AURÉLIEN GARIVIER

Modèles contextuels et alphabets infinis en théorie de l'information

Thèses d'Orsay, 2006

<http://www.numdam.org/item?id=BJHTUP11_2006__0708__P0_0>

63657

UNIVERSITÉ
PARIS-SUD 11

N° d'ordre: 8461

UNIVERSITE PARIS-SUD
FACULTE DES SCIENCES D'ORSAY

THESE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITE PARIS XI

Spécialité : Mathématiques

par

Aurélien GARIVIER

MODELES CONTEXTUELS ET ALPHABETS INFINIS EN THEORIE DE L'INFORMATION

Soutenue le 23 novembre 2006 devant la commission d'examen :

M.	Stéphane BOUCHERON	(Examinateur)
M.	Fabrice GAMBOA	(Examinateur)
Mme	Elisabeth GASSIAT	(Examinatrice)
M.	Laszlo GYORFI	(Rapporteur)
M.	John KIEFFER	(Rapporteur)
M.	Pascal MASSART	(Président)
M.	Eric MOULINES	(Examinateur)

La simplicité n'a pas besoin d'être simple, mais du complexe resserré et synthétisé.

Alfred Jarry (1873 - 1907)
Les Minutes de sable mémorial

Remerciements

Je dois le bon accomplissement de cette thèse à plusieurs personnes : chacune, dans son rôle, m'a littéralement porté jusqu'à son terme.

En premier lieu à ma directrice, Elisabeth, qui m'a tout simplement montré par sa détermination et son professionnalisme, mais aussi par sa patience, sa générosité et par la variété de ses centres d'intérêt, comment être chercheur. A Stéphane Boucheron qui, choisissant un rôle moins formel, m'a indiqué le chemin et accompagné avec l'indécrottable bonne humeur qui ne le quitte jamais. A Pascal Massart, qui a guidé mes premiers pas à Orsay et dans les statistiques, et dont le volontarisme et la bienveillance dans sa charge de directeur d'èquipe auront allégé les peines de tant de doctorants ! Aux professeurs de l'Université Paris Sud, et en particulier à Pierre Pansu et Wendelin Werner, toujours disponibles et de bon conseil, qui sont pour ce métier les meilleurs modèles que l'on puisse imaginer ; je tiens ici également à mentionner l'aide sympathique et néanmoins décisive de Jean-Louis Nicolas, de l'ENS Lyon, dans un de mes travaux. Merci aussi à mes rapporteurs, John Kieffer et László Györfi (qui a très gentiment fait le voyage depuis Budapest pour ma soutenance), ainsi qu'à Eric Moulines et Fabrice Gamboa, membres de mon jury, dont le jugement expert m'encourage à aller plus loin.

D'autres personnes m'ont beaucoup apporté, que ce soit sur un plan scientifique ou non. Antoine Chambaz me fournit une transition parfaite : j'ai eu la chance d'apprendre en travaillant avec le chercheur, et de profiter des conseils avisés et généreux de celui dont j'avais repris la poste et le bureau. Et c'était une grande chance de croiser quelques générations de thésards de l'université Paris Sud : Magalie, Cedric, Mina, Christian, Gilles, Karine, Khaled, Marie, Marion, Ismaël(s), Katia, Besma, Sophie, Neil, Laurent... la variété des caractères et des origines ont fait de ces trois années une véritable ouverture sur le monde, venir travailler à Orsay aura toujours été un plaisir.

Enfin, il y a ceux qui ont rendu tout cela possible, et qui y donnent un sens. Mes parents, à leurs façons, très différentes, ont posé pour moi toutes les fondations que l'on peut espérer. Elena, soutien indéfectible et indestructible, a taillé l'ébauche et précisé sa destination. Florian, s'épanouissant avec Laurence, m'a accompagné et poussé à la façonner par moi-même. Et A.

Table des matières

Introduction

0.1 Présentation

L'objectif de cette thèse est l'examen de quelques idées récentes en théorie de l'information et leur exploitation pour l'étude de problèmes statistiques de choix de modèles. La considération d'alphabets infinis, ou de processus à mémoire infinie, constitue le fil conducteur de ces travaux.

Cette introduction situe et présente les résultats obtenus dans le cadre de cette thèse, en collaboration lorsque cela est indiqué. Il y est fait référence aux théorèmes et aux preuves qui se trouvent dans les chapitres suivants. Chacun de ceux-ci consitue une unité suffisamment autonome pour être lu indépendamment des autres. Le chapitre 1 est issu d'une collaboration avec Stéphane Boucheron et Elisabeth Gassiat. Les chapitres 2 à 5 correspondent à des publication soumises. Les articles reprenant le contenu des chapitres 3 et 4 ont été acceptés et sont en cours de publication par le journal *IEEE Transactions on Information Theory*. Le chapitre 5 est le fruit d'une collaboration avec Antoine Chambaz et Elisabeth Gassiat.

Nous commençons par un exposé succinct du cadre classique de la théorie du codage universel sans perte qui se limite aux aspects indispensables à la compréhension de nos travaux. Il nous permet de présenter les résultats que nous avons obtenus sur les processus sans mémoire à valeur dans des *alphabets infinis* ; nous avons considéré d'une part des classes dites *enveloppes*, et d'autre part le codage des *motifs*. Nous exposons ensuite des développements récents sur le codage de processus à mémoire finie appelés sources à arbres de contexte. Nous montrons ici l'*efficacité* d'un algorithme en découlant, la “Context Tree Weighting method” (CTW), sur des classes de processus à mémoire infinie : les processus de renouvellement. Des inégalités de mélange, nécessaires à la construction de l'algorithme CTW, seront ensuite exploitées pour un problème statistique : nous complétons un résultat de *consistance* pour l'estimateur *BIC* (Bayesian Information Criterion) des arbres de contexte. Enfin, des inégalités du même type sont prouvées pour des chaînes de Markov cachées à *émission poissonienne et gaussienne* ; sans interprétation dans la théorie du codage, elles sont appliquées à la construction d'*estimateurs de l'ordre*, c'est-à-dire du nombre d'états cachés, dont est montrée la consistance.

0.2 Théorie du codage

0.2.1 Le problème du codage universel

0.2.1.1 Codage d'une source connue

La théorie du codage a été initiée par Shannon en 1948 dans son célèbre article [Sha48]. On peut l'introduire ainsi :

- On dispose d'un ensemble A de symboles que l'on appellera *alphabet.* Outre les exemples linguistiques, on pensera que A peut désigner l'alphabet binaire, l'ensemble $\{A, C, T, G\}$ des bases de nucléotides qui forment l'ADN, l'ensemble des valeurs possibles pour un pixel d'image numérique ou encore l'ensemble $\mathbb{N}$ des entiers naturels.
- La *source* P émet successivement des symboles $X_1, X_2, \ldots$ de l'alphabet A, dont la concaténation forme des *messages* $X_1^n = X_1 \ldots X_n$. Cela signifie en langage probabiliste que $(X_n)_n$ est un processus de loi P, que l'on supposera ici toujours stationnaire.
- L'objectif est de transmettre ces messages sans perte par un canal binaire, c'est-à-dire de construire des fonctions de codage ϕ_n injectives prenant en argument des messages de taille n à valeur dans l'ensemble des suites binaires :
$$\phi_n : A^n \to \{0,1\}^* = \bigcup_{k \geqslant 0} \{0,1\}^k.$$
La longueur du code $\phi_n(x_1^n)$ sera notée dans la suite $L(x_1^n)$.
- On veut alors construire ϕ_n de sorte que la longueur moyenne du code soit la plus petite possible, autrement dit minimiser $\mathbb{E}_P[L(X_1^n)]$.

Supposons pour l'instant que les propriétés statistiques de la source P sont connues. Ce premier problème est alors résolu par les arguments suivants, très bien présentés dans les livres de synthèse [CK81] et [CT91] :

1. L'*inégalité de Kraft* d'abord montrée pour les codes préfixes (c'est-à-dire tels qu'aucun mot de code n'est le préfixe d'un autre), puis pour tous les codes non ambigus (pour lesquels une concaténation quelconque de mots de code $\phi_n(w_1) \ldots \phi_n(w_n)$ permet de retrouver les messages $w_1 \ldots w_n$, voir [CT91, Bou00]) :

   Proposition 1. *[Kra49] :*
$$\sum_{x \in A^n} 2^{-L(x)} \leqslant 1.$$

A toute longueur de code L on peut donc associer une sous-probabilité $q_L(\cdot) = 2^{-L(\cdot)}$.

2. On déduit aisément de l'inégalité de Kraft une borne inférieure pour la longueur moyenne de toute fonction de codage. Soit P^n la loi marginale de P sur A^n, définie par $P^n(x_1^n) = P\left(X_1^n = x_1^n\right)$; c'est un élément de l'ensemble $\mathcal{M}_1\left(A^n\right)$ des mesures de probabilité sur A^n. Alors L vérifie l'*inégalité de Shannon* :
$$\mathbb{E}_P\left[L\left(X_1^n\right)\right] \geqslant H\left(X_1^n\right) \overset{\text{déf}}{=} \sum_{x\in A^n} P^n(x)\log\frac{1}{P^n(x)}. \tag{1}$$
La fonction log désigne ici et pour la suite le logarithme à base 2. Cette borne inférieure $H\left(X_1^n\right)$ est appelée *n-entropie.* On dit souvent qu'elle mesure l'incertitude sur X_1^n, ou l'information contenue dans X_1^n : c'est le nombre minimal de bits qu'il faut en moyenne pour transmettre un message de taille n de la source.
3. Il existe des codes qui atteignent presque la limite entropique. Parmi eux, le *codage arithmétique* [Ris76] permet de construire à partir de toute probabilité q_n sur A^n – appelée *probabilité de codage* – une fonction de codage de longueur de code $L_{q_n}\left(X_1^n\right) \leqslant -\log q_n\left(x\right) + c$, où c est une petite constante qui dépend des conditions d'implémentation. Si q_n coïncide avec la loi marginale de X_1^n, L_q a donc une moyenne optimale à cette petite constante près.
Le codage arithmétique présente un double intérêt : sur un plan algorithmique, il peut s'implanter en ligne (c'est-à-dire en traitant les symboles X_i de façon séquentielle) et son temps d'exécution est linéaire en n ; sur un plan théorique, il ramène le problème de construction de codes à des considérations probabilistes et statistiques sur le choix de la probabilité de codage q_n. Aussi, dans la suite, on parlera de la longueur de code associée à la distribution de codage q_n pour désigner $-\log q_n(.)$, négligeant par commodité le petit terme constant.

Il est facile de se convaincre que lorsque la longueur du message n augmente, le nombre moyen de bits nécessaire au codage de chaque caractère ne peut que diminuer. De fait, on vérifie que pour toutes variables aléatoires Y et Z on a $H(Y,Z) \leqslant H(Y) + H(Z)$ et donc, grâce à la stationnarité de P, que la n-entropie $H\left(X_1^n\right)$ est sous-additive :
$$H\left(X_1^{n+m}\right) \leqslant H\left(X_1^n\right) + H\left(X_1^m\right).$$
Par conséquent, le lemme de Fekete implique l'existence d'une limite
$$H(X) = \lim_{n\to\infty}\frac{1}{n}H\left(X_1^n\right) = \inf_{n\in\mathbb{N}_+}\frac{1}{n}H\left(X_1^n\right).$$

Cette limite $H(X)$ est appelée *taux entropique* de la source P : c'est le nombre minimal de bits par caractère que l'on doit utiliser pour coder des messages suffisamment longs. On montre que $H(X) = \lim_{n\to\infty} H\left(X_{n+1}|X_1^n\right)$: c'est donc aussi le nombre minimal de bits qu'il faut au bout d'un temps suffisamment long pour coder le symbole suivant en ayant vu tous les précédents. Enfin, on intreprète également $H(X)$ comme la limite de $\frac{1}{n}\log|T_n|$, où T_n désigne l'ensemble des "messages typiques" de taille n pour la source P, cf [CT91].

0.2.1.2 Universalité faible et forte, Redondances

Bien sûr, la discussion précédente (qui suppose une seule source P connue) ne suffit pas pour répondre dans la pratique à la plupart des besoins. Il apparaît par exemple nécessaire de pouvoir construire des codes efficaces pour plusieurs sources à la fois : on utilise gzip avec tous les fichiers, qu'ils contiennent de l'anglais, du français ou des données quelconques. D'ailleurs, il est très fréquent que l'on ne connaisse pas les statistiques de la source à coder.

Supposons donc que l'on ne connaisse pas P, mais que l'on sache seulement que c'est un élément d'une classe Λ de sources stationnaires. On cherche une probabilité de codage q_n dont la longueur de code $-\log q_n(x)$ approche au mieux la longueur idéale $-\log P^n(x)$, pour autant de messages $x \in A^n$ que possible et pour toutes les sources P à la fois. Notons $\Lambda^n = \{P^n : P \in \Lambda\}$; on appelle *regret* $R(q_n, P, x) = \log\frac{P^n(x)}{q_n(x)}$ la différence entre la longueur de code obtenue avec q_n et celle, appelée *longueur oracle*, que l'on atteindrait si l'on connaissait la source P.

Deux approches sont habituellement considérées pour caractériser l'efficacité de la probabilité de codage q_n avec la source P :

- une analyse en espérance, par la *redondance moyenne* :

$$\bar{R}(q_n, P) = \mathbb{E}_P\left[\log\frac{P^n\left(X_1^n\right)}{q_n\left(X_1^n\right)}\right] \overset{\text{déf}}{=} D(P^n, q_n)$$

  égale à l'information de *Küllback-Leibler* entre P^n et q_n. Il est souvent fructueux d'interpréter celle-ci comme une *pseudo-distance* (positive, vérifiant un certain type d'inégalité triangulaire, mais pas symétrique) : c'est donc un terme que nous reprendrons dans la suite de cette introduction.
- une analyse dans le pire des cas, par la *redondance individuelle*

$$R^*(q_n, P) = \max_{x\in A^n}\log\frac{P^n(x)}{q_n(x)}.$$

On dit que la classe Λ est *faiblement universelle* s'il existe une suite de probabilités de codages $(q_n)_n$ dont la redondance moyenne par symbole tend vers 0 pour toute source de Λ :

$$\sup_{P\in\Lambda}\lim_n \frac{1}{n}D(P^n, q_n) = 0.$$

Si cette limite est atteinte uniformément sur Λ, c'est-à-dire si

$$\lim_n \sup_{P\in\Lambda} \frac{1}{n} D(P^n, q_n) = 0,$$

la classe Λ est dite *fortement universelle.*

Pour les alphabets finis, on sait que la classe des processus stationnaires ergodiques est faiblement universelle. En témoigne la performance d'algorithmes couramment utilisés comme les codes de Lempel-Ziv [ZL77, ZL78, Wel84]. Shields a prouvé [Shi93] qu'elle n'est pas fortement universelle, contrairement à de nombreuses classes à mémoire courte (se référer par exemple au cours donné à Saint-Flour par Catoni [Cat01]).

Pour les alphabets infinis, la situation est complètement différente : même la classe des processus sans mémoire sur un alphabet dénombrable n'est pas faiblement universelle. Cela peut être vu comme une conséquence de la *propriété de Kieffer*, qui caractérise les classes faiblement universelles par une condition nécessaire et suffisante facilement vérifiable. Cette proposition a été prouvée par Kieffer en 1978, la condition nécessaire et suffisante ayant ensuite été simplifiée par Györfi & al. dans l'article [GPvdM94] :

Théorème 1. *[Kie78] Une classe Λ de sources stationnaires sur un alphabet dénombrable A est faiblement universelle si et seulement si il existe une probabilité Q telle que pour toute source P de Λ d'entropie finie, $D(P^1, Q) < \infty$.*

On soulignera que cette condition d'universalité faible se vérifie uniquement sur P^1. Ainsi, les classes faiblement universelles sont celles dont toutes les sources ont une loi marginale à "distance" de Küllback-Leibler finie d'une même loi de probabilité sur A.

Quand le codage universel est possible, on cherche à quantifier l'universalité de la probabilité de codage q_n, c'est-à-dire l'uniformité de son efficacité sur la classe Λ. Nous présentons ici l'approche minimax sous forme d'un jeu entre un émetteur (qui choisit la source afin d'obtenir le code le plus long possible) et un messager (qui doit coder le message en utilisant un minimum de bits). Le messager choisit d'abord sa fonction de codage q_n, puis l'émetteur choisit la source la plus mal approchée. On définit ainsi la *redondance moyenne minimax* :

$$\bar{R}\,(\Lambda^n) = \inf_{q_n\in\mathcal{M}_1(A^n)} \sup_{P\in\Lambda} \bar{R}(q_n, P^n)$$

et la *redondance individuelle minimax* :

$$R^*\,(\Lambda^n) = \inf_{q_n\in\mathcal{M}_1(A^n)} \sup_{P\in\Lambda} R^*(q_n, P^n).$$

Il apparaît immédiatement que $\bar{R}\,(\Lambda^n) \leqslant R^*\,(\Lambda^n)$. Pour la plupart des classes considérées jusque là, redondances moyenne et individuelle sont du même ordre de grandeur. Il peut en être différemment : nous construisons avec la proposition 13 un exemple de classe pour laquelle la redondance minimax moyenne est finie, mais pas la redondance minimax individuelle.

0.2.1.3 Redondance moyenne minimax et approche maximin

Ces deux quantités s'étudient différemment. Pour minorer la redondance moyenne minimax, on introduit souvent un autre type de jeu entre l'émetteur et le récepteur. L'émetteur choisit d'abord une loi de probabilité π (appelée *prior* : l'approche est d'inspiration bayésienne) sur l'ensemble de sources Λ, puis tire ses messages de la source $\mathcal{P}$ qu'il tire au hasard selon la loi π. Le messager, qui connaît π, cherche à minimiser la longueur de code moyenne. On définit ainsi la *redondance (moyenne) maximin* :

$$R^{-}(\Lambda^n) = \sup_{\pi \in \mathcal{M}_1(\Theta)} \inf_{q_n \in \mathcal{M}_1(A^n)} \mathbb{E}_\pi \left[\bar{R}(q_n, \mathcal{P})\right].$$

Il apparaît immédiatement que $R^{-}(\Lambda^n) \leqslant \bar{R}(\Lambda^n)$: pour tout prior π et pour toute loi de codage q_n on a en effet

$$\max_{P \in \Lambda} \bar{R}_n(q_n, P) \;\geqslant\; \mathbb{E}_\pi \left[\bar{R}(q_n, \mathcal{P})\right].$$

On montre en fait que redondances moyenne minimax et maximin coïncident. Ceci peut être vu comme une application du théorème minimax de Sion [Sio58]. Cette égalité est attribuée dans [DLG80] à Gallager (elle se trouverait dans ses notes de cours du MIT). Nous donnons ici l'énoncé de Haussler, dont la preuve se trouve dans son article de 1997 :

Théorème 2. *[Hau97] Si la classe Λ est tendue, alors $R^{-}(\Lambda^n) = \bar{R}(\Lambda^n)$ et la borne minimax est atteinte par un mélange $q_n(.) = \int P^n(.)\mathrm{d}\pi(P)$.*
Sinon, $R^{-}(\Lambda^n)$ est infinie (et donc $\bar{R}(\Lambda^n)$ l'est aussi).

L'introduction de la redondance maximin a pour intérêt principal le fait qu'il y a un moyen simple de la minorer : il suffit d'exhiber un prior et d'étudier la meilleure redondance moyenne atteignable, qui s'interprète comme la capacité d'un canal [CT91]. Comme ce raisonnement sera utilisé plusieurs fois dans la suite, nous le détaillons un peu dans un cadre assez restreint, mais suffisant pour les applications que nous en ferons. Rappellons d'abord [CT91] que si X et Y forment un couple de variables aléatoires de loi $P_{(X,Y)}$ dont les marginales sont P_X et P_Y, à valeur dans un ensemble discret $\mathcal{X}$,

- l'*entropie conditionnelle de X sachant Y* est définie comme l'espérance de l'entropie de X conditionnellement à Y :

$$H(X|Y) = \sum_{y \in \mathcal{X}} P_Y(y) \sum_{x \in \mathcal{X}} P(X = x|Y = y) \log \frac{1}{P(X = x|Y = y)}.$$

- l'*information mutuelle* $I(X,Y) = D\left(P_{(X,Y)}, P_X \otimes P_Y\right)$ est l'information de Küllback-Leibler entre la loi du couple (X,Y) et le produit tensoriel des lois marginales et vérifie :

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

- l'entropie condionnelle de X sachant Y est contrôlée par la probabilité pour X et Y d'être différents – c'est le *lemme de Fano* [Fan61] :

$$H(X|Y) \leqslant P_{X,Y}(X \neq Y)\log|\mathcal{X}| + 1.$$

Supposons que W soit une variable aléatoire de loi uniforme sur un ensemble $\mathcal{W}$, et supposons qu'une source sans mémoire $P_\theta \in \Lambda$ soit associée à chaque paramètre $\theta \in \mathcal{W}$. On note alors X le processus sur A, de loi $\mathbb{P}$, tel que conditionnellement à $W = \theta$ la loi de X est P_θ. On a donc pour $x_1^n \in A^n$

$$\mathbb{P}(X_1^n = x_1^n) = \frac{1}{|\mathcal{W}|}\sum_{\theta\in\mathcal{W}} P_\theta^n(x_1^n).$$

En appliquant l'inégalité de Shannon (1) à X_1^n, il apparaît que la longueur moyenne de code pour n'importe quelle distribution de codage q_n est supérieure à la n-entropie de notre source composée $\mathbb{P}$, soit :

$$\mathbb{E}\left[\log\frac{1}{q_n(X_1^n)}\right] = \frac{1}{|\mathcal{W}|}\sum_{\theta\in\mathcal{W}}\mathbb{E}_{P_\theta}\left[\log\frac{1}{q\left(X_1^n\right)}\right] \geqslant H\left(X_1^n\right).$$

Cela implique que la redondance maximin de Λ^n est plus grande que l'information mutuelle entre X_1^n et W :

$$\begin{aligned}
R^-\left(\Lambda^n\right) &\geqslant \frac{1}{|\mathcal{W}|}\sum_{\theta\in\mathcal{W}}\bar{R}_n(q_n, P_\theta)\\
&= \frac{1}{|\mathcal{W}|}\sum_{\theta\in\mathcal{W}}\mathbb{E}_{P_\theta}\left[\log\frac{P_\theta\left(X_1^n\right)}{q_n\left(X_1^n\right)}\right]\\
&= \frac{1}{|\mathcal{W}|}\sum_{\theta\in\mathcal{W}}\mathbb{E}_{P_\theta}\left[\log\frac{1}{q_n\left(X_1^n\right)}\right] - \frac{1}{|\mathcal{W}|}\sum_{\theta\in\mathcal{W}}\mathbb{E}_{P_\theta}\log\frac{1}{P_\theta\left(X_1^n\right)}\\
&\geqslant H\left(X_1^n\right) - H\left(X_1^n|W\right)\\
&= I\left(X_1^n, W\right).
\end{aligned}$$

On se ramène ainsi à la minoration de l'information mutuelle $I\left(W, X_1^n\right) = H(W) - H\left(W|X_1^n\right)$ entre le paramètre aléatoire W et le message X_1^n (jusqu'ici, on n'a d'ailleurs pas vraiment besoin de supposer la loi de W uniforme). Cette minoration peut être effectuée en contrôlant l'incertitude $H\left(W|X_1^n\right)$ que l'on a sur le paramètre quand on a vu le message. En particulier, si l'on est capable de construire un estimateur consistant $\hat{\theta} = \hat{\theta}\left(X_1^n\right)$ du paramètre θ à partir du message X_1^n, le lemme de Fano permet de voir que $H\left(W|X_1^n\right)$ est négligeable devant $H(W)$, et donc que $I(W, X)$ est de l'ordre de $H(W) = \log|\mathcal{W}|$.

La redondance maximin $R^-\left(\Lambda^n\right)$ apparaît ainsi intuitivement comme le logarithme du nombre maximal de sources de Λ que l'on est capable de *distinguer* grâce à un message X_1^n. Le schéma de preuve décrit ci-dessus sera utilisé à trois reprises dans la thèse : d'une part pour les théorèmes 16 et 17 concernant deux classes enveloppes en codage des entiers (avec une petite simplification dans le premier cas qui court-circuite l'utilisation du lemme de Fano), d'autre part pour la démonstration du théorème 20 au sujet de la redondance de motifs.

Citons d'ores et déjà une application très importante de ce principe, d'abord présentée par Gallager [Gal76], Davisson & Leon-Garcia [DLG80] et Ryabko [Rya79, Rya81], puis améliorée par Rissanen [Ris84] et par Merhav et Feder [MF95]. Elle permet de minorer avec une grande généralité la redondance maximin pour les classes paramétriques :

Théorème 3. *[MF95] Supposons que $\Lambda = \left\{P_\theta : \theta \in \Theta = [0,1]^k\right\}$. Supposons que pour tout message x_1^n, $\theta \longmapsto P_\theta\left(x_1^n\right)$ soit mesurable, et qu'il existe un estimateur $\hat{\theta} = \hat{\theta}\left(X_1^n\right)$ et une fonction α tendant vers 0 à l'infini tels que pour tout $\theta \in \Theta$ et pour n assez grand on ait*

$$P_\theta\left(\left\|\hat{\theta} - \theta\right\| \geqslant \frac{c}{\sqrt{n}}\right) \leqslant \alpha(c).$$

Alors il existe un sous-ensemble $\mathcal{N}$ de Θ de mesure de Lebesgue nulle tel que pour toute distribution de codage q_n et pour tout $\theta \notin \mathcal{N}$,

$$\limsup_{n\to\infty} \frac{1}{\log n} \bar{R}\left(q_n, P_\theta^n\right) \geqslant \frac{k}{2}.$$

En particulier, on voit qu'avec les hypothèses de ce théorème la redondance maximin $R^-\left(\Lambda^n\right)$ est au moins de l'ordre de $\frac{k}{2}\log n$ quand n tend vers l'infini. Ce résultat couvre de très nombreux cas utilisés en pratique. On retient notamment que pour un alphabet de taille m, la redondance dans la classe des processus sans mémoire est au moins $\frac{m-1}{2}\log n$. Il s'applique aussi aux cas des chaînes de Markov et de leurs généralisations (voir les VLMC plus loin). En fait, l'énoncé "la redondance est $\frac{k}{2}\log n$ où k est le nombre de degrés de liberté" est presque devenu proverbial bien au delà du cadre strict de la théorie de l'information, notamment par ses applications au principe MDL (voir plus bas). Cela est justifié par le fait qu'il est souvent optimal : il arrive que l'on sache également montrer des bornes supérieures pour les redondances minimax moyennes et individuelles qui sont du même ordre de grandeur. Par exemple, le cas des classes $\mathcal{I}_m$ des processus où les symboles sont indépendants, identiquement distribués (iid) sur l'alphabet $\{1,\ldots,m\}$ a été traité dans [CB94, XB97]. En posant $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$, Clarke & Barron puis Xie & Barron prouvent que

$$\bar{R}\left(\mathcal{I}_m^n\right) - \frac{m-1}{2}\log\frac{n}{2\pi e} \to \log\frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)},$$

la convergence ayant lieu à une vitesse dépendant de la taille m de l'alphabet. Les auteurs exhibent même des probabilités de codage à la fois asymptotiquement maximin et minimax.

Pour les classes de processus plus compliquées, il n'est pas facile de montrer des résultats aussi précis. Le moyen le plus simple d'obtenir des bornes supérieures est souvent de passer par la redondance individuelle.

0.2.1.4 Redondance individuelle minimax et NML

La situation de la redondance individuelle est en effet grandement simplifiée par le fait qu'on sait décrire le codeur qui atteint la borne minimax (quand il existe) : c'est la distribution *Normalized Maximum Likelihood* (NML) de Shtarkov [Sht87] définie pour $x_1^n \in A^n$ par

$$q_n^{NML}\left(x_1^n\right) = \frac{\hat{P}\left(x_1^n\right)}{\sum_{x_1^n\in A^n}\hat{P}\left(x_1^n\right)},$$

où $\hat{P}\left(x_1^n\right) = \sup_{P\in\Lambda} P^n\left(x_1^n\right)$. En effet, pour toute probabilité de codage q_n il existe un message $x_1^n \in A^n$ tel que $q_n(x_1^n) \leqslant q_n^{NML}(x_1^n)$. Donc

$$\begin{aligned}
R^*\left(q_n, \Lambda^n\right) &= \sup_{P\in\Lambda} R^*(q_n, P)\\
&\geqslant -\log q_n(x_1^n) - \inf_{P\in\Lambda} -\log P^n(x_1^n)\\
&= \log\frac{\hat{P}\left(x_1^n\right)}{q_n(x_1^n)}\\
&\geqslant \log\frac{\hat{P}\left(x_1^n\right)}{q_n^{NML}(x_1^n)} = \log\sum_{x_1^n\in A^n}\hat{P}\left(x_1^n\right),
\end{aligned}$$

et on voit que

$$R^*(\Lambda^n) = R^*(q_n^{NML},\Lambda) = \log\sum_{x_1^n\in A^n}\hat{P}\left(x_1^n\right).$$

Ainsi, l'étude de la redondance individuelle se déduit du comportement asymptotique du logarithme d'une somme. Quand on peut les mettre en oeuvre, les arguments combinatoires sont pour cela les plus précis (voir dans [Szp01] et les références qu'il contient); ils permettent par exemple d'obtenir le résultat suivant :

$$R^*_n\left(\mathcal{I}^n_m\right) \leqslant \frac{m-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)} + o_m(1)$$

où $o_m(1)$ tend vers 0 (quand n tend vers l'infini) à une vitesse dépendant de m. Cette borne supérieure rejoint donc asymptotiquement la borne inférieure de Rissanen pour $R^-\left(\mathcal{I}^n_m\right)$.

0.2.1.5 Alphabets infinis dénombrables

Cette borne supérieure tend vers l'infini quand la taille de l'alphabet m augmente. Et de fait, comme nous l'avons déjà dit, on déduit aisément du critère de Kieffer que la classe $\mathcal{I}_\infty$ des processus iid sur un alphabet infini n'est pas même faiblement universelle. Il convient donc dans ce cas de se restreindre à l'étude de classes plus petites; c'est le travail qui est entrepris dans le chapitre 1, issu d'un travail mené en collaboration avec Stéphane Boucheron et Elisabeth Gassiat.

Après quelques remarques et propriétés d'ordre général, un intérêt particulier est porté aux classes de processus sans mémoire sur l'ensemble des entiers strictement positifs $\mathbb{N}_+$. Il est rappelé que $R^*(\Lambda^n)$ est une fonction de n (croissante) sous-additive, et donc finie si et seulement si $\sum_{x\in\mathbb{N}_+}\hat{P}(x) < \infty$. Pour tout nombre positif u, posons $\bar{F}_{\Lambda^1}(u) = \sum_{x>u}\hat{P}(x)$. Nous prouvons la proposition suivante :

Proposition 2.

$$R^*(\Lambda^n) \leqslant \inf_{u:u\leqslant n}\left[n\bar{F}_{\Lambda^1}(u)\log e + \frac{u-1}{2}\log\frac{en}{u} + O(1)\right].$$

Il en découle que si $R^*(\Lambda^n)$ est fini il est aussi asymptotiquement négligeable devant n.

Ce chapitre contient une discussion plus spécifique des classes dites *enveloppes* définies, à partir d'une fonction $f : \mathbb{N}_+ \longrightarrow [0,1]$, comme la classe Λ_f de tous les processus sans mémoire P tels que pour tout entier positif x, $P(x) \leqslant f(x)$. Il est montré que, pour ces classes enveloppes, redondance moyenne minimax et redondance individuelle minimax sont toutes les deux finies si et seulement si $\sum_{k\in\mathbb{N}_+}\hat{P}(k) < \infty$. Deux exemples sont étudiés en détail :

- la *classe algébrique* $\Lambda_{M\cdot^{-\alpha}}$ associée à la fonction lentement décroissante $f_{\alpha,M} : x \mapsto \frac{M}{x^\alpha}$ pour $M > 1$ et $\alpha > 1$. Nous obtenons les résultats suivants pour ses différents types de redondance :
  Théorème 4. *Supposons que* $M\sum_{k\geqslant 1}\frac{1}{k^\alpha} \geqslant 2^\alpha$. *Alors*

$$n^{1/\alpha}A(\alpha)\log\lfloor M\zeta(\alpha)\rfloor \leqslant R^-\left(\Lambda^n_{M\cdot^{-\alpha}}\right)$$

et

$$R^*\left(\Lambda^n_{M\cdot^{-\alpha}}\right) \leqslant \left(\frac{2Mn}{\alpha-1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1),$$

où

$$A(\alpha) = \frac{1}{\alpha}\int_0^\infty \frac{1}{u^{1-1/\alpha}}\left(1 - e^{-1/(\zeta(\alpha)u)}\right)du\,.$$

Ce théorème laisse un petit trou d'ordre $(\log n)^{1-\frac{1}{\alpha}}$ entre la borne inférieure et la borne supérieure. Il est difficile de savoir si ce trou peut être comblé, et quelle borne il faudrait améliorer. Notons toutefois qu'en passant à la limite simultanément en $\alpha \to \infty$ et $M = H^\alpha$, la classe $\Lambda_{M\cdot^{-\alpha}}$ converge vers la classe des processus sans mémoire sur l'alphabet $\{1,\ldots,H\}$ pour laquelle redondances individuelle minimax et redondance maximin sont toutes deux d'ordre $\frac{H-1}{2}\log n$. C'est (à un facteur 2 près) ce que l'on obtient en prenant la limite dans la borne supérieure du théorème.

- la *classe exponentielle* $\Lambda_{Ce^{-\alpha\cdot}}$ définie par la fonction à décroissance plus rapide $x \mapsto Ce^{-\alpha x}$. On montre :

  Théorème 5. *Supposons que $C > e^{2\alpha}$. Alors*

  $$\frac{1}{8\alpha}\log^2 n\,(1-o(1)) \leqslant R^-(\Lambda^n_{Ce^{-\alpha\cdot}}) \leqslant R^*(\Lambda^n_{Ce^{-\alpha\cdot}}) \leqslant \frac{1}{2\alpha}\log^2 n + O(1)$$

  Il est ici remarquable que borne inférieure et borne supérieure ne diffèrent asymptotiquement que par un facteur 4.

Dans les deux cas, la preuve de la borne inférieure utilise l'approche maximin esquissée ci-dessus, alors que celle de la borne supérieure est une conséquence de la proposition 2 sur la redondance du codeur de Shtarkov. Dans un travail en cours de rédaction [BGG06], nous travaillons à une minoration générale pour les classes enveloppe qui devrait permettre de retrouver les cas particuliers évoqués ici.

Sur un plan plus pratique, le chapitre 1 contient également la description d'un algorithme séquentiel linéaire en temps destiné à coder sur certaines classes de processus sans mémoire sur $\mathbb{N}_+$. Son principe de fonctionnement est suggéré par la preuve de la proposition 2 : il faut traiter séparément les petits symboles, plus probables, et les symboles plus grands qui apparaissent moins souvent. Pour les premiers, un codage de type Krichevsky-Trofimov ([KT81], voir la section 0.2.2.1 ci-dessous) est utilisé, alors qu'on fait appel sur les seconds au code d'Elias [Eli75]. Nous montrons que cet algorithme a une redondance au plus égale à $\left(\frac{2Mn}{\alpha-1}\right)^{\frac{1}{\alpha}}\log n\,(1+o(1))$ sur la classe $\Lambda_{M\cdot^{-\alpha}}$. Une version adaptative en α est donnée pour les processus P tels que le nombre de symboles différents paraissant dans un message de taille n ait une espérance d'ordre $n^{\frac{1}{\alpha}}$ pour un certain réel positif α inconnu.

0.2.1.6 Codage par motifs

Une autre approche, radicalement différente, est possible pour le codage de sources sans mémoire sur les alphabets infinis. Elle consiste à renoncer à une partie de l'information du message et à n'en transmettre que la *structure*, en oubliant le contenant – ou autrement dit, à ne coder que les répétitions en omettant la valeur des symboles.

Cette approche a d'abord été proposée par Äberg & al. dans l'article [ÄSMS97] comme solution au problème du codage *multi-alphabet*, où l'on cherche à transmettre un message x_1^n en sachant a priori qu'il ne contiendra qu'un petit sous-ensemble de caractères de l'alphabet A. Son étude a ensuite été poursuivie par Shamir [Sha03a, Sha04, Sha03b, Sha06] et par Jevtić, Orlitsky, Santhanam and Zhang [OS04, JOS05, OSZ04, OSVZ04], qui évoquent différents problèmes pratiques. Il arrive en effet que l'alphabet doive de toutes façons être transmis à part (par exemple, pour la transmission d'un texte dans une langue inconnue), ou bien que l'alphabet soit de taille très grande par rapport au message (citons le cas des images où chaque pixel est codé sur $m = 2^{24}$ bits, ou les textes littéraires si l'on choisit les mots pour symboles). Dans ce dernier cas, l'alphabet est certes fini mais les bornes en $\frac{k}{2}\log n$ présentées ci-dessus sont soit inutiles, soit extrêmement pessimistes si k n'est pas négligeable devant n.

Pour présenter la notion de motif, considérons l'exemple de la chaîne

$$x_1^{11} = \text{abracadabra}.$$

L'information qu'elle contient peut être séparée en deux morceaux :

- un *dictionnaire* $\Delta = \Delta(x)$, défini comme la suite des caractères présents dans x dans leur ordre d'apparition ; dans notre exemple, $\Delta = (a, b, r, c, d)$.
- un *motif* $\psi = \psi(x)$, tel que ψ_i est l'indice du symbole X_i dans l'alphabet Δ ; ici, $\psi = 12314151231$.

Si l'on se contente de coder les motifs (en laissant de côté le dictionnaire), le codage universel pour la classe $\mathcal{I}_\infty$ des processus sans mémoire sur $\mathbb{N}$ est possible. Supposons en effet que la source sans mémoire $P \in \mathcal{I}_\infty$ génère le message $X = (X_n)_n$ tel que pour tout entier positif k, $P(X_1 = k) = \theta_k$. Soit $\mathcal{P}^n$ l'ensemble de tous les motifs des messages à n symboles. On voit par exemple que $\mathcal{P}^1 = \{1\}$, $\mathcal{P}^2 = \{11, 12\}$, $\mathcal{P}^3 = \{111, 112, 121, 122, 123\}$. Le processus X induit un processus de motifs $\Psi = (\Psi_n)_{n\in\mathbb{N}}$ tel que $\Psi_1^n = \psi(X_1^n)$. Le processus Ψ n'est bien sûr pas sans mémoire. Notons P_Ψ sa loi, image de P par l'application ψ : $P_\Psi(\Psi_1^n = \psi) = \sum_{\psi(x)=\psi} P(X_1^n = x)$. Comme les motifs comportent moins d'information que les chaînes initiales, il paraît plus simple de les coder.

On peut définir la *n-entropie de motif* $H(\Psi_1^n) = \mathbb{E}_{P_\Psi}[-\log P_\Psi(\Psi_1^n)]$. Pour un processus stationnaire ergodique P, Orlitsky & al. [OSVZ04] montrent que le *taux d'entropie de motif* $H(\Psi) = \lim_{n\to\infty} \frac{1}{n} H(\Psi_1^n)$ existe et est égal à l'entropie de la source $H(X)$ (que celle-ci soit finie ou non). Ce résultat a été trouvé indépendamment par Gemelos and Weissman [GW04].

La *redondance de motif moyenne* de la distribution de codage q_n sur $\mathcal{P}^n$ peut être définie par analogie comme la différence entre la longueur de code moyenne sous la distribution P_Ψ et la n-entropie de motif :

$$\begin{aligned}\bar{R}_\Psi(q_n, P) &= \mathbb{E}_{P_\Psi}\left[-\log q_n(\Psi_1^n)\right] - H(\Psi_1^n) \\ &= \sum_{\psi\in\Psi^n} P_\Psi(\psi)\log\frac{P_\Psi(\psi)}{q_n(\psi)}.\end{aligned}$$

La *redondance de motif moyenne* de q_n est définie comme le maximum de $\bar{R}_\Psi(q_n, P)$ sur toutes les sources sans mémoire $P \in \mathcal{I}_\infty$, et la *redondance de motif moyenne minimax* $\bar{R}_\Psi(\mathcal{I}_\infty^n)$ est la borne inférieure en q_n des $\bar{R}_\Psi(q_n, P)$.

De façon analogue, Orlitsky & al. définissent le regret de la distribution de codage q_n sur le motif $\psi \in \mathcal{P}^n$ comme la différence entre la longueur de code $-\log q_n(\psi)$ et l'oracle $\inf_{\theta\in\Theta} - \log P_\Psi(\psi)$. La redondance individuelle est le regret maximal

$$R^*_\Psi(q_n, \mathcal{I}^n_\infty) = \max_{\psi\in\mathcal{P}^n}\left[-\log q_n(\psi) - \inf_{\theta\in\Theta} - \log P_\Psi(\psi)\right],$$

et la redondance minimax individuelle est sa borne inférieure en q_n. Elle est (comme d'habitude) atteinte par la probabilité de codage NML de Shtarkov

$$q_{NML}(\psi) = \frac{\max_{\theta\in\Theta} P_\Psi(\psi)}{\sum_{\psi\in\mathcal{P}^n}\max_{\theta\in\Theta} P_\Psi(\psi)},$$

qui atteint le regret constant sur chaque motif

$$R^*_\Psi(\mathcal{I}^n_\infty) = \log\sum_{\psi\in\mathcal{P}^n}\max_{\theta\in\Theta} P_\Psi(\psi).$$

Dans l'article [OSZ04], Orlitsky, Santhanam et Zhang montrent que la redondance de motif individuelle minimax est finie, et même plus précisément que

$$R^*_\Psi(\mathcal{I}^n_\infty) \leqslant \left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n}.$$

Les auteurs ne prouvent toutefois pas que ce résultat est optimal, ils démontrent seulement que $R^*_\Psi(\mathcal{I}^n_\infty) \geqslant \left(\frac{3}{2}\log e\right) n^{\frac{1}{3}}\left(1 + o(1)\right)$. De son côté, Shamir montre dans [Sha03b] que la redondance de motif maximin $R^-_\Psi(\mathcal{I}^n_\infty)$ est d'ordre de grandeur supérieur à $n^{1/3-\epsilon}$, borne que l'on peut facilement renforcer à $R^-_\Psi(\mathcal{I}^n_\infty) \geqslant 5 \times 10^{-8}\left(\frac{n}{\log^2 n}\right)^{1/3}$ pour n assez grand, en faisant varier ϵ avec n.

Il existe donc un trou entre bornes inférieures et bornes supérieures connues, qui a été réduit par Shamir dans [Sha04] grâce à la nouvelle borne $R^*_\Psi(\mathcal{I}^n_\infty) = O\left(n^{\frac{2}{5}+\epsilon}\right)$ pour tout $\epsilon > 0$.

La contribution du chapitre 2 est de proposer une nouvelle borne inférieure pour $R^-_\Psi(\mathcal{I}^n_\infty)$ par l'approche maximin. Nous prouvons le théorème suivant :

Théorème 6. *Pour tout entier n assez grand,*

$$R^-_\Psi(\mathcal{I}^n_\infty) \geqslant 1.84 \left(\frac{n}{\log n}\right)^{1/3}.$$

Outre une amélioration de l'ordre de grandeur (et de la constante) dans la borne inférieure de la redondance maximin, ce résultat est intéressant en ce que la technique qu'il met en oeuvre est assez différente de celle que l'on trouve dans [Sha03b]. En effet, la preuve de ce résultat utilise des résultats avancés de combinatoire sur les partitions des nombres entiers. Pour comprendre cela, il convient d'introduire quelques définitions.

- Le nombre d'occurences du symbole j dans ψ est appelé sa multiplicité $\mu_j(\psi)$, et $\mu(\psi) = (\mu_j(\psi))\, 1 \leqslant j \leqslant n$ est appelée la *multiplicité du motif* ψ : pour nous en tenir à notre exemple, $\mu = (5, 2, 2, 1, 1, 0, \ldots)$. On a bien sûr $\sum_{j=1}^n \mu_j = n$.
- Le *profil* $\phi = (\phi_\mu)_{\mu \geqslant 1}$ du motif ψ donne pour chaque valeur μ sa fréquence dans $\mu(\psi)$ (formellement, ϕ est donc "la multiplicité de la multiplicité de ψ"). Le profil du message $x = abracadabra$ est donc $(2, 2, 0, 0, 1, 0, \ldots)$, puisque deux symboles (c et d) y apparaissent une fois, deux symboles (b et r) y apparaissent deux fois et un symbole (a) apparaît cinq fois. Il est facile de voir que la probabilité sous P_Ψ d'un motif *ne dépend que de son profil.*
- Soit Φ^n l'ensemble de tous les profils possibles pour des motifs de taille n – par exemple $\Phi^3 = \{(3, 0, 0), (1, 1, 0), (0, 0, 1)\}$. Comme cela est expliqué dans l'article [OSZ04], on se convainc aisément de la relation : $\sum_{\mu=1}^n \mu\phi_\mu = n$, et il apparît que Φ^n est en bijection avec l'ensemble

$$\Theta^n = \left\{\theta = (\theta_j)_{j\in\mathbb{N}^+} : \theta_1 \geqslant \theta_2 \geqslant \ldots \text{ et } \sum_{j=1}^{\infty} \theta_j = n\right\}$$

des partitions (non ordonnées) de l'entier n.

C'est cette bijection qui permet aux auteurs d'obtenir leurs bornes pour $R^*_\Psi(n, \theta)$, via l'équivalent en $\frac{1}{4n\sqrt{3}} e^{\pi\sqrt{\frac{2n}{3}}}$ du nombre de partitions de l'entier n prouvé par Hardy de Ramanujan. Dans la preuve du théorème 6, c'est un raffinement de cette idée qui est utilisé. Entrons un peu dans les détails.

- Nous utilisons l'approche maximin en cherchant à minorer l'information mutuelle $I(W, \Psi)$ entre un paramètre aléatoire W, et le motif aléatoire Ψ dont la distribution conditionnellement à $W = \theta$ est P_θ.

