mathbf{a}/\mathbf{r}) \mathbf{da}, \tag{6.6}$$

soit de façon globale :

$$\hat{\mathbf{a}}_{ls}(\mathbf{r}) = \int_{\mathbb{R}^p} \mathbf{a} p(\mathbf{a}/\mathbf{r}) \mathbf{da}. \tag{6.7}$$

L'estimateur des moindres carrés est donc la moyenne conditionnelle ou moyenne *a posteriori*, comme dans le cas scalaire.

Estimateur du maximum *a posteriori*

Si on choisit l'estimateur MAP, on cherche la valeur $\mathbf{a} = \hat{\mathbf{a}}(\mathbf{r})$ qui maximise la densité multivariable $p(\mathbf{a}/\mathbf{r})$. Si le maximum est intérieur au domaine de variation de $\mathbf{a}$, et si les dérivées partielles $\partial \ln p(\mathbf{a}/\mathbf{r}) / \partial a_i$ existent, l'équation MAP est constituée de p équations élémentaires :

$$\frac{\partial \ln p(\mathbf{a}/\mathbf{r})}{\partial a_i} \Big|_{\mathbf{a} = \hat{\mathbf{a}}_{map}(\mathbf{r})} = 0, \; i = 1, \ldots, p, \tag{6.8}$$

que l'on peut écrire de façon compacte :

$$\nabla_{\mathbf{a}} \ln p(\mathbf{a}/\mathbf{r}) \Big|_{\mathbf{a} = \hat{\mathbf{a}}_{map}(\mathbf{r})} = 0, \tag{6.9}$$

où $\nabla_{\mathbf{a}}$ représente le vecteur gradient par rapport à $\mathbf{a}$.

On doit bien entendu sélectionner parmi les solutions, celle qui correspond au *maximum maximorum*.

6.1.2 Estimation de vecteurs déterministes

Pour un vecteur déterministe, on choisit l'estimateur du maximum de vraisemblance, c'est-à-dire la valeur $\mathbf{a} = \hat{\mathbf{a}}(\mathbf{r})$ qui maximise la vraisemblance multivariable $p(\mathbf{r}/\mathbf{a})$. Si le maximum est intérieur au domaine de variation de $\mathbf{a}$, et si les dérivées partielles $\partial \ln p(\mathbf{r}/\mathbf{a})/\partial a_i$ existent, l'équation du maximum de vraisemblance est constituée de p équations élémentaires :

$$\frac{\partial \ln p(\mathbf{r}/\mathbf{a})}{\partial a_i}\Big|_{\mathbf{a}=\hat{\mathbf{a}}_{ml}(\mathbf{r})} = 0,\ i = 1, \ldots, p, \tag{6.10}$$

que l'on peut aussi écrire de façon compacte :

$$\nabla_{\mathbf{a}} \ln p(\mathbf{r}/\mathbf{a})\Big|_{\mathbf{a}=\hat{\mathbf{a}}_{ml}(\mathbf{r})} = 0, \tag{6.11}$$

où $\nabla_{\mathbf{a}}$ représente le vecteur gradient par rapport à $\mathbf{a}$.

On doit encore sélectionner parmi les solutions, celle qui correspond au *maximum maximorum*.

6.2 Performance

Dans le cas d'estimation de vecteurs déterministes, on mesure les performances en calculant l'écart entre l'estimation et la solution théorique (le biais) et la dispersion des estimations.

6.2.1 Biais de l'estimateur

Dans le cas multivariable, le biais est un vecteur, calculé simplement comme la différence entre l'espérance de l'estimateur et le vecteur théorique :

$$\mathbf{b}(\mathbf{a}) = E[\hat{\mathbf{a}}(\mathbf{r})] - \mathbf{a}. \tag{6.12}$$

L'estimateur est non biaisé, si le biais est le vecteur nul : $\mathbf{b}(\mathbf{a}) = 0$, c'est-à-dire si chaque composante est nulle :

$$(\mathbf{b}(\mathbf{a}))_i = E[\hat{a}_i(\mathbf{r})] - a_i = 0,\ \forall i = 1, \ldots, p. \tag{6.13}$$

6.2.2 Dispersion de l'estimateur

Dans le cas d'un scalaire, la dispersion était mesurée par la variance de l'écart entre l'estimateur et la valeur théorique. Dans le cas multivariable, la dispersion pourra être mesurée par la matrice de variance-covariance $\Gamma_{\mathbf{e}}$ de l'erreur $\mathbf{e} = \hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}$:

$$\Gamma_{\mathbf{e}} = E\Big[\Big((\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}) - E[\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}]\Big)\Big((\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}) - E[\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a}]\Big)^T\Big]. \tag{6.14}$$

Les éléments diagonaux, d'indice ii, de la matrice donneront la dispersion de chacune des composantes a_i. Les éléments ij, hors de la diagonale, renseignent sur le couplage des erreurs d'estimation entre a_i et a_j.

6.2.3 Dispersion dans le cas gaussien

On suppose que l'erreur $(\hat{\mathbf{a}}(\mathbf{r}) - \mathbf{a})$, notée $\mathbf{e}(\mathbf{r})$ pour simplifier, est une variable aléatoire gaussienne de dimension p. Sa densité de probabilité s'écrit alors :

$$p(\mathbf{e}) = (2\pi^{p/2}|\det(\Gamma_{\mathbf{e}})|^{1/2})^{-1} \exp\Big(-\frac{1}{2}\mathbf{e}^T\Gamma_{\mathbf{e}}^{-1}\mathbf{e}\Big) \tag{6.15}$$

De cette équation, on déduit que les valeurs de $\mathbf{e}$ qui ont une même densité de probabilité satisfont l'équation :

$$\mathbf{e}^T\Gamma_{\mathbf{e}}^{-1}\mathbf{e} = c^2, \tag{6.16}$$

où c est une constante.

Dans le cas $p = 2$, la courbe définie par (6.16) est une ellipse. On peut alors calculer la probabilité que l'erreur $\mathbf{e}$ soit située à l'intérieur de l'ellipse. Pour cela, il faut intégrer $p(\mathbf{e})$ sur le domaine correspondant à l'ellipse. L'aire de l'ellipse définie par $\mathbf{e}^T\Gamma_{\mathbf{e}}^{-1}\mathbf{e} = c^2$ est simplement :

$$\mathcal{A} = |\det(\Gamma_{\mathbf{e}})|^{1/2}\pi c^2. \tag{6.17}$$

Entre c et $c + dc$, on a une couronne élémentaire dont l'aire vaut :

$$d\mathcal{A} = |\det(\Gamma_{\mathbf{e}})|^{1/2}2\pi c dc. \tag{6.18}$$

Sur un point de la couronne, la densité de probabilité est égale (au premier ordre) à :

$$p(\mathbf{e}) = (2\pi|\det(\Gamma_{\mathbf{e}})|^{1/2})^{-1} \exp\Big(-\frac{c^2}{2}\Big). \tag{6.19}$$

La probabilité des points extérieurs à l'ellipse est donc :

$$\begin{array}{rcl} P & = & \Pr(\mathbf{e} \in \text{extérieur de l'ellipse}) \\ & = & \int_c^{+\infty}(2\pi^{p/2}|\det(\Gamma_{\mathbf{e}})|^{1/2})^{-1}\exp(-\frac{x^2}{2})d\mathcal{A} \\ & = & \int_c^{+\infty}(2\pi|\det(\Gamma_{\mathbf{e}})|^{1/2})^{-1}\exp(-\frac{x^2}{2})|\det(\Gamma_{\mathbf{e}})|^{1/2}2\pi x dx \\ & = & \int_c^{+\infty}\exp\Big(-\frac{x^2}{2}\Big)x dx \\ & = & \int_{c^2/2}^{\infty}\exp(-u)du \\ & = & \exp(-\frac{c^2}{2}). \end{array} \tag{6.20}$$

La probabilité de $\mathbf{e}$ à l'intérieur de l'ellipse vaut donc :

$$\Pr(\mathbf{e} \in \text{ellipse}) = 1 - P = 1 - \exp\Big(-\frac{c^2}{2}\Big). \tag{6.21}$$

Les ellipses $\mathbf{e}^T\Gamma_{\mathbf{e}}^{-1}\mathbf{e} = c^2$ sont appelées ellipses de concentration car elles donnent une mesure de la concentration de la densité de l'erreur, qui ne dépend que du scalaire c.

Dans le cas général (p quelconque), $\mathbf{e}^T\Gamma_{\mathbf{e}}^{-1}\mathbf{e} = c^2$ définit une ellipsoïde. On peut étendre les calculs prédédents, et on trouve :

$$\Pr(\mathbf{e} \in \text{extérieur de l'ellipsoïde}) = \frac{p}{2^{p/2}\Gamma(p/2+1)}\int_c^{+\infty} x^{p-1}\exp(-\frac{x^2}{2})dx, \tag{6.22}$$

où Γ est la fonction Eulérienne de première espèce (fonction factorielle) définie par :

$$\Gamma(u) = \int_0^{+\infty} t^{u-1}\exp(-t)dt. \tag{6.23}$$

On parle alors d'ellipsoïdes de concentration.

Troisième partie

Théorie de l'information

Objectifs

La théorie de l'information répond à deux questions importantes :

1. Quel est le taux de compression ultime ? Nous verrons qu'il s'agit de l'entropie H.
2. Quel est le taux de transmission ultime d'une communication ? On verra qu'il s'agit de la capacité du canal C.

La théorie de l'information a des liens avec de nombreuses disciplines scientifiques et technologiques :

— mathématiques : les quantités fondamentales de la théorie de l'information sont définies dans un formalisme statistique, elles permettent aussi de caractériser les distributions, le comportement de longues séquences de varaibles aléatoires, etc.
— Thermodynamique physique : mécanique statistique, seconde loi de la thermodynamique,
— les communications : capacité d'un canal de transmission, codage et compression, codes détecteurs et correcteurs d'erreur, etc.
— l'informatique : complexité de Kolmogorov, *minimum description length* de Rissanen.

Repères historiques

La notion d'entropie a été initialement introduite en physique statistique par Boltzman. En 1930, Hartley propose une mesure logarithmique de l'information, définie comme le logarithme de la taille de l'alphabet. Dans les années 40, Shannon introduit les définitions actuelles de l'entropie et de l'information mutuelle.

Information : une mesure de l'incertitude

Définition 6.2.1 *Soit une variable aléatoire discrète X à valeurs x dans $\mathcal{X}$ de probabilité $p(x)$, on appelle incertitude ou information d'un évènement x, la quantité $I(x) = -\log p(x)$*

Cette définition est liée à une définition objective de l'information, non fondée sur son contenu, mais sur son incertitude : l'information apportée par un événement est égale à l'incertitude sur cet événement avant l'expérience. De plus, elle est cohérente avec le bon sens : un événement imprévu apporte beaucoup plus d'information qu'un événement prévisible. En effet, on remarque que :

— Pour un événement x certain, c'est-à-dire tel que $p(x) = 1$, l'incertitude vaut $I(x) = -\log 1 = 0$.
— Pour un événement de probabilité nulle, l'incertitude $I(x) \to +\infty$.

Organisation

Cette troisième partie traite de la théorie de l'information pour des sources discrètes. Elle est organisée en deux chapitres : le premier introduit les grandeurs fondamentales de la théorie de l'information, le second chapitre est consacré aux principes de base du codage.

Le lecteur curieux pourra compléter ses connaissances notamment pour les sources continues et le théorème de Shannon concernant la capacité d'un canal bruité en consultant des ouvrages spécialisés.

Chapitre 7

Grandeurs fondamentales de la théorie de l'information

7.1 Entropie

7.1.1 Définitions

Définition 7.1.1 *Soit une variable aléatoire discrète X à valeurs x dans $\mathcal{X}$, et de distribution de probabilité $p(x)$, l'entropie $H(X)$ est égale à l'incertitude ou information moyenne :*

$$H(X) = E[I(x)] = -E[\log p(x)] = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{7.1}$$

Dans l'entropie, la base usuelle du logarithme est le base 2. Dans ce cas, l'unité d'entropie est le BIT pour *BInary uniT*. Dans la suite, on ne notera pas $\log_2$ mais simplement $\log$. On a alors quelques égalités usuelles :

— $\log 2 = 1$,
— $\log 2^n = n$,
— $\log 3 \approx 1.58$.

