Z. Wahrscheinlichkeitstheorie verw. Gebiete
65, 181–237 (1983)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete

Approximation dans les espaces métriques et théorie de l'estimation

Lucien Birgé

U.E.R. de Sciences Économiques, Université de Paris X Nanterre, 200 Avenue de la République
F-92001 Nanterre Cedex, France

Summary. We investigate the relations between the speed of estimation and the metric structure of the parameter space Θ, especially in the case when its metric dimension is infinite. Given some distance d on Θ (generally Hellinger distance in the case of n i.i.d. variables), we consider the minimax risk for n observations: $R_n(q)=\inf_{T_n}\sup_{\theta\in\Theta}\mathbb{E}_\theta[d^q(\theta, T_n)]$, T_n being any estimate of θ. We shall look for functions r such that for positive constants $C_1(q)$ and $C_2(q)$ $C_1 r^q(n)\leqq R_n(q)\leqq C_2 r^q(n)$. $r(n)$ is the speed of estimation and we shall show under fairly general conditions (including i.i.d. variables and regular cases of Markov chains and stationnary gaussian processes) that $r(n)$ is determined, up to multiplicative constants, by the metric structure of Θ. We shall also give a construction for some sort of "universal" estimates the risk of which is bounded by $C_2 r^q(n)$ in all cases where the preceding theory applies.

1. Introduction

Le sujet que nous allons aborder ici est celui de la vitesse d'estimation d'un problème statistique. Etant donné un espace de paramètres Θ et des suites $P_{\theta,n}$ de lois de probabilités indexées par θ, nous voulons déterminer quelle est la vitesse optimale de convergence vers θ d'une suite $\{\hat{\theta}_n\}$ d'estimateurs, le but étant de traiter globalement des problèmes considérés comme distincts et de tenter de présenter une théorie unifiée de la vitesse d'estimation d'un problème statistique en liant cette dernière aux propriétés des mesures $P_{\theta,n}$. Dans tous les problèmes classiques cette vitesse apparaît sous forme d'un facteur de normalisation, lequel s'introduit naturellement dans chaque cas particulier lorsque l'on veut faire des calculs de risque ou de lois limites d'estimateurs.

Dans le cas paramétrique ($\Theta\subset\mathbb{R}^k$) régulier, on étudie des estimateurs $\hat{\theta}_n$ tels que $\sqrt{n}(\hat{\theta}_n-\theta)$ ait une loi limite gaussienne ou bien l'on s'efforce de minimiser $\limsup_n \sup_\theta \mathbb{E}_\theta[(\sqrt{n}\,\|\hat{\theta}_n-\theta\|)^2]$. Le facteur $\sqrt{n}$ apparaît comme corollaire de l'utilisation du théorème central limite ou de considérations analogues. Dans ce cas la vitesse d'estimation est $n^{-1/2}$, au sens où la distance moyenne de $\hat{\theta}_n$ à θ (pour n observations) est de l'ardre de $n^{-1/2}$ [1].

Par contre, lorsque Θ n'est plus dans $\mathbb{R}^k$ (entre autre dans les problèmes d'estimation de densités) il en va tout autrement: on trouve des vitesses en $n^{-\alpha}$, $\alpha<\frac{1}{2}$ (cf. [8], [12] et [33]) suivant les exemples, calculées dans chaque cas particulier et non justifiées par des considérations générales. De nombreux types d'estimateurs ont été proposés (noyaux, spines, séries orthogonales, etc. ... sans qu'aucun d'eux n'obtienne la faveur générale. C'est pourquoi l'objet principal du travail qui suit est de résoudre partiellement les deux problèmes suivants:

i) fournir une théorie de la vitesse d'estimation fondée sur les propriétés de séparation des mesures $P_{\theta,n}$ et décrite par la structure métrique de l'espace des paramètres.

ii) donner une construction générale de «bons» estimateurs atteignant la vitesse convenable, ceci d'une manière unique pour tous les cas qui relèvent de la théorie précédente.

Nous nous efforcerons ici de conserver un cadre assez large, recouvrant une grande partie des problèmes classiques d'estimation non paramétrique pour divers types de processus (variables indépendantes, chaînes de Markov et processus gaussiens stationaires). Etant donnée la généralité des hypothèses utilisées (fort différentes de celles communément employées), il ne sera pas question ici d'optimalité ni de lois limites pour les estimateurs vu que nous utiliserons le risque minimax et qu'il est impossible de le déterminer exactement même asymptotiquement, sans hypothèses supplémentaires [2]. Par contre, les estimateurs que nous allons construire auront de très bonnes propriétés de robustesse. De plus nous n'obtiendrons pas seulement des évaluations asymptotiques des vitesses d'estimation mais aussi des encadrements du risque minimax pour des valeurs finies du nombre des observations.