- Nous choisissons de prendre W uniformément distribué sur l'ensemble $\mathcal{W} \subset \Theta^c$ des partitions de l'entier $c = c(n)$ formées de sommants au maximum égaux à $d = d(n)$; pour obtenir les résultats optimaux, nous prenons $c = \left(\frac{n}{\frac{16}{3}\lambda \log n}\right)^{2/3}$ et $d = 0.8\sqrt{c}$.
- Pour $\theta \in \mathcal{W}$, on définit la source sans mémoire P_θ par la relation :

$$P_\theta(X_1 = j) = \frac{\theta_j}{c}.$$

- Dans l'article [DN90], qui fait référence à [Sze51], Dixmier et Nicolas montrent que le logarithme du cardinal de $\mathcal{W}$ est alors équivalent à $2.07236\sqrt{c}$: la restriction aux petits sommants ne diminue pas trop (à l'échelle logarithmique) le nombre de partitions. Elle permet en revanche d'assurer que tous les symboles ont une "petite" probabilité :

$$\forall \theta \in \mathcal{W}, \forall k \in \mathbb{N}_+, P_\theta(k) \leqslant \frac{d}{c} \leqslant 0.8 \left(\frac{\frac{16}{3}\lambda \log n}{n}\right)^{1/3}.$$

- Cela permet de montrer, via l'inégalité de Bernstein, que l'on peut estimer correctement θ à partir de Ψ_1^n avec une probabilité qui tend vers 1 quand n tend vers l'infini – et, comme annoncé, on conclut grâce à l'inégalité de Fano.

0.2.2 Codeurs universels efficaces

Nous revenons maintenant aux alphabets finis, et présentons une méthode de codage universel efficace en théorie comme en pratique pour les processus stationnaires ergodiques. Pour cela, nous commençons par le cas bien connu des processus sans mémoire. Nous verrons qu'il peut servir de "brique élémentaire" pour le codage de sources plus complexes.

0.2.2.1 Processus iid : le mélange de Krichevsky-Trofimov

Nous avons vu qu'il est possible de coder dans les classes $\mathcal{I}_m$ avec une redondance de l'ordre de $O\left(\frac{m-1}{2}\log n\right)$. Exposons maintenant un moyen particulièrement élégant et efficace de le faire en pratique. Soit

$$\Theta_m = \left\{\theta = (\theta_1, \ldots, \theta_m) : \sum_{i=1}^m \theta_i = 1\right\}$$

le simplexe sur l'alphabet $A = \{1, 2, \ldots, m\}$. Pour $\theta \in \Theta_m$ et $a \in A$, on définit la probabilité p_θ sur A par $p_\theta(a) = \theta_a$, et pour $x_1^n \in A^n$, notons $p_\theta(x_1^n) = \prod_{i=1}^n p_\theta(x_i)$.

Par ailleurs, notons $N_a(x_1^n) = \sum_{i=1}^n \mathbb{1}_{x_i=a}$ le nombre d'occurences du symbole $a \in A$ dans le message x.

Il existe de nombreux codeurs qui approchent la redondance optimale. Le théorème 2 suggère l'utilisation de probabilités de codages qui *mélangent* les différentes lois p_θ ; dans cette famille, le *mélange de Krichevsky-Trofimov* $\mathcal{KT}$ est particulièrement remarquable. Il utilise pour ce mélange la loi ν de Dirichlet de paramètre 1/2 : pour $x_1^n \in A^n$,

$$\mathcal{KT}(x_1^n) = \int_{\theta \in \Theta_m} p_\theta(x_1^n)\ \nu(\mathrm{d}\theta). \tag{2}$$

Il se trouve (voir par exemple [XB97]) que cette loi de Dirichlet est ici aussi le prior de Jeffrey (proportionnel à la racine carrée du déterminant de la matrice d'information de Fischer) qui est asymptotiquement maximin. D'un point de vue algorithmique, l'utilisation d'une loi de Dirichlet – bien connue dans la littérature pour être conjuguée aux lois binomiales – permet une évaluation simple $\mathcal{KT}(x_1^n)$:

$$\mathcal{KT}(x_1^n) = \frac{\prod_{a:N_a(x_1^n)\geqslant 1}\left(N_a(x_1^n) - \frac{1}{2}\right)\left(N_a(x_1^n) - \frac{3}{2}\right)\dots\frac{1}{2}}{\left(n-1+\frac{m}{2}\right)\left(n-2+\frac{m}{2}\right)\dots\frac{m}{2}}, \tag{3}$$

et même des probabilités conditionnelles

$$\mathcal{KT}(a|x_1^n) = \frac{N_a\left(x_1^n\right) + \frac{1}{2}}{n+\frac{m}{2}} \tag{4}$$

ce qui, combiné au codage arithmétique, permet la construction d'algorithmes séquentiels.

Sur un plan théorique, Krichevsky et Trofimov ont montré directement dès le début des années 1980 l'inégalité suivante :

Proposition 3. *[KT81, DMPW81, Cat01] Pour tout entier positif n et pour tout message $x_1^n \in A^n$:*

$$-\log \mathcal{KT}(x_1^n) \leqslant \inf_{\theta\in\Theta_m} -\log p_\theta(x_1^n) + \frac{m-1}{2}\log n + \log m.$$

Il faut souligner le fait remarquable que cette majoration est trajectorielle et uniforme. On sait depuis [XB97] que le mélange de Krichevsky-Trofimov n'est par contre pas asymptotiquement minimax, ce qui constitue bien sûr plus une curiosité qu'un véritable handicap en pratique...

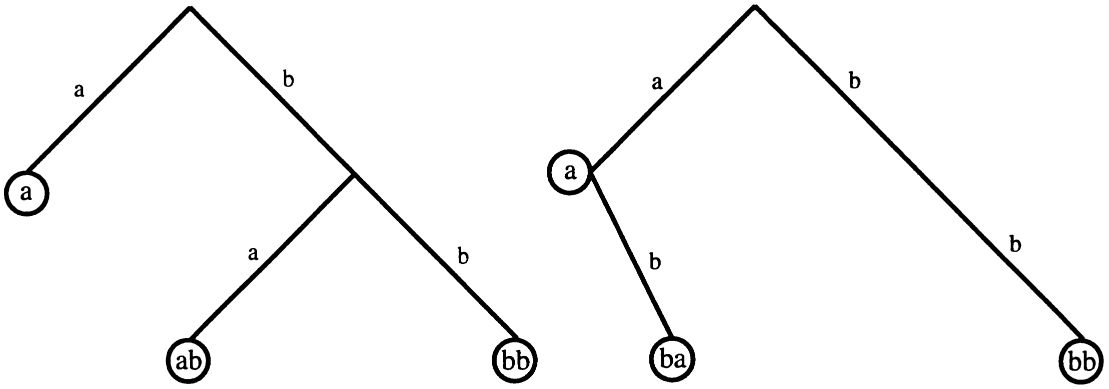


FIG. 1 – $T_1 = \{a, ab, bb\}$ est un arbre de contexte, contrairement à $T_2 = \{a, ba, bb\}$.

0.2.2.2 Modèles à mémoire

Nous allons voir que les résultats précédents se généralisent aux cas de modèles paramétriques avec mémoire finie : on est capable de construire des codeurs efficaces algorithmiquement dont la redondance est presque optimale. Mais notons d'ores et déjà que cela ne fournit pas une réponse immédiate à tous les problèmes concrets de codage : prenons en effet l'exemple d'une chaîne de Markov d'ordre r sur l'alphabet $A = \{1, \ldots, m\}$ sous sa loi stationnaire. Elle est caractérisée par $m^r (m-1)$ paramètres : les redondances moyenne et individuelle sont donc équivalentes à $\frac{m^r(m-1)}{2} \log n$ pour cette classe. On voit vite apparaître un problème : la redondance minimax augmente très vite avec l'ordre markovien – et il ne suffit pas de savoir l'approcher pour obtenir de bons résultats sur les messages de taille finie. Si l'on veut utiliser des modèles à mémoire pour modéliser les langages naturels, ou les séquences biologiques (ADN, etc.), on est donc contraint de choisir une petite mémoire (un petit ordre markovien) pour pouvoir coder efficacement – mais alors, la modélisation est de très mauvaise qualité. Nous allons voir qu'il est possible de faire beaucoup mieux en introduisant des modèles *plus souples* pouvant avoir de la mémoire sans être de trop grande dimension.

Les *chaînes de Markov d'ordre variables* (*VLMC*), ou *sources à arbres de contexte*, répondent à ce besoin en autorisant la taille de la mémoire à dépendre du passé. Pour préciser cette idée, nous devons d'abord introduire la notion d'*arbre de contexte*. Un ensemble T de mots sur l'alphabet A peut être représenté par un arbre dont les arêtes sont indexées par des lettres de A, en faisant correspondre à chaque mot un chemin d'un noeud à la racine (voir la figure 1). Si aucun mot $s_1 \in T$ n'est le suffixe d'un autre mot $s_2 \in T$, alors chaque mot de T correspond à une feuille de l'arbre et on dit que T vérifie la *propriété de l'arbre* (tree property). Les noeuds intermédiaires correspondent alors à des suffixes propres de mots de T, et en particulier la racine correspond au mot vide $\emptyset$.

Si en outre tout mot semi-infini $x = x_{-\infty}^{n}$ admet exactement un suffixe dans T, on parle alors d'un *arbre de contexte* ou d'un dictionnaire complet de suffixes ; on note alors $\mathcal{T}(x)$ ce suffixe, les éléments de T sont appelés *contextes*, et on dit que x_i apparaît *en contexte* s si le suffixe de $x_{-\infty}^{i-1}$ qui appartient à T est s.

Dans la suite, T désignera à la fois, sans ambiguïté, un dictionnaire complet de suffixes et l'arbre de contexte lui correspondant. Ainsi, la taille $|T|$ de cet arbre est égale au nombre de ses feuilles, qui est le nombre de mots dont est constitué le dictionnaire T. L'ensemble des arbres de contexte de taille t sur l'alphabet A sera appelé $\mathcal{CT}_t(A)$, ou $\mathcal{CT}_t$ si le contexte ne nécessite pas de préciser l'alphabet. Noter par exemple que si $|A| \geqslant 2$, alors $\mathcal{CT}_1(A) = \{\emptyset\}$. On notera également $\mathcal{CT}(A) = \bigcup_{t \in \mathbb{N}_+} \mathcal{CT}_t(A)$ l'ensemble de tous les arbres de contexte. Par ailleurs, la profondeur de l'arbre, qui est la longueur maximale d'un mot de T, sera notée $\mathtt{depth}(T)$. Le seul arbre de contexte de profondeur nulle est $\{\emptyset\}$. Enfin, on notera $\mathcal{T}^*(s, x)$ le sous-mot (non nécessairement contigu) de x_1^n qui apparaît en contexte s. Par exemple, si $x_1^n = aabbbbabaa$ et si $x_0 = a$ alors $\mathcal{T}_1^*(a, x) = aabba$, $\mathcal{T}_1^*(ab, x) = ba$, $\mathcal{T}_1^*(bb, x) = bba$.

Etant donné un arbre de contexte T, on appelle *source à arbre de contexte de modèle* T un processus $P_{T,p}$ stationnaire tel que

$$P_{T,p}\left(X_1 = a | X_{-\infty}^{0} = x_{-\infty}^{0}\right) = P_{T,p}\left(X_1 = a | T\left(X_{-\infty}^{0}\right) = T\left(x_{-\infty}^{0}\right)\right). \qquad (5)$$

$P_{T,p}$ est défini par un $|T|$-uplet $p = (p_s)_{s \in T}$ de lois de probabilité sur A, qui est paramétré par $\theta \in \Theta_T = \left\{(\theta_s)_{s \in T} : \theta_s \in \Theta_m\right\}$. Le modèle est ainsi de dimension $|T|(m-1)$, le théorème 3 s'y applique et on a donc :

$$R_n^-(\Theta_T) \geqslant \frac{|T|(m-1)}{2} \log n.$$

Notons que $P_{T,p}$ est une chaîne de Markov d'ordre $\mathtt{depth}(T)$ (voir figure 2), et qu'une chaîne de Markov d'ordre r est un VLMC dont le modèle est défini par l'arbre m-aire complet de profondeur r (voir figure 3).

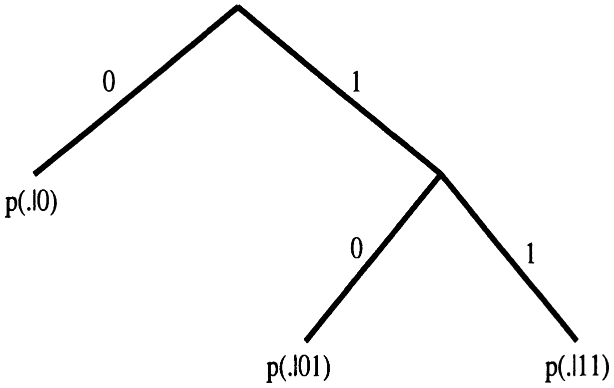
Par ailleurs, un regroupement des facteurs par contexte d'apparition dans l'équation (5) permet d'écrire une expression compacte et utile pour la vraisemblance :

$$P_{T,p}\left(x_1^n | x_{-\infty}^{0}\right) = \prod_{i=1}^{n} p_{T\left(x_{-\infty}^{i-1}\right)}(x_i) = \prod_{s \in T} p_s\left(\mathcal{T}(s, x)\right). \qquad (6)$$

Celle-ci est ainsi simplement le produit des vraisemblances des sous-mots $\mathcal{T}(s, x)$ sous les probabilités $(p_s)_{s \in T}$.

Il est donc très naturel, pour approcher les sources à arbres de contexte, d'utiliser le mélange de Krichevsky-Trofimov en définissant par analogie :

$$\mathcal{KT}_T\left(x_1^n | x_{-\infty}^{0}\right) = \prod_{s \in T} \mathcal{KT}\left(\mathcal{T}(s, x)\right). \qquad (7)$$



Chaîne de Markov de matrice de transition

$$\longrightarrow \quad M = \begin{pmatrix} p\,(.|0) \\ p\,(.|0) \\ p\,(.|01) \\ p\,(.|11) \end{pmatrix}$$

FIG. 2 – Une source à arbre de contexte vue comme chaîne de Markov

Chaîne de Markov de matrice de transition

$$M = \begin{pmatrix} p\,(.|000) \\ p\,(.|100) \\ \vdots \\ p\,(.|111) \end{pmatrix} \quad \longrightarrow$$

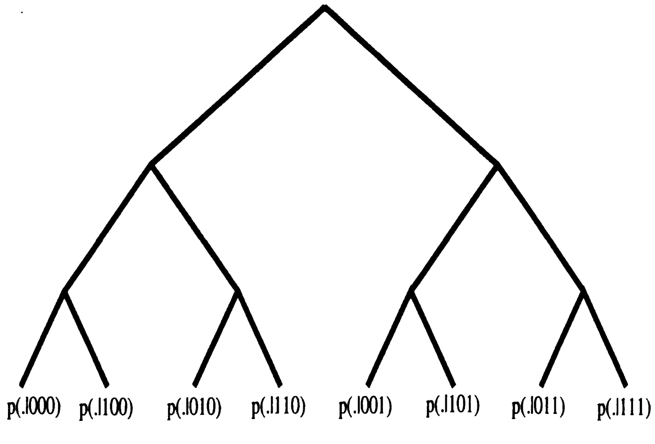


FIG. 3 – Une chaîne de Markov vue comme source à arbre de contexte

Remarque 1. *Dans cette introduction, on suppose qu'un passé $x_{-\infty}^0$ est connu par tous, de sorte que l'on peut déterminer le contexte de tous les symboles (en son absence, il serait impossible de connaître celui des quelques premiers symboles de x_1^n). Cette hypothèse est faite ici pour simplifier la discussion, il s'agit en fait d'un problème mineur qui peut être résolu de différentes manières, comme cela est expliqué dans le chapitre 3.*

La Proposition 3 d'approximation uniforme permet de démontrer l'efficacité de $\mathcal{KT}_T$ dans le modèle défini par T. La propriété suivante, prouvée dans l'article [CS00] pour les chaînes de Markov, se généralise très facilement aux VLMC (on trouvera dans [Cat01] une version très fine de ce résultat).

Proposition 4. *Pour tout entier n et pour toute chaîne semi-infinie $x_{-\infty}^n$,*

$$\begin{aligned}-\log_2 \mathcal{KT}_T\left(x_1^n|x_{-\infty}^0\right) \leqslant \inf_{\theta\in\Theta_T} &-\log_2 P_{T,p}\left(x_1^n|x_{-\infty}^0\right)\\ &+\frac{m-1}{2}|T|\log_2^+\frac{n}{|T|}+|T|\log m+m-1.\end{aligned}$$

La mesure de Krichevsky-Trofimov $\mathcal{KT}_T$ permet donc un codage asymptotiquement minimax sur le modèle défini par l'arbre de context T. Nous allons voir que l'idée de mélange peut être poussée encore plus loin.

0.2.2.3 Double mélange : l'algorithme CTW

Munissons en effet l'ensemble $\mathcal{CT}(A)$ d'une mesure de probabilité π telle que

$$\pi(T) = 2^{-2|T|+1}. \tag{8}$$

Il y a plusieurs manières de voir que π est une mesure de probabilité. On peut le vérifier directement par des arguments combinatoires de fonctions génératrices. C'est aussi la mesure qu'on obtient pour la composante connexe de la racine dans un branchement critique sur l'arbre m-aire infini (pour lequel chaque arête est présente avec probabilité $\frac{1}{2}$), cf [WT95] pour le cas binaire et [Cat01] pour des généralisations. Enfin, on peut évoquer le lemme de Kraft (Proposition 1) pour un code des arbres, voir [Bou00]. Cette mesure est uniforme sur les modèles $\mathcal{CT}_t$ de même taille t.

On peut grâce à π définir une probabilité de codage par le *double mélange* dit *Context Tree Weighting* :

$$\mathcal{CTW}\left(x_1^n|x_{-\infty}^0\right) = \sum_{T\in\mathcal{CT}} \pi(T)\mathcal{KT}_T\left(x_1^n|x_{-\infty}^0\right). \tag{9}$$

A notre connaissance, l'idée d'utiliser un double mélange apparaît pour la première fois en 1984 dans [Rya84], où Ryabko la met en oeuvre pour combiner les différents ordres markoviens. Le double mélange $\mathcal{CTW}$ est né de la collaboration de Willems, Shtarkov et Tjalkens au début des années 1990. Son fonctionnement, ainsi que les bases de son implémentation algorithmique, sont remarquablement introduits dans l'article [WT95]. Il apparaît immédiatement comme conséquence de la proposition 3 que $\mathcal{CTW}$ permet de coder dans tous les modèles à arbres de contexte à la fois.

Proposition 5. *[WT95] Pour tout entier n et pour toute chaîne semi-infinie $x_{-\infty}^{n}$,*

$$-\log_2 \mathcal{CTW}\left(x_1^n | x_{-\infty}^0\right) \leqslant \inf_{T\in\mathcal{CT}} \inf_{\theta\in\Theta_T} -\log_2 P_{T,p}\left(x_1^n | x_{-\infty}^0\right) + \frac{m-1}{2}|T| \log_2^+ \frac{n}{|T|} + |T|\left(2+\log m\right) + m - 2.$$

On parle ici d'*inégalité oracle* : la probabilité de codage $\mathcal{CTW}$ atteint asymptotiquement la vitesse minimax de chaque modèle T. Cette inégalité oracle est possible grâce au fait que la mesure π charge suffisamment chaque arbre T pour que le rapport entre $\mathcal{CTW}\left(x_1^n|x_{-\infty}^0\right)$ et $\mathcal{KT}_T\left(x_1^n|x_{-\infty}^0\right)$, supérieur à $\pi(T)$, soit négligeable devant $\frac{\inf_{\theta\in\Theta_T} P_{T,p}\left(x_1^n|x_{-\infty}^0\right)}{\mathcal{KT}_T\left(x_1^n|x_{-\infty}^0\right)}$. Autrement dit, le coût du mélange entre les différents modèles est négligeable par rapport au coût d'estimation dans chaque modèle. Si l'on veut un codeur universel pour une certaine classe de sources à arbres de contexte, il ne coûte pas beaucoup plus cher d'en construire un qui soit adaptatif pour toutes les classes de VLMC.

Ainsi, le codeur $\mathcal{CTW}$ est remarquablement efficace pour les processus à mémoire finie, dont il est un mélange. Il a connu un grand succès depuis son introduction en 1993, succès qui s'explique également par une implémentation algorithmique efficace (cf [WT95, Wil94, Cat01]).

Mais quelle est sa performance sur des classes plus riches de processus ?

0.2.2.4 Classes massives et processus de renouvellement

Pour les classes plus massives, non paramétriques, la question de savoir quel est l'ordre de grandeur de la redondance minimax est restée plus longtemps ouverte. Si l'on s'intéresse à la classe des processus stationnaires ergodiques, on sait depuis [Shi93] que l'on ne peut pas trouver de redondance non triviale (négligeable devant n). En fait, Shields et Weiss ont montré en 1996 que cela est impossible même si l'on se restreint à la classe des B-processus (c'est-à-dire des codages stationnaires de processus iid).

Théorème 7. *[Shi93, SW95] Soit $\{P_\theta : \theta \in \Theta\}$ la classe des processus stationnaires ergodiques (ou des B-processus) sur l'alphabet $\{0,1\}$. Soit f une fonction* *de* $\mathbb{N}_+$ *dans* $\mathbb{R}$ *telle que* $\lim_{n\to\infty}\frac{f(n)}{n}=0$. *Alors*

$$\lim_{n\to\infty}\frac{\bar{R}_n(\Theta)}{f(n)}=\infty.$$

Ce résultat laisse donc un trou entre les classes “simples”, paramétriques, ayant une redondance en $\frac{k}{2}\log n$ et les classes trop riches n'admettant pas de redondance non triviale.

En 1996, Csiszár et Shields ont exhibé une classe de *complexité intermédiaire* : les *processus de renouvellement.* Ce sont des processus stationnaires sur l'alphabet binaire $A=\{0,1\}$ pour lesquels les distances entre deux symboles 1 successifs sont des variables aléatoires indépendantes identiquement distribuées sur $\mathbb{N}_+$. De façon analogue, les processus de renouvellement markoviens sont des processus binaires pour lesquels les distances entre symboles 1 successifs forment une chaîne de Markov. Notons $\mathcal{R}$ la classe des processus de renouvellement sur l'alphabet binaire, et $\mathcal{MR}$ la classe des processus de renouvellement markovien. Csiszár et Shields montrent que les différents types de redondances sur ces classes sont respectivement d'ordre $\sqrt{n}$ et $n^{\frac{2}{3}}$:

Théorème 8. *[CS96] Il existe deux constantes positives c et C telles que pour tout entier n,*

$$\begin{array}{lcl} c\sqrt{n}\leqslant & R_n^-(\mathcal{R})\leqslant R_n^*(\mathcal{R}) & \leqslant C\sqrt{n}, \\ cn^{\frac{2}{3}}\leqslant & R_n^-(\mathcal{MR})\leqslant R_n^*(\mathcal{MR}) & \leqslant Cn^{\frac{2}{3}}. \end{array}$$

Comme le soulignent les auteurs, ces résultats fournissent les premiers exemples de classes de complexité intermédiaire, à la fois massives (paramétrées par toutes les lois sur les entiers) et suffisamment restreintes pour admettre des redondances non triviales. Elles constituent donc un excellent test pour examiner la performance de l'algorithme Context Tree Weighting en dehors des classes sur lesquelles celui-ci a été construit.

Dans le chapitre 3, nous prouvons que $\mathcal{CTW}$ est presque adaptatif sur les classes $\mathcal{R}$ et $\mathcal{MR}$: il atteint la redondance minimax à un facteur $\log n$ près.

Théorème 9. *Il existe des constantes C_1 et C_2 telles que la redondance ponctuelle de l'algorithme CTW sur la classe $\mathcal{R}$ des processus de renouvellement vérifie :*

$$C_1\sqrt{n}\log_2 n\leqslant R_n^*(\mathcal{CTW},\mathcal{R})\leqslant C_2\sqrt{n}\log_2 n \;\; \textit{pour tout } n\in\mathbb{N}.$$

De plus, il existe des constantes C_3 et C_4 telles que la redondance ponctuelle de l'algorithme CTW sur la classe $\mathcal{MR}$ des processus de renouvellement markovien vérifie :

$$C_3 n^{\frac{2}{3}}\log_2 n\leqslant R_n^*(\mathcal{CTW},\mathcal{MR})\leqslant C_4 n^{\frac{2}{3}}\log_2 n \;\; \textit{pour tout } n\in\mathbb{N}.$$

La preuve de ce résultat fait apparaître le fait que CTW réalise un *équilibre* entre deux types de redondances : l'*approximation* d'un processus à mémoire infinie par des sources à arbres de contexte finis d'une part, et la difficulté d'*estimation* dans les gros modèles d'autre part.

Il apparaît également indispensable d'utiliser dans le mélange des arbres très déséquilibrés et *profonds* en comparaison de la taille n du message. Ainsi, les nombreuses méthodes à arbres de contexte qui limitent la profondeur des arbres considérés à $O(\log n)$, comme par exemple [Ris99], ne peuvent avoir ce type de performance. L'algorithme CTW est donc non seulement efficace pour les processus à courte mémoire, mais il peut aussi l'être sur des classes plus complexes. En outre, ce résultat prouve qu'un codage presque minimax est possible avec un codeur algorithmiquement efficace – la preuve de [CS96] s'appuie sur le codeur NML, qui a une complexité de calcul prohibitive.

0.3 Deux problèmes de choix de modèle

La Proposition 4, assurant que la redondance minimax est de l'ordre de $|T|\frac{|A|-1}{2}\log n$ pour la classe des VLMC définie par l'arbre de contexte T, a des implications au delà de la théorie du codage. Elle peut notamment servir comme heuristique de pénalisation dans des problèmes de choix de modèles, comme nous allons l'expliquer ici.

0.3.1 Choix de modèles

Un grand nombre de travaux statistiques actuels traitent de la question de savoir quel modèle, parmi une collection définie a priori, représente le mieux des données. Pour mieux expliquer cela, nous présentons ici un des cas particuliers qui nous ont intéressé dans cette thèse et qui a été fréquemment considéré, entre autres, en linguistique ou en génomique. Supposons que l'on dispose d'un message x_1^n sur l'alphabet A, et que l'on doive décider de quel modèle de VLMC il provient. Chaque modèle est donc caractérisé par un arbre de contexte T, et il convient de retrouver "le bon" parmi tous les arbres de $\mathcal{CT}$. Plusieurs approches ont été proposées ; citons parmi elle la méthode de *pénalisation de la vraisemblance* : elle consiste à choisir le modèle $\hat{T}$ qui minimise le maximum de vraisemblance pénalisé

$$\inf_{\theta\in\Theta_T} -\log P_{T,\theta}\left(x_1^n\right) + \mathrm{pen}(n,T).$$

Le rôle de la pénalité pen est de prévenir le risque de sur-adaptation aux données : plus un modèle est riche, plus il est à même de capturer non seulement l'information, mais aussi le bruit que contient le message. Dans notre exemple, il est clair que si T est un sous-ensemble de T' alors $\inf_{\theta\in\Theta_T} -\log P_{T,\theta}\left(x_1^n\right) \geqslant$

$\inf_{\theta \in \Theta'_T} -\log P_{T',\theta}(x_1^n)$ pour *tout* message x_1^n. Il faut donc *pénaliser* la richesse de chaque modèle, et la question est dès lors de savoir comment. Précisément, nous considérons le problème suivant : supposons que x_1^n soit la réalisation d'un processus de loi P_0, et que T_0 définisse le plus petit des modèles contenant P_0. Que doit valoir $\mathrm{pen}(n, T)$ pour que l'estimateur du maximum de vraisemblance pénalisé soit fortement consistant, c'est-à-dire presque sûrement égal à T_0 à partir d'un certain rang $n > 0$?

0.3.2 Principe MDL

Pour comprendre comment un résultat de la théorie du codage peut être utile dans ce contexte, introduisons le principe "Minimum Description Length" (MDL). On fait souvent remonter son origine très loin dans l'histoire des sciences, notamment à la fameuse maxime du théologien franciscain Guillaume d'Ockham (1285 - 1349) :

Entia non sunt multiplicanda, praeter necessitatem.

Pour ce qui nous concerne plus précisément, nous reprendrons ici la formulation de Rissanen [Ris78] :

Il faut choisir le modèle qui donne la plus courte description des données.

Il conviendrait donc, dans notre exemple, de choisir le modèle T pour lequel un codeur efficace dans la classe de VLMC définie par T donne le code le moins long. Bien sûr, on obtient ainsi autant d'estimateurs que de codeurs : il resterait donc une part d'arbitraire si l'on n'était capable d'exhiber une longeur de code objective pour chaque message dans le modèle. Les éléments de la théorie du codage universel présentés précédemment permettent justement de s'affranchir de toute considération informatique, et ce de deux manières :

- en prenant comme probabilités de codage les lois Krichevsky-Trofimov, puisque celles-ci sont presque (asymptotiquement) minimax dans les classes de VLMC ; on obtient ainsi une longueur de code pour le message x_1^n égale à $-\log \mathcal{KT}_T(x_1^n)$ pour le modèle de T, et l'application du principe MDL suggère de définir l'estimateur de Krichevsky-Trofimov comme :

$$\widehat{T_{\mathcal{KT}}} = \arg\min_{T \in \mathcal{CT}} -\log \mathcal{KT}_T(x_1^n) .$$

- grâce au théorème 3 et la proposition 4, on sait qu'un codeur universel efficace pour la classe de T a une redondance asymptotiquement de l'ordre de $|T|\frac{|A|-1}{2}\log n$. On admet donc que la longueur de code du message x_1^n dans le modèle de T est la somme de la longueur de code oracle

$\inf_{\theta\in\Theta_T} -\log P_{T,\theta}(x_1^n)$ et du surcoût de redondance $|T|\frac{|A|-1}{2}\log n$. Le principe MDL conduit donc à définir l'estimateur

$$\widehat{T_{BIC}} = \arg\min_{T\in\mathcal{CT}} \inf_{\theta\in\Theta_T} -\log P_{T,\theta}(x_1^n) + |T|\frac{|A|-1}{2}\log n,$$

qui s'interprète comme un estimateur du maximum de vraisemblance pénalisé avec $\mathrm{pen}(n,T) = |T|\frac{|A|-1}{2}\log n$, et qui correspond au *critère BIC* (Bayesian Information Criterion) de Schwarz[Sch78].

Notons que dans deux cas le minimum est toujours atteint, éventuellemnent pour plusieurs modèles, et que l'estimateur peut alors être choisi quelconque parmi les minimisants.

0.3.3 Consistance de l'estimateur BIC pour les arbres de contexte

Mais il reste bien sûr à valider cette heuristique, en montrant la consistance des estimateurs $\widehat{T_{BIC}}$ et $\widehat{T_{\mathcal{KT}}}$. C'est le.travail qu'ont entrepris Csiszár et Talata dans leur article [CT06]. Ils ont montré la consistance de ces estimateurs si la minimisation est restreinte aux arbres peu profonds : si $D(n) = o(\log n)$ et si $\mathcal{F}_n = \{T \in \mathcal{CT} : \mathrm{depth}(T) < D(n)\}$, alors les estimateurs

$$\widetilde{T_{BIC}} = \arg\min_{T\in\mathcal{F}_n} \inf_{\theta\in\Theta_T} -\log P_{T,\theta}(x_1^n) + |T|\frac{|A|-1}{2}\log n$$

et

$$\widetilde{T_{\mathcal{KT}}} = \arg\min_{T\in\mathcal{F}_n} -\log \mathcal{KT}_T(x_1^n)$$

sont fortement consistants. Cette preuve, très semblable à celle qu'avaient utilisée Csiszàr et Shields dans leur article [CS00] pour montrer la consistance de l'estimateur BIC de l'ordre markovien, s'appuie sur les résultats de typicalité à grande échelle montrés par Csiszár dans [Csi02].

La restriction aux modèles peu profonds est nécessaire pour l'estimateur de Krichevsky-Trofimov : en son absence, $\widehat{T_{\mathcal{KT}}}$ se trompe en attribuant à des modèles de plus en plus complexes une suite de tirages à pile ou face (voir [CS00]). De tels phénomènes sont d'ailleurs bien connus pour les estimateurs bayesiens, voir [DF93]. On sait aussi que cette restriction n'est pas nécessaire quand il s'agit d'identifier l'ordre markovien (autrement dit, si l'on se restreint aux arbres de contexte complets à une certaine profondeur) : Csiszár et Shields l'ont montré en prouvant que les ordres "élevés" (plus grands que $D(n)$) n'étaient pas souvent choisis par le critère BIC. En revanche, les arbres de contexte proposent *bien plus de modèles par dimension* que les chaînes de Markov, ce qui a empêché les auteurs de [CT06] d'appliquer un raisonnement semblable pour les VLMC. Il est d'ailleurs habituel que la pluralité des modèles influe sur le choix de la pénalité.

La question était donc de savoir si une restriction sur la profondeur de l'arbre était indispensable pour l'estimation du modèle d'arbre de contexte. Dans le chapitre 4, il est montré que non : nous prouvons que $\widehat{T_{BIC}}$ est consistant sans condition.

Théorème 10. *Soit T_0 un arbre de contexte et P_0 un VLMC dans le modèle défini par T_0. Si X est un processus stationnaire de loi P_0, alors*

$$\widehat{T_{BIC}}(X_1^n) = T_0$$

à partir d'un certain rang n, presque sûrement.

Plus précisément, il est prouvé que le critère BIC ne choisit pas souvent les arbres "gros" (dont le nombre de feuilles est supérieur à $D(n)$). Comme la profondeur d'un arbre de contexte est toujours inférieure à sa taille, cela est suffisant pour obtenir la conclusion.

L'intérêt de ce résultat est double : d'abord, il répond à une question ouverte depuis les travaux de Csiszár et Talata, laissée jusque là en suspens. Mais le chapitre 4 contient également un algorithme qui calcule $\widehat{T_{BIC}}\,(x_1^n)$ en un temps linéaire en n, grâce à la notion d'arbre compact de suffixes et aux algorithmes d'Ukkonen, de McCreight ou de Weiner (bien décrits dans [GK97]). Ainsi, la maximisation sur tous les modèles $\mathcal{CT}$ n'induit pas de véritable surcoût de calcul, et possède les mêmes qualités de consistance sur les arbres finis – cela est d'autant plus intéressant que, comme annoncé précédemment, le chapitre 3 montre la nécessité d'utiliser des arbres très profonds pour le traitement des processus à longue mémoire.

0.3.4 Estimation d'ordre pour les HMM à émission infinie

Nous allons montrer maintenant comment la théorie de l'information peut aider à résoudre des problèmes d'estimation d'ordre plus complexes, dans un contexte où l'on perd l'interprétation du principe MDL en termes de longueur de codes mais où des inégalités analogues conduisent à des estimateurs d'ordre consistants.

Les *chaînes de Markov cachées* (HMM, pour *Hidden Markov Models*) ont, depuis leur introduction formelle en 1966 par Baum et Petrie, connu un immense succès pour la modélisation de situations aussi variées que la reconnaissance de la parole [LRS83], le traitement du signal [KV94], la météorologie [HG94] et bien sûr les biostatistiques [DEKM98]. On trouvera dans [EM02] et dans [CMR05] une très bonne exposition de ces modèles et des recherches qu'ils motivent.

Il est très fréquent dans les applications que l'*ordre* du modèle, c'est-à-dire le nombre d'états cachés, soit inconnu et doive donc être estimé. C'est un problème difficile ; une approche par maximum de vraisemblance pénalisé a été initiée dans [Fin91, Kie93], alors que des procédures bayésiennes sont proposées dans [LN94]. Comme Gassiat & Boucheron [GB03] l'ont fait pour les alphabets finis, nous montrons dans le chapitre 5 l'application de la méthodologie de Finesso et Liu & Narayan à quelques situations où l'alphabet d'émission est infini. Dans ce travail, mené en collaboration avec Antoine Chambaz et Elisabeth Gassiat, nous nous inspirons d'inégalités analogues à celles qui ont été utilisées précédemment pour construire des estimateurs d'ordre de HMM à émission gaussienne et poissonienne, et montrer leur consistance.

On dit que le processus $\{X_n\}_{n\geqslant 1}$ est une *chaîne de Markov à k états cachés* s'il existe une chaîne de Markov $\{Z_n\}_{n\geqslant 0}$ à valeur dans l'espace $\{1,\dots,k\}$ telle que, conditionnellement à Z_1^n, les variables X_n soient indépendantes, et telles que pour tout i la loi (dite d'émission) de X_i ne dépende que de Z_i – ici, cette dépendance se fera via le paramètre m_{Z_i}, qui sera la moyenne de la loi d'émission.

Nous noterons $(p_j^o : j \leqslant k) \in \mathbb{R}_+^k$ la distribution initiale de la chaîne de Markov, et $\mathcal{S}_k$ l'ensemble des probabilités de transition $\mathbf{p} = (p_{jj'} : j, j' \leqslant k) \in \mathbb{R}_+^{k^2}$ telles que pour tout $j \leqslant k$, $\sum_{j'=1}^k p_{jj'} = 1$. Ainsi, l'espace des paramètres de notre problème est

$$\Theta_k = \left\{\theta = (\mathbf{p}, \mathbf{m}) : \mathbf{p} \in \mathcal{S}_k, \mathbf{m} = (m_1, \dots, m_k) \in \mathbb{R}^k\right\},$$

de dimension $\dim \mathcal{S}_k + k = k\,(k-1) + k = k^2$. Nous étudions deux exemples de lois d'émission :

- émission gaussienne (**GE**) : pour tout $n \geqslant 1$, X_n a densité $\phi_{m_{Z_n},\sigma^2}$ conditionnellement Z_n, où ϕ_{m,σ^2} est la densité gausienne de moyenne m et variance σ^2 par rapport à la mesure de Lebesgue. Le niveau de bruit σ^2 est supposé fixé.
- émission poissonienne (**PE**) : pour tout $n \geqslant 1$, X_n a densité $\pi_{m_{Z_n}}$ conditionnellement à Z_n, où π_m est la densité de la loi de Poisson de moyenne m par rapport à la mesure de comptage sur $\mathbb{N}$.

Comme pour l'inégalité de Krichevsky-Trofimov, nous allons obtenir une inégalité d'approximation en mélangeant les différents paramètres. Pour tout $\theta \in \Theta_k$, soit g_θ la densité de $X_1^n = (X_1, \dots, X_n)$ sous θ. Pour chaque $k \geqslant 1$, on utilise la loi a priori ν_k sur Θ_k telle que, pour un certain $\tau > 0$, on ait sous ν_k :

- $\mathbf{p}$ et $\mathbf{m}$ sont indépendantes,
- la distribution initiale $p_{j'}^o = 1/k$ pour tout $j' \leqslant k$ est déterministe,
- les vecteurs $(p_{jj'} : j' \leqslant k)$ $(j \leqslant k)$ sont indépendants et suivent une loi de Dirichlet de paramètre $(1/2, \dots, 1/2)$,
- les moyennes $m_1, \dots, m_k$ sont indépendantes, identiquement distribuées selon la loi normale $\mathcal{N}_{0,\tau}$ pour l'exemple **GE**, et avec la loi Gamma$(\tau, 1/2)$

pour l'exemple **PE**.
On définit ainsi le mélange

$$q_k(X_1^n) = \int_{\Theta_k} g_\theta(X_1^n) d\nu_k(\theta).$$

Notons $X_{(n)}$ et $|X|_{(n)}$ les maximaux respectifs de $X_1, \ldots, X_n$ et de $|X_1|, \ldots, |X_n|$, et pour $k, n \geqslant 1$,

$$\begin{aligned} c_{kn} &= \log k - k \log \frac{\Gamma(k/2)}{\Gamma(1/2)} + \frac{k^2(k-1)}{4n} + \frac{k}{12n}, \\ d_{kn} &= \frac{k}{2} \log\left(\frac{\tau^2}{k\sigma^2} + \frac{1}{n}\right), \\ e_{kn} &= \frac{k}{2}\Big(1 + \tau - \log(k\tau)\Big). \end{aligned}$$

Nous montrons des inégalités d'approximation du maximum de vraisemblance par ces mélanges.

Théorème 11. *Pour tous les entiers k et n supérieurs ou égaux à 1, on a*
Pour le cas gaussien (GE) :

$$0 \leqslant \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leqslant \frac{k^2}{2} \log n + \frac{k}{2\tau^2} |X|_{(n)}^2 + c_{kn} + d_{kn}.$$

Pour le cas poissonien (PE) :

$$0 \leqslant \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leqslant \frac{k^2}{2} \log n + k\tau X_{(n)} + c_{kn} + e_{kn}.$$

On remarque l'apparition, par rapport aux inégalités connues quand l'alphabet d'émission est fini, des maxima $|X|_{(n)}^2$ et $X_{(n)}$. On n'interprète d'ailleurs plus ces inégalités dans le cadre de la théorie du codage. Il n'en demeure pas moins que, tout comme la Proposition 4 permettait de montrer (en partie) la consistance du critère BIC pour les VLMC dans le chapitre 4, ces inégalités de mélange permettent ici de montrer la consistance d'estimateurs de l'ordre k. Nous en étudions ici deux : l'un est basé sur le maximum de vraisemblance et l'autre est l'analogue dans notre contexte de l'estimateur de Krichevsky-Trofimov. Tous deux sont pénalisés par une fonction $\text{pen}(n, k)$ qui sera supposée positive, croissante en n et k et telle que pour tout $k \geqslant 1$, $\text{pen}(n, k) = o(n)$. Posons donc

$$\begin{aligned} \widehat{k}_{\text{ML}} &= \arg\min_{k \geqslant 1} \left\{ - \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) + \text{pen}(n, k) \right\} \quad \text{et} \\ \widehat{k}_{\text{MIX}} &= \arg\min_{k \geqslant 1} \left\{ - \log q_k(X_1^n) + \text{pen}(n, k) \right\}. \end{aligned}$$

Pour fixer la pénalité, on utilise le réel $\alpha > 2$ et une suite de nombres positifs $\{\varphi_n\}_{n\geqslant 1}$ qui tend vers l'infini suffisamment lentement pour que $\varphi_n = o(n)$. Pour $n, k \geqslant 1$, on introduit les sommes partielles $C_{kn} = \sum_{\ell=1}^{k} c_{\ell n}$, $C'_{kn} = \sum_{\ell=1}^{k} c'_{\ell n}$, $D_{\ell n} = \sum_{\ell=1}^{k} d_{\ell n}$ et $E_{kn} = \sum_{\ell=1}^{k} e_{\ell n}$, toutes bornées en n. Nous montrons les résultats de consistance suivants :

Théorème 12. *Soit* $S_{kn} = D_{kn} + k(k+1)\varphi_n \log n$ *dans le cas gaussien* **GE**, *et* $S_{kn} = E_{kn} + k(k+1)\frac{\log n}{\sqrt{\log\log n}}$ *dans le cas poissonien* **PE**.
Si

$$\mathrm{pen}(n,k) = \sum_{\ell=1}^{k} \frac{\ell^2+\alpha}{2} \log n + C_{kn} + S_{kn},$$

alors presque sûrement à partir d'un certain $n \geqslant 3$ *on a* $\widehat{k}_{\text{ML}} = k_0$.

En outre, si

$$\mathrm{pen}(n,k) = \sum_{\ell=1}^{k-1} \frac{\ell^2+\alpha}{2} \log n + S_{kn}.$$

alors presque sûrement à partir d'un certain $n \geqslant 3$ *on a* $\widehat{k}_{\text{MIX}} = k_0$.

Des résultats analogues sont présentés dans le chapitre 5 pour les *mélanges* gaussiens et poissoniens. Dans ce cas, légèrement plus simple, la principale différence est que la dimension du modèle à k composantes de mélange est $2k-1$ et non plus k^2.

Notons que contrairement à nombre d'estimateurs proposés jusque là, tous ces estimateurs ne nécessitent aucune borne a priori ni sur l'ordre k_0, ni sur les paramètres d'émission m_j. Remarquons également que dans le cas gaussien la pénalité est déterminée sans connaître σ^2. C'est d'ailleurs à cause de cela que pour k fixé, $\mathrm{pen}(n,k)$ est asymptotiquement plus grand que $\log n$ quand n tend vers l'infini (ainsi, notre pénalité est significativement supérieure à celle du critère BIC). Ce problème n'apparaît pas pour le cas poissonien grâce au comportement des maxima en $o(\log n)$.

Chapitre 1

Redundancy rates for classes on countably infinite alphabets

1.1 Introduction

This chapter describes universal lossless coding strategies for compresssing sources with countably infinite alphabets. It contains both comments on basic or general results and some original theorems and proofs on some specific classes of processes. It is actually a preliminar version of a current work in progress with Stéphane Boucheron and Elisabeth Gassiat (see further [BGG06]). In particular, we work at generalizing our lower bounds on the minimax redundancy of enveloppe classes, and at clarifying the adaptivity results for our final algorithm.

Let us first agree on a few definitions. A source on the countable alphabet $\mathcal{X}$ is a probability distribution on the set of infinite sequences of symbols from $\mathcal{X}$ (provided with the σ-algebra generated by cylindrical events). Henceforth, Λ will denote various classes of sources. Let $X = (X_n)_{n \in \mathbb{N}}$ be a stationary (shift-invariant) source over alphabet $\mathcal{X}$ defined by distribution $P \in \Lambda$. We denote by P^n the probability distribution of $X_1^n = X_1 \ldots X_n$ (the first n coordinate projections), and let $\Lambda^n = \{P^n : P \in \Lambda\}$. For any countable set $\mathcal{X}$, let $\mathfrak{M}_1(\mathcal{X})$ be the set of all probability measures on $\mathcal{X}$.

From Shannon noiseless coding Theorem [Sha48], the entropy of P^n, $H(X_1^n) = \mathbb{E}_{P^n}\left[-\log P(X_1^n)\right]$ (where log denotes the binary logarithm) provides a tight lower bound on the expected number of bits needed to encode outcomes of P^n. On the other hand, thanks to arithmetic coding [Ris76], any distribution $Q^n \in \mathfrak{M}_1(\mathcal{X}^n)$ defines a prefix code, encoding string x_1^n using $\lceil -\log Q^n(x_1^n) \rceil + 1$ bits. If the arithmetic code derived from distribution Q^n is used to encode outcomes from P^n, the *expected redundancy* of Q^n (with respect to P^n) is defined as the difference between the expected code length $\mathbb{E}_P\left[-\log Q^n(X_1^n)\right]$ and $H(X_1^n)$. It is equal to the

Kullback-Leibler divergence (or relative entropy) $D(P^n, Q^n) = \mathbb{E}_{P^n}\left[\log \frac{P^n(X_1^n)}{Q^n(X_1^n}\right]$.

