Université Paris-Sud
Laboratoire de Mathématiques

Mémoire de Stage de Master

Redondance bayésienne et minimax, Sources stationnaires sans mémoire en alphabet infini

Dominique Bontemps

14 septembre 2007

Sous la direction de Mme Elisabeth Gassiat

Remerciements

Pour l'écriture de ce document je suis particulièrement redevable à Elisabeth Gassiat, mon maître de stage, qui m'a guidé dans la compréhension des différents articles que j'ai étudiés, et dans l'obtention des applications nouvelles qui sont présentées ici.

Merci aussi à mes camarades qui ont écoutées mes questions techniques et m'ont aidé à les résoudre, en particulier à Laurent Veysseire et à Fabrice Mathurin.

Résumé

On dispose d'une famille de lois de probabilité $(\boldsymbol{P}_\theta)$ appelées *sources*, qui produisent des signaux indépendants identiquement distribués sur un alphabet A dénombrable, et on désire trouver un code qui compresse sans perte ces signaux de manière optimale. La qualité de la compression est mesurée par la *redondance*, ou entropie relative.

Le but de ce mémoire est d'explorer dans l'optique du codage / compression de données les applications possibles des encadrements de la *redondance bayésienne* et de la *redondance minimax* fournis par l'article « Mutual Information, Metric Entropy and Cumulative Relative Entropy Risk » de Haussler et Opper ([HO97]). Ceux-ci exploitent des transformées de Laplace de la *distance de Hellinger*, et peuvent être reliées à l'*entropie métrique* de la famille des sources pour cette même distance.

En combinant ces résultats avec des techniques issues des travaux de Xie et Barron sur le mélange de Dirichlet en alphabet fini, nous serons en mesure de fournir un résultat asymptotique nouveau, en exprimant un équivalent de la redondance minimax de certaines classes de lois en alphabet infini : les *classes enveloppe à décroissance exponentielle*.

Table des matières

1 Introduction

Ce mémoire est motivé principalement par le codage (compression) de sources sans mémoire en alphabet infini, mais le modèle utilisé est plus général.

Supposons que nous disposons d'une famille paramétrée de sources $\Lambda = (\boldsymbol{P}_\theta)_{\theta\in\Theta}$ qui produisent un signal $\boldsymbol{X} = (X_n)_{n\geq 1} \in A^{\mathbb{N}^*}$ à valeur dans un alphabet A fini ou dénombrable [1]. En pratique, si A est fini de taille k il sera identifié à $\{1, 2, \ldots, k\}$, et si A est infini il sera identifié à $\mathbb{N}^*$.
Sous la loi $\boldsymbol{P}_\theta$ la loi de X_1 est notée P_θ, et la loi de $X_{1:n} = (X_1, \ldots, X_n)$ est P_θ^n. Nous nous plaçons dans l'hypothèse de sources stationnaires sans mémoire : le processus $\boldsymbol{X}$ est i.i.d. sous $\boldsymbol{P}_\theta$, et $P_\theta^n = {P_\theta}^{\otimes n}$.
D'autre part nous notons ν la mesure de comptage sur A [2] ; pour tout $\theta \in \Theta$, P_θ est absolument continue devant ν, et nous noterons $\mathrm{d}P_\theta$ sa densité par rapport à ν [3].
Enfin, nous supposons que si $\theta \neq \theta^*$ sont deux éléments de Θ, $\boldsymbol{P}_\theta$ et $\boldsymbol{P}_{\theta^*}$ sont distincts. Nous munissons Θ de la topologie de la convergence faible des lois $\boldsymbol{P}_\theta$, et de la tribu borélienne associée ; par la suite les priors définis sur Θ seront des mesures par rapport à cette tribu.

Dans ce contexte, nous sommes intéressés par les problématiques suivantes : premièrement, trouver un code qui atteigne un « bon » taux de compression quel que soit $\theta \in \Theta$ (on ne connaît pas la « vraie » loi $\boldsymbol{P}_{\theta^*}$, on sait seulement que $\theta^* \in \Theta$) ; deuxièmement, estimer quel est le meilleur taux de compression que l'on puisse espérer sur toute la famille Λ.
Les codes concernés sont les codes binaires (le message est codé par une suite de bits 0 ou 1) uniquement décodables : chaque message codé correspond à un unique message non codé.
Donner un sens mathématique à ces questions nécessite de définir la notion de redondance, et de rappeler la correspondance entre codage et loi de probabilité.

Avant de rentrer plus avant dans la problématique du codage de données, il nous faut remarquer que le modèle développé ici rentre dans une problématique statistique plus large d'estimation avec perte logarithmique (ou d'entropie relative), dont les applications

[1] Dans un modèle plus général, on demande que le processus $\boldsymbol{X}$ soit à valeurs dans un espace métrique $\mathcal{Y}$ séparable et complet. Voir [HO97] pour plus de précisions.

[2] Dans le cas où $\boldsymbol{X}$ est à valeurs dans un espace métrique $\mathcal{Y}$, il faut supposer qu'il existe une mesure σ-finie ν qui domine toute la classe $\Lambda^1 = (P_\theta)_{\theta\in\Theta}$ des premières marginales. Toutes les mesures sur $\mathcal{Y}$ considérées par la suite sont alors supposées absolument continues par rapport à ν. Le choix d'un tel ν n'a pas d'influence sur les résultats. [HO97] montre que cette condition est en fait plus faible que la simple existence de la redondance minimax, qui est la quantité que nous voulons étudier dans ce travail.

[3] Ceci afin de préserver la notation intégrale, valable lorsque $\boldsymbol{X}$ est à valeurs dans un ensemble $\mathcal{Y}$ non nécessairement dénombrable.

s'étendent de la théorie de l'information (capacité d'un canal informatique) aux mathématiques financières (gestion de portefeuilles), aux théories de l'apprentissage (estimation de la loi d'une source) et des jeux, en incluant bien sûr la compression de données y compris de sources avec mémoire. Les outils présentés et utilisés dans ce mémoire sont utilisables et sont même souvent apparus dans ces autres domaines. Nous invitons à se reporter à [HO97], qui a fournit la base de notre travail et qui est rédigé dans cette optique plus large.

1.1 Longueur de code

Nous nous intéressons maintenant aux liens entre longueur de code et loi de probabilité. Ces liens sont illustrés par les théorèmes ci-après. Il s'agit de résultats classiques dont la démonstration n'est pas donnée ici — les personnes intéressées peuvent se reporter par exemple à [CT91].

Nous utilisons les notations suivantes : Si $\mathcal{X}$ est un ensemble fini ou dénombrable, $\mathcal{X}^* = \bigcup_{m\in\mathbb{N}} \mathcal{X}^m$ est l'ensemble des mots formés à partir de l'alphabet $\mathcal{X}$. Un code est une application $B \subset A^* \to \{0,1\}^*$, et la longueur de code $l(\cdot)$ est l'application qui à un mot de code associe le nombre de bits qui le composent.

Théorème 1 (association d'une sous-probabilité à un code)**.** *Si f est un code uniquement décodable de $B \subset A^*$ dans $\{0,1\}^*$, alors*

$$\sum_{x\in B} 2^{-l(f(x))} \leq 1.$$

Ceci permet d'associer à un code f une sous probabilité $Q_f(x) = 2^{-l(f(x))}$.

Théorème 2 (code de Shannon)**.** *Soit Q une sous-probabilité sur l'ensemble $B \subset A^*$. Alors il existe un code f uniquement décodable tel que*

$$\forall x \in B, \quad l(f(x)) = \lceil -\log_2 Q(x) \rceil.$$

En réalité on connaît bien un code qui réalise ce théorème : c'est le code de Shannon qui est en outre un code préfixe [4].

Pour la suite, l'optique d'un codage binaire nous amène à utiliser le logarithme en base 2 ; nous le noterons « log », et nous utiliserons « ln » pour le logarithme népérien [5].

Ces deux théorèmes permettent de faire le lien entre longueur de code et probabilité d'un message. En particulier ils permettent d'associer un code à une loi de probabilité. Remarquons qu'un code associé à une sous-probabilité peut être amélioré en un code associé à une probabilité.

Lorsqu'on s'intéresse à la longueur moyenne de code on voit apparaître l'entropie de la source.

[4] pour des précisions sur les codes préfixes, voir [CT91].

[5] Si cette manière de faire a clairement son intérêt lorsqu'on parle de codage, elle a aussi son inconvénient lorsqu'on veut utiliser des résultats portant sur l'entropie relative qui ne sont pas spécifiques au codage de données, comme ce sera le cas dans nos sections 2 et 3.

Définition 1 (entropie binaire)**.** *Soit P une loi de probabilité sur l'alphabet A. L'entropie de P, ou entropie binaire, est définie par*

$$H(P) = -\sum_{x\in A} P(x)\log P(x) = \mathbb{E}_P[-\log P(X)]$$

Théorème 3 (Shannon-Kraft-Mc Millan)**.** *Soit P^n est loi de probabilité sur A^n. Si f est un code uniquement décodable sur A^n, alors*

$$\mathbb{E}_{P^n} l(f(X_{1:n})) \geq H(P^n).$$

Réciproquement, il existe un code uniquement décodable f (en fait préfixe) tel que

$$\mathbb{E}_{P^n} l(f(X_{1:n})) \leq H(P^n) + 1.$$

On sait construire de tels codes via le théorème 2 (codes de Shannon ou de Huffman) : il s'agit du code associé à la « vraie » loi de probabilité P^n. Cependant ils présentent plusieurs difficultés :

- la connaissance de la vraie loi de codage P^n est requise pour construire un code qui atteint la borne du théorème 3 ;
- ces codes peuvent être trop lourds à implémenter en terme de temps de calculs, et en pratique un soin particulier doit être apporté à chercher des codes raisonnablement implémentables ;
- si la suite des probabilités $(P^n)_{n\in\mathbb{N}}$ à partir desquels on construit les codes n'est pas compatible, il ne peut qu'on ne puisse pas réaliser d'implémentation en ligne qui réutilise le code de $X_{1:n}$ pour coder $X_{1:n+1}$.

La nécessité de fournir des codes qui se comportent bien même lorsque la loi sous-jacente de la source est inconnue nous pousse à introduire la notion de redondance d'un code.

1.2 Redondance d'un code

Désormais nous utilisons la correspondance code/probabilité de la manière suivante : nous recherchons une loi de probabilité Q^n sur A^n telle que $-\log Q^n(X_{1:n})$ soit minimale en un certain sens.
Plus précisément, nous voulons comparer la longueur de code obtenue avec Q^n avec la longueur de code optimale obtenue avec la « vraie » probabilité comme dans le théorème 3. Selon que veut minimiser cet écart en moyenne ou en norme infinie, cela donne lieu à deux mesures de la qualité d'un code : redondance ou risque individuel (regret).

Definition 2. *Soit P^n une loi de probabilité (source) sur A^n. Soit Q^n une autre loi de probabilité sur A^n.*

(*i*) *Le* regret*, ou risque individuel, de la loi Q^n par rapport à la loi P^n est*

$$\sup_{x_{1:n}\in A^n} \log P^n(x_{1:n}) - \log Q^n(x_{1:n}).$$

(*ii*) *La* redondance, *ou redondance moyenne*[6], *de la loi* Q^n *par rapport à la loi* P^n *est définie par*

$$R_n(Q^n, P^n) = \mathbb{E}_{P^n}[-\log Q^n(X_{1:n}) + \log P^n(X_{1:n})] = D(P^n \,\|\, Q^n)$$

où $D(P \,\|\, Q) = (\log e)\, D_{KL}(P \,\|\, Q)$ *désigne la version de la divergence de Kullbach-Liebler, ou entropie relative, qui utilise le logarithme en base* 2[7].

Le regret comme la redondance admettent une écriture sous forme de regret cumulatif (règle de chaînage) ; Dans le cas de la redondance, qui nous intéresse plus particulièrement dans ce mémoire, cela s'écrit :

$$R_n(Q^n, P^n) = \sum_{k=1}^{n} \mathbb{E}_{P^k} D(P(X_k = \cdot|X_{1:k-1}) \,\|\, Q(X_k = \cdot|X_{1:k-1}))$$

où P^k désigne la loi marginale $P_{X_{1:k}}$.
Si la source P^n est une source sans mémoire $P^{\otimes n}$, la formule se simplifie en

$$R_n(Q^n, P^n) = \sum_{k=1}^{n} \mathbb{E}_{P^k} D(P \,\|\, Q(X_k = \cdot|X_{1:k-1})).$$

Nous désirons mesurer le comportement d'un code par rapport à toute la famille $\Lambda = (\boldsymbol{P}_\theta)_{\theta\in\Theta}$ dans laquelle se trouve la vraie loi inconnue de la source. Deux manières de faire : une approche bayésienne et une approche minimax.

Définition 3. (*i*) *Le* regret minimax *de la classe* Λ *est définie par*

$$R_n^*(\Lambda) = \inf_{Q^n} \sup_{P\in\Lambda} \sup_{x_{1:n}\in A^n} \log P^n(x_{1:n}) - \log Q^n(x_{1:n}).$$

(*ii*) *Soit* μ *une loi de probabilité sur* Θ. *La* redondance bayésienne *de la classe* Λ *par rapport au prior* μ *est définie par*

$$R_{n,\mu}^{\text{Bayes}}(\Lambda) = \inf_{Q^n} \int_\Theta D(P_\theta^n \,\|\, Q^n)\, \mathrm{d}\mu(\theta).$$ [8]

(*iii*) *La* redondance minimax *de la classe* Λ *est définie par*

$$R_n^{\text{minimax}}(\Lambda) = \inf_{Q^n} \sup_{\theta\in\Theta} D(P_\theta^n \,\|\, Q^n).$$

à chaque fois l'infimum en Q^n *est pris sur toutes les distributions de probabilité sur* A^n.

[6] Selon un autre terminologie, ce que nous avons appelé *regret* est appelé *redondance individuelle*, et ce que nous avons appelé *redondance* est appelée *redondance moyenne*. Nous avons choisi cette manière de dire pour alléger dans cet exposé le terme *redondance moyenne minimax* en *redondance minimax*.

[7] Rappelons la définition de la divergence de Kullback-Liebler (entropie relative) dans sa version « habituelle » :

$$D_{KL}(P \,\|\, Q) = -\int \mathrm{d}P\, \ln \frac{\mathrm{d}Q}{\mathrm{d}P}.$$

[8] Ailleurs on utilise le terme *redondance bayésienne* pour désigner la *redondance (moyenne) maximin*, qui vaut $\sup_\mu R_{n,\mu}^{\text{Bayes}}$.

Nous avons cité le regret minimax par soucis de complétude, mais dans ce mémoire nous nous focalisons sur la redondance bayésienne et la redondance minimax. Le lien entre les deux est illustré dans le résultat suivant :

Théorème 4. *Soit $(\Theta^*, X_{1:n})$ un vecteur aléatoire à valeurs dans $\Theta \times A^n$, tel que la loi de Θ^* est μ et la loi de $X_{1:n}$ sachant $\Theta^* = \theta^*$ est $P_{\theta^*}^n$. Alors*

(*i*) *L'unique distribution qui réalise l'infimum dans la définition de $R_{n,\mu}^{\text{Bayes}}$ est le mélange de Bayes $M_{n,\mu}$ défini par*

$$M_{n,\mu}(x_{1:n}) = \int_\Theta P_\theta^n(x_{1:n}) \,\mathrm{d}\mu(\theta),$$

et

$$R_{n,\mu}^{\text{Bayes}} = \int_\Theta D(P_\theta^n \,\|\, M_{n,\mu}) \,\mathrm{d}\mu(\theta) = I(\Theta^*, X_{1:n}),$$

où $I(X,Y)$ désigne l'information mutuelle entre les variables X et Y [9].

(*ii*)

$$R_n^{\text{minimax}} = \sup_\mu R_{n,\mu}^{\text{Bayes}},$$

où le supremum est pris sur tous les mesures de probabilités (priors) sur Θ. De plus

$$R_n^{\text{minimax}} = \inf_\mu \sup_{\theta \in \Theta} D(P_\theta^n \,\|\, M_{n,\mu}).$$

La deuxième partie de ce théorème — redondance minimax et redondance maximin sont égales — est la base de la suite de ce mémoire. On pourra se référer à [Hau96] pour sa démonstration dans un cadre très général.

Nous allons dans la section 2 présenter des encadrements de la redondance bayésienne, d'où sont tirées dans la section 3 des encadrements de la redondance minimax. Enfin, dans notre section 4 nous en déduirons diverses applications nouvelles à un type particulier de classes Λ — les classes enveloppe — en nous attachant plus particulièrement aux classes enveloppe à décroissance exponentielle.