7.1.2 Propriétés

L'entropie est positive ou nulle : $H(X) \geq 0$. En effet :

$$\begin{array}{c} 0 \leq p(x) \leq 1 \\ 0 \leq -\log p(x) \\ \text{d'où } E[-\log p(x)] \geq 0, \end{array} \tag{7.2}$$

avec égalité $E[-\log p(x)] = 0$ si et seulement si $\exists x / p(x) = 1$. Cette égalité, admise pour le moment, sera montrée par la suite.

7.1.3 Exemples

Soit la variable aléatoire binaire X :

$$X = \begin{cases} 1 & \text{avec une probabilité } p, \\ 0 & \text{avec une probabilité } 1-p. \end{cases} \tag{7.3}$$

L'entropie s'écrit alors :

$$H(X) = -p\log p - (1-p)\log(1-p) \triangleq H(p). \tag{7.4}$$

Pour $p = 1$, on trouve $H(1) = -1\log 1 - 0\log 0 = 0$ bit. Pour $p = 0.5$, $H(0.5) = -0.5\log 0.5 - 0.5\log 0.5 = 1$ bit. Pour $p = 0$, on a $H(0) = 0$ en utilisant la convention $\epsilon \log \epsilon \overset{\rightarrow}{\epsilon} 0\ 0$. Par symétrie, on a aussi $H(1) = H(0)$.

7.2 Entropies jointes et conditionnelles

7.2.1 Définitions

On peut définir également l'entropie d'un vecteur aléatoire. Dans cette partie, on donnera les définitions d'entropie jointe et conditionnelle dans le cas d'un vecteur aléatoire à deux dimensions. L'extension à un vecteur aléatoire de dimension quelconque est immédiate.

Définition 7.2.1 *Soient deux variables aléatoires X à valeurs x dans $\mathcal{X}$ et de distribution de probabilité $p(x)$, et Y à valeurs y dans $\mathcal{Y}$ et de distribution de probabilité $p(y)$, l'entropie jointe $H(X,Y)$ est égale :*

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = -E[\log p(x,y)]. \tag{7.5}$$

Définition 7.2.2 *Soient deux variables aléatoires X à valeurs x dans $\mathcal{X}$ et de distribution de probabilité $p(x)$, et Y à valeurs y dans $\mathcal{Y}$ et de distribution de probabilité $p(y)$, l'entropie conditionnelle $H(Y/X)$ est égale :*

$$\begin{array}{rcl} H(Y/X) & = & \sum_{x\in\mathcal{X}} p(x) H(Y/X=x) \\ & = & -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y/x)\log p(y/x) \\ & = & -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y/x)\log p(y/x) \\ & = & -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y/x) \\ & = & -E[\log p(y/x)]. \end{array} \tag{7.6}$$

Bien sûr, on peut également définir l'entropie conditionnelle $H(X/Y)$. En suivant un calcul similaire, on trouve :

$$H(X/Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y/x) = -E[\log p(x/y)]. \tag{7.7}$$

7.2.2 Relations entre entropies

Théorème 7.2.1 *Soient deux variables aléatoires X et Y, les entropies simple, jointe et conditionnelles sont liées par la relation :*

$$H(X,Y) = H(X) + H(Y/X) = H(Y) + H(X/Y). \tag{7.8}$$

Démonstration

Par définition, l'entropie jointe vaut :

$$\begin{array}{rcl} H(X,Y) & = & -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ & = & -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(p(x)p(y/x)\right) \\ & = & -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y/x) \\ & = & H(X)+H(Y/X) \end{array} \tag{7.9}$$

La seconde égalité se démontre de façon similaire en utilisant $p(x,y) = p(y)p(x/y)$.

7.2.3 Propriétés et cas particulier

On montre que :

$$\begin{array}{rcl} H(X/Y) & \leq & H(X), \\ H(Y/X) & \leq & H(Y), \end{array} \tag{7.10}$$

avec égalités si les variables X et Y sont indépendantes.

Si X et Y sont deux variables aléatoires indépendantes, c'est-à-dire vérifiant $p(x,y) = p(x)p(y)$ (ou $p(x/y) = p(x)$ et $p(y/x) = p(y)$), on a les relations :

$$\begin{array}{rcl} H(X) & = & H(X/Y), \\ H(Y) & = & H(Y/X), \\ H(X,Y) & = & H(X)+H(Y). \end{array} \tag{7.11}$$

7.2.4 Exemple 1

Soient deux variables aléatoires X et Y prenant chacune quatre valeurs, $\{x_1,x_2,x_3,x_4\}$ et $\{y_1,y_2,y_3,y_4\}$ respectivement, avec les probabilités jointes données dans le tableau ci-dessous.

	x_1	x_2	x_3	x_4	$p(y_i)$
y_1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
y_2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
y_3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
y_4	$1/4$	0	0	0	$1/4$
$p(x_i)$	$1/2$	$1/4$	$1/8$	$1/8$	

TABLE 7.1 – Probabilités jointes $p(x_i,y_j)$ de l'exemple 1.

Calcul des entropies simples, $H(X)$ et $H(Y)$

On utilise les probabilités marginales, calculées en faisant la somme des probabilités jointes, en ligne pour y et en colonne pour x. On a donc :

$$\begin{array}{rcl} H(X) & = & -\sum_{i=1}^{4} p(x_i)\log p(x_i) \\ & = & -(\frac{1}{2}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{4}+\frac{1}{8}\log\frac{1}{8}+\frac{1}{8}\log\frac{1}{8}) \\ & = & \frac{1}{2}1+\frac{1}{4}2+\frac{1}{8}3+\frac{1}{8}3 \\ & = & \frac{7}{4}\text{bits}. \end{array} \tag{7.12}$$

Le calcul de $H(Y)$ est similaire et on trouve $H(Y) = 2$bits

Calcul des entropies conditionnelles, $H(X/Y)$ et $H(Y/X)$

Calculons $H(X/Y)$ en utilisant la relation :

$$\begin{array}{rcl} H(X/Y) & = & -E[\log p(x/y)] \\ & = & -\sum_{i=1}^{4}\sum_{j=1}^{4} p(x_i,y_j)\log p(x_i/y_j) \\ & = & -\sum_{i=1}^{4}\sum_{j=1}^{4} p(y_j)p(x_i/y_j)\log p(x_i/y_j) \\ & = & -\sum_{j=1}^{4} p(y_j)\Big(\sum_{i=1}^{4} p(x_i/y_j)\log p(x_i/y_j)\Big). \end{array} \tag{7.13}$$

On doit donc calculer les probabilités conditionnelles $p(x_i/y_j)$. Pour $y = y_1$, on calcule :

$$\begin{array}{rcl} p(x_1/y_1) & = & p(x_1,y_1)/p(y_1) \\ & = & \frac{1/8}{1/4} \\ & = & \frac{1}{2}, \end{array} \tag{7.14}$$

et de façon similaire $p(x_2/y_1) = 1/4$, $p(x_3/y_1) = 1/8$ et $p(x_4/y_1) = 1/8$. On effectue les mêmes calculs pour les autres y_i. On peut maintenant appliquer la formule (7.13), et on trouve en notant $H(U) = H(p(u_1), p(u_2), p(u_3), p(u_4))$:

$$\begin{array}{rcl} H(X/Y) & = & \frac{1}{4}H(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}) + \frac{1}{4}H(\frac{1}{4},\frac{1}{2},\frac{1}{8},\frac{1}{8}) + \frac{1}{4}H(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}) + \frac{1}{4}H(1,0,0,0) \\ & = & \frac{1}{4}\frac{7}{4} + \frac{1}{4}\frac{7}{4} + \frac{1}{4}2 + \frac{1}{4}0 \\ & = & \frac{11}{8}\text{bits}. \end{array} \tag{7.15}$$

En procédant de façon similaire, on trouve $H(Y/X) = 13/8$ bits.

Vérification des relations entre entropies

On peut vérifier les relations entre entropies :

$$\begin{array}{rcl} H(X,Y) & = & \frac{27}{8} \\ & = & H(X) + H(Y/X) = \frac{7}{4} + \frac{13}{8} = \frac{27}{8} \\ & = & H(Y) + H(X/Y) = 2 + \frac{11}{8} = \frac{27}{8}. \end{array} \tag{7.16}$$

On vérifie également que :

$$\begin{array}{rcl} H(X/Y) = \frac{11}{8} & \leq & H(X) = \frac{14}{8}, \\ H(Y/X) = \frac{13}{8} & \leq & H(Y) = 2. \end{array} \tag{7.17}$$

On remarque que $H(X/Y) \neq H(Y/X)$. Enfin, on a :

$$\begin{array}{rcl} H(X/Y) & < & H(X) \\ H(Y/X) & < & H(Y), \end{array} \tag{7.18}$$

car X et Y ne sont pas des variables aléatoires indépendantes.

7.2.5 Exemple 2

On considère la langue française comme une source d'information X, qui utilise 27 symboles : 26 lettres plus le caractère *espace* sont les éléments x_i de l'alphabet $\mathcal{X}$.

Si chaque symbole x_i était équiprobable, on aurait donc $p(x_i) = 1/27$ et par conséquent :

$$\begin{aligned} H(X) &= -\sum_{i=1}^{27} p(x_i) \log p(x_i) \\ &= -\sum_{i=1}^{27} \tfrac{1}{27} \log \tfrac{1}{27} \\ &= \log 27 \\ &= 4.75 \text{ bits/lettre.} \end{aligned} \tag{7.19}$$

En réalité, on sait que les lettres ne sont pas équiprobables. Si on estime les probabilités $p(x_i)$ à partir des fréquences relatives d'un texte, on trouve par exemple :

$$\begin{aligned} p(\textit{espace}) &= 0.184, \\ p(e) &= 0.148, \\ p(s) &= 0.077, \\ p(n) &= 0.071, \\ p(t) &= 0.068, \text{ etc.} \end{aligned}$$

Ces probabilités sont très différentes de $1/27 \approx 0.037$. Avec ces valeurs, le calcul de l'entropie donne maintenant :

$$H(X) = 3.98 \text{ bits/lettre.} \tag{7.20}$$

La perte d'entropie est due au fait que l'incertitude n'est pas identique pour toutes les lettres.

On peut encore aller plus loin. En effet, on sait que la probabilité d'une lettre dépend fortement de la lettre précédente. Par exemple, après la lettre q, la probabilité d'avoir un u est très proche de 1. Pour avoir une meilleure estimation de l'entropie de la langue française (ou d'une autre langue), on doit tenir compte de cette dépendance, et considérer non pas des lettres isolées mais des groupes de deux lettres, trois lettres ou plus. Ici, nous considérerons la variable aléatoire Z correspondant aux paquets de 2 lettres. Si deux lettres successives étaient indépendantes, on aurait :

$$H(Z) = 2H(X) \approx 7.9 \text{ bits.} \tag{7.21}$$

En fait puisque deux lettres successives ne sont pas indépendantes, on peut écrire (en notant X_1 et X_2 les sources associées respectivement à la première et à la seconde lettre) :

$$H(Z) = H(X_1) + H(X_2/X_1) < H(X_1) + H(X_2) \approx 7.9 \text{ bits.} \tag{7.22}$$

Enfin, on peut se demander si toutes les langues ont des entropies identiques. A partir de statistiques sur les lettres (comme ci-dessus), on peut calculer l'entropie de l'anglais, et on trouve :

$$H(\text{anglais}) \approx 4.1 \text{bits/lettre.} \tag{7.23}$$

et en considérant des paquets de deux lettres, on trouve $H(X_2/X_1) = 3.6$ bits, d'où $H(Z) \approx 7.7$ bits. On remarque que l'entropie de l'anglais est un peu supérieure à l'entropie du français. Autrement dit, une lettre en anglais apporte plus d'information qu'une lettre en français. Concrètement, on observe qu'une traduction anglaise est plus courte que le texte français correspondant.