Universal coding attempts to develop sequences of coding probabilities $(Q^n)_n$ (or equivalently prefix codes) so as to minimize expected redundancy over a whole class of sources. As a matter of fact, several distinct notions of universality have been considered in the literature. A function $\rho(n)$ is said to be a strong (resp. weak) universal redundancy rate for a class of sources Λ if there exists a sequence of coding probabilities $(Q_n)_n$ such that for all n, $\bar{R}(Q^n, \Lambda^n) \triangleq \sup_{P\in\Lambda} D(P^n, Q^n) \leqslant \rho(n)$ (resp. for all $P \in \Lambda$, there exists a constant $C(P)$ such that for all n, $D(P^n, Q^n) \leqslant C(P)\rho(n)$). Finally a class Λ of sources is said to be weakly universal if there exists a single sequence of coding probabilities $(Q^n)_n$ such that $\sup_{P\in\Lambda} \lim_n \frac{1}{n} D(P^n, Q^n) = 0$. As far as finite alphabets are concerned, it is well-known that the class of stationary ergodic sources is weakly universal. This is witnessed by the performance of Lempel-Ziv codes [ZL77, ZL78]. It is also known that the class of stationary ergodic sources over a finite alphabet does not admit any non-trivial weak universal redundancy rate [Shi93]. On the other hand, fairly large classes of sources admitting strong universal redundancy rates have been exhibited (see [Cat01] for a review). In this chapter, we will focus on strong universal redundancy rates for classes of sources over infinite alphabets. In the latter setting, even weak universality should not be taken for granted : even the class of *memoryless* processes on $\mathbb{N}_+$ is not *weakly universal.*

[Kie78] characterized weakly universal classes; the argument was simplified by [GPvdM94]. Remember that the entropy rate of a stationary source is defined as $\lim_n H(P^n)/n$.

Proposition 6. *A class Λ of stationary sources over a countable alphabet $\mathcal{X}$ is weakly universal if and only if there exists a probability distribution $Q \in \mathfrak{M}_1(\mathcal{X})$ such that for every $P \in \Lambda$ with finite entropy rate $H(P)$, Q satisfies $\mathbb{E}_P \log \frac{1}{Q(X_1)} < \infty$ or equivalently $D(P^1, Q) < \infty$.*

The characterization of strong universality has not yet reached such a degree of maturity. The *maximal redundancy* of Q^n with respect to Λ is defined by

$$\bar{R}(Q^n, \Lambda^n) = \sup_{P\in\Lambda} D(P^n, Q^n)\,.$$

The infimum of $\bar{R}(Q^n, \Lambda^n)$ is called the *minimax redundancy* with respect to Λ :

$$\bar{R}(\Lambda^n) = \inf_{Q^n \in \mathfrak{M}_1(\mathcal{X}^n)} \bar{R}(Q^n, \Lambda^n).$$

It is the smallest strong universal redundancy rate for Λ, when finite it is often called the information radius of Λ^n.

A convenient way to derive lower bounds on $\bar{R}(\Lambda^n)$ consists of endowing Λ with a prior probability distribution π, denoting a random element W from Λ selected according to π and using the relation $\mathbb{E}_\pi[D(W^n,Q^n)] \leqslant \bar{R}(Q^n,\Lambda^n)$. The sharpest lower bound is obtained by optimizing over the prior probability distribution (taking the least favorable prior), it is called the maximin bound

$$\sup_{\pi\in\mathfrak{M}_1(\Lambda)} \inf_{Q^n\in\mathfrak{M}_1(\mathcal{X}^n)} \mathbb{E}_\pi[D(W^n,Q^n)].$$

It has been proved in a series of chapters [Gal68, Dav73, Hau97] and could also have been derived from a general minimax theorem by [Sio58] that such a lower bound is tight.

Theorem 13. *Let Λ denote a class of sources over some countably finite alphabet. For each n, the minimax redundancy over Λ coincides with*

$$\bar{R}(\Lambda^n) = \sup_{\pi\in\mathfrak{M}_1(\Lambda)} \inf_{Q^n\in\mathfrak{M}_1(\mathcal{X}^n)} \mathbb{E}_\pi[D(W^n,Q^n)]\,.$$

If the set $\{P^n : P\in\Lambda\}$ is not pre-compact with respect to the topology of weak convergence, then both sides are infinite. Otherwise the maximin and minimax average redundancies are finite and coincide; moreover, the minimax redundancy is achieved by a mixture coding distribution $Q^n(.) = \int P^n(.)\mathrm{d}\pi(P)$ where π is the least favorable prior.

Another approach to universal coding considers *individual sequences* [FMG92, CBL06]. Let the *regret* of a coding distribution Q^n on string $x\in\mathbb{N}_+^n$ with respect to Λ be $\sup_{P\in\Lambda}\log P^n(x)/Q^n(x)$. Taking the maximum with respect to $x\in\mathbb{N}_+^n$, and then optimizing over the choice of Q^n, we get the *minimax regret* :

$$R^*(\Lambda^n) = \inf_{Q^n\in\mathfrak{M}_1(\mathcal{X}^n)} \max_{x\in\mathbb{N}_+^n} \sup_{P\in\Lambda} \log\frac{P^n(x)}{Q^n(x)}\,.$$

In order to provide proper insight into the stake, let us recall the precise asymptotic bounds on minimax redundancy and regret for memoryless sources over finite alphabets [CB90, CB94, BRY98, XB97, XB00, OS04, Cat01].

Proposition 7. *Let $\mathcal{X}$ be an alphabet of m symbols, and $I_\mathcal{X}$ denote the class of memoryless processes on $\mathcal{X}$ then*

$$\bar{R}(I_\mathcal{X}^n) - \frac{m-1}{2}\log\frac{n}{2\pi e} \to \log\left(\frac{\Gamma(1/2)^m}{\Gamma(m/2)}\right)\,.$$

Remark 2. *The set of memoryless sources over alphabet $\mathcal{X}=\{1,...,m\}$ is conveniently parameterized by $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\theta}\in\mathbb{R}_+^{m-1}, \sum_{i=1}^{m-1}\boldsymbol{\theta}[i]\leqslant 1\}$. For any string x_1^n from $\mathcal{X}$, let $n_j = \sum_{i=1}^n 1_{x_i=j}$ then*

$$P_{\boldsymbol{\theta}}^n(x_1^n) = \prod_{j=1}^{m-1}(\boldsymbol{\theta}[j])^{n_j}\left(1-\sum_{j=1}^{m-1}\boldsymbol{\theta}[j]\right)^{n_m}\,.$$

Jeffrey's prior $w_J()$ assigns probability proportional to the square root of Fisher Information $\sqrt{J(\boldsymbol{\theta})} = 1/\sqrt{\left(1-\sum_{j=1}^{m-1}\boldsymbol{\theta}[j]\right)\prod_{j=1}^{m-1} 1/\sqrt{\boldsymbol{\theta}[j]}}$ to parameter $\boldsymbol{\theta} \in \Theta$. Jeffrey's prior is asymptotically maximin (least favorable) : letting

$$Q^n_{w_J}(x_1^n) = \int_\Theta P^n_{\boldsymbol{\theta}}(x_1^n)\mathrm{d}w_J(\boldsymbol{\theta})$$

then

$$\lim_n \mathbb{E}_{w_J}\left[D(P^n_{\boldsymbol{\theta}}, Q^n_{w_J})\right] - \frac{m-1}{2}\log\frac{n}{2\pi e} = \log\left(\frac{\Gamma(1/2)^m}{\Gamma(m/2)}\right) .$$

Moreover, a sequence of modifications of Jeffrey's prior asymptotically achieves minimax redundancy.

Remark 3. *This phenomenon holds not only for the class of memoryless sources over a finite alphabet but also for classes of sources smoothly parameterized by finite dimensional sets [CB90, CB94, BRY98, XB97, XB00, OS04, Cat01].*

The maximin redundancy under prior π reflects the average rate of convergence of the posterior distribution in probability mass function estimation under the least favorable prior. In the case of memoryless sources over a finite alphabet, the limiting behavior of the maximin average redundancy can be considered as a consequence of the (Laplace-) Bernstein-Von Mises Theorem asserting the asymptotic normality of the rescaled posterior measure and of the relation between the entropy and the variance of Gaussian measures [CB90, CB94, BRY98, vdV98]. Hence, Theorem 7 can be considered as an information-theoretical refinement of a classical result in parametric statistics .

Let us now pay attention to the minimax regret. For a source class Λ, for every $x \in \mathcal{X}^n$, let $f_{\Lambda^n}(x) = \sup_{P\in\Lambda} P^n(x)$, $f_\Lambda()$ serves as a maximum likelihood. If $\sum_{x\in\mathbb{N}_+^n} f_{\Lambda^n}(x) < \infty$, the *Normalized Maximum Likelihood* coding probability is well-defined and given by

$$Q^n_{\mathrm{NML}}(x) = \frac{f_{\Lambda^n}(x)}{\sum_{x\in\mathbb{N}_+^n} f_{\Lambda^n}(x)} .$$

[Sht87] showed that the *Normalized Maximum Likelihood* coding probability achieves the same regret over all strings of length n and that this regret coincides with the *minimax regret* :

$$R^*(\Lambda^n) = \log\sum_{x\in\mathbb{N}_+^n} f_{\Lambda^n}(x).$$

The connection between minimax regret and minimax redundancy for memoryless sources over a finite alphabet is well-understood [BRY98, XB00].

Proposition 8. *Let $\mathcal{X}$ be an alphabet of m symbols, and $I_{\mathcal{X}}$ denote the class of memoryless processes on $\mathcal{X}$.*
The minimax individual redundancy over $I_{\mathcal{X}}$ is achieved by Shtarkov Normalized Maximum Likelihood probability. For all n, if $l = \min\{m, n\}$:

$$R^*(I^n_{\mathcal{X}}) \leqslant \frac{l-1}{2} \log \frac{n}{l} + \frac{l}{2} \log e + \log \binom{m}{l} + O(1).$$

Asymptotically

$$R^*(I^n_{\mathcal{X}}) - \frac{m-1}{2} \log \frac{n}{2\pi} \to \log \left(\frac{\Gamma(1/2)^m}{\Gamma(m/2)} \right).$$

Remark 4. *The maximum regret achieved by the mixture defined by Jeffrey's prior is within a non-null constant from the minimax regret.*

Remark 5. *The relation between $R^*(I^n_{\mathcal{X}})$ and $\bar{R}(I^n_{\mathcal{X}})$ when $\mathcal{X}$ is a finite alphabet of size m can be related to another classical results from asymptotic parameterized statistics. Let $\hat{\boldsymbol{\theta}} \in \Theta$ denote the parameter that achieves maximum likelihood on x_1^n. Simple algebra shows that*

$$\begin{aligned} D(Q^n_{w_J}, Q^n_{\text{NML}}) &= R^*(\Lambda^n) - \bar{R}(\Lambda^n) \\ &\quad - \int_{\Theta} w_J(\boldsymbol{\theta}) \mathbb{E}_{P^n_{\boldsymbol{\theta}}} \left[\log \frac{P^n_{\hat{\boldsymbol{\theta}}}(X^n_1)}{P^n_{\boldsymbol{\theta}}(X^n_1)} \right] \mathrm{d}\boldsymbol{\theta}. \end{aligned}$$

Then, by Wilks' Theorem [vdV98], the last summand converges toward one half the expectation of a χ^2_{m-1} distributed random variable (using natural logarithms).
Moreover, [XB97] have shown that :

$$D\left(Q^n_{w_J}, Q^n_{\text{NML}}\right) \to 0.$$

This holds for a variety of classes of sources smoothly parameterized by finite-dimensional sets [BRY98] .

In this chapter, we will bound redundancy rates and minimax regret for classes of memoryless sources over a countably infinite alphabet. Source classes defined by an envelope distribution will be of primary interest in the sequel.

Definition 1. *For a mapping f from $\mathbb{N}_+$ to $[0, 1]$, let $\bar{F}(u) = \sum_{k>u} f(u)$. Let the set of probability distributions Λ^1_f be defined as :*

$$\Lambda^1_f = \{p \in \mathfrak{M}_1(\mathbb{N}_+) : \forall k \in \mathbb{N}_+, p(k) \leqslant f(k)\}.$$

The source class defined by envelope function f, Λ_f is the class of stationary memoryless sources with marginal distributions in Λ^1_f.

Note that the envelope function satisfies $f(k) = f_{\Lambda^1_f}(k)$, it is the maximum likelihood defined over the source class Λ^1_f on a single observation.

In Section 1.3, we will focus our attention on two kinds of envelope source classes illustrating two different asymptotic behaviors. First, The "power-law" source class $\Lambda_{M \cdot^{-\alpha}}$ is associated to the slowly decreasing function $f_{\alpha,M} : x \mapsto \frac{M}{x^\alpha}$ for $M > 1$ and $\alpha > 1$. Second, the "exponential" source class $\Lambda_{C\mathrm{e}^{-\alpha\cdot}}$ is defined by the faster decreasing envelope function $x \mapsto C\mathrm{e}^{-\alpha x}$.

We will be mostly concerned with the following questions.

1. Characterizing those sources classes that have finite minimax regret.
2. Relating minimax regret and minimax redundancy without resorting to Wilks' Theorem.
3. Relating minimax redundancy and the complexity of source classes
4. Developing adaptive coding schemes for collection of source classes that are too large to enjoy even a weak redundancy rate.

The chapter is organized as follows. In Section 1.2, some structural properties of minimax redundancies and regrets for classes of stationary memoryless sources are described. Those properties include monotonicity and sub-additivity. Proposition 12 provides a simple upper-bound on minimax regret using the envelope tail. Theorem 14 characterizes those source classes that admit finite regret. This characterization emphasizes the role of Shtarkov Normalized Maximum Likelihood coding probability. Proposition 13, describes a simple source class for which the minimax regret is infinite, while the minimax redundancy is finite. Finally, Proposition 15 shows that such a contrast is not possible for the so-called envelope classes.

Section 1.3 focuses on the two kinds of envelope classes we consider in this chapter. In Subsection 1.3.1, lower-bounds on minimax redundancy and upper-bounds on minimax regret for power-lax classes are described. Up to a factor $\log n$ those bounds are matching. In Subsection 1.3.2, lower-bounds on minimax redundancy and upper-bounds on minimax regret for exponential classes are described. Up to a multiplicative constant, those bounds are matching and grow like $\log^2 n$.

In Section 1.4, we turn back to effective coding techniques geared toward source classes defined by power-law envelopes. In Subsection 1.4.1, we elaborate on the ideas embodied in Proposition 12 from Section 1.2, and combine mixture coding and Elias penultimate code [Eli75] to match the upper-bounds on minimax redundancy described in Section 1.3. One of the messages from Section 1.3, is that the union of envelope classes defined by power laws, does not admit a weak redundancy rate that grows at a rate slower than $n^{1/\beta}$ for any $\beta > 1$. In Subsection 1.4.2, we propose an adaptive coding scheme for the union of envelope classes defined by power laws. This adaptive coding scheme combines the censoring coding technique developed in the preceding subsection and an estimation of tail-heaviness.

1.2 Properties of the minimax redundancy and minimax regret

We start with two propositions (10 and 11) that are both easy and useful. In order to prove them, we will use the following proposition which emphasizes the role of the NML coder with respect to the minimax regret. At best, this is a comment on Shtarkov original work [Sht87].

Proposition 9. *Let Λ be a class of stationary memoryless sources over a countably infinite alphabet, the minimax regret with respect to Λ^n, $R^*(\Lambda^n)$ is finite if and only if the normalized maximum likelihood (Shtarkov) coding probability Q^n_{NML} is well-defined and given by*

$$Q^n_{\mathrm{NML}}(x_1^n) = \frac{P_{x_1^n}(x_1^n)}{\sum_{y_1^n \in \mathcal{X}^n} P_{x_1^n}(x_1^n)}$$

with $P_{x_1^n}(x_1^n) = \sup_{P \in \Lambda} P^n(x_1^n)$.

Note that the definition of Q^n_{NML} does not assume either that the maximum likelihood is achieved or that it is uniquely defined.

Proof. That the minimax regret is finite and equal to

$$\log \left(\sum_{y_1^n \in \mathcal{X}^n} P_{y_1^n}(y_1^n) \right)$$

if $\sum_{y_1^n \in \mathcal{X}^n} P_{y_1^n}(y_1^n) < \infty$ is the fundamental observation of [Sht87].

On the other hand, if $R^*(\Lambda^n) < \infty$, there exists a probability distribution Q^n on $\mathcal{X}^n$ and a finite number r such that for all $x_1^n \in \mathcal{X}^n$,

$$P_{x_1^n}(x_1^n) \leqslant r Q^n(x_1^n) ,$$

summing over the x_1^n gives

$$\sum_{x_1^n \in \mathcal{X}^n} P_{x_1^n}(x_1^n) \leqslant r < \infty .$$

■

Proposition 10. *Let Λ denote a class of sources, then the minimax redundancy $\bar{R}(\Lambda^n)$ and the minimax regret $R^*(\Lambda^n)$ are non-decreasing functions of n.*

Proof. As far as $\bar{R}$ is concerned, by Theorem 13, it is enough to check that the maximin lower bound is non-decreasing. Let π be any prior, and $I(W; X_1^n) = H(X_1^n) - H(X_1^n|W)$ be the mutual information between W and X_1^n.

$$I(W; X_1^{n+1}) = I(W; (X_1^n, X_{n+1})) \geqslant I(W; X_1^n).$$

The conclusion follows, as $\bar{R}(\Lambda^n) = \inf_\pi I(W; X_1^n)$.

Let us now consider the minimax regret. It is enough to consider the case where $R^*(\Lambda^n)$ is finite. Thus we may rely on Proposition 9. Let n be a positive integers and let ϵ be a small positive real. For every string $x_1^{n+1} \in \mathcal{X}^{n+1}$, there is some let $Q \in \Lambda$, be such that $Q(x_1^n) \geqslant P_{x_1^n}(x_1^n)(1-\epsilon)$. Then

$$\begin{aligned} P_{x_1^{n+1}}(x_1^{n+1}) &\geqslant Q(x_1^n) \times Q(x_{n+1} \mid x_1^n) \\ &\geqslant P_{x_1^n}(x_1^n)(1-\epsilon) \times Q(x_{n+1} \mid x_1^n) . \end{aligned}$$

Summing over all possible x_{n+1}, we get

$$\sum_{x_{n+1}} P_{x_1^{n+1}}(x_1^{n+1}) \geqslant P_{x_1^n}(x_1^n)(1-\epsilon) .$$

Summing now over all x_1^n,

$$\sum_{x_1^{n+1}} P_{x_1^{n+1}}(x_1^{n+1}) \geqslant \sum_{x_1^n} P_{x_1^n}(x_1^n)(1-\epsilon) .$$

So that

$$R^*(\Lambda^{n+1}) \geqslant R^*(\Lambda^n)(1-\epsilon)$$

As ϵ can be chosen arbitrary small, the result follows.

■

Note that this lemma does not depend on the assumption that sources are memoryless. Moreover, it can be easily completed when dealing with memoryless sources.

Proposition 11. *If Λ is a class of stationary memoryless sources, then the functions $n \mapsto \bar{R}(\Lambda^n)$ and $n \mapsto R^*(\Lambda^n)$ are sub-additive.*

Proof. Here again, given Theorem 13, in order to establish sub-additivity for $\bar{R}$, it is enough to check the property for the maximin lower bound. Let n, m be two positive integers, and π be any prior on Λ. As the source is memoryless, X_1^n and X_{n+1}^{n+m} are independent conditionally on W and thus

$$\begin{aligned}
& I\left(X_{n+1}^{n+m}; W|X_1^n\right) \\
&= H\left(X_{n+1}^{n+m}|X_1^n\right) - H\left(X_{n+1}^{n+m}|X_1^n, W\right) \\
&= H\left(X_{n+1}^{n+m}|X_1^n\right) - H\left(X_{n+1}^{n+m}|W\right) \\
&\leqslant H\left(X_{n+1}^{n+m}\right) - H\left(X_{n+1}^{n+m}|W\right) \\
&= I\left(X_{n+1}^{n+m}; W\right)
\end{aligned}$$

Hence, using the fact that process $(X_n)_{n\in\mathbb{N}_+}$ is stationary :

$$\begin{aligned}
I\left(X_1^{n+m}; W\right) &= I\left(X_1^n; W\right) + I\left(X_{n+1}^{n+m}; W|X_1^n\right) \\
&\leqslant I\left(X_1^n; W\right) + I\left(X_{n+1}^{n+m}; W\right) \\
&= I\left(X_1^n; W\right) + I\left(X_1^m; W\right).
\end{aligned}$$

Let us now check the sub-additivity of minimax regret. Let n and m be two positive integers. For every string $x_1^{n+m} \in \mathcal{X}^{n+m}$:

$$\begin{aligned}
P_{x_1^{n+m}}\left(x_1^{n+m}\right) &= P_{x_1^{n+m}}\left(x_1^n\right) P_{x_1^{n+m}}\left(x_{n+1}^{n+m}\right) \\
&\leqslant P_{x_1^n}\left(x_1^n\right) P_{x_{n+1}^{n+m}}\left(x_{n+1}^{n+m}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
& R^*\left(\Lambda^{n+m}\right) \\
&= \log \sum_{x_1^{n+m}\in\mathcal{X}^{n+m}} P_{x_1^{n+m}}\left(x_1^{n+m}\right) \\
&= \log \left[\sum_{x_1^n\in\mathcal{X}^n} P_{x_1^{n+m}}\left(x_1^n\right) \sum_{x_{n+1}^{n+m}\in\mathcal{X}^n} P_{x_1^{n+m}}\left(x_{n+1}^{n+m}\right)\right] \\
&= \log \sum_{x_1^n\in\mathcal{X}^n} P_{x_1^{n+m}}\left(x_1^n\right) + \log \sum_{x_{n+1}^{n+m}\in\mathcal{X}^m} P_{x_1^{n+m}}\left(x_{n+1}^{n+m}\right) \\
&\leqslant \log \sum_{x_1^n\in\mathcal{X}^n} P_{x_1^n}\left(x_1^n\right) + \log \sum_{x_{n+1}^{n+m}\in\mathcal{X}^m} P_{x_{n+1}^{n+m}}\left(x_{n+1}^{n+m}\right) \\
&= R^*\left(\Lambda^n\right) + R^*\left(\Lambda^m\right).
\end{aligned}$$

■

Remark 6. *One can manufacture counter-examples that witness the fact that sub-additivity of redundancies does not hold in full generality. Here is a very simple example with stationary ergodic sources.*

Let ϵ be small positive number. Let $\Lambda = \{P_1, P_2\}$, where P_1 and P_2 are two stationary Markov sources on alphabet $\mathcal{X}$ with probability transitions

$$\left\{\begin{array}{lcll} P_1\left(X_{n+1}=1|X_n=1\right) & = & 1-\epsilon, & \\ P_1\left(X_{n+1}=0|X_n=0\right) & = & 1-\epsilon & \textit{and} \\ P_2\left(X_{n+1}=0|X_n=1\right) & = & 1-\epsilon, & \\ P_2\left(X_{n+1}=1|X_n=0\right) & = & 1-\epsilon. & \end{array}\right.$$

Hence, the kernel of P_1 almost always faithfully transmits bits while the kernel of P_2 aims at flipping them. The stationary distribution of both sources is uniform on $\mathcal{X}$.

Let W be a random variable on $\{1,2\}$, $(X_n)_n$ be the process on $\mathcal{X}$ with distribution P_j conditionally on $W = j$ for $j = 1, 2$. It is obvious that $I(W; X_1) = 0$ (seeing one symbol offers no information on W), while $\lim_{\epsilon\to 0} I(W; X_1X_2) = 1$ (for example, if $X_1 = X_2$ then with high probability the source is P_1).

The Fekete Lemma, see [PS98], leads to the following corollary :

Corollary 1. *Let Λ denote a class of stationary memoryless sources over a countable alphabet. For both minimax redundancy $\bar{R}()$ and minimax regret R^*,*

$$\lim_{n\to\infty} \frac{\bar{R}\left(\Lambda^n\right)}{n} = \inf_{n\in\mathbb{N}_+} \frac{\bar{R}\left(\Lambda^n\right)}{n} \leqslant \bar{R}\left(\Lambda^1\right),$$

and

$$\lim_{n\to\infty} \frac{R^*\left(\Lambda^n\right)}{n} = \inf_{n\in\mathbb{N}_+} \frac{R^*\left(\Lambda^n\right)}{n} \leqslant R^*\left(\Lambda^1\right),$$

Hence, in order to prove that $\bar{R}\left(\Lambda^n\right) < \infty$ (resp. $R^*\left(\Lambda^n\right) < \infty$), it is enough to check that $\bar{R}\left(\Lambda^1\right) < \infty$ (resp. $R^*\left(\Lambda^1\right) < \infty$,).

Proposition 12. *If Λ is a class of memoryless sources, let f denote the envelope of Λ^1, and let the tail function $\bar{F}_{\Lambda^1}$ be defined by $\bar{F}_{\Lambda^1}(u) = \sum_{k\geqslant u} f(k)$, then :*

$$R^*(\Lambda^n) \leqslant \inf_{u:u\leqslant n} \left[n\bar{F}_{\Lambda^1}(u) \log e + \frac{u-1}{2} \log \frac{en}{u} + O(1) \right].$$

Proof. Any integer u defines a decomposition of a string $x \in \mathbb{N}_+^n$ into two non-contiguous substrings : one substring $k_1^Z = k_1 \ldots k_Z$ made of the Z symbols from x that are larger than u, and one substring y_1^{n-Z} made of the remaining smaller symbols.

$$
\begin{aligned}
&\sum_{x_1^n \in \mathbb{N}_+^n} f_{\Lambda^n}(x_1^n) \\
&\overset{(a)}{=} \sum_{Z=0}^{n} \binom{n}{Z} \sum_{k_1 \dots k_Z > u} \sum_{y_1^{n-Z} \in \{1,2,\dots,u\}^{n-Z}} f_{\Lambda^n}\left(k_1^Z y_1^{n-Z}\right) \\
&\overset{(b)}{\leqslant} \sum_{Z=0}^{n} \binom{n}{Z} \sum_{k_1 \dots k_Z > u} \prod_{i=1}^{Z} f_{\Lambda^1}(k_i) \sum_{y_1^{n-Z} \in \{1,2,\dots,u\}^{n-Z}} f_{\Lambda^{n-z}}\left(y_1^{n-Z}\right) \\
&\overset{(c)}{\leqslant} \left(\sum_{Z=0}^{n} \binom{n}{Z} \bar{F}_{\Lambda^1}(u)^Z\right) \left(\sum_{y_1^n \in \{1,2,\dots,u\}^n} f_{\Lambda^n}\left(y_1^n\right)\right) \\
&\overset{(d)}{\leqslant} \left(1 + \bar{F}_{\Lambda^1}(u)\right)^n 2^{\frac{u-1}{2} \log \frac{n}{u} + \frac{u}{2} + O(1)} .
\end{aligned}
$$

Equation (a) is obtained by reordering the symbols in the strings, Inequalities (b) and (c) follow respectively from Proposition 11 and Proposition 10. Inequality (d) is a direct consequence of Proposition 7.

Hence,

$$
\begin{aligned}
R^*(\Lambda^n) &\leqslant n \log\left(1 + \bar{F}_{\Lambda^1}(u)\right) + \frac{u-1}{2} \log \frac{en}{u} + O(1) \\
&\leqslant n \bar{F}_{\Lambda^1}(u) \log e + \frac{u-1}{2} \log \frac{en}{u} + O(1)
\end{aligned}
$$

■

The following Theorem combines Propositions 9, 10, 11 and 12. The main message of the theorem may be rephrased as follows : a class of memoryless sources admits a non-trivial strong minimax individual redundancy rate if and only if Shtarkov NML coding probability is well-defined for $n = 1$.

Theorem 14. *Let Λ be a class of stationary memoryless sources over a countably infinite alphabet, the minimax individual redundancy with respect to Λ^n is finite if and only if the normalized maximum likelihood (Shtarkov) coding probability is well-defined and :*

$$R^*(\Lambda^n) < \infty \Leftrightarrow \sum_{k \in \mathbb{N}_+} f_{\Lambda^1}(x) < \infty,$$

moreover, when $R^*(\Lambda^n) < \infty$, it is achieved by the NML *coding probability and the following holds :*

$$R^*(\Lambda^n) < \infty \Leftrightarrow R^*(\Lambda^n) = o(n) .$$

Proof. Let us start by checking the converse part. If $\sum_{k\in\mathbb{N}_+} f_{\Lambda^1}(x) = \infty$, then $R^*(\Lambda^1) = \infty$ and from Proposition 10, $R^*(\Lambda^n) = \infty$ for every positive integer n.

For the direct part of the statement, let us assume that $\sum_{x\in\mathbb{N}_+^n} f_{\Lambda^1}(x) < \infty$. This assumption entails that $\bar{F}_{\Lambda^1}(u) \to 0$ as $u \to \infty$. Let us choose a sequence $(u_n)_{n\in\mathbb{N}_+}$ going to infinity slowly enough so as to ensure that $\frac{u_n}{n} = o(1)$. From Proposition 12, as $x \mapsto x\log x$ goes to 0 with x, it holds that :

$$R^*(\Lambda^n) \leqslant n\bar{F}_{\Lambda^1}(u_n)\log e + \frac{u_n - 1}{2}\log\frac{en}{u_n} + O(1) = o(n).$$

■

Remark 7. *When dealing with smoothly parameterized classes of sources over finite alphabets, the maximal individual redundancy and minimax average redundancy are usually of the same order of magnitude [BRY98, XB00], even though the coding probabilities achieving the minimax average and minimax individual redundancy differ [XB00]. A comparable phenomenon also holds for the massive classes defined by renewal sources investigated by [CS96]. This can not be taken for granted when dealing with classes of stationary memoryless sources over a countable alphabet.*

Up to this point, the behavior of minimax redundancies and regrets have seemed quite similar. However, the remark relating the asymptotic gap between the two quantities for smoothly parameterized source classes to Wilk's Theorem suggests that the two quantities could behave in dramatically different ways. The following proposition exhibits a class of sources for which minimax redundancy and minimax regret behave in such a contrasted way.

Proposition 13. *Let f be a positive, strictly decreasing function defined on $\mathbb{N}$ such that $f(1) < 1$. For $k \in \mathbb{N}$, let p_k be the probability mass function on $\mathbb{N}$ defined by :*

$$p_k(l) = \begin{cases} 1 - f(k) & \textit{if } \; l = 0; \\ f(k) & \textit{if } \; l = k; \\ 0 & \textit{otherwise.} \end{cases}$$

Let $\Lambda^1 = \{p_1, p_2, \ldots\}$, let Λ be the class of stationary memoryless sources with first marginal Λ^1. Then

$$f(k)\log k \to_{k\to\infty} \infty \Leftrightarrow \bar{R}(\Lambda^n) = \infty \textit{ for every positive integer } n.$$

Remark 8. *When $f(k) = \frac{1}{\log k}$, the minimax redundancy is finite and strong universal coding is still possible. However, as $\sum_k f(k) = \infty$, minimax individual redundancy is infinite by Theorem 14.*

Proof. Let us first prove the direct part. Assume that $f(k)\log k \to_{k\to\infty} \infty$. In order to prove that $\bar{R}(\Lambda^1)=\infty$, we resort to Theorem 13 and describe an appropriate collection of Bayesian games.

Let m be a positive integer and let W be uniformly distributed on $\{p_1, p_2, \dots, p_m\}$. Let X be distributed according to p_k conditionally on $W=k$. Let Z be the random variable equal to 1 if $X=W$ and equal to 0 otherwise. Obviously, $H(W|X, Z=1)=0$; moreover, as f is assumed to be non-increasing, $P(Z=0|W=k)=1-f(k)\leqslant 1-f(m)$ and thus :

$$\begin{aligned} H(W|X) &= H(Z|X)+H(W|Z,X) \\ &\leqslant 1+P(Z=0)H(W|X,Z=0) \\ &\quad +P(Z=1)H(W|X,Z=1) \\ &\leqslant 1+(1-f(m))\log m. \end{aligned}$$

Hence,

$$\begin{aligned} \bar{R}(\Lambda^1) &\geqslant I(W,X) \\ &\geqslant \log m-(1-f(m))\log m-1 \\ &= f(m)\log m-1 \end{aligned}$$

which grows to infinity with m, so that as forecast, $\bar{R}(\Lambda^1)=\infty$.

Let us now prove the converse part. Assume that the sequence $(f(k)\log k)_{k\in\mathbb{N}_+}$ is upper-bounded by some constant C. In order to check that $\bar{R}(\Lambda^n)<\infty$, for all n it is enough to exhibit a probability distribution Q over $\mathcal{X}=\mathbb{N}$ such that $\sup_{p\in\Lambda^1} D(p,Q)<\infty$.

Let Q be defined by $Q(k)=Z/(k(\log k)^2)$ for $k\geqslant 2$, $Q(0), Q(1)>0$ where Z is a normalizing constant that ensures that Q is a probability distribution over $\mathcal{X}$.

Then for any $k\geqslant 2$:

$$\begin{aligned} D(p_k,Q) &= (1-f(k))\log\frac{(1-f(k))}{Q(0)}+f(k)\log\left(\frac{f(k)k(\log k)^2}{Z}\right) \\ &\leqslant -\log Q(0)+C+f(k)\left(2\log^{(2)}(k)-\log(Z)\right) \\ &\leqslant 3C+\log\frac{1}{Z\,Q(0)}. \end{aligned}$$

This is enough to conclude that

$$\bar{R}(\Lambda^1)<\infty.$$

■

The following theorem shows that as far as envelope classes are concerned, minimax redundancy and minimax regret are either both finite or both infinite.

Theorem 15. *Let f define a non-negative function from $\mathbb{N}_+$, to $[0,1]$ let Λ_f be the class of stationary memoryless sources defined by envelope f. Then*

$$\bar{R}\left(\Lambda_f^n\right) < \infty \Leftrightarrow R^*\left(\Lambda_f^n\right) < \infty$$

and

$$R^*\left(\Lambda_f^n\right) < \infty \Leftrightarrow \sum_{k\in\mathbb{N}_+} f(k) < \infty.$$

Proof. The second statement has already been established (see Theorem 14). In order to check the non-trivial part of the first statement, it is enough to check that if $\sum_{k\in\mathbb{N}_+} f(k) = \infty$ then the envelope class Λ_f contains an infinite collection of mutually singular sources, which immediately implies that $\bar{R}\left(\Lambda_f^n\right) = \infty$ by Kieffer's criterion.

Let the infinite sequence $(p_i)_{i\in\mathbb{N}}$ be defined recursively by $p_0 = 0$ and

$$p_{i+1} = \min\left\{p \ : \ \sum_{k=p_i+1}^{p} f(k) > 1\right\}.$$

The memoryless source P_i is defined by its first marginal P_i^1 which is given by

$$P_i^1(m) = \frac{f(m)}{\sum_{k=p_i+1}^{p_{i+1}} f(k)} \text{ for } m \in \{p_i+1, ..., p_{i+1}\}.$$

Taking any prior with infinite Shannon entropy over the $(\{P_i^1 \ ; \ i \in n\mathbb{N}_+\}$ shows that

$$\bar{R}\left(\{P_i^1 \ ; \ i \in \mathbb{N}_+\}\right) = \infty.$$

■

1.3 Redundancies and regrets for some envelope classes

Theorem 15 asserts that the summability of the envelope defining a class of memoryless sources characterizes the (strong) universal compressibility of that class. This Theorem is a qualitative statement. It is rather natural to wonder whether the relationship between the different kinds of redundancies is tight for general envelope classes as is the case when the envelope has finite support.

In this section, we investigate the case of envelopes which decline either like power laws or exponentially fast. The derivations of lower-bounds on minimax redundancies follow a common pattern : a finite collection of sources from the envelope class is explicitly constructed and endowed with a prior probability. This defines a joint probability distribution over couples of sources and sequences of symbols from the alphabet. The mutual information between source and observations is then bounded from below. This may actually define a sequence of lower bounds on the maximin average redundancy. The final result is obtained by taking the supremum over those lower bounds. The ultimate goal is to check whether the upper-bound derived from Proposition 12 is tight.

1.3.1 Power-law envelope classes

Let us agree on the classical notation :

$$\zeta(\alpha) = \sum_{k \geqslant 1} \frac{1}{k^\alpha}\,, \text{ for } \alpha > 1\,.$$

Theorem 16. *Let α denote a real number larger than* 1, *and M be such that $M\zeta(\alpha) > 2^\alpha$. The source class $\Lambda_{M\cdot^{-\alpha}}$ is the envelope class associated with the decreasing function*

$$f_{\alpha,M} : x \mapsto \frac{M}{x^\alpha}$$

for $M > 1$ and $\alpha > 1$. Then :

$$n^{1/\alpha} A(\alpha) \log \left\lfloor (M\zeta(\alpha))^{\frac{1}{\alpha}} \right\rfloor \leqslant \bar{R}(\Lambda^n_{M\cdot^{-\alpha}})$$

and

$$R^*(\Lambda^n_{M\cdot^{-\alpha}}) \leqslant \left(\frac{2Mn}{\alpha-1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1),$$

where

$$A(\alpha) = \frac{1}{\alpha}\int_0^\infty \frac{1}{u^{1-1/\alpha}} \left(1 - e^{-1/(\zeta(\alpha)u)}\right) \mathrm{d}u\,.$$

Remark 9. *There is small gap of order $(\log n)^{1-\frac{1}{\alpha}}$ between the lower- and the upper-bound. We are not in position to claim that one of the two bounds is tight, let alone which one is tight. Note however that as $\alpha \to \infty$ and $M = H^\alpha$, class $\Lambda_{M\cdot^{-\alpha}}$ converges to the class of memoryless sources on alphabet $\{1, \ldots, H\}$ for which the minimax individual redundancy is $\frac{H-1}{2}\log n$. This is (up to a factor 2) the limit we obtain by taking the limits in our upper-bound of $R^*(\Lambda^n_{M\cdot^{-\alpha}})$.*

Remark 10. *The proof of the lower bound on redundancy deserves a comment (which is also relevant for all similar lower bounds in this thesis). It provides for each length n, a probability distribution over the envelope class that witnesses the lower bound. The probability distribution varies with n. The construction of a single probability distribution over $\Lambda_{M,\cdot^{-\alpha}}$ that achieves the lower bound for infinitely many n is an open question.*

Remark 11. *In [BGG06] (work in progress), we aim to prove a general lower bound for the minimax redundancy of envelope classes. The technique we use is closer to the proof of Theorem 17. However, it is interesting to see that the following rudimentary argument gives an almost optimal result.*

Proof. For the upper-bound on minimax regret, note that

$$\bar{F}_{\alpha,M}(u) = \sum_{k>u} 1 \wedge \frac{M}{k^{\alpha}} \leqslant \frac{M}{(\alpha-1)\,u^{\alpha-1}}.$$

Hence, choosing $u_n = \left(\frac{2Mn}{(\alpha-1)\log n}\right)^{\frac{1}{\alpha}}$, resorting to Proposition 12, we get :

$$R^*(\Lambda^n_{M\cdot^{-\alpha}}) \leqslant \left(\frac{2Mn}{\alpha-1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1).$$

Let us now turn to the lower bound. We first define a finite set Θ of parameters such that $P^n_{\boldsymbol{\theta}} \in \Lambda^n_{\alpha,M}$ for any $\boldsymbol{\theta} \in \Theta$ and then we use the fact that

$$\bar{R}(\Lambda^n_{\alpha,M}) \geqslant I(W; X_1^n),$$

where W is uniformly distributed on Θ, and the distribution of X_1^n conditionally on $W = \boldsymbol{\theta} \in \Theta$ is the product probability distribution $P^n_{\boldsymbol{\theta}}$.

Let m be a positive integer such that $m^{\alpha} < M\zeta(\alpha)$. For any sufficiently large integer p, we have

$$m^{\alpha} \leqslant M \sum_{k=1}^{p} \frac{1}{k^{\alpha}}.$$

Henceforth, let $c_p^{-1} = \sum_{k=1}^{p} \frac{1}{k^{\alpha}}$, so that the previous condition translates into $c_p m^{\alpha} \leqslant M$.

The set $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ will consist of memoryless sources on the finite alphabet $\{1, ..., p \times m\}$. Each parameter $\boldsymbol{\theta}$ is a tuple $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \{1, ..., m\}^p$. For any such $\boldsymbol{\theta}$, $P^1_{\boldsymbol{\theta}}$ is a probability distribution on $\{1, \ldots, p \times m\}$ with support $\cup_{k \leqslant p}\{(k-1)p + \theta_k\}$, namely :

$$P_{\boldsymbol{\theta}}((k-1)m + \theta_k) \;=\; a_k \quad \text{for } k = 1, \ldots, p\,,$$

where for $k = 1, \dots, p$:

$$a_k = \frac{c_p}{k^\alpha}.$$

The condition $m^\alpha c_p \leqslant M$ ensures that $P^1_{\boldsymbol{\theta}} \in \Lambda^1_{\alpha,M}$.

The prior distribution over $\Lambda^n_{\alpha,M}$ is uniform over $\{P^n_{\boldsymbol{\theta}} \; : \; \boldsymbol{\theta} \in \Theta\}$. Now, the entropy of W is just the logarithm of the cardinality of Θ, that is $p \log m$, and the entropy of W conditionally to $X_1^n = x_1^n$ is upper bounded by the logarithm of the cardinality of the support of the distribution of W conditionally to $X_1^n = x_1^n$.

Let C_n be the number of distinct symbols in $\{X_1, \dots, X_n\}$. Then the entropy of W conditionally to $X_1^n = x_1^n$ is-upper bounded by $(p - C_n)\log m$. Thus

$$\bar{R}(\Lambda^n_{\alpha,M}) \geqslant I(W; X_1^n) \geqslant \left(\frac{1}{m^p} \sum_{\boldsymbol{\theta} \in \Theta} E_{\boldsymbol{\theta}} \left[C_n \right] \right) \log m.$$

The expectation does not depend on the value of $\boldsymbol{\theta}$. For any $x \in \mathbb{N}_+$, define $U_x = \sum_{i=1}^n 1_{X_i = x}$. Then

$$C_n = \sum_{x \geqslant 1} 1_{U_x \geqslant 1}. \tag{1.1}$$

Thus,

$$E_\theta \left[C_n \right] = \sum_{k=1}^{p} \left(1 - (1 - a_k)^n \right).$$

This leads to :

$$\bar{R}(\Lambda^n_{\alpha,M}) \geqslant \left(\sum_{k=1}^{p} \left(1 - (1 - a_k)^n \right) \right) \times \log m\,.$$

In order to optimize the bound we choose $m^\alpha = \lfloor M\zeta(\alpha) \rfloor$.

The sum can be lower-bounded by :

$$\begin{aligned}
\sum_{k=1}^{p} \left(1 - (1 - a_k)^n\right) &\geqslant \sum_{k=1}^{p} \left(1 - \exp(-n a_k)\right) \qquad \text{as } 1 - x \leqslant \exp(-x) \\
&= \sum_{k=1}^{p} \left(1 - \exp\left(-\frac{n c_p}{k^\alpha}\right)\right) \\
&\geqslant \int_1^p \left(1 - \exp\left(-\frac{n c_p}{x^\alpha}\right)\right) \mathrm{d}x \\
&= \frac{n^{\frac{1}{\alpha}}}{\alpha} \int_{1/n}^{p^\alpha/n} \frac{1}{u^{1-\frac{1}{\alpha}}} \left(1 - \exp\left(-\frac{c_p}{u}\right)\right) \mathrm{d}u \\
&\geqslant \frac{n^{\frac{1}{\alpha}}}{\alpha} \int_{0}^{p^\alpha/n} \frac{1}{u^{1-\frac{1}{\alpha}}} \left(1 - \exp\left(-\frac{1}{\zeta(\alpha) u}\right)\right) \mathrm{d}u.
\end{aligned}$$

We conclude by letting p go to infinity. ■

1.3.2 Exponential envelope classes

Theorem 17. *Let C and α denote positive real numbers satisfying $C > e^{2\alpha}$. Class $\Lambda_{Ce^{\alpha\cdot}}$ is the envelope associated with function $f_\alpha : x \mapsto Ce^{-\alpha x}$. Then*

$$\frac{1}{8\alpha}\log^2 n\,(1-o(1)) \leqslant \bar{R}(\Lambda^n_{Ce^{-\alpha\cdot}}) \leqslant R^*(\Lambda^n_{Ce^{-\alpha\cdot}}) \leqslant \frac{1}{2\alpha}\log^2 n + O(1)$$

Proof. For the upper-bound, note that as $u \to \infty$,

$$\bar{F}_\alpha(u) = \sum_{k>u} C\mathrm{e}^{-\alpha k} = \frac{C}{1-\mathrm{e}^{-\alpha}}\,\mathrm{e}^{-\alpha(u+1)}\,.$$

Hence, by choosing the optimal value $u_n = \frac{1}{\alpha}\log n$ in Proposition 12 we get :

$$R^*(\Lambda^n_{Ce^{-\alpha\cdot}}) \leqslant \frac{1}{2\alpha}\log^2 n + O(1).$$

Let us now turn to the lower bound. The proof we give here follows the maximin approach as it is presented in the introduction of this thesis : the minimax redundancy is the logarithm of the maximal number of distinguishable sources.

Let $p = \left\lfloor \frac{1}{2\alpha}\log\frac{nC}{4\log n}\right\rfloor$. For $1 \leqslant j \leqslant p$, let

$$u_j = \left\lfloor \sqrt{\frac{n\,(Ce^{-2\alpha j}\wedge 1))}{\log n}}\right\rfloor .$$

Observe that, for n large enough, $ne^{-2\alpha p} \geqslant \frac{4\log n}{C}$ and $u_p = 2$.