Ce travail doit beaucoup à celui de L. Le Cam. C'est lui qui le premier a obtenu des résultats généraux de majoration du risque dans le cas de n variables indépendantes équidistribuées en fonction de la structure métrique de l'espace des paramètres et qui a eu l'idée de construire des estimateurs à partir de tests. Il a aussi donné une minoration universelle de la vitesse d'estimation pour la distance de Hellinger h qui peut s'écrire ainsi: s'il existe dans Θ s_n et t_n tels que $0<a\leqq nh^2(s_n, t_n)\leqq b$, alors

$$\inf_{T_n}\sup_{\theta\in\Theta} \mathbb{E}_\theta[nh^2(\theta, T_n)]\geqq C \tag{1.1}$$

où C est une constante qui ne dépend que de a et b.

D'autres résultats de minoration ont été obtenus dans un cadre moins général, celui de l'estimation des densités, et pour des familles d'un type

[1] D'autres facteurs de normalisation peuvent apparaître même lorsque θ est réel (cf. [10] et [22])

[2] Voir [30] pour un exemple de calcul asymptotique du risque minimax exact

particulier par Farrell [12], Wahba [33] et Bretagnolle-Huber [8]. La comparaison qu'ils ont faite avec les performances obtenues par les estimateurs classiques (à noyaux par exemple) montre que ces bornes inférieures sont adéquates (cf. également [17] et [18]).

Farrell [13] et C. Huber [14] ont fait des études analogues pour les densités spectrales et obtiennent les mêmes valeurs que dans le cas des variables indépendantes, ce qui amène Farrell dans [13] à s'interroger sur le pourquoi du phénomène. Ibragimov et Khas'minskii ont également étudié le cas des équations différentielles stochastiques et pour des structures identiques de l'espace des paramètres ils obtiennent les mêmes vitesses que dans le cas de l'estimation des densités (cf. [16]). Nous pourrons expliquer ces analogies, dans des problèmes d'estimation en apparence bien différents, en montrant que sous des hypothèses convenables la vitesse d'estimation est déterminée par certaines propriétés métriques de l'espace des paramètres.

Cadre statistique

$(\mathscr{X}, \mathscr{A})$ étant un ensemble mesurable (l'espace des observations) on considère une famille $\{P_{\theta,n}\}_{\theta\in\Theta}$ de lois de probabilités sur l'espace produit $(\mathscr{X}^n, \mathscr{A}^n)$; l'exemple le plus simple est celui des variables indépendantes équidistribuées où $P_{\theta,n}$ est la loi produit P_θ^n. Θ est un sous-ensemble d'un espace métrique (E, d) et nous supposerons que les lois $P_{t,n}$ sont aussi définies lorsque t appartient à E (E peut bien entendu être égal à Θ). Nous noterons Θ_n (resp. E_n) l'ensemble des mesures $P_{\theta,n}$ où θ parcourt Θ (resp. E) et nous identifierons le plus souvent Θ et Θ_n, E et E_n. Par cette identification Θ_n devient un espace métrique, mais nous utiliserons aussi la distance $d_n = n^{\frac{1}{2}} d$ sur Θ_n, pour des raisons de normalisation. Cette identification nous amènera à noter $d_n(s, t)$ pour $d_n(P_{s,n}, P_{t,n})$. Nous nous poserons le problème de l'estimation de θ dans l'expérience $\mathscr{E}_n = \{\mathscr{X}^n, \mathscr{A}^n, E_n, d_n, \Theta_n\}$ et lorsque n sera égal à 1, nous abandonnerons les indices pour parler de $\mathscr{E} = \{\mathscr{X}, \mathscr{A}, E, d, \Theta\}$.

Les fonctions de perte que nous utiliserons seront de la forme $\ell \circ d_n$ (ou $\ell \circ d$), ℓ étant une fonction croissante de $\mathbb{R}^+$ dans $\mathbb{R}^+$ telle que $\ell(0) = 0$. En fait, nous nous limiterons, le plus souvent, à considérer les seules fonctions vérifiant $\ell(2x) \leqq C\ell(x)$, $C > 0$, celles qui ont une croissance plus rapide donnant lieu à des résultats partiels seulement. Ceci est dû au fait que nous n'introduirons pas de facteur normalisant dans ℓ (du genre $\ell(n^{-\alpha}x)$) comme dans [16, 17] et [18]. En procédant ainsi on pourrait aussi obtenir des résultats pour des fonctions ℓ plus générales.