7.3 Entropies relatives et information mutuelle

7.3.1 Définitions

Définition 7.3.1 *On appelle entropie relative ou divergence de Kullback-Leibler (KL) entre deux distributions $p(x)$ et $q(x)$ de la même variable aléatoire X la quantité :*

$$D(p//q) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p\Big[\log \frac{p(x)}{q(x)}\Big]. \tag{7.24}$$

On montre que $D(p//q) \geq 0$ et ne s'annulle que si les deux distributions sont égales : $p(x) = q(x)$.

Remarque. D'un point de vue mathématique, la quantité $D(p//q)$ n'est pas une distance, car elle n'est pas symétrique : $D(p//q) \neq D(q//p)$. Le terme *divergence* provient de l'anglais et peut se traduire par *écart*.

Définition 7.3.2 *Soient deux variables aléatoires discrètes X et Y de probabilité jointe, $p(x, y)$, et de probabilités marginales $p(x)$ et $p(y)$, on appelle information mutuelle $I(X, Y)$ l'entropie relative entre la distribution jointe entre la distribution jointe et le produit des distributions marginales :*

$$\begin{aligned} I(X,Y) &= D(p(x,y)//p(x)p(y)) \\ &= E\Big[\log \frac{p(x,y)}{p(x)p(y)}\Big] \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \end{aligned} \tag{7.25}$$

7.3.2 Relations avec les entropies

A partir de la définition de l'information mutuelle, on tire :

$$\begin{aligned} I(X,Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y) \log p(y) \\ &= -H(Y/X) + H(Y). \end{aligned} \tag{7.26}$$

Avec un calcul similaire, on trouve également :

$$I(X,Y) = H(X) - H(X/Y). \tag{7.27}$$

De plus, en utilisant la relation entre entropies conditionnelles, simples et jointes, on arrive à la relation :

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{7.28}$$

Cas de variables indépendantes. Si les variables X et Y sont indépendantes, on a alors $p(x, y) = p(x)p(y)$ et $I(X, Y) = \sum_{x,y} p(x, y) \log 1 = 0$.

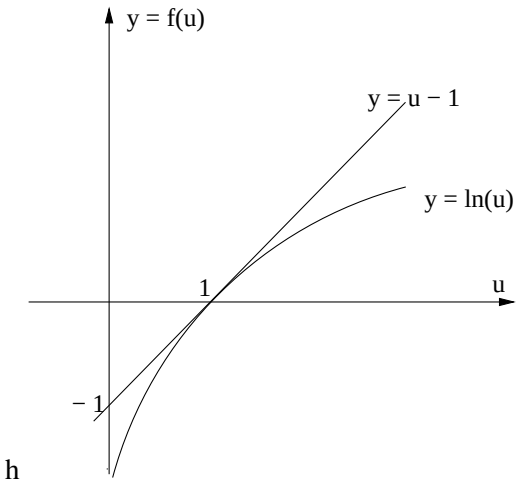


FIGURE 7.1 – La droite $y = u - 1$ est tangente à la fonction $\ln u$ au point -1. La fonction $\ln$ étant concave, on en déduit $\ln u \leq u - 1$.

7.4 Inégalité de Jensen

Cette inégalité est un résultat très utile en théorie de l'information qui établit la positivité de la divergence de KL. Elle permet aussi de déduire que l'entropie est maximale pour une distribution équiprobable.

7.4.1 Théorème

Théorème 7.4.1 *Soient $p(x)$ et $q(x)$ deux distributions d'une même variable aléatoire X, alors $D(p//q) \geq 0$ avec $D(p//q) = 0$ si et seulement si $p(x) = q(x)$.*

Preuve. Calculons l'opposé de la divergence de KL :

$$\begin{aligned} -D(p//q) &= -\sum_{x\in\mathcal{X}} p(x) \log \Big[\frac{p(x)}{q(x)}\Big] \\ &= +\sum_{x\in\mathcal{X}} p(x) \log \Big[\frac{q(x)}{p(x}\Big]. \end{aligned} \tag{7.29}$$

Le logarithme étant une fonction concave, on peut écrire :

$$\ln u \leq u - 1, \tag{7.30}$$

car la droite $y = u - 1$ est tangente à $\ln u$ en $u = 1$ (Fig. 7.1). On en déduit que $\log_2 u = \ln u / \ln 2 \leq (u-1)/\ln 2$. En reportant dans (7.29), on a alors :

$$\begin{aligned} -D(p//q) &\leq +\sum_{x\in\mathcal{X}} \frac{p(x)}{\ln 2}\Big(\frac{q(x)}{p(x)} - 1\Big) \\ -D(p//q) &\leq \frac{1}{\ln 2}\Big(\sum_{x\in\mathcal{X}} q(x) - \sum_{x\in\mathcal{X}} p(x)\Big). \end{aligned} \tag{7.31}$$

La somme d'une distribution sur tout l'espace étant égale à 1, on trouve finalement :

$$D(p//q) \geq 0. \tag{7.32}$$

Si $p = q$, alors $\log(p/q) = 0$ et $D(p//q) = 0$. Réciproquement, si $p \neq q$, alors $\log(q/p) < (q/p) - 1$ et par conséquent $D(p//q) > 0$.

7.4.2 Conséquences

Voici quelques résultats que l'on peut déduire de ce théorème.

Théorème 7.4.2 *Soient deux variables aléatoires X et Y, l'information mutuelle $I(X,Y)$ vérifie $I(X,Y) \geq 0$ avec $I(X,Y) = 0$ si et seulement si X et Y sont indépendantes.*

La démonstration est immédiate en utilisant la définition de l'information mutuelle et l'inégalité de Jensen.

Théorème 7.4.3 *Soient deux variables aléatoires X et Y, on a $H(X/Y) \leq H(X)$ avec égalités si et seulement si X et Y sont indépendantes.*

Pour montrer ce résultat, on écrit :

$$\begin{aligned} I(X,Y) &\geq 0 \\ H(X) - H(X/Y) &\geq 0 \\ H(X) &\geq H(X/Y). \end{aligned} \tag{7.33}$$

De plus, $I(X,Y) = H(X) - H(X/Y) = 0$ si et seulement si X et Y sont indépendantes.

Théorème 7.4.4 *Soient une variable aléatoire X de distribution $p(x)$ sur l'ensemble $\mathcal{X}$ de cardinal card$(\mathcal{X}) = N$, et $q(x) = 1/N$ la distribution uniforme sur $\mathcal{X}$, alors $H(X) \leq \log N$, et l'égalité $H(X) = \log N$ est obtenue pour $p(x) = q(x) = 1/N$, c'est-à-dire pour la distribution uniforme.*

Partons de l'opposé de la divergence KL entre p et q et appliquons l'inégalité de Jensen :

$$\begin{aligned} +\sum_{x\in\mathcal{X}} p(x) \log \frac{q(x)}{p(x)} &\leq 0 \\ H(X) + \sum_{x\in\mathcal{X}} p(x) \log q(x) &\leq 0 \\ H(X) + \sum_{x\in\mathcal{X}} p(x) \log(1/N) &\leq 0 \\ H(X) - \log N \sum_{x\in\mathcal{X}} p(x) &\leq 0 \\ H(X) - \log N &\leq 0, \end{aligned} \tag{7.34}$$

d'où finalement :

$$H(X) \leq \log N, \tag{7.35}$$

avec égalité si et seulement si $p(x) = q(x) = 1/N$.

7.5 Exercice : entropies d'une expérience

7.5.1 Enoncé

Soient une balance et neuf pièces de monnaie. La balance, de type Roberval, ne permet de faire des pesées comparatives. Huit des pièces sont identiques, la neuvième est fausse et se distingue pas sa masse différente des pièces vraies. On veut déterminer quelle pièce est fausse et si elle est plus lourde ou moins lourde que les vraies.

7.5.2 Existence d'une solution en trois pesées

Calculer l'entropie de l'expérience. Chaque pièce pouvant être fausse, et sa différence de masse étant inconnue, on a donc $9 \times 2 = 18$ situations possibles, toutes aussi probables. On a donc :

$$H(\textit{expérience}) = \log 18 \approx 4.16\text{bits}. \tag{7.36}$$

Calculer l'entropie maximale d'une pesée. Une pesée comparative peut proposer trois résultats : le plateau gauche est plus lourd (événement G), le plateau droit est plus lourd (événement D), équilibre (événement E). L'entropie maximale d'une pesée est atteinte si les trois probabilités sont égales : $P_D = P_G = P_E = 1/3$. On peut donc écrire :

$$H(\textit{pesée}) \leq \log 3 \approx 1.58\text{bits}. \tag{7.37}$$

En déduire que trois pesées suffisent pour résoudre ce problème. En trois pesées, on peut donc acquérir une information :

$$H(\textit{3 pesées}) \leq 3H(\textit{pesée}) \approx 4.75\text{bits}. \tag{7.38}$$

On remarque que l'entropie de trois pesées peut être supérieure à l'entropie de l'expérience. Par conséquent, en choisissant judicieusement chaque pesée, on peut résoudre le problème en trois pesées.

7.5.3 Détermination de la première pesée

On place n pièces dans chaque plateau de la balance. Calculer les probabilités P_D, P_G et P_E. On a donc n pièces dans chaque plateau et $9 - 2n$ pièces à l'écart de la balance. Calculons P_G :

$$\begin{aligned} P_G &= \text{Pr[(pièce fausse et plus lourde à gauche)} \\ &\quad \text{ou (pièce fausse et plus légère à droite)]} \qquad (7.39) \\ &= \frac{n}{18} + \frac{n}{18} \\ &= \frac{n}{9}. \qquad (7.40) \end{aligned}$$

De même, on trouve $P_D = n/9$. On en déduit donc :

$$\begin{aligned} P_E &= 1 - P_G - P_D \\ &= 1 - \frac{2n}{9}. \qquad (7.41) \end{aligned}$$

On aurait aussi pu calculer la probabilité d'être à l'équilibre. Cet événement se produit si la pièce fausse est écartée, c'est-à-dire se trouve parmi les $9 - 2n$. On a alors :

$$P_E = \frac{9-2n}{9}. \tag{7.42}$$

Par symétrie, $P_D = P_G$ et en utilisant $P_D + P_E + P_G = 1$, on déduit :

$$\begin{aligned} P_D = P_G &= (1 - P_E)/2 && (7.43)\\ &= 1/2 - \frac{9-2n}{18} && (7.44)\\ &= \frac{n}{9}. && (7.45) \end{aligned}$$

Pour quelles valeurs de P_D, P_G et P_E, l'entropie d'une pesée est-elle maximales ? Trois événements étant possibles, d'après les conséquences du théorème de Jensen, on sait que l'entropie d'une pesée est inférieure ou égale à $\log 3 \approx 1.58$ bits. L'égalité se produit lorsque les trois événements ont des probabilités équiprobables.

En déduire le nombre de pièces n à placer sur chaque plateau pour que l'entropie de la pesée soit maximale. On cherche n telle que $P_D = P_G = P_E = 1/3$, c'est-à-dire :

$$\frac{n}{9} = 1 - \frac{2n}{9}, \tag{7.46}$$

d'où :

$$n = 3. \tag{7.47}$$

La pesée apporte une information maximale de $\log 3 \approx 1.58$ bits si on place 3 pièces dans chaque plateau.

7.5.4 Détermination de la seconde pesée

Le résultat de la première pesée est l'équilibre

On sait donc que les 6 pièces sur les plateau sont vraies et la pièce fausse se trouve parmi les trois pièces restantes.

On peut donc considérer le problème avec trois pièces, dont l'entropie vaut $H = \log 6 = 2.58$ bits, et qui peut donc être résolu en 2 pesées.

En répétant le raisonnement du paragraphe précédent, on place 1 pièce dans chaque plateau : cette expérience possède une entropie de 1.58 bits. Si le résultat de cette seconde pesée est l'équilibre, la pièce fausse se trouve hors du plateau : une troisième pesée déterminera si elle est plus lourde ou plus légère. Cette troisième pesée ne peut donc avoir que deux résultats équiprobables de probabilité $P_D = P_G = 1/2$, l'équilibre étant impossible ($P_E = 0$). On a donc :

$$H(\text{troisième pesée}) = \log 2 = 1\text{bit}. \tag{7.48}$$

Pour ces trois pesées, on a donc une entropie totale :

$$\begin{aligned} H(\text{trois pesées}) &= H(\text{première pesée}) + H(\text{deuxième pesée}) && (7.49)\\ &\quad + H(\text{troisième pesée})\\ &= \log 3 + \log 3 + \log 2\\ &= \log(3^2 \times 2)\\ &= \log 18, \end{aligned}$$

c'est-à-dire qui est égale à l'entropie de l'expérience.