Let

$$\Theta = \left\{(\theta_1,\ldots,\theta_p) : \forall j, \theta_j \in \left\{\frac{1}{u_j},\frac{2}{u_j}\ldots,1\right\}\right\}.$$

1. **The cardinality of Θ :** $|\Theta| = \prod_{j=1}^p u_j$. Let $a = \lceil\frac{\log C}{2\alpha}\rceil$ be the first integer

j such that $Ce^{-2\alpha j} < 1$. Then

$$\begin{aligned}
\log|\Theta| &= \sum_{j=1}^{p} \log \left\lfloor \sqrt{\frac{n\left(Ce^{-2\alpha j} \wedge 1\right)}{\log n}} \right\rfloor \\
&= \sum_{j=1}^{p} \log \sqrt{\frac{n\left(Ce^{-2\alpha j} \wedge 1\right)}{\log n}} + O(\log n) \\
&= \frac{p}{2}\left(\log n - \log\log n\right) - \frac{1}{2}\sum_{j=a}^{p} \log\left(Ce^{-2\alpha j}\right) + O(\log n) \\
&= \frac{p}{2}\left(\log n - \log\log n\right) - \frac{(p-a+1)\log C}{2} \\
&\qquad -\frac{\alpha}{2}\left((p(p+1)) - a(a-1)\right) + O(\log n) \\
&= \frac{1}{8\alpha}\log^2 n + O(\log n).
\end{aligned}$$

2. **Definition of the source :** Let the alphabet be $\mathcal{A} = \{1, 2, \ldots, 2p\}$.
   Let $c_p = \sum_{j=1}^{p} e^{-2\alpha j} = e^{-2\alpha}\frac{1-e^{-2\alpha p}}{1-e^{-2\alpha}}$. Condition $C > e^{2\alpha}$ implies that $c_p > \frac{1}{C}$ for n large enough.
   For $\theta \in \Theta$ let p_θ be the distribution on $\mathcal{A}$ such that for $1 \leqslant j \leqslant p$:
   - $p_\theta(2j-1) = \theta_j \frac{e^{-2\alpha j}}{c_p}$;
   - $p_\theta(2j) = (1-\theta_j)\frac{e^{-2\alpha j}}{c_p}$.

   Note that $\forall i \in \mathcal{A}, p_\theta(i) \leqslant Ce^{-\alpha i}$, and hence $p_\theta \in \Lambda_f$.
   Let P_θ be the memoryless process with marginals p_θ. Let W be a random variable with uniform distribution on Θ. We conclude from the previous computation that

   $$H(W) = \log|\Theta| = \frac{1}{8\alpha}\log^2 n + O(\log n).$$

   We consider the memoryless random process $(X_n)_{n\geqslant 0}$ such that if $W = \theta$ then the distribution of X is P_θ.
3. **Estimation of θ :** For $i \in \mathcal{A}$, let $N_i = \sum_{k=1}^{n} \mathbb{1}_{X_k = i}$ be the number of occurrences of symbol i in X_1^n. For $1 \leqslant j \leqslant p$, let $Z_j = N_{2j-1} + N_{2j}$.
   We define the estimator $\hat{\theta} = \left(\hat{\theta}_j\right)_{1\leqslant j\leqslant p}$ of θ by :

   $$\hat{\theta}_j = \frac{N_{2j-1}}{Z_j}.$$

   Let j be a positive integer at most equal to p. Conditionally on W, Z_j has a binomial distribution $\mathcal{B}(n, e^{-2\alpha j}/c_p)$, whose variance is upper-bounded by $ne^{-2\alpha j}/c_p$. Hence, Bernstein's deviation inequality (see [MAS06] for a comprehensive book on deviation inequalities and their applications) implies that :

$$\begin{aligned}
P\left(Z_j < \frac{ne^{-2\alpha j}}{2c_p}\middle| W = \theta\right) &\leqslant e^{-\frac{\left(\frac{ne^{-2\alpha j}}{2c_p}\right)^2}{2\left(\frac{ne^{-2\alpha j}}{c_p}+\frac{ne^{-2\alpha j}}{6c_p}\right)}} \\
&\leqslant e^{-\frac{ne^{-2\alpha j}}{10c_p}} \\
&\leqslant e^{-\frac{ne^{-2\alpha p}}{10c_p}} \\
&\leqslant e^{-\frac{4\log n}{10Cc_p}} \leqslant n^{-\frac{1}{3Cc_p}}.
\end{aligned}$$

Now, conditionally on W and Z_j, variable N_{2j-1} has a binomial distribution $\mathcal{B}(Z_j, \theta_j)$. Hence by Höffdings inequality :

$$\begin{aligned}
P\left(\left|\frac{N_{2j-1}}{z} - \theta_j\right| > \frac{1}{2u_j}\middle| Z_j = z, W = \theta\right) &\leqslant 2e^{-\frac{2z}{4u_j^2}} \\
&\leqslant 2e^{-\frac{z\log n}{2n(Ce^{-2\alpha j}\wedge 1)}} \\
&\leqslant 2e^{-\frac{z\log n}{2nCe^{-2\alpha j}}}.
\end{aligned}$$

If follows that if $z \geqslant \frac{ne^{-2\alpha j}}{2c_p}$:

$$\begin{aligned}
P\left(\left|\frac{N_{2j-1}}{z} - \theta_j\right| > \frac{1}{2u_j}\middle| Z_j = z, W = \theta\right) &\leqslant 2e^{-\frac{\frac{ne^{-2\alpha j}}{2c_p}\log n}{2nCe^{-2\alpha j}}} \\
&\leqslant n^{-\frac{1}{4Cc_p}}
\end{aligned}$$

and hence

$$\begin{aligned}
P\left(\hat{\theta}_j \neq \theta_j \middle| W = \theta\right) &\leqslant P\left(Z_j < \frac{ne^{-2\alpha j}}{2c_p}\middle| W = \theta\right) \\
&\quad + \sum_{z=\left\lceil\frac{ne^{-2\alpha j}}{2c_p}\right\rceil}^{n} P\left(\left|\frac{N_{2j-1}}{z} - \theta_j\right| > \frac{1}{2u_j}\middle| \begin{array}{c} Z_j = z, \\ W = \theta \end{array}\right) \\
&\qquad \times P(Z_j = z | W = \theta) \\
&\leqslant n^{-\frac{1}{3Cc_p}} + n^{-\frac{1}{4Cc_p}}.
\end{aligned}$$

Thus, the probability of error in the estimation of θ by $\hat{\theta}$ conditionally on $W = \theta$ is upper-bounded as :

$$\begin{aligned}
P\left(\hat{\theta} \neq \theta \middle| W = \theta\right) &\leqslant \sum_{j=1}^{p} P\left(\hat{\theta}_j \neq \theta_j \middle| W = \theta\right) \\
&\leqslant p\left(n^{-\frac{1}{3Cc_p}} + n^{-\frac{1}{4Cc_p}}\right) = o(1)
\end{aligned}$$

as n goes to infinity, and

$$P\left(\hat{\theta} \neq \theta\right) = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} P\left(\hat{\theta} \neq \theta \middle| W = \theta\right) = o(1).$$

4. **Conclusion :**

$$\begin{aligned} H(W|X_1^n) &\leqslant H(W|\hat{\theta}) \\ &\leqslant P(W \neq \hat{\theta}) \log |\Theta| + \log 2 \\ &= o\left(\log^2 n\right). \end{aligned}$$

Hence,

$$\begin{aligned} \bar{R}(n, \Theta) &\geqslant I(X_1^n; W) \\ &= H(W) - H(W|X_1^n) \\ &\geqslant \frac{1}{8\alpha} \log^2 n \left(1 - o(1)\right). \end{aligned}$$

■

1.4 A censoring coding algorithm for infinite alphabets

In this section we describe an encoding algorithm that performs well on some envelope classes. The design of the algorithm builds on the following observation. The proof of Proposition 12 suggests to handle separately small (and hopefully frequent) symbols and large (hopefully rare) symbols. Given Theorem 15, such an algorithm should perform quite well as soon as the tail behavior of the envelope provides an adequate description of the sources in the class.

Algorithm `CensoringCode` follows this principle : it is parameterized by a cutoff value K and handles differently symbols that are smaller and larger than K. For the small symbols, the theoretical discussions used an NML-like inequality; for a practical algorithm it is natural to think at a universal coder like the Krichevsky-Trofimov mixture. For large symbols, we use the well-known efficient Elias code.

To discuss the efficiency of this algorithm, we focus on the example of the classes $\Lambda_{f_{\alpha,M}}$. But it is easy to extend the results presented here to other envelope classes.

Remark 12. *We anticipate a little on chapter 2 to remark that for* $1 < \alpha \leqslant 2$, pattern coding *leads to an almost minimax adaptive procedure. In fact, consider* *an algorithm successively coding the dictionary and the pattern of a message X_1^n. As the message and pattern have the same entropy rate, this coder suffers from two causes of redundancy : first the number of bits required to code the dictionary, which is of order $n^{1/\alpha}\log n$ (see Eq. (1.1)), and second the pattern redundancy, which is of order at most $O\left(\sqrt{n}\right)$ and hence insignificant.*

1.4.1 Algorithm for known parameters α, M

We present here a linear-time online algorithm almost achieving the minimax upper-bound (16) on envelope classes $\Lambda_{M\cdot^{-\alpha}}$. We assume that two subroutines `ArithCode` and `EliasCode` are available :

```
Algorithm 1 ArithCode(integer $j$ and integer array counts)
Ensure: arithmetic encoding of $j$ according to frequency table counts
  We assume this code uses exactly
      $-\log \frac{counts[j]}{\sum_i counts[i]}$ bits.
```

```
Algorithm 2 EliasCode
Require: integer $j$
Ensure: output Elias penultimate encoding for $j$ using
      $l(j) = \lfloor \log j + 2\log(1+\log j) + 1 \rfloor$ bits.
```

The idea of this algorithm is to compress string x_1^n into two code-strings : `C1`(x_1^n) contains the code for all "small" symbols (and escape characters to locate the others), while `C2`(x_1^n) contains the code of the "large" symbols. In processing the i-th symbol x_i, we use a *cutoff* K_i : if x_i is smaller than K_i then it is considered small, otherwise it is considered large and called *censored.* Considering the proof of the upper-bound in Theorem 16), we are led to choosing $K_i = \lambda i^{\frac{1}{\alpha}}$ (it will appear at the end that the best choice is $\lambda = \left(\frac{2M}{\alpha-1}\right)^{\frac{1}{\alpha}}$).

To describe `C1`(x_1^n) and `C2`(x_1^n) more in detail, let us define string $\tilde{x}_1^n$, by

$$\tilde{x}_i = \begin{cases} x_i & \text{if } x_i \leqslant K_i; \\ 0 & \text{else.} \end{cases}$$

- Code-string `C1`(x_1^n) contains an arithmetic code for the censored string $\tilde{x}_1^n$. The conditional probabilities used by subroutine `ArithCode` to code symbol

$\tilde{x}_i$ are those given by the Krichevsky-Trofimov mixture on alphabet $A_i = \{0, \ldots, K_i\}$. Note that we have to maintain special counters to take into account the fact that alphabet A_i is growing with i;

- Code-string $\texttt{C2}(x_1^n)$ contains the concatenation of the Elias codes of all censored symbols.

Algorithm 3 CensoringCode

```
$K \leftarrow 0$
$counts \leftarrow [1/2, 1/2, \ldots]$
for $i$ from 1 to $n$ do
  $cutoff \leftarrow \left\lfloor \left(2\frac{Mi}{\alpha-1}\right)^{1/\alpha} \right\rfloor$
  if $cutoff > K$ then
    for $j \leftarrow K+1$ to $cutoff$ do
      $counts[0] \leftarrow counts[0] - counts[j] + 1/2$
    end for
    $K \leftarrow cutoff$
  end if
  if $x[i] \leqslant cutoff$ then
    $\texttt{C1} \leftarrow \texttt{C1} \cdot \mathrm{ArithCode}(x[i], counts[0 : cutoff])$
  else
    $\texttt{C1} \leftarrow \texttt{C1} \cdot \mathrm{ArithCode}(0, counts[0 : cutoff])$
    $\texttt{C2} \leftarrow \texttt{C2} \cdot \mathrm{EliasCode}(x[i])$
    $counts[0] \leftarrow counts[0] + 1$
  end if
  $counts[x[i]] \leftarrow counts[x[i]] + 1$
end for
return $C_1 \cdot C_2$
```

Theorem 18. *The expected redundancy of procedure* CensoringCode *on the envelope class* $\Lambda_{M\cdot^{-\alpha}}$ *is not larger than*

$$\left(\frac{2Mn}{\alpha-1}\right)^{\frac{1}{\alpha}} \log n \, (1 + o(1)) \, .$$

Remark 13. *This redundancy is to be compared to the lower bound and to the theoretical (non-algorithmic) upper-bound of Theorem 16. These three quantities only differ by a factor of order at most* $\log n$.

To prove this theorem, we need a few definitions and two lemmas. Remember that $K_i = \lambda i^{\frac{1}{\alpha}}$, and let y_1^n be the string of length n defined by :

$$y_i = \begin{cases} x_i & \text{if } x_i \leqslant K_n; \\ 0 & \text{else.} \end{cases}$$

Note that as $(K_i)_{1\leqslant i\leqslant n}$ is an increasing sequence, y_1^n may contain a few less symbols '0' than $\tilde{x}_1^n$. For $0 \leqslant j \leqslant K_n$, let $n_j\left(y_1^i\right)$ be the number of occurrences of symbol j in string y_1^i, and $n_j = n_j\left(y_1^n\right)$. Let also κ be the Krichevsky-Trofimov mixture on alphabet $A_n = \{0, \dots, K_n\}$. Hence,

$$\kappa(y_1^n) = \frac{\prod_{j=0}^{K_n}\left(\frac{1}{2}\right)\left(1+\frac{1}{2}\right)\dots\left(n_j-2+\frac{1}{2}\right)\left(n_j-1+\frac{1}{2}\right)}{\left(\frac{1+K_n}{2}\right)\left(1+\frac{1+K_n}{2}\right)\dots\left(n-2+\frac{1+K_n}{2}\right)\left(n-1-\frac{1+K_n}{2}\right)}.$$

Lemma 1. *For every string $x_1^n \in \mathbb{N}_+^n$, the length of the first code part $C1(x_1^n)$ is not larger than $-\log\kappa(y_1^n)$.*

Proof of Lemma 1. Let $c_i[j]$ be the value of variable counts$[j]$ at the beginning of the i-th loop in function CensoringCode : for all $j \geqslant 1$, $c_i[j] = n_i(y_1^{j-1})+\frac{1}{2}$. For $1 \leqslant j \leqslant K_n$ let s_j be the number of occurrences of symbol j have been censored during the execution. If $i(j)$ is the smallest integer such that $K_{i(j)}$ is larger than j, then $s_j = n_j\left(y_1^{i(j)}\right)$.

Let $s_0 = |\{i : y_i = 0\}|$, the number of symbols in x_1^n that are larger than K_n. Note that the value of counts$[0]$ at the end of the execution is $s_0 + 1/2$. Let $T_0 = \prod_{i=1,\tilde{x}[i]=0}^{n} c_i[0]$ be the numerator of the probability assigned to symbol 0 during the computation. Note that at the end of each loop i, it holds that $c_i[0] - \frac{1}{2} = \sum_{j>K_i} N_i(j)$. It follows that :

$$\begin{aligned} T_0 &= \prod_{j=1}^{\infty}\prod_{i:x_i=j>K_i} c_i[0] && (1.2)\\ &= \prod_{i:y_i=0} c_i[0]\prod_{j=1}^{K_n}\prod_{i:y_i=j>K_i} c_i[0] && (1.3)\\ &\geqslant \left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\cdots\left(s_0-1+\frac{1}{2}\right)\prod_{j=1}^{K_n}\left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\cdots\left(s_j-1+\frac{1}{2}\right). && (1.4)\end{aligned}$$

Equality 1.2 is obtained by reordering the factors in T_0, in Equality 1.3 we put together all symbols larger than K_n (they remain censored during the whole computation) and Inequality 1.4 holds since $|\{i : y_i = j > K_i\}| = s_j$.

Moreover, it holds that $\sum_{0\leqslant j\leqslant K_i} c_i[j] = i-1+\frac{K_i+1}{2}$. Hence, it appears that

$$\begin{aligned}
\left|\mathtt{C1}\left(x_1^n\right)\right| &= \sum_{i=1}^{n} -\log\frac{c_i\left[\tilde{x}[i]\right]}{\sum_{0\leqslant j\leqslant K_i} c_i[j]} \\
&= -\log\frac{T_0\prod_{j=1}^{K_n}\left(s_j+\frac{1}{2}\right)\left(s_j+1+\frac{1}{2}\right)\dots\left(n_j-1+\frac{1}{2}\right)}{\left(\frac{K_1+1}{2}\right)\left(1+\frac{K_2+1}{2}\right)\dots\left(n-1+\frac{K_n+1}{2}\right)} \\
&\leqslant -\log\frac{T_0\prod_{j=1}^{K_n}\left(s_j+\frac{1}{2}\right)\left(s_j+1+\frac{1}{2}\right)\dots\left(n_j-1+\frac{1}{2}\right)}{\left(\frac{K_n+1}{2}\right)\left(1+\frac{K_n+1}{2}\right)\dots\left(n-1+\frac{K_n+1}{2}\right)} \\
&\leqslant -\log\frac{\left(\frac{1}{2}\right)\left(1+\frac{1}{2}\right)\dots\left(s_0-1+\frac{1}{2}\right)\prod_{j=1}^{K_n}\left(\frac{1}{2}\right)\left(1+\frac{1}{2}\right)\dots\left(n_j-1+\frac{1}{2}\right)}{\left(\frac{K_n+1}{2}\right)\left(1+\frac{K_n+1}{2}\right)\dots\left(n-1+\frac{K_n+1}{2}\right)} \\
&= -\log\kappa(y_1^n)
\end{aligned}$$

The first inequality holds since $\left(K_i\right)_i$ is a non-decreasing sequence and the second inequality is a consequence of Inequality (1.4).

Lemme 2. *For every source $P\in\Lambda_{M\cdot^{-\alpha}}$, the expected length of the second code part satisfies :*

$$\mathbb{E}_P\left[\,\left|\mathit{C2}\left(X_1^n\right)\right|\,\right] \leqslant \frac{M}{(\alpha-1)\,\lambda^{\alpha-1}} n^{\frac{1}{\alpha}}\log n\left(1+o(1)\right).$$

Proof of Lemma 2.

Let $1\leqslant a<b$ and $\beta>0$. First note that :

$$\begin{aligned}
\int_a^b \frac{1}{x^\beta}\mathrm{d}x &= \left[\frac{1}{(1-\beta)\,x^{\beta-1}}\right]_a^b, && (1.5)\\
\int_a^b \frac{\log x}{x^\beta}\mathrm{d}x &= \left[\frac{\log x}{(1-\beta)\,x^{\beta-1}}\right]_a^b - \frac{1}{1-\beta}\int_a^b\frac{\mathrm{d}x}{x^\beta} \\
&= \left[\frac{\log x-\frac{1}{1-\beta}}{(1-\beta)\,x^{\beta-1}}\right]_a^b, && (1.6)\\
\int_a^b \frac{\log\left(1+\log x\right)}{x^\beta}\mathrm{d}x &= \left[\frac{\log\left(1+\log x\right)}{(1-\beta)\,x^{\beta-1}}\right]_a^b - \frac{1}{1-\beta}\int_a^b\frac{\mathrm{d}x}{x^\beta\left(1+\log x\right)} \\
&\leqslant \left[\frac{\log\left(1+\log x\right)-\frac{1}{1-\beta}}{(1-\beta)\,x^{\beta-1}}\right]_a^b. && (1.7)
\end{aligned}$$

The expected length of the second part of the code is :

$$\begin{aligned}
\mathbb{E}[|\mathrm{C2}(X_1^n)|] &= \mathbb{E}\left[\sum_{j=1}^{n} l(x_j)\mathbb{1}_{x_j>K_j}\right] \\
&= \sum_{j=1}^{n}\sum_{x=K_j+1}^{\infty} l(x)P(x_j) \\
&\leqslant \sum_{j=1}^{n}\sum_{x=K_j+1}^{\infty} l(x)\frac{M}{x^\alpha} \\
&\leqslant M\sum_{j=1}^{n}\sum_{x=K_j+1}^{\infty} \frac{\log(x)+2\log(1+\log x)+1}{x^\alpha}.
\end{aligned}$$

In the last inequality, we use the fact that $l(x) \leqslant \log x + 2\log(1+\log x)+1$. Using (1.5), (1.6) and (1.7) we get :

$$\begin{aligned}
\sum_{x=K_j+1}^{\infty} \frac{\log x+2\log(1+\log x)+1}{x^\alpha} &\leqslant \int_{K_j}^{\infty} \frac{\log x+2\log(1+\log x)+1}{x^\alpha}\mathrm{d}x \\
&\leqslant \frac{\log K_j + 2\log(1+\log K_j)+\frac{4}{\alpha-1}}{(\alpha-1)K_j^{\alpha-1}}.
\end{aligned}$$

Thus, as we choose $K_j = \lambda j^{1/\alpha}$, we substitute β by $1-\frac{1}{\alpha}$ in Equations (1.5), (1.6) and (1.7) to obtain :

$$\begin{aligned}
\mathbb{E}[|\mathrm{C2}(X_1^n)|] &\leqslant \frac{M}{\alpha-1}\sum_{j=1}^{n}\frac{\log K_j+2\log(1+\log K_j)+\frac{4}{\alpha-1}}{K_j^{\alpha-1}} \\
&= \frac{M}{\alpha-1}\sum_{j=1}^{n}\frac{\frac{1}{\alpha}\log j+\log\lambda+2\log\left(1+\log\left(\lambda j^{1/\alpha}\right)\right)+\frac{4}{\alpha-1}}{\lambda^{\alpha-1}j^{1-\frac{1}{\alpha}}} \\
&\leqslant \frac{M}{\alpha(\alpha-1)\lambda^{\alpha-1}}\int_{x=1}^{n+1}\frac{\left(\log x+\alpha\log\lambda+2\alpha\log\left(1+\log\left(\lambda x^{1/\alpha}\right)\right)+\frac{4\alpha}{\alpha-1}\right)}{x^{1-\frac{1}{\alpha}}}\mathrm{d}x \\
&= \frac{M}{(\alpha-1)\lambda^{\alpha-1}}n^{\frac{1}{\alpha}}\log n\,(1+o(1)).
\end{aligned}$$

Proof of Th. 18 Remember that $A_n = \{0,\dots,K_n\}$. Let us first note that for every string $x_1^n \in \mathbb{N}_+^n$,

$$\max_{p\in\mathfrak{M}_1(A_n)} p^{\otimes n}(y_1^n) \geqslant \max_{p\in\mathfrak{M}_1(\mathbb{N}_+)} p^{\otimes n}(x_1^n) \geqslant \max_{P\in\Lambda_{M\cdot-\alpha}} P^n(x_1^n).$$

Together with Lemma 1 and the redundancy of the Krichevsky-Trofimov mixture (see [KT81]), this implies that :

$$\begin{aligned}
\left|\mathtt{C1}\left(X_1^n\right)\right| &\leqslant -\log \kappa\left(y_1^n\right) \\
&\leqslant \inf_{p \in \mathfrak{M}_1(\{0,\dots,K_n\})} -\log p^{\otimes n}\left(y_1^n\right) + \frac{K_n}{2}\log n + O(1) \\
&\leqslant -\log f_{\Lambda^n_{M\cdot^{-\alpha}}}\left(x_1^n\right) + \frac{\lambda n^{\frac{1}{\alpha}}}{2}\log n + O(1).
\end{aligned}$$

Let $L(x)$ be the length of the code produced by algorithm `CensoringCode` on the input string x, and Q^n the corresponding coding distribution, that is $Q^n(\cdot) = 2^{-L(\cdot)}$. Then

$$\begin{aligned}
\bar{R}(Q^n, \Lambda^n_{M\cdot^{-\alpha}}) &= \sup_{P \in \Lambda_{M\cdot^{-\alpha}}} \mathbb{E}_P\left[L(X_1^n) + \log P^n(X_1^n)\right] \\
&\leqslant \sup_{P \in \Lambda_{M\cdot^{-\alpha}}} \mathbb{E}_P\left[L(X_1^n) + \log f_{\Lambda^n_{M\cdot^{-\alpha}}}(X_1^n)\right] \\
&\leqslant \sup_{P \in \Lambda^n_{M\cdot^{-\alpha}}} \mathbb{E}_P\left[\left|\mathtt{C1}\left(X_1^n\right)\right| + \log f_{\Lambda^n_{M\cdot^{-\alpha}}}\left(X_1^n\right) + \left|\mathtt{C2}\left(X_1^n\right)\right|\right] \\
&\leqslant \frac{\lambda n^{\frac{1}{\alpha}}}{2}\log n + \frac{M}{(\alpha-1)\,\lambda^{\alpha-1}} n^{\frac{1}{\alpha}}\log n\,(1+o(1))\,.
\end{aligned}$$

The optimal value is $\lambda = \left(\frac{2M}{\alpha-1}\right)^{\frac{1}{\alpha}}$, for which we get :

$$\bar{R}(Q^n, \Lambda^n_{M\cdot^{-\alpha}}) \leqslant \left(\frac{2Mn}{\alpha-1}\right)^{\frac{1}{\alpha}} \log n\,(1+o(1))\,.$$

1.4.2 An adaptive version

In order to build adaptive versions of algorithm `CensoringCode`, one possibility is to estimate the cutoff K_{n+1} using the symbols x_1^n seen so far. We tried to do this using the number C_n of *disctinct* symbols, hoping (see Eq. (1.1)) that for some constant C and some positive α :

$$E(C_n) \sim C n^{1/\alpha}. \tag{1.8}$$

This is a correct estimation if $P^1(k)$ is really of order $\frac{1}{k^\alpha}$. Unfortunately, sparse distributions may lead this project into troubles. If, for example, $\left(Y_n\right)_n$ is a sequence of geometrically distributed random variables, and if $X_n = \left\lfloor 2^{\frac{Y_n}{\alpha}} \right\rfloor$, then the distribution of the X_n just fits in $\Lambda_{M\cdot^{-\alpha}}$ but obviously $C_n(X_1^n) = C_n\left(Y_1^n\right) = O\left(\log n\right)$.

The results of this section are hence expressed for some classes $\mathcal{W}_\alpha$, included in $\Lambda_{\cdot^{-\alpha}}$, such that every source P in $\mathcal{W}_\alpha$ satifies (1.8). We shall try in the future (see

[BGG06]) to specify class $\mathcal{W}_\alpha$ more precisely. In particular, we are considering conditions like $0 < \liminf k^\alpha P^1(k) \leqslant \limsup k^\alpha P^1(k) < \infty$ or $P(k) = \frac{L(k)}{k^\alpha}$ where $L(\cdot)$ is a slowly varying function, and we try to use other kinds of estimators.

1.4.2.1 A concentration inequality for C_n

If we want to use the number C_n of disctinct symbols in X_1^n to estimate cutoff K_n, we need a concentration inequality for C_n. To prove it, we refer to [DR98]. A family $(X_k)_{1 \leqslant k \leqslant n}$ is said to be *negatively associated* if for every two non-overlapping index sets $I, J \subset \{1, \ldots, n\}$ and for all functions $f : \mathcal{R}^{|I|} \to \mathcal{R}$ and $g : \mathcal{R}^{|J|} \to \mathcal{R}$ both non-decreasing or non-decreasing :

$$\mathbb{E}\left(f\left(X_i, i \in I\right) g\left(X_j, j \in J\right)\right) \leqslant \mathbb{E}\left(f\left(X_i, i \in I\right)\right) \mathbb{E}\left(g\left(X_j, j \in J\right)\right).$$

Let $B_{j,i}$ be the random variable equal to 1 if $X_i = j$ and equal to 0 otherwise. Let also $O_j = \mathbb{1}_{\sum_i B_{i,j} \geqslant 1}$ equal to 1 if and only if at least one of the $(X_i)_{1 \leqslant i \leqslant n}$ is equal to symbol j. The distribution of O_j is the Bernoulli $\mathcal{B}\left(1 - (1-p_j)^n\right)$ and $Var[O_j] \leqslant \mathbb{E}[O_j] = 1 - (1-p_j)^n$.

Note that the $(O_j)_j$ are not independent, but it is proved in [DR98] that the family $(B_{j,i})_{1 \leqslant i \leqslant n, j}$ is negatively associated. Thus, by applying Proposition 8.2 in [DR98] to the non-overlapping sets $I_j = \{(j,i); 1 \leqslant i \leqslant n\}$ and to the non-decreasing function $h(b_1, b_2, \ldots) = \mathbb{1}_{\sum_k b_k \geqslant 1}$, it follows that family $(O_j)_j$ is also negatively associated. Now,

$$\log \mathbb{E}\left[e^{\lambda C_n}\right] = \log \mathbb{E}\left[e^{\lambda \sum_{j=1}^\infty O_j}\right] = \log \mathbb{E}\left[\prod_{j=1}^{\infty} e^{\lambda O_j}\right] \leqslant \sum_{j=1}^{\infty} \log \mathbb{E}\left[e^{\lambda O_j}\right].$$

The last inequality uses the negative association of the $(O_j)_j$ with the increasing function $x \to e^{\lambda x}$. The infinite summations cause no difficulty as $\sum_{j=1}^\infty O_j$ is always upper-bounded by n. Hence, the classical proof for Bernstein's inequality (see [BBLM05], Chapter 1) applies and :

$$P\left(|C_n - \mathbb{E}[C_n]| > t\right) \leqslant 2e^{-\frac{t^2}{2\left(\sum_j\left[1-(1-p_j)^n\right]+\frac{t}{3}\right)}} = 2e^{-\frac{t^2}{2\left(\mathbb{E}[C_n]+\frac{t}{3}\right)}}.$$

As a consequence

$$P\left(C_n \leqslant \frac{1}{2}\mathbb{E}[C_n]\right) \quad \leqslant \quad 2e^{-\frac{3}{8}\mathbb{E}[C_n]}.$$

Noting that $C_n \geqslant 1$, we can derive the following inequality that will prove useful later on :

$$\begin{aligned}
\mathbb{E}_P\left[\frac{1}{C_n^{\alpha-1}}\right] &= \mathbb{E}_P\left[\frac{1}{C_n^{\alpha-1}}\mathbb{1}_{C_n>\frac{1}{2}\mathbb{E}[C_n]}\right]+\mathbb{E}_P\left[\frac{1}{C_n^{\alpha-1}}\mathbb{1}_{C_n\leqslant\frac{1}{2}\mathbb{E}[C_n]}\right] && (1.9)\\
&\leqslant \frac{1}{\left(\frac{1}{2}\mathbb{E}[C_n]\right)^{\alpha-1}}+P\left(C_n\leqslant\frac{1}{2}\mathbb{E}[C_n]\right). && (1.10)
\end{aligned}$$

Note that no assumption on $\mathbb{E}\left[C_n\right]$ is necessary to derive these results : whether P is in $\mathcal{W}_\alpha$ or not, they remain true.

1.4.2.2 Adaptive choice of the cutoff

We consider here a modified version of `CensoringCode` that operates similarly, except that

- the string x_1^n is first scanned completely to determine $C_n\left(x_1^n\right)$ (this prevents the algorithm from being sequential) ;
- the constant cutoff
$$\hat{K}_n=\mu C_n$$
is used for all symbols x_i, $1\leqslant i\leqslant n$, where μ is some positive constant.
- the value of K_n is transmitted before `C1` and `C2`, using the Elias code.

Note that we consider here a non-sequential algorithm in order to clarify the proof of its adaptivity, focusing on the estimation of the cutoff. In fact, let us still denote by `C1`(x) and `C2`(x) the two parts of the code-string. Let $\hat{L}$ be the code length corresponding to this new algorithm, and $\hat{Q}^n$ the corresponding distribution on $\mathbb{N}_+^n$. For any distribution P :

$$\begin{aligned}
\overline{R}(\hat{Q}^n,P) &= \mathbb{E}_P\Bigg[\ell(\hat{K}_n)+|\mathtt{C1}\,(X_1^n)|-\sum_{k=1}^{\hat{K}_n}P^1(k)\log\frac{1}{P^1}(k)\\
&\qquad +|\mathtt{C2}\,(X_1^n)|-\sum_{k>\hat{K}_n}P^1(k)\log\frac{1}{P^1}(k)\Bigg]\\
&\leqslant \mathbb{E}_P\Bigg[\ell(\hat{K}_n)+\frac{\hat{K}_n}{2}\log\frac{n}{\hat{K}_n+1}+\frac{\hat{K}_n+1}{2}\log e+C\\
&\qquad +n\sum_{k\geqslant\hat{K}_n+1}P^1(k)\left(\log\frac{P^1(k)}{Q^n(0)}+\ell(k)\right)\Bigg]\\
&\leqslant \frac{E_P(\hat{K}_n)}{2}(\log ne)+C'+E_P(\ell(\hat{K}_n))+nE_P\left(\sum_{k\geqslant\hat{K}_n+1}P^1(k)\ell(k)\right).
\end{aligned}$$

As function ℓ is increasing and equivalent to log at infinity, we obtain that for some positive η :

$$\begin{aligned}\overline{R}(\hat{Q}^n, P) &\leqslant \frac{E_P(\hat{K}_n)}{2}(\log n + \eta) + nME_P\left(\frac{1}{\hat{K}_n^{\alpha-1}}\int_1^\infty \frac{\ell(u\hat{K}_n+1)}{u^\alpha}du\right)\\ &\leqslant \frac{E_P(\hat{K}_n)}{2}(\log n + \eta) + nME_P\left(\frac{1}{\hat{K}_n^{\alpha-1}}\right)(\log n + \eta)\int_1^\infty \frac{\ell(u+1)}{u^\alpha}du\\ &= O\left(n^{\frac{1}{\alpha}}\log n\right) \quad \text{if } P \in \mathcal{W}_\alpha,\end{aligned}$$

by Inequalities (1.8) and (1.10). We can conclude :

Theorem 19. *The modified version of algorithm* ***CensoringCode*** *presented above is almost adaptive on the class $\bigcup_{\alpha>0}\mathcal{W}_\alpha$ in so far that for every $P \in \mathcal{W}_\alpha$, the redundancy is $O\left(n^{\frac{1}{\alpha}}\log n\right)$ although the algorithm has no information on α.*

Remark 14. *In order to be adaptive on class $\bigcup_{\alpha>0}\mathcal{W}_\alpha$, we strongly suspect that there is no need in the modified* ***CensoringCode*** *algorithm to scan the entire string x_1^n first, and to transmit $\hat{K}_n$: we believe that a sequential estimation of the cutoff is also satisfying. But before proving such a result, we need to specify more precisely the classes $\mathcal{W}_\alpha$, which is what we are working on presently.*

Chapitre 2

A lower bound for the maximin redundancy in pattern coding

2.1 Introduction

2.1.1 Universal coding

Let P be a stationary *source* on an alphabet A, both known by the coder and the decoder. Let $X = (X_n)_{n\in\mathbb{N}}$ be a random process with distribution P. For a positive integer n, we denote by X_1^n the vector of the n first components of X and by P^n the distribution of X_1^n on A^n. We denote the logarithm with base 2 by log and the natural logarithm by ln. Shannon's classical bound [Sha48] states that no coding function on that source can use (in average) fewer bits than the *n-th order entropy* $H(X_1^n) = \mathbb{E}\left[-\log P_n\left(X_1^n\right)\right]$; moreover, this code length can be nearly approached, see [CT91]. One important idea in the proof of this result is the following : every code on the strings of length n is associated with a *coding distribution* q_n on A^n in such a way that the code length for x is $-\log q_n(x)$, *and reciprocally* any distribution q_n on A^n can be associated with a coding function whose code length is approximately $-\log q_n(x)$. When P is ergodic, its *entropy rate* $H(X) = \lim_{n\to\infty} \frac{1}{n}H(X_1^n)$ exists. It is a lower bound on the number of bits required per character, and it can be reached up to a factor $1/n$.

If P is only known to be an element P_θ of some class $\mathcal{C} = \{P_\theta : \theta \in \Theta\}$, *universal coding* consists in finding a single code, or equivalently a single sequence of coding distributions $(q_n)_n$, approaching the entropy rate for all sources $P_\theta \in \mathcal{C}$ at the same time. Such versatility has a price : for any given source P_θ, there is an additional cost called *(expected) redundancy* $R(q_n, \theta)$ of the coding distribution q_n that is defined as the difference between the expected code length $\mathbb{E}_\theta\left[-\log q_n(X_1^n)\right]$ and the n-th order entropy $H(X_1^n)$. Two criteria measure the *universality* of q_n :

- First, a deterministic approach judges the performance of q_n in the worst case by the *maximal redundancy* $R^+(q_n, \Theta) = \sup_{\theta \in \Theta} R(q_n, \theta)$. The lowest achievable maximal redundancy is called *minimax redundancy* :

$$R^+(n, \Theta) = \min_{q_n} \max_{\theta} R(q_n, \theta).$$

- Second, a Bayesian approach consists in providing Θ with a prior distribution π, and then considering the expected redundancy $\mathbb{E}_\pi[R(q_n, \theta)]$ (the expectation is here taken over θ). Let q_n^π be the coding distribution minimizing $\mathbb{E}_\pi[R(q_n, \theta)]$. The *maximin redundancy* $R^-(n, \Theta)$ of class $\mathcal{C}$ is the supremum of all $\mathbb{E}_\pi[R(q_n^\pi, \theta)]$ over all possible prior distributions π :

$$R^-(n, \Theta) = \max_{\pi} \min_{q_n} \mathbb{E}_\pi[R(q_n, \theta)].$$

It is obvious that $R^-(n, \Theta)$ is always at most equal to $R^+(n, \Theta)$, and a result by Haussler [Hau97] states that mild hypotheses are sufficient to ensure that $R^-(n, \Theta) = R^+(n, \Theta)$. Class $\mathcal{C}$ is said to be *strongly universal* if $R^+(n, \Theta) = o(n)$: then universal coding is possible uniformly on $\mathcal{C}$. An important theorem by Rissanen [Ris84] asserts that if the parameter set Θ is k-dimensional, the existence of a $\sqrt{n}$–consistent estimator for θ implies that

$$R^-(n, \Theta) = R^+(n, \Theta) = \frac{k}{2} \log n + O(1). \tag{2.1}$$

This well-known bound has many applications in Information Theory, often related to the Minimum Description Length Principle. It is remembered as a "thumb rule" that redundancy is $1/2 \log n$ for each parameter of the model. This result actually covers a large variety of cases, among others : memoryless processes, Markov chains, Context tree sources, hidden Markov chains. However, further generalization have been investigated. Shields (see [Shi93]) proved that no coder can achieve a non-trivial redundancy rate on all stationary ergodic processes. Csiszár and Shields [CS96] gave an example of non-parametric, intermediate complexity class (*renewal processes*) for which $R^-(n, \Theta)$ and $R^+(n, \Theta)$ are both of order $O(\sqrt{n})$. If alphabet A is not known, or if its size k is not insignificant compared to n, then equation (2.1) is useless. If the alphabet A is infinite, Kieffer [Kie78] showed that no universal coding is possible even for the class of memoryless processes.

2.1.2 Dictionary and Pattern

To address those problems, the idea appeared of coding separately the *structure* of string x and the characters present in x. It was first introduced by Äberg in

[ÄSMS97] as a solution to the *multi-alphabet* coding problem, where the message x contains only a small subset of the known alphabet A. It was further studied and motivated in a series of articles by Shamir [Sha03a, Sha04, Sha03b, Sha06] and by Jevtić, Orlitsky, Santhanam and Zhang [OS04, JOS05, OSZ04, OSVZ04] for practical applications : the alphabet is unknown and has to be transmitted separately anyway (for instance, transmission of a text in an unknown language), or the alphabet is very large in comparison to the message (consider the case of images with $k = 2^{24}$ colors, or texts when taking words as the alphabet units).

To explain the notion of pattern, we shall use for example the string

$$x = \text{“abracadabra”}$$

made of $n = 11$ characters. The information it conveys can be separated in two blocks :

- a *dictionary* $\Delta = \Delta(x)$ defined as the sequence of different characters present in x in order of appearance; in the example $\Delta = (a, b, r, c, d)$.
- a *pattern* $\psi = \psi(x)$ defined as the sequence of positive integers pointing to the *indices* of each letter in Δ; here, $\psi = 1231415123$.

Let $\mathcal{P}^n$ be the set of all possible patterns of n-strings. For instance, $\mathcal{P}^1 = \{1\}$, $\mathcal{P}^2 = \{11, 12\}$, $\mathcal{P}^3 = \{111, 112, 121, 122, 123\}$. Taking the same notations as in [OSZ04], we call multiplicity $\mu_j(\psi)$ of symbol j in pattern $\psi \in \mathcal{P}^n$ the number of occurrences of j in ψ; the *multiplicity of pattern* ψ is the vector made of all symbol's multiplicities : $\mu(\psi) = (\mu_j(\psi))\, 1 \leqslant j \leqslant n$ - in the former example, $\mu = (5, 2, 2, 1, 1, 0, \ldots)$. Note that $\sum_{j=1}^{n} \mu_j = n$. Moreover, the *profile* $\phi = (\phi_\mu)_{\mu \geqslant 1}$ of pattern ψ provides, for every multiplicity μ, its frequence in $\mu(\psi)$. It can be formally defined as the multiplicity of ψ's multiplicity : $\mu(\mu(\psi))$. The profile of string “abracadabra” is $(2, 2, 0, 0, 1, 0, \ldots)$ as two symbols (c and d) appear once, two symbols (b and r) appear twice and one symbol (a) appears five times. We denote by Φ^n the set of possible profiles for patterns of length n, so that $\Phi^1 = \{(1)\}$, $\Phi^2 = \{(2, 0), (0, 1)\}$, $\Phi^3 = \{(3, 0, 0), (1, 1, 0), (0, 0, 1)\}$. Note that $\sum_{\mu=1}^{n} \mu\phi_\mu = n$. As explained in [OSZ04], there is one-to-one mapping between Φ^n and the set of *unordered partitions* of integer n. This point will be used and specified in Section 2.3.

2.1.3 Pattern coding

Any process X from a source P_θ induces a pattern process $\Psi = (\Psi_n)_{n \in \mathbb{N}}$ with marginal distributions on $\mathcal{P}^n$ defined by $P_\theta(\Psi_1^n = \psi) = \sum_{\psi(x)=\psi} P_\theta(X_1^n = x)$. Thus, we can define a *n-th block pattern entropy* $H(\Psi_1^n) = \mathbb{E}_\theta[-\log P_\theta(\Psi_1^n)]$. For stationary ergodic P_θ, Orlitsky & al. [OSVZ04] prove that the *pattern entropy rate* $H(\Psi) = \lim_{n\to\infty} \frac{1}{n} H(\Psi_1^n)$ exists and is equal to $H(X)$ (whether this quantity is finite or not). This result was independently discovered by Gemelos and Weissman [GW04].

In the following, we shall consider only the case of *memoryless* sources P_θ, with marginal distributions p_θ on a (possibly infinite) alphabet $\mathcal{A}$. Hence, Θ will be the set parameterizing all probability distributions on $\mathcal{A}$.

Of course, the process they induce on $(\mathcal{P}^n)_{n\in\mathbb{N}}$ is not memoryless. But as patterns convey less information than the initial strings, coding them seems to be an easier task. The *expected pattern redundancy* of a coding distribution q_n on $\mathcal{P}^n$ can be defined by analogy as the difference between the expected code length under distribution P_θ and the n-th block pattern entropy :

$$\begin{aligned} R_\Psi(q_n,\theta) &= \mathbb{E}_\theta\left[-\log q_n(\Psi_1^n)\right] - H(\Psi_1^n) \\ &= \sum_{\psi\in\Psi^n} P_\theta(\psi)\log\frac{P_\theta(\psi)}{q_n(\psi)}. \end{aligned}$$

As the alphabet is unknown, the *maximal pattern redundancy* $R_\Psi^+(q_n,\Theta)$ must be defined as the maximum of $R_\Psi^+(q_n,\theta)$ over *all alphabets A and all memoryless distributions on A*. Of course, the *minimax pattern redundancy* $R_\Psi^+(n,\Theta)$ is defined as the lower-bound of $R_\Psi^+(q_n,\Theta)$ in q_n. Similarly, the *worst expected pattern redundancy* $R_\Psi^-(q_n,\Theta)$ is the supremum of all $\mathbb{E}_\pi[R_\Psi(q_n,\theta)]$ over all possible alphabets A and all prior distributions π on the set of memoryless distributions on A. The *maximin pattern redundancy* $R_\Psi^-(n,\Theta)$ is defined as the lower-bound of $R_\Psi^-(q_n,\Theta)$ in q_n.

2.2 Theorem

The true order of magnitude of $R_\Psi^-(q_n,\Theta)$ and $R_\Psi^+(n,\Theta)$ is not known yet. However, Orlistky & al in [OSZ04] and Shamir in [Sha03b] proved that $\omega(n^{1/3-\epsilon}) \leqslant R_\Psi^-(q_n,\Theta) \leqslant R_\Psi^+(n,\Theta) \leqslant O(\sqrt{n})$. There is hence a large gap between upper- and lower-bounds. This gap has been reduced in an article by Shamir [Sha04] where the upper-bound is improved to $O\left(n^{2/5}\right)$. The following theorem contributes to the evaluation of $R_\Psi^-(q_n,\Theta)$, by providing a better lower-bound.

Theorem 20. *For all integers n large enough :*

$$R_\Psi^-(n,\Theta) \geqslant 1.84\left(\frac{n}{\log n}\right)^{1/3}.$$

To situate the interest of Theorem 20, let us present precisely the results that have been proved on that topic. In [OSZ04], Orlitsky & al. study a slightly different notion of redundancy. The *regret* of a coding distribution q_n on pattern

$\psi \in \mathcal{P}^n$ is the difference between the code length $-\log q_n(\psi)$ and the best achievable code length $\inf_{\theta\in\Theta} -\log P_\theta(\psi)$. The *individual redundancy* is the maximal regret obtained by q_n on a pattern :

$$R_\Psi^*(q_n, \Theta) = \max_{\psi\in\mathcal{P}^n} \left[-\log q_n(\psi) - \inf_{\theta\in\Theta} -\log P_\theta(\psi) \right].$$

The minimizer of the individual redundancy is known to be Shtarkov's *Normalized Maximum Likelihood* [Sht87] defined by :

$$\mathrm{NML}(\psi) = \frac{\max_{\theta\in\Theta} P_\theta(\psi)}{\sum_{\psi\in\mathcal{P}^n} \max_{\theta\in\Theta} P_\theta(\psi)}.$$

It achieves a constant regret on all patterns (the *minimax individual redundancy*) equal to the logarithm of the normalizing factor :

$$R_\Psi^*(n, \Theta) = \log \sum_{\psi\in\mathcal{P}^n} \max_{\theta\in\Theta} P_\theta(\psi).$$

Obviously, $R_\Psi^*(n, \Theta) \geqslant R_\Psi^+(n, \Theta)$, and for all classes studied so far, minimax individual and minimax average redundancies have similar asymptotic behaviors. The authors prove that

$$\frac{3\log e}{2} n^{1/3} (1 + o(1)) \leqslant R^*(n, \Psi) \leqslant \pi\sqrt{\frac{2}{3}} \log e \sqrt{n}.$$

In [Sha03b], it is proven that for all positive ϵ and n large enough, $R_\Psi^-(n, \Theta) \geqslant \frac{3\log e}{4\pi^{1/3}} n^{1/3-\epsilon} (1 + o(1))$. When I discussed about Theorem 20 with Gil Shamir, he suggested to me that a look into his proof shows the possibility to strengthen this result by letting ϵ depend on n. Actually, we can choose $\epsilon = \frac{2\log\log n + 3\log\frac{1601}{3\log e}}{\log n}$, since the quantity $P_\theta(A)$ referred to in Lemma 2 on p.22 and studied in Appendix B on p.50 of that paper still goes to zero with this choice. This leads to the existence of a positive constant $\eta \approx 5 \times 10^{-8}$ such that

$$R_\Psi^-(n, \Theta) \geqslant \eta \left(\frac{n}{\log^2 n} \right)^{1/3}$$

as soon as n is large enough. This proof was elaborated independently from Shamir's lower bounds ; both of them use the channel capacity inequality. However, it is interesting to note that they rely on different ideas (unordered partitions and Bernstein's inequality here, sphere packing arguments or inhomogeneous grids there) but lead to comparable results. In [Sha03b], Shamir insists on the order of magnitude of the minimax average redundancy versus the alphabet size. The approach presented here, focused on infinite alphabets, leads to a slightly better order of magnitude and to a much larger constant for this case.