Etant donné un estimateur T_n de θ, fondé sur les n observations $X_1, \ldots, X_n$, le risque de T_n pour la fonction de perte $\ell \circ d_n$, calculé au point θ s'écrit

$$\mathbb{E}_\theta[\ell(d_n(\theta, T_n))].$$

Si $\mathscr{T}_n$ désigne l'ensemble des estimateurs possibles, c'est-à-dire des fonctions mesurables de $(\mathscr{X}, \mathscr{A})$ dans Θ muni de la tribu borélienne induite par d, le risque minimax sera défini par

$$R_n(\ell \circ d_n) = \inf_{T_n \in \mathcal{T}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\ell[d_n(\theta, T_n)]].$$

Lorsque $\ell(x) = x^q$, au lieu de $R_n(d_n^q)$ nous considérerons le plus souvent $R_n(d^q) = n^{-\frac{q}{2}} R_n(d_n^q)$ dont l'expression sera plus simple. Le calcul exact de $R_n(d^q)$ étant impossible dans le cadre général que nous allons nous fixer, nous nous contenterons d'encadrements uniformes en n ce qui nous amène à poser:

Définition 1.1. Nous dirons que deux fonctions f et g sont équivalentes s'il existe deux constantes C_1 et C_2 strictement positives et telles que

$$C_1 f \leqq g \leqq C_2 f.$$

Ce que nous noterons $f \asymp g$. Nous rechercherons des fonctions r de $\mathbb{N}^*$ dans $\mathbb{R}^+$ telles que $r(n) \asymp R_n(d^q)$, au moins pour $n \geqq n_0$. Dans ce cas si $C_1 r(n) \leqq R_n(d^q) \leqq C_2 r(n)$, $C_1 r(n)$ et $C_2 r(n)$ seront dites bornes adéquates pour le risque minimax et $r(n)$ sera la vitesse d'estimation du problème (définie à des constantes multiplicatives près). Pour la déterminer, nous allons construire des suites d'estimateurs $\hat{\theta}_n$, dits d-estimateurs, tels que:

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta[d_n^q(\theta, \hat{\theta}_n)] \leqq r_2(n), \tag{1.2}$$

puis nous calculerons des minorations $r_1(n)$ du risque minimax $R_n(d_n^q)$ et nous vérifierons que le rapport $\frac{r_2(n)}{r_1(n)}$ reste borné. Les suites $\{\hat{\theta}_n\}_{n \geqq 1}$ qui vérifient (1.2) avec $r_2(n) \leqq C r_1(n)$ seront dites suites adéquates d'estimateurs. Il apparaîtra clairement que les bornes obtenues pour $\frac{r_2(n)}{r_1(n)}$ sont très grandes. Il y a à cela plusieurs explications liées à la formulation même du problème et il semble peu vraisemblable de pouvoir améliorer sensiblement ces bornes en conservant ce niveau de généralité. Cependant, il est probable que les performances réelles des d-estimateurs seront bien plus proches du risque minimax que ce qu'indique ce rapport. En effet nos calculs de $r_2(n)$ sont très grossiers en raison de la forme des hypothèses et amènent à une surestimation du risque des d-estimateurs, d'autre part les minorations $r_1(n)$ ne sont pas du tout précises.

Il est certain que l'on pourrait améliorer l'estimation en précisant davantage les hypothèses et que le fait de traiter simultanément les variables indépendantes, les chaînes de Markov et les processus gaussiens avec un même type d'estimateur est un sérieux handicap. En pratique, dans un cas précis, on pourra utiliser des estimateurs de type classique mieux adaptés au problème (estimateurs à noyaux par exemple, dans le cas de l'estimation d'une densité, cf. [16] et [17]) mais il n'est pas du tout clair qu'ils auront les mêmes propriétés de robustesse que les d-estimateurs. On a ici l'avantage d'obtenir une théorie générale qui permet de déterminer a priori la vitesse du problème (en particulier de permettre un bon choix de «fenêtres» pour des estimateurs à noyaux).

Quelques exemples d'applications

Les techniques développées ici sont avant tout adaptées au cas de suites de variables indépendantes équidistribuées pour lequel les hypothèses nécessaires à la construction des estimateurs seront naturellement vérifiées. En particulier nous considérerons divers exemples, plus ou moins classiques, d'estimation de densités par rapport à la mesure de Lebesgue. Les classes de fonctions que nous étudierons plus spécialement sont les suivantes:

i) les ellipsoïdes $E(a)$ de $\mathbb{L}^2(0;1)$, c'est-à-dire les ensembles de fonctions f sur $[0;1]$ qui vérifient pour une certaine suite décroissante $a=\{a_k\}_{k\geqq 1}$

$$\sum_{k=1}^{+\infty}\frac{x_k^2+y_k^2}{a_k^2}\leqq 1, \qquad x_0=1,$$

où x_k et y_k sont les coefficients de $\sin kx$ et $\cos kx$ dans le développement de f en série trigonométrique.