Le résultat de la première pesée est G

On sait alors que les trois pièces écartées sont vraies, et que la pièce fausse se trouve dans un des deux plateaux, à gauche si elle est plus lourde, à droite si elle est plus légère.

Montrer qu'il ne faut pas mélanger les pièces des 2 plateaux et considérer un problème à 6 pièces. Si on remet ensemble ces 6 pièces, le problème a une entropie de :

$$H(\text{problème à 6 pièces}) = \log 12 \approx 3.58\text{bits} \tag{7.50}$$

supérieure à l'information que l'on peut acquérir en deux pesées :

$$H(2 \text{ pesées}) \leq 2 \log 3 \approx 3.16\text{bits}. \tag{7.51}$$

Il faut donc distinguer les pièces de chaque plateau. En effet, en les mélangeant, on perd les informations *"si la pièce fausse est à gauche elle est plus lourde"* et *"si la pièce fausse est à droite elle est plus légère"*, soit une entropie de 1 bit. Si l'on commet cette erreur, le bilan reste malgré tout cohérent, même si on ne peut plus résoudre en trois pesées. En effet, on retrouve l'entropie du problème :

$$\begin{aligned} H(\text{problème}) &= H(\text{problème à 6 pièces}) + H(\text{première pesée}) \\ &\quad - H(\text{perdue en mélangeant}) \\ &= \log 12 + \log 3 - \log 2 \\ &= \log 18 \approx 4.16\text{bits}. \end{aligned} \tag{7.52}$$

On enlève une pièce de chaque plateau. Montrer que cette pesée ne convient pas. On enlève une pièce de chaque plateau, mais on place les pièces écartées près du plateau d'où elles proviennent afin de conserver la mémoire de la première pesée. Dans ce cas, on sait que deux résultats seuls sont possibles : E [1] ou G (car $P_D = 0$), avec les probabilités :

$$\begin{aligned} P_E &= 1/3 \\ P_G &= 2/3, \end{aligned}$$

ce qui correspond à une entropie :

$$\begin{aligned} H(\text{seconde pesée}) &= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \\ &= \frac{\log 3}{3} + \frac{2\log 3}{3} - \frac{2\log 2}{3} \\ &= \log 3 - \frac{2}{3} \\ &\approx 0.92\text{bits}. \end{aligned} \tag{7.53}$$

Les deux premières pesées apportent donc $H = 1.58 + 0.92 = 2.5$ bits. Avec la troisième pesée qui apporte au maximum 1.58 bits, on a donc au plus 4.08 bits, ce qui est plus faible que l'entropie de l'expérience. Cette pesée n'est donc pas judicieuse.

1. P_E est la probabilité que la pièce fausse soit parmi les 2 pièces écartées sur les 6 considérées)

Afin que $p_D \neq 0$, on enlève une pièce de chaque plateau, on permute une pièce de chaque plateau. Calculer les probabilités des résultats de la pesée et son entropie. Calculons P_E :

$$\begin{aligned} P_E &= \text{Pr(pièce fausse est une des deux pièces enlevées)} \\ &= 2/6. \end{aligned}$$

Calculons maintenant P_G :

$$\begin{aligned} P_G &= \text{Pr[(pièce fausse est la pièce restée dans le plateau gauche)} \\ &\quad \text{ou (pièce fausse est la pièce restée dans le plateau droit)]} \\ &= 2/6. \end{aligned} \tag{7.54}$$

De même, on trouve $P_D = 1/3$.

L'entropie de cette pesée est donc maximale et vaut $H(\text{seconde pesée}) = 1.58$ bits.

Il est facile de vérifier qu'une troisième pesée permettra de déterminer la fausse pièce et son poids, quel que soit le résultat de cette pesée.

Chapitre 8

Codage et compression de données

Comme nous l'avons indiqué dans l'introduction, un des objectifs de la théorie de l'information est de fournir des méthodes de compression de l'information. Intuitivement, on comprend qu'un code qui représente les symboles les plus fréquents par des mots-codes les plus courts réalise cet objectif. C'est ce que nous allons voir de façon plus formelle dans ce chapitre.

8.1 Exemples de codes

8.1.1 Définitions

Définition 8.1.1 *Un code C d'une variable aléatoire X est une application de $\mathcal{X}$ vers $\mathcal{D}$, l'ensemble des chaînes de longueur finie réalisées à partir d'un alphabet à D lettres.*

On notera $C(x)$ le mot-code associé au symbole x et $l(x)$ la longueur de ce mot-code.

Définition 8.1.2 *Soient une variable aléatoire X, prenant les valeurs x avec une probabilité $p(x)$, et C un code sur X. La longueur moyenne $L(C)$ du code C est égale à :*

$$L(C) = \sum_{x \in \mathcal{X}} l(x)p(x) \tag{8.1}$$

Sans perte de généralité, lorsque l'alphabet possède D lettres, on pourra supposer $\mathcal{D} = \{0, 1, \ldots, D-1\}$. Si le code utilise un alphabet à deux lettres, c'est un code binaire.

8.1.2 Exemples

Exemple 1

On considère X et les deux codes binaires définis par le tableau ci-dessous.
L'entropie de la source X vaut :

$$H(X) = -\sum_{i=1}^{4} p(x_i) \log p(x_i) = 1.75 \text{ bits.} \tag{8.2}$$

Dans cette équation, l'unité "bit" signifie Binary unIT.

x	$p(x)$	$C_1(x)$	$C_2(x)$
x_1	$1/2$	0	00
x_2	$1/4$	10	01
x_3	$1/8$	110	10
x_4	$1/8$	111	11

TABLE 8.1 – Probabilités et mots-codes de la source X de l'exemple 1.

La longueur moyenne $L(C_1)$ des mots-codes de C_1 est égale à :

$$\begin{aligned} L(C_1) &= \sum_{i=1}^{4} l_1(x_i)p(x_i) \\ &= 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} \\ &= 1.75 \text{ digits binaires.} \end{aligned}$$

De façon similaire, la longueur moyenne $L(C_2)$ vaut :

$$\begin{aligned} L(C_2) &= \sum_{i=1}^{4} l_2(x_i)p(x_i) \\ &= 2 \text{ digits binaires.} \end{aligned}$$

On n'utilisera pas le terme "bit" pour les digits binaires, afin d'éviter la confusion avec l'unité d'information. On remarque que

— la longueur moyenne des mots-codes est plus petite pour C_1 que pour C_2 : C_1 est donc plus *efficace* que C_2.
— le nombre de digits binaires de la longueur moyenne $L(C_1)$ est égale à l'entropie,
— toute suite de mots-codes, de C_1 comme de C_2, correspond à une suite unique de symboles x_i.

Exemple 2

On considère X et les deux codes définis par le tableau ci-dessous.

x	$p(x)$	$C_1(x)$	$C_2(x)$
x_1	$1/3$	0	0
x_2	$1/3$	10	1
x_3	$1/3$	01	2

TABLE 8.2 – Probabilités et mots-codes de la source X de l'exemple 2.

L'entropie de la source X vaut :

$$H(X) = -\sum_{i=1}^{3} p(x_i) \log p(x_i) = 1.58 \text{ bits.} \tag{8.3}$$

La longueur moyenne $L(C_1)$ est égale à :

$$\begin{aligned} L(C_1) &= \sum_{i=1}^{3} l_1(x_i)p(x_i) \\ &= 1 \times \frac{1}{3} + 2 \times \frac{1}{3} + 2 \times \frac{1}{3} \\ &\approx 1.67 \text{ bits.} \end{aligned}$$

De façon similaire, la longueur moyenne $L(C_2)$ vaut :

$$\begin{aligned} L(C_2) &= \sum_{i=1}^{3} l_2(x_i)p(x_i) \\ &= 1 \text{ digit ternaire.} \end{aligned}$$

On remarque que

— la longueur moyenne des codes, utilisant des alphabets différents (ici binaire et ternaire), est difficilement comparable, puisque les unités sont différentes : des digits binaires ou ternaires !

— il est également difficile de comparer ces longueurs moyennes avec l'entropie dont l'unité est binaire (bit = binary unit), dûe au choix du logarithme en base 2,

— les suites de mots-codes de C_1 peuvent être ambigües. Par exemple 1001010 peut correspondre aux mots-codes des suites x_2, x_3, x_1, x_2 ou x_2, x_1, x_2, x_2 ou x_2, x_3, x_3, x_1.

8.1.3 Codes réguliers, déchiffrables et instantanés

Les exemples précédents ont montré que les codes ne possèdent pas tous de bonnes propriétés : il ne suffit pas qu'un code soit une application de $\mathcal{C}$ dans $\mathcal{D}$. Dans ce paragraphe, on s'attachera à définir de façon précise ces propriétés.

Définition 8.1.3 *Un code C est dit régulier (ou non singulier) si chaque lettre $x \in \mathcal{X}$ a une représentation unique $C(x)$, autrement dit si l'application $x \to C(x)$ est injective :*

$$x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j). \tag{8.4}$$

Comme nous l'avons vu dans l'exemple 2 du paragraphe prédédent (C_1), cette propriété n'est pas suffisante. En effet, il faut que cette propriété d'injectivité soit également vraie pour le code des suites de symboles x_i, sinon les suites de mots-codes sont ambigües.

Définition 8.1.4 *L'extension X^n de X est l'ensemble des suites de n symboles de $\mathcal{X}$.*

Définition 8.1.5 *L'extention C^n du code C est l'application des chaînes de longueur n d'éléments de $\mathcal{X}$ en chaînes finies d'éléments de $\mathcal{D}$, définies ainsi :*

$$C^n(x^{(1)}, \ldots, x^{(n)}) = C(x^{(1)})C(x^{(2)}) \ldots C(x^{(n)}), \tag{8.5}$$

où $C(x^{(1)})C(x^{(2)}) \ldots C(x^{(n)})$ est la concaténation des mots-codes $C(x^{(1)})$, $C(x^{(2)}) \ldots$ et $C(x^{(n)})$.

Par exemple, si $C(x_1) = 00$ et $C(x_2) = 11$, on a $C^2(x_1, x_2) = C(x_1)C(x_2) = 0011$.

Définition 8.1.6 *Un code est déchiffrable (ou à décodage unique) si toutes ses extentions sont régulières.*

Cette définition permet d'éviter les problèmes d'ambiguités du décodage, mais ce n'est pas encore satisfaisant. En effet, considérons la source X et les deux codes binaires C_1 et C_2 du tableau 8.3.

x	$C_1(x)$	$C_2(x)$
x_1	0	0
x_2	11	01
x_3	100	011
x_4	101	0111

TABLE 8.3 – Deux codes binaires pour la source X.

Il est facile de vérifier que ces deux codes sont déchiffrables. Cependant, si on considère la suite de lettres $x_1, x_2, x_3, x_1, x_3, x_4$, dont les codes sont données dans le tableau 8.4, on remarque que les mots-codes de C_2 ne peuvent être interprétés que tardivement, après le début du mot-code suivant.

C^6	x_1	x_2	x_3	x_1	x_3	x_4
C_1^6	0	11	100	0	100	101
C_2^6	0	01	011	0	011	0111

TABLE 8.4 – Certains mots-codes de C_2 ne peuvent être décodés que tardivement, après acquisition de la première lettre du mot-code suivant.

On introduit donc la notion de code instantané ou irréductible.

Définition 8.1.7 *Un code est dit instantané ou irréductible s'il vérifie la condition du préfixe, c'est-à-dire s'il n'existe aucun couple (x_i, x_j) pour lequel le mot code $C(x_i)$ est le début du mot-code $C(x_j)$.*

On voit clairement que le code C_2 de l'exemple prédédent ne satisfait pas cette condition : $C_2(x_i)$, $i > 1$, commence toujours avec $C_2(x_{i-1})$! En revanche, on peut facilement vérifier que C_1 vérifie la condition du préfixe.