2.3 Proof

We use here standard technique for lower-bounds based on Davisson [Dav73] and Rissanen [Ris84] : the n-th order maximin redundancy is bounded from below by (and asymptotically equivalent to) the capacity of the channel joining an input variable W with distribution π on Θ to the output variable Ψ_1^n with conditional probabilities $P_\theta(\Psi_1^n)$. Let $H(\Psi_1^n|W)$ be the conditional entropy of Ψ_1^n given W, and let $I\left(\Psi_1^n; W\right) = H(\Psi_1^n) - H(\Psi_1^n|W)$ denote the mutual information of these two random variables, see [CT91]. Then from [Dav73] and [Ris84] we know that inequality

$$R_\Psi^-(n, \Theta) \geqslant I\left(\Psi_1^n; W\right)$$

holds for all alphabets $\mathcal{A}$ and all prior distributions π on the set of memoryless distributions on $\mathcal{A}$: it is sufficient to give a lower-bound for the mutual information $I\left(\Psi_1^n; W\right)$ between parameter W and observation Ψ. In words, $R_\Psi^-(n, \Theta)$ is larger than the logarithm of the number of memoryless sources that can be *distinguished* from one observation of Ψ_1^n.

Given the positive integer n, let $c = c_n$ be an integer growing with n to infinity in a way defined later, $\lambda \in \mathcal{R}^+$ to be specified later, $d = \lambda\sqrt{c}$ and $\mathcal{A} = \{1, \ldots, c\}$. We denote by $\Theta^{c,d}$ the set of all unordered partitions of c made of summands at most equal to d :

$$\Theta^{c,d} = \left\{\theta = (\theta_j)_{j \in \mathbb{N}^+} : d \geqslant \theta_1 \geqslant \theta_2 \geqslant \ldots \text{ and } \sum_{j=1}^{\infty} \theta_j = c\right\}.$$

Then $\Theta^c \triangleq \Theta^{c,c}$ is the set of all unordered partitions of c. Let also $\Phi^{c,d}$ be the subset of Φ^c containing the profiles of all patterns $\psi \in \mathcal{P}^c$ whose symbols appear at most d times :

$$\Phi^{c,d} = \left\{\phi = (\phi_1, \ldots, \phi_d) \in \mathbb{N}^d : \sum_{\mu=1}^{d} \mu\phi_\mu = c\right\}.$$

There is a one-to-one mapping χ_c between Θ^c and Φ^c defined by

$$\left\{\begin{array}{lcl} \chi_c(\theta)_\mu & = & |\{i : \theta_i = \mu\}| ; \\ \chi_c^{-1}(\phi)_j & = & \left\{\begin{array}{ll} 0 & \text{if } \sum_{i=1}^d \phi_\mu < j, \\ \max\{\mu : \sum_{i=\mu}^d \phi_\mu \geqslant j\} & \text{else.} \end{array}\right. \end{array}\right.$$

It is immediately verified that $\chi\left(\Theta^{c,d}\right) = \Phi^{c,d}$.

In [DN90], citing [Sze51], Dixmier and Nicolas show the existence of an increasing function $f : \mathcal{R}^+ \longrightarrow [0, \pi\sqrt{\frac{2}{3}}]$ such that $\ln\left|\Theta^{c,d}\right| = f(\lambda)\sqrt{c}\,(1 + o(c))$ as $c \longrightarrow \infty$.

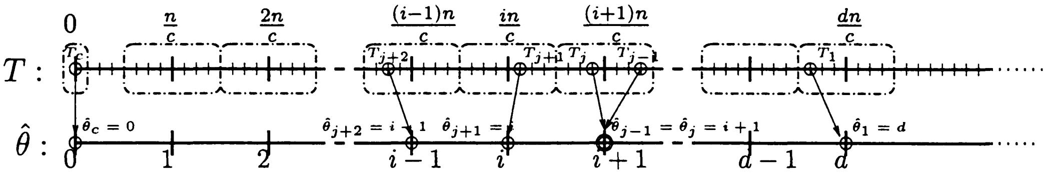


FIG. 2.1 – The profile of pattern ψ forms a partition of n that can be "shrunk" to θ, the parameter partition of c, with high probability.

For $\theta \in \Theta^{c,d}$, let p_θ be the distribution on $\mathcal{A}$ defined by $p_\theta(i) = \frac{\theta_i}{c}$, and let P_θ be the memoryless process with marginal distribution p_θ. Let W be a random variable with uniform distribution on the set $\Theta^{c,d}$. Let $X = (X_n)_{n \in \mathbb{N}^+}$ be a random process such that conditionally on the event $\{W = \theta\}$, then the distribution of X is P_θ, and let $\Psi = (\Psi_n)_{n \in \mathbb{N}^+}$ be the induced pattern process.

We want to bound $I(\Psi_1^n; W) = H(W) - H(W|\Psi_1^n)$ from below. As $H(\Theta^{c,d}) = \log |\mathcal{S}| = f(\lambda) \log e\sqrt{c}(1 + o(c))$, we need to find an upper-bound for $H(W|\Psi_1^n)$. The idea of the proof is the following. From Fano's inequality, upper-bounding $H(W|\Psi_1^n)$ reduces to finding an good estimator $\hat{\theta}$ for W : conditionally on $W = \theta$, string X_1^n is a memoryless process with distribution P_θ and we aim at recovering parameter θ from its pattern Ψ_1^n. Each parameter $\theta = (\theta_j)_{j \geqslant 1}$ is here an unordered partition with small summands of integer c. Let T_j be the number of occurrences of j-th most frequent symbol in ψ. Then $T = (T_j)_{j \geqslant 1}$ constitutes a random unordered partition of n. We show that by "shrinking" T by a factor c/n we we build a unordered partition $\hat{\theta}$ of c that is equal to parameter θ with high probability, see Figure 2.1. Note that only partitions with small summands are considered : this allows to have a better uniform control on the probabilities of deviation of each symbol's frequency, while the cardinality of $\Theta^{c,d}$ remains of same (logarithmic) order as that of Θ^c. The value of $\lambda = d/\sqrt{c}$ is chosen at the end to maximize the constant in Theorem 20.

Let us now give the details of the proof. If $W = \theta$ and if we observe string $X_1^n = x$ having pattern $\Psi_1^n = \psi \in \mathcal{P}^n$, we construct an estimator $\hat{\theta} = \left(\hat{\theta}_j\right)_{1 \leqslant j \leqslant c}$ of θ in the following way : let $\phi(\psi)$ be the profile of ψ, and $T = (T_j)_{j \geqslant 1} = \chi_n^{-1}(\phi(\psi))$ be the corresponding partition of n. For $j \geqslant c$, let $\hat{\theta}_j = \left[\frac{T_j c}{n}\right]$, where $[x]$ denotes the nearest integer of x. Observe that as alphabet $\mathcal{A}$ contains only c different symbols, for all $j > c$ we have $T_j = \hat{\theta}_j = \theta_j = 0$.

The distribution of T is difficult to study, but is very related to much simpler random variables. For $1 \leqslant i \leqslant n$ and $j \geqslant 1$, let $U_j^i = \mathbb{1}_{X_i = j}$; as U_j^i has a Bernoulli distribution with parameter $\frac{\theta_j}{c}$, and as process X is memoryless, we observe that $U_j \triangleq \sum_{i=1}^n U_j^i$, the number of occurrences of symbol j in x, has a binomial distribution $\mathcal{B}\left(n, \frac{\theta_j}{c}\right)$. Let $\tilde{\theta}_j = \left[\frac{U_j c}{n}\right]$, and $\tilde{\theta} = \left(\tilde{\theta}_j\right)_{j \geqslant 1}$; $\tilde{\theta}$ would be an estimator of θ if we had access to x, but here estimators may only be constructed from ψ. However, there is a strong connexion between $\hat{\theta}$ and $\tilde{\theta}$: the symbols in x are in one-to-one correspondence with the symbols in ψ. Hence, T is just the order statistics of U : $T_j = U_{(j)}$ and thus $\hat{\theta}_j = \tilde{\theta}_{(j)}$.

Now, if $\left|\frac{U_j c}{n} - \theta_j\right| < \frac{1}{2}$ then $\tilde{\theta}_j = \theta_j$. Thus, if $\forall j \in \{1, \dots, c\}, \left|\frac{U_j c}{n} - \theta_j\right| < \frac{1}{2}$ then $\tilde{\theta} = \theta$, hence $\tilde{\theta}$ is an increasing sequence and is equal to its order statistics $\hat{\theta}$. It appears thus that

$$\left\{\hat{\theta} = \theta\right\} \supset \bigcap_{j=1}^{c} \left\{\left|\frac{U_j c}{n} - \theta_j\right| < \frac{1}{2}\right\},$$

and hence :

$$\begin{aligned} P_\theta(\hat{\theta} \neq \theta) &\leqslant P_\theta\left(\bigcup_{j=1}^{c} \left\{\left|\frac{U_j c}{n} - \theta_j\right| \geqslant \frac{1}{2}\right\}\right) \\ &\leqslant \sum_{j=1}^{c} P_\theta\left(\left|\frac{U_j}{n} - \frac{\theta_j}{c}\right| \geqslant \frac{1}{2c}\right). \end{aligned}$$

We chose parameter set $\Theta^{c,d}$ so that all summands in partition θ are small in comparison to c. Consequently, the variance of the $\left(U_j^i\right)_{i,j}$ is uniformly bounded : $\mathrm{Var}[U_j^i] = \frac{\theta_j}{c}\left(1 - \frac{\theta_j}{c}\right) \leqslant \frac{d}{c}$. We can use Bernstein's inequality for the $\left(U_j^i\right)_{1 \leqslant i \leqslant n}$ to obtain :

$$\begin{aligned} P_\theta\left(\left|\frac{U_j}{n} - \frac{\theta_j}{c}\right| \geqslant \frac{1}{2c}\right) &\leqslant 2e^{-\frac{n/4c^2}{2(d/c+1/6c)}} \\ &= 2e^{-\frac{n}{8c(d+1/6)}}. \end{aligned}$$

Thus,

$$P(\hat{\theta} \neq \theta) = \frac{1}{\Theta^{c,d}} \sum_{\theta \in \Theta^{c,d}} P_\theta(\hat{\theta} \neq \theta) \leqslant 2ce^{-\frac{n}{8c(d+1/3)}}.$$

Now, using Fano's inequality [CT91] :

$$\begin{aligned} H(W|\Psi_1^n) &\leqslant H(W|\hat{\theta}) \\ &\leqslant P(W \neq \hat{\theta}) \log\left|\Theta^{c,d}\right| + \log 2 \\ &\leqslant 2ce^{-\frac{n}{8\lambda c^{3/2}}} f(\lambda)\sqrt{c}\,(1 + o(1)). \end{aligned}$$

Hence,

$$\begin{aligned} \mathcal{R}^{-}_{\Psi}(n, \Theta) &\geqslant I(\Psi_1^n; W) \\ &\geqslant f(\lambda) \log e\sqrt{c}\,(1 + o(1)) - 2ce^{-\frac{n}{8\lambda c^{3/2}}} f(\lambda) \log e\sqrt{c}\,(1 + o(1)) \\ &= f(\lambda) \log e\sqrt{c}\left(1 - 2ce^{-\frac{n}{8\lambda c^{3/2}}} + o(1)\right). \end{aligned}$$

By choosing $c = \left(\frac{n}{\frac{16}{3}\lambda\log n}\right)^{2/3}$ we get :

$$\begin{aligned}
\mathcal{R}^-_\Psi(n,\Theta) &\geqslant f(\lambda)\log e\left(\frac{n}{\frac{16}{3}\lambda\log n}\right)^{1/3}\left(1-2\left(\frac{n}{\frac{16}{3}\lambda\log n}\right)^{2/3}e^{-\frac{2}{3}\log n}+o(1)\right)\\
&= \frac{f(\lambda)}{\lambda^{1/3}}\log e\left(\frac{3n}{16\log n}\right)^{1/3}(1+o(1))\,.
\end{aligned}$$

By looking into the table of f given at page 151 of [DN90], we see that function $\lambda \longrightarrow f(\lambda)/\lambda^{1/3}$ reaches its maximum around $\lambda = 0.8$; for that choice, $f(\lambda) \approx 2.07236$ and we obtain :

$$\mathcal{R}^-_\Psi(n,\Theta) \geqslant 1.843\left(\frac{n}{\log n}\right)^{1/3}(1+o(1))\,.$$

Chapitre 3

Redundancy of the Context-tree Weighting Method on Renewal Processes

3.1 Introduction

Given a class $\mathcal{C} = \{P_\theta : \theta \in \Theta\}$ of sources on the alphabet $\{0,1\}$, the aim of *universal coding* is to find a single code (or equivalently a single sequence of coding distributions $(q_n)_n$) that uniformly approaches the entropy rate over all sources $P_\theta \in \mathcal{C}$. In order to price universality, we adopt the *individual sequences* framework [WMF94]. Let the *oracle code length* of a string $x \in \{0,1\}^n$ be the minimum number of bits $\inf_{P_\theta \in \mathcal{C}} - \log_2 P_\theta(x)$ required to encode x with coding probabilities from $\mathcal{C}$. The *regret* $R^*(q_n, \mathcal{C}, x)$ of q_n with respect to $\mathcal{C}$ over $x \in \{0,1\}^n$ is : $R^*(q_n, \mathcal{C}, x) = -\log q_n(x) - \inf_{P_\theta \in \mathcal{C}} - \log P_\theta(x)$. The performance index of q_n with respect to $\mathcal{C}$ is the maximal pointwise regret (or *individual redundancy*) :

$$R^*(q_n, \mathcal{C}) = \sup_{x \in \{0,1\}^n} R^*(q_n, \mathcal{C}, x) .$$

The *minimax individual redundancy* over model $\mathcal{C}$ on strings of length n, $R^*_n(\mathcal{C})$ is the infimum individual redundancy :

$$R^*_n(\mathcal{C}) = \inf_{q_n} R^*(q_n, \mathcal{C})$$

In this chapter, we show that a particular code called *Context Tree Weighting method* almost achieves the minimax individual redundancy on the classes of *renewal* and *Markov renewal* processes, see Theorems 22 and 23 formally stated in Section 3.

Let us first recall some background. A theorem by Rissanen [Ris84] states that if the model is smoothly parameterized and the parameter set Θ is k-dimensional, then the minimax individual redundancy is roughly $\frac{1}{2}\log n$ for each degree of freedom :

$$R_n^*(C) = \frac{k}{2}\log n + O(1) \tag{3.1}$$

This result handles most parametric models we are aware of. Complemented with efficient on-line implementations of *arithmetic coding*, Statistics-motivated universal coders offer an alternative to pattern-matching coders (as for example Lempel-Ziv coders) [RL81]. Moreover, standard game-theoretical [Hau97] arguments suggest that *mixture*-based techniques [KT81],[XB97] are optimal. In sharp contrast with dictionary techniques, mixture-based statistical methods asymptotically achieve minimax redundancy over many parametric classes of sources. Mixture-based universal coders culminate with the *Context-tree weighting* (CTW) algorithm, introduced in 1993 by F.M.J. Willems, Y.M. Shtarkov and T.J. Tjalkens (see [WT95] and Section 3.2 for relevant background).

It is natural to ask which non-parametric classes of sources have non-trivial (that is $o(n)$) minimax individual redundancy. In [Shi93], Shields proved that the class of all stationary ergodic processes is too large to have non-trivial minimax individual redundancy. But a few years later, in [CS96], Csiszár and Shields considered the class of *renewal processes*, that is processes on alphabet $\{0,1\}$ for which the distances between successive occurrences of 1 (the so-called inter-arrival times) are identically independently distributed. They characterized the minimax individual redundancy rate over this source class and showed it is of order $\sqrt{n}$. This provided a first example of an *intermediate*, massive complexity class (the natural parameterization is a subset of the unit sphere of $\ell_1(\mathbb{N})$). The same authors provided similar results on Markovian renewal processes, for which the inter-arrival times form a Markov chain. The upper-bounds in [CS96] were obtained using Shtarkov's Normalized Maximum Likelihood coder [Sht87]. The authors left open the question of building computationally efficient universal coders for the class of renewal processes.

We show in this chapter that the context-tree weighting method is almost *adaptive* on these classes : up to a constant times $\log n$, it achieves the minimax redundancy (Theorems 22 and 23). The nice behavior of CTW over the class of renewal processes could not be taken for granted : when using CTW on finite context-tree sources, the target source belongs to the very set of sources of which CTW mixtures are made. The model selection problem raised by universal compression over finite-context-tree sources is related to problems pertaining to order identification (see [CS00] and [Gar05b]).

When dealing with renewal processes using CTW, a correct trade-off between approximation and estimation errors has to be found. In contrast with Markov order estimation problems [CS00], it is necessary to handle large context-trees with *fast-growing* depth. Hence, analyzing universal coding for renewal processes using CTW is very much like investigating non-parametric estimation procedures (see [Tsy04]). Moreover, Theorem 21 shows that many algorithms that have been proposed so far cannot achieve minimax individual redundancy over renewal processes.

The chapter is organized as follows : Section 2 provides notations, concepts and relevant background on the context-tree weighting algorithm. Section 3 contains the main results of this chapter. Our theorem on the redundancy of the context-tree weighting method over the class of renewal processes is stated ; its counterpart for Markovian renewal processes follows. Section 4 is devoted to the proofs of the main theorems. Some technical lemmas are proved in the Appendix.

3.2 Notations and background

3.2.1 Strings

A *string* x on alphabet $\{0,1\}$ is an element of $\{0,1\}^* = \bigcup_{n=0}^{\infty}\{0,1\}^n$. If $x \in \{0,1\}^n$, its *length* is $|x| = n$. The empty string ϵ is the only element of $\{0,1\}^0$. The number of occurrences of a symbol $a \in \{0,1\}$ in string x is denoted by $N_a(x) = \sum_{i=1}^{n} \mathbb{1}_{x_i=a}$.

If string x is the concatenation of strings $s \in \{0,1\}^n$ and $z \in \{0,1\}^m$, then s is called a *prefix* of x, and z a *suffix* of x. Let $\text{Pref}(x) = \left\{x_1^j | 0 \leqslant j \leqslant n\right\}$ be the set of all prefixes of x and $\text{Suff}(x) = \left\{x_j^n | 1 \leqslant j \leqslant n+1\right\}$ be the set of all suffixes of x. The empty string ϵ is a prefix and a suffix of all strings. For $1 \leqslant i \leqslant l \leqslant n$, the *substring* $x_i.x_{i+1}\dots x_l$ is denoted by x_i^j.A string w is said to be a subsequence of string x if the characters of w appear in order within x, but possibly with gaps between occurrences of each character.

3.2.2 Binary context tree sources

For $\theta \in [0,1]$, let P_θ be the Bernoulli probability distribution with parameter θ on alphabet $\{0,1\}$ such that $P_\theta(1) = 1 - P_\theta(0) = \theta$, and let P_θ be the memoryless source with marginal P_θ ; for any positive integer n and for a string $x_1^n \in \{0,1\}^n$, $P_\theta(X_1^n = x_1^n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)}$.

When attempting to model dependent sources in a parsimonious way, *variable length* Markov chains [BW99] are now widely used. They allow the size of the memory to be data-dependent : the probability of a symbol given the past $x_{-\infty}^{-1}$ only involves a finite context :

$$p\left(.|x_{-\infty}^{-1}\right) = p\left(.|x_{-k}^{-1}\right),$$

whose length k depends on $x_{-\infty}^{-1}$. We need a few definitions to properly define context tree sources. Any model for finite context-tree sources is defined thanks to a complete suffix dictionary.

Définition 2. *A* complete suffix dictionary *is a finite set S of finite strings such that :*

$$\forall x_{-\infty}^{-1} \in \{0,1\}^{\mathbb{Z}-}, \exists \textit{ unique } k \in \mathbb{N} : x_{-k}^{-1} \in S.$$

Let $|S|$ denote the cardinality of S, and $l(S) = \max_{s \in S} |s|$ the length of its longest element.

Note that $\{\epsilon\}$ is the only complete suffix dictionary that contains ϵ.

Any complete suffix dictionary S defines a *suffix function* S^* mapping every semi-infinite string onto its unique (relevant) suffix in S. In a semi-infinite string x, symbol x_{k+1} is said to show up in context $s \in S$ if $S^*(x_{-\infty}^k) = s$. Every complete suffix dictionary can be visualized as a tree whose edges are labeled with letters, whose leaves are in one-to-one correspondence with the elements of S and whose internal nodes have exactly two outgoing edges, see figure 3.1. Henceforth, such a tree will be called a *Context Tree.* The set of all context trees with n leaves will be denoted by $\mathcal{CT}_n$ and the set of all finite context trees will be denoted by $\mathcal{CT} = \bigcup_{n \in \mathbb{N}} \mathcal{CT}_n$.

We can now formally introduce the concept of *context tree source* :

Définition 3 (Context-tree Source). *Let S be a context tree, and for each context $s \in S$ let $\theta_s \in [0,1]$ be a Bernoulli parameter. Under mild conditions, the transition kernel over* $\{0,1\}^{\mathbb{Z}-} \times \{0,1\}$

$$P_{S,\theta}\left(X_n = x_n \left| X_{-\infty}^{n-1} = x_{-\infty}^{n-1}\right.\right) = \theta_{S^*\left(x_{-\infty}^{n-1}\right)}(x_n)$$

defines a unique stationary source which is called a context tree source.

Note that a context-tree source is Markov of finite order. In the sequel, we shall use the more explicit notation $p_{S,\theta}(.|s)$ for $\theta_s(.)$, or even $p(.|s)$ when there is no ambiguity about the context tree under consideration.

To provide explicit formulas for the likelihood in context-tree models, let us define for all positive integers n the function $\mathcal{S}^* : S \times \{0,1\}^{\{-\infty \ldots n\}} \to \{0,1\}^*$ mapping every context s and each semi-infinite string x to the subsequence of x_1^n containing those symbols that occurs in context s :

$$\mathcal{S}^*(s,x) = \bigodot_{\substack{i=1..n \\ S^*\left(x_{-\infty}^{i-1}\right)=s}} x_i.$$

We have :

$$\begin{aligned} P_{S,\theta}\left(X_1^n = x_1^n \,\middle|\, X_{-\infty}^0 = x_{-\infty}^0\right) &= \prod_{i=1}^{n} p_{S,\theta}\left(x_i | S^*\left(x_{-\infty}^{i-1}\right)\right) \\ &= \prod_{s\in S} P_{\theta_s}\left(\mathcal{S}^*(s,x)\right). \end{aligned} \qquad (3.2)$$

In terms of coding theory, this means that a context tree source codes separately each subsequence $\mathcal{S}^*(s,x)$ with the memoryless coding distributions defined by θ_s, and that the code length of x is the sum of these different contributions.

If we are only provided with a finite string instead of the whole past (as will occur in practice), the context of the first symbols may not be well-defined. In order to circumvent this problem, the function S^* is extended : let $S^+ : \{0,1\}^k \rightarrow D \cup \{nil\}$ associate to every finite string $x_1^k \in \{0,1\}^k$ its suffix in D if it exists, or *nil* otherwise. Let also $\mathcal{S} : D \cup \{nil\} \times \{0,1\}^n \rightarrow \{0,1\}^*$ associate to each context s (in S or *nil*) and to each string x_1^n the subsequence of x_1^n containing those symbols that occurs in context s :

$$\mathcal{S}\left(s, x_1^n\right) = \bigodot_{\substack{i=1..n \\ S^+\left(x_1^{i-1}\right)=s}} x_i.$$

Note that if $k > l(S)$, then x_k cannot appear in state *nil* and thus

$$\left|\mathcal{S}\left(nil, x_1^n\right)\right| \leqslant l(S). \qquad (3.3)$$

3.2.3 KT-mixtures and context trees

The Krichevsky-Trofimov mixture [KT81] is defined as the *mixture* of all memoryless distributions P_θ, with respect to the Dirichlet Beta distribution ν with parameter 1/2 :

$$\mathcal{KT}(x) = \int P_\theta(x)\ \nu(\mathrm{d}\theta). \qquad (3.4)$$

The conjugacy between Beta and Binomial distributions which is at the core of classical Bayesian techniques gives an explicit, easily computable expression :

$$\mathcal{KT}(x) = \frac{\prod_{a:N_a(x)\geqslant 1}\left(N_a(x)-\frac{1}{2}\right)\left(N_a(x)-\frac{3}{2}\right)\dots\frac{1}{2}}{n\,(n-1)\dots 1}. \qquad (3.5)$$

The most remarkable property of the Krichevsky-Trofimov mixture (see [Cat01] for other alphabets) is that while it defines a consistent sequence of coding probabilities, up to a quantity that only depends on the target source, it almost achieves asymptotically minimax redundancy over all memoryless sources :

Proposition 14. *[WT95] If $x \in \{0,1\}^n$, then*

$$-\log_2 \mathcal{KT}(x) \leqslant \inf_{\theta\in\Theta} -\log_2 P_\theta(x) + \frac{1}{2}\log_2 n + 1. \tag{3.6}$$

To efficiently code in the class $(P_{S,\theta})_{\theta\in[0,1]^S}$ of sources based on a given context tree $S \in \mathcal{CT}$, the discussion following Equation (3.2) suggests to use a $\mathcal{KT}$ mixture and to define the following kernel and probability distribution, with or without knowledge of the past :

$$\begin{aligned} \mathcal{KT}_S(x_1^n|x_{-\infty}^0) &= \prod_{s\in S} \mathcal{KT}\left(S^*(s,x,n)\right) \text{ , and} \qquad (3.7)\\ \mathcal{KT}_S(x_1^n) &= P_{1/2}\left(\mathcal{S}(nil,x_1^n)\right)\prod_{s\in S}\mathcal{KT}\left(\mathcal{S}(s,x_1^n)\right). \qquad (3.8)\end{aligned}$$

Just as for $\mathcal{KT}$, there is an easy and efficient way (see [WT95] and [Cat01]) way to compute $\mathcal{KT}_S(x_1^n)$. As expected, $\mathcal{KT}_S$ is a probability measure close to all the $(P_{S,\theta})_{\theta\in\Theta^S}$ in the sense of the following proposition :

Proposition 15. *[WT95] Let $\gamma : \mathcal{R}_+ \to \mathcal{R}_+$ be defined by*

$$\gamma(z) = \begin{cases} z & \text{if} \quad 0 \leqslant z \leqslant 1,\\ 1+\frac{1}{2}\log_2 z & \text{if} \quad z > 1.\end{cases}$$

For every context tree S, for all positive integer n and for all semi-infinite words $x_{-\infty}^n$:

$$\begin{aligned} -\log_2 \mathcal{KT}_S(x_1^n|x_{-\infty}^0) &\leqslant \inf_{\theta\in\Theta^S} -\log_2 P_{S,\theta}(x_1^n|x_{-\infty}^0) + |S|\gamma\left(\frac{n}{|S|}\right),\\ &\textit{and}\\ -\log_2 \mathcal{KT}_S(x_1^n) &\leqslant \inf_{\theta\in\Theta^S}\inf_{x_{-\infty}^0\in\{0,1\}^{\mathbb{Z}_-}} -\log_2 P_{S,\theta}(x_1^n|x_{-\infty}^0) + |S|\gamma\left(\frac{n}{|S|}\right) + l(S).\end{aligned}$$

For the sake of self-reference, the proof of this proposition is given in the Appendix.

3.2.4 Double Mixture and universal coding : the Context Tree Weighting Algorithm

We follow here the presentation of [Cat01], which generalizes the arguments of [Wil94]. Consider the critical branching process on the rooted, infinite binary tree for which every node has two offsprings with probability 1/2, and none otherwise. A realization of that process is exactly tree S with probability

$$\pi(S) = 2^{-2|S|+1}. \tag{3.9}$$

For every integer n, the probability measure π on the set $\mathcal{CT}$ of all context trees is uniform over all elements of subset $\mathcal{CT}_n$. A decisive advantage of this measure is that, when considered as a Bayesian prior, it assigns a penalty to every model that is just a little smaller than the one prescribed by the Minimum Description Length principle [BRY98] and induced by the KT-mixture. This property will prove useful in the following. Also note that weight of the family $\mathcal{F}(S)$ of all trees containing a certain tree S as a rooted subtree is the probability that a realization of the branching process contains tree S, that is :

$$\pi\left(\mathcal{F}(S)\right) = 2^{-|S|+1}. \tag{3.10}$$

Now consider the double mixture probability distribution on $\{0,1\}^n$ defined for all $n \in \mathbb{N}$ by :

$$\mathcal{CTW}\left(x_1^n\right) = \sum_{S \in \mathcal{CT}} \pi(S)\ \mathcal{KT}_S\left(x_1^n\right). \tag{3.11}$$

It approximates all context tree sources, namely :

$$\mathcal{CTW}\left(x_1^n\right) \geqslant \sup_{S\in\mathcal{CT}} \pi(S)\ \mathcal{KT}_S\left(x_1^n\right) \geqslant \sup_{S\in\mathcal{CT}}\ \sup_{\theta=(\theta_s)_{s\in S}}\ \sup_{x_{-\infty}^0\in\{0,1\}^{\mathbb{Z}-}} P_{S,\theta}(x_1^n)2^{-|S|\gamma\left(\frac{n}{|S|}\right)-l(S)}\ \pi(S),$$

so that we have :

Proposition 16. *[Cat01] For all $n \in \mathbb{N}^*$ and all $x_1^n \in \{0,1\}^n$,*

$$-\log_2 \mathcal{CTW}\left(x_1^n\right) \leqslant \inf_{S\in\mathcal{CT}}\ \inf_{\theta=(\theta_s)_{s\in S}}\ \inf_{x_{-\infty}^0\in\{0,1\}^{\mathbb{Z}-}} -\log_2 P_{S,\theta}(x_1^n)+|S|\gamma\left(\frac{n}{|S|}\right)+l(S)+2|S|-1.$$

In a word, this means that the pointwise regret of $\mathcal{CTW}$ on the class of all context tree sources is not bigger than $\frac{1}{2}|S|\log_2\left(\frac{n}{|S|}\right) + l(S) + 3|S| - 1$. This is almost the asymptotic lower bound determined in [Ris84] (see also [WMF94] for individual sequences). We may thus assert that CTW is adaptive in the class of context tree sources.

3.3 CTW redundancy over Renewal and Markovian-Renewal processes

Given a distribution Q on $\mathbb{N}^*$, the *renewal process* P_Q is defined as the stationary process on $\{0,1\}^{\mathbb{Z}}$ such that the distances between two successive 1 (the so-called inter-arrival times) are i.i.d. with distribution Q. Renewal processes can easily be viewed as *infinite* context sources, involving a long range memory.

More precisely, let m_Q be the expectation of a random variable with distribution Q. String x_1^n can be parsed into renewal intervals :

$$x_1^n = 0^{t_0-1}1\, 0^{t_1-1}1\, 0^{t_2-1}1 \;\ldots\; 0^{t_N-1}1\, 0^{t_{N+1}-1}.$$

For $t \in \mathbb{N}^*$, let $R_Q(t) = \sum_{u=t}^{\infty} Q(u)$. Then (see [Var82]) the likelihood of x_1^n for the renewal process with inter-arrival times distribution as Q is :

$$P_Q^{\mathcal{R}}(x) = \left(\frac{1}{m_Q} R_Q(t_0)\right) \prod_{i=1}^{N} Q(t_i)\, R_Q\left(t_{N+1}\right). \tag{3.12}$$

See [CS96] for details, just note that t_0 has a "special treatment" because it is not an inter-arrival time (return time) but a first-passage time. This likelihood is maximized by the renewal process $\widehat{P}^{\mathcal{R}} = P_{\widehat{Q}}^{\mathcal{R}}$ whose inter-arrival time distribution will be denoted $\widehat{Q} = \widehat{Q}(x)$.

The maximum individual redundancy of the coding distribution q_n with respect to class $\mathcal{R}$ turns out to be :

$$R_n^*(q_n | \mathcal{R}) = \max_{x \in \{0,1\}^n} \log_2 \frac{\widehat{P}^{\mathcal{R}}(x)}{q_n(x)}.$$

Theorem 1 in [CS96] states that the minimax individual redundancy over renewal sources $\inf_{q_n} R_n^*\,(q_n | \mathcal{R})$, achieved by the Normalized Maximum Likelihood coder, grows like $\sqrt{n}$. The renewal sources model is both too large to enjoy parametric minimax individual redundancy (that is $O(\log n)$) and small enough to have a non-trivial minimax individual redundancy.

We shall prove that the context-tree weighting method is almost adaptive with respect to the class of renewal sources. This amounts to comparing $\mathcal{CTW}$'s code length (the code length derived from the CTW coding probability, see (3.11)) for a certain string $x \in \{0,1\}^n$ with the "oracle code length" $\log_2 \widehat{P}^{\mathcal{R}}(\cdot)$ defined by the class of renewal sources.

For a given context tree S, let us define $k_0(S)$ as the largest k such that 0^k belongs to S, that is $0^{k_0(S)} = \mathcal{S}(0^\infty)$. For the tree with one empty context, we agree on $k_0\left(\{\epsilon\}\right) = 0$. The following theorem characterizes the approximation capabilities of fixed context tree models with $\mathcal{KT}$-mixtures with respect to renewal processes :

Theorem 21. *There exists a positive constant C such that for any context tree S and all integers $n \in \mathbb{N}$:*

$$R_n^*(\mathcal{KT}_S | \mathcal{R}) \geqslant C \left(\frac{n \log_2 (k_0(S)+2)}{k_0(S)+2} + \frac{k_0(S)}{2} \max\left\{\log_2 \frac{n}{k_0(S)+2}, 1\right\}\right)$$

and

$$R_n^*(\mathcal{KT}_S|\,\mathcal{R}) \leqslant \frac{n\log_2\left[e\left(k_0(S)+1\right)\right]}{k_0(S)} + |S|\gamma\left(\frac{n}{|S|}\right) + l(S).$$

Remark 15. *Since the inception of context-tree universal coding, several universal coders have been built that use trees with bounded or slowly growing depth ([Ris83, WT95, Ris99]). This theorem shows that there is no hope to achieve minimax individual redundancy over renewal sources using those methods.*

In the upper and in the lower bounds, two leading terms appear : $\frac{n\log_2 k_0(S)}{k_0(S)}$ is an *approximation term* due to the limited memory of context tree models based on S, and (at least) $k_0(S)\log_2\frac{n}{|k_0(S)|}$ is an *estimation term* coming from all the $\mathcal{KT}$ mixtures at each leaf of S. They have antagonistic influences and need to be balanced.

The next theorem, characterizing CTW's redundancy over renewal processes, follows from Theorem 21 when balancing the approximation and estimation errors :

Theorem 22. *There exist constants C_1 and C_2 such that the pointwise redundancy of the context-tree weighting method over the class of renewal processes $\mathcal{R}$ satisfies :*

$$C_1\sqrt{n}\log_2 n \leqslant R_n^*(\mathcal{CTW}|\,\mathcal{R}) \leqslant C_2\sqrt{n}\log_2 n \quad \textit{for all } n\in\mathbb{N}.$$

Proof. Choosing $k=\lfloor\sqrt{n}\rfloor$ and $S=\{0^k\}\cup\{10^{j-1}|j=0..k\}$, we have $|S|=k+1$ (this amounts to choosing leaves for the trees S_i in Fig. 3.1).

From definition (3.11), $\mathcal{CTW}(x)\geqslant\pi(S)\mathcal{KT}_S(x)$ and hence

$$\begin{aligned}-\log_2\left(\mathcal{CTW}(x)\right) &\leqslant -\log_2\mathcal{KT}_S(x)-\log_2\pi(S)\\ &\leqslant -\log_2\widehat{P}^{\mathcal{R}}(x)+\frac{n\log_2 k}{k}+(k+1)\gamma\left(\frac{n}{k+1}\right)+k+2k+1\\ &\leqslant -\log_2\widehat{P}^{\mathcal{R}}(x)+C_2\sqrt{n}\log_2 n\end{aligned}$$

for some positive constant C_2.

And as $\mathcal{CTW}(x)\leqslant\max_{S\in\mathcal{CT}}\mathcal{KT}_S(x)$, the lower bound is obtained by noting that for all n, $\frac{n\log_2(k+2)}{k+2}+\frac{k}{2}\max\left\{\log_2\frac{n}{k+2},1\right\}$ is minimal around $k=\sqrt{n}$, and can be lower-bounded by $C_1\sqrt{n}\log_2 n$. ■

As in [CS96], the method extends to the class $\mathcal{MR}$ of *Markovian* renewal times, when the sequence of inter-arrival times form an integer-valued Markov chain.

Theorem 23. *There exist constants C_3 and C_4 such that the pointwise redundancy of the context-tree weighting method on Markov renewal sources $\mathcal{MR}$ satisfies :*

$$C_3 n^{2/3}\log_2 n \leqslant R_n^*(\mathcal{CTW}|\,\mathcal{MR}) \leqslant C_4 n^{2/3}\log_2 n \quad \textit{for all } n\in\mathbb{N}.$$

Remark 16. *Let us point out another distinctive feature of the context-tree weighting method. In [CS96], the* weak redundancy-rate *is defined in the following way : a non-negative function $n \longrightarrow \rho(n)$ is said to be a weak redundancy rate bound for the class C if there is a sequence $(Q_n)_{n\in\mathbb{N}}$ of coding distributions such that for each source P in C, there is a constant $K(P)$ such that for all n, the expected redundancy $\bar{R}(Q_n|P) = \sum_{x\in\{0,1\}^n} P(x)\log_2 \frac{P(x)}{Q_n(x)}$ is upper-bounded by $K(P)\rho(n)$.*

As any finite *(Markov) renewal process is a context-tree source (it will become apparent in the following proof),* $\log n$ *is a weak redundancy-rate bound for these classes. Hence, Proposition 16 shows that the context-tree weighting method also achieves the weak redundancy-rate for finite renewal and Markov renewal processes.*

3.4 Proofs

3.4.1 Theorem 21, proof of the upper bound

Let $n \in \mathbb{N}$, let $x \in \{0,1\}^n$. We suppose in the following that x contains at least two symbols '1'. (otherwise, x has such a low empirical entropy that the result is obvious). Then, there exist positive integers $N \leqslant n$, $(t_0, \dots, t_{N+1}) \in (\mathbb{N}^*)^{N+2}$ such that $\sum_{i=0}^{N+1} t_i = n$ and let

$$x_1^n = 0^{t_0-1}1\; 0^{t_1-1}1\; 0^{t_2-1}1\; \dots\; 0^{t_N-1}1\; 0^{t_{N+1}-1}.$$

Given a context tree S, we will write k for $k_0(S)$. The string x is split into three substrings : a prefix b, a factor m and a suffix e. Let $b = 0^{t_0-1}1$ if $t_0 \leqslant k$ and $b = 0^k$ else (so that $|b| \leqslant k$), $e = 0^{t_{N+1}-1}$ and m be such that we can write

$$x = b\; m\; e.$$

In [Var82], Vardi studies maximum likelihood estimates of the renewal distribution. In the present case, he shows (Theorem 1 and in the following remark) the existence of a unique maximizer $\widehat{Q}$ of the likelihood expression (3.12), and provides an algorithm to compute it numerically. Let thus $\widehat{P}^{\mathcal{R}} = P^{\mathcal{R}}_{\widehat{Q}}$ be the renewal source distribution that maximizes likelihood on x, then the oracle code length for x with respect to $\mathcal{R}$ is $-\log_2 \widehat{P}^{\mathcal{R}}(x)$. Henceforth, we write Q (resp. R) as a shorthand for $\widehat{Q}$ (resp. $R_{\widehat{Q}}$), and $0/0 = 0$. We will use the term "loss" of a distribution q versus a distribution p for the difference of code length $\log \frac{1}{q(x)} - \log \frac{1}{p(x)}$.

We shall now specify the parameters of a tree source P_S that faithfully simulates the behavior of $P^{\mathcal{R}}_Q$ for inter-arrival times smaller than k. We will see that the loss between $P_S(x)$ and $\widehat{P}^{\mathcal{R}}(x)$, due to the other (larger) inter-arrival times,

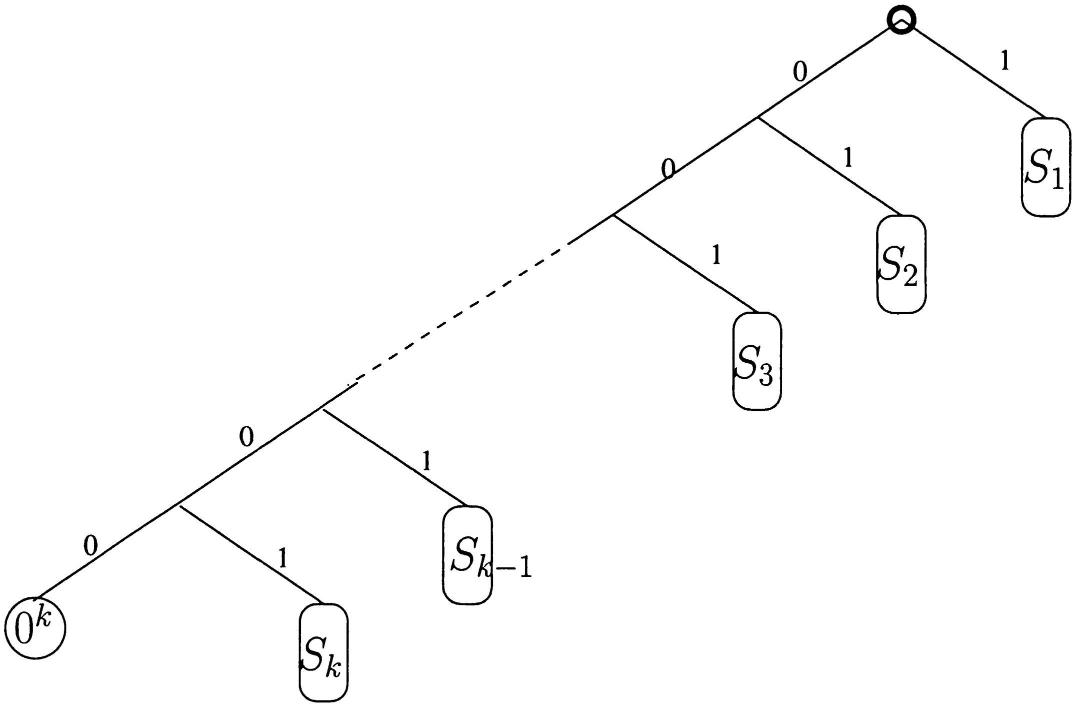


FIG. 3.1 – Decomposition of the context tree S

is bounded by a decreasing function of k. On the other hand, the classical $\mathcal{KT}$-approximation argument will upper-bound the loss between $\mathcal{KT}_S(x)$ and $P_S(x)$ with an increasing function of k.

For each j with $1 \leqslant j \leqslant k$, let $S_j = \{0,1\}^*10^{j-1} \cap S$ (S_j is the set of strings from S ending with 10^{j-1}). As S is a complete suffix dictionary, S_j is not empty and the collection $(S_j)_{j \leqslant n}$ defines a partition of S, see Figure (3.1) :

$$S = \{0^k\} \cup \bigcup_{j=1}^{k} S_j.$$

Note that for the tree of size one this decomposition writes $S = \{\epsilon = 0^0\}$.

- For each $j, 1 \leqslant j \leqslant k$, and $s \in S_j$, let :

  $$p_S\left(1|s\right) = p_S\left(1|10^{j-1}\right) = \frac{Q(j)}{R(j)};$$

  Hence : $p_S\left(0|10^{j-1}\right) = \frac{R(j+1)}{R(j)}$.
- Let $w = \mathcal{S}(0^k, x)$, $B = |w|$ and $A = \sum_{i=1}^{N+1} \mathbb{1}_{t_i > k}$ be the number of occurrences of 1 in w. As $|w| \leqslant |x| = n$, we observe that $A \leqslant n/(k+1)$, and we choose :

  $$p_S\left(1|0^k\right) = \frac{A}{B}.$$

For all $1 \leqslant t \leqslant k$, source p_S perfectly models the renewal process :

$$\begin{aligned} P_S\left(0^{t-1}1|1\right) &= \prod_{u=1}^{t-1} p_S\left(0|10^{u-1}\right)\ p_S\left(1|10^{t-1}\right) \\ &= \prod_{u=1}^{t-1} \frac{R\,(u+1)}{R\,(u)}\ \frac{Q(t)}{R\,(t)} = Q(t)\,. \end{aligned}$$

But on the other hand, for $t \geqslant k$:

$$\begin{aligned} P_S\left(0^{t-1}1|1\right) &= \left(\prod_{u=1}^{k-1} p_S\left(0|10^{u-1}\right)\right)\ P_S\left(0^{t-k}1|0^k\right) \\ &= R\,(k)\ P_S\left(0^{t-k}1|0^k\right) \\ &\geqslant Q(t)\ P_S\left(0^{t-k}1|0^k\right)\ . \end{aligned}$$

Whatever the value of t, we can write $p_S\left(0^{t-1}1|1\right) \geqslant Q(t)\ P_S\left(\mathcal{S}(0^k,0^{t-1})1|0^k\right)$ and hence

$$p_S\,(m|1) \ \geqslant\ \prod_{i=1}^{N} Q(t_i)\ P_S\left(\mathcal{S}(0^k,m)|0^k\right)\ .$$

It appears similarly that

$$P_S\,(e|1) \ \geqslant\ R\,(t_{N+1})\ P_S\left(\mathcal{S}(0^k,e)|0^k\right),$$

so that, as $w = \mathcal{S}(0^k, me)$:

$$\begin{aligned} P_S\,(me|1) &\geqslant \prod_{i=1}^{N} Q(t_i)\ R\,(t_{N+1})\ P_S\left(w|0^k\right) \\ &\geqslant \widehat{P}^{\mathcal{R}}(x)\ P_S\left(w|0^k\right))\,. \end{aligned} \tag{3.13}$$

And using the $\mathcal{KT}_S$ approximation of P_S :

$$\begin{aligned} -\log_2 \mathcal{KT}_S(x) &= -\log_2 P_{1/2}(b) - \log_2 \mathcal{KT}_S(me|1) && (3.14) \\ &\leqslant -\log_2|b| - \log_2 P_S(me|1) + |S|\gamma\left(\frac{n}{|S|}\right) && (3.15) \\ &\leqslant l(S) - \log_2 \widehat{P}^{\mathcal{R}}(x) - \log_2 p_S\left(w|0^k\right) + |S|\gamma\left(\frac{n}{|S|}\right). && (3.16) \end{aligned}$$

where (3.14) follows from Definition (3.8), while (3.15) follows from Proposition (15) and (3.16) from (3.3) and(3.13).