ii) les ensembles Λ_ω de fonctions sur un intervalle fermé, ayant un module de continuité majoré par un module de continuité donné ω.

iii) les fonctions définies sur un parallélépipède K de $\mathbb{R}^s$ ayant des dérivées partielles jusqu'à l'ordre r, celles d'ordre r étant Hölderiennes d'ordre α, $0<\alpha\leqq 1$, ou bornées $(\alpha=0)$. Elles forment la classe $\Lambda^s_{\alpha,r}$.

iv) $\Lambda^{s,p}_{\alpha,r}$ est l'analogue de $\Lambda^s_{\alpha,r}$, mais la condition de Hölder s'écrit en norme $\mathbb{L}^p$, $1\leqq p\leqq 2$. C'est-à-dire pour $s=1$, $K=[0;1]$

$$\int_0^1 |f^{(r)}(x+h)-f^{(r)}(x)|^p\leqq C h^{p\alpha}.$$

Vu le type de résultats cherchés, il est indispensable que les distances utilisées reflètent la structure des ensembles de probabilités Θ_n, en particulier celle des tests entre les points de Θ_n. C'est ce qui confère un rôle fondamental à la distance de Hellinger h:

$$h^2(P,Q)=\tfrac{1}{2}\int(\sqrt{dP}-\sqrt{dQ})^2.$$

Nous pourrons utiliser également d'autres distances, mais toujours majorées par un multiple de h. En particulier lorsque Θ est une famille de densités uniformément bornées supérieurement, nous utiliserons les distances déduites des normes $\mathbb{L}^p$ pour $1\leqq p\leqq 2$.

La théorie que nous allons développer s'applique également à d'autres modèles que celui des variables indépendantes, tels que les chaînes de Markov ou les processus gaussiens. Mais, pour les traiter, nous aurons besoin d'hypothèses plus restrictives et nous utiliserons la propriété suivante:

Définition 1.2. Une famille F de fonctions positives satisfait à la condition $LB(\gamma_1,\gamma_2)$ s'il existe deux nombres $\gamma_1,\gamma_2>0$, tels que pour tout f dans F

$$\gamma_1\leqq f\leqq\gamma_2.$$

Une telle famille sera dite log-bornée.

Pour ce qui concerne les chaînes de Markov, nous supposerons que l'on dispose d'une probabilité μ sur $\mathscr{X}$ et que les noyaux de transition ont tous des densités $p_\theta(x, y)$ par rapport à μ, la mesure initiale ν_θ étant par ailleurs arbitraire. Nous imposerons que la famille $\{p_\theta(x, y)\}_{\theta\in\Theta}$ soit log-bornée[3] et choisirons pour d la distance (analogue à celle de Hellinger)

$$d^2(s, t)=\tfrac{1}{2}\int(\sqrt{p_s(x, y)}-\sqrt{p_t(x, y)})^2\,\mu(dx)\mu(dy).$$

Nous étudierons plus particulièrement l'exemple de fonctions $p_s(x, y)$ sur $[0; 1]^2$, lipschitziennes en x et höldériennes en y.

Pour les modèles de processus gaussiens, nous supposerons que le processus a une densité spectrale f_θ, la famille $\{f_\theta\}_{\theta\in\Theta}$ étant log-bornée. Nous pourrons ici reprendre certains des exemples développés dans le cas des variables indépendantes en particulier celui où $\Theta=\Lambda_\omega$. Nous utiliserons comme distance sur Θ la distance $\mathbb{L}^2$ entre les densités spectrales.

Les hypothèses

La construction des estimateurs fait intervenir une famille de tests entre des sous-ensembles (boules le plus souvent) de Θ et les hypothèses correspondantes porteront à la fois sur les propriétés de recouvrement de Θ par ces ensembles et sur celles de ces tests. Rappelons tout d'abord qu'un (η, d)-réseau S de Θ (ou η-réseau lorsque d est clairement spécifié) est un sous-ensemble de E tel que tout point de Θ soit à distance inférieure ou égale à η d'un point de S.

Définition 1.3. Soit S un réseau de Θ. Un recouvrement $\{B_s\}_{s\in S}$ de Θ sera dit centré sur S si $s\in B_s$ pour tout s dans S. Il sera dit (η, d)-recouvrement (centré sur S) si l'on a en outre $d(s, s')\leqq\eta$ pour tout s' dans B_s.

Etant donnée une expérience $\mathscr{E}=(\mathscr{X}, \mathscr{A}, E, d, \Theta)$, la construction d'un d-estimateur s'appuiera sur:

i) un (η, d)-réseau S de Θ et un recouvrement $\{B_s\}_{s\in S}$ de Θ centré sur S.

ii) une famille de fonctions de tests non randomisés (fonctions indicatrices d'ensembles) $\{\varphi_{s,t}\}_{(s,t)\in S\times S}$ de B_s contre B_t telles que $\varphi_{s,t}=1-\varphi_{t,s}$.