[9] L'information mutuelle est logiquement définie ici par $I(X,Y) = D(P_{(X,Y)} \,\|\, P_X \otimes P_Y)$, et elle hérite du facteur $(\log e)$ qui vient de l'utilisation du logarithme en base 2 dans la définition que nous prenons de $D(P \,\|\, Q)$.

2 Encadrer la Redondance bayésienne

Cette section, ainsi que la suivante, est essentiellement basée sur l'article [HO97] qui utilise la métrique de Hellinger et certaines intégrales de Laplace pour fournir des encadrements originaux de la redondance bayésienne et de la redondance minimax. Nous présenterons les résultats eux-mêmes mais laisserons de côté les preuves que l'on peut trouver dans l'article cité [1].

Nous allons dans un premier temps définir les objets mathématiques utilisés : I-divergence, D-divergence et condition de α-affinité ; puis nous exposerons les résultats ; enfin nous en donnerons un exemple d'application aux familles dénombrables.

2.1 α-affinité, I-divergence et D-divergence

Nous allons maintenant définir divers quantités utilisées dans le courant de ce mémoire. Les notations suivront celles de [HO97].

2.1.1 α-affinité

Une condition qui est amenée à jouer un rôle central dans les résultats les plus intéressants de ce mémoire est la condition de α-affinité.

Définition 4 (α-affinité)**.** *Soit $\alpha > 1$, et P et Q deux distributions sur A. l'α-affinité entre P et Q est définie par*

$$\rho_\alpha(P,Q) = \int_A (\mathrm{d}P(x))^\alpha (\mathrm{d}Q(x))^{1-\alpha}\, \mathrm{d}\nu(x).$$

Heuristiquement, l'utilité de faire appel à l'α-affinité peut s'expliquer ainsi : la qualité d'un code dépend de la divergence de Kullbach entre la « vraie » loi de la source et la distribution associée au code. Nous remplacerons la divergence de Kullbach par la distance de Hellinger, mais en faisant cela on perd des informations : il se peut que la divergence de Kullbach soit infinie alors que la distance de Hellinger est finie. Des conditions sur l'α-affinité permettent de palier à ce problème [2].

Le paramètre α étant appelé à jouer un autre rôle dans notre exposé, il est remplacé dans la pratique par $1 + \lambda$.

[1] Si nous avons choisi de ne pas inclure — à quelques exceptions près — les démonstrations de ces résultats c'est pour deux raisons : d'une part elles alourdiraient notre texte, et d'autre part nous pensons que la démonstration originale est très propre. Cependant nous effectuerons ponctuellement certains calculs laissés par les auteurs à la vérification du lecteur.

[2] Ceci est visible en particulier dans le lemme 5 de [HO97].

Les conditions concernant la $(1+\lambda)$-affinité qui apparaîtront par la suite concernent l'existence d'un $\lambda > 0$ pour lequel l'une deux quantités suivantes est finie :

Définition 5. *Soit $\lambda > 0$, et μ un prior.*

(i)
$$R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} = \inf_Q \int_\Theta \rho_{1+\lambda}(P_\theta, Q)\,\mathrm{d}\mu(\theta)$$

(ii)
$$R^{\text{minimax}}_{1,\rho_{1+\lambda}} = \inf_Q \sup_{\theta\in\Theta} \rho_{1+\lambda}(P_\theta, Q)$$

dans les deux cas l'infimum en Q est pris sur toutes les distributions de probabilité sur A.

Les deux quantités sont reliés par la formule $R^{\text{minimax}}_{1,\rho_{1+\lambda}} \geq \sup_\mu R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}$ qui découle de la relation toujours vraie maximin $\leq$ minimax. Mais ici, à la différence de la redondance, il n'y a pas de résultat d'égalité.
Laquelle des deux quantités interviendra dans les théorèmes, cela dépendra de l'objet que l'on veut encadrer : la redondance bayésienne ou la redondance minimax.

Ces quantités méritent que l'on s'y arrête un peu plus avant. Commençons par la version bayésienne.

Condition portant sur risque bayésien

Proposition 1. *Soit μ un prior, et $\lambda > 0$.*

(i) *Supposons que $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$. L'infimum dans la définition de $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}$ est alors réalisé par l'unique mesure $Q = Q_\mu$ définie par*

$$\mathrm{d}Q_\mu(x) = \frac{\left(\int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda}\,\mathrm{d}\mu(\theta)\right)^{1/(1+\lambda)}}{C_{\lambda,\mu}},$$

où
$$C_{\lambda,\mu} = \int_A \left(\int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda}\,\mathrm{d}\mu(\theta)\right)^{1/(1+\lambda)} \mathrm{d}\nu(x)\ ;$$

en outre
$$R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} = C^{1+\lambda}_{\lambda,\mu}.$$

(ii) *Réciproquement, si $C_{\lambda,\mu} < \infty$, alors $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$.*

Démonstration. Supposons que $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$, et posons $\alpha = 1+\lambda$. La définition de $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}$ s'écrit

$$R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} = \inf_Q \int_\Theta \int_A (\mathrm{d}P_\theta(x))^\alpha (\mathrm{d}Q(x))^{1-\alpha}\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(\theta),$$

avec Q variant dans l'ensemble C des mesures de probabilités sur A. Dans le cas particulier d'un alphabet dénombrable, ν est la mesure de comptage et il n'y a pas à justifier que Q est absolument continue devant ν[3].

Sur C nous pouvons définir la fonction f à valeurs dans $[0,\infty]$ par

$$f(Q) = \int_A \left(\int_\Theta (\mathrm{d}P_\theta(x))^\alpha \, \mathrm{d}\mu(\theta) \right) \frac{1}{(\mathrm{d}Q(x))^{\alpha-1}} \, \mathrm{d}\nu(x),$$

et $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}$ s'écrit plus simplement $\inf_{Q\in C} f(Q)$. Comme $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$, il existe $Q \in C$ tel que $f(Q) < \infty$, et pour tout $x \in A$

$$\int_\Theta (\mathrm{d}P_\theta(x))^\alpha \, \mathrm{d}\mu(\theta) < \infty.$$

Quitte à restreindre A, on peut supposer que

$$\int_\Theta (\mathrm{d}P_\theta(x))^\alpha \, \mathrm{d}\mu(\theta) > 0.$$

Pour Q tel que $f(Q) < \infty$, nous posons

$$h_Q(x) = \frac{\mathrm{d}Q(x)}{\left(\int_\Theta (\mathrm{d}P_\theta(x))^\alpha \, \mathrm{d}\mu(\theta)\right)^{1/\alpha}}.$$

Alors

$$\int_\Theta (\mathrm{d}P_\theta(x))^\alpha \, \mathrm{d}\mu(\theta) = \left(\frac{\mathrm{d}Q(x)}{h_Q(x)}\right)^\alpha,$$

et

$$f(Q) = \int_A \left(\frac{1}{h_Q(x)}\right)^\alpha \mathrm{d}Q(x) \, \mathrm{d}\nu(x) < \infty.$$

La fonction $J : x \mapsto x^\alpha$ est convexe et nous pouvons appliquer le théorème de Jensen avec la mesure $\mathrm{d}Q(x)\,\mathrm{d}\nu(x)$:

$$\begin{aligned} f(Q) &\geq J\left(\int_A \frac{\mathrm{d}Q(x)}{h_Q(x)} \, \mathrm{d}\nu(x)\right) \\ &= \left(\int_A \left(\int_\Theta (\mathrm{d}P_\theta(x))^\alpha \, \mathrm{d}\mu(\theta)\right)^{1/\alpha} \mathrm{d}\nu(x)\right)^\alpha \\ &= C^{1+\lambda}_{\lambda,\mu}. \end{aligned}$$

En conséquence $C_{\lambda,\mu} < \infty$ et Q_μ est bien définie. En outre

$$\begin{aligned} f(Q_\mu) &= \int_A \left(\int_\Theta (\mathrm{d}P_\theta(x))^\alpha \, \mathrm{d}\mu(\theta)\right) C^{\alpha-1}_{\lambda,\mu} \left(\int_\Theta (\mathrm{d}P_\theta(x))^\alpha \, \mathrm{d}\mu(\theta)\right)^{\frac{1-\alpha}{\alpha}} \mathrm{d}\nu(x) \\ &= C^{\lambda}_{\lambda,\mu} \int_A \left(\int_\Theta (\mathrm{d}P_\theta(x))^\alpha \, \mathrm{d}\mu(\theta)\right)^{1/\alpha} \mathrm{d}\nu(x) \\ &= C^{1+\lambda}_{\lambda,\mu}. \end{aligned}$$

[3] Si ce n'était pas le cas, il faudrait en outre changer les « pour tout x » en « pour presque tout x ».

En conséquence $f(Q) \geq f(Q_\mu)$ pour tout $Q \in C$, Q_μ réalise le minimum de f et $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} = C^{1+\lambda}_{\lambda,\mu}$.

Réciproquement $C_{\lambda,\mu} < \infty$ entraîne $f(Q_\mu) < \infty$ et $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$. □

Condition portant sur risque minimax

Si nous nous intéressons maintenant à la condition « il existe $\lambda > 0$ tel que $R^{\text{minimax}}_{1,\rho_{1+\lambda}} < \infty$ », il faut remarquer qu'elle est moins forte qu'une autre condition que l'on rencontre dans différents travaux, et qui est en particulier vérifiée pour les classes enveloppe traitées dans la section 4 : la condition d'enveloppe intégrable $\int_A (\sup_{\theta\in\Theta} \mathrm{d}P_\theta(x))\,\mathrm{d}\nu(x) < \infty$ [4]. Voyons cela de plus près.

Supposons qu'il existe une fonction g sur A d'intégrale finie par rapport à ν, et telle que pour tout $\theta \in \Theta$ et tout $x \in A$, $\mathrm{d}P_\theta(x) \leq g(x)$ [5]. Cette condition est en particulier vérifiée si A est finie [6]. On note $S(\Theta) = \int_A g(x)\,\mathrm{d}\nu(x)$. Soit Q la distribution sur A qui admet pour densité par rapport à ν la fonction $g/S(\Theta)$, et $\lambda > 0$.
On a alors, pour tout $\theta \in \Theta$,

$$\begin{aligned}\rho_{1+\lambda}(P_\theta, Q) &\leq \int (\mathrm{d}P_\theta) S(\Theta)^\lambda \,\mathrm{d}\nu \\ &\leq S(\Theta)^\lambda,\end{aligned}$$

d'où $R^{\text{minimax}}_{1,\rho_{1+\lambda}} < \infty$.

On pourra trouver dans [HO97] une discussion plus complète au sujet des différentes conditions qui apparaissent chez les auteurs traitant des sujets liés à l'entropie relative.

2.1.2 I-divergence et D-divergence

L'objectif final est déterminer la redondance de la classe Λ en fonction de ses propriétés géométriques par rapport à la distance de Hellinger. Pour passer de la divergence de Kullbach — qui dans la définition de la redondance mesure la « distance » entre deux probabilités — à la distance de Hellinger, nous avons besoin d'introduire deux familles de divergences liées à l'α-affinité : l'I-divergence et la D-divergence.

Définition 6. *Soit P et Q deux mesures de probabilité sur A.*

(i) Soit $\alpha \neq 1$. L'I-divergence d'ordre α entre P et Q est définie par

$$I_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \ln \rho_\alpha(P, Q) = \frac{1}{\alpha - 1} \ln \int (\mathrm{d}P)^\alpha (\mathrm{d}Q)^{1-\alpha}.$$

[4] A fortiori la condition d'enveloppe intégrable entraîne que pour tout $\lambda > 0$ et tout prior μ, $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$.

[5] Cette façon de faire résout la question de la mesurabilité de $\sup_{\theta\in\Theta} \mathrm{d}P_\theta$ dans le cas où $\boldsymbol{X}$ est à valeurs dans un ensemble $\mathcal{Y}$ non nécessairement dénombrable. Pour obtenir une quantité qui ne dépend pas du choix arbitraire de g ni de ν, on peut aussi prendre l'infimum sur les fonctions g mesurables qui dominent la famille $(\mathrm{d}P_\theta)_{\theta\in\Theta}$.

[6] Dans le cas où le processus est à valeurs dans un espace complet $\mathcal{Y}$ (voir la note 1 page 5), l'hypothèse correspondante est que $\mathcal{Y}$ est un partie bornée de $\mathbb{R}^k$.

(*ii*) *Soit* $0 < \alpha < 1$. *La D-divergence d'ordre* α *entre* P *et* Q *est définie par*

$$D_\alpha(P,Q) = \frac{1}{1-\alpha}\left(1-\rho_\alpha(P,Q)\right) = \frac{1}{1-\alpha}\left(1-\int (\mathrm{d}P)^\alpha (\mathrm{d}Q)^{1-\alpha}\right).$$

(*iii*) *Pour* $\alpha = 1$, *on définit*

$$D_1(P,Q) = I_1(P\,\|\,Q) = (\ln 2)\, D(P\,\|\,Q)\ [7] = \int \left(\mathrm{d}Q - \mathrm{d}P - \mathrm{d}P \ln \frac{\mathrm{d}Q}{\mathrm{d}P}\right).$$

En écrivant

$$D_\alpha(P,Q) = \frac{1}{1-\alpha}\int \left(\alpha \mathrm{d}P + (1-\alpha)\mathrm{d}Q - (\mathrm{d}P)^\alpha (\mathrm{d}Q)^{1-\alpha}\right),$$

on obtient facilement $D_\alpha(P,Q) \geq 0$ à partir de l'inégalité $\alpha x + (1-\alpha)y - x^\alpha y^{1-\alpha} \geq 0$ valable pour tout $x, y \geq 0$ et $0 < \alpha < 1$.
A partir de $-\ln x \geq 1 - x$ on a $I_\alpha(P\,\|\,Q) \geq D_\alpha(P,Q)$, d'où $I_\alpha(P\,\|\,Q) \geq 0$.
Remarquons que les deux quantités sont proches lorsque $\rho_\alpha(P,Q)$ est proche de 1.

Un cas particulier important est le carré de la distance de Hellinger, obtenu lorsque $\alpha = 1/2$:

$$D_{1/2}(P,Q) = D^2_{HL}(P,Q) = \int \left(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\right)^2 \mathrm{d}\nu.$$

A la différence des autres distances ou divergences dont on a parlé jusqu'ici, la distance de Hellinger D_{HL} est une distance au sens propre (une métrique) : elle est symétrique et vérifie l'inégalité triangulaire. De nombreux auteurs [8] l'ont utilisée pour encadrer le risque de procédures d'estimation en statistique.
L'intérêt de cette valeur particulière de α vis à vis des autres possibles sera discuté plus en détail à la fin de la partie 2.2 (page 21).

Les résultats que nous présenterons dans la section 3 ferons le lien entre la redondance minimax de la classe Λ et ses propriétés métriques pour la distance de Hellinger.

2.2 Les résultats

Nous allons donner ici des résultats concernant la redondance bayésienne $R^{\mathrm{Bayes}}_{n,\mu}(\Lambda)$ en fonctions du logarithme de la transformée de Laplace de l'I-divergence, puis de la D-divergence. Commençons donc par fixer un prior μ sur Θ et un α dans $(0,1)$.

Les résultats originaux de [HO97] donnent non seulement des encadrements de $R^{\mathrm{Bayes}}_{n,\mu}$, mais aussi des encadrements de la redondance ponctuelle $R_{n,P^{\mathrm{Bayes}}_\mu}(\theta)$ de la stratégie de

[7] La présence du facteur $\ln 2$ est rendue nécessaire par la définition de la divergence de Kullbach donnée dans l'introduction (définition 2), qui utilise le logarithme en base 2. Le même problème surviendra dans l'énoncé des différents résultats des sections 2 et 3. On remarquera au passage que l'article [HO97] a été rédigé en utilisant le logarithme népérien et évite donc ce problème.
[8] [HO97] cite en particulier des travaux de Le Cam, Birgé, Hasminskii et Ibragimov, et van der Geer.

Bayes $P_\mu^{\text{Bayes}} = M_{n,\mu}$, valables pour θ variant dans des ensembles de grande probabilité pour μ. Rappelons que

$$R_{n,P_\mu^{\text{Bayes}}}(\theta) = D(P_\theta^n \,\|\, M_{n,\mu}),$$

où P_θ est la vraie loi de la source.

Cependant nous ne présenterons ici que les points qui concernent la redondance bayésienne, qui nous intéresse davantage en vue des sections suivantes.