In order to bound the loss in state 0^k, we take advantage of the fact that w contains "few" occurrences of 1. If $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$ is the binary Shannon entropy of the Bernoulli distribution with success probability x, then

$$-\log_2 P_S(w|0^k) = Bh\left(\frac{A}{B}\right) \leqslant A\log_2\left(e\,\frac{B}{A}\right) \leqslant A\log_2\left(e\,\frac{n}{A}\right) \leqslant \frac{n}{k+1}\log_2\left[e\ \ (k+1)\right],$$

where we successively used the upper-bound of Lemma 3 (see Appendix) for h, the fact that $x \longrightarrow x\log_2\frac{e\,n}{x}$ is an increasing function of x on $]0,n]$, and that $A \leqslant n/(k+1)$.

Hence

$$-\log_2 \mathcal{KT}_S(x) \ \leqslant\ -\log_2 \widehat{P}^{\mathcal{R}}(x) + l(S) + \frac{n}{k+1}\log_2\left[e\,(k+1)\right] + |S|\gamma\left(\frac{n}{|S|}\right).$$

3.4.2 Theorem 21, proof of the lower bound

Let n be an integer larger than 4, S be a context source and $k = k_0(S)$.

- If $k \geqslant n/2$, then we consider string $x = 0^n$. Obviously, x has a maximum likelihood equal to 1, but the k first symbols of x appear in different states of S. Hence, $-\log \mathcal{KT}_S(x) \geqslant \sum_{j=0}^{k-1} \mathcal{KT}\left(\mathcal{S}(0^j,x) = 0\right) \geqslant k$.
- If $k < n/2$, we bound the maximal regret of $\mathcal{KT}_S$ on $\{0,1\}^n$ from below by considering a particular string x made of runs of 0's of equal length separated by 1's. String x has very high probability under some renewal process, but $\mathcal{KT}_S$ does not catch its regularity.

  Let q (resp. l) be the quotient (resp. remainder) of the integer division of $n-1$ by $2k+2$: $n-1 = q\times(2k+2)+l$. Let

  $$x = 0^l 1 \left(0^{2k+1}1\right)^q$$

  be the string of length $l+1+2kq = n$ we shall consider. If Q is the distribution on $\mathbb{N}^*$ concentrated on $2k+2$, then from Equation (3.12) we can bound from below the maximum likelihood by

  $$\widehat{P}^{\mathcal{R}}(x) \geqslant P_Q^{\mathcal{R}}(x) = \left(\frac{1}{m_Q}R_Q(l+1)\right)\prod_{i=1}^{N} Q(2k+2)\,R_Q\,(0) \geqslant \frac{1}{2k+2},$$

  and hence the maximal log-likelihood satisfies :

  $$-\log_2 \widehat{P}^{\mathcal{R}}(x) \leqslant 1 + \log_2(k+1)\,. \qquad (3.17)$$

As $S = \{0^k\} \cup \{S_j : j = 1..k\}$, we can write :

$$\begin{aligned} -\log_2 \mathcal{KT}_S(x) &= -\log_2 \left(P_{1/2}(0^l 1) \mathcal{KT}(\mathcal{S}(0^k, x)) \prod_{j=1}^{k} \prod_{s \in S_j} \mathcal{KT}(\mathcal{S}(s, x)) \right) \\ &\geqslant -\log_2 \mathcal{KT}(\mathcal{S}(0^k, x)) \\ &\qquad + \sum_{j=1}^{k} -\log_2 \prod_{s \in S_j} \mathcal{KT}(\mathcal{S}(s, x)). \end{aligned} \tag{3.18}$$

When scanning x, all symbols appear in context 0^k except maybe the first occurrence of 1 and the first k occurrences of 0 in each run. In fact, if $y = \mathcal{S}(0^k, x)$ then $B = |y| \geqslant \frac{n-l-1}{2} = (k+2)\,q$, y contains $A \in \{q, q+1\}$ symbols '1' and hence, whatever $p_S(.|0^k)$ is it cannot be better than the maximum likelihood :

$$\begin{aligned} -\log_2 \mathcal{KT}\,(y) &\geqslant -\log_2 p_S\,(y|0^k) \geqslant Bh\left(\frac{A}{B}\right) \\ &\geqslant A \log_2 \frac{B}{A} \geqslant q \log_2 (k+2)\,. \end{aligned} \tag{3.19}$$

On the other hand, for each $1 \leqslant j \leqslant k$ and each $s \in S_j$, the subsequence $\mathcal{S}(s, x)$ appearing in state s is sequence of 0's : $\bigodot_{s \in S_j} \mathcal{S}(s, x) = 0^q$. Coding these subsequences with a $\mathcal{KT}$-mixture causes redundancy : Lemma 4 from the Appendix ensures that for all j :

$$\prod_{s \in S_j} \mathcal{KT}\,(\mathcal{S}(s, x)) \leqslant \mathcal{KT}\left(\bigodot_{s \in S_j} \mathcal{S}(s, x) \right) = \mathcal{KT}\,(0^q) \leqslant \frac{1}{\sqrt{q}},$$

so that

$$\sum_{j=1}^{k} -\log_2 \prod_{s \in S_j} \mathcal{KT}\,(\mathcal{S}(s, x)) \geqslant \frac{k}{2} \log_2 q. \tag{3.20}$$

Combining (3.17), (3.18), (3.19) and (3.20), we get :

$$-\log_2 \mathcal{KT}_S(x) + \log_2 \widehat{P}^{\mathcal{R}}(x) \geqslant q \log_2 k + \frac{k}{2} \log_2 q - 1 - \log_2 (k+1)\,.$$

As $q > \frac{n}{2k+2} - 1$, this concludes the proof.

3.4.3 Theorem 23, proof of the upper bound

The structure of this proof parallels the proof of Theorem 21 : to bound $\mathcal{CTW}(x)$ from below, we will use a particular tree that only depends on n and thestructure of which captures "small" Markovian dependencies. We shall provide it with parameters inherited from the maximum likelihood estimate for x, so that all probabilities of those transitions will be evaluated correctly. Then we shall balance approximation loss and estimation loss.

Let again $n \in \mathbb{N}$, $2 \leqslant N \leqslant n$, $(t_0, \dots, t_{N+1}) \in (\mathbb{N}^*)^{N+2}$ such that $\sum_{i=0}^{N+1} t_i = n$ and let us suppose that we can write

$$x_1^n = 0^{t_0-1}1\;0^{t_1-1}1\;0^{t_2-1}1\;\dots\;0^{t_N-1}1\;0^{t_{N+1}-1}.$$

(the result is obvious strings x containing at most two 1). Let also $k \in \mathbb{N}$, and $b = 0^{t_0-1}10^{t_1-1}1$ if $t_0+t_1 \leqslant k$, and the first k symbols of this string otherwise. Let m be such that $bm = 0^{t_0-1}10^{t_1-1}10^{t_2-1}1\;0^{t_3-1}1\;\dots\;0^{t_N-1}1$ and $e = 0^{t_{N+1}}$. String x can be decomposed as

$$x = b\;m\;e.$$

For a Markov kernel $Q(.|.)$ on $\mathbb{N}^*$ and $t \in \mathbb{N}^*$, we introduce the notation $R_Q(t|.) = \sum_{u=t}^{\infty} Q(u|.)$. Let $\widehat{P}^{\mathcal{MR}} = P^{\mathcal{MR}}_{\widehat{Q}}$ be a Markovian renewal distribution with underlying inter-arrival Markov kernel $\widehat{Q}(.|.) = \widehat{Q}_x(.|.)$ that realizes the maximum likelihood on x, then :

$$\widehat{P}^{\mathcal{MR}}_Q(x) = \widehat{P}^{\mathcal{MR}}_Q(b) \prod_{i=2}^{N} \widehat{Q}(t_i|t_{i-1})\; R_{\widehat{Q}}\left(t_{N+1}\,|t_N\right).$$

In the following, we write Q (resp. R) as a shorthand for $\widehat{Q}$ (resp. $R_{\widehat{Q}}$), and 0/0=0.

Let S_k be the suffix tree of depth $l(S_k) = 2k$ containing the following k^2+k+1 contexts :

$$S_k = \left\{10^{j-1}10^{i-1} : 1 \leqslant i, j \leqslant k\right\} \cup \left\{s_j = 0^k10^{j-1} : 1 \leqslant j \leqslant k\right\} \cup \left\{s_0 = 0^k\right\}.$$

Let P_{S_k} the context-tree source in the model of S_k defined by :

$$\begin{cases} p_{S_k}\left(1|10^{j-1}10^{i-1}\right) = \frac{Q(i|j)}{R(i|j)} & ,1 \leqslant i, j \leqslant k; \\ p_{S_k}\left(1|0^k10^{j-1}\right) = \frac{A_j}{B_j} & ,1 \leqslant j \leqslant k; \\ p_{S_k}\left(1|0^k\right) = \frac{A_0}{B_0}. & \end{cases}$$

As for the renewal case, S_k catches the probability dependencies for transitions

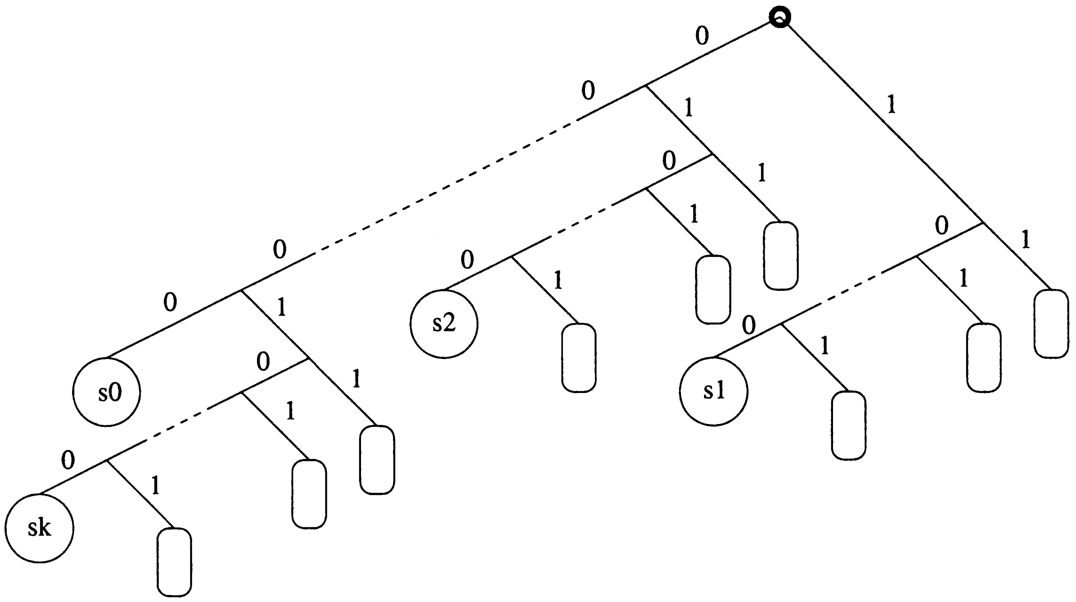


FIG. 3.2 – Context tree adapted to Markov Renewal processes

involving "small" inter-arrival times : for all $1 \leqslant i, j \leqslant k$,

$$\begin{aligned} P_{S_k}(0^{i-1}1|10^{j-1}1) &= \prod_{u=1}^{i-1} p_S\left(0|10^{j-1}10^{u-1}\right)\, p_S\left(1|10^{j-1}10^{i-1}\right) \\ &= \prod_{u=1}^{i-1} \frac{R(u+1|j)}{R(u|j)}\,\frac{Q(i|j)}{R(i|j)} = Q(i|j). \end{aligned}$$

Hence we can write the same kind of decomposition as in the renewal case, distinguishing the small transitions from those involving some s_j (with $0 \leqslant j \leqslant k$), where bounds of memory cause a loss of information :

$$\begin{aligned} -\log_2 \mathcal{KT}_{S_k}(x) &= -\log_2 P_{1/2}(b) - \log_2 \mathcal{KT}_{S_k}(me|b) \\ &\leqslant -\log_2 |b| - \log_2 P_{S_k}(me|b) + |S_k|\gamma\left(\frac{n}{|S_k|}\right) \\ &\leqslant -\log_2 k - \log_2 \widehat{P}^{RM}(x) + \sum_{j=0}^{k} -\log_2 p\left(S_k(s_j, x)|s_j\right) \\ &\quad +(k^2+k+1)\gamma\left(\frac{n}{k^2+k+1}\right). \end{aligned}$$

But coding the part of x that is not properly modeled by S_k in many different subsequences cannot be worse than coding them altogether : if $B = \sum_{j=0}^{k} B_j$ and

$A = \sum_{j=0}^{k} A_j$, the concavity of function h ensures that :

$$\sum_{j=0}^{k} -\log_2 p(S_k(s_j,x)|s_j) = \sum_{j=0}^{k} B_j h\left(\frac{A_j}{B_j}\right) \leqslant Bh\left(\frac{A}{B}\right) \leqslant A\log_2\left(e\frac{B}{A}\right) \leqslant 2\frac{n}{k}\log_2 e\; k.$$

where the last inequality comes from the fact that $x \longrightarrow x\log_2 \frac{e\,n}{x}$ is an increasing function of x on $]0,n]$, and that $A \leqslant 2n/k$ and $B \leqslant n$.

The conclusion comes by balancing the approximation and the estimation losses, which amounts to setting $k = \lfloor n^{1/3} \rfloor$ in inequality :

$$\begin{aligned} -\log_2 \mathcal{CTW}(x) \;&\leqslant\; -\log_2 \widehat{P}^{\mathcal{MR}}(x) + 2k + 2\frac{n}{k}\log_2 e\; k + (k^2+k+1) \\ &\quad +(k^2+k+1)\gamma\left(\frac{n}{k^2+k+1}\right). \end{aligned}$$

3.4.4 Theorem 23, proof of the lower bound

To alleviate excessive technicalities, for each integer k we give the proof for some n in the interval $[4k^3 - k^2, 4k^3 + 3k^2]$. This is sufficient to prove the theorem, as $\left(R_n^*(\mathcal{CTW}, \mathcal{MR})\right)_n$ is a non-decreasing sequence. Indeed, if x_1^n maximizes $\frac{\widehat{P}^{\mathcal{MR}}(x_1^n)}{\mathcal{CTW}(x_1^n)}$ and if $P = \widehat{P}^{\mathcal{MR}}$ is the maximum likelihood distribution for x_1^n, then for all x_{n+1} in $\{0,1\}$ we have

$$\begin{aligned} R_{n+1}^*(\mathcal{CTW}, \mathcal{MR}) \;&\geqslant\; \log_2 \frac{P(x_1^n) \times P(x_{n+1}|x_1^n)}{\mathcal{CTW}(x_1^n) \times \mathcal{CTW}(x_{n+1}|x_1^n)} \\ &=\; R_n^*(\mathcal{CTW}, \mathcal{MR}) + \log_2 \frac{P(x_{n+1}|x_1^n)}{\mathcal{CTW}(x_{n+1}|x_1^n)}, \end{aligned}$$

and there is at least one x_{n+1} such that $P(x_{n+1}|x_1^n) \geqslant \mathcal{CTW}(x_{n+1}|x_1^n)$.

For $k \in \mathbb{N}^*$, let q be the one-to-one mapping on $W_k = \{k, k+1, \ldots, 3k-1\}$ defined by

$$q : i \longrightarrow \begin{cases} i+k+1 & \text{if} \quad k \leqslant i < 2k-1 \\ \quad 2k & \text{if} \quad\; i = 2k-1 \\ \; i-k & \text{if} \quad 2k \leqslant i < 3k, \end{cases}$$

and let Q be the deterministic Markovian kernel with support W_k defined by $\forall k \leqslant i < 3k, Q(q(i)|i) = 1$.

We construct a string $x = 0^{t_0-1}10^{t_1-1}1\ldots$ having inter-arrival time transitions compatible with q : hence its probability under $P_Q^{\mathcal{MR}}$ reduces to $P_Q^{\mathcal{MR}}(0^{t_0-1}10^{t_1-1}1) \times 1$. In order to keep this probability high, we choose for (t_0, t_1) the pair (i^*, j^*) that maximizes $P_Q^{\mathcal{MR}}(0^{i-1}10^{j-1}1)$ for $1 \leqslant i < 3k$ and $k \leqslant j < 3k$. As there are $3k \times 2k = 6k^2$ such pairs, and as they cover all the possible beginnings, we have $P_Q^{\mathcal{MR}}(0^{i^*-1}10^{j^*-1}1) \geqslant \frac{1}{6k^2}$.

Let $b^* = 0^{i^*-1}10^{j^*-1}1 \bigodot_{r=0}^{q^r(j^*)=2k-1} 0^{q^r(j^*)-1}1$ and

$$x = b^* \left(\bigodot_{i=0}^{k-1} 10^{2k+i-1}10^{k+i-1} \right)^k ,$$

so that $n = |x| \in [4k^3 - k^2, 4k^3 + 3k^2]$, and let S be a context tree.

- The "oracle cost" for x is negligible : $\widehat{P}^{\mathcal{MR}}(x) \geqslant P_Q^{\mathcal{MR}}(x) = P_Q^{\mathcal{MR}}(0^{i^*-1}10^{j^*-1}1) \times 1 \geqslant \frac{1}{6k^2}$. Hence
$$-\log_2 \widehat{P}^{\mathcal{MR}}(x) \leqslant 3 + 2\log_2 k.$$
- If S does not contain 10^{2k-1} as an internal node, then all the tails of the blocks 10^{2k+i-1} are coded in the same state. In fact, if $s = 0^{k_0(S)}$ and $w = \mathcal{S}(s,x)$ then w contains $B \geqslant k\sum_{i=0}^{2k-1} i \geqslant k^3$ symbols, among which $k^2 \leqslant a \leqslant 2k^2$ symbols '1', and hence
$$\begin{aligned} -\log_2 \mathcal{KT}_S(x) &\geqslant -\log_2 P_S(x) \geqslant -\log_2 P_S(w) \geqslant Bh\left(\frac{A}{B}\right) \\ &\geqslant A\log_2 \frac{B}{A} \geqslant A\log_2 \frac{k^3}{A} \geqslant k\log_2 k \geqslant \epsilon n^{2/3}\log n. \end{aligned}$$
for ϵ small enough, independent of n.
- If S does contain 10^{2k-1} as an internal node, then for all $1 \leqslant i < j \leqslant 2k$ and all $w \in \{0,1\}^*$, $S^+(w10^{i-1}) \neq S^+(w10^{j-1})$: the blocks 10^{k+i-1} are coded in different states. So that if we let $S_i = S \cap \{0,1\}^*10^{i-1}$, we have :
$$-\log_2 \mathcal{KT}_S(x) \geqslant \prod_{i=1}^{k}\prod_{s\in S_i} \mathcal{KT}\left(\mathcal{S}(s,x)\right).$$
Now, for all $1 \leqslant i \leqslant k$:
  - if S contains a internal node $0^{k-1}10^i$, then for all $1 \leqslant j \leqslant k$, $10^{2k+i-1}10^j$ appears at least k times followed by a '0'. Thus, thanks to Lemma 4, the cost of those '0' is at least $\log_2 k$ for every j. Hence :
$$-\log_2 \prod_{s\in S_i} \mathcal{KT}\left(\mathcal{S}(s,x)\right) \geqslant k\log_2 k.$$
  - if S does not contain $0^{k-1}10^i$ as an internal node, then defining $s = 0^{j_i}10^i$ (j_i being the largest integer j such that 0^j10^i belongs to S) and $w = \mathcal{S}(s,x)$, w contains $B \geqslant \sum_{i=k}^{2k-1} i \geqslant k^2$ symbols, among which $k \leqslant A \leqslant 2k$ symbols '1', and hence
$$-\log_2 \prod_{s\in S_i} \mathcal{KT}\left(\mathcal{S}(s,x)\right) \geqslant Bh\left(\frac{A}{B}\right) \geqslant A\log_2\frac{B}{A} \geqslant A\log_2\frac{k^2}{A} \geqslant k\log_2 k.$$

Hence, summing up all contributions for $1 \leqslant i \leqslant k$, then for some constant C_3 small enough independent of n :

$$-\log_2 \mathcal{KT}_S(x) \geqslant k^2 \log_2 k \geqslant C_3 n^{2/3} \log n.$$

3.5 Proofs of the technical lemmas

Proof. Proposition 15.
We suppose here that no memory of the past is available, otherwise the proof is slightly shorter :

$$\begin{aligned}
-\log_2 \mathcal{KT}_S(x_1^n) &= \sum_{s \in S \cup \{nil\}} -\log_2 \mathcal{KT}(\mathcal{S}(s, x_1^n)) && (3.21)\\
&\leqslant \left[\sum_{s \in S} \inf_{\theta(s) \in \Theta} -\log_2 P_{\theta(s)}(\mathcal{S}(s, x_1^n))) + \gamma\left(|\mathcal{S}(s, x_1^n)|\right)\right] \\
&\qquad + P_{1/2}(\mathcal{S}(nil, x_1^n)) && (3.22)\\
&\leqslant \inf_{\theta \in \Theta^S} \left[\inf_{x_{-\infty}^0 \in \{0,1\}^{\mathbb{Z}-}} \sum_{s \in \mathcal{S}} -\log_2 P_{\theta(s)}(\mathcal{S}^*(s, x_{-\infty}^n)))\right] \\
&\qquad + |S| \sum_{s \in \mathcal{S}} \frac{1}{|S|} \gamma\left(|\mathcal{S}(s, x_1^n)|\right) + l(S) && (3.23)\\
&\leqslant \inf_{\theta \in \Theta^S} \inf_{x_{-\infty}^0 \in \{0,1\}^{\mathbb{Z}-}} -\log_2 P_{S,\theta}(x_1^n) \\
&\qquad + |S| \gamma\left(\frac{n}{|S|}\right) + l(S). && (3.24)
\end{aligned}$$

Equality (3.21) follows from (3.8), Inequality (3.22) follows from Proposition 14, Inequality (3.23) from (3.3) and Inequality (3.24) from (3.2), from the concavity of γ and from the inequality : $\sum_{s \in S} |\mathcal{S}(s, x_1^n)| \leqslant n$. ■

Lemma 3.

$$\forall p \in]0,1], p \log_2 \frac{1}{p} \leqslant h(p) \leqslant p \log_2 \frac{e}{p}.$$

Proof. Let $f(x) = -\log(1-x) - x - x^2$.

Then $f'(x) = \frac{1}{1-x} - 1 - 2x$ and $\forall x \in \left[0, \frac{1}{2}\right], f'(x) \leqslant 0$.

As $f(0) = 0$, we have $\forall x \in \left[0, \frac{1}{2}\right], -\log(1-x) \leqslant x + x^2$, or $-\log_2(1-x) \leqslant (x + x^2) \log_2 e$.

Hence if $p \leqslant \frac{1}{2}$ then $-(1-p)\log_2(1-p) \leqslant (1-p)(p+p^2)\log_2 e \leqslant p \log_2 e$, and $h(p) \leqslant p \log_2 \frac{e}{p}$.

But $p \longrightarrow p \log_2 \frac{e}{p}$ is an increasing function of p on $]0,1]$, whereas h is decreasing on $\left]\frac{1}{2}, 1\right]$. ■

Lemma 4. *For all $N \in$ and all $(k_1, \ldots, k_N) \in \mathbb{N}^N$:*

$$\prod_{i=1}^{N} \mathcal{KT}\left(0^{k_i}\right) \leqslant \mathcal{KT}\left(0^{K}\right) \leqslant \frac{1}{\sqrt{K}}$$

where $K = \sum_{i=1}^{N} k_i$.

Proof. It suffices to show the first inequality for $N = 2$. By definition (3.4) :

$$\begin{array}{rcl} \mathcal{KT}\left(0^{k_1}\right)\mathcal{KT}\left(0^{k_2}\right) & = & \frac{1}{2} \quad \frac{3}{4} \cdots \frac{2k_1-1}{2k_1} \quad \frac{1}{2} \quad \frac{3}{4} \quad \cdots \quad \frac{2k_2-1}{2k_2} \\ & \leqslant & \frac{1}{2} \quad \frac{3}{4} \cdots \frac{2k_1-1}{2k_1} \quad \frac{2k_1+1}{2k_1+2} \quad \frac{2k_1+3}{2k_1+4} \quad \cdots \quad \frac{2K-1}{2K} \\ & = & \mathcal{KT}\left(0^{K}\right). \end{array}$$

Moreover :

$$-\log \mathcal{KT}\left(0^{K}\right) = \sum_{i=1}^{K} -\log \frac{2i-1}{2i} \geqslant \sum_{i=1}^{K} \frac{1}{2i} \geqslant \frac{1}{2} \log K.$$

■

Chapitre 4

Consistency of the unlimited BIC Context Tree Estimator

4.1 Introduction

Modeling discrete data from a finite alphabet A by Markov chains suffers from a major disadvantage : the dimension of these models grows exponentially with the order, so that very few models (compared to the data length) are effectively available. Offering much more flexibility, *Variable Length Markov Chains* (VLMC) [BW99], also called *context tree sources* ([Ris83, CT06, Gar05a]), have become popular tools for data compression : they allow the probability distribution of the next symbol to depend on a *varying* number of predecessors. In other words, the conditional probabilities are determined by substrings called *contexts* which may be of different sizes. A formal definition is given in Section 4.2. It turns out that VLMC can be efficiently used for data compression. In particular, thanks to arithmetic coding [Ris76] and after the pioneering work [Ris83] of Rissanen in 1983 in which algorithm `context` was introduced, a family of universal coders has been developed on the same architecture, see e.g. [WT95, BW99].

In their article [CT06], Imre Csiszár and Zsolt Talata take an MDL (Minimum Description Length, [BRY98]) point of view : they choose $\widehat{T_n}$ as the context-tree model minimizing (over some family $\mathcal{F}_n$) the "ideal code length" which can be either the BIC criterion :

$$\widehat{T_{BIC}}(x_1^n) = \arg\min_{T \in \mathcal{F}_n} \min_{p=(p_s)_{s\in T}} -\log_2 P_{T,p}(x_1^n) + \frac{|T|\,(|A|-1)}{2}\log_2 n$$

or the $\mathcal{KT}$ criterion :

$$\widehat{T_{\mathcal{KT}}}(x_1^n) = \arg\min_{T\in\mathcal{F}_n} -\log_2 \widehat{\mathcal{KT}_T}(x_1^n).$$

Thee definition of these estimators will be clarified in Section 4.2, Equations 4.4–4.6.

In the case where x_1^n is a realization of a probability source P that is some variable length Markov chain in the model of T_0, they prove that both $\widehat{T_{BIC}}$ and $\widehat{T_{KT}}$ are strongly consistent estimators of T_0 as soon as the minimization is limited to "small" trees, i.e. $\mathcal{F}_n = \{T : \text{depth}(T) < D(n)\}$ for some function $D(n) = o(\log_2(n))$. As an example in [CS00] shows, this restriction is necessary for $\widehat{T_{KT}}$. On the other hand, it was known from [CS00] that if only Markovian models (i.e. full trees of finite depth) are considered, then this restriction is not necessary for $\widehat{T_{BIC}}$.

We prove here that the same happens for the whole class of context tree sources : one may minimize the *BIC* criterion on all models without losing the strong consistency. Note that contrary to [CT06], we do not consider in this chapter the case of infinite context trees. In [CT06], a linear-time algorithm was provided when the minimization is restricted to trees having depth smaller than $o(\log_2 n)$. We propose here a modification of this algorithm working without this restriction. In one step, it uses the same principle as [CT06], called *context tree maximizing*, first presented in an old version of [WT95].

The interest we see for this result is twofold : first, it is the answer to a natural question that remained open in [CT06]. The second point may be more important : it shows that the resulting "unlimited" algorithm behaves as well on short-memory processes like context-tree sources, while it may have better performance on long-memory processes (see [Gar05a]) for which the use of large contexts is crucial.

The structure of this chapter is the following : in section 4.2, we introduce formal notations and definitions, and formulate the strong consistency result about the consistency of the unbounded BIC estimator. Section 4.3 contains the proof of this result. Section 4.4 provides an algorithm computing the BIC context tree estimator with a linear time complexity. Section 4.5 contains some remarks and conclusions.

4.2 The consistency result

Let A be a finite alphabet. In this chapter, we denote by $|A|$ the cardinality A. If $x \in A^n$, let $|x| = n$ and $x_i^j = x_i x_{i+1} \dots x_j \in A^{j-i+1}$. The empty string, denoted by $\emptyset$, is the only element of A^0. The set of all strings on alphabet A is $A^* = \bigcup_{n\in\mathbb{N}} A^n$, the set of all non-empty strings on A is $A^+ = \bigcup_{n\in\mathbb{N}^+} A^n$ We denote by "." the concatenation operator so that for $1 \leqslant i \leqslant n$, $x = x_1^i . x_{i+1}^n = \bigodot_{i=1}^n x_i$. We say that $v \in A^*$ is a *suffix* (resp. *proper suffix*) of x if there exists $u \in A^*$ (resp. $u \in A^+$) such that $x = uv$. The natural logarithm is denoted by ln, the binary logarithm by log, and $\log^+(x) = \max(\log x, 0)$.

Let n be a positive integer, $y \in A^n$ a finite string, and denote by $N_a(y) = \sum_{i=1}^n \mathbb{1}_{y_i=a}$ the number of occurrences of letter $a \in A$ in y. Let also Θ_A be the simplex $\left\{\theta = (\theta_a)_{a\in A} \middle| \forall a \in A, 0 \leqslant \theta_a \leqslant 1 \text{ and } \sum_{a\in A} \theta_a = 1\right\}$, and for $\theta \in \Theta_A$ let P_θ be the probability distribution over A defined by $\forall a \in A, P_\theta(a) = \theta_a$; θ naturally defines a memoryless distribution P_θ on $A^{\mathbb{Z}}$ satisfying $P_\theta(y) = \prod_{i=1}^n P_\theta(y_i) = \prod_{a\in A} P_\theta(a)^{N_a(y)}$. Information theory helps us interpret

$$-\log_2 \hat{P}(y) = -\log_2 \prod_{a\in A} \left(\frac{N_a(y)}{n}\right)^{N_a(y)}$$

as an *oracle memoryless code length* for the string y. As it is a function of the counts $(N_a(y))_{a\in A}$, we shall also denote it $-\log_2 \hat{P}((N_a(y))_{a\in A})$ with some abuse of notation. Moreover, the Krichevski-Trofimov mixture [KT81] is defined as the mixture of all i.i.d. distributions P_θ, with respect to the Dirichlet distribution ν with parameter 1/2 :

$$\mathcal{KT}(y) = \int_{\theta\in\Theta_A} P_\theta(y)\, \nu(\mathrm{d}\theta). \tag{4.1}$$

Some calculation gives an explicit, easily computable expression :

$$\mathcal{KT}(y) = \frac{\prod_{a:N_a(y)\geqslant 1} \left(N_a(y) - \frac{1}{2}\right)\left(N_a(y) - \frac{3}{2}\right)\dots\frac{1}{2}}{\left(n - 1 + \frac{|A|}{2}\right)\left(n - 2 + \frac{|A|}{2}\right)\dots\frac{|A|}{2}}. \tag{4.2}$$

It has the remarkable property to uniformly (and asymptotically essentially optimally) approach all memoryless distributions on A^n (see the proof of Lemma 5 in the Appendix). As a matter of fact, $-\log_2 \mathcal{KT}(y)$ is interpreted in information theory as a *universal memoryless code length* for y, and the difference $-\log_2 \mathcal{KT}(y) + \log_2 \hat{P}(y) - \frac{|A|-1}{2}\log_2 n$ is bounded independently of n and y.

A set of strings T on alphabet A can be represented by a *trie*, i.e. a tree with edges labeled by letters of A such that every string of T corresponds to a path from a node to the root. If no word $s_1 \in T$ is a suffix of another $s_2 \in T$, then T is said to have the *tree property* : then the leaves of the trie are exactly the elements of T, and its internal nodes correspond to proper suffixes of the elements of T. In particular, the root corresponds to the empty string $\emptyset$. If T is a tree such that every semi-infinite word has (exactly one) suffix in T, it is called a *context tree* (or sometimes a *complete suffix dictionary*). For a semi-infinite string $x = x_{-\infty}^n$ on alphabet A and a given a word $s \in T$, symbol x_i is said to appear *in context* s if the suffix of $x_{-\infty}^{i-1}$ that belongs to T is s. We denote by $\mathcal{T}(s, x)$ the sequence of symbols in x_1^n appearing in context s.

For $l \in \mathbb{N}$, a string $s \in A^l$ is a *context* of a stationary process P if

1. $P(X_1^l = s) > 0$,

2. for $k \in \mathbb{N}$ and any $w \in A^k$

$$P\left(X_1 = a | X^0_{-k-l+1} = w.s\right) = P\left(X_1 = a | X^0_{-l+1} = s\right),$$

3. and if no proper suffix of s has the previous property.

A context-tree source $P_{T,p}$ is a stationary ergodic stochastic process on A defined by

- a *suffix function* T mapping every left semi-infinite string to a context of $P_{T,p}$. As there will be no ambiguity, we also denote by T the range of the suffix function, i.e. the set of all contexts of $P_{T,p}$. Note that it is a context tree.
- a family $p = \left(p_s\right)_{s \in T}$ of probability distributions defining the conditional distribution after each context. Formally, the probability of seeing symbol $a \in A$ after any semi-infinite string $x^0_{-\infty}$ ending by $s = T(x^0_{-\infty})$ is $p_s(a)$:

$$P_{T,p}\left(X_1 = a | X^0_{-\infty} = x^0_{-\infty}\right) = p_{T(x^0_{-\infty})}(a). \tag{4.3}$$

In words, a context-tree source is a variable-length Markov chain, the memory of which is allowed to vary with the past. At each moment i, the suffix function T gives the shortest suffix $x^i_{-\infty}$ is necessary to determine the distribution of the next symbol $p\left(x_{i+1} | T\left(x^i_{-\infty}\right)\right)$. This relevant suffix of the past is the context of x_{i+1}.

Now let T be a context tree. In order to code the finite string x_1^n, to determine its log-probability under some context tree distribution or to compute a maximum likelihood, there is a small difficulty at the beginning due to the fact that one has no context for some initial symbols, see [Wil94, Cat01]. Several solutions are possible : one may either consider a stationary source, or compute a modified likelihood with a distribution given a priori for the first symbols, or consider that fake initial data is available to both the coder and the decoder. So few characters are implied that the solution chosen does not have any impact on the consistency results, especially when they are asymptotic as in [CT06] and in Theorem 24. In the algorithm of Section 4.4, we choose to add an initial "start" symbol before x. This amounts to computing a modified likelihood with special contexts for the very first symbols. For the ease of the theoretical introduction that follows now and in Section 4.3, we suppose that x_1^n is preceeded by a semi-infinite substring $x^0_{-\infty}$ known by everyone, so that the context of all symbols in x_1^n can be determined.

Recall that for $s \in T$, we write $T(s, x)$ for the (non-contiguous) subsequence of x appearing in context s, so that equation (4.3) implies :

$$\begin{aligned}\hat{P}_T(x_1^n) &= \prod_{s\in T} \hat{P}\left(\mathcal{T}(s,x)\right) \quad \text{, and} && (4.4)\\ \mathcal{KT}_T(x_1^n) &= \prod_{s\in T} \mathcal{KT}\left(\mathcal{T}(s,x)\right). && (4.5)\end{aligned}$$

Hence

$$-\log_2 \hat{P}_T(x_1^n) = \sum_{s\in T} -\log_2 \hat{P}\left(\mathcal{T}(s,x)\right)$$

(resp. $-\log_2 \mathcal{KT}_T(x_1^n) = \sum_{s\in T} -\log_2 \mathcal{KT}\left(\mathcal{T}(s,x)\right)$), and one can say that the oracle (resp. universal) code length of x in the model defined by T is the sum over its leaves s of the memoryless oracle (resp. universal) code lengths of the subsequences $\mathcal{T}(s,x)$. The BIC context tree estimator $\widehat{T_{BIC}}(x_1^n)$ is defined as a minimizer of the penalized oracle code length, i.e. of the BIC criterion :

$$BIC_T(x_1^n) = -\log_2 \hat{P}_T(x_1^n) + \frac{|T|\left(|A|-1\right)}{2}\log_2 n. \tag{4.6}$$

Note that from Lemma 5, the penalty term is the asymptotic value (up to a constant) of the difference between the universal code length $-\log_2 \mathcal{KT}_T(x)$ and the oracle code length $-\log_2 \hat{P}_T(x)$. It is also the minimax redundancy rate of any coder universal in the class of context tree sources that belongs to the model defined by T, see [Ris86].

The following theorem is the main result of this chapter. It extends Theorem 2.6 in [CT06] for finite context trees by removing the restriction on the hypothetical tree depths in the minimization of the BIC criterion.

Theorem 24. *Let T_0 be a context tree and P_0 a context tree source in the model defined by T_0. If X is a stationary process with distribution P_0, then*

$$\widehat{T_{BIC}}(X_1^n) = T_0$$

eventually almost surely as $n \to \infty$.

The proof of Theorem 24 follows; then, section 4.4 provides an algorithm computing the unbounded BIC estimation with a linear time complexity.

4.3 Proof of Theorem 24

Theorem 2.6 in [CT06] implies that if $D(n) = o(\log_2 n)$, then

$$\underset{\text{depth}(T)<D(n)}{\arg\min}\; BIC_T(X) = T_0$$

eventually almost surely as $n \to \infty$. In words, if the criterion is minimized only on "not too deep" models, then the resulting estimator is consistent. Hence, it is sufficient to prove that deep models are never chosen after a certain time. We prove a stronger statement : namely that only "not too big" models are eventually chosen. More precisely, let $k_n = \left\lceil \frac{\log_2 n}{\log_2 \log_2 \log_2 n} \right\rceil$; we show that eventually, P_0-almost surely :

$$|\widehat{T_{BIC}}(X_1^n)| \leqslant k_n.$$

This will lead to the conclusion, as any context tree T satisfies $|T| > \text{depth}(T)$. Note that the general frame of this proof is inspired from Proposition 2 in [CS00].

For a given context tree T, let $B_T^n = \left\{ x \in A^n : \widehat{T_{BIC}}(x_1^n) = T \right\}$. Using similar reasoning as in Barron's Lemma on the competitive optimality of the Shannon code [Bar85] helps us control the probability of this event :

$$\begin{aligned}
P_0\left(X_1^n \in B_T^n\right) &\leqslant P_0\Big(-\log_2 P_0(X_1^n) + \frac{|T_0|\left(|A|-1\right)}{2} \log_2 n \\
&\qquad\qquad \geqslant -\log_2 \hat{P}_T(X_1^n) + \frac{|T|\left(|A|-1\right)}{2} \log_2 n \Big) &\quad (4.7)\\
&= \sum_{x \in A^n} P_0(X_1^n = x) \mathbb{1}_{\left\{\log_2 P_0(x) \leqslant \log_2 \hat{P}_T(x) + \frac{(|T_0|-|T|)(|A|-1)}{2} \log_2 n \right\}} &\quad (4.8)\\
&\leqslant \sum_{x \in A^n} P_0(X_1^n = x) \mathbb{1}_{\left\{\log_2 P_0(x) \leqslant \log_2 \mathcal{KT}_T(x) + \frac{(|T_0|-|T|)(|A|-1)}{2} \log_2 n + \frac{|T|(|A|-1)}{2} \log_2 \frac{n}{|T|} + C|T|\right\}} \\
&\leqslant \sum_{x \in A^n} \mathcal{KT}_T(x) 2^{\frac{|T_0|(|A|-1)}{2} \log_2 n + C|T| - \frac{|T|(|A|-1)}{2} \log_2 |T|} &\quad (4.9)\\
&\leqslant 2^{\frac{|T_0|(|A|-1)}{2} \log_2 n - |T|\left(-C + \frac{(|A|-1)}{2} \log_2 |T|\right)}, &\quad (4.10)
\end{aligned}$$

where (4.9) comes from Lemma 5 proved in the Appendix, and (4.10) from the fact that $\mathcal{KT}_T$ is a probability measure on A^n. It follows that :

$$\begin{aligned}
\sum_{1+k_n \leqslant |T| < n} 2^{-|T|\left(-C + \frac{(|A|-1)}{2} \log_2 |T|\right)} &= \sum_{t=1+k_n}^{n-1} |\mathcal{T}_t| 2^{-t\left(-C + \frac{(|A|-1)}{2} \log_2 t\right)} \\
&\leqslant \sum_{t=1+k_n}^{\infty} 16^t 2^{-t\left(-C + \frac{(|A|-1)}{2} \log_2 t\right)} &\quad (4.11)\\
&= \sum_{t=1+k_n}^{\infty} 2^{-t\left(-C - 4 + \frac{(|A|-1)}{2} \log_2 t\right)} \\
&\leqslant 2^{-k_n\left(-(C+4) + \frac{(|A|-1)}{2} \log_2 k_n\right)} &\quad (4.12)
\end{aligned}$$

for n large enough.

We used some tree counting arguments proved in Lemma 6 to upper bound in (4.11) the cardinal of the set $\mathcal{T}_t$ of context trees of size t by 16^t. Inequality (4.12) comes from the upper-bound :

$$\begin{aligned}
\sum_{t=L+1} 2^{-t\left(-C-4+\frac{(|A|-1)}{2}\log_2 t\right)} &\leqslant \int_L^\infty e^{-x\left(-(C+4)\ln 2+\frac{(|A|-1)}{2}\ln x\right)}\mathrm{d}x \\
&\leqslant \int_L^\infty \left(-(C+4)\ln 2+\frac{(|A|-1)}{2}(1+\ln x)\right) \\
&\qquad \times e^{-x\left(-(C+4)\ln 2+\frac{(|A|-1)}{2}\ln x\right)}\mathrm{d}x \\
&= e^{-L\left(-(C+4)\ln 2+\frac{(|A|-1)}{2}\ln L\right)} \\
&= 2^{-L\left(-(C+4)+\frac{(|A|-1)}{2}\log_2 L\right)},
\end{aligned}$$

valid as soon as the integer L is large enough to ensure inequality : $-(C+4)\ln 2+\frac{(|A|-1)}{2}(1+\ln L)\geqslant 1$.

Now, $\widehat{T_{BIC}}(x_1^n)$ has at most one leave for each context present in x_1^n : more contexts cannot improve the likelihood while they would increase the penalty term in the BIC criterion. Hence (4.10) implies :

$$\begin{aligned}
P_0\left(\bigcup_{|T|>k_n} B_T^n\right) &\leqslant \sum_{1+k_n\leqslant|T|<n} P_0\left(B_T^n\right) \\
&\leqslant 2^{\frac{|T_0|(|A|-1)}{2}\log_2 n} \sum_{1+k_n\leqslant|T|<n} 2^{-|T|\left(-C+\frac{(|A|-1)}{2}\log_2|T|\right)} \\
&\leqslant 2^{\frac{|T_0|(|A|-1)}{2}\log_2 n-\frac{\log_2 n}{\log_2\log_2\log_2 n}\left(-C-4+\frac{(|A|-1)}{2}\log_2\left(\frac{\log_2 n}{\log_2\log_2\log_2 n}\right)\right)} \\
&= n^{\frac{|T_0|(|A|-1)}{2}-\frac{\left(-(C+4)+\frac{(|A|-1)}{2}(\log_2\log_2 n-\log_2\log_2\log_2\log_2 n)\right)}{\log_2\log_2\log_2 n}} \\
&\leqslant n^{-2} \qquad \text{for n large enough.}
\end{aligned}$$

The Borel-Cantelli lemma thus implies that eventually almost-surely, $\widehat{T_{BIC}}\left(X_1^n\right)$ has at most k_n contexts.

4.4 An algorithm for the BIC Context Tree estimator

In [CT06], Imre Csiszár and Zsolt Talata give a linear time algorithm to compute the BIC and KT context tree estimators when the minimization is restricted to trees whose depths is smaller than $o(\log_2 n)$. As stated by Theorem 24, and contrary to the case of KT, this restriction is not necessary for finite context trees in the case of the BIC estimator. Hence, the question naturally arises whether minimizing the BIC criterion among *all* context trees can still be done in linear time. This does not appear as an immediate adaptation of the algorithm given in [CT06] ; in particular, the algorithmical notion of *compact suffix tree* needs to be introduced. The interest of this algorithm is discussed in the conclusion. In short : it is more "natural" to remove an unnecessary condition, Theorem 24 shows that it is asymptotically as good as any restricted version for finite context trees, while one can conjecture that it is more efficient for infinite context tree sources.

Actually, the computation of $\widehat{T_{BIC}}(x_1^n)$ can be done efficiently, in linear time, by a bottom up recursive algorithm. Note that this algorithm can be adapted to the case of $\widehat{T_{KT}}(x_1^n)$, but this is less interesting since the corresponding (unlimited) estimator is not always consistent.