Les langues usuelles ne sont pas des codes instantanés. Par exemple, en français :
— paillasse, paillasson, paille, pailler, paillis, paillotte, etc.
— soixante, soixante-dix, soixante-dix-sept, etc.

Ces définitions des codes peuvent être schématisées dans le diagramme 8.1.

8.1.4 Exercice

Enoncé

On considère la source X et les cinq codes définis dans le tableau suivant.

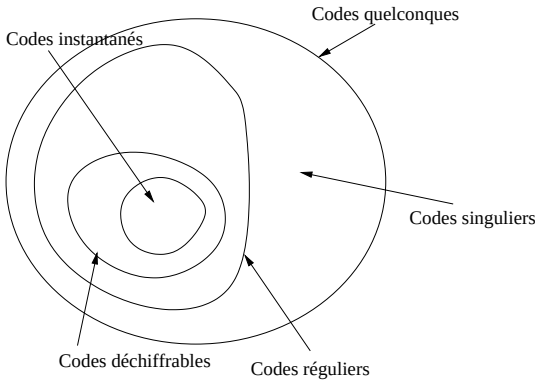


FIGURE 8.1 – Représentation des différents types de codes.

x	C_1	C_2	C_3	C_4	C_5
x_1	0	0	10	0	00
x_2	1	010	00	10	01
x_3	0	01	11	110	10
x_4	1	10	110	111	11

Etudier chaque code et indiquer en argumentant s'il est régulier, déchiffrable ou instantané.

Réponses

Code C_1. Le code C_1 est singulier car l'application n'est pas injective. Plusieurs symboles de X ont le même mot-code : $C(x_1) = C(x_3)$ et $C(x_2) = C(x_4)$. Ce code est sans intérêt.

Code C_2. Ce code est régulier. En revanche, il n'est pas déchiffrable. En effet, $C(x_2) = 010$ peut être décodé comme x_2 ou x_3, x_1 ou x_1, x_4 ! Ce code est également sans intérêt.

Code C_3. Ce code est régulier. Il est aussi déchiffrable. En revanche, il n'est pas instantané. En effet, la condition du préfixe n'est pas vérifiée : $C(x_4)$ commence comme $C(x_3)$.

Code C_4. Ce code est régulier, déchiffrable et instantané. On peut facilement vérifier la condition du préfixe.

Code C_5. Ce code est régulier, déchiffrable et instantané. On peut remarquer que tout code régulier dont les mots-codes ont même longueur est déchiffrable et instantané.

8.2 Construction de codes

La construction de codes instantanés de longueur moyenne minimale est un bon objectif pour la compression d'information.

Bien sûr, on ne peux pas attribuer des mots-codes courts à chaque lettre fréquente tout en respectant la condition du préfixe. Pour cela, on montre dans ce paragraphe que l'ensemble des longueurs des mots-codes doit satisfaire la condition de Kraft.

8.2.1 Inégalité de Kraft

Théorème 8.2.1 *Tout code instantané d'une source X à m mots sur un alphabet de taille D dont les longueurs des mots-codes sont $l_1, l_2, \ldots, l_m$ doit satisfaire l'inégalité :*

$$\sum_{i=1}^{m} D^{-l_i} \leq 1. \tag{8.6}$$

Réciproquement, étant donné un ensemble de longueurs de mots-codes qui satisfait à l'inégalité de Kraft (8.6), il existe un code instantané dont les mots-codes ont ces longueurs.

Preuve. On considère un arbre D-aire, dans lesquel chaque nœud a D descendants. A chaque niveau, les D branches de l'arbre sont associées aux D lettres de l'alphabet du code. La condition du préfixe implique qu'aucun mot-code ne contienne le début d'un mot-code existant : dans l'arbre, chaque mot-code élimine donc toutes les branches descendantes de l'arbre.

Appelons l_{max} la longueur du mot-code le plus long. Tous les noeuds du niveau l_{max} de l'arbre sont donc soit des mots-codes (de longueur maximale), d'autres des descendants de mots-codes, et d'autres des nœuds de branches inutilisées. Cela signifie que le nombre de mots-codes doit être inférieur ou égal à $D^{l_{max}}$.

Au niveau de la longueur l_i, un mot-code de longueur l_i a $D^{l_{max}-l_i}$ descendants de longueur l_{max}, qu'il faut éliminer pour satisfaire la condition du préfixe.

L'ensemble des descendants éliminés de tous les mots-codes doit évidemment être inférieur ou égal au nombre maximal de mots-codes de longueur l_{max}, $D^{l_{max}}$:

$$\sum_{i=1}^{m} D^{l_{max}-l_i} \leq D^{l_{max}}, \tag{8.7}$$

c'est-à-dire :

$$\sum_{i=1}^{m} D^{-l_i} \leq 1. \tag{8.8}$$

Réciproquement, si on se donne des longueurs de mots-codes $l_1, l_2, \ldots, l_m$ qui satisfont l'inégalité de Kraft, on peut construire un arbre D-aire associé à un code qui satisfait la condition du préfixe, c'est-à-dire un code instantané. Il suffit d'étiqueter $C(x_1)$ le premier noeud de niveau l_1 et de supprimer tous ses descendants, etc.

8.2.2 Extension et remarque

Ce résultat peut s'étendre pour un code infini, mais la preuve ne sera pas produite dans ce document.

L'inégalité de Kraft donne deux résultats essentiels :

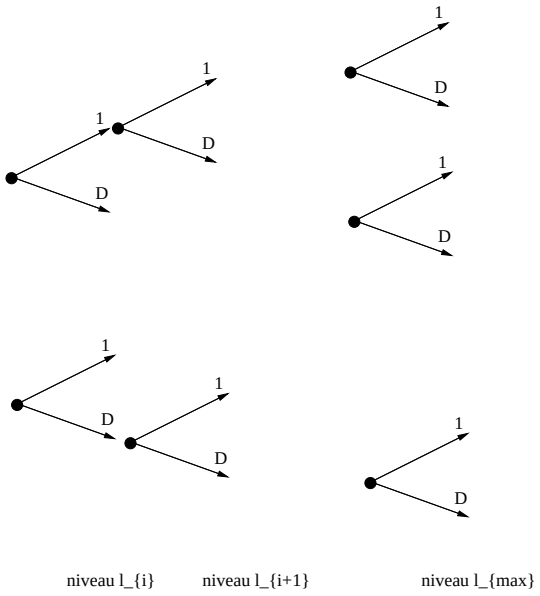


FIGURE 8.2 – Arbre D-aire.

— Si l'inégalité (8.6) n'est pas vérifiée par un code, alors on peut affirmer que ce code n'est pas instantané. Attention, si l'inégalité est vérifiée, on peut seulement dire qu'il est *peut-être* instantané.
— Pour un ensemble de longueurs l_i vérifiant (8.6), on peut construire (*il existe*) un code instantané avec ces longueurs. Attention, cela ne veux pas dire que tous les codes avec ces longueurs sont instantanés : c'est très facile de construire un code avec ces longueurs qui ne soit pas instantané, ni déchiffrable ni même régulier.

8.2.3 Exemples

Reprenons les codes de l'exercice (paragraphe 8.1.4), qui sont reproduits ci-dessous.

x	C_1	C_2	C_3	C_4	C_5
x_1	0	0	10	0	00
x_2	1	010	00	10	01
x_3	0	01	11	110	10
x_4	1	10	110	111	11

Calculons la quantité $\sum_i D^{-l_i}$ pour chacun des codes afin de vérifier si l'inégalité de Kraft est vérifiée.

Code C_1. Le code est binaire, donc $D = 2$. Les longueurs sont $l_1 = l_2 = l_3 = l_4 = 1$, d'où :

$$\sum_i D^{-l_i} = 4 \times 2^{-1} = 2 > 1. \tag{8.9}$$

L'inégalité de Kraft n'est pas vérifiée. Le code C_1 ne peut pas être instantané.

Code C_2. Le code est binaire, donc $D = 2$. On calcule :

$$\sum_i D^{-l_i} = 2^{-1} + 2^{-3} + 2^{-2} + 2^{-2} = 1.125 > 1. \tag{8.10}$$

L'inégalité de Kraft n'est pas vérifiée. Le code C_2 ne peut pas être instantané.

Code C_3. Le code est binaire, donc $D = 2$. On calcule :

$$\sum_i D^{-l_i} = 2^{-2} + 2^{-2} + 2^{-2} + 2^{-3} = 0.875 \leq 1. \tag{8.11}$$

L'inégalité de Kraft est vérifiée. Le code C_3 est *peut-être* instantané.

Code C_4. Le code est binaire, donc $D = 2$. On calcule :

$$\sum_i D^{-l_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 1 \leq 1. \tag{8.12}$$

L'inégalité de Kraft est vérifiée. Le code C_4 est *peut-être* instantané.

Code C_5. Le code est binaire, donc $D = 2$. On calcule :

$$\sum_i D^{-l_i} = 2^{-2} + 2^{-2} + 2^{-2} + 2^{-2} = 1 \leq 1. \tag{8.13}$$

L'inégalité de Kraft est vérifiée. Le code C_5 est *peut-être* instantané.

8.3 Codes optimaux

Au paragraphe précédent, nous avons montré que tout code instantané satisfait l'inégalité de Kraft. Le problème consiste maintenant à élaborer une méthode de construction de codes instantanés de longueur moyenne minimale.

On peut formuler le problème de la façon suivante. Soient x_1, x_2, ..., x_m, les m symboles de $\mathcal{X}$ et l_1, l_2, ..., l_m les longueurs des mots-codes $C(x_1)$, $C(x_2)$, ..., $C(x_m)$ satisfaisant l'inégalité de Kraft, cherchons le code $C(x)$ qui minimise $L(C) = \sum_i l_i p(x_i)$.

Compte tenu de l'inégalité de Kraft, on peut l'exprimer ainsi. Cherchons un code C qui minimise $L(C) = \sum_i l_i p(x_i)$ sous la contrainte $\sum_i D^{-l_i} \leq 1$.

8.3.1 Longueur optimale

En fait, pour simplifier, on ne tient pas compte de la nature entière des l_i et on prendra $\sum_i D^{-l_i} = 1$. Le problème de minimisation sous contrainte revient alors à chercher un code C qui minimise $L(C) = \sum_i l_i p(x_i)$ sous la contrainte $\sum_i D^{-l_i} = 1$.

On peut associer à ce problème la fonction de coût :

$$J = \sum_i p(x_i) l_i + \lambda \sum_i D^{-l_i}, \tag{8.14}$$

où λ est un multiplieur de Lagrange. En dérivant par rapport à l_i, on trouve que les minimas sont obtenus pour :

$$\frac{\partial J}{\partial l_i} = p(x_i) - \lambda D^{-l_i} \ln D = 0, \tag{8.15}$$

c'est-à-dire pour :

$$D^{-l_i} = \frac{p(x_i)}{\lambda \ln D}. \tag{8.16}$$

En reportant cette relation dans la contrainte $\sum_i D^{-l_i} = 1$, on a :

$$\begin{aligned} \sum_i \frac{p(x_i)}{\lambda \ln D} &= 1 \\ \frac{1}{\lambda \ln D} \sum_i p(x_i) &= 1 \\ \lambda &= \frac{1}{\ln D}. \end{aligned} \tag{8.17}$$

On en déduit que, à l'optimum (minimum de J), on doit avoir des longueurs minimales notées l_i^* :

$$p(x_i) = D^{-l_i^*} = \exp(-l_i^* \ln D), \tag{8.18}$$

d'où :

$$\begin{aligned} \ln p(x_i) &= -l_i^* \ln D, \\ l_i^* &= -\frac{\ln p(x_i)}{\ln D}, \\ l_i^* &= -\frac{\log p(x_i)}{\log D}. \end{aligned} \tag{8.19}$$

Finalement, la longueur moyenne minimale L^* vaut :

$$\begin{aligned} L^* &= \sum_i p(x_i) l_i^* \\ &= \sum_i p(x_i) \Big(-\frac{\log p(x_i)}{\log D} \Big) \\ &= \frac{1}{\log D} \Big(-\sum_i p(x_i) \log p(x_i) \Big) \\ &= \frac{H(X)}{\log D}. \end{aligned} \tag{8.20}$$

8.3.2 Théorème

On peut donc énoncer le théorème suivant.