Théorème 5. *Soit μ une mesure a priori sur Θ, et soit $0 < \alpha < 1$. Pour tout $\theta \in \Theta$, soit Q_θ une distribution de probabilité arbitraire sur A* [9] *sachant θ. Pour tout $n \geq 1$,*

$$\begin{aligned} -\int_\Theta \mathrm{d}\mu(\theta) \log \int_\Theta \mathrm{d}\mu(\tilde{\theta}) \exp[-n(1-\alpha) I_\alpha(P_\theta \,\|\, P_{\tilde{\theta}})] \\ \leq R_{n,\mu}^{\text{Bayes}}(\Lambda) \\ \leq -\int_\Theta \mathrm{d}\mu(\theta) \log \int_\Theta \mathrm{d}\mu(\tilde{\theta}) \exp[-n I_1(P_\theta \,\|\, Q_{\tilde{\theta}})]. \end{aligned}$$

En utilisant la relation $I_\alpha(P \,\|\, Q) \geq D_\alpha(P, Q)$, on peut remplacer dans la minoration I_α par D_α, et ainsi obtenir la forme de la qui nous intéressera en fin de compte. En revanche dans la majoration, si on voit bien apparaître la même structure (intégrale du logarithme d'une transformée de Laplace) que dans la minoration, la divergence de Kullbach est encore présente. L'objet des résultats qui viennent sera de la transformer en une divergence D_α. L'introduction d'une nouvelle famille Q_θ permettra justement de faire cette transformation ; bien sûr on peut aussi obtenir un cas particulier important en choisissant $Q_\theta = P_\theta$.

Une première variante du théorème utilisant la divergence D_α est obtenue de la manière suivante. Pour $0 < \alpha < 1$ et $x > 0$, on définit

$$b_\alpha(x) = \frac{(1-\alpha)(x - \ln x - 1)}{\alpha + (1-\alpha)x - x^{1-\alpha}},$$

ainsi que $b_\alpha(0) = \infty$. On peut montrer [10] que $b_\alpha(x)$ décroît strictement en x, tend vers 1 à l'infini, et vers l'infini en 0. Soit

$$B_\alpha(\Theta) = \sup_{x \in A,\ \theta^*, \theta \in \Theta} b_\alpha\left(\frac{\mathrm{d}P_{\theta^*}(x)}{\mathrm{d}P_\theta(x)}\right).$$

Corollaire 1. *Pour tout $0 < \alpha < 1$ et $n \geq 1$,*

$$\begin{aligned} -\int_\Theta \mathrm{d}\mu(\theta) \log \int_\Theta \mathrm{d}\mu(\tilde{\theta}) \exp[-n(1-\alpha) D_\alpha(P_\theta, P_{\tilde{\theta}})] \\ \leq R_{n,\mu}^{\text{Bayes}}(\Lambda) \\ \leq -\int_\Theta \mathrm{d}\mu(\theta) \log \int_\Theta \mathrm{d}\mu(\tilde{\theta}) \exp[-n B_\alpha(\Theta) D_\alpha(P_\theta, P_{\tilde{\theta}})]. \end{aligned}$$

[9] Le théorème original est rédigé avec $A = \mathcal{Y}$ un espace métrique séparable et complet, comme signalé dans l'introduction.

[10] Nous le ferons à l'occasion du lemme 1

Ce corollaire découle du lemme suivant, dont nous donnons la démonstration [11].

Lemme 1. *Pour toutes distributions P et Q sur A et pour tout $0 < \alpha < 1$,*

$$I_1(P \,\|\, Q) \leq \left(\sup_{x \in A} b_\alpha \left(\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)} \right) \right) D_\alpha(P, Q).$$

Démonstration. Avant de démontrer le lemme lui-même, nous commençons par montrer que $b_\alpha(x)$ décroît strictement en x, tend vers 1 à l'infini, et vers l'infini en 0.

Les limites s'obtiennent facilement en raisonnant par équivalents :

$$b_\alpha(x) \underset{x \to 0^+}{\sim} \frac{-(1-\alpha)\ln x}{\alpha} \xrightarrow[x \to 0^+]{} \infty \; ;$$

$$b_\alpha(x) \underset{x \to \infty}{\sim} \frac{(1-\alpha)x}{(1-\alpha)x} = 1.$$

En revanche pour la décroissance stricte de b_α nous raisonnons par dérivées successives :

$$b'_\alpha(x) = \frac{1-\alpha}{(\alpha + (1-\alpha)x - x^{1-\alpha})^2} f(x),$$

avec $f(x) = \alpha - \alpha x^{-1} - \alpha x^{1-\alpha} + \alpha x^{-\alpha} + (1-\alpha)\ln x - (1-\alpha)x^{-\alpha}\ln x$ et $b'_\alpha(1) = f(1) = 0$.

$$f'(x) = x^{-1-\alpha} g(x),$$

avec $g(x) = -\alpha(1-\alpha)x + (1-\alpha)x^\alpha + (\alpha - \alpha^2 - 1) + \alpha x^{\alpha-1} + \alpha(1-\alpha)\ln x$ et $f'(1) = g(1) = 0$.

$$g'(x) = \alpha(1-\alpha)x^{\alpha-2} h(x),$$

avec $h(x) = -x^{2-\alpha} + x - 1 + x^{1-\alpha}$. On a $g'(1) = h(1) = 0$ et

$$h'(x) = x^{-\alpha} i(x),$$

avec $i(x) = -(2-\alpha)x + x^\alpha + (1-\alpha)$. On a $h'(1) = i(1) = 0$, $\lim\limits_{x \to 0^+} i(x) = 1 - \alpha > 0$ et

$$i'(x) = x^{\alpha-1} j(x),$$

avec $j(x) = \alpha - (2-\alpha)x^{1-\alpha}$. j est croissante et s'annule seulement au point $\lambda_\alpha = \left(\frac{\alpha}{2-\alpha}\right)^{1/(1-\alpha)}$. En remarquant que les termes mis en facteur dans les calculs de dérivées ci-dessus sont tous strictement positifs sur $(0, \infty)$, on peut alors résumer ce qu'on obtient dans le tableau de variation suivant, dans lequel les signes + et − désignent des valeurs strictement positives ou strictement négatives :

[11] Cette démonstration est simple, et c'est l'occasion de corriger une erreur mineure qui s'est glissée dans le texte original. Nous en profitons pour vérifier les propriétés de b_α laissées dans [HO97] au soin du lecteur. Cependant ces preuves n'ont rien d'essentiel à la compréhension du sujet et peuvent être laissées de côté lors de la lecture.

x	0		λ_α		1		∞
$j(x)$		$+$	0		$-$		
$i(x)$			$+$		0	$-$	
$h(x)$			$-$		0	$-$	
$g(x)$			$+$		0	$-$	
$f(x)$ $b'_\alpha(x)$			$-$		0	$-$	

Ceci démontre bien la décroissance stricte de b_α. Passons maintenant à la démonstration du lemme lui-même.

Si $\mathrm{d}P = \mathrm{d}Q$ presque partout, alors $I_1(P \,\|\, Q) = 0$ et la relation est vraie. On peut donc se restreindre au cas où $D_\alpha(P,Q) > 0$. Soit $S = \{x \in A : \mathrm{d}P(x) = 0\}$.

$$\begin{aligned}
\frac{I_1(P \,\|\, Q)}{D_\alpha(P,Q)} &= \frac{(1-\alpha)\int_{A-S} \mathrm{d}P \left(\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)} - \ln \frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)} - 1\right) + (1-\alpha)\int_S \mathrm{d}Q}{\int_{A-S} \mathrm{d}P \left(\alpha + (1-\alpha)\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)} - \left(\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)}\right)^{1-\alpha}\right) + (1-\alpha)\int_S \mathrm{d}Q} \\
&\leq \frac{\left(\sup_{x\in A} b_\alpha\left(\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)}\right)\right)\int_{A-S} \mathrm{d}P \left(\alpha + (1-\alpha)\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)} - \left(\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)}\right)^{1-\alpha}\right) + (1-\alpha)\int_S \mathrm{d}Q}{\int_{A-S} \mathrm{d}P \left(\alpha + (1-\alpha)\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)} - \left(\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)}\right)^{1-\alpha}\right) + (1-\alpha)\int_S \mathrm{d}Q} \\
&\leq \sup_{x\in A} b_\alpha\left(\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)}\right),
\end{aligned}$$

puisque $b_\alpha \geq 1$. □

Nous allons maintenant présenter un autre résultat utilisant de manière plus fine les divergences D_α grâce à un recours à l'α-affinité.

Théorème 6. *Soit μ une mesure a priori sur Θ, et soit $0 < \alpha < 1$. On suppose que $R^{\mathrm{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$. Pour tout $n \geq 1$, on a alors*

$$\begin{aligned}
&-\int_\Theta \mathrm{d}\mu(\theta) \ln \int_\Theta \mathrm{d}\mu(\tilde\theta) \exp[-n(1-\alpha)D_\alpha(P_\theta, P_{\tilde\theta})] \\
&\quad\leq (\ln 2)\, R^{\mathrm{Bayes}}_{n,\mu}(\Lambda) \\
&\quad\leq -\int_\Theta \mathrm{d}\mu(\theta) \ln \int_\Theta \mathrm{d}\mu(\tilde\theta) \exp\left[-(n\ln n)\frac{4(1-\alpha)(1+o(1))}{\alpha\lambda} D_\alpha(P_\theta, P_{\tilde\theta})\right] \\
&\qquad + R^{\mathrm{Bayes}}_{1,\mu,\rho_{1+\lambda}} + o(1),
\end{aligned}$$

où, pour α et λ fixés, $o(1)$ est à chaque fois une fonction $f(n)$ qui tend vers 0 à l'infini. En outre le même résultat reste valable si on remplace $D_\alpha(P_\theta, P_{\tilde\theta})$ par $I_\alpha(P_\theta \,\|\, P_{\tilde\theta})$.

Vis-à-vis du corollaire 1, l'intérêt du théorème 6 est de supprimer le recours à $B_\alpha(\Theta)$ qui est infini si les densités $\mathrm{d}P_\theta$ ne sont pas bornées et maintenues loin de 0 en tout

point $x \in A$. Cependant le théorème ne prétend pas fournir les meilleurs constantes ni le résultat le plus fin. Nous verrons dans la section suivante que cela suffit déjà pour obtenir des résultats intéressants concernant l'ordre de grandeur de la redondance minimax.

Concernant la condition $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$, nous avons déjà remarqué plus haut [12] qu'elle était vérifiée si la famille Λ admettait une enveloppe intégrable ; ce sera en particulier le cas dans notre section 4. Cependant le théorème s'applique à de nombreux autres cas. Pour caractériser les couples (Θ, μ) couverts ou non par le théorème 6, définissons la fonction $f_{\Theta,\mu}$ à valeurs dans $[0, \infty]$ par

$$f_{\Theta,\mu}(\lambda) = \frac{1}{\lambda} \ln R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}$$

pour $\lambda > 0$ et

$$f_{\Theta,\mu}(0) = (\ln 2)\, R^{\text{Bayes}}_{1,\mu}.$$

Nous montrons ci-dessous que pour tous Θ et μ, si $f_{\Theta,\mu}(\lambda) < \infty$ pour un $\lambda > 0$, alors pour tout $0 < \lambda' < \lambda$, $f_{\Theta,\mu}(\lambda') < \infty$ et

$$\lim_{\lambda \to 0} f_{\Theta,\mu}(\lambda) = f_{\Theta,\mu}(0).$$

En conséquence il y a trois cas possibles :

1. $f_{\Theta,\mu}(\lambda) < \infty$ pour un $\lambda > 0$. Dans ce cas $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$ et donc le théorème 6 s'applique et peut être utilisé pour obtenir des bornes sur $R^{\text{Bayes}}_{n,\mu}$.
2. $f_{\Theta,\mu}(\lambda) = \infty$ pour tout $\lambda > 0$ et $f_{\Theta,\mu}(0) = \infty$ [13]. Dans ce cas $R^{\text{Bayes}}_{n,\mu} = \infty$ pour tout n. le théorème 6 ne s'applique pas mais encadrer $R^{\text{Bayes}}_{n,\mu}$ est trivial.
3. $f_{\Theta,\mu}(\lambda) = \infty$ pour tout $\lambda > 0$ mais $f_{\Theta,\mu}(0) < \infty$. C'est le seul cas non trivial où le théorème 6 ne s'applique pas. Selon la terminologie de [HO97], le couple (Θ, μ) est alors dit irrégulier ; un exemple est proposé plus bas.

Démonstration. Nous donnons une preuve partielle, n'étant pas encore parvenu à justifier une étape des calculs qui suivent.

Rappelons que $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} = C^{1+\lambda}_{\lambda,\mu}$, avec

$$C_{\lambda,\mu} = \int_A \left(\int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda}\, \mathrm{d}\mu(\theta) \right)^{1/(1+\lambda)} \mathrm{d}\nu(x).$$

Soit $\lambda > 0$ tel que $f_{\Theta,\mu}(\lambda) < \infty$. Alors pour ν-presque tout $x \in A$,

$$\left(\int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda}\, \mathrm{d}\mu(\theta) \right)^{1/(1+\lambda)} = \| \theta \mapsto \mathrm{d}P_\theta(x) \|_{1+\lambda} < \infty$$

[12] Lors de la définition de $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}$, page 13.

[13] Selon [HO97] la fonction $f_{\Theta,\mu}$ est en outre croissante — et en conséquence $f_{\Theta,\mu}(0) = \infty$ implique $f_{\Theta,\mu}(\lambda) = \infty$ pour tout $\lambda > 0$ — mais nous n'avons pas pu le vérifier.

en utilisant la norme de $L^{1+\lambda}(\mu)$; pour tout $0 < \lambda' < \lambda$,

$$\|\theta \mapsto \mathrm{d}P_\theta(x)\|_{1+\lambda'} \leq \|\theta \mapsto \mathrm{d}P_\theta(x)\|_{1+\lambda}$$

puisque μ est une mesure de probabilités. On en tire que $C_{\lambda',\mu} \leq C_{\lambda,\mu}$ et $f_{\Theta,\mu}(\lambda') < \infty$.

Montrons que $\lim_{\lambda' \to 0} C_{\lambda',\mu} = 1$. Soit $x \in A$ tel que $\|\theta \mapsto \mathrm{d}P_\theta(x)\|_{1+\lambda} < \infty$. En dominant par la fonction $1 \wedge (\mathrm{d}P_\theta(x))^{1+\lambda}$ on a

$$\lim_{\lambda' \to 0} \int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda'} \,\mathrm{d}\mu(\theta) = \int_\Theta \mathrm{d}P_\theta(x) \,\mathrm{d}\mu(\theta).$$

En dominant l'intégrale sur A par $\|\theta \mapsto \mathrm{d}P_\theta(x)\|_{1+\lambda}$, on obtient

$$\lim_{\lambda' \to 0} C_{\lambda',\mu} = \int_A \left(\int_\Theta \mathrm{d}P_\theta(x) \,\mathrm{d}\mu(\theta) \right) \mathrm{d}\nu(x) = 1$$

par le théorème de Fubini.

En conséquence $\lim_{\lambda \to 0} R^{\mathrm{Bayes}}_{1,\mu,\rho_{1+\lambda}} = 1$, et la règle de l'Hôpital donne

$$\lim_{\lambda \to 0} f_{\Theta,\mu}(\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda} \left(R^{\mathrm{Bayes}}_{1,\mu,\rho_{1+\lambda}} \right) \Big|_{\lambda=0},$$

si cette dernière dérivée existe. Vérifions que c'est le cas et calculons-la.

Soit $\lambda' < \lambda$, et $0 < \epsilon < \lambda - \lambda'$. Soit $x \in A$ tel que $\|\theta \mapsto \mathrm{d}P_\theta(x)\|_{1+\lambda} < \infty$. Posons $M = \sup_{t \geq 1} t^{-\epsilon} \ln t$. En séparant les cas supérieurs et inférieurs à 1, on vérifie que

$$(\mathrm{d}P_\theta(x))^{1+s} |\ln(\mathrm{d}P_\theta(x))| \leq e^{-1} + M(\mathrm{d}P_\theta(x))^{1+\lambda}$$

pour tout $s \in (0, \lambda - \epsilon)$, grâce à la relation $\min_{0<t<1} t \ln t = -e^{-1}$. Grâce à cette domination nous pouvons dériver en λ' :

$$\frac{\mathrm{d}}{\mathrm{d}\lambda'} \left(\int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda'} \,\mathrm{d}\mu(\theta) \right) = \int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda'} \ln(\mathrm{d}P_\theta(x)) \,\mathrm{d}\mu(\theta).$$

Notons maintenant B la quantité

$$\begin{aligned} B(\lambda', x) &= \frac{\mathrm{d}}{\mathrm{d}\lambda'} \left(\int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda'} \,\mathrm{d}\mu(\theta) \right)^{1/(1+\lambda')} \\ &= \frac{1}{(1+\lambda')^2} \left(\int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda'} \,\mathrm{d}\mu(\theta) \right)^{1/(1+\lambda')} \left[-\ln \left(\int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda'} \,\mathrm{d}\mu(\theta) \right) \right. \\ &\quad \left. + (1+\lambda') \left(\int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda'} \,\mathrm{d}\mu(\theta) \right)^{-1} \int_\Theta (\mathrm{d}P_\theta(x))^{1+\lambda'} \ln(\mathrm{d}P_\theta(x)) \,\mathrm{d}\mu(\theta) \right]. \end{aligned}$$

En utilisant le théorème de Jensen avec la fonction convexe $x \ln x$ on obtient que $B(\lambda', x) \geq 0$.