- First build the *compact context tree* $\mathcal{CT}(x_1^n)$, defined as the compact suffix tree of the string obtained by processing $@x_1^{n-1}$ from right to left (where @ is a character that does not belong to A). We identify each node s of $\mathcal{CT}(x)$ with the string obtained by reading the labels from s to the root of $\mathcal{CT}(x)$. Thus, every leaf of $\mathcal{CT}(x)$ is identified with a prefix of $@x$. More generally, a context s followed in $@x$ by a least two different characters is represented by an internal node $n_s \in \mathcal{CT}(x)$. Moreover, $\mathcal{CT}_x(s, x)$ denotes the concatenation of all symbols in string x appearing in context s.
  We denote by $\mathcal{CT}(x|s)$ the subtree of $\mathcal{CT}(x)$ rooted in s. The compactness ensures that every internal node of $\mathcal{CT}(x)$ has an arity between 2 and |A|+1. Each node s of the tree is supplied with :
  - a counter `counts` $\in \mathbb{N}^{|A|}$ that will indicate the number of occurrences of each letter $a \in A$ in context s. At this step, all counters are set to zero except that of the leaves, which can be set corectly within the construction of $\mathcal{CT}(x)$, and which hence contain one '1' and $|A| - 1$ '0'.
  - two real numbers `selfcost` and `subcost`, all initialized to 0.

  The resulting tree can be proven to be of size $O(n)$. Moreover, by the efficient algorithms described for instance in [GK97], it can be constructed in linear time.
- Then, we can use the Context Tree Maximizing algorithm, as described for instance in [CT06], to find the tree minimizing the criterion. It is presented here in a slightly unusual way that, in the author's opinion, offers an interesting insight of its greedy mechanism. $\widehat{T_{BIC}}(x_1^n)$ can be computed by processing $\mathcal{CT}(x_1^n)$ from bottom to top, doing for each node s the following operations :
  - update the parent's counts, increasing them by c.counts for every child c of node s.

– compute

$$\begin{cases} s.\texttt{selfcost} & = & -\log_2 \hat{P}(s.\texttt{counts}) + \frac{|A|-1}{2}\log_2 n \\ s.\texttt{subcost} & = & \sum_{t \text{ child of } s} \min(t.\texttt{selfcost}, t.\texttt{subcost}) \end{cases}$$

except for leaves which have an infinite subcost. We call *active* a node whose subcost is smaller than its selfcost. These computations take a constant time for each of the $O(n)$ nodes.

– Finally, select the tree T iteratively obtained at the end of the following process :

Step 1 - start with T equal to the root r of $\mathcal{CT}(x)$

Step 2 - if T contains an active node n, replace n in T by its children, and go back to Step 2;

Step 3 - when T does no longer contain any active node, stop. Note that the tree T selected by this procedure is the largest such that for every leaf $s \in T$, s is not active but all proper suffixes of s present in $\mathcal{CT}(x)$ are active.

Obviously, this process requires $O(n)$ operations. Taking $s = \emptyset$ in Lemma 7 shows that the selected tree minimizes the BIC criterion among all subtrees of $\mathcal{CT}(x)$. It remains to be proved that the BIC criterion is minimized by some subtree of $\mathcal{CT}(x)$. In fact, if T is any tree minimizing the BIC criterion then every leaf $s \in T$:

– either does not appear in $@x_1^{n-1}$, in which case $T(s,x) = \emptyset$;

– or there exists a string u such that $u.s \in \mathcal{CT}(x)$, and by construction $T(s,x) = \mathcal{CT}_x(u.s,x)$.

Hence, there is always $T' \subset \mathcal{CT}(x)$ with at most as many leaves as T such that $\hat{P}_{T'}(x) = \hat{P}_T(x)$, and as the penalty grows with $|T|$, $\widehat{T_{BIC}}(x_1^n)$ can be chosen as a subtree of $\mathcal{CT}(x)$.

4.5 Conclusion

As for Markov Chains (cf [CS00]), and contrary to the Krichevski-Trofimov MDL estimator, no bound on hypothetical tree depth (that is on the memory length) is required for the BIC context tree estimator in the case of finite context tree sources. Computing this estimator by minimizing the BIC criterion on all possible trees is still possible in linear time thanks to the concept of compact suffix tree. As former "restricted" BIC criterion minimizing algorithms, Theorem 24 shows that this algorithm will eventually almost surely identify the good tree on finite context tree sources, with the same time complexity as the restricted minimizing procedure.

Moreover, it might behave better on infinite memory sources. In fact, some processes, like the Renewal and Markov Renewal processes introduced by Imre Csiszâr and Paul C. Shields in [CS96], need context trees much deeper than $O(\log_2 n)$ to be correctly approximated by context trees, see [Gar05a]. Hence, we may believe that compression algorithms based on an unbounded tree estimation, while keeping as good results on short memory processes and a linear time complexity, can have a much better behaviour on these sources. However, the typical size of selected trees on infinite depth context trees remains to be studied, and this seems to be a very challenging task.

4.6 Proof of the technical lemmas

Lemma 5. *There is a constant C depending only on the alphabet size $|A|$ such that :*

$$\forall n \geqslant 1, \forall x \in A^n, -\log_2 \mathcal{KT}_T(x) \leqslant -\log_2 \hat{P}_T(x) + \frac{|A|-1}{2}|T| \log_2^+ \frac{n}{|T|} + C|T|$$

For an explicit constant C see [Cat01].

Proof. This result is proved in [CS00] for Markov chains. The adaption to the case of context tree sources is straightforward, we give it here for the sake of self-containment. It is well-known (see [KT81]) that Stirling's formula implies the existence of K depending only on $|A|$ such that :

$$\forall k \in \mathbb{N}^*, \forall x \in A^k, -\log_2 \mathcal{KT}(x) \leqslant -\log_2 \hat{P}(x) + \frac{|A|-1}{2} \log_2 n + K.$$

Applying this inequality to subsequence $\mathcal{T}(s,x)$ of $x \in A^n$ appearing in context $s \in T$, we get :

$$\forall s \in T, \mathcal{T}(s,x) \neq \emptyset \implies -\log_2 \mathcal{KT}\left(\mathcal{T}(s,x)\right) \leqslant -\log_2 \hat{P}\left(\mathcal{T}(s,x)\right) + \frac{|A|-1}{2} \log_2 |\mathcal{T}(s,x)| + K,$$

and thus

$$\begin{aligned}
-\log_2 \mathcal{KT}_T(x) &= \sum_{s \in T, \mathcal{T}(s,x) \neq \emptyset} -\log_2 \mathcal{KT}\left(\mathcal{T}(s,x)\right) \\
&\leqslant \sum_{s \in T, \mathcal{T}(s,x) \neq \emptyset} -\log_2 \hat{P}\left(\mathcal{T}(s,x)\right) + \frac{|A|-1}{2} \log_2 |\mathcal{T}(s,x)| + K \\
&\leqslant -\log_2 \hat{P}_T(x) + K|T| + \frac{|A|-1}{2} \sum_{s \in T, \mathcal{T}(s,x) \neq \emptyset} \log_2 |\mathcal{T}(s,x)| .
\end{aligned}$$

The function

$$f : t \rightarrow \begin{cases} t-1 & \text{if} \quad 0 \leqslant t \leqslant 1 \\ \log_2 t & \text{if} \quad t > 1 \end{cases}$$

is concave, hence we have :

$$\begin{aligned} \sum_{s \in T, \mathcal{T}(s,x) \neq \emptyset} \log_2 |\mathcal{T}(s,x)| &\leqslant |T| + |T| \sum_{s \in T} \frac{1}{|T|} f\left(|\mathcal{T}(s,x)|\right) \\ &\leqslant |T| + |T| f\left(\frac{n}{|T|}\right) \\ &\leqslant |T| + |T| \log_2^+ \frac{n}{|T|}. \end{aligned}$$

Hence Lemma 5 holds with $C = K + \frac{|A|-1}{2}$.

■

We give here the inductive definition of a general ordered tree (GOT) [HU79]. A GOT is either :

- a structure *leaf* ;
- or a structure *internal node* consisting of a list L of GOTs. The elements of L are called the nodes *children*, their number is called the node's *arity*.

A binary tree is a GOT such that every node has at most 2 children. It is *complete* if every node has either no child (it is a leaf) or exactly 2 children : then the first one is called *left* child and the second one is called *right* child.

Lemma 6. *The number of general ordered trees with t leaves containing no node of arity 1 is upper-bounded by* 16^t.

Proof. The "first son, next brother" mapping Ψ (see e.g. [Ans]) injectively maps the set of general ordered, not necessarily binary trees with n nodes to the set of uncomplete binary trees with n nodes : each node N in the ordered tree corresponds to a node N' in the binary tree ; the left child of N' is the node corresponding to the first child of N, and the right child of N' is the node corresponding to N's next brother – that is, the node following N in the list the parent of N.

Now, let T be any GOT with t leaves, and without node of arity 1. An immediate induction shows that it has at most $i \leqslant t-1$ internal nodes (with equality iff it is complete binary), and thus at most $2t-1$ nodes. Hence, by adding an arbitrary tree with $t-1-i$ nodes as a right child of $\Psi(T)$, T can be injectively associated a binary tree $\tilde{T}$ with exactly $2t-1$ nodes.

$\tilde{T}$ can in turn bijectively be associated a complete binary tree with $2t-1$ internal nodes, just adding children to all its nodes with arity smaller than 2. Hence, there is an injective application mapping the set of GOT with t leaves and without arity 1 node to the set B_n of complete binary trees with $n = 2t-1$ internal nodes. This injection is illustrated by an example in Figure 4.1. The

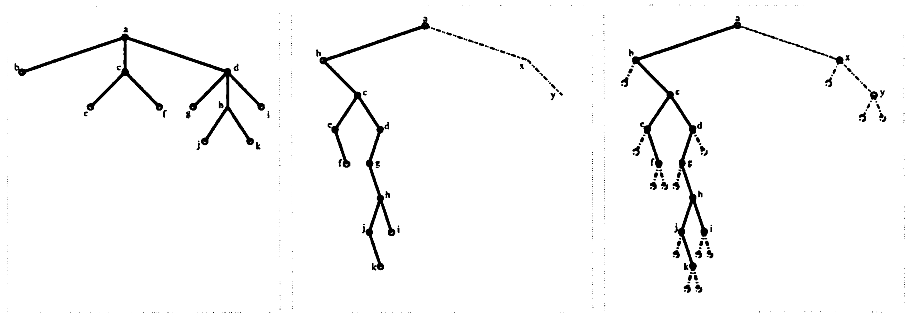


FIG. 4.1 – From a context tree with $t = 7$ leaves to a complete binary tree with $2t-1 = 13$ internal nodes (nodes x and y are added in order that the binary tree would have exactly $2t-1$ internal nodes).

conclusion follows since the cardinality of B_n is equal to the n^{th} Catalan number $C_n = \frac{1}{n+1}\binom{2n}{n} \leqslant 4^n \leqslant 16^t$. ■

Lemma 7. *For every node s of $\mathcal{CT}(x)$, if $y = \mathcal{CT}_x(s, x)$ then*

$$\min\left\{s.\mathtt{selfcost},\, s.\mathtt{subcost}\right\} = \min_{T\subset\mathcal{CT}(x|s)} -\log_2 \hat{P}_T(y) + \frac{|T|\left(|A|-1\right)}{2}\log_2 n.$$

Proof. This lemma is very similar to Proposition 3.9 in [CT06], itself inspired from early versions of [WT95]. We proceed similarly by induction on the size of the subtree $\mathcal{CT}(x|s)$.

If $|\mathcal{CT}(x|s)| = 1$, s is a leaf and the statement is obvious. Otherwise, for each $a \in A \cup \{@\}$ there is at most one (possibly empty) string w_a such that $w_a.a$ is a child of s in $\mathcal{CT}(x|s)$. Thus,

$$\begin{aligned}
s.\mathtt{subcost} &= \sum_{a\in A} \min\left\{s.\mathtt{selfcost},\, s.\mathtt{subcost}\right\}\\
&= \sum_{a\in A} \min_{T_a\subset\mathcal{CT}(x|w_a.a.s)} -\log_2 \hat{P}_{T_a}(y) + \frac{|T_a|\left(|A|-1\right)}{2}\log_2 n\\
&= \min_{T\subset\mathcal{CT}(x|s),\mathrm{depth}(T)\geqslant 1} -\log_2 \hat{P}_{T_a}(y) + \frac{|T|\left(|A|-1\right)}{2}\log_2 n.
\end{aligned}$$

The first equality follows from the induction hypothesis, as $\mathcal{CT}(x|w_a.a.s) \subsetneq \mathcal{CT}(x|s)$. Moreover, any family of subtrees $(T_a)_{a\in A}$ uniquely corresponds to a subtree $\tilde{T}$ of $\mathcal{CT}(x|s)$ rooted in s with depth at least 1. The second equality follows, as the size of $\tilde{T}$ is $\sum_{a\in A}|T_a|$. Hence,

$$\begin{aligned}\min\left\{\mathtt{S.selfcost}, \mathtt{S.subcost}\right\} &= \min\Bigg\{ -\log_2 \hat{P}_{\emptyset}(y) + \frac{|A|-1}{2}\log_2 n, \\ &\qquad \min_{T\subset \mathcal{CT}(x|s), \mathrm{depth}(T)\geqslant 1} -\log_2 \hat{P}_T(y) + \frac{|T|\,(|A|-1)}{2}\log_2 n\Bigg\} \\ &= \min_{T\subset \mathcal{CT}(x|s)} -\log_2 \hat{P}_T(y) + \frac{|T|\,(|A|-1)}{2}\log_2 n.\end{aligned}$$

■

Chapitre 5

A MDL approach to HMM with Poisson and Gaussian emissions. Application to order identification

5.1 Introduction

Formally introduced as *probabilistic functions of Markov Chains* in 1966 by Baum & Petrie, hidden Markov models (HMM) have known since then a growing interest as they proved useful in various applications, from speech recognition [LRS83] to blind deconvolution of unknown communication channels [KV94], bioinformatics [Kos01] or meteorology [HG94]. For a mathematical survey on HMM, see [EM02].

In most practical cases, the *order* of the model (*ie* the true number of hidden states) is unknown and has to be estimated. For that purpose, two approaches have been proposed : penalized maximum likelihood estimators as in [Fin91, Kie93] and Bayesian procedures as in [LN94]. As Gassiat & Boucheron [GB03] did for finite emission alphabet HMM, we follow the methodologies of Finesso and Liu & Narayan for some instances of infinite (continuous and discrete) emission alphabets. Following the work of these authors, we address the issue of order identification for HMM with Poisson and Gaussian emissions : we prove MDL-inspired mixture inequalities which lead to consistent penalized estimators requiring no prior bound on the order nor on the parameters of the mixture components.

In 1978, Rissanen introduced the Minimum Desription Length (MDL) principle with motto :

> *"Choose the model that gives the shortest description of data."*

More precisely, given any k-dimensional model (*ie* parametric family of densities indexed by Θ of dimension $k \geqslant 1$) :

$$\mathcal{M} = \{g_\theta : \theta \in \Theta\},$$

let E_θ be the expectation with respect to a random variable X_1^n with distribution P_θ, which density is g_θ (with respect to Lebesgue measure). For any density q such that $q(x_1^n) = 0$ implies $g_\theta(x_1^n) = 0$, the Kullback-Leibler divergence between g_θ and q is

$$K_n(g_\theta, q) = E_\theta \log \frac{g_\theta(X_1^n)}{q(X_1^n)} = E_\theta \left[- \log q(X_1^n) - (- \log g_\theta(X_1^n))\right].$$

In Information theory, $- \log q(X_1^n)$ is interpreted as the code length for X_1^n when using coding distribution q, so $E_\theta[- \log g_\theta(X_1^n)]$ is the *ideal code length* for X_1^n. In this perspective, $K_n(g_\theta, q)$ is the average additionnal cost (or *redundancy*) for compressing some source in $\mathcal{M}$ without knowing which one.

When assuming that the maximum likelihood estimator $\widehat{\theta}(x_1^n)$ achieves a $\sqrt{n}$-rate and that the distribution of $\widehat{\theta}(X_1^n)$ has uniformly summable tail probabilities : there exists a summable sequence $\{\delta_n\}$ of positive numbers such that, for every $\theta \in \Theta$,

$$P_\theta \left\{ \sqrt{n} \left\| \widehat{\theta}(X^n) - \theta \right\| \geqslant \log n \right\} \leqslant \delta_n,$$

Rissanen proved [Ris86] that

$$\liminf_{n \to \infty} \frac{K_n(g_\theta, q)}{\frac{k}{2} \log n} \geqslant 1 \tag{5.1}$$

for all $\theta \in \Theta$ except on a set with Lebesgue measure 0 (that depends on q and k). Here, k is the dimension of the parameter space Θ. This result has a minimax counterpart for i.i.d sequences [CB90] : under mild assumptions,

$$K_n^* = \min_q \sup_{\theta \in \Theta} K_n(g_\theta, q) \geqslant \frac{k}{2} \log \frac{n}{2\pi e} + O(1). \tag{5.2}$$

Both (5.1) and (5.2) put forward a leading term $\frac{k}{2} \log n$ that has taken a great importance in Information theory and Statistics. The coding density q is said optimal if it achieves equality in inequation (5.1). Three optimal coding distributions are often encountered in Information theory (we refer to [BRY98, HY01] for surveys) :

- two-stage coding, that yields description length

$$- \log q(x_1^n) = - \log g_{\widehat{\theta}(x_1^n)}(x_1^n) + \frac{k}{2} \log n;$$

- mixture coding, where q is a mixture of all densities g_θ $(\theta \in \Theta)$;
- predictive coding, where q is based on an iterative prediction scheme of x_j given x_1^{j-1} $(j = 1, \dots, n)$, namely $q_{\text{PDL}}(x_1^n) = \prod_{j=1}^n g_{\widehat{\theta}(x_1^{j-1})}(x_j)$.

We want to highlight that the quantity $-\log g_{\widehat{\theta}(x_1^n)}(x_1^n) + \frac{k}{2}\log n$, also called Bayesian Information Criterion (BIC), has been considerably studied since its first introduction in [Sch78] for purpose of model dimension estimation.

Now, let us consider the following problem : given a family of models $(\mathcal{M}_i)_{i \in I}$, which one best represents some given data x_1^n ? The MDL methodology suggests to choose model $\widehat{\mathcal{M}} = \mathcal{M}_{\widehat{i}}$ that yields the shortest description length of x_1^n.

Let k_i be the dimension of model $\mathcal{M}_i$ for every $i \in I$. Each of the three optimal coding distributions presented above selects a model :

- two-stage coding chooses
$$\widehat{\mathcal{M}}_{\text{BIC}} = \underset{\mathcal{M}_i\ (i \in I)}{\arg\min} \left\{ -\log g_{\widehat{\theta}_i(x_1^n)}(x_1^n) + \frac{k_i}{2}\log n \right\} ,$$
where $\widehat{\theta}_i$ is the maximum likelihood estimator over model $\mathcal{M}_i$;
- mixture coding chooses
$$\widehat{\mathcal{M}}_{\text{MIX}} = \underset{\mathcal{M}_i\ (i \in I)}{\arg\min} \left\{ -\log q_i(x_1^n) \right\} ,$$
where q_i is a particular mixture to be specified later – we will actually introduce a penalized version of this estimation procedure;
- predictive coding chooses
$$\widehat{\mathcal{M}}_{\text{PDL}} = \underset{\mathcal{M}_i\ (i \in I)}{\arg\min} \left\{ -\log q_{\text{PDL},i}(x_1^n) \right\} ,$$
where $q_{\text{PDL},i}$ is the predictive distribution relative to $\mathcal{M}_i$.

The challenging task is to prove that such estimators are consistent : if x_1^n is output by a source of density g_{θ_0} such that $g_{\theta_0} \in \mathcal{M}_{i_0}$ and $g_{\theta_0} \in \mathcal{M}_i$ implies $\mathcal{M}_{i_0} \subset \mathcal{M}_i$, then $\widehat{\mathcal{M}} = \mathcal{M}_{i_0}$ eventually almost surely. This has been successfully accomplished for Markov Chains by Csiszár & Shields [CS00], and for Context Tree Models (or Variable Length Markov Chains) by Csiszár & Talata [CT06] and Garivier [Gar05b].

Organization of the chapter

We start in Section 5.2 by stating and proving inequalities that compare BIC criterion and a particular mixture coding distribution (see Theorems 25 and 26). Four related models are involved, namely HMM mixture models and i.i.d models, with Poisson or Gaussian emissions. These inequalities are used in Section 5.3 for order identification purposes in the four models cited above. We notably obtain the consistency of two order estimators based on two-stage and mixture coding without assuming the existence of any prior bound on orders (see Theorems 27 and 28). We give a hint in Section 5.4 why order estimation based on raw (that is *not* penalized) predictive coding procedure cannot likely be proven consistent along the same lines than the two others (see Theorem 29). The proof of two lemmas and a useful result due to Leroux [Ler92b] are postponed to Appendix 5.5 and Appendix 5.6.

5.2 Mixture inequalities

Mixture inequalities for HMM mixture model

Let σ^2 be a positive number. The Gaussian density with mean m and variance σ^2 (with respect to the Lebesgue measure on the real line) is denoted by ϕ_{m,σ^2}. The Poisson density with mean m (with respect to the counting measure on the set of nonnegative integers) is denoted by π_m.

Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables with values in the measured space $(\mathcal{X}, \mathcal{A}, \mu)$ and defined on a measurable set upon which all random variables will be defined. Let us denote by $\{Z_n\}_{n\geqslant 0}$ a sequence of hidden random variables such that, conditionally on $Z_1^n = (Z_1, \dots, Z_n)$, $X_1, \dots, X_n$ are independent and the distribution of each X_i only depends on Z_i (all $i \leqslant n$).

For every $k \geqslant 1$, let $(p_j^o : j \leqslant k) \in \mathbb{R}_+^k$ be an initial distribution and let $\mathcal{S}_k$ be the set of all $\mathbf{p} = (p_{jj'} : j, j' \leqslant k) \in \mathbb{R}_+^{k^2}$ such that, for all $j \leqslant k$, $\sum_{j'=1}^k p_{jj'} = 1$, then the parameter set

$$\Theta_k = \left\{\theta = (\mathbf{p}, \mathbf{m}) : \mathbf{p} \in \mathcal{S}_k, \mathbf{m} = (m_1, \dots, m_k) \in \mathbb{R}^k\right\}.$$

Under parameter $\theta = (\mathbf{p}, \mathbf{m}) \in \Theta_k$ (some $k \geqslant 1$), $\{Z_n\}_{n\geqslant 0}$ is a Markov chain with values in $\{1, \dots, k\}$, initial distribution $P_\theta\{Z_0 = j'\} = p_{j'}^o$ and transition probabilities $P_\theta\{Z_{i+1} = j' | Z_i = j\} = p_{jj'}$ (all $j, j' \leqslant k$). Therefore, $\{X_n\}_{n\geqslant 1}$ is a HMM under parameter θ.

We shall consider two examples of emission distributions :

Gaussian emission (GE) For every $n \geqslant 1$, X_n has density $\phi_{m_{Z_n},\sigma^2}$ conditionally on Z_n.

Poisson emission (PE) For every $n \geqslant 1$, X_n has density $\pi_{m_{Z_n}}$ conditionally on Z_n.

For all parameter $\theta \in \Theta_k$ (any $k \geqslant 1$), let g_θ be the density of $X_1^n = (X_1, \dots, X_n)$ under θ. For every $k \geqslant 1$, let ν_k be a prior probability on Θ_k such that, for some chosen $\tau > 0$, under ν_k :

- $\mathbf{p}$ and $\mathbf{m}$ are independent,

- $p^o_{j'} = 1/k$ for all $j' \leqslant k$ are determinist,
- the vectors $(p_{jj'} : j' \leqslant k)$ $(j \leqslant k)$ are independently Dirichlet$(1/2, \dots, 1/2)$ distributed,
- $m_1, \dots, m_k$ are independent, identically distributed with density $\phi_{0,\tau}$ in example **GE** and with density Gamma$(\tau, 1/2)$ in example **PE**.

The related mixture statistics is defined by

$$q_k(X_1^n) = \int_{\Theta_k} g_\theta(X_1^n) d\nu_k(\theta). \tag{5.3}$$

It is worth noting that q_k is a positive function of $x_1^n \in \mathcal{X}^n$ in examples **GE** and **PE**.

The main results of this section are comparisons between the maximum log-likelihood and the mixture statistics in examples **GE** and **PE**.

Let $X_{(n)}$ and $|X|_{(n)}$ be the maxima of $X_1, \dots, X_n$ and $|X_1|, \dots, |X_n|$, respectively. Let us also introduce, for all $k, n \geqslant 1$,

$$\begin{aligned} c_{kn} &= \log k - k \log \frac{\Gamma(k/2)}{\Gamma(1/2)} + \frac{k^2(k-1)}{4n} + \frac{k}{12n}, \\ d_{kn} &= \frac{k}{2} \log\left(\frac{\tau^2}{k\sigma^2} + \frac{1}{n}\right), \\ e_{kn} &= \frac{k}{2}\Big(1 + \tau - \log(k\tau)\Big). \end{aligned}$$

Theorem 25 (HMM **mixture models**). *Under the assumptions described above, for every integers $k, n \geqslant 1$,*

GE

$$0 \leqslant \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leqslant \frac{k^2}{2} \log n + \frac{k}{2\tau^2} |X|^2_{(n)} + c_{kn} + d_{kn}. \tag{5.4}$$

PE

$$0 \leqslant \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leqslant \frac{k^2}{2} \log n + k\tau X_{(n)} + c_{kn} + e_{kn}. \tag{5.5}$$

Particular case of i.i.d mixture models

The i.i.d mixture model is a particular case of the HMM model. Here, the sequence $\{Z_n\}_{n \geqslant 0}$ is made of mutually independent random variables. In other words, for all $j, j' \leqslant k$, $p^o_{j'} = p_{jj'}$.

For every $k \geqslant 1$, let us introduce the set $\mathcal{S}'_k$ of all $\mathbf{p} = (p^o_j : j \leqslant k) \in \mathbb{R}^k_+$ such that $\sum_{j=1}^k p^o_j = 1$, then the parameter set

$$\Theta'_k = \left\{\theta = (\mathbf{p}, \mathbf{m}) : \mathbf{p} \in \mathcal{S}'_k, \mathbf{m} = (m_1, \dots, m_k) \in \mathbb{R}^k\right\}.$$

Again, g_θ is the density of X_1^n under parameter $\theta \in \Theta'_k$. For every $k \geqslant 1$, a new mixing probability ν_k on Θ'_k is chosen such that, under ν'_k :

- $\mathbf{p}$ and $\mathbf{m}$ are independent,
- $\mathbf{p}$ is Dirichlet$(1/2, \ldots, 1/2)$ distributed,
- $m_1, \ldots, m_k$ are independent, identically distributed with density $\phi_{0,\tau}$ in example **GE** and with density Gamma$(\tau, 1/2)$ in example **PE**.

Equality (5.3) defines a mixture statistics $q_k(X_1^n)$ in this framework. The main result of this section is another comparison between the maximum log-likelihood and the mixture statistics in examples **GE** and **PE**.

Let us introduce, for all $n, k \geqslant 1$,

$$c'_{kn} = -\log \frac{\Gamma(k/2)}{\Gamma(1/2)} + \frac{k(k-1)}{4n} + \frac{1}{12n}.$$

Theorem 26 (i.i.d mixture models). *Under the assumptions described above, for every integers $k, n \geqslant 1$,*

GE

$$0 \leqslant \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leqslant \frac{2k-1}{2} \log n + \frac{k}{2\tau^2} |X|^2_{(n)} + c'_{kn} + d_{kn}. \tag{5.6}$$

PE

$$0 \leqslant \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leqslant \frac{2k-1}{2} \log n + k\tau X_{(n)} + c'_{kn} + e_{kn}. \tag{5.7}$$

Comment

In inequations (5.4), (5.5), (5.6) and (5.7), the upper bounds write as a sum of $\frac{1}{2}\dim(\Theta_k) \log n$, a bounded term and a random term which involves the maximum of $|X_1|, \ldots, |X_n|$. The following lemmas guarantee that these random terms are bounded in probability at rate $\log n$ in example **GE** and slower than $\log n$ in example **PE**.

Lemma 8. *Let $\{Y_n\}_{n \geqslant 1}$ be a sequence of independent Gaussian random variables with variance σ^2. The mean of Y_n is denoted by m_n. If $\sup_{n \geqslant 1} |m_n|$ is finite, then for n large enough,*

$$P\left\{|Y|^2_{(n)} \geqslant 5\sigma^2 \log n\right\} \leqslant \frac{1}{n^{3/2}}.$$

Lemma 9. *Let $\{Y_n\}_{n \geqslant 1}$ be a sequence of independent Poisson random variables. The mean of Y_n is denoted by m_n. If $\sup_{n \geqslant 1} m_n$ is finite, then for n large enough,*

$$P\left\{Y_{(n)} \geqslant \frac{\log n}{\sqrt{\log \log n}}\right\} \leqslant \frac{1}{n^2}.$$

The proofs of Lemmas 8 and 9 are postponed to Section 5.5 of the Appendix.

Proof of Theorems 25 and 26

In the first place, let us introduce some notations.

For all $\theta \in \Theta_k$ (any $k \geqslant 1$) and for all $x_1^n \in \mathcal{X}^n$, $z_0^n = (z_0, \dots, z_n) \in \{1, \dots, k\}^{n+1}$, we denote by $g_\theta(x_1^n|z_1^n)$ the density of X_1^n at x_1^n conditionally on $Z_1^n = z_1^n$. The mixture density $q_k(x_1^n|z_1^n)$ at x_1^n conditionally on $Z_1^n = z_1^n$ is defined as in (5.3), with substitution of $g_\theta(x_1^n|z_1^n)$ to $g_\theta(X_1^n)$.

Similarly, we denote by $g_\theta(x_1^n|z_0)$ the density of X_1^n at x_1^n conditionally on $Z_0 = z_0$, and $q_k(\cdot|z_0)$ the corresponding conditional mixture density. Besides, if $P_\theta\{z_1^n|z_0\}$ is a shortcut for $P_\theta\{Z_1^n = z_1^n|Z_0 = z_0\}$, then the mixture density at z_1^n $q_k(z_1^n|z_0)$ is defined as in (5.3), with replacement of $g_\theta(X_1^n)$ by $P_\theta\{z_1^n|z_0\}$. Finally, for every $j \leqslant k$, let us set

$$n_j = \sum_{i=1}^n \mathbb{1}\{z_i = j\}, \quad I_j = \{i \leqslant n : z_i = j\} \quad \text{and} \quad \bar{x}_j = n_j^{-1} \sum_{i \in I_j} x_i.$$

Proof of Theorem 25. Let us set $x_1^n \in \mathcal{X}^n$. The left-hand inequalities of (5.4) and (5.5) are obvious.

Quite straightforwardly, using twice inequality $\sum_{j \leqslant k} \alpha_j / \sum_{j \leqslant k} \beta_j \leqslant \max_{j \leqslant k} \alpha_j/\beta_j$ (valid for all nonnegative $\alpha_1, \dots, \alpha_k$ and positive $\beta_1, \dots, \beta_k$) yields

$$\begin{aligned}
\sup_{\theta \in \Theta_k} \log \frac{g_\theta(x_1^n)}{q_k(x_1^n)} &= \log k + \sup_{\theta \in \Theta_k} \log \frac{\sum_{z_0 \leqslant k} g_\theta(x_1^n|z_0) p_{z_0}^o}{\sum_{z_0 \leqslant k} q_k(x_1^n|z_0)} \\
&\leqslant \log k + \sup_{\theta \in \Theta_k} \max_{z_0 \leqslant k} \log \frac{g_\theta(x_1^n|z_0) p_{z_0}^o}{q_k(x_1^n|z_0)} \\
&\leqslant \log k + \sup_{\theta \in \Theta_k} \max_{z_0 \leqslant k} \log \frac{g_\theta(x_1^n|z_0)}{q_k(x_1^n|z_0)} \\
&\leqslant \log k + \sup_{\theta \in \Theta_k} \max_{z_0 \leqslant k} \log \frac{\sum_{z_1^n \in \{1,\dots,k\}^n} g_\theta(x_1^n|z_0^n) P_\theta\{z_1^n|z_0\}}{\sum_{z_1^n \in \{1,\dots,k\}^n} q_k(x_1^n|z_0^n) q_k(z_1^n|z_0)} \\
&\leqslant \log k + \sup_{\theta \in \Theta_k} \max_{z_0^n \in \{1,\dots,k\}^{n+1}} \log \frac{g_\theta(x_1^n|z_1^n)}{q_k(x_1^n|z_1^n)} \cdot \frac{P_\theta\{z_1^n|z_0\}}{q_k(z_1^n|z_0)}. \qquad (5.8)
\end{aligned}$$

Now, as shown in ([DMPW81]) (see equations (52)-(61) therein),

$$\begin{aligned}
\sup_{\theta \in \Theta_k} \max_{z_0^n \in \{1,\dots,k\}^{n+1}} \log \frac{P_\theta\{z_1^n|z_0\}}{q_k(z_1^n|z_0)} &\leqslant k \log \frac{\Gamma(n + k/2)\Gamma(1/2)}{\Gamma(k/2)\Gamma(n+1/2)} \\
&\leqslant k \left(\frac{k-1}{2} \log n - \log \frac{\Gamma(k/2)}{\Gamma(1/2)} + \frac{k(k-1)}{4n} + \frac{1}{12n} \right), \qquad (5.9)
\end{aligned}$$

where the second inequality is derived from the following Robbins-Stirling approximation formula, valid for all $z \in \mathbb{R}_+^*$,

$$\sqrt{2\pi} e^{-z} z^{z-1/2} \leqslant \Gamma(z) \leqslant \sqrt{2\pi} e^{-z+1/12z} z^{z-1/2}.$$

So, the second ratio in the right-hand term of inequality (5.8) is controlled. The last step of the proof is dedicated to bounding the first ratio. The same scheme of proof applies to both examples **GE** and **PE**. It is nevertheless simpler to address each of them at a time.

GE Conditionally on $Z_1^n = z_1^n$ the maximum likelihood estimator of m_j is $\bar{x}_j$ for every $j \leqslant k$, so that the following bound holds for every $x_1^n \in \mathcal{X}^n$ and $z_1^n \in \{1, \dots, k\}^n$:

$$g_\theta(x_1^n|z_1^n) \leqslant \prod_{j=1}^{k} \prod_{i \in I_j} \phi_{\bar{x}_j, \sigma^2}(x_i) = \frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{j=1}^{k} \exp\left(-\frac{\sum_{i\in I_j} x_i^2}{2\sigma^2} - \frac{n_j(\bar{x}_j)^2}{2\sigma^2}\right). \tag{5.10}$$

Besides, simple calculations yield

$$\begin{aligned} q_k(x_1^n|z_1^n) &= \prod_{j=1}^{k} \frac{1}{(\sigma\sqrt{2\pi})^{n_j}} \int \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{m^2}{2\tau^2} - \frac{1}{2\sigma^2}\sum_{i\in I_j}(x_i - m)^2\right) dm \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{j=1}^{k} \frac{1}{\sqrt{1 + \frac{n_j\tau^2}{\sigma^2}}} \exp\left(-\frac{\sum_{i\in I_j} x_i^2}{2\sigma^2} + \frac{n_j^2}{2\sigma^2(n_j + \frac{\sigma^2}{\tau^2})}(\bar{x}_j)^2\right). \end{aligned} \tag{5.11}$$

We now get, as a by-product of inequalities (5.10) and (5.11),

$$\frac{g_\theta(x_1^n|z_1^n)}{q_k(x_1^n|z_1^n)} \leqslant \prod_{j=1}^{k} \sqrt{1 + \frac{n_j\tau^2}{\sigma^2}} \exp\left(\sum_{j=1}^{k} \frac{n_j}{2\sigma^2(1 + n_j\tau^2/\sigma^2)}(\bar{x}_j)^2\right).$$

By convexity, the first factor in the right-hand side expression above satisfies

$$\prod_{j=1}^{k} \sqrt{1 + \frac{n_j\tau^2}{\sigma^2}} \leqslant \left(1 + \frac{n\tau^2}{k\sigma^2}\right)^{k/2}, \tag{5.12}$$

while the ratios $n_j/(1 + n_j\tau^2/\sigma^2)$ are upper bounded by σ^2/τ^2 for all $j \leqslant k$. Therefore,

$$\sup_{\theta\in\Theta_k} \max_{z_0^n \in \{1,\dots,k\}^{n+1}} \log \frac{g_\theta(x_1^n|z_1^n)}{q_k(x_1^n|z_1^n)} \leqslant \frac{k}{2}\log\left(1 + \frac{n\tau^2}{k\sigma^2}\right) + \frac{k}{2\tau^2}|x|_{(n)}^2. \tag{5.13}$$

Combining inequalities (5.8), (5.9) and (5.13) yields the result.

PE As justified above, for each $j \leqslant k$, for every $x_1^n \in \mathcal{X}^n$ and $z_1^n \in \{1, \dots, k\}^n$:

$$g_\theta(x_1^n|z_1^n) \leqslant \prod_{j=1}^{k} \prod_{i\in I_j} \pi_{\bar{x}_j}(x_i) = P_n \prod_{j=1}^{k} \exp\Big(- n_j\bar{x}_j(1 - \log\bar{x}_j)\Big) \tag{5.14}$$

if $P_n = 1/\prod_{i=1}^{n}(x_i)!$. In particular, the factor associated with some $j \leqslant k$ for which $\bar{x}_j = 0$ equals one. Furthermore, it is readily seen that

$$\begin{aligned} q_k(x_1^n|z_1^n) &= P_n \prod_{j=1}^{k} \sqrt{\frac{\tau}{2\pi}} \int m^{n_j\bar{x}_j - 1/2} \exp\Big(- (n_j + \tau)m\Big) dm \\ &= P_n \prod_{j=1}^{k} \sqrt{\frac{\tau}{2\pi}} \frac{\Gamma(n_j\bar{x}_j + 1/2)}{(n_j + \tau)^{n_j\bar{x}_j + 1/2}}. \end{aligned} \tag{5.15}$$

Here, the factor associated with some $j \leqslant k$ for which $\bar{x}_j = 0$ equals $\sqrt{\tau/(n_j + \tau)}$.

At this stage, the ratio $g_\theta(x_1^n|z_1^n)/q_k(x_1^n|z_1^n)$ is naturally decomposed into the product of k ratios : for each $j \leqslant k$, the right-hand side factor of (5.14) divided by the right-hand side factor of (5.15) is upper bounded by

$$\sqrt{\frac{e}{\tau}} \times \exp\left(\frac{1}{2}\log n_j + \left(n_j\bar{x}_j + \frac{1}{2}\right)\log\left(1 + \frac{\tau}{n_j}\right)\right)$$

whether $\bar{x}_j = 0$ or not. This simple calculation relies again on the lower bound for $\Gamma(n_j\bar{x}_j + 1/2)$ yielded by the Robbins-Stirling approximation formula.

Consequently, it holds that

$$\begin{aligned} \log \frac{g_\theta(x_1^n|z_1^n)}{q_k(x_1^n|z_1^n)} &\leqslant \frac{k}{2}(1 - \log\tau) + \sum_{j=1}^{k}\left[\frac{1}{2}\log n_j + \tau\left(x_{(n)} + \frac{1}{2}\right)\right] \\ &\leqslant \frac{k}{2}\log\frac{n}{k} + k\tau x_{(n)} + \frac{k}{2}\left(1 + \tau - \log\tau\right) \end{aligned} \tag{5.16}$$

(the second inequality follows by convexity). Combining inequalities (5.8), (5.9) and (5.16) (we emphasize that the right-hand term in (5.16) does not depend on z_0^n nor on θ) gives the result.

■

Remark 17. *We want to point out that inequality (5.12) can not be improved, since it is optimal when all n_j's are roughly equal.*

The scheme of proof for Theorem 26 is similar to the one of Theorem 25.

Proof of Theorem 26. Let $x_1^n \in \mathcal{X}^n$. Straightforwardly, for every $\theta \in \Theta_k$,

$$g_\theta(x_1^n) = \sum_{z_1^n \in \{1,\ldots,k\}^n} g_\theta(x_1^n|z_1^n) \prod_{j=1}^{k} (p_j^o)^{n_j} \leqslant \sum_{z_1^n \in \{1,\ldots,k\}^n} g_\theta(x_1^n|z_1^n) \prod_{j=1}^{k} \left(\frac{n_j}{n}\right)^{n_j}.$$

Besides, it is readily seen that

$$\begin{aligned} q_k(x_1^n) &= \sum_{z_1^n \in \{1,\ldots,k\}^n} q_k(x_1^n|z_1^n) \int_{S_k'} \prod_{j=1}^{k} (p_j^o)^{n_j} d\nu_k'(\mathbf{p}) \\ &= \sum_{z_1^n \in \{1,\ldots,k\}^n} \frac{\Gamma(k/2)}{\Gamma(n+k/2)} q_k(x_1^n|z_1^n) \prod_{j=1}^{k} \frac{\Gamma(n_j+1/2)}{\Gamma(1/2)}. \end{aligned}$$

Consequently, using the same argument that yielded inequality (5.8) implies that

$$\log \frac{g_\theta(x_1^n)}{q_k(x_1^n)} \leqslant \sup_{z_1^n \in \{1,\ldots,k\}^n} \Bigg(\log \frac{\Gamma(n+k/2)\Gamma(1/2)^k}{\Gamma(k/2)} + \\ \log \prod_{j=1}^{k} \frac{\left(\frac{n_j}{n}\right)^{n_j}}{\Gamma(n_j+1/2)} + \log \frac{g_\theta(x_1^n|z_1^n)}{q_k(x_1^n|z_1^n)} \Bigg) .$$

Handling the second term in the right-hand side of the display above has already been done in the proof of Theorem 25. As for the first term, it holds that it is bounded by

$$\log \frac{\Gamma(n+k/2)\Gamma(1/2)}{\Gamma(k/2)\Gamma(n+1/2)} \leqslant \frac{k-1}{2} \log n + c_{kn}'$$

(by virtue of [DMPW81], equations (52-61) again and the Robbins-Stirling approximation formula). This completes the proof. ■

5.3 Application to order identification

Let k_0 be the sole integer such that the common distribution P_0 of all X_n $(n \geqslant 1)$ satisfies

$$P_0 \in \{P_\theta : \theta \in \Theta_{k_0}\} \setminus \{P_\theta : \theta \in \Theta_{k_0-1}\}$$

(with convention $\Theta_0 = \emptyset$). By definition, k_0 is the order of P_0. In examples **GE** and **PE**, k_0 is the minimal number of Gaussian or Poisson densities needed for describing the distribution P_0. Our goal in this section is to estimate k_0.

There is a large amount of literature dedicated to the issue of order estimation. The particular case of i.i.d order estimation for mixtures of continuous densities is notoriously challenging (refer to [Cha03] for a comprehensive bibliography). It has been addressed through various methods that can be classified into three categories : ad hoc [Hen85, DCG97, JPM01], maximum likelihood [Ler92a, Ker00, Gas02, Cha03] or Bayesian [IJS01, CR05]. Actually, Bayesian literature on order selection in mixture models is essentially devoted to determining coherent noninformative priors, see for instance [ML03] and to implementing procedures, see for instance [MR96]. The particular case of HMM order estimation was addressed for instance in [Fin91, Kie93, Gas02] from a maximum likelihood point of view and in [LN94] from a Bayesian one. Both approaches are combined in [GB03], and we refer to the comprehensive bibliography therein for further references.

Here, we study two estimators : one is based on maximum likelihood and the other is Bayesian flavored.

Let us denote by $\mathrm{pen}(n, k)$ a so-called penalization term, which is a positive valued increasing function of $n, k \geqslant 1$ such that, for each $k \geqslant 1$, $\mathrm{pen}(n, k) = o(n)$. This device is required for defining our estimators :

$$\begin{aligned} \widehat{k}_{\mathrm{ML}} &= \arg\min_{k\geqslant 1}\left\{-\sup_{\theta\in\Theta_k}\log g_\theta(X_1^n) + \mathrm{pen}(n,k)\right\} \quad \text{and} \\ \widehat{k}_{\mathrm{MIX}} &= \arg\min_{k\geqslant 1}\left\{-\log q_k(X_1^n) + \mathrm{pen}(n,k)\right\}. \end{aligned}$$

Convenient choices of the penalty term involve the following quantities. Let us set $\alpha > 2$, $\{\varphi_n\}_{n\geqslant 1}$ a sequence of positive numbers that increases slowly to infinity with $\varphi_n = o(n)$. For every $n, k \geqslant 1$, we introduce the cumulative sums : $C_{kn} = \sum_{\ell=1}^k c_{\ell n}$, $C'_{kn} = \sum_{\ell=1}^k c'_{\ell n}$, $D_{\ell n} = \sum_{\ell=1}^k d_{\ell n}$ and $E_{kn} = \sum_{\ell=1}^k e_{\ell n}$. All of them are bounded functions of n.

Theorem 27 (consistency of $\widehat{k}_{\mathrm{ML}}$). *Under the assumptions described above, $\widehat{k}_{\mathrm{ML}} = k_0$ eventually almost surely as soon as for every $n \geqslant 3, k \geqslant 1$,*

$$\mathrm{pen}(n,k) = \sum_{\ell=1}^{k} \frac{D(\ell)+\alpha}{2}\log n + R_{kn} + S_{kn},$$

where $D(k) = \dim(\Theta_k) = k^2$ and $R_{kn} = C_{kn}$ for HMM *mixtures models, $D(k) = \dim(\Theta'_k) = (2k-1)$ and $R_{kn} = C'_{kn}$ for i.i.d mixtures models and*

GE

$$S_{kn} = D_{kn} + k(k+1)\varphi_n \log n,$$

PE

$$S_{kn} = E_{kn} + k(k+1)\frac{\log n}{\sqrt{\log\log n}}.$$

Similarly,

Theorem 28 (consistency of $\widehat{k}_{\mathrm{MIX}}$). *Under the assumptions described above, $\widehat{k}_{\mathrm{MIX}} = k_0$ eventually almost surely as soon as for every $n \geqslant 3, k \geqslant 1$,*

$$\mathrm{pen}(n,k) = \sum_{\ell=1}^{k-1} \frac{D(\ell)+\alpha}{2}\log n + S_{kn},$$

where $D(k) = \dim(\Theta_k) = k^2$ *for* HMM *mixtures models,* $D(k) = \dim(\Theta'_k) = (2k-1)$ *for i.i.d mixtures models and*

GE

$$S_{kn} = k(k+1)\varphi_n \log n,$$

PE

$$S_{kn} = k(k+1)\frac{\log n}{\sqrt{\log\log n}}.$$

Theorems 27 and 28 thus guarantee that $\widehat{k}_{\text{ML}}$ and $\widehat{k}_{\text{MIX}}$ are consistent estimators of k_0. We emphasize that *no prior bound on* k_0 *is required.* This is particularly interesting for $\widehat{k}_{\text{ML}}$, since such a bound was generally needed when studying its overestimation properties in the mixture of continuous densities order estimation setting, as in [Ker00, Gas02, Cha03]. This feature illustrates the fact that our contribution to order estimation is mainly related to the overestimation phenomenon. In the HMM framework, few results are known for infinite alphabet emissions.