Théorème 8.3.1 *La longueur moyenne $L(C)$ des mots-codes $C(x)$ d'une source X par un code instantané C utilisant un alphabet à D lettres vérifie :*

$$L(C) \geq \frac{H(X)}{\log D}. \tag{8.21}$$

8.4 Bornes

Dans ce paragraphe, nous montrons que l'on peut facilement encadrer la longueur moyenne d'un code instantané. De plus, nous montrons l'intérêt pratique d'un code par paquets de symboles.

8.4.1 Codes mot à mot

Soient l_i^*, $i = 1, 2, \ldots, m$, les valeurs optimales (mais pas forcément entières) des longueurs, il est clair que la longueur réelle l_i des mots-codes $C(x_i)$ est un entier qui vérifie :

$$l_i^* \leq l_i < l_i^* + 1. \tag{8.22}$$

En multipliant par $p(x_i)$ puis sommant sur tous les mots-codes, on a :

$$\sum_i p(x_i) l_i^* \leq \sum_i p(x_i) l_i < \sum_i p(x_i)(l_i^* + 1), \tag{8.23}$$

soit :

$$\frac{H(X)}{\log D} \leq L < \frac{H(X)}{\log D} + 1. \tag{8.24}$$

En codant symbole par symbole les éléments de la source X, on peut toujours construire un code instantané telle que la longueur moyenne est comprise entre la longueur minimale théorique L^* et $L^* + 1$.

8.4.2 Codes par paquets

Considérons maintenant le codage par paquets de n symboles de la source X, c'est-à-dire par éléments de sa n-ième extention X^n. On notera $l(x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ la longueur du mot-code associé à l'élément $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ de probabilité $p(x^{(1)}, x^{(2)}, \ldots, x^{(n)})$.

La longueur moyenne des mots-codes (de paquets de n symboles) est alors :

$$L_n = \sum l(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) p(x^{(1)}, x^{(2)}, \ldots, x^{(n)}). \tag{8.25}$$

En appliquent le théorème sur la longueur minimale, on peut donc écrire :

$$\frac{H(X^n)}{\log D} \leq L_n < \frac{H(X^n)}{\log D} + 1. \tag{8.26}$$

Si les symboles successifs sont indépendants et identiquement distribués (iid), c'est-à-dire que les $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ sont indépendants et identiquement distribués, en appliquant les relations entre entropie jointe et entropies simples, on a $H(X^n) = nH(X)$. Ainsi, la relation (8.26) devient :

$$\frac{nH(X)}{\log D} \leq L_n < \frac{nH(X)}{\log D} + 1, \tag{8.27}$$

soit, après division par n :

$$\frac{H(X)}{\log D} \leq \frac{L_n}{n} < \frac{H(X)}{\log D} + \frac{1}{n}, \tag{8.28}$$

où L_n/n représente la longueur moyenne par symbole des mots-codes par paquets de n.

Si les symboles successifs ne sont pas indépendants, on peut seulement écrire :

$$H(X^n) \leq nH(X). \tag{8.29}$$

La longueur moyenne par symbole devient alors :

$$\frac{H(X^n)}{n \log D} \leq \frac{L_n}{n} < \frac{H(X^n)}{n \log D} + \frac{1}{n}. \tag{8.30}$$

Conclusion. On remarque qu'en codant par paquet de n symboles, la longueur moyenne par symbole est encadrée entre la longueur minimale théorique $H(X)/\log D$ et $H(X)/\log D + 1/n$. En augmentant la taille n du paquet, on peut donc réduire l'intervalle $1/n$ et se rapprocher de la borne. Si les symboles successifs ne sont pas iid, alors la borne est plus faible mais son calcul exact est délicat. L'intervalle reste toujours égal à $1/n$.

8.4.3 Comparaison de deux codes

Pour comparer des codes sur des alphabets différents ou bien des codes symbole à symbole ou par paquets de symboles, on introduit l'efficacité d'un code.

Définition 8.4.1 *On appelle efficacité ρ d'un code C sur un alphabet à D lettres d'une source X dont la longueur moyenne des mots-codes est $L(C)$, le rapport :*

$$\rho = \frac{H(X)}{L(C) \log D}. \tag{8.31}$$

La borne inférieure de la longueur moyenne étant $H(X)/\log D$, on voit que l'efficacité est inférieure ou égale à 1.

8.5 Théorème de Mac Millan

Le théorème de Kraft fournit une condition nécessaire pour les codes instantanés. On peut se demander s'il existe une condition similaire pour les codes simplement déchiffrables.

8.5.1 Théorème

McMillan a montré le théorème suivant.

Théorème 8.5.1 *Les longueurs, l_i, $i = 1, \ldots, m$, de mots-codes d'un code déchiffrable à D lettres doivent satisfaire l'inégalité de Kraft :*

$$\sum_i D^{-l_i} \leq 1.$$

Réciproquement, étant donné un ensemble de longueurs de mots-codes, l_i, $i = 1, \ldots, m$, satisfaisant l'inégalité de Kraft, il existe un code déchiffrable avec ces longueur de mots-codes.

8.5.2 Commentaires

Ce résultat peut sembler surprenant à première vue. Il montre en effet que l'inégalité de Kraft est une condition nécessaire à la fois pour les codes instantanés et déchiffrables.

Attention à utiliser ce résultat avec justesse, comme le théorème de Kraft :
— Si l'inégalité de Kraft est satisfaite, on peut conclure simplement que le code est peut-être déchiffrable et/ou instantané.
— Si l'inégalité de Kraft n'est pas satisfaite, on peut conclure que le code n'est ni déchiffrable ni (*a fortiori*) instantané.

8.6 Codes de Shannon et d'Huffman

Dans ce paragraphe, nous proposons deux méthodes systématiques de construction de codes instantanés dont la longueur moyenne L satisfait l'encadrement $H(X)/\log D \leq L < H(X)/\log D + 1$. Ces deux méthodes peuvent aussi être utilisées pour construire des codes par paquets de n symboles. Dans ce cas, la longueur moyenne par symbole L_n/n satisfait $H(X^n)/n\log D \leq L_n/n < H(X^n)/n\log D + 1/n$.

8.6.1 Code de Shannon

Principe

L'idée est d'attribuer des mots-codes courts aun symboles fréquents. Au paragraphe 8.3.1, nous avons vu que la longueur minimale l_i^* satisfaisait :

$$l_i^* = -\frac{\log p(x_i)}{\log D}.$$

Dans la cas binaire, cette relation se simplifie et $l_i^* = \log[1/p(x_i)]$. En général, l_i^* n'est pas un entier, et on peut choisir pour x_i un mot-code $C(x_i)$ de longueur l_i vérifiant :

$$l_i^* \leq l_i < l_i^* + 1. \tag{8.32}$$

On peut facilement vérifier que la condition de Kraft est vérifiée avec les l_i. En effet,

$$\begin{aligned} -l_i^* \log D \geq -l_i \log D &> -(l_i^* + 1)\log D \\ \ln p(x_i) \geq -l_i \ln D &> \ln p(x_i) + \ln D \\ p(x_i) \geq \exp[-l_i \log D] &> p(x_i)D \\ \sum_i p(x_i) \geq \sum_i D^{-l_i} &> D\sum_i p(x_i), \end{aligned}$$

d'où finalement :

$$\sum_i D^{-l_i} \leq 1. \tag{8.33}$$

Avec ce choix, il est facile de remarquer que, si l'on sait construire un code C avec ces longueurs, la longueur moyenne $L(C)$ vérifie :

$$l_i^* p(x_i) \leq \quad l_i p(x_i) \quad < (l_i^* + 1)p(x_i),$$

$$\begin{aligned}
\sum_i l_i^* p(x_i) \leq \quad & \textstyle\sum_i l_i p(x_i) \quad & < \sum_i (l_i^* + 1) p(x_i), \\
\sum_i -\frac{\log p(x_i)}{\log D} p(x_i) \leq \quad & \textstyle\sum_i l_i p(x_i) \quad & < \sum_i (-\frac{\log p(x_i)}{\log D} + 1) p(x_i), \\
\frac{H(X)}{\log D} \leq \quad & L(C) \quad & < \frac{H(X)}{\log D} + 1.
\end{aligned}$$

Après avoir attribué une longueur à chacun des symboles x_i, d'après le théorème de Kraft, on sait qu'il existe un code instantané qui satisfait cette condition. Il est facile de construire un code instantané à l'aide un arbre D-aire qui satisfait ces longueurs et la condition du préfixe.

Exemple

On considère la source X dont les 5 symboles x_i, $i = 1, \ldots, 5$ dont les probabilités sont données dans le tableau 8.6.1.

x_i	p_i	$l_i^* = -\log p_i$	l_i	$C(x_i)$
x_1	0.25	2	2	00
x_2	0.25	2	2	01
x_3	0.2	2.3	3	100
x_4	0.15	2.7	3	101
x_5	0.15	2.7	3	110

TABLE 8.5 – Exemple de construction d'un code binaire de Shannon

On veut construire un code binaire, c'est-à-dire $D = 2$. On calcule (voir tableau 8.6.1) les longueurs optimales l_i^* (colonne 3) puis les longueurs l_i choisies (colonne 4), vérifiant (8.32). Une fois les longueurs choisies, on construit un arbre binaire ($D = 2$) dont on étiquette les nœuds selon ce choix (Fig. 8.3). Les poids faibles des mots-codes sont les extrémités des branches. L'ensemble des mots-codes de l'arbre est reporté dans la dernière colonne du tableau 8.6.1.

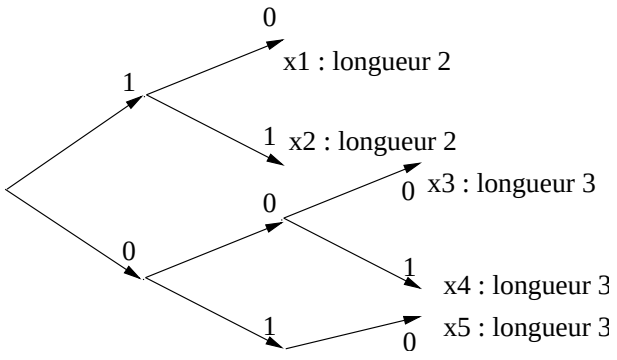


FIGURE 8.3 – Arbre binaire pour la construction d'un code de Shannon.

Pour vérifier l'efficacité du code, on peut calculer si la longueur moyenne est proche de la borne inférieure. Pour cela, on calcule l'entropie de la source X : $H(X) \approx 2.29$ bits, puis la

longueur moyenne :

$$\begin{aligned} L(C) &= \sum_{i=1}^{5} p_i l_i \\ &= 0.25 \times 2 + 0.25 \times 2 + 0.2 \times 3 + 0.15 \times 3 + 0.15 \times 3 \\ &= 2.5 \text{ bits.} \end{aligned} \tag{8.34}$$

On remarquue que $L(c)$ satisfait bien l'encadrement :

$$\begin{aligned} H(X)/\log D \leq L(C) < H(X)/\log D + 1 \\ 2.29/\log 2 \leq L(C) < 2.29/\log 2 + 1 \\ \text{soit : } 2.29 \leq 2.5 < 3.29, \end{aligned}$$

et se trouve même assez proche de la borne inférieure. Malgré tout, on se rend compte que l'arbre n'est pas utilisé au mieux en respectant les longueurs l_i. En effet, la dernière branche de l'arbre n'est pas utilisé. On pourrait gagner en longueur moyenne en affectant un mot-code à deux lettres au symbole x_3 (Fig. 8.4).