A ce stade la logique voudrait que l'on majore B par une fonction ν-intégrable indépendante de s, au moins pour s variant dans un voisinage de l'origine. *Nous n'y sommes*

pas parvenus. Supposons cependant que nous l'ayons fait. Alors nous pouvons dériver sous le signe intégrale, et

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(C_{\lambda,\mu}\right) = \int_A B(\lambda, x)\,\mathrm{d}\nu(x).$$

Comme d'autre part

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(R^{\mathrm{Bayes}}_{1,\mu,\rho_{1+\lambda}}\right) = C^{1+\lambda}_{\lambda,\mu}\,\ln C_{\lambda,\mu} + (1+\lambda)\,C^{\lambda}_{\lambda,\mu}\,\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(C_{\lambda,\mu}\right),$$

nous pouvons calculer, en utilisant la relation $\int_A \mathrm{d}P_\theta(x)\,\mathrm{d}\nu(x) = 1$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(R^{\mathrm{Bayes}}_{1,\mu,\rho_{1+\lambda}}\right)\Big|_{\lambda=0} &= 0 + \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(C_{\lambda,\mu}\right)\Big|_{\lambda=0} \\
&= \int_A B(0,x)\,\mathrm{d}\nu(x) \\
&= \int_A \left(\int_\Theta \mathrm{d}P_\theta(x)\ln \mathrm{d}P_\theta(x)\,\mathrm{d}\mu(\theta)\right) \\
&\qquad - \left(\int_\Theta \mathrm{d}P_\theta(x)\,\mathrm{d}\mu(\theta)\right)\ln\left(\int_\Theta \mathrm{d}P_\theta(x)\,\mathrm{d}\mu(\theta)\right)\,\nu(x) \\
&= \int_{A\times\Theta} \mathrm{d}P_\theta(x)\ln\frac{\mathrm{d}P_\theta(x)}{\int_A \mathrm{d}P_\theta(x)\,\mathrm{d}\nu(x)\cdot\int_\Theta \mathrm{d}P_\theta(x)\,\mathrm{d}\mu(\theta)}\,\mathrm{d}(\nu\otimes\mu)(x,\theta) \\
&= (\ln 2)\,I(\Theta^*, X_1) \\
&= (\ln 2)\,R^{\mathrm{Bayes}}_{1,\mu}.
\end{aligned}$$

□

Exemple 1. Soit $A = \{1,2,3,\ldots\}$, $\Theta = \{3,4,5,\ldots\}$ et pour tout $\theta \in \Theta$ et tout $x \in A$, $P_\theta(X = x)$ vaut $1 - (1/\ln\theta)$ si $x = 1$, $1/\ln\theta$ si $x = \theta$ et 0 dans les autres cas. On choisit $\mu(\theta) = c/(\theta \ln^2\theta)$, avec $c = \sum_{i=3}^{\infty} 1/(i\ln^2 i)$. Alors (Θ, μ) est irrégulier.

Démonstration. Soit $\lambda > 0$. Calculons $C_{\lambda,\mu}$:

$$\begin{aligned}
C_{\lambda,\mu} &= \sum_{x\geq 1}\left(\sum_{\theta\geq 3}(\mathrm{d}P_\theta(x))^{1+\lambda}\right)^{1/(1+\lambda)} \\
&= \left(c\sum_{\theta\geq 3}\frac{(1-(1/\ln\theta))^{1+\lambda}}{\theta\ln^2\theta}\right)^{1/(1+\lambda)} + \sum_{x\geq 3}\frac{c^{\frac{1}{1+\lambda}}}{x^{\frac{1}{1+\lambda}}(\ln x)^{\frac{3+\lambda}{1+\lambda}}} \\
&= \infty \qquad \text{quelque soit } \lambda > 0.
\end{aligned}$$

La dernière égalité se justifie ainsi : dans la deuxième ligne le premier terme est sommable mais pas le second.

Vérifions maintenant que la redondance bayésienne est finie ; Soit P_μ le mélange bayésien.

$$\begin{aligned} P_\mu(1) &= c\sum_{\theta\geq 3}\frac{1-(1/\ln\theta)}{\theta\ln^2\theta} \\ P_\mu(2) &= 0 \\ P_\mu(x) &= \frac{c}{x\ln^2 x} \qquad \text{pour tout } x\geq 3. \end{aligned}$$

$$\begin{aligned} (\ln 2)\, D(P_\theta \,\|\, P_\mu) &= \left(1-\frac{1}{\ln\theta}\right)\ln\frac{1-(1/\ln\theta)}{P_\mu(1)} + \frac{1}{\ln\theta}\ln\frac{\theta\ln^2\theta}{c} \\ &= \left(1-\frac{1}{\ln\theta}\right)\ln\left(1-\frac{1}{\ln\theta}\right) - \left(1-\frac{1}{\ln\theta}\right)\ln P_\mu(1) + 1 + \frac{2\ln\ln\theta}{\ln\theta} - \frac{1}{\ln\theta}\ln c. \\ R^{\text{Bayes}}_{1,\mu} &= c\sum_{\theta\geq 3}\frac{1}{\theta\ln^2\theta} D(P_\theta \,\|\, P_\mu) \\ (\ln 2)\, R^{\text{Bayes}}_{1,\mu} &= c\sum_{\theta\geq 3}\left[\left(\left(1-\frac{1}{\ln\theta}\right)\ln\left(1-\frac{1}{\ln\theta}\right) - \frac{1}{\ln\theta}\ln c\right)\frac{1}{\theta\ln^2\theta}\right. \\ &\qquad \left. + \left(1-\left(1-\frac{1}{\ln\theta}\right)\ln P_\mu(1)\right)\frac{1}{\theta\ln^2\theta} + 2\frac{\ln\ln\theta}{\theta\ln^3\theta}\right] \\ &< \infty. \end{aligned}$$

En effet chacun des termes de la deuxième ligne se distribue en un terme sommable : le premier est équivalent à $-(1+\ln c)\frac{1}{\theta\ln^3\theta}$, le deuxième est équivalent à $(1-\ln P_\mu(1))\frac{1}{\theta\ln^2\theta}$ et le dernier est négligeable devant $\frac{1}{\theta\ln^2\theta}$. □

Choix $\alpha = 1/2$.

Nous donnons ici un argument heuristique pour justifier l'intérêt du choix $\alpha = 1/2$. Si l'on considère les intégrales qui apparaissent dans le théorème 5, la méthode de Laplace nous enseigne que ce qui compte est le comportement près du maximum de la fonction dont on prend la transformée de Laplace, c'est-à-dire lorsque $I_\alpha(P_\theta \,\|\, P_{\hat\theta})$ est proche de 0. Ce phénomène est particulièrement visible dans le premier exemple d'application que donne [HO97]. Mais il se trouve que lorsque P et Q deviennent proches, au sens où $\mathrm{d}P/\mathrm{d}Q \to 1$ uniformément, alors

$$\frac{I_1(P\,\|\,Q)}{(1-\alpha)I_\alpha(P\,\|\,Q)} \to \frac{1}{\alpha(1-\alpha)}.$$

Pour diminuer l'écart entre la minoration et la majoration il faut donc choisir α de manière à minimiser $\frac{1}{\alpha(1-\alpha)}$, c'est-à-dire prendre $\alpha = 1/2$. A cela s'ajoute l'avantage non négligeable déjà mentionné : $D_{1/2}$ est le carré de la distance de Hellinger, et Θ muni de cette distance devient un espace métrique.

2.3 Exemple d'application : Les familles dénombrables

Dans le cas où la famille de lois Λ est paramétrique et suffisamment régulière on dispose déjà d'estimations de la redondance bayésienne en fonction de l'information de Fisher, grâce notamment aux résultats présentés dans [CB94]. Sous ce rapport les théorèmes 5 et 6 n'apportent pas d'éléments nouveaux [14]. En réalité le vrai intérêt de ces théorèmes est qu'ils s'appliquent aussi à des familles Λ paramétriques mais non régulières, et à des familles non paramétriques. En plus de fournir des encadrements de la redondance bayésienne pour ces familles, cela permettra aussi de déduire les résultats présentés dans la prochaine section.

Nous nous sommes intéressés dans ce mémoire à deux exemples d'application des théorèmes 5 et 6 : les familles dénombrables et les classes enveloppe. L'objet de ce paragraphe est d'étudier les classes dénombrables, tandis que les classes enveloppe, qui sont un exemple de classes « régulières » mais non paramétriques, seront l'objet de la section 4.

Les familles dénombrables ont un comportement particulier qui mérite qu'on s'y arrête. En effet, à condition que la loi μ soit d'entropie finie, la proposition 2 [15] affirme que la redondance bayésienne reste bornée quand n varie. Les différents résultats qui viennent sont intéressants sans être réellement nouveaux [16].

Proposition 2. *On suppose que $\Theta = \{\theta_i\}$ est dénombrable. Soit $(\Theta^*, X_{1:n})$ un vecteur aléatoire à valeurs dans $\Theta \times A^n$, tel que la loi de Θ^* est μ et la loi de $X_{1:n}$ sachant $\Theta^* = \theta_i$ est $P_{\theta_i}^n$. Soit $H(\Theta^*) = -\sum_i \mu(\theta_i) \log \mu(\theta_i)$ l'entropie de la loi μ, à valeurs dans $[0, \infty]$. Pour tout $n \in \mathbb{N}$,*

$$R_{n,\mu}^{\text{Bayes}} = I(\Theta^*, X_{1:n}) \leq H(\Theta^*),$$

et

$$\lim_{n\to\infty} R_{n,\mu}^{\text{Bayes}} = H(\Theta^*).$$

Ce comportement est à rapprocher du celui, déjà connu par ailleurs, de la redondance minimax dans le cas où Θ est fini. Le lien est la relation $\sup_\mu H(\Theta^*) = \log|\Theta|$.

Corollaire 2. *On suppose que Θ est dénombrable. Si Θ est infini, on pose $|\Theta| = \infty$. Alors, pour tout $n \in \mathbb{N}$,*

$$R_n^{\text{minimax}} \leq \log|\Theta|,$$

et

$$\lim_{n\to\infty} R_n^{\text{minimax}} = \log|\Theta|.$$

[14] [HO97] préfère se concentrer sur la redondance ponctuelle de la stratégie de Bayes $R_{n,P_\mu^{\text{Bayes}}}(\theta)$: le théorème 3 — qui correspond à notre théorème 6 — permet d'obtenir un ordre de grandeur de cette quantité déjà connu grâce à [CB90] et [CB94], offrant ainsi un résultat intéressant sans être novateur. Il nous faut mentionner qu'à cette occasion [HO97] illustre l'intérêt de choisir $\frac{1}{2}$ pour valeur de la constante α : cette valeur optimise la minoration de $R_{n,P_\mu^{\text{Bayes}}}(\theta)$ obtenue.

[15] Il s'agit du corollaire 3 de [HO97].

[16] [HO97] cite divers ouvrages contenant des résultats semblables.

Concernant la vitesse de convergence, on peut tirer du théorème 5 le corollaire suivant, valable lorsque Θ est de cardinal fini :

Proposition 3. *On reprend les hypothèses et les notations de la proposition 2. Pour tout* $n \in \mathbb{N}$,

$$H(\Theta^*) - I(\Theta^*, X_{1:n}) \leq (|\Theta| - 1) \left(\max_{1 \leq i < j \leq |\Theta|} \sum_{x \in A} \sqrt{P_{\theta_i}(x) P_{\theta_j}(x)} \right)^n .$$

Comme P_{θ_i} et P_{θ_j} sont supposés différents pour $i \neq j$, $\sum_{x \in A} \sqrt{P_{\theta_i}(x) P_{\theta_j}(x)} < 1$ et la proposition 3 garantit une vitesse de convergence exponentielle.

3 Encadrer la Redondance minimax

Dans cette section nous allons utiliser les résultats de la section précédente sur la redondance bayésienne pour obtenir des encadrements de la redondance minimax. Nous avons déjà exprimé la redondance en fonction d'intégrales de Laplace d'une D_α-divergence, avec le cas particulier $\alpha = 1/2$ qui correspond à la distance de Hellinger ; ici nous irons plus loin dans cette direction, en reliant la redondance minimax aux propriétés métriques de la famille de lois envisagée, plus précisément à l'entropie métrique introduite par Kolmogorov et Tikhomirov dans [KT61].

3.1 Entropie métrique

La distance de Hellinger entre les lois marginales d'ordre 1 P_θ issues de la classe Λ nous fournit une métrique sur l'ensemble Θ, définie par

$$h(\theta^*, \theta) = D_{HL}(P_{\theta^*}, P_\theta),$$

sous l'hypothèse déjà exprimée que P_θ et P_{θ^*} sont distincts si $\theta \neq \theta^*$. Les caractéristiques métriques de l'espace (Θ, h) qui nous intéressent sont les nombres « packing » et « covering »[1] et l'entropie métrique associée.

Définition 7. *Soit (S, ρ) un espace métrique séparable et complet. Soit $\epsilon > 0$.*

(i) Nous notons $\mathcal{D}_\epsilon(S, \rho)$ le cardinal de la plus petite partition finie de S de diamètre au plus ϵ, ou nous posons $\mathcal{D}_\epsilon(S, \rho) = \infty$ si une telle partition finie n'existe pas. S est dit totalement borné *si $\mathcal{D}_\epsilon(S, \rho) < \infty$ pour tout $\epsilon > 0$.*

(ii) L'entropie métrique de (S, ρ) est définie par

$$\mathcal{H}_\epsilon(S, \rho) = \ln \mathcal{D}_\epsilon(S, \rho).$$

(iii) Une ϵ-couverture de S est une partie A de S telle que pour tout $x \in S$ il existe $y \in A$ vérifiant $\rho(x, y) < \epsilon$. Le nombre covering $\mathcal{N}_\epsilon(S, \rho)$ est le cardinal de la plus petite ϵ-couverture finie de S, ou nous posons $\mathcal{N}_\epsilon(S, \rho) = \infty$ si une telle ϵ-couverture finie n'existe pas.

(iv) Une partie ϵ-séparée de S est un sous-ensemble A de S tel que pour tous $x, y \in A$ distincts, $\rho(x, y) > \epsilon$. Le nombre packing $\mathcal{M}_\epsilon(S, \rho)$ est le cardinal de la plus grande partie ϵ-séparée de S, ou nous posons $\mathcal{M}_\epsilon(S, \rho) = \infty$ s'il existe des parties ϵ-séparées de cardinal arbitrairement grand.

[1] Nous avons choisi de ne pas traduire les termes anglais « packing number » et « covering number ».

L'intérêt de ces différentes définitions est donné par le lemme suivant, qui permet de choisir la quantité la plus convenable pour estimer l'entropie métrique :

Lemme 2. *Pour tout $\epsilon > 0$,*

$$\mathcal{M}_{2\epsilon}(S,\rho) \leq \mathcal{D}_{2\epsilon}(S,\rho) \leq \mathcal{N}_{\epsilon}(S,\rho) \leq \mathcal{M}_{\epsilon}(S,\rho).$$

Ce résultat est très classique ; nous n'en donnons pas la démonstration, qui peut se trouver par exemple dans [vdVW96].

Liée à l'entropie métrique on trouve la notion de dimension d'un espace métrique.

Définition 8. *Les dimensions métriques supérieure et inférieure de S sont définies respectivement par :*

$$\overline{\dim}(S) = \limsup_{\epsilon \to 0} \frac{\mathcal{H}_{\epsilon}(S,\rho)}{\ln(1/\epsilon)},$$
$$\underline{\dim}(S) = \liminf_{\epsilon \to 0} \frac{\mathcal{H}_{\epsilon}(S,\rho)}{\ln(1/\epsilon)}.$$

Lorsque $\overline{\dim}(S) = \underline{\dim}(S)$, alors cette quantité est notée $\dim(S)$ et appelée dimension métrique de S. En conséquence

$$\dim(S) = \lim_{\epsilon \to 0} \frac{\mathcal{H}_{\epsilon}(S,\rho)}{\ln(1/\epsilon)}.$$

Nous désirons donc utiliser les résultats de la section 2, avec $\alpha = \frac{1}{2}$. Le lien entre la redondance bayésienne et la redondance minimax est fournie par le théorème 4 au point (ii). Pour passer à l'entropie métrique, commençons par définir

$$b(\epsilon) = \sup \left\{ \frac{I_1(P_\theta \,\|\, P_{\theta^*})}{D_{HL}^2(P_\theta, P_{\theta^*})} \;:\; \theta, \theta^* \in \Theta \text{ et } D_{HL}^2(P_\theta, P_{\theta^*}) \leq \epsilon \right\}.$$

Cette quantité est à comparer avec $B_{1/2}(\Theta)$ utilisé dans le corollaire 1. De fait c'est là le cœur de l'utilisation de l'entropie métrique : utiliser des petites boules de diamètre ϵ dans lesquelles le passage de D_{KL} à D_{HL} ne soit pas trop pénalisant.