Ils devront satisfaire aux hypothèses:

H1 (η, k_0, δ): Il existe un nombre positif δ tel que l'on ait pour tout point s de S

$$\operatorname{Card}\{S\cap\mathscr{B}(s, 2^j\eta)\}\leqq 2^{j\delta}, \qquad \text{si } 2^j\geqq 2k_0, \tag{1.3}$$

où $\mathscr{B}(s, x)$ désigne la boule ouverte de centre s et de rayon x.

H2 (d_0, A_1, A_2): Il existe un réel $d_0>0$ et des constantes A_1 et A_2 indépendants de s et t tels que si $d(s, t)\geqq d_0$,

$$\sup_{s'\in B_s}\mathbb{E}_{s'}[\varphi_{s,t}]\leqq A_1\exp[-A_2 d^2(s, t)]. \tag{1.4}$$

[3] Ces hypothèses sont à rapprocher de celles utilisées par Donsker et Varadhan dans [11]

La première hypothèse est d'ordre purement métrique et s'apparente beaucoup à A3 dans [24]. Pour ce qui est de la seconde c'est une propriété de grandes déviations reliant les erreurs des tests à la distance des ensembles à tester tout à fait analogue à celles utilisées dans [21] et [23]. On remarquera que si des tests randomisés satisfont à (1.4) on peut les remplacer par des tests non randomisés, quitte à changer A_1 en $2A_1$.

Sous ces hypothèses, lorsque $\{B_s\}_{s\in S}$ est un (η, d)-recouvrement, nous construirons des estimateurs $\hat{\theta}$ dont le risque vérifiera

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta[d^q(\theta,\hat{\theta})] \leqq C\eta^q \qquad \text{si } \eta^2 \geqq \delta \geqq \tfrac{1}{2}.$$

Cette inégalité fait apparaître un lien entre le risque et la structure métrique de Θ définie par la taille de ses réseaux. Lorsque on étudiera des suites d'expériences, on sera amené à considérer des (η, d_n)-réseaux de Θ_n qui sont également des $(\eta n^{-\frac{1}{2}}, d)$-réseaux de Θ, ce qui nous amène à définir une fonction $\tilde{d}$ par

$$\tilde{d}(\varepsilon) = \inf_S \sup_{\substack{j\geqq 3\\ j\in\mathbb{N}}} \sup_{\theta\in\Theta} \left\{\frac{\log \operatorname{Card}\{S\cap\mathscr{B}(\theta, 2^j\varepsilon)\}}{j\log 2}\right\}, \tag{1.5}$$

l'inf étant pris sur tous les ε-réseaux S. $\tilde{d}$ est relative au couple (Θ, d) mais permet aussi d'exprimer H1 pour n observations. Lorsque $\tilde{d}$ est bornée ce qui est le cas sur $\mathbb{R}^k$ muni de la distance euclidienne, nous dirons que l'espace (Θ, d) est de dimension métrique finie.

Telle qu'elle est définie, la fonction $\tilde{d}$ est assez laborieuse à manipuler, mais elle est fort heureusement liée à des notions classiques de théorie de l'approximation telles que l'ε-entropie, ce qui fait qu'il sera souvent possible non pas de calculer $\tilde{d}$ mais d'en obtenir des majorants qui seront des fonctions décroissantes de ε. Cela sera suffisant pour établir les résultats qui nous intéressent concernant les vitesses. En particulier nous verrons que pour les classes Λ_ω et $\Lambda^s_{\alpha,r}$ on a respectivement $\tilde{d}(2\omega(t)) \leqq \frac{3}{2}t^{-1}$ et $\tilde{d}(\varepsilon) \leqq C\varepsilon^{-\frac{s}{\alpha+r}}$ où d est soit la distance de Hellinger soit la distance $\mathbb{L}^p$, $1\leqq p\leqq 2$.

De même qu'on a défini $\tilde{d}$ relativement à (Θ, d), le couple (Θ_n, d_n) fournit une fonction $\tilde{d}_n$ et il est immédiat que $\tilde{d}_n(\eta) = \tilde{d}(\eta n^{-\frac{1}{2}})$ ce qui permettra de se ramener toujours à $\tilde{d}$. En fait l'application de (1.3) amène à choisir $\eta^2 \geqq \delta$ mais le plus petit possible, aussi serons-nous amenés à chercher dans le cas de n observations les solutions de l'équation $\tilde{d}_n(\eta) = \eta^2$ ou encore si $\varepsilon = n^{-\frac{1}{2}}\eta$, $n\varepsilon^2 = \tilde{d}(\varepsilon)$.