The conditions required on the penalty function in Theorems 27 and 28 should be discussed too. It is worth noting that, in example **GE**, the penalty is designed without knowing σ^2. This is why $\text{pen}(n,k)$ is larger than $\log n$ when n grows to infinity, for every $k \geqslant 1$, which is not satisfactory in reference to the BIC criterion. Assuming known an upper-bound for σ^2 would allow to choose a penalty function that satisfies $\text{pen}(n,k) = O(\log n)$ for every $k \geqslant 1$. In example **PE**, the behavior of the random term in inequalities (5.5) and (5.7) allows to choose penalties satisfying $\text{pen}(n,k) = O(\log n)$ for every $k \geqslant 1$.

It is also important to compare the dependency of $\text{pen}(n,k)$ with respect to k with that of the BIC criterion. We do not get a single term $\frac{1}{2}D(k)$ on the $\log n$ scale, but rather a cumulative sum of terms $\frac{1}{2}[D(\ell)+\alpha]$ for ℓ ranging from 1 to k.

The few lines of comment above are devoted to $\widehat{k}_{\text{ML}}$. It is well understood that Bayesian estimators naturally take into account the uncertainty on the parameter by integrating it out [JB92], thus providing an example of auto-penalization. This is illustrated by the equivalence between marginal likelihood and BIC criterion that holds, for instance, in regular models :

$$-\log q_k(X_1^n) = -\log \sup_{\theta\in\Theta_k} g_\theta(X_1^n) + \frac{1}{2}D(k)\log n + O_P(1),$$

as n goes to infinity, valid for every $k \geqslant 1$. It is proven in [CR05] that efficient order estimation can be achieved by comparing marginal likelihoods (implicitely, without additional penalization) even in non-regular models (and for instance for mixtures of continuous densities). However, Csiszár & Shields [CS00] provide an example where $\widehat{k}_{\text{ML}}$ is consistent while $\widehat{k}_{\text{MIX}}$ is not when its penalty term is set to zero. Here, we (over-) penalize $q_k(X_1^n)$ so that the proofs of Theorems 27 and 28 mainly rely on the mixture inequalities stated in Theorems 25 and 26.

Proof of Theorem 27. In the i.i.d framework, showing that $\widehat{k}_{\text{ML}} \geqslant k_0$ eventually almost surely is a rather simple consequence of the strong law of large numbers and $\min_{k<k_0} \inf_{\theta\in\Theta'_k} K(g_{\theta_0}, g_\theta) > 0$ for any $\theta_0 \in \Theta'_{k_0} \setminus \Theta'_{k_0-1}$ (see [Ler92a] for a proof of the latter), where

$$K(g_{\theta_0}, g_\theta) = \int_{x_1\in\mathcal{X}} g_{\theta_0}(x_1) \log \frac{g_{\theta_0}(x_1)}{g_\theta(x_1)} d\mu(x_1)$$

is the P_{θ_0}-almost sure limit of $n^{-1}[\log g_{\theta_0}(X_1^n) - \log g_\theta(X_1^n)]$.

In the HMM framework, it is a consequence of Lemma 10 (see Appendix 5.6), which contains a Shannon-Breiman-McMillan theorem for HMM that holds in examples **GE** and **PE** (see Theorem 2 in [Ler92b]) and a useful by-product of the proof of Theorem 3 in the same paper.

The difficult part is to get that $\widehat{k}_{\text{ML}} \leqslant k_0$ eventually almost surely.

Let $P_0 = P_{\theta_0}$ for $\theta_0 \in \Theta_{k_0} \setminus \Theta_{k_0-1}$. Let us consider a positive valued sequence $\{t_n\}_{n\geqslant 3}$ to be chosen conveniently later on. Let $k > k_0$. Obviously, if $\widehat{k}_{\text{ML}} = k$, then $\log g_{\theta_0}(X_1^n) \leqslant \sup_{\theta\in T_k} g_\theta(X_1^n) + \text{pen}(n, k_0) - \text{pen}(n, k)$. Here, T_k equals Θ_k for HMM mixture models and equals Θ'_k for i.i.d mixture models. Consequently, using inequalities (5.4), (5.5), (5.6) or (5.7) (with $\tau = 1/2$ in example **GE** and $\tau = 2$ in example **PE**), $\widehat{k}_{\text{ML}} = k$ yields

$$\log g_{\theta_0}(X_1^n) \leqslant \log q_k(X_1^n) + \Delta_{nk} \tag{5.17}$$

with

$$\Delta_{nk} = \text{pen}(n, k_0) - \text{pen}(n, k) + \frac{D(k)}{2} \log n + a_{kn} + b_{kn} + 2kU_n,$$

where $U_n = |X|_{(n)}^2$, $b_{kn} = d_{kn}$ in example **GE** and $U_n = X_{(n)}$, $b_{kn} = e_{kn}$ in example **PE**, while $a_{kn} = c_{kn}$ for HMM mixture models and $a_{kn} = c'_{kn}$ for i.i.d mixture models.

Furthermore, because q_k defines a probability measure, the event defined by inequality (5.17) has P_0-probability

$$\int_{x_1^n\in\mathcal{X}^n} \frac{g_{\theta_0}(x_1^n)}{q_k(x_1^n)} \mathbb{1}\left\{\log \frac{g_{\theta_0}(x_1^n)}{q_k(x_1^n)} \leqslant \Delta_{nk}\right\} q_k(x_1^n) d\mu(x_1^n) \leqslant \exp(\Delta_{nk}).$$

Now, if we choose $t_n = \varphi_n \log n$ in example **GE** and $t_n = \log n/\sqrt{\log\log n}$ in example **PE**, then $U_n \leqslant t_n$ implies

$$\Delta_{nk} \leqslant \frac{\alpha}{2}(k_0 - k) \log n. \tag{5.18}$$

Therefore

$$P_0\left\{\widehat{k}_{\text{ML}} > k_0 \text{ and } U_n \leqslant t_n\right\} \leqslant \sum_{k>k_0} \exp\left\{-\frac{\alpha}{2}(k-k_0)\log n\right\} = O(n^{-\alpha/2}).$$

In conclusion, by virtue of the Borel-Cantelli lemma, the previous bound and Lemmas 8 and 9 guarantee that $\widehat{k}_{\text{ML}} \leqslant k_0$ eventually almost surely. This completes the proof. ■

Remark 18. *The specific form chosen for the penalty term in each setting is meant for ensuring inequality (5.18).*

Proof of Theorem 28. In that case, proving that $\widehat{k}_{\text{MIX}} \geqslant k_0$ eventually almost surely is a little more involved (but still well known). The key-point is to show the existence of a random variables sequence $\{\varepsilon_n\}_{n\geqslant 1}$ that converges to 0 P_0-almost surely and, for each $k < k_0$, $\widehat{k}_{\text{MIX}} = k$ yields

$$\frac{1}{n}\left[\sup_{\theta\in\Theta_k} \log g_\theta(X_1^n) - \log g_{\theta_0}(X_1^n)\right] \geqslant \varepsilon_n \text{ i.o.}$$

(i.o. stands for "infinitely often"). Then, the conclusion follows again from the strong law of large numbers again in the i.i.d framework and from Lemma 10 in the HMM framework.

Let us set $k < k_0$. Because $\text{pen}(n,k) = o(n)$ and $\text{pen}(n,k_0) = o(n)$, $\widehat{k}_{\text{MIX}} = k$ yields

$$0 \geqslant \frac{1}{n}\log\frac{q_{k_0}(X_1^n)}{q_k(X_1^n)} + o(1).$$

By adding the same quantity to both sides, we get

$$\frac{1}{n}\left[\sup_{\theta\in\Theta_k} \log g_\theta(X_1^n) - \log g_{\theta_0}(X_1^n)\right] \geqslant \frac{1}{n}\log\frac{\sup_{\theta\in\Theta_k} g_\theta(X_1^n)}{q_k(X_1^n)} - \frac{1}{n}\log\frac{g_{\theta_0}(X_1^n)}{q_{k_0}(X_1^n)} + o(1).$$

Now, by virtue of inequalities (5.4), (5.5), (5.6) and (5.7) and Lemmas 8 and 9, P_0-almost surely

$$\frac{1}{n}\log\frac{\sup_{\theta\in\Theta_k} g_\theta(X_1^n)}{q_k(X_1^n)} \xrightarrow[n\to\infty]{} 0.$$

The same inequalities and lemmas also guarantee that, P_0-almost surely,

$$\frac{1}{n}\left(\log\frac{g_{\theta_0}(X_1^n)}{q_{k_0}(X_1^n)}\right)_+ \xrightarrow[n\to\infty]{} 0$$

(the positive part of $t \in \mathbb{R}$ is denoted by $(t)_+$). The final step is a variant of the so-called Barron's lemma taken from ([Fin91], Theorem 4.4.1) : another application of the Borel-Cantelli lemma implies that, P_0-almost surely,

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{g_{\theta_0}(X_1^n)}{q_{k_0}(X_1^n)} \geqslant \liminf_{n\to\infty} \frac{-2\log n}{n} = 0.$$

Therefore, the key-point holds, hence the conclusion for underestimation.

From now on, we use the same notations than in the preceding proof except when notified. Let $k > k_0$. If $\widehat{k}_{\text{MIX}} = k$, then $-\log q_k(X_1^n) + \text{pen}(n,k) \leqslant -\log q_{k_0}(X_1^n) + \text{pen}(n,k_0)$ and using inequalities (5.4), (5.5), (5.6) and (5.7) implies that

$$\log g_{\theta_0}(X_1^n) \leqslant \log q_k(X_1^n) + \Delta_{nk}$$

with

$$\Delta_{nk} = \text{pen}(n,k_0) - \text{pen}(n,k) + \frac{D(k_0)}{2}\log n + a_{k_0 n} + 2k_0 U_n,$$

where $\{a_{k_0 n}\}_{n\geqslant 1}$ is a bounded sequence. The definition of the penalty guarantees that $U_n \leqslant t_n$ implies that inequality (5.18) still holds in this setting. Consequently,

$$P_0\left\{\widehat{k}_{\text{MIX}} > k_0 \text{ and } U_n \leqslant t_n\right\} \leqslant \sum_{k>k_0} \exp\left\{-\frac{\alpha}{2}(k-k_0)\log n\right\} = O(n^{-\alpha/2}).$$

The result follows by virtue of the Borel-Cantelli lemma, the previous bound and Lemmas 8 and 9 : $\widehat{k}_{\text{ML}} \leqslant k_0$ eventually almost surely. This completes the proof. ■

5.4 Predictive MDL

In this section, we are interested in predictive MDL applied to mixture order identification in the i.i.d framework. This is an alternative to methods based on maximum likelihood (a variant of two-stage MDL) or Bayesian estimation (a variant of mixture MDL), see [HY01], considered in Section 5.3.

For each $k \geqslant 1$, let $\widehat{\theta}_k(x_1^n) \in \Theta_k$ be the maximum likelihood estimator of θ over Θ_k based on the observation of $x_1^n \in \mathcal{X}^n$. The predictive MDL statistics is

$$r_k(X_1^n) = \prod_{i=1}^{n} g_{\widehat{\theta}_k(X_1^{i-1})}(X_i),$$

where $\widehat{\theta}_k(x_1^0)$ equals some fixed parameter in the interior of Θ_k.

This statistics yields another order estimator, which is the smallest maximizer of $r_k(X_1^n)$ over integers $k \geqslant 1$. We merely aim at giving a hint why the scheme of proof used in Section 5.3 does not apply for this estimator. In words, if the true distribution P_{θ_0} is of order k_0, then for all $k > k_0$, the null measure set outside of which Rissanen's lower bound over Θ_k is verified will include θ_0.

Let us denote by $\mathcal{F}$ a set of densities with respect to a measure μ over $\mathbb{R}^d$. For any two densities $g_1, g_2 \in \mathcal{F}$, $H^2(g_1, g_2) = \int(\sqrt{g_1} - \sqrt{g_2})^2 d\mu$ is the square Hellinger distance between g_1 and g_2. We recall that $K(g_1, g_2) = \int g_1 \log(g_1/g_2) d\mu$ is the Kullback-Leibler divergence between g_1 and g_2.

Let $X_1, \dots, X_n$ be i.i.d random variables whose distribution P has density $f \in \mathcal{F}$ with respect to μ. We introduce, for every $\varepsilon > 0$,

$$\mathcal{F}_\varepsilon = \{g \in \mathcal{F} : H^2(g, f) \leqslant \epsilon^2\}$$

and

$$\mathcal{D}_\varepsilon = \left\{ \frac{\sqrt{g/f} - 1}{H(g, f)} : g \in \mathcal{F}_\varepsilon \right\}.$$

We recall that the bracket-entropy $H_{[]}(u, \mathcal{S}, L^2(P))$ of the set $\mathcal{S} \subset L^2(P)$ is the logarithm of the minimal number of brackets of length u needed to cover $\mathcal{S}$ (see [vdV98]). Let us state three assumptions (see [vdV98] for definitions of Glivenko-Cantelli and Donsker classes).

Envelope There exists $C \in L^2(P)$ such that, for all $g_1, g_2 \in \mathcal{F}_\varepsilon$, $|\log g_1 - \log g_2| \leqslant CH(g_1, g_2)$ P-almost surely.

Glivenko The set $\{\log g : g \in \mathcal{F}\}$ is P-Glivenko-Cantelli.

Donsker There exist $\delta, \varepsilon > 0$ such that

$$\int_0^\delta \sqrt{H_{[]}(u, \mathcal{D}_\varepsilon, L^2(P))} du < \infty$$

and, if $\{g_p\}_{p \geqslant 1}$ is a sequence of elements of $\mathcal{F}_\varepsilon$ such that $H(f, g_p) = o(1)$, then $K(f, g_p) = 2H^2(f, g_p)(1 + o(1))$.

Let $\widehat{f}$ be the maximum likelihood estimator of f on $\mathcal{F}$.

We emphasize that, by virtue of Theorems 5.7 and 5.52 in [vdV98], $\widehat{f}$ is consistent in Hellinger distance when assumption **Glivenko** holds, that is

$$H^2(\widehat{f}, f) = o_P(1),$$

and, under assumption **Envelope**,

$$nH^2(\widehat{f}, f) = O_P(1).$$

Under assumption **Donsker**, we can define the set $\mathcal{D}_0$ of all limit points of sequences $\{d_\varepsilon\}$, $d_\varepsilon \in \mathcal{D}_\varepsilon$, $\varepsilon \to 0$, as a compact subset of the unit sphere in $L^2(P)$. Moreover, $\mathcal{D}_\varepsilon$ and $\mathcal{D}_0$ are P-Donsker classes. Notice that for all $g \in \mathcal{F}_\varepsilon$,

$$2 \int \frac{1 - \sqrt{g/f}}{H(g, f)} f d\mu = H(g, f) \leqslant \varepsilon,$$

so that for all $d \in \mathcal{D}_0$, $E(d(X_1)) = 0$. Let us introduce the centered empirical process $\mathbb{G}_n$ on $\mathcal{D}_0$ defined, for every $d \in \mathcal{D}_0$, by

$$\mathbb{G}_n(d) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} d(X_i).$$

This process has unit variance and covariance $\langle d_1, d_2 \rangle_{L^2(P)}$. We denote by W the centered Gaussian process on $\mathcal{D}_0$ with the same covariance.

Theorem 29 (expansion of $nK(f, \widehat{f})$). *Under assumptions* **Envelope, Glivenko** *and* **Donsker**, *the following expansion holds :*

$$nK(f, \widehat{f}) = \frac{1}{2} \sup_{d \in \mathcal{D}_0} \Big(\mathbb{G}_n(d) \mathbb{1}\{\mathbb{G}_n(d) \geqslant 0\} \Big)^2 + o_P(1).$$

Therefore, $nK(f, \widehat{f})$ converges in distribution to

$$\frac{1}{2} \sup_{d \in \mathcal{D}_0} \Big(W(d) \mathbb{1}\{W(d) \geqslant 0\} \Big)^2.$$

Theorem 29 applies to mixture of densities from a smooth enough parametric family $\{\gamma_{\alpha_i,\beta} : \alpha \in A \subset \mathbb{R}^m\}$ with possibly unknown common nuisance parameter $\beta \in B \subset \mathbb{R}^q$. In that case, any k-mixture density g_θ writes as

$$g_\theta = \sum_{j=1}^{k} p_j \gamma_{\alpha_j,\beta},$$

where $\mathbf{p} = (p_j : j \leqslant k) \in \mathcal{S}'_k$ (introduced in Section 5.2), $\mathbf{m} = (\alpha_1, \ldots, \alpha_k, \beta) \in A^k \times B$ and $\theta = (\mathbf{p}, \mathbf{m}) \in \Theta'_k = \mathcal{S}'_k \times A^k \times B$. Here, the number of parameters is $D_k = km + q + k - 1$.

In [AGM04], general assumptions are given on parametric family $\{\gamma_{\alpha,\beta} : \alpha \in A, \beta \in B\}$ so that assumptions **Envelope, Glivenko** and **Donsker** be satified. In particular, they hold for Binomial, Poisson and multidimensional Gaussian mixtures.

Let $\mathcal{F} = \{g_\theta : \theta \in \Theta'_k\}$ for some integer $k \geqslant 2$.

If $f \in \mathcal{F} \setminus \{g_\theta : \theta \in \Theta'_{k-1}\}$ (all p_j are positive and $\alpha_1, \ldots, \alpha_k$ are mutually distinct), then $\mathcal{D}_0$ is a linear space, $\sup_{d \in \mathcal{D}_0} (W(d) \mathbb{1}\{W(d) \geqslant 0\})^2$ has a chi-square distribution with D_k degrees of freedom and

$$E\left[\sup_{d \in \mathcal{D}_0} (W(d) \mathbb{1}\{W(d) \geqslant 0\})^2 \right] = D_k,$$

as in identifiable parametric situations.

Now if, on the contrary, $f \in \{g_\theta : \theta \in \Theta'_{k_0}\}$ for some $k_0 < k$, then

$$E\left[\sup_{d \in \mathcal{D}_0} (W(d)\mathbb{1}\{W(d) \geqslant 0\})^2\right] < D_k$$

(see for instance [Del01]).

Proof of Theorem 29. Let us define the log-likelihood

$$\ell_n(g) = \sum_{i=1}^n \log g(X_i)$$

and, for every $g \in \mathcal{F}$,

$$d_g = \frac{\sqrt{g/f} - 1}{H(g, f)}.$$

Using the same tricks as in Section 3 of [Gas02] yields, for some $\varepsilon_n = o(1)$,

$$\begin{aligned} \ell_n(\widehat{f}) - \ell_n(f) &= \sup_{g \in \mathcal{F}_{\varepsilon_n}} \left(2H(f,g) \sum_{i=1}^n d_g(X_i) - H^2(f,g) \sum_{i=1}^n d_g^2(X_i) \right) (1 + o_P(1)) \\ &= \sup_{g \in \mathcal{F}_{\epsilon_n}} \left(2H(f,g) \sum_{i=1}^n d_g(X_i) - 2nH^2(f,g) \right) (1 + o_P(1)). \end{aligned}$$

Because $nH^2(f, \widehat{f}) = O_P(1)$ and (as in Section 3 of [Gas02]), for all $d \in \mathcal{D}_0$, there is a submodel $g_{c,d}$ with normalized score d and $c = H(f, g_{c,d})$, it holds that

$$nH^2\left(f, \widehat{f}\right) = \frac{1}{4} \sup_{d \in \mathcal{D}_0} \left(\mathbb{G}_n(d) 1_{\mathbb{G}_n(d) \geqslant 0}\right)^2 (1 + o_P(1)),$$

so that Theorem 29 follows from assumption **Donsker**. ■

5.5 Proofs of Lemmas 8 and 9

Proof of Lemma 8. Let $m = \sup_{n \geqslant 1} |m_n|$ and $t_n = \sqrt{5\sigma^2 \log n}$ (all $n \geqslant 1$). Let n be large enough, so that $t_n \geqslant m$. For every $i \leqslant n$,

$$\begin{aligned} P\{|Y_i| \leqslant t_n\} &= P\{|m_i + Y_i - m_i| \leqslant t_n\} \\ &\geqslant P\{|Y_i - m_i| \leqslant t_n - |m_i|\} \\ &\geqslant P\{|Y_i - m_i| \leqslant t_n - m\} \\ &= \int_{-t_n+m}^{t_n-m} \phi_{0,\sigma^2}(y) dy \\ &= \left(1 - \sigma \frac{\phi_{0,\sigma^2}(t_n)}{t_n}\right)(1 + o(1)). \end{aligned}$$

Hence, by virtue of the independence of $Y_1, \dots, Y_n$,

$$\begin{aligned}
P\left\{|Y|^2_{(n)} \geqslant t_n^2\right\} &= 1 - \prod_{i=1}^{n} P\{|Y_i| \leqslant t_n\} \\
&\leqslant 1 - \left(1 - \sigma \frac{\phi_{0,\sigma^2}(t_n)}{t_n}(1 + o(1))\right)^n \\
&= 1 - \exp\left\{ - \frac{n \exp\left(-\frac{t_n^2}{2\sigma^2}\right)}{t_n\sqrt{2\pi}}(1 + o(1))\right\} \\
&= - \frac{n \exp\left(-\frac{5\sigma^2 \log n}{2\sigma^2}\right)}{\sqrt{5\sigma^2 \log n}\sqrt{2\pi}}(1 + o(1)) \\
&\leqslant n^{-3/2},
\end{aligned}$$

as soon as n is large enough. ■

Proof of Lemma 9. Let $m = \sup_{n \geqslant 1} m_n$ and $t_n = \log n / \sqrt{\log \log n}$ (all $n \geqslant 3$). Let Y be a Poisson random variable with mean m. The logarithmic moment generating function Ψ of $(Y - m)$ satisfies $\Psi(\lambda) = \log E e^{\lambda(Y-m)} = m(e^\lambda - \lambda - 1)$ (all $\lambda \geqslant 0$). Its Legendre transform Ψ^* is given for all $t \geqslant 0$ by

$$\Psi^*(t) = \sup_{\lambda \geqslant 0}\{\lambda t - \Psi(\lambda)\} = (t + m)\log\frac{t+m}{m} - t.$$

Now, it is obvious that $P\{Y_i \geqslant t\} \leqslant P\{Y \geqslant t\}$ (for each $i \leqslant n$ and $t > m$). Therefore, by using the Chernoff bounding method,

$$P\{Y_{(n)} \geqslant t_n\} \leqslant nP\{Y \geqslant t_n\} = nP\{Y - m \geqslant t_n - m\} \leqslant n \exp\{-\Psi^*(t_n - m)\}. \tag{5.19}$$

Besides,

$$\Psi^*(t_n - m) = t_n \log \frac{t_n}{m} - t_n - m = (\log n)\sqrt{\log\log n}(1 + o(1)) \geqslant 3 \log n$$

as soon as n is large enough. We conclude by plugging this lower bound into inequality (5.19). ■

5.6 A useful lemma for HMM mixture models

Lemma 10 (Leroux). *For* HMM *mixture models, both in examples* **GE** *and* **PE**, *for every* $k \geqslant 1$ *and* $\theta_0, \theta \in \Theta_k$, *there exists a constant* $K_\infty(g_{\theta_0}, g_\theta) < \infty$ *such* *that, P_{θ_0}-almost surely, $n^{-1}[\log g_{\theta_0}(X_1^n) - \log g_\theta(X_1^n)]$ tends to $K_\infty(g_{\theta_0}, g_\theta)$ as n goes to infinity. Besides, for any $\theta_0 \in \Theta_{k_0} \setminus \Theta_{k_0-1}$,*

$$\min_{k<k_0} \inf_{\theta\in\Theta_k} K_\infty(g_{\theta_0}, g_\theta) > 0.$$

Sketch of proof of Lemma 10. The Shannon-Breiman-McMillan part of the lemma is a straightforward consequence of Theorem 2 in [Ler92b]. The second part of the lemma is a by-product of the proof of Theorem 3 of the same paper. Indeed, Leroux proved that, for each $\theta \in \Theta_{k_0}$ such that $g_\theta \neq g_{\theta_0}$, there exists an open neighborhood $\mathcal{O}_\theta$ of θ (for the euclidean topology of the one-point compactification of Θ_{k_0}) and $\varepsilon > 0$ such that $\inf_{\theta'\in\mathcal{O}_\theta} K_\infty(g_{\theta_0}, g_\theta) > \varepsilon$. Because Θ_{k_0-1} is precompact, it is covered by the finite union of $\mathcal{O}_{\theta_1}, \dots, \mathcal{O}_{\theta_I}$ (each of them associated with $\varepsilon_i > 0$) and therefore

$$\inf_{\theta\in\Theta_{k_0-1}} K_\infty(g_{\theta_0}, g_\theta) \geqslant \min_{i\leqslant I} \inf_{\theta\in\mathcal{O}_{\theta_i}} K_\infty(g_{\theta_0}, g_\theta) \geqslant \min_{i\leqslant I} \varepsilon_i > 0.$$

■

Bibliographie

[ADC86] Robert Azencott and Didier Dacunha-Castelle. *Series of irregular observations.* Springer-Verlag, New-York., 1986.

[AGM04] Jean-Marc Azais, Elisabeth Gassiat, and Cécile Mercadier. Asymptotic distribution and power of the likelihood ratio test for mixtures : bounded and unbounded case. *Submitted*, 2004.

[Ans] Answers.com. Binary tree.

[Bar85] Andrew R. Barron. *Logically Smooth Density Estimation.* PhD thesis, Stanford University, Dept. of Engineering, 1985.

[BBLM05] Stéphane Boucheron, Oliver Bousquet, Gabor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *Annals of Probability*, 33(2) :514–560, 2005.

[BGG06] Stéphane Boucheron, Aurélien Garivier, and Elisabeth Gassiat. Infinite alphabets : redundancy rates for enveloppe classes. Work in progress, to be submitted, 2006.

[Bou00] Stéphane Boucheron. Théorie de l'information, notes de cours. `http://www.proba.jussieu.fr/pageperso/boucheron/mpriit.php`, 2000.

[BP66] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37 :1554–1563, 1966.

[BRY98] Andrew R. Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6) :2743–2760, 1998.

[BW99] Peter Bühlmann and Abraham J. Wyner. Variable length Markov chains. *Ann. Statist.*, 27(2) :480–513, 1999.

[Cat01] Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization.* Lecture Notes in Mathematics , Vol. 1851. Springer-Verlag, Berlin, 2001.

[CB90] Bertrand S. Clarke and Andrew R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory*, 36 :453–471, 1990.

[CB94] B.S. Clarke and A.R Barron. Jeffrey's prior is asymptotically least favorable under entropy risk. *J. Stat. Planning and Inference*, 41 :37–60, 1994.

[CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006. ISBN 0521841089.

[CGG05] Antoine Chambaz, Elisabeth Gassiat, and Aurélien Garivier. A MDL approach to HMM with Poisson and Gaussian emissions. Application to order identification. *JSPI*, submitted, 2005.

[Cha03] Antoine Chambaz. Testing the order of a model. *Ann. Statist.*, Accepted, 2003.

[CK81] Imre Csiszár and János Körner. *Information theory.* Akadémiai Kiadó (Publishing House of the Hungarian Academy of Sciences), Budapest, 1981. Coding theorems for discrete memoryless systems.

[CMR05] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models.* Springer Series in Statistics. Springer, New York, 2005. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

[Cox93] D. Cox. An analysis of Bayesian inference for nonparametric regression. *Annals of Statistics*, 21 :903–923, 1993.

[CR05] Antoine Chambaz and Judith Rousseau. Nonasymptotic bounds for Bayesian order identification with application to mixtures. *Submitted.*, 2005.

[CS96] Imre Csiszár and Paul C. Shields. Redundancy rates for renewal and other processes. *IEEE Transactions on Information Theory*, 42, Nov. 1996.

[CS00] Imre Csiszár and Paul C. Shields. The consistency of the BIC Markov order estimator. *Ann. Statist.*, 28(6) :1601–1619, 2000.

[Csi02] Imre Csiszár. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory*, 48(6) :1616–1628, 2002. Special issue on Shannon theory : perspective, trends, and applications.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley Series in Telecommunications. John Wiley & Sons Inc., New York, 1991. A Wiley-Interscience Publication.

[CT06] Imre Csiszár and Zsolt Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE-IT*, 52(3), march 2006.

[Dav73] Lee D. Davisson. Universal noiseless coding. *IEEE Trans. Information Theory*, IT-19 :783–795, 1973.

[DCG97] Didier Dacunha-Castelle and Elisabeth Gassiat. The estimation of the order of a mixture model. *Bernoulli*, pages 279–299, 1997.

[DEKM98] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis.* Cambridge University Press, 1998.

[Del01] Céline Delmas. *Distribution du maximum d'un champ aléatoire et applications statistiques.* PhD thesis, Université Paul Sabatier, Toulouse, 2001.

[DF93] P. Diaconis and D. A. Freedman. Nonparametric binary regression : a Bayesian approach. *Ann. Statist.*, 21(4) :2108–2137, 1993.

[DG90] Luc Devroye and László Györfi. No empirical probability measure can converge in the total variation sense for all distributions. *Ann. Statist.*, 18(3) :1496–1499, 1990.

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York).* Springer-Verlag, New York, 1996.

[DLG80] Lee D. Davisson and Alberto Leon-Garcia. A source matching approach to finding minimax codes. *IEEE Trans. Inform. Theory*, 26(2) :166–174, 1980.

[DMPW81] Lee D. Davisson, Robert J. McEliece, Michael B. Pursley, and Mark S. Wallace. Efficient universal noiseless source codes. *IEEE Trans. Inf. Theory*, 27 :269–279, 1981.

[DN90] Jacques Dixmier and Jean-Louis Nicolas. *A tribute to Paul Erdös*, chapter Partitions sans petits sommants, pages 121–152. Cambridge University Press, 1990.

[DR98] D. Dubhashi and D. Ranjan. Balls and bins : A study in negative dependence. *Random Struct. & Algorithms*, 13(2) :99–124, 1998.

[DZ98] A. Dembo and O. Zeitouni. *Large deviation techniques and applications.* Springer, 1998.

[Eli75] Peter Elias. Universal codeword sets and representations of the integers. *IEEE Trans. Information Theory*, IT-21 :194–203, 1975.

[EM02] Yariv Ephraim and Neri Merhav. Hidden markov processes. *IEEE Trans. Inform. Theory*, 48 :1518–1569, 2002.

[EVKV02] Michelle Effros, Karthik Visweswariah, Sanjeev R. Kulkarni, and Sergio Verdú. Universal lossless source coding with the burrows Wheeler transform. *IEEE Trans. Inform. Theory*, 48(5) :1061–1081, 2002.

[Fan61] Robert M. Fano. *Transmission of information : A statistical theory of communications.* The M.I.T. Press, Cambridge, Mass., 1961.

[Fin91] Lorenzo Finesso. *Consistent estimation of the order for Markov and hidden Markov chains.* PhD thesis, University of Maryland, 1991.

[FMG92] Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *IEEE Trans. Inform. Theory*, 38(4) :1258–1270, 1992.

[Fre63] David A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.*, 34 :1386–1403, 1963.

[Fre99] David A. Freedman. On the Bernsteinvon Mises theorem with infinite dimensional parameters. *Annals of Statistics*, 27 :1119–1140, 1999.

[Gal68] R. G. Gallager. *Information theory and reliable communication.* John Wiley & sons, 1968.

[Gal76] Robert G. Gallager. Source coding with side information and universal coding, Sept. 1976.

[Gar03] Aurélien Garivier. Rapport de DEA : Parsing et codage universel. Master's thesis, Université Paris Sud Orsay, 2003.

[Gar05a] Aurélien Garivier. Redundancy of the context tree weighting algorithm on renewal and other processes. *IEEE-IT*, accepted, 2005.

[Gar05b] Aurélien Garivier. Consistency of the unlimited BIC context tree estimator. *IEEE Trans. Inform. Theory*, accepted, 2005.

[Gar06] Aurélien Garivier. A new lower-bound for the pattern maximin redundancy. *IEEE-IT*, submitted, 2006.

[Gas02] Elisabeth Gassiat. Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. H. Poincaré Probab. Statist.*, 38 :897–906, 2002.

[GB03] Elisabeth Gassiat and Stéphane Boucheron. Optimal error exponents in hidden Markov models order estimation. *IEEE Trans. Inform. Theory*, 49 :964–980, 2003.

[GK97] Robert Giegerich and Stefan Kurtz. From Ukkonen to McCreight and Weiner : A unifying view of linear-time suffix tree construction. *Algorithmica*, 19(3) :331–353, 1997.

[GPvdM94] Lászlo Györfi, István Páli, and Edward C. van der Meulen. There is no universal source code for an infinite source alphabet. *IEEE Trans. Inform. Theory*, 40(1) :267–271, 1994.

[GW04] George M. Gemelos and Tsachy Weissman. On the entropy rate of pattern processes. Technical report hpl-2004-159, HP Laboratories Palo Alto, San Antonio, Texas, USA, sept 2004.

[Hau97] David Haussler. A general minimax result for relative entropy. *IEEE Trans. Inform. Theory*, 43(4) :1276–1280, 1997.

[Hen85] Jogi Henna. On estimating of the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Statist. Math.*, 37 :235–240, 1985.

[HG94] James P. Hughes and Peter Guttorp. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Ressources Research*, 30 :1535–1546, 1994.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning.* Springer-Verlag, New-York, 2001.

[HU79] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to automata theory, languages, and computation.* Addison-Wesley Publishing Co., Reading, Mass., 1979. Addison-Wesley Series in Computer Science.

[HY01] Mark H. Hansen and Bin Yu. Model selection and the principle of minimum description length. *JASA*, 96 :746–774, 2001.

[IJS01] Hemant Ishwaran, Lancelot F. James, and Jiayang Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.*, 96 :1316–1332, 2001.

[JB92] William Jefferys and James Berger. Ockam's razor and bayesian analysis. *American Scientist*, 80 :64–72, 1992.

[JOS05] Nikola Jevtić, Alon Orlitsky, and Narayana P. Santhanam. A lower bound on compression of unknown alphabets. *Theoret. Comput. Sci.*, 332(1-3) :293–311, 2005.

[JPM01] Lancelot F. James, Carey E. Priebe, and David J. Marchette. Consistent estimation of mixture complexity. *Ann. Statist.*, 29 :1281–1296, 2001.

[JST01] Philippe Jacquet, Wojciech Szpankowski, and Jim Tang. Average profile of the Lempel-Ziv parsing scheme for a Markovian source. *Algorithmica*, 31(3) :318–360, 2001. Mathematical analysis of algorithms.

[Ker00] Christine Keribin. Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, 62(1) :49–66, 2000.

[Kie78] John C. Kieffer. A unified approach to weak universal source coding. *IEEE Trans. Inform. Theory*, 24(6) :674–682, 1978.

[Kie93] John C. Kieffer. Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Inform. Theory*, 39 :893–902, 1993.

[Kon98] Ioannis Kontoyiannis. Asymptotic recurrence and waiting times for stationary processes. *J. Theoret. Probab.*, 11(3) :795–811, 1998.

[Kos01] Timo Koski. *Hidden Markov Models For Bioinformatics.* Kluwer Academic Publishers Group., 2001.

[Kra49] L. G. Kraft. A device for quantizing, grouping, and coding amplitude modulated pulses. Master's thesis, Dept. of Electrical Engineering, M.I.T., Cambridge, MA, 1949.

[KSY04] John C. Kieffer, Wojciech Szpankowski, and En-hui Yang. Problems on sequences : information theory and computer science interface. *IEEE Trans. Inform. Theory*, 50(7) :1385–1392, 2004.

[KT81] Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Trans. Inform. Theory*, 27(2) :199–207, Mar 1981.

[KV94] Ghassan K. Kaleh and Robert Vallet. Joint parameter estimation and symbol detection for linear or nonlinear unknown channels. *IEEE Trans. Commun.*, 42 :2406–2413, 1994.

[KY00] John C. Kieffer and En-hui Yang. Grammar-based codes : a new class of universal lossless source codes. *IEEE Trans. Inform. Theory*, 46(3) :737–754, 2000.

[Ler92a] Brian G. Leroux. Consistent estimation of a mixing distribution. *Ann. Statist.*, 20(3) :1350–1360, 1992.

[Ler92b] Brian G. Leroux. Maximum-likelihood estimation for hidden markov models. *Stochastic Processes Their Applic.*, 40 :127–143, 1992.

[LN94] Chuang-Chun Liu and Prakash Narayan. Order estimation and sequential universal data compression of a hidden markov source by the method of mixtures. *IEEE Trans. Inf. Theory*, 40(4) :1167–1180, 1994.

[LRS83] Stephen E. Levinson, Lawrence R. Rabiner, and Man Mohan Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, 62 :1035–1074, 1983.

[LS97] Guy Louchard and Wojciech Szpankowski. On the average redundancy rate of the Lempel-Ziv code. *IEEE Trans. Inform. Theory*, 43(1) :2–8, 1997.

[MAS06] Pascal MASSART. *Concentration inequalities and model selection.* Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics, Springer, 2006.

[MF95] Neri Merhav and Meir Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Trans. Inform. Theory*, 41(3) :714–722, May 1995.

[ML03] Elías Moreno and Brunero Liseo. A default Bayesian test for the number of components in a mixture. *J. Statist. Plann. Inference*, 111(1-2) :129–142, 2003.

[MP00] Geoffrey J. McLachlan and David Peel. *Finite mixture models.* Wiley-Interscience, New York, 2000.

[MR96] Kerrie L. Mengersen and Christian P. Robert. Testing for mixtures : a Bayesian entropy approach. In J. O. Berger, J. M. Bernardo, and A. P.. Dawid, editors, *Bayesian Statistics 5*, pages 255–276. Oxford Science Publications, 1996.

[OS04] Alon Orlitsky and Narayana P. Santhanam. Speaking of infinity. *IEEE Trans. Inform. Theory*, 50(10) :2215–2230, 2004.

[OSVZ04] Alon Orlitsky, Narayana P. Santhanam, K. Viswanathan, and Junan Zhang. Limit results on pattern entropy of stationary processes. In *Proceedings of the 2004 IEEE Information Theory workshop*, San Antonio, Texas, USA, oct 2004.

[OSZ04] Alon Orlitsky, Narayana P. Santhanam, and Junan Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Trans. Inform. Theory*, 50(7) :1469–1481, 2004.

[PS98] George Pólya and Gabor Szegő. *Problems and theorems in analysis. II.* Classics in Mathematics. Springer-Verlag, Berlin, 1998. Theory of functions, zeros, polynomials, determinants, number theory, geometry, Translated from the German by C. E. Billigheimer, Reprint of the 1976 English translation.

[Ris76] Jorma Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM J. Res. Dev.*, 20(3), 1976.

[Ris78] Jorma Rissanen. Modelling by shortest data description. *Automatica*, 14 :465–471, 1978.

[Ris83] Jorma Rissanen. A universal data compression system. *IEEE Transactions on Information Theory*, 29, september 1983.

[Ris84] Jorma Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, 30(4) :629–636, 1984.

[Ris86] Jorma Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3) :1080–1100, 1986.

[Ris99] Jorma Rissanen. Fast universal coding with context models. *IEEE Trans. Inform. Theory*, 45(4) :1065–1071, 1999.

[RL81] Jorma Rissanen and Glen Jr. Langdon. Universal modeling and coding. *IEEE Trans. Inform. Theory*, 27(1) :12–23, Jan 1981.

[Rya79] B. Ya. Ryabko. Coding of a source with unknown but ordered probabilities. *Problems Inform. Transmission*, 15(2) :71–77, 1979.

[Rya81] Boris Ya. Ryabko. Comments on : "A source matching approach to finding minimax codes" [IEEE Trans. Inform. Theory **26** (1980), no. 2, 166–174; MR 81c :94021] by L. D. Davisson and A. Leon-Garcia. *IEEE Trans. Inform. Theory*, 27(6) :780–781, 1981.

[Rya84] B. Ya. Ryabko. Twice-universal coding. *Problemy Peredachi Informatsii*, 20(3) :24-28, 1984.

[Sch78] Gedeon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27 :379–423, 623–656, 1948.

[Sha03a] Gil I. Shamir. On the entropy of patterns of i.i.d. sequences. In *Proceedings of 41st Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, USA, oct 2003.

[Sha03b] Gil I. Shamir. Universal lossless compression with unknown alphabets - the average case. *IEEE Trans. Inform. Theory*, submitted in June 2003.

[Sha04] Gil I. Shamir. A new redudancy bound for universal lossless compression of unknown alphabets. In *Proceedings of the 38th Annual Conference on Information Sciences and Systems - CISS*, pages 1175–1179, Princeton, New-Jersey, USA, march 2004.

[Sha06] Gil I. Shamir. On the MDL principle for i.i.d. sources with large alphabets. *IEEE Trans. Inform. Theory*, 52 :1939 – 1955, may 2006.

[Shi93] Paul C. Shields. Universal redundancy rates do not exist. *IEEE Trans. Inform. Theory*, 39(2) :520–524, 1993.

[Sht87] Yuri Shtar'kov. Universal sequential coding of individual messages. *Problemy Peredachi Informatsii*, 23(3) :3–17, 1987.

[Sio58] Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8 :171–176, 1958.

[SW95] Paul Shields and Benjamin Weiss. Universal redundancy rates for the class of B-processes do not exist. *IEEE Trans. Inform. Theory*, 41(2) :508–512, 1995.

[Sze51] George Szekeres. An asymptotic formula in the theory of partitions. *Quart. J. Math. Oxford*, 2 :85–108, 1951.

[Szp01] Wojciech Szpankowski. *Average case analysis of algorithms on sequences.* Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2001. With a foreword by Philippe Flajolet.

[TSM85] Donald M. Titterington, Adrian F. M. Smith, and Udi E. Makov. *Statistical analysis of finite mixture distributions.* John Wiley & Sons Ltd., Chichester., 1985.

[Tsy04] Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications].* Springer-Verlag, Berlin, 2004.

[Var82] Yehuda Vardi. Nonparametric estimation in renewal processes. *Ann. Statist.*, 10(3) :772–785, 1982.

[vdV98] Aad W. van der Vaart. *Asymptotic statistics.* Cambridge University Press., 1998.

[Wel84] Terry A. Welch. A technique for high performance data compression. *IEEE Computer*, 17(6) :8–19, 1984.

[Wil94] Frans M.J. Willems. The Context-tree Weighting Method : Extensions. In *IEEE Int. Symp. Information Theory*, Trondheim, Norway, 1994.

[WMF94] Marcelo J. Weinberger, Neri Merhav, and Meir Feder. Optimal sequential probability assignment for individual sequences. *IEEE Transactions on Information Theory*, 40, Mar 1994.

[WT95] Yuri M. Willems, Frans M. J.; Shtarkov and Tjalling J. Tjalkens. The Context-tree Weighting Method : basic properties. *IEEE Transactions on Information Theory*, 41, May 1995.

[XB97] Qun Xie and A.R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Trans. Inform. Theory*, 43(2) :646–657, Mar 1997.

[XB00] Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Trans. Inform. Theory*, 46 :431–445, 2000.

[ZL77] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Information Theory*, IT-23(3) :337–343, 1977.

[ZL78] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, 24(5) :530–536, 1978.

[ÄSMS97] Jan Äberg, Yuri M. Shtarkov, and Ben J. M. Smeets. Multialphabet coding with separate alphabet description. In IEEE Computer Society Press, editor, *Proceedings of Compression and Complexity of SEQUENCES*, pages 56–65, 1997.

Table des figures

N° d'impression 2771
3ème trimestre 2006

Résumé : Ce travail de thèse explore quelques aspects contemporains de la théorie de l'information allant de la théorie du codage à certains problèmes de choix de modèles. Nous y considérons d'abord le problème du codage de sources sans mémoire émettant dans un alphabet infini dénombrable. Comme il est impossible d' y apporter une solution générale et efficace, deux approches sont utilisées : dans un premier temps nous établissons des conditions dans lesquelles le taux entropique peut être approché, et nous nous restreignons à des classes pour lesquelles les queues de probabilités sont contrôlées. Dans un second temps, il n'est posé aucune restriction sur la source, il est possible de fournir une solution partielle en codant seulement une partie de l'information – le motif – qui capture les répétitions contenues dans le message.

Pour arriver à l'étude de processus plus complexes, nous revenons sur le cas de sources à mémoire finie sur un alphabet fini, qui a donné lieu a beaucoup de travaux, ainsi qu'à des algorithmes efficaces comme la Context Tree Weighting (CTW) Method. Nous prouvons ici que cet algorithme est également efficace sur une classe non paramétrique de sources à mémoire infinie : les sources de renouvellement.

Nous montrons ensuite que les idées sous-jacentes à la méthode CTW permettent de construire un estimateur consistant de la structure de mémoire d'un processus quand celle-ci est finie : nous complètons l'étude de l'estimateur BIC pour les chaînes de Markov à longueur variable. Dans une dernière partie, il est montré qu'une telle approche est généralisable dans un cadre plus large de sources émettant dans un alphabet infini, dénombrable ou non. On obtient ainsi des estimateurs consitants de l'ordre de chaînes de Markov cachées à émission poissonienne et gaussienne.

Abstract : This thesis explores some contemporary aspects of information theory, from source coding to issues of model selection. We first consider the problem of coding memoryless sources on a countable, infinite alphabet. As it is impossible to provide a solution which is both efficient and general, two approaches are considered : we first establish conditions under which the entropic rate can be reached, and we consider restricted classes for which tail probabilities are controlled. The second approach does not set any condition on the sources but provides a partial solution by coding only a part of the information – the pattern – which captures the repetitions in the message.

In order to study more complex processes, we come back to the case of finite memory sources on a finite alphabet : it has given rise to many works and efficient algorithms like the Context Tree Weighting (CTW) Method. We show here that this method is also efficient on a non-parametric class of infinite memory sources : the renewal processes.

We show then that the ideas on which CTW is based lead to a consistent estimator of the memory structure of a process, when this structure is finite. In fact, we complete the study of the BIC context tree estimator for Variable Length Markov Chains. In the last part, it is shown how similar ideas can be generalized for more complex sources on a (countable or not) infinite alphabet. We obtain consistent estimators for the order of hidden Markov models with Poisson and Gaussian emission.

Keywords : BIC, Context trees, CTW, HMM, infinite alphabet, MDL, minimax, pattern coding, renewal processes, source coding, VLMC.
AMS Classification : 62M05, 62P30, 62L12, 68P30, 68Q17, 94A15, 94A24, 94A29, 94A45.