Les outils techniques qui vont nous permettre d'encadrer la redondance minimax peuvent maintenant être exprimés : ce sont les deux lemmes qui suivent.

Lemme 3. *On suppose que (Θ, h) est totalement borné. Alors, pour tout $n \geq 1$,*

(i)

$$\begin{aligned}(\ln 2)\, R_n^{\text{minimax}} &\geq \sup_{\epsilon > 0} \left\{ -\ln \left(\frac{1}{\mathcal{M}_{\epsilon}(\Theta, h)} + \exp\left(-\frac{n\epsilon^2}{2} \right) \right) \right\} \\ &\geq \sup_{\epsilon > 0} \min \left\{ \mathcal{H}_{\epsilon}(\Theta, h), \frac{n\epsilon^2}{8} \right\} - \ln 2.\end{aligned}$$

(*ii*)

$$(\ln 2)\, R_n^{\text{minimax}} \leq \inf_{\epsilon>0} \left\{ \mathcal{H}_\epsilon(\Theta, h) + b(\epsilon) n \epsilon^2 \right\} \leq \inf_{\epsilon>0} \left\{ \mathcal{H}_\epsilon(\Theta, h) + B_{1/2}(\Theta) n \epsilon^2 \right\}.$$

Si de plus il existe $\lambda > 0$ *tel que* $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$,

$$(\ln 2)\, R_n^{\text{minimax}} \leq \inf_{\epsilon>0} \left\{ \mathcal{H}_\epsilon(\Theta, h) + \frac{4(1+o(1)) n \ln n\, \epsilon^2}{\lambda} \right\} + R_{1,\rho_{1+\lambda}}^{\text{minimax}} + o(1),$$

où $o(1)$ *est à chaque fois une fonction* $f(n)$ *qui tend vers* 0 *à l'infini.*

Si ce lemme montre bien l'intérêt de faire recours à l'entropie métrique, il n'est pas toujours utilisable en l'état. En effet la dépendance en ϵ de $\mathcal{H}_\epsilon(\Theta, h)$ peut être non continue, et même si elle est continue il n'est pas toujours facile de voir quelle sorte de bornes est obtenue via le lemme 3. Il est alors utile de poser les définitions suivantes : Soit Θ fixé et soit $f_l(x)$ et $f_u(x)$ des fonctions continues et croissantes de $(0, \infty)$ dans $(0, \infty)$, vérifiant

$$\liminf_{\epsilon\to 0} \frac{\mathcal{H}_\epsilon(\Theta, h)}{f_l(1/\epsilon)} \geq 1 \text{ et } \limsup_{\epsilon\to 0} \frac{\mathcal{H}_\epsilon(\Theta, h)}{f_u(1/\epsilon)} \leq 1.$$

Pour tout $s > 0$, soit $\epsilon_l(s)$ l'unique solution de l'équation $f_l(1/\epsilon) = s\,\epsilon^2$, et $\epsilon_u(s)$ l'unique solution de l'équation $f_u(1/\epsilon) = s\,\epsilon^2$. Soit

$$F_l(s) = f_l\left(\frac{1}{\epsilon_l(s)}\right) = s\,\epsilon_l^2(s), \tag{3.1}$$

$$F_u(s) = f_u\left(\frac{1}{\epsilon_u(s)}\right) = s\,\epsilon_u^2(s). \tag{3.2}$$

Ceci nous permet d'énoncer le lemme suivant

Lemme 4. *Soit* F_l *et* F_u *les fonctions définies ci-dessus.*

(*i*)

$$\liminf_{n\to\infty} \frac{R_n^{\text{minimax}}}{(\log e)\, F_l(n/8)} \geq 1$$

(*ii*) *Si* $\lim_{\epsilon\to 0} b(\epsilon) < \infty$, *alors, pour toute fonction* $g(n)$ *vérifiant* $g(n) \xrightarrow[n\to\infty]{} \infty$,

$$\limsup_{n\to\infty} \frac{R_n^{\text{minimax}}}{(\log e)\, F_u(n g(n))} \leq 1.$$

De même, s'il existe $\lambda > 0$ *tel que* $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$, *alors, pour toute fonction* $g(n)$ *comme ci-dessus,*

$$\limsup_{n\to\infty} \frac{R_n^{\text{minimax}}}{(\log e)\, F_u(n g(n) \ln n)} \leq 1.$$

Dans la pratique il est courant que $F_l(n)$ et $F_u(n \ln n)$ soient proches asymptotiquement. Dans ce cas le taux de croissance asymptotique de R_n^{minimax} est essentiellement obtenue en résolvant l'équation $\mathcal{H}_\epsilon(\Theta, h) = n\epsilon^2$. Nous pouvons utiliser cette approche pour obtenir les résultats présentés dans le paragraphes qui vient.

3.2 Les théorèmes sur la redondance minimax

Les résultats présentés ici sont deux théorèmes qui utilisent les lemmes du paragraphe précédent : le premier donne de manière globale le comportement de la redondance minimax en fonction de la dimension métrique de Θ, tandis que le second entre plus en détail dans l'encadrement de la redondance minimax. Nous donnerons en corollaires des versions légèrement modifiées des mêmes résultats pour traiter le cas où on n'a pas à sa disposition l'entropie métrique de Θ, mais seulement des minorations ou des majorations ; ce sont en effet ces versions que nous utiliserons en pratique par la suite.

Nous utilisons la notation $a(x) \underset{x\to l}{\asymp} b(x)$ — avec $l \in \overline{\mathbb{R}}$ — pour signifier qu'il existe $0 < c, d < \infty$ et un voisinage V de l tels que $c \cdot b(x) \leq a(x) \leq d \cdot b(x)$ pour tout $x \in V$.

Théorème 7. *On suppose qu'il existe $\lambda > 0$ tel que $R^{\text{minimax}}_{1,\rho_{1+\lambda}} < \infty$.*

(i) Si Θ est fini, alors

$$R^{\text{minimax}}_n \xrightarrow[n\to\infty]{} \log|\Theta|.$$

(ii) Si $\dim(\Theta, h) = 0$, alors

$$R^{\text{minimax}}_n \in o(\log n).$$

(iii) Si $\dim(\Theta, h) = D$, avec $0 < D < \infty$, alors

$$R^{\text{minimax}}_n \sim \frac{D}{2}\log n.$$

(iv) Si $\dim(\Theta, h) = \infty$ mais (Θ, h) est totalement borné, alors

$$R^{\text{minimax}}_n \in O(n) \quad \textit{mais} \quad R^{\text{minimax}}_n \notin O(\log n).$$

(v) Si (Θ, h) n'est pas totalement borné, alors

$$\textit{si } R^{\text{minimax}}_1 < \infty, \textit{ alors } R^{\text{minimax}}_n \asymp n, \textit{ sinon } R^{\text{minimax}}_n = \infty \textit{ pour tout } n.$$

La condition $R^{\text{minimax}}_{1,\rho_{1+\lambda}} < \infty$ ne peut pas être enlevée. L'exemple suivant, tiré de [Gar06], en donne confirmation :

Exemple 2. Soit f une fonction positive strictement décroissante sur $\mathbb{N}^*$, et telle que $f(1) < 1$. Pour $k \in \mathbb{N}^*$, Soit p_k la probabilité sur $\mathbb{N}$ définie par

$$p_k(l) = \begin{cases} 1 - f(k) & \text{si } l = 0 \ ; \\ f(k) & \text{si } l = k \ ; \\ 0 & \text{sinon.} \end{cases}$$

Soit $\Lambda^1 = \{p_1, p_2, \ldots\}$, et soit Λ la famille des lois stationnaires sans mémoire dont les premières marginales sont dans Λ^1.

Le premier résultat est la proposition 13 de [Gar06] [2] :

$$\limsup_{k\to\infty} f(k)\ln k = \infty \Leftrightarrow \forall n \in \mathbb{N},\ R^{\text{minimax}}_n(\Lambda) = \infty.$$

[2] Nous avons légèrement corrigé l'énoncé en remplaçant la limite simple par la limite supérieure. La démonstration reste inchangée.

Nous voulons maintenant comparer le comportement de cette famille avec l'énoncé du théorème 7. Nous affirmons tout d'abord que Λ^1 est totalement bornée pour la distance de Hellinger si et seulement si $\lim_{k\to\infty} f(k) = 0$. Cela semble contredire le théorème 7 si on choisit f qui tend vers 0 moins vite que $1/(\ln k)$, par exemple $f(k) = 1/\sqrt{\ln k}$. En réalité, c'est la condition $R^{\text{minimax}}_{1,\rho_{1+\lambda}} < \infty$ qui n'est pas vérifiée. En effet, nous montrons que $R^{\text{minimax}}_{1,\rho_{1+\lambda}} = \infty$ dès que $\limsup_{k\to\infty} k^{\frac{\lambda}{1+\lambda}} f(k) = \infty$, ce qui est en particulier vérifié pour tout $\lambda > 0$ dès que $\limsup_{k\to\infty} f(k) \ln k = \infty$.

Démonstration. Commençons par calculer $D_{HL}(p_k, p_l)$ pour $k < l$.

$$\begin{aligned} D_{HL}(p_k,p_l)^2 &= \sum_{m\geq 0} \left(\sqrt{p_k(m)} - \sqrt{p_l(m)}\right)^2 \\ &= \left(\sqrt{1-f(k)} - \sqrt{1-f(l)}\right)^2 + f(k) + f(l) \\ &= 2\left(1 - \sqrt{(1-f(k))\cdot(1-f(l))}\right) \end{aligned}$$

Comme $f(l) < f(k)$, on obtient

$$\sqrt{2f(l)} < D_{HL}(p_k, p_l) < \sqrt{2f(k)}.$$

Soit maintenant $\epsilon > 0$. Si $f(k) \to 0$, alors il existe l_ϵ tel que $f(l_\epsilon) < \frac{\epsilon^2}{2}$. En conséquence la boule $B_{D_{HL}}(P_{l_\epsilon}, \epsilon)$ contient tous les P_k pour $k \geq l_\epsilon$, et $\mathcal{N}_\epsilon(\Lambda^1, D_{HL}) \leq l_\epsilon$. Comme c'est vrai pour tout $\epsilon > 0$, Λ^1 est totalement bornée.

Réciproquement, si Λ^1 est totalement bornée, alors pour tout $\epsilon > 0$, $\mathcal{D}_\epsilon$ est fini et il existe $k \neq l$ tels que $D_{HL}(p_k, p_l) < \epsilon$. En conséquence $\lim f(k) \leq \frac{1}{2} D_{HL}(p_k, p_l)^2 < \frac{\epsilon^2}{2}$, et $f(k) \to 0$.

Pour montrer que $R^{\text{minimax}}_{1,\rho_{1+\lambda}} = \infty$, il suffit de prouver que $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}$ n'est pas borné lorsque le prior μ varie. En effet $R^{\text{minimax}}_{1,\rho_{1+\lambda}} \geq \sup_\mu R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}$. Par la proposition 1, nous savons en outre que cela revient à montrer que $C_{\lambda,\mu}$ n'est pas borné en μ. Soit $m \geq 2$ et μ la distribution uniforme sur $\{1, 2, \ldots, m\}$.

$$\begin{aligned} C_{\lambda,\mu} &= \sum_{l\geq 0} \left(\frac{1}{m}\sum_{k=1}^m p_k(l)^{1+\lambda}\right)^{1/(1+\lambda)} \\ &= \left(\frac{1}{m}\sum_{k=1}^m (1-f(k))^{1+\lambda}\right)^{1/(1+\lambda)} + \sum_{l=1}^m \left(\frac{1}{m} f(l)^{1+\lambda}\right)^{1/(1+\lambda)} \\ &\geq \frac{1}{m^{1/(1+\lambda)}} \sum_{l=1}^m f(l) \\ &\geq m^{\frac{\lambda}{1+\lambda}} f(m). \end{aligned}$$

Si $\limsup_{k\to\infty} k^{\frac{\lambda}{1+\lambda}} f(k) = \infty$, $C_{\lambda,\mu}$ n'est donc pas borné en μ. □

Nous donnons maintenant le corollaire annoncé :

Corollaire 3. *On suppose qu'il existe $\lambda > 0$ tel que $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$.*

(i) Si $\overline{\dim}(\Theta, h) \leq D$, avec $0 < D < \infty$, alors

$$R_n^{\text{minimax}} \leq (1 + o(1))\frac{D}{2}\log n.$$

(ii) Si $\underline{\dim}(\Theta, h) \geq D$, avec $0 < D < \infty$, alors

$$R_n^{\text{minimax}} \geq (1 + o(1))\frac{D}{2}\log n.$$

Démonstration. La démonstration est celle que donne [HO97]. Nous la reproduisons parce qu'elle permet de voir la démarche utilisée en se basant sur les lemmes 3 et 4. Nous traiterons uniquement la majoration, la minoration étant tout à fait semblable.

Puisque $\overline{\dim}(\Theta, h) \leq D$, nous pouvons choisir

$$f_u(x) = D \ln x$$

En résolvant l'équation $D\ln(1/\epsilon) = n\epsilon^2$ on trouve, comme nous le vérifions plus bas,

$$\epsilon_u(n) \sim \sqrt{\frac{D}{2n}\ln n}, \tag{3.3}$$

et en conséquence l'équation (3.2) entraîne

$$F_u(n) \sim \frac{D}{2}\ln n.$$

On choisit, pour appliquer le lemme 4, $g(n) = \ln n$, d'où il vient

$$(\log e)\, F_u(n \ln^2 n) \sim \frac{D}{2}\log n \ ;$$
$$\limsup_{n\to\infty} \frac{R_n^{\text{minimax}}}{(D/2)\log n} \leq 1.$$

Il nous reste maintenant à justifier la relation (3.3). En posant $x_n = \epsilon_u(n)^{-2}$, l'équation à résoudre devient $x_n \ln x_n = 2n/D$, soit encore $f(x_n) = 2n/D$ en posant $f(x) = x\ln x$. Comme $2n/D > 0$, $x_n > 1$. f est strictement croissante sur $[1, \infty)$, donc l'équation admet une unique solution.

Si n est assez grand, $2n/D > f(e) = e$ et $x_n > e$, d'où $x_n < f(x_n)$. D'autre part

$$\begin{aligned} f\left(\frac{2n}{D\ln n}\right) &= \frac{2n}{D\ln n}(\ln n + \ln 2 - \ln D - \ln\ln n) \\ &= \frac{2n}{D} - \frac{2n}{D}\frac{\ln\ln n + \ln D - \ln 2}{\ln n} \\ &< f(x_n) \text{ si } n \text{ est assez grand,} \end{aligned}$$

d'où $x_n > \frac{2n}{D\ln n}$.

Par ailleurs

$$\frac{f(x)}{f(y)} = \frac{x}{y}\frac{\ln(x/y)+\ln y}{\ln y} = \frac{x}{y} + \frac{1}{\ln y} f\left(\frac{x}{y}\right),$$

d'où

$$\frac{x}{y} = \frac{f(x)}{f(y)} - \frac{1}{\ln y} f\left(\frac{x}{y}\right).$$

En prenant $x = \frac{2n}{D\ln n}$ et $y = x_n$, on obtient

$$\frac{1}{\ln n} < \frac{x_n D \ln n}{2n} < 1,$$

d'où

$$-e^{-1} \le f\left(\frac{x_n D\ln n}{2n}\right) \le 0$$

et

$$-\frac{1}{\ln x_n} f\left(\frac{x_n D\ln n}{2n}\right) \xrightarrow[n\to\infty]{} 0.$$

Comme

$$f\left(\frac{2n}{D\ln n}\right) f(x_n)^{-1} = 1 - \frac{\ln\ln n + \ln D - \ln 2}{\ln n} \xrightarrow[n\to\infty]{} 1,$$

on a

$$\frac{x_n D\ln n}{2n} \xrightarrow[n\to\infty]{} 1 \quad \text{et} \quad x_n \sim \frac{2n}{D\ln n}.$$

□

Si le théorème 7 est assez précis lorsque la dimension métrique de Θ pour la distance de Hellinger est finie, en particulier lorsque Θ est une partie de $\mathbb{R}^n$, ce n'est plus le cas lorsque la dimension de Θ est infinie. Cependant les lemmes 3 et 4 permettent de caractériser le taux d'accroissement asymptotique de la redondance minimax en fonction de l'entropie métrique de Θ dans de nombreux cas. En pratique l'entropie métrique croît habituellement en $(1/\epsilon)^\alpha \ln(1/\epsilon)^\beta$, et le théorème 8 permet de traiter ce cas.