En pratique $\tilde{d}$ est toujours remplacé par une fonction majorante δ décroissante de $\mathbb{R}^+$ dans $\mathbb{R}^+ \cup \{\infty\}$ qui prend au moins une valeur finie, à laquelle nous associerons la fonction $L\delta$ de $\mathbb{N}^*$ dans $\mathbb{R}^*$ par

$$L\delta(n) = \inf\{x \mid nx^2 > \delta(x)\}. \tag{1.6}$$

Nous en déduirons des majorations du risque du type $R_n(d^q) \leqq C[L\delta(n)]^q$. En particulier pour Λ_ω $R_n(d^q) \leqq C_1[\omega(t_n)]^q$ avec $t_n\omega^2(t_n) = \dfrac{3}{8n}$, et pour $\Lambda^s_{\alpha,r}$ $R_n(d^q) \leqq C_2 n^{\frac{-q(r+\alpha)}{s+2(r+\alpha)}}$. Ces bornes se révéleront d'ailleurs adéquates.

L'introduction de H2 est liée à l'utilisation de la distance de Hellinger dans le cas de n variables indépendantes équidistribuées. Nous rappellerons d'abord deux notions introduites par Le Cam. L'affinité de Hellinger entre deux probabilités s'écrit:

$$\rho(P, Q) = \int \sqrt{dP\,dQ} = 1 - h^2(P, Q).$$

Si $\mathscr{P}$ et $\mathscr{Q}$ sont deux ensembles de probabilités, leur parenté π s'écrit

$$\pi(\mathscr{P}, \mathscr{Q}) = \inf_{\varphi} \left[\sup_{P \in \mathscr{P}} \int \varphi\, dP + \sup_{Q \in \mathscr{Q}} \int (1-\varphi)\, dQ\right]. \tag{1.7}$$

φ décrivant l'ensemble des fonctions de tests, c'est-à-dire des fonctions mesurables à valeur dans $[0; 1]$. π représente donc le minimax de la somme des erreurs des tests entre $\mathscr{P}$ et $\mathscr{Q}$. Si l'on se restreint à des φ indicatrices d'ensembles, il suffit de remplacer dans (1.7) π par π' et $\pi' \leqq 2\pi$.

Lorsque $\mathscr{P}$ et $\mathscr{Q}$ sont réduits à un seul point respectivement P et Q, on notera $\pi(P, Q)$. Il est facile de voir que π est alors réalisé en choisissant $\varphi = 1_{\{dQ > dP\}}$ (voir [10]) et que l'on a les relations

$$1 - \tfrac{1}{2}\|P - Q\|_1 = \pi(P, Q), \qquad \pi^2 \leqq \rho^2 \leqq \pi(2 - \pi), \qquad h^2 \leqq \tfrac{1}{2}\|P - Q\|_1 \leqq \sqrt{2}\, h, \tag{1.8}$$

où $\|P - Q\|_1$ désigne la norme en variation totale, c'est-à-dire la norme $\mathbb{L}^1(P + Q)$.

Il a été démontré par Le Cam ([21] et [23]) que si $\mathscr{P}$ et $\mathscr{Q}$ sont deux ensembles convexes disjoints de probabilités, on a l'inégalité

$$\pi(\mathscr{P}^n, \mathscr{Q}^n) \leqq \sup_{P \in \mathscr{P},\, Q \in \mathscr{Q}} \rho^n(P, Q), \tag{1.9}$$

où $\mathscr{P}^n$ désigne l'ensemble des probabilités produits $P_1 \otimes P_2 \otimes \ldots \otimes P_n$ où P_i est un élément de $\mathscr{P}$, $1 \leqq i \leqq n$.

Des résultats voisins mais moins forts ont été obtenus par Birgé [6]. Ils permettent de majorer séparément chacune des erreurs de certains tests non randomisés par la même quantité $\sup_{P \in \mathscr{P}, Q \in \mathscr{Q}} \rho^n(P, Q)$ ce qui entraîne:

$$\pi(\mathscr{P}^n, \mathscr{Q}^n) \leqq 2 \sup_{P \in \mathscr{P},\, Q \in \mathscr{Q}} \rho^n(P, Q). \tag{1.10}$$

Mais alors que (1.9) ne montre que l'existence de tests ayant de telles performances, les résultats de [6] permettent, dans le cas où $\mathscr{P}$ et $\mathscr{Q}$ sont deux boules pour la distance de Hellinger d'obtenir explicitement les tests satisfaisant à (1.10). Cela sera intéressant en vue d'une construction effective des d-estimateurs.