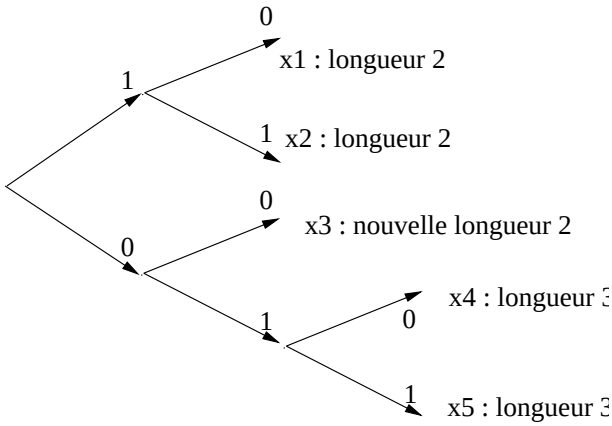


FIGURE 8.4 – Arbre binaire qui améliore le code de Shannon.

Avec ce code C', on trouve alors la longueur moyenne :

$$\begin{aligned} L(C') &= 0.25 \times 2 + 0.25 \times 2 + 0.2 \times 2 + 0.15 \times 3 + 0.15 \times 3 \\ &= 2.3 \text{ bits.} \end{aligned} \tag{8.35}$$

Cette longueur moyenne est vraiment très proche de la borne inférieure.

8.6.2 Code d'Huffman

Nous avons vu au paragraphe 8.3.1 les bornes des longueurs moyennes applicables aux codes instantanés et déchiffrables. La construction du code de Shannon, très simple dans son principe, ne conduit pas à des codes optimaux. Dans ce paragraphe, nous présentons une autre méthode de construction systématique de codes optimaux, due à Huffman.

Lemme préliminaire

Comme précédemment, la source X à coder est constituée de m symboles, $x_1, x_2, \ldots, x_n$, que l'on a ordonnés de sorte que les probabilités soient décroissantes $p(x_1) \geq p(x_2) \geq \ldots \geq p(x_n)$.

Lemme 8.6.1 *Pour toute distribution, $p(x_i)$, $i = 1, \ldots, m$, il existe un code instantané optimal, c'est-à-dire de longueur moyenne minimale, qui satisfait les trois propositions suivantes :*

- *si $p(x_j) > p(x_k)$, alors $l_k \geq l_j$,*
- *les deux mots-codes les plus longs ont même longueur,*
- *les deux mots-codes les plus longs diffèrent seulement par les bits de poids faibles.*

Preuve. Montrons d'abord la première proposition : si $p(x_j) > p(x_k)$, alors $l_k \geq l_j$.
Soient C un code optimal et C' un code dans lequel les mots-codes $C(x_j)$ et $C(x_k)$ sont permuttés par rapport au code C : on a donc $l'_j = l_k$ et $l'_k = l_j$. On peut alors écrire :

$$\begin{aligned} L(C') - L(C) &= \sum_i p(x_i) l'_i - \sum_i p(x_i) l_i \\ &= p(x_j) l_k + p(x_k) l_j - p(x_j) l_j - p(x_k) l_k \\ &= (p(x_j) - p(x_k))(l_k - l_j). \end{aligned} \tag{8.36}$$

Or $p(x_j) - p(x_k)$ étant positif par hypothèse, et C étant optimal, les longueurs moyennes doivent vérifier $L(C') - L(C) \geq 0$. D'après la relation (8.36), on en déduit que $(l_k - l_j) \geq 0$, c'est-à-dire $l_k \geq l_j$.

Montrons maintenant la seconde proposition : les deux mots-codes les plus longs ont même longueur.
Si les deux mots-codes les plus longs (d'un code optimal) n'avaient pas même longueur, on pourrait supprimer le dernier bit du plus long des deux sans détruire la condition du préfixe, ce qui fournirait un code meilleur, ce qui est contraire à l'hypothèse que le code est optimal.

La troisième proposition ne sera pas démontrée. Mais elle est évidente si l'on respecte la condition du préfixe.

Principe de construction du code d'Huffman

A l'aide du lemme précédent, Huffman propose la méthode de construction suivante :

1. On ordonne les m symboles x_i de façon à ce que leur probabilités soient décroissantes,
2. On traite les D (2 dans le cas binaire) symboles de probabilités les plus faibles et on attribue à chacun le poids faible de leur mot-code : dans le cas binaire, 0 pour l'un et 1 pour l'autre.
3. On considère ce groupe de D symboles, affecté d'une probabilité égale à la somme des probabilités. On forme une source auxiliaire possédant $m - D + 1$ symboles, que l'on ordonne selon les probabilités décroissantes.
4. Si le nombre de symboles est égal à D, on étiquette les symboles et on arrête, sinon on recommence à l'étape 1.

Exemple 1

On considère la source X_0 à 5 symboles dont les probabilités, classées de manière décroissante, sont données dans le tableau 8.5.

Symboles X_0	Proba. X_0	Proba. X_1	Proba. X_2	Proba. X_3	Mots-codes
x_1	0.25	0.30	0.45	0.55 (0)	01
x_2	0.25	0.25	0.30 (0)	0.45 (1)	10
x_3	0.20	0.25 (0)	0.25 (1)	-	11
x_4	0.15 (0)	0.20 (1)	-	-	000
x_5	0.15 (1)	-	-	-	001

FIGURE 8.5 – Exemple de construction d'un code binaire d'Huffman

□ Commençons par construire un code d'Huffman binaire C_2. Pour cela, considèrons les deux symboles de probabilités les plus faibles, et attribuons le poids faible 0 à l'un et 1 à l'autre (l'affectation n'a pas d'importance, puisqu'il auront tous les deux la même longueur). Considérons maintenant l'ensemble de ces deux symboles, dont la probabilité vaut $0.15 + 0.15 = 0.30$. On forme donc la source secondaire X_1, constituée de $m - D + 1 = 5 - 2 + 1 = 4$ symboles que l'on classe selon les probablitées décroissantes. Dans le tableau 8.5, les flèches indiquent la correspondance entre les probabilités des symboles de deux sources successives. On répète la procédure jusqu'à ce que la source auxiliaire ne compte plus que 2 ($D = 2$) symboles. Les mots-codes sont ensuite attribués en partant de chaque symbole, et en recopiant les bits rencontrés du poids faibles (colonne de gauche) vers les poids forts (colonne la plus à droite).

Calculons la longueur moyenne du code C_2 ainsi obtenu.

$$\begin{aligned} L(C_2) &= 0.25 \times 2 + 0.25 \times 2 + 0.2 \times 2 + 0.15 \times 3 + 0.15 \times 3 \\ &= 2.3 \text{ bits.} \end{aligned} \tag{8.37}$$

□ Construisons maintenant un code d'Huffman ternaire C_3. Pour cela, on considère les trois symboles de probabilités les plus faibles, et on leurs attribue 0, 1 ou 2 pour poids faible. Puis on construit une source auxiliaire X_1 à $m - D + 1 = 5 - 3 + 1 = 3$ symboles. L'ensemble des étapes, comme expliqué précédemment, est détaillé dans le tableau 8.6.

Calculons la longueur moyenne du code C_3 ainsi obtenu.

$$\begin{aligned} L(C_3) &= 0.25 \times 1 + 0.25 \times 1 + 0.2 \times 2 + 0.15 \times 2 + 0.15 \times 2 \\ &= 1.5 \text{ digits ternaires.} \end{aligned} \tag{8.38}$$

□ Pour comparer les deux codes, nous utilisons l'efficacité, définie au paragraphe 8.4.3.

$$\rho = \frac{H(X}{L(C) \log D}.$$

On calcule déjà l'entropie $H(X) \approx 2.29$ bits. Pour le code C_2, on a :

$$\rho_2 = \frac{H(X)}{L(C_2) \log D}$$

Symboles X_0	Proba. X_0	Proba. X_1	Mots-codes
x_1	0.25	0.50 (0)	1
x_2	0.25	0.25 (1)	2
x_3	0.20 (0)	0.25 (2)	00
x_4	0.15 (1)	-	01
x_5	0.15 (2)	-	02

FIGURE 8.6 – Exemple de construction d'un code ternaire d'Huffman

$$\begin{aligned} &= \frac{2.29}{2.30 \times \log 2} \\ &= 0.996. \end{aligned} \tag{8.39}$$

Pour le code C_3, on a :

$$\begin{aligned} \rho_3 &= \frac{H(X)}{L(C_2) \log D} \\ &= \frac{2.29}{1.5 \times \log 3} \\ &= 0.966. \end{aligned} \tag{8.40}$$

On voit que, pour cette source, le code binaire C_2 est plus efficace que le code ternaire C_3.

Exemple 2

Un second exemple avec une source à 4 symboles (d'entropie $H(X) = 1.75$ bits) est proposé dans ce paragraphe. Le tableau 8.7 donne les étapes du code binaire ; le tableau 8.8 donne les étapes du code ternaire.

La longueur moyenne du code binaire C_2 vaut $L(C_2) = 1.75$ bits. L'efficacité du code vaut $\rho_2 = 1$.

La longueur moyenne du code ternaire C_3 vaut $L(C_3) = 1.5$ digits ternaires. L'efficacité du code vaut $\rho_3 = 0.738$.

On remarque que la construction du code ternaire d'Huffman peut être améliorée car un mot-code à un digit n'est pas utilisé. Ce code, noté C'_3 est présenté dans le tableau 8.9. Sa longueur moyenne vaut $L(C'_3) = 1.25$ digits ternaires. Son efficacité est $\rho'_3 = 0.886$.

En conclusion, on remarque que, pour cette source, le code binaire est le meilleur. Il a l'efficacité maximale. On pouvait s'y attendre car les probabilités $p(x_i)$ des mots-codes sont exactement des puissances (négatives) de 2, ce qui correspond à la condition optimale (8.18) trouvée au paragraphe 8.3.1 :

$$p(x_i) = D^{-l_i^*}.$$

Symboles X_0	Proba. X_0	Proba. X_1	Proba. X_2	Mots-codes
x_1	0.50	0.50	0.50 (0)	0
x_2	0.25	0.25 (0)	0.50 (1)	10
x_3	0.125 (0)	0.25 (1)	-	110
x_4	0.125 (1)	-	-	111

FIGURE 8.7 – Exemple de construction d'un code binaire d'Huffman

Symboles X_0	Proba. X_0	Proba. X_1	Mots-codes
x_1	0.50	0.50 (0)	0
x_2	0.25 (0)	0.50 (1)	10
x_3	0.125 (1)	-	11
x_4	0.125 (2)	-	12

FIGURE 8.8 – Exemple de construction d'un code ternaire d'Huffman

Symboles X_0	Proba. X_0	Mots-codes
x_1	0.50	0
x_2	0.25	1
x_3	0.125	20
x_4	0.125	21

FIGURE 8.9 – Code ternaire amélioré d'Huffman

Quatrième partie

Travaux dirigés

Objectifs

Les énoncés d'exercices seront résolus pendant les séances de travaux dirigés. Ils ont été choisis pour illustrer les différents points du cours (détection et estimation) sans présenter de difficultés calculatoires. Ils sont complétés par la résolution du devoir de l'année précédente, lors de la dernière séance de travaux dirigés. Il n'y a pas de séances de travaux dirigés de théorie de l'information, car les exercices sont intégrés dans les séances de cours.

Deux séances de rappel de probabilités précèdent ces exercices.

Des exercices supplémentaires sont accessibles dans les ouvrages dont les références sont données dans l'introduction de ce document.

Exercices de détection

1.1 Détection binaire 1

On considère un problème de détection binaire dans lequel l'observation r vaut :

$$r = \begin{cases} n, & \text{si } H_0, \\ s+n, & \text{si } H_1, \end{cases} \tag{1.41}$$

où s et n sont deux variables aléatoires indépendantes.

On connaît les densités de probabilité de s et du bruit n :

$$p_s(u) = \begin{cases} a\exp(-au), & \text{si } u \geq 0, \\ 0, & \text{sinon}, \end{cases} \tag{1.42}$$

et

$$p_n(u) = \begin{cases} b\exp(-bu), & \text{si } u \geq 0, \\ 0, & \text{sinon.} \end{cases} \tag{1.43}$$

1. Ecrire le rapport de vraisemblance.
2. En distinguant les trois cas $b-a>0$, $b-a<0$ et $b=a$, montrer que le test du rapport de vraisemblance peut s'écrire :

$$r \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \gamma \tag{1.44}$$

   et déterminer γ, en fonction des probabilités *a priori*, P_i, et des coûts, C_{ij}.
3. Exprimer la probabilité de fausse alarme $P_F = \Pr(\text{décider } H_1/H_0 \text{ vraie})$, sous forme d'une intégrale dépendant de γ.
4. En déduire le test de Neyman-Pearson obtenu en imposant $P_F = \alpha$.