Théorème 8. *On suppose qu'il existe $\lambda > 0$ tel que $R^{\text{minimax}}_{1,\rho_{1+\lambda}} < \infty$. Soit $l(x)$ une fonction continue et croissante, définie sur $(0,\infty)$, et telle que pour tous $\gamma \ge 0$ et $C > 0$,*

(i)

$$\lim_{x\to\infty} \frac{l\left(Cx(l(x))^\gamma\right)}{l(x)} = 1$$

(ii)

$$\lim_{x\to\infty} \frac{l\left(Cx(\ln x)^\gamma\right)}{l(x)} = 1.$$

Alors

1.

$$\text{Si } \mathcal{H}_\epsilon(\Theta,h) \sim l\left(\frac{1}{\epsilon}\right), \text{ alors } R^{\text{minimax}}_n \sim (\log e)\, l(\sqrt{n}).$$

2. *Si, pour un certain* $\alpha \geq 0$,

$$\mathcal{H}_\epsilon(\Theta, h) \asymp \left(\frac{1}{\epsilon}\right)^\alpha l\left(\frac{1}{\epsilon}\right),$$

alors

(*i*) *Si* $\lim_{\epsilon\to 0} b(\epsilon) < \infty$, *alors*

$$R_n^{\text{minimax}} \asymp n^{\alpha/(\alpha+2)} \left[l(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)}$$

(*ii*) *sinon*

$$\liminf_{n\to\infty} \frac{R_n^{\text{minimax}}}{n^{\alpha/(\alpha+2)} \left[l(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)}} > 0$$

et

$$\limsup_{n\to\infty} \frac{R_n^{\text{minimax}}}{n^{\alpha/(\alpha+2)} \left[l(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)} (\log n)^{\alpha/(\alpha+2)}} < \infty.$$

Les conditions sur la fonction l portent sur sa vitesse de croissance, qui ne doit pas être trop grande. Une fonction logarithme convient, et permet de traiter les entropies de la forme $\mathcal{H}_\epsilon(\Theta, h) \sim C \cdot (1/\epsilon)^\alpha \ln(1/\epsilon)^\beta$. Cependant il faut faire attention à la condition $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$, qui ne peut pas plus être supprimée ici que dans le théorème 7. En revanche les bornes inférieures fournies par le théorème restent valables sans cette condition mais rien ne garantit alors qu'elles sont fines.

Ici aussi il nous faut énoncer une version qui sépare les minorations et les majorations, pour traiter le cas de classes Λ dont on ne connaît pas l'entropie métrique à un équivalent près.

Corollaire 4. *Soit* $l(x)$ *une fonction comme dans le théorème 8.*

1. *Soit* $g(\epsilon)$ *une fonction et* $\gamma > 0$ *un réel tels que*

$$g(\epsilon) \underset{\epsilon\to 0}{\sim} \gamma\, l\left(\frac{1}{\epsilon}\right).$$

(*i*) *Si* $\mathcal{H}_\epsilon(\Theta, h) \geq g(\epsilon)$ *pour tout* $\epsilon > 0$, *alors*

$$\liminf_{n\to\infty} \frac{R_n^{\text{minimax}}}{\gamma\,(\log e)\, l(\sqrt{n})} \geq 1.$$

(*ii*) *Si* $\mathcal{H}_\epsilon(\Theta, h) \leq g(\epsilon)$ *pour tout* $\epsilon > 0$, *et s'il existe en outre* $\lambda > 0$ *tel que* $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$, *ou bien si* $\lim_{\epsilon\to 0} b(\epsilon) < \infty$, *alors*

$$\limsup_{n\to\infty} \frac{R_n^{\text{minimax}}}{\gamma\,(\log e)\, l(\sqrt{n})} \leq 1.$$

2. (*i*) *Si pour* $\alpha > 0$ *il existe* $c > 0$ *tel que*

$$\liminf_{\epsilon \to 0} \mathcal{H}_\epsilon(\Theta, h) \left(\frac{1}{\epsilon}\right)^{-\alpha} l\left(\frac{1}{\epsilon}\right)^{-1} \geq c,$$

alors

$$\liminf_{n \to \infty} \frac{R_n^{\text{minimax}}}{n^{\alpha/(\alpha+2)} \left[l(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)}} > 0.$$

(*ii*) *Si pour* $\alpha > 0$ *il existe* $C < \infty$ *tel que*

$$\limsup_{\epsilon \to 0} \mathcal{H}_\epsilon(\Theta, h) \left(\frac{1}{\epsilon}\right)^{-\alpha} l\left(\frac{1}{\epsilon}\right)^{-1} \leq C,$$

alors

(*a*) *Si* $\lim_{\epsilon \to 0} b(\epsilon) < \infty$, *alors*

$$\limsup_{n \to \infty} \frac{R_n^{\text{minimax}}}{n^{\alpha/(\alpha+2)} \left[l(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)}} < \infty,$$

(*b*) *S'il existe* $\lambda > 0$ *tel que* $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$, *alors*

$$\limsup_{n \to \infty} \frac{R_n^{\text{minimax}}}{n^{\alpha/(\alpha+2)} \left[l(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)} (\ln n)^{\alpha/(\alpha+2)}} < \infty.$$

Démonstration. Ici encore il ne s'agit que d'une adaptation de la preuve fournie par [HO97]. Seule la partie 1 apporte des aménagements, aussi nous ne nous occupons que d'elle. Comme plus haut, nous ne traiterons que la majoration, la minoration étant semblable (remarquons seulement que la condition $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$ n'est plus nécessaire).

Nous supposons donc que $\mathcal{H}_\epsilon(\Theta, h) \leq g(\epsilon)$, avec $g(\epsilon) \sim \gamma\, l(1/\epsilon)$. Nous pouvons donc poser

$$f_u(x) = \gamma\, l(x),$$

ce qui entraîne, comme nous le montrons plus bas,

$$\epsilon_u(n) \sim \sqrt{\frac{\gamma\, l(\sqrt{n})}{n}}, \tag{3.4}$$

d'où

$$F_u(n) \sim \gamma\, l(\sqrt{n}).$$

En choisissant $g(n) = \ln n$ dans le lemme 4, il vient

$$F_u(n \ln^2 n) \sim \gamma\, l(\sqrt{n} \ln n) \sim \gamma\, l(\sqrt{n}),$$

ce qui permet de conclure.

Vérifions la relation (3.4). En posant $x_n = \epsilon_u(n)^{-1}$, l'équation à résoudre devient

$$x_n^2 l(x_n) = n/\gamma.$$

Comme n/γ est strictement positif et tend vers l'infini, nous observons que $l(x_n) > 0$ et $x_n \to \infty$. En partant de l'équation $x_n\sqrt{l(x_n)} = \sqrt{n/\gamma}$, nous pouvons écrire

$$\frac{x_n\sqrt{l(x_n)}}{\sqrt{l(x_n\sqrt{l(x_n)})}} = \frac{\sqrt{n/\gamma}}{\sqrt{l(\sqrt{n/\gamma})}},$$

d'où

$$x_n\sqrt{\frac{l(\sqrt{n/\gamma})}{n/\gamma}} = \sqrt{\frac{l(x_n\sqrt{l(x_n)})}{l(x_n)}}.$$

Comme $x_n \to \infty$, le terme de droite tend vers 1 par propriété de l, et donc

$$\epsilon_u(n) \sim \sqrt{\frac{l(\sqrt{n/\gamma})}{n/\gamma}} \sim \sqrt{\frac{\gamma\, l(\sqrt{n})}{n}}.$$

□

Maintenant que nous avons énoncé les outils dont nous avons besoin, nous allons pouvoir dans la section qui vient les appliquer au sujet qui nous intéresse : les classes enveloppe.

4 Classes enveloppe

Nous allons dans un premier temps définir ce que sont les classes enveloppe et citer divers résultats déjà acquis les concernant. Ensuite nous verrons quelles applications aux classes enveloppe on peut tirer des théorèmes énoncés dans les sections précédentes, et nous obtiendrons en particulier une majoration originale de la redondance des classes à décroissance exponentielle. Enfin, nous ferons appel à d'autres outils pour obtenir la minoration correspondante, complétant ainsi notre résultat en un équivalent de la redondance minimax des classes à décroissance exponentielle.

4.1 État des lieux

C'est sur les classes enveloppe que se concentre notre intérêt pour les résultats de [HO97] appliqués au codage de données. Ici l'alphabet est infini dénombrable et identifié à $\mathbb{N}^*$.

Définition 9. *Soit f une fonction de $\mathbb{N}^*$ dans $[0,1]$. La classe de sources stationnaires sans mémoire d'enveloppe f est définie par*

$$\Lambda_f = \{\boldsymbol{P} = P^{\otimes\infty} : \ \forall k \in \mathbb{N}^*, \ P(k) \leq f(k)\}.$$

Bien entendu il faut que $\sum_{k\geq 1} f(k) \geq 1$ pour que Λ_f soit non vide. On peut aussi accepter des valeurs de f supérieures à 1 mais elles ne changent rien à Λ_f. Le résultat suivant donne un premier résultat important sur le comportement des classes enveloppe en fonction de leur enveloppe. Il regroupe la proposition 5 et le théorème 3 de [BGG06].

Théorème 9. *Soit f une fonction de $\mathbb{N}^*$ dans $[0,1]$, et Λ_f la classe enveloppe associée. Alors, pour tout $n \in \mathbb{N}$,*

$$R_n^{\text{minimax}}(\Lambda_f) < \infty \Leftrightarrow R_n^*(\Lambda_f) < \infty \Leftrightarrow \sum_{k\geq 1} f(k) < \infty.$$

La première équivalence est propre aux classes enveloppe alors que la seconde est valable pour toutes les classes de sources stationnaires sans mémoire sur un alphabet dénombrable. On remarque que le comportement est le même pour toutes les valeurs de n (mais ce n'est pas une nouveauté).

Les classes enveloppes intéressantes seront donc celles pour lesquelles f décroît assez vite. Heuristiquement, on peut dire que la redondance de la classe enveloppe dépend essentiellement de la vitesse de décroissance de l'enveloppe. [BGG06] fournit un résultat général pour encadrer la redondance des classes enveloppe [1] et l'applique à deux types de

[1] C'est leur théorème 5.

classes enveloppe qui nous intéresseront plus particulièrement : les classes à décroissance polynomiale et les classes à décroissance exponentielle.

Définition 10. *Nous utilisons la notation classique* $\zeta(\alpha) = \sum_{k\geq 1} k^{-\alpha}$, *pour* $\alpha > 1$.

(*i*) *Soit* $\alpha > 1$ *et* C *tel que* $C > 1 \wedge \frac{2^\alpha}{\zeta(\alpha)}$. *La classe à décroissance polynomiale* $\Lambda_{C\cdot^{-\alpha}}$ *est la classe associée à la fonction enveloppe* $f_{\alpha,C} : \ x \mapsto 1 \wedge Cx^{-\alpha}$.

(*ii*) *Soit* $\alpha > 1$ *et* C *tel que* $C > e^{2\alpha}$. *La classe à décroissance exponentielle* $\Lambda_{Ce^{-\alpha\cdot}}$ *est la classe associée à la fonction enveloppe* $g_{\alpha,C} : \ x \mapsto 1 \wedge Ce^{-\alpha x}$.

Théorème 10. *Dans les deux cas ci-dessous on suppose que* α *et* C *vérifient les conditions données dans la définition 10.*

(*i*)

$$n^{1/\alpha} A(\alpha) \log\lfloor C\zeta(\alpha)\rfloor \leq R_n^{\text{minimax}}(\Lambda_{C\cdot^{-\alpha}}) \leq R_n^*(\Lambda_{C\cdot^{-\alpha}}) \leq \left(\frac{2Cn}{\alpha-1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1),$$

où

$$A(\alpha) = \frac{1}{\alpha}\int_1^\infty \frac{1 - e^{-1/(\zeta(\alpha)u)}}{u^{1-1/\alpha}}\mathrm{d}u.$$

(*ii*)

$$\frac{1}{8\alpha}\log^2 n(1-o(1)) \leq R_n^{\text{minimax}}(\Lambda_{Ce^{-\alpha\cdot}}) \leq R_n^*(\Lambda_{Ce^{-\alpha\cdot}}) \leq \frac{1}{2\alpha}\log^2 n + O(1).$$

Ces deux résultats nous serviront de base de comparaison pour les applications aux classes enveloppe des théorèmes d'encadrement de la redondance minimax fournis dans la section 3.

4.2 Application des résultats précédents aux classes enveloppes

Comme l'indique le théorème 9, les classes enveloppe qui présentent un intérêt sont celles pour lesquelles l'enveloppe est intégrable. En conséquence la condition $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$ est réalisée pour tout $\lambda > 0$, et nous pourrons bien appliquer les différents théorèmes qui l'exigent.

Étudions de plus près le comportement des classes enveloppe. Pour cela, fixons f une enveloppe sommable. Une quantité importante qui permet de caractériser le comportement de Λ_f est la queue de la série, que nous notons $\bar{f}$:

$$\bar{f}(n) = \sum_{k\geq n+1} f(k). \tag{4.1}$$

Considérons deux lois P_1 et P_2 correspondant aux premières marginales de deux éléments de Λ_f. Alors,

$$\begin{aligned} D_{HL}(P_1, P_2) &= \sqrt{\sum_{k\geq 1} \left(\sqrt{P_1(k)} - \sqrt{P_2(k)}\right)^2} \\ &\leq \sqrt{\sum_{k\geq 1} \left(2\sqrt{f(k)}\right)^2} \\ &= 2\bar{f}(1). \end{aligned}$$

Nous pouvons faire deux remarques concernant l'expression de $D_{HL}(P_1, P_2)$:

– L'espace (Λ_f^1, D_{HL}) est en isométrie avec la partie $A_f \cap \{\|x\| = 1\}$ de ℓ^2, où A_f est défini par

$$A_f = \{(x_k)_{k\in\mathbb{N}^*} \in \ell^2 : \ \forall k \in \mathbb{N}^*, \ 0 \leq x_k \leq \sqrt{f(k)}\}, \tag{4.2}$$

grâce à la transformation $P \mapsto (\sqrt{P(k)})_{k\in\mathbb{N}^*}$. L'entropie métrique $\mathcal{H}_\epsilon$ sera calculée sur A_f.

– Il suffit qu'il existe deux entiers k_1 et k_2 dans le support de f tels que $\sum_{k\neq k_1,k_2} f(k) \geq 1$ pour que l'on puisse construire deux lois P_1 et P_2 telles que $D_{HL}(P_1, P_2)$ soit aussi petit que l'on veut mais $D_{KL}(P_1 \,\|\, P_2) = \infty$. En conséquence $b(\epsilon) = \infty$ quelque soit $\epsilon > 0$, et on ne peut utiliser les parties du lemme 3 et du théorème 8 qui supposent que $b(\epsilon)$ est fini. Cette situation est en fait vérifiée pour la plupart des enveloppes f, et en particulier pour les enveloppes $f_{\alpha,C}$ et $g_{\alpha,C}$ des classes à décroissance polynomiales ou exponentielles dès que C est assez grand.

Les idées que nous utilisons pour calculer l'entropie métrique des classes enveloppes sont principalement deux. La première est que les coefficients correspondant à k grand varient peu et contribuent peu à la distance de Hellinger, et que nous pourrons donc tronquer la suite des coefficients de manière à se ramener à un rectangle en dimension finie. La seconde est que le nombre de points ϵ-écartés dans une partie de $\mathbb{R}^n$ est plus petit que le nombre de boules de rayon $\epsilon/2$ que l'on peut mettre dans le même ensemble élargi de $\epsilon/2$. Voyons comment exploiter ces idées pour majorer l'entropie métrique.

Soit $\epsilon > 0$. Commençons par définir le point où nous voulons tronquer les coefficients. Pour cela posons

$$N_\epsilon = \inf\left\{n \geq 1 : \ \bar{f}(n) \leq \frac{\epsilon^2}{16}\right\}. \tag{4.3}$$

Vérifions que notre troncature « fonctionne ». Soit $x = (x_n)_{n\geq 1}$ un élément de A_f et

$\tilde{x} = (x_n \mathbb{1}_{n \leq N\epsilon})_{n \geq 1}$ sa version tronquée. Alors

$$\begin{aligned}
\|x - \tilde{x}\| &= \sqrt{\sum_{n \geq N_\epsilon + 1} x_n^2} \\
&\leq \sqrt{\sum_{n \geq N_\epsilon + 1} \sqrt{f(n)}^2} \\
&= \sqrt{\bar{f}(N_\epsilon)} \\
&\leq \frac{\epsilon}{4}.
\end{aligned}$$

Soit maintenant S une $\epsilon/4$-couverture de $\{x \in A_f : \forall n \geq N_\epsilon, x_n = 0\}$. Soit $x \in A_f$. Alors il existe $y \in S$ tel que $\|\tilde{x} - y\| \leq \epsilon/4$, d'où $\|x - y\| \leq \epsilon/2$, et S est une $\epsilon/2$-couverture de A_f.