Métrisons l'espace des probabilités sur $\mathscr{X}$ par h. Si $\mathscr{P}$ et $\mathscr{Q}$ sont les boules fermées $\bar{\mathscr{B}}(P, \varepsilon)$ et $\bar{\mathscr{B}}(Q, \varepsilon)$ et $h(P, Q) \geqq 2\varepsilon$,

$$\sup_{P' \in \mathscr{P},\, Q' \in \mathscr{Q}} \rho(P', Q') = 1 - \inf_{P' \in \mathscr{P},\, Q' \in \mathscr{Q}} h^2(P', Q') \leqq 1 - [h(P, Q) - 2\varepsilon]^2,$$

d'où en utilisant (1.9):

$$\pi(\bar{\mathscr{B}}^n(P, \varepsilon), \bar{\mathscr{B}}^n(Q, \varepsilon)) \leqq \exp[-n(h(P, Q) - 2\varepsilon)^2]. \tag{1.11}$$

Les tests explicites que l'on peut déduire de [6] vérifient une inégalité analogue à cela près que π est remplacé par l'une ou l'autre des erreurs des tests, mais cela sera suffisant pour l'application à H2.

Considérons une expérience $\mathscr{E}_n$ avec n variables indépendantes équidistribuées, $d_n = n^{\frac{1}{2}} h$. Soit S un (η, d_n)-réseau dans Θ_n et un η-recouvrement associé en prenant $B_s = \bar{\mathscr{B}}(s, \eta)$. Alors $B_s = \{P^n | h(P_s, P) \leqq n^{-\frac{1}{2}} \eta\}$. Il s'ensuit que B_s est inclus dans la puissance n-ième de la boule $\bar{\mathscr{B}}(s, \varepsilon)$, $\varepsilon = n^{-\frac{1}{2}} \eta$, de (Θ, h). On peut alors appliquer l'inégalité (1.11) ce qui implique si $d(s,t) = h(s,t) \geqq 2\varepsilon$,

$$\pi(B_s, B_t) \leqq \exp\left[-n(d(s,t) - 2\varepsilon)^2\right].$$

Ceci entraîne alors (1.4) avec $A_1 = 2 + \eta$, $\eta > 0$ arbitrairement petit, $A_2 = \frac{1}{4}$ et $k_0 = 4$. L'utilisation des tests déduits de [6] permet de prendre $A_1 = 1$ dans la mesure où on obtient une majoration de chacune des deux erreurs. Ainsi H2 sera satisfaite sur (Θ_n, d_n).

Si l'on utilise pour d une distance autre que h qui vérifie $d \leqq Mh$, ce qui sera le cas dans de nombreux exemples pour les distances $\mathbb{L}^p$, $1 \leqq p \leqq 2$, on obtient le même résultat avec $A_2 = \dfrac{1}{4M^2}$ et $k_0 = 4$. Dans le cas des variables indépendantes équidistribuées l'hypothèse H2 sera donc aisément vérifiée. Nous étudierons également des extensions de l'inégalité (1.4) lorsque B_s est une boule pour la distance $\mathbb{L}^\infty$.

Les démonstrations de minoration, contrairement à ce qui précède, utiliseront la structure locale de Θ, au voisinage d'un point θ_0 à définir. Elles reposent sur l'utilisation de l'information de Kullback définie par

$$\begin{aligned} K(P,Q) &= \int \log \frac{dP}{dQ}\, dP \qquad \text{si } P \ll Q \\ &= +\infty \qquad\qquad\qquad \text{sinon,} \end{aligned}$$

et s'inspirent directement des techniques utilisées par Ibragimov et Khas'minskii dans [16] et [17].

Rappelons que $K(P^n, Q^n) = nK(P,Q)$ et considérons l'hypothèse

H3(η', γ): Il existe un sous-ensemble fini S' de Θ de cardinal $m+1$, tel que l'on ait:

$$d(s,t) \geqq \eta' \quad \text{pour tous } s \text{ et } t, \quad s \neq t \quad \text{dans } S' \tag{1.12}$$

$$\sup_{s,t \in S'} K(P_s, P_t) \leqq \gamma \log m - \log 2. \tag{1.13}$$

Cette hypothèse va entraîner une minoration du risque de la forme $R(\ell, d) \geqq (1-\gamma)\ell\left(\dfrac{\eta'}{2}\right)$ comme nous le verrons. Le lien avec les majorations pourra se faire si dans H3 $\log m$ est de l'ordre du δ qui intervient dans H2 et η' de l'ordre de η. Comme δ sera choisi voisin de η^2, ceci nécessite que les informations de Kullback entre les points de S' soient aussi de l'ordre de η^2. En fait il faudra que l'on ait simultanément sur S'

$$d(s,t) \geqq \eta' \quad \text{et} \quad K(s,t) \leqq \eta^2$$

où η n'est pas beaucoup plus grand que η' c'est-à-dire que K et d^2 soient comparables sur S'. La comparaison de $K(P, Q)$ et $d^2(P, Q)$ est très facile lorsque les rapports de vraisemblance $\frac{dP}{dQ}$ sont log-bornés uniformément en P et Q. Dans le cas contraire, on ne peut rien dire. En particulier K peut être infini alors que P et Q sont arbitrairement proches en distance de Hellinger (ou en distance $\mathbb{L}^p$). Et dans certains modèles où $K(P, Q)$ est infini pour tous les couples P et Q, il est impossible d'obtenir des minorations par cette méthode.