1.2 Détection binaire 2

On considère un système de transmission numérique dont les deux états 0 et 1 sont représentés au niveau du signal par des tensions α et β, respectivement. La transmission est perturbée par un bruit additif n de densité de probabilité (ddp) connue. On mesure donc :

$$r = \begin{cases} \alpha+n, & \text{si } H_0, \\ \beta+n, & \text{si } H_1. \end{cases} \tag{1.45}$$

On considère dans cet exercice deux modèles de bruit, et on répondra à toutes les questions pour les deux modèles :

$$\begin{array}{c}\text{Modèle } M1 \text{ : la ddp du bruit est gaussienne } n \sim N(0,\sigma^2),\\ \text{Modèle } M2 \text{ : la ddp du bruit est uniforme dans } [-L,+L].\end{array} \tag{1.46}$$

1. Calculer le rapport de vraisemblance $\Lambda(r)$, et représenter le graphiquement.
2. On veut utiliser le critère de Bayes.
   (a) Quelles données supplémentaires sont nécessaires ?
   (b) Donner le test du rapport de vraisemblance qui minimise le critère de Bayes sous la forme
   $$r \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \gamma \tag{1.47}$$
   et déterminer γ.
   (c) Calculer formellement les probabilités de fausse alarme, P_F, et de détection, P_D.
3. On pose $\alpha = 0$, $\beta = 2$, $\sigma = 1$, $L = \sqrt{3}$ et $\eta = 1$. Calculer les valeurs des probabilités de fausse alarme, P_F, et de détection, P_D.
4. Tracer les courbes $P_D(\gamma)$ et $P_F(\gamma)$ pour le modèle $M2$.
5. On impose $P_F = P_{F_0} = 0.01$ pour concevoir un test de Neyman-Pearson. Pour les deux modèles :
   (a) Calculer la valeur du seuil γ qui correspond à cette contrainte,
   (b) Calculer la probabilité de détection P_D.
   (c) Placer ce point dans la courbe COR $P_D(P_F)$.
6. On suppose maintenant que les coûts sont égaux à $C_{ij} = 1 - \delta_{ij}$. En utilisant le cours, écrire le risque de Bayes en fonction de P_1, P_F et P_M (la probabilité d'oubli).
   (a) Comment peut-on interpréter le risque ?
   (b) Calculer le risque de Bayes pour $P_1 = 0$, $P_1 = 0.25$, $P_1 = 0.5$, $P_1 = 0.75$ et $P_1 = 1$. Tracer les points $\mathcal{R}_{Bayes}(P_1)$.
   (c) P_F et P_M dépendant de P_1, on note P_F^* et P_M^* les probabilités calculées pour $P_1 = P_1^*$. Si $P_1 \neq P_1^*$ que devient le risque et quelle est sa représentation graphique $\mathcal{R}(P_1)$? Tracer le risque pour $P_1^* = 0.25$.
   (d) Donner la condition MINIMAX. Tracer le risque $\mathcal{R}_{minimax}$ correspondant. Calculer pour les deux modèles P_F et P_D.
7. Proposer une réalisation matérielle de ce détecteur.

1.3 Détection binaire dans un espace à deux dimensions

On considère le problème de détection binaire, dans lequel on réalise N observations indépendantes, représentées par un vecteur $\mathbf{r}$ de $\mathbb{R}^N$. Chaque mesure élémentaire r_i suit la densité de probabilité conditionnelle :

$$p(r_i/H_k) = \frac{1}{\sqrt{2\pi}\sigma_k} exp\Big[-\frac{(r_i - m_k)^2}{2\sigma_k^2}\Big], \tag{1.48}$$

avec $k = 0$ ou $k = 1$.

1. Calculer le test du rapport de vraisemblance et exprimer le en fonction des grandeurs :

$$I_1 = \sum_{i=1}^{N} r_i, \tag{1.49}$$

et

$$I_2 = \sum_{i=1}^{N} r_i^2. \tag{1.50}$$

2. Tracer les régions de décision dans le plan (I_1, I_2) dans le cas particulier : $m_0 = 1$, $m_1 = 2$, $\sigma_1 = 1$, $\sigma_0 = 2$.
3. Proposer une réalisation électronique de ce détecteur.

1.4 Détection ternaire

On veut détecter la source S_k qui émet le signal reçu par un récepteur. Trois sources S_i peuvent émettre le signal avec la même probabilité *a priori*. On observe donc :

$$r = s_k, \text{ si } S_k \text{ émet} \tag{1.51}$$

On veut trouver le test qui minimise l'erreur totale.

1. Dans l'expression classique du risque de Bayes, comment choisir les valeurs des probabilités *a priori* P_i et des coûts C_{ij} ?
2. Exprimer le test du rapport de vraisemblance qui minimise le critère de Bayes à partir des densités de probabilités conditionnelles $p(r/H_i)$.
3. Les trois sources émettent des signaux gaussiens :

$$\begin{array}{l} \text{Source } S_0 : p(r/S_0) \sim N(0, \sigma_a^2), \\ \text{Source } S_1 : p(r/S_1) \sim N(m, \sigma_a^2), \\ \text{Source } S_2 : p(r/S_2) \sim N(0, \sigma_b^2), \end{array} \tag{1.52}$$

   avec $m > 0$ et $\sigma_b > \sigma_a$.

   (a) Tracer approximativement les densités de probabilités conditionnelles.

   (b) Calculer les valeurs de r qui définissent les frontières des trois régions de décision, et donner analytiquement les critères de décision des différentes sources.

4. On se place dans le cas particulier $\sigma_a = m$ et $\sigma_b^2 = 2\sigma_a^2$. Calculer en fonction de m les valeurs trouvées à la question précédente. Tracer les régions correspondant aux trois régions sur l'axe réel représentant la mesure r.
5. Calculer pour les valeurs trouvées à la question précédente, les probabilités d'erreur $P(\text{erreur}/S_i)$. En déduire la probabilité totale d'erreur.

Exercices d'estimation

1.5 Prédiction d'un signal aléatoire

Soit un signal centré gaussien $x(t)$ stationnaire au second ordre. On note $E[x(t)x(t-\tau] = \Gamma_{xx}(\tau)$ sa fonction d'auto-corrélation. On veut estimer $x(t+\theta)$ ($\theta \geq 0$) à partir de l'observation de $x(t)$.

On pose $x_1 = x(t)$ et $x_2 = x(t+\theta)$; x_1 et x_2 sont donc conjointement gaussiens, mais pas indépendants (sauf si $\Gamma_{xx}(\tau) = \Gamma_{xx}(0)\delta(\tau)$). Pour le vecteur aléatoire $\mathbf{x} = (x_1, x_2)^T$ La densité de probabilité conjointe s'écrit donc :

$$p(\mathbf{x}) = p(x_1, x_2) = \frac{1}{(2\pi|\det \mathbf{\Gamma}|^{1/2})} \exp\Big[-\frac{1}{2}\mathbf{x}^T\mathbf{\Gamma}^{-1}\mathbf{x}\Big]. \tag{1.53}$$

1. Calculer la matrice de variances-covariances $\mathbf{\Gamma}$ de $\mathbf{x}$, et son inverse $\mathbf{\Gamma}^{-1}$.
2. Calculer la densité de probabilité *a posteriori* $p(x_2/x_1)$. En déduire l'estimateur du maximum *a posteriori* (MAP), $\hat{x}_2^{map}(x_1)$.
3. Mettre la densité $p(x_2/x_1)$ sous la forme d'une densité de probabilité gaussienne. En déduire l'estimateur des moindres carrés $\hat{x}_2^{ls}(x_1)$.
4. Comparer à l'estimateur MAP. Ce résultat était-il prévisible ?
5. Retrouver ce résultat en appliquant le théorème de la projection orthogonale (cours de filtrage optimal de P.-O. Amblard) à l'estimation de $x_2 = kx_1$.

1.6 Estimation d'un paramètre déterministe

Un dispositif reçoit le signal vectoriel :

$$\mathbf{r} = a\mathbf{s} + \mathbf{n}, \tag{1.54}$$

dans lequel a est un scalaire inconnu, $\mathbf{s}$ est un vecteur réel certain connu, centré et $\mathbf{n}$ est un vecteur réel aléatoire gaussien de matrice de variances-covariances $\mathbf{\Gamma} = E[\mathbf{n}\mathbf{n}^T]$.

1. Calculer la densité de probabilité du vecteur gaussien $\mathbf{n}$.
2. Calculer la densité de probabilité $p(\mathbf{r}/a)$.
3. En déduire l'estimateur de maximum de vraisemblance $\hat{a}_{ml}(\mathbf{r})$.
4. Calculer directement le biais et la variance de l'estimateur.
5. Montrer que l'estimateur $\hat{a}_{ml}(\mathbf{r})$ est efficace.
6. En utilisant les bornes de Cramer-Rao, calculer la variance de l'estimateur.

1.7 Bornes de Cramer-Rao d'un estimateur biaisé

On considère $\hat{a}(\mathbf{r})$ un estimateur biaisé de a, tel que $E[\hat{a}(\mathbf{r})] = a + b(a)$, où $b(a)$ est le biais, fonction de a. En calculant la dérivée par rapport à a de l'espérance $E[\hat{a}(\mathbf{r}) - (a + b(a))]$ et en suivant le principe de la démonstration faite en cours, montrer que la borne de Cramer-Rao d'un estimateur biaisé est égale à :

$$E[(\hat{a}(\mathbf{r}) - a)^2] \geq \frac{\left(1 + \frac{db(a)}{da}\right)^2}{E\left[\left(\frac{\partial \ln p(\mathbf{r}/a)}{\partial a}\right)^2\right]}. \tag{1.55}$$

1.8 Estimation d'un processus de Poisson

On considère un processus de Poisson stationnaire $x(t)$. Les événements sont des impulsions que l'on peut supposer infiniment brèves et d'amplitude constante. La probabilité d'observer n impulsions pendant un temps τ est égale à :

$$\Pr(n/\tau) = \frac{(k\tau)^n}{n!} \exp(-k\tau). \tag{1.56}$$

Le paramètre k du processus est une variable déterministe inconnue que l'on désire estimer. Pour cela, on observe le signal pendant un temps de mesure T.

1. Est-il nécessaire d'enregistrer la dates d'arrivée des impulsions ou suffit-il de compter ces impulsions ?
2. Calculer l'estimateur du maximum de vraisemblance $\hat{k}_{ml}(n)$.
3. Calculer le biais de cet estimateur.
4. En utilisant les inégalités de Cramer-Rao, calculer la borne inférieure de la variance de $(\hat{k}_{ml}(n) - k)$.
5. Montrer que l'estimateur est efficace. En déduire la valeur exacte de la variance.

1.9 Estimation de la durée d'une expérience

On mesure une quantité scalaire y qui est la somme de N échantillons x_k prélevés sur un bruit blanc gaussien, centré et de variance égale à σ_x^2 :

$$y = \sum_{k=1}^{N} x_k. \tag{1.57}$$

L'expérience commence au temps $t = 0$. Le signal $x(t)$, échantillonné selon la période d'échantillonnage T_e, fournit les échantillons $x_k = x(kT_e)$ ($k \in \mathbb{N}^*$). On désire estimer la durée T de l'expérience. Pour cela, on estimera le nombre N de mesures, afin d'en déduire ensuite $\hat{T}$ par la relation $\hat{T} = \hat{N}T_e$.

1. Calculer la densité de probabilité de l'observation y, sachant N, $p(y/N)$, en fonction de σ_x.
2. Calculer l'estimateur du maximum de vraisemblance $\hat{N}_{ml}(y)$.

3. Calculer le biais de cet estimateur. Cet estimateur est-il efficace ?
4. Calculer la variance de l'estimateur (a) par un calcul direct, (b) en utilisant les bornes de Cramer-Rao.
5. Comment peut-on améliorer l'estimateur en tenant compte de la nature entière de N ?