Rassemblons toutes ces points :

$$\begin{aligned}
\mathcal{D}_\epsilon(\Lambda_f^1, D_{HL}) &= \mathcal{D}_\epsilon(A_f \cap \{\|x\| = 1\}, \|\cdot\|_{\ell^2}) \\
&\leq \mathcal{D}_\epsilon(A_f, \|\cdot\|_{\ell^2}) \\
&\leq \mathcal{N}_{\epsilon/2}(A_f, \|\cdot\|_{\ell^2}) \\
&\leq \mathcal{N}_{\epsilon/4}\left(\prod_{1 \leq k \leq N_\epsilon} \left[0, \sqrt{f(k)}\right], \|\cdot\|_{\mathbb{R}^{N_\epsilon}}\right) \\
&\leq \mathcal{M}_{\epsilon/4}\left(\prod_{1 \leq k \leq N_\epsilon} \left[0, \sqrt{f(k)}\right], \|\cdot\|_{\mathbb{R}^{N_\epsilon}}\right) \\
&\leq \frac{\mathrm{Vol}\left(\prod_{1 \leq k \leq N_\epsilon} \left[-\frac{\epsilon}{8}, \sqrt{f(k)} + \frac{\epsilon}{8}\right]\right)}{\mathrm{Vol}\left(B_{\mathbb{R}^{N_\epsilon}}(0, \frac{\epsilon}{8})\right)} \\
&\leq \left(\frac{\epsilon}{8}\right)^{-N_\epsilon} \frac{1}{\mathrm{Vol}\left(B_{\mathbb{R}^{N_\epsilon}}(0,1)\right)} \prod_{k=1}^{N_\epsilon} \left(\sqrt{f(k)} + \frac{\epsilon}{4}\right).
\end{aligned}$$

Remarquons au passage que la classe Λ_f^1 est totalement bornée.

Nous pouvons alors écrire le résultat suivant :

Proposition 4.

$$\mathcal{H}_\epsilon \leq N_\epsilon \ln(1/\epsilon) + 3N_\epsilon \ln 2 + A(\epsilon) + B(\epsilon),$$

où

$$A(\epsilon) = -\ln \mathrm{Vol}\left(B_{\mathbb{R}^{N_\epsilon}}(0,1)\right) = \ln \frac{\Gamma(\frac{N_\epsilon}{2} + 1)}{\pi^{\frac{N_\epsilon}{2}}}$$

et

$$B(\epsilon) = \sum_{k=1}^{N_\epsilon} \ln\left(\sqrt{f(k)} + \frac{\epsilon}{4}\right).$$

En outre

$$A(\epsilon) \sim \frac{N_\epsilon}{2} \ln N_\epsilon,$$

et

$$-N_\epsilon \ln(1/\epsilon) - 2N_\epsilon \ln 2 \leq B(\epsilon) \leq \frac{\epsilon}{4} N_\epsilon.$$

L'encadrement concernant $B(\epsilon)$ nous montre que $B(\epsilon)$ est un terme qui diminue plutôt l'entropie, à la différence de $A(\epsilon)$ qui y contribue ; on peut remarquer que si $\ln N_\epsilon \asymp \ln(1/\epsilon)$, alors l'ordre de grandeur du majorant est $N_\epsilon \ln(1/\epsilon)$, ce qui correspond au point 2 du théorème 8.

Démonstration. Il reste à justifier nos affirmations sur $A(\epsilon)$ et $B(\epsilon)$. Concernant le premier, nous utilisons la formule de Stirling (dans sa version modifiée par [WW96, chapitre XII]) :

$$\Gamma(x) = \sqrt{2\pi}\, x^{x-1/2} e^{-x} e^{\frac{\beta}{12x}}, \quad \text{avec } \beta \in [0,1]. \tag{4.4}$$

Nous pouvons alors écrire

$$\begin{aligned} A(\epsilon) &= -\frac{N_\epsilon}{2} \ln \pi + \ln \Gamma\left(\frac{N_\epsilon}{2} + 1\right) \\ &= -\frac{N_\epsilon}{2} \ln \pi + \left(\frac{N_\epsilon + 1}{2}\right) \ln\left(\frac{N_\epsilon}{2} + 1\right) - \left(\frac{N_\epsilon}{2} + 1\right) + \frac{\ln(2\pi)}{2} + \frac{\beta}{12\left(\frac{N_\epsilon}{2} + 1\right)}. \end{aligned}$$

Comme $N_\epsilon \to \infty$, l'équivalence annoncée est vraie.

L'encadrement de $B(\epsilon)$, quant à lui, ne pose pas de difficultés. □

Voyons maintenant ce que ce résultat donne lorsqu'il est appliqué aux classes à décroissance polynomiale ou exponentielle.

Corollaire 5. *On choisit $\Lambda_f = \Lambda_{C \cdot^{-\alpha}}$ une classe enveloppe à décroissance polynomiale. Alors il existe une constante $K > 0$ telle que*

$$R_n^{\text{minimax}} \leq K n^{1/\alpha} \ln n.$$

Démonstration. Commençons par calculer N_ϵ :

$$N_\epsilon \leq \inf \left\{ n \geq 1 : \sum_{k \geq n+1} C k^{-\alpha} \leq \frac{\epsilon^2}{16} \right\} ;$$

or

$$\sum_{k \geq n+1} k^{-\alpha} \leq \int_n^\infty x^{-\alpha} \mathrm{d}x = \frac{1}{(\alpha - 1) n^{\alpha - 1}},$$

et en conséquence

$$N_\epsilon \leq \inf\left\{n \geq 1 :\ n^{\alpha-1} \geq \frac{16C}{(\alpha-1)\epsilon^2}\right\} \leq \left\lceil \left(\frac{16C}{(\alpha-1)\epsilon^2}\right)^{\frac{1}{\alpha-1}} \right\rceil.$$

Nous pouvons donc écrire que $N_\epsilon \leq g(\epsilon)$, avec

$$g(\epsilon) \sim \left(\frac{16C}{(\alpha-1)\epsilon^2}\right)^{\frac{1}{\alpha-1}} = \left(\frac{16C}{(\alpha-1)}\right)^{\frac{1}{\alpha-1}} (1/\epsilon)^{\frac{2}{\alpha-1}}.$$

En utilisant la remarque après la proposition 4 il vient alors

$$\mathcal{H}_\epsilon \leq K(1/\epsilon)^{\frac{2}{\alpha-1}} \ln(1/\epsilon)$$

pour un $K > 0$ et pour ϵ au voisinage de 0.

Nous pouvons alors utiliser le corollaire 4 avec $l = \ln$ pour obtenir

$$\limsup_{n\to\infty} \frac{R_n^{\text{minimax}}}{n^{1/\alpha}\ln n} < \infty,$$

puisque

$$\frac{2}{\alpha-1}\left(\frac{2}{\alpha-1}+2\right)^{-1} = \frac{1}{\alpha}.$$

□

Ce résultat est moins bon que celui du théorème 10 issu de [BGG06]. On peut penser que la majoration de $\mathcal{H}_\epsilon$ peut être améliorée, cependant nous remarquons que la puissance $\frac{2}{\alpha-1}$ qui porte sur $(1/\epsilon)$ est correcte, puisque c'est celle qui est demandée pour produire le terme principal $n^{1/\alpha}$. Malheureusement elle entraîne nécessairement un facteur $(\ln n)^{1/\alpha}$ entre la majoration et la minoration du théorème 8, ce qui est moins bon, dès que $\alpha < 2$, que l'écart constaté avec le théorème 10. Cependant nous ne savons pas à ce stade si le théorème 8 permet d'améliorer la minoration fournie par le théorème 10 : si notre majoration de l'entropie n'est pas trop brutale nous pouvons espérer un ordre de grandeur en $n^{1/\alpha}(\ln n)^{1-1/\alpha}$ pour la minoration de R_n^{minimax}, ce qui serait très bon !

Corollaire 6. *On choisit $\Lambda_f = \Lambda_{Ce^{-\alpha\cdot}}$ une classe enveloppe à décroissance exponentielle. Alors*

$$R_n^{\text{minimax}} \leq (1+o(1))\frac{1}{4\alpha\log e}\log^2 n.$$

Démonstration. Cette fois-ci nous avons

$$\begin{aligned} N_\epsilon &\leq \inf\left\{ n \geq 1 : \sum_{k\geq n+1} Ce^{-\alpha k} \leq \frac{\epsilon^2}{16} \right\} \\ &\leq \inf\left\{ n \geq 1 : \frac{Ce^{-\alpha\,(n+1)}}{1-e^{-\alpha}} \leq \frac{\epsilon^2}{16} \right\} \\ &\leq \inf\left\{ n \geq 1 : -\alpha\,(n+1) \leq -\ln\frac{16C}{(1-e^{-\alpha})\epsilon^2} \right\} \\ &\leq \left\lceil \frac{2}{\alpha}\ln(1/\epsilon) + \frac{1}{\alpha}\ln\frac{16C}{1-e^{-\alpha}} - 1 \right\rceil, \end{aligned}$$

et nous pouvons écrire

$$N_\epsilon \leq \frac{2}{\alpha}\ln(1/\epsilon) + \frac{1}{\alpha}\ln\frac{16C}{1-e^{-\alpha}} \sim \frac{2}{\alpha}\ln(1/\epsilon).$$

Cela nous permet de majorer $B(\epsilon)$:

$$\begin{aligned} B(\epsilon) &\leq \sum_{k=1}^{N_\epsilon} \ln\left(\frac{\epsilon}{4} + \sqrt{C}e^{-\frac{\alpha}{2}k}\right) \\ &\leq \int_0^{N_\epsilon} \ln\left(\frac{\epsilon}{4} + \sqrt{C}e^{-\frac{\alpha}{2}x}\right) \mathrm{d}x \\ &= \int_0^{N_\epsilon} \frac{\ln C}{2} - \frac{\alpha}{2}x + \ln\left(1 + \frac{\epsilon\cdot e^{\frac{\alpha}{2}x}}{4\sqrt{C}}\right) \mathrm{d}x. \end{aligned}$$

D'autre part

$$\begin{aligned} N_\epsilon &\leq \frac{2}{\alpha}\ln(1/\epsilon) + \frac{1}{\alpha}\ln\frac{16C}{1-e^{-\alpha}} \\ e^{\frac{\alpha}{2}N_\epsilon} &\leq \frac{1}{\epsilon}\cdot\sqrt{\frac{16C}{1-e^{-\alpha}}} \\ \frac{\epsilon\cdot e^{\frac{\alpha}{2}N_\epsilon}}{4\sqrt{C}} &\leq \frac{1}{\sqrt{1-e^{-\alpha}}}, \end{aligned}$$

et donc

$$B(\epsilon) \leq -\frac{\alpha}{4}N_\epsilon^2 + \left(\frac{\ln C}{2} + \frac{1}{\sqrt{1-e^{-\alpha}}}\right)N_\epsilon \sim -\frac{\alpha}{4}N_\epsilon^2 \sim -\frac{1}{\alpha}\ln^2(1/\epsilon).$$

Par ailleurs

$$A(\epsilon) \sim \frac{2}{\alpha}\ln(1/\epsilon)\cdot\ln\ln(1/\epsilon) = o(\ln^2(1/\epsilon)),$$

et, en rassemblant les différents termes de la proposition 4, on obtient

$$\mathcal{H}_\epsilon \leq g(\epsilon)$$

avec $g(\epsilon) \sim \frac{1}{\alpha}\ln^2(1/\epsilon) = \frac{1}{\alpha\log^2 e}\log^2(1/\epsilon)$.

Il ne reste plus qu'à appliquer le corollaire 4 avec $l = \log^2$ pour conclure. □

Cette fois-ci la comparaison avec le théorème 10 montre que cette majoration est très intéressante. Mieux, nous serons en mesure d'obtenir dans le paragraphe 4.3 une minoration semblable qui complétera notre résultat en une équivalence. Ce bon comportement des classes à décroissance exponentielle peut s'analyser ainsi : la décroissance de f est suffisamment rapide pour garantir une entropie qui croît plus lentement qu'un polynôme, ce qui permet d'utiliser la partie 1 du corollaire 4 qui donne une vraie équivalence.

Et la minoration ?

Concernant la minoration de l'entropie métrique nous pouvons utiliser une démarche analogue à celle suivie pour la majoration. Malheureusement nous ne pouvons pas nous débarrasser aussi facilement de la condition $\|x\| = 1$. Nous avons essayé de résoudre le problème dans le cas des classes à décroissance exponentielle, pour lesquelles nous avions de bons espoirs, mais nous avons été bloqués par le calcul de volumes un peu compliqués (l'intersection d'une couronne et d'un rectangle) en grande dimension.

Aussi nous avons eu recours à une toute autre méthode, dérivée du travail de [XB97], pour minorer la redondance des classes à décroissance exponentielle.

4.3 Minoration de la redondance de la classe à décroissance exponentielle

L'approche que nous développons dans ce paragraphe est basée sur la recherche d'un prior μ sur la classe Λ_f^1, où $\Lambda_f = \Lambda_{Ce^{-\alpha\cdot}}$ est une classe enveloppe à décroissance exponentielle, tel que $R_{n,\mu}^{\text{Bayes}}(\Lambda_f)$ soit arbitrairement proche de la majoration $\frac{1}{4\alpha \log e} \log^2 n$ obtenue dans le paragraphe précédent. Le prior que nous choisissons est dérivé du prior de Dirichlet, et nous aurons besoin d'une version modifiée des résultats de [XB97] qui utilise ce prior pour les sources stationnaires sans mémoire en alphabet fini. Voyons de plus près ce dont il s'agit.

Considérons un alphabet $A = \{a_1, a_2, \ldots, a_k\}$ fini de cardinal $k \in \mathbb{N}^*$. Alors un processus $\boldsymbol{P}$ stationnaire sans mémoire à valeurs dans A est caractérisé par les valeurs $P(a_i)$, $1 \leq i \leq k$, prises par sa première marginale sur les éléments de A, avec la contrainte $\sum_{1\leq i\leq k} P(a_i) = 1$. Nous pouvons alors paramétrer la classe des sources sans mémoire sur l'alphabet A sous la forme $(\boldsymbol{P}_\theta)$, avec $\theta \in S_k$ le simplex de $\mathbb{R}^k$

$$S_k = \{(\theta_1, \theta_2, \ldots, \theta_k) \in [0,1]^k : \sum_{1\leq i\leq k} \theta_i = 1\},$$

en posant $P_\theta(a_i) = \theta_i$.

Une autre notation du simplex est obtenue en posant $\theta_1 = 1 - \sum_{2\leq i\leq k} \theta_i$, avec $(\theta_2, \ldots, \theta_k)$ variant dans l'ensemble

$$S'_k = \{(\theta_2, \ldots, \theta_k) \in [0,1]^{k-1} : \sum_{2\leq i\leq k} \theta_i \leq 1\}.$$

Cela permet de définir facilement la mesure de Lebesgue $\mathrm{d}\theta$ sur le simplex de $\mathbb{R}^k$ par restriction de la mesure de Lebesgue sur $[0,1]^{k-1}$.

Si maintenant nous considérons $X_{1:n}$ la suite des n premiers symboles produits par la source $\boldsymbol{P}_\theta$, nous pouvons poser

$$T_i = \sum_{j=1}^{n} \mathbb{1}_{X_j=a_i} \quad \text{pour } 1 \le i \le k,$$

et la probabilité de la séquence $X_{1:n}$ sous $\boldsymbol{P}_\theta$ est alors

$$P_\theta^n(X_{1:n}) = \theta_1^{T_1}\theta_2^{T_2}\cdots\theta_k^{T_k}.$$

Le mélange de Dirichlet [2] consiste à choisir le prior μ de densité proportionnelle à $\theta_1^{-1/2}\theta_2^{-1/2}\cdots\theta_k^{-1/2}$ par rapport à la mesure de Lebesgue. La loi de Bayes sur A^n associée vérifie donc

$$P_\mu^n(X_{1:n}) = \frac{\int_{S'_k} \theta_1^{T_1-1/2}\theta_2^{T_2-1/2}\cdots\theta_k^{T_k-1/2}\mathrm{d}\theta}{\int_{S'_k} \theta_1^{-1/2}\theta_2^{-1/2}\cdots\theta_k^{-1/2}\mathrm{d}\theta}.$$

Cette expression peut être simplifiée grâce aux expressions des intégrales de Dirichlet utilisant la fonction Γ :

$$\begin{aligned} D_k(\lambda_1,\ldots,\lambda_k) &= \int_{S'_k} \theta_1^{\lambda_1-1}\theta_2^{\lambda_2-1}\cdots\theta_k^{\lambda_k-1}\mathrm{d}\theta \\ &= \frac{\Gamma(\lambda_1)\Gamma(\lambda_2)\cdots\Gamma(\lambda_k)}{\Gamma\left(\sum_{i=1}^k \lambda_i\right)}. \end{aligned}$$

Proposition 5. *Soit θ un élément du simplex S_k, et P_μ^n le mélange de Bayes sur A^n défini ci-dessus. Alors*

$$D(P_\theta^n \,\|\, P_\mu^n) \ge \frac{k-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma(1/2)^k}{\Gamma(k/2)} - \frac{5k}{3}\log e.$$

Démonstration. La démonstration, comme la proposition elle-même, est une adaptation de la proposition 1 de [XB97].