Ceci laisse à penser que l'information de Kullback n'est pas l'outil le plus adéquat pour ce genre de problème; fort heureusement nous verrons que K se comporte bien dans de nombreux cas et nous serons ainsi à même de traiter un certain nombre d'exemples classiques ou nouveaux.

Un autre inconvénient de l'utilisation de K et aussi de l'aspect local de ces minorations est que tout repose sur l'existence d'un point θ_0 de Θ au voisinage duquel on choisira l'ensemble S'. Il s'agira donc de vérifier qu'il existe un tel point ce qui pratiquement reviendra à l'exhiber. De ce fait, contrairement, à ce qui se produit pour les majorations, les conditions suffisantes pour démontrer l'existence de bornes inférieures du risque sont délicates à vérifier et il n'y a pas d'hypothèses générales simples analogues à H1 et H2 permettant de comparer majorations et minorations. On devra se contenter de vérifier à chaque fois un certain nombre de relations telles que (1.12) et (1.13) sur un certain voisinage d'un point θ_0 qu'il faut exhiber. Fort heureusement, les procédés de théorie de l'approximation qui permettent de majorer d fournissent en même temps des voisinages «naturels» sur lesquels on peut effectuer ces calculs de minoration.

Le chapitre qui suit sera consacré à la mise en place des résultats théoriques qui permettent d'obtenir des bornes pour le risque dans le cadre d'un modèle à une observation. On peut d'ailleurs toujours se ramener à ce cas en considérant $(X_1, \ldots, X_n)$ comme une observation n-dimensionnelle. En particulier nous donnerons la construction et les performances des d-estimateurs.

Dans le chapitre 3, nous retraduirons ces résultats dans le cadre d'une expérience avec n observations et nous définirons les conditions d'uniformité en n qui permettent de déterminer les vitesses d'estimation et d'obtenir des suites adéquates de d-estimateurs. Nous nous intéresserons tout spécialement au cas de l'estimateur du maximum de vraisemblance sur un réseau et à la manière dont on peut déduire les minorations de certains résultats de dimension concernant les «systèmes de petites perturbations», analogues à ceux utilisés par Bretagnolle-Huber [8].

Le chapitre 4 sera consacré à l'application de ce qui précède aux variables indépendantes, avec deux exemples très détaillés d'estimation de densités. Ces exemples sont en fait génériques et tous les autres cas que nous considérerons se traitent de manière tout à fait analogue.

Au chapitre 5, nous verrons comment étendre l'utilisation du maximum de vraisemblance sur des réseaux à certains cas de chaînes de Markov ou de processus gaussiens stationnaires. Pour cela nous serons amenés à ajouter des hypothèses $LB(\gamma_1, \gamma_2)$ destinées à éviter un comportement trop pathologique des processus. Sous ces hypothèses, nous montrerons dans chacun des deux cas, que les tests de rapports de vraisemblance au seuil 0 entre $P_{s,n}$ et $P_{t,n}$ ont des erreurs majorées par

$$A_1 \exp[-nA_2 d^2(s,t)] \quad \text{et que} \quad K(P_{s,n}, P_{t,n}) \leqq CnK(P_{s,1}, P_{t,1}).$$

C'est-à-dire qu'en gros tout se passe comme pour les variables indépendantes et que la vitesse de séparation entre $P_{s,n}$ et $P_{t,n}$ est bien celle qu'il faut. Dans ces conditions, la vitesse d'estimation ne dépend plus que de la structure métrique (la fonction $\tilde{d}$) de l'espace des paramètres. Ceci explique que pour les processus gaussiens on obtienne les mêmes vitesses que pour les variables indépendantes lorsque les densités spectrales et les densités varient dans le même espace. De même pour les processus de Markov, si $\Lambda = \{p_\theta(x,y)\}_{\theta \in \Theta}$, on obtient la même vitesse que lorsqu'on estime une densité $f_\theta(x,y)$ bi-dimensionnelle appartenant à Λ à partir de variables indépendantes. Ceci permet donc de répondre aux questions posées par Farrell [