La formule est invariante lorsqu'on change la base du logarithme, aussi nous choisirons le logarithme népérien pour cette démonstration. Nous notons $\mathbb{E}_\theta$ l'espérance sous la loi $\boldsymbol{P}_\theta$. Alors

$$\begin{aligned} D(P_\theta^n \,\|\, P_\mu) &= \mathbb{E}_\theta \ln \frac{P_\theta^n(X_{1:n})}{P_\mu^n(X_{1:n})} \\ &= \mathbb{E}_\theta \sum_{i=1}^k T_i \ln\theta_i - \mathbb{E}_\theta \ln \frac{D_k(T_1+\frac{1}{2},\cdots,T_k+\frac{1}{2})}{D_k(\frac{1}{2},\cdots,\frac{1}{2})} \\ &= \ln\frac{\Gamma(1/2)^k}{\Gamma(k/2)} + \overbrace{\sum_{i=1}^k n\theta_i\ln\theta_i - \mathbb{E}_\theta \ln\frac{\prod_{i=1}^k \Gamma(T_i+\frac{1}{2})}{\Gamma(n+\frac{k}{2})}}^{\text{(A)}}. \end{aligned} \tag{4.5}$$

[2] Dans ce cas il s'agit aussi du mélange de Jeffrey.

Nous utilisons maintenant la formule de Stirling (4.4), en notant β_0 le coefficient correspondant à l'équation de $\Gamma(n+\frac{k}{2})$ et β_i celui correspondant à $\Gamma(T_i+\frac{1}{2})$.

$$\begin{aligned}(\mathrm{A}) &= \sum_{i=1}^{k} n\theta_i \ln\theta_i - \mathbb{E}_\theta \frac{\prod_{i=1}^{k}\left(\sqrt{2\pi}(T_i+\frac{1}{2})^{T_i}\right)}{\sqrt{2\pi}(n+\frac{k}{2})^{n+(k-1)/2}} - \sum_{i=1}^{k}\mathbb{E}_\theta \frac{\beta_i}{12(T_i+\frac{1}{2})} + \frac{\beta_0}{12(n+\frac{k}{2})} \\ &\geq -\frac{k-1}{2}\ln 2\pi - \frac{k}{6} + \overbrace{\sum_{i=1}^{k}\left(n\theta_i\ln\theta_i - \mathbb{E}_{\theta_i}T_i\ln\left(T_i+\frac{1}{2}\right)\right)}^{(\mathrm{B})} + \overbrace{\left(n+\frac{k-1}{2}\right)\ln\left(n+\frac{k}{2}\right)}^{(\mathrm{C})}.\end{aligned}$$

Nous minorons maintenant les termes (B) et (C) séparément. Pour le second nous avons

$$\begin{aligned}(\mathrm{C}) &= \left(n+\frac{k-1}{2}\right)\ln n + \left(n+\frac{k-1}{2}\right)\ln\left(1+\frac{k}{2n}\right) \\ &\geq n\ln n + \frac{k-1}{2}\ln n.\end{aligned}$$

Nous pouvons écrire le terme (B) sous la forme

$$(\mathrm{B}) = -n\ln n + \sum_{i=1}^{k}\overbrace{(n\theta_i\ln n\theta_i - \mathbb{E}_{\theta_i}T_i\ln T_i)}^{(\mathrm{B}_1)} - \sum_{i=1}^{k}\overbrace{\mathbb{E}_{\theta_i}T_i\ln\left(1+\frac{1}{2T_i}\right)}^{\leq 1/2}.$$

En partant de l'équation $t\ln t \leq t(t-1)$, et en posant $t=\frac{T_i}{n\theta_i}$, nous avons

$$\begin{aligned}\frac{T_i}{n\theta_i}(\ln T_i - \ln n\theta_i) &\leq \frac{T_i}{n\theta_i}\frac{T_i-n\theta_i}{n\theta_i} \\ T_i\ln T_i - T_i\ln n\theta_i &\leq \frac{(T_i-n\theta_i)^2}{n\theta_i} + (T_i-n\theta_i) \\ \mathbb{E}_{\theta_i}T_i\ln T_i - n\theta_i\ln n\theta_i &\leq \frac{\operatorname{Var}T_i}{n\theta_i} = 1-\theta_i,\end{aligned}$$

grâce à la relation $\mathbb{E}_{\theta_i}T_i = n\theta_i$. En conséquence

$$(\mathrm{B}_1) \geq -(1-\theta_i) \geq -1,$$

$$(\mathrm{B}) \geq -n\ln n - \frac{3k}{2},$$

et

$$(\mathrm{A}) \geq \frac{k-1}{2}\ln\frac{n}{2\pi} - \frac{5k}{3}.$$

Il ne reste plus qu'à rassembler les différents éléments de l'équation (4.5) pour obtenir le résultat annoncé. □

Une fois acquis ce résultat en dimension finie, nous voudrions l'exploiter pour traiter le cas de la classe $\Lambda_{Ce^{-\alpha\cdot}}$ en dimension infinie. Soit $m \in \mathbb{N}^*$ fixé tel que

$$\sum_{i \geq m} Ce^{-\alpha i} < 1,$$

et soit

$$\Theta_k = \left\{ \theta = (\theta_1, 0, \ldots, 0, \theta_m, \ldots, \theta_k, 0, \ldots) : \theta_1 = 1 - \sum_{i=m}^{k} \theta_i \text{ et } \forall m \leq i \leq k,\ 0 \leq \theta_i \leq Ce^{-\alpha i} \right\}.$$

Nous choisissons ensuite le prior μ_k sur Θ_k proportionnel au prior de Dirichlet μ sur S_{k-m+2} portant sur les coordonnées $(\theta_1, \theta_m, \ldots, \theta_k)$:

$$\begin{aligned} \mathrm{d}\mu_k(\theta_1, 0, \ldots, 0, \theta_m, \ldots, \theta_k, 0, \ldots) &= \frac{\theta_1^{-1/2}\theta_m^{-1/2} \cdots \theta_k^{-1/2}\, \mathrm{d}\theta}{\int_{\Theta_k} \theta_1^{-1/2}\theta_m^{-1/2} \cdots \theta_k^{-1/2}\, \mathrm{d}\theta} \\ &= \frac{\int_{S_{k-m+2}} \theta_1^{-1/2}\theta_m^{-1/2} \cdots \theta_k^{-1/2}\, \mathrm{d}\theta}{\int_{\Theta_k} \theta_1^{-1/2}\theta_m^{-1/2} \cdots \theta_k^{-1/2}\, \mathrm{d}\theta} \mathrm{d}\mu(\theta_1, \theta_m, \ldots, \theta_k), \end{aligned}$$

où $\mathrm{d}\theta$ est la mesure de Lebesgue sur le simplex S_{k-m+2} indexé par $(\theta_1, \theta_m, \ldots, \theta_k)$, et Θ_k est identifié à sa projection sur ce simplex. Par la suite nous poserons

$$\begin{aligned} C(k) &= \int_{\Theta_k} \theta_1^{-1/2}\theta_m^{-1/2} \cdots \theta_k^{-1/2}\, \mathrm{d}\theta, \\ D(k) &= \int_{S_{k-m+2}} \theta_1^{-1/2}\theta_m^{-1/2} \cdots \theta_k^{-1/2}\, \mathrm{d}\theta \\ &= \frac{\Gamma(1/2)^{k-m+2}}{\Gamma(\frac{k-m+2}{2})}. \end{aligned}$$

En pratique cela revient à se restreindre à l'alphabet $A_k = \{a_1, a_m, \ldots, a_k\} \subset A$. Soit donc $x_{1:n}$ un élément de A_k^n. Alors

$$\begin{aligned} P_{\mu_k}^n(x_{1:n}) &= \int_{\Theta_k} P_\theta(x_{1:n})\, \mathrm{d}\mu_k(\theta) \\ &= \frac{D(k)}{C(k)} \int_{\Theta_k} P_\theta(x_{1:n})\, \mathrm{d}\mu(\theta) \\ &\leq \frac{D(k)}{C(k)} \int_{S_{k-m+2}} P_\theta(x_{1:n})\, \mathrm{d}\mu(\theta) \\ &= \frac{D(k)}{C(k)} P_\mu^n(x_{1:n}), \end{aligned}$$

et si θ est un élément de Θ_k,

$$\begin{aligned} D(P_\theta^n \,\|\, P_{\mu_k}^n) &= \int_{A_k^n} \log \frac{P_\theta^n(x_{1:n})}{P_{\mu_k}^n(x_{1:n})} \,\mathrm{d}P_\theta^n(x_{1:n}) \\ &\geq \int_{A_k^n} \log \frac{P_\theta^n(x_{1:n})}{P_{\mu}^n(x_{1:n})} \,\mathrm{d}P_\theta^n(x_{1:n}) + \log C(k) - \log D(k) \\ &= D(P_\theta^n \,\|\, P_{\mu}^n) + \log C(k) - \log D(k) \\ &\geq \frac{k-m+1}{2} \log \frac{n}{2\pi} + \log C(k) - \frac{5(k-m+2)}{3} \log e, \end{aligned}$$

en vertu de la proposition 5. En conséquence

$$\begin{aligned} R_{n,\mu_k}^{\text{Bayes}} &= \int_{\Theta_k} D(P_\theta^n \,\|\, P_{\mu_k}^n) \,\mathrm{d}\mu_k(\theta) \\ &\geq \log C(k) + \frac{k}{2} \log n - \frac{10 \log e + 3 \log 2\pi}{6} k - \frac{m-1}{2} \log n \\ &\quad + \left(\frac{m-1}{2} \log 2\pi + \frac{5(m-2)}{3} \log e \right). \end{aligned} \tag{4.6}$$

Calculons maintenant $C(k)$. Remarquons tout d'abord que nous avons choisi m de telle sorte que tous les choix de $(\theta_m, \ldots, \theta_k)$ dans le rectangle $[0, Ce^{-\alpha m}] \times \cdots \times [0, Ce^{-\alpha k}]$ sont possibles, ce qui permet d'écrire les intégrales sur Θ_k comme intégrales sur ce rectangle. Nous avons alors

$$\begin{aligned} C(k) &= \int_{\Theta_k} \frac{\mathrm{d}\theta_m \cdots \mathrm{d}\theta_k}{\sqrt{\theta_1 \theta_m \cdots \theta_k}} \\ &\geq \prod_{i=m}^{k} \int_0^{Ce^{-\alpha i}} \frac{\mathrm{d}\theta_i}{\sqrt{\theta_i}} \\ &= \prod_{i=m}^{k} 2\sqrt{C} e^{-\frac{\alpha}{2} i}, \end{aligned}$$

d'où

$$\begin{aligned} \log C(k) &\geq (k-m+1) \left(1 + \frac{\log C}{2} \right) - \frac{\alpha \log e}{2} \sum_{i=m}^{k} i \\ &= (k-m+1) \left(1 + \frac{\log C}{2} \right) - \frac{\alpha \log e}{4} \left(k(k+1) - m(m+1) \right). \end{aligned}$$

Nous pouvons alors injecter ce résultat dans la relation 4.6, et obtenir $R_{n,\mu_k}^{\text{Bayes}} \geq g(n,k)$,

avec

$$g(n,k) = -\frac{\alpha \log e}{4} k^2 + \frac{k}{2} \log n + \left[1 + \frac{\log C}{2} - \frac{\alpha \log e}{4} - \frac{5 \log e}{3} - \frac{\log 2\pi}{2}\right] k \\ - \frac{m-1}{2} \log n + \left[\frac{\alpha(m^2+m-1)\log e}{4} - (m-1)\left(1 + \frac{\log C}{2}\right) \right. \\ \left. + \frac{m-1}{2} \log 2\pi + \frac{5(m-2)\log e}{3}\right].$$

En choisissant $k_n = \frac{1}{\alpha \log e} \log n$, seuls les deux premiers termes comptent, et nous obtenons

$$g(n, k_n) \sim \frac{1}{4\alpha \log e} \log^2 n.$$

Il ne nous reste plus qu'à énoncer le résultat obtenu :

Théorème 11. *On considère $\Lambda_{Ce^{-\alpha \cdot}}$ une classe enveloppe à décroissance exponentielle. Alors, lorsque $n \to \infty$,*

$$R_n^{\text{minimax}}(\Lambda_{Ce^{-\alpha \cdot}}) \geq (1 + o(1)) \frac{1}{4\alpha \log e} \log^2 n,$$

et en conséquence

$$R_n^{\text{minimax}}(\Lambda_{Ce^{-\alpha \cdot}}) \sim \frac{1}{4\alpha \log e} \log^2 n.$$

5 Et au-delà ?

Avec notre théorème 11 nous avons illustré l'intérêt du recours à l'entropie métrique et aux résultats de [HO97] pour calculer la redondance de classes de sources sans mémoire.

Dans le cas des classes enveloppe à décroissance rapide, de type exponentielle, nous avons constaté que l'entropie métrique restait « petite » et assurait des résultats fins sur l'encadrement de la redondance minimax. En revanche, des classes enveloppe à décroissance plus lente, de type polynomiale, entraîne une écart entre l'ordre de grandeur de la majoration et celui de la minoration. On peut cependant espérer que la minoration fournisse le bon ordre de grandeur, mais il reste un travail à accomplir pour minorer efficacement l'entropie métrique et ainsi tester cette hypothèse.

Par ailleurs dans ce travail a été développée, en collaboration étroite avec Elisabeth Gassiat, une technique pour minorer la redondance minimax, issue des travaux de Xie et Barron sur le mélange de Dirichlet. Peut-être pourrons-nous réutiliser cela pour d'autres classes de sources ?

Enfin il reste à traduire ce travail théorique sur la redondance en la découverte éventuelle d'un nouveau code adapté aux classes enveloppe à décroissance exponentielle.

Index

Bibliographie

[BGG06] Stéphane Boucheron, Aurélien Garivier et Elisabeth Gassiat: *Coding on countably infinite alphabets.* article soumis, dec. 2006.

[CB90] Bertrand S. Clarke et Andrew R. Barron: *Information-theoretic asymptotics of Bayes methods.* IEEE Transactions on Information Theory, 36(3) :453–471, 1990.

[CB94] Bertrand S. Clarke et Andrew R. Barron: *Jeffrey's prior is asymptotically least favorable under entropy risk.* Journal of Statistical Planning and Inference, 41 :37–60, 1994.

[CT91] Thomas M. Cover et Joy A. Thomas: *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley & sons, 1991, ISBN 0-471-06259-6.

[Gar06] Aurélien Garivier: *Modèles contextuels et alphabets infinis en théorie de l'information.* Thèse de doctorat, Université Paris-Sud 11, nov. 2006.

[Hau96] David Haussler: *A general minimax result for relative entropy.* Rapport technique UCSC-CRL-96-26, 1996. http ://citeseer.ist.psu.edu/haussler96general.html.

[HO97] David Haussler et Manfred Opper: *Mutual information, metric entropy and cumulative relative entropy risk.* The Annals of Statistics, 25(6) :2451–1492, 1997.

[KT61] A. N. Kolmogorov et V. M. Tikhomirov: *ϵ-entropy and ϵ-capacity of sets in functional spaces.* Amer. Math. Soc. Trans. Ser. 2, 17 :277–364, 1961.

[vdVW96] Aad W. van der Vaart et Jon A. Wellner: *Weak Convergence and Empirical Processes.* Springer Series in Statistics. Springer-Verlag, 1996, ISBN 0-387-94640-3.

[WW96] E. T. Whittaker et G. N. Watson: *A Course of Modern Analysis.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, reprint of the fourth (1927) edition édition, 1996, ISBN 0-521-58807-3. An Introduction to the general theory of infinite processes and of analytic functions ; with an account of the principal transcendental functions.

[XB97] Qun Xie et Andrew R. Barron: *Minimax redundancy for the class of memoryless sources.* IEEE Transactions on Information Theory, 43(2) :646–657, 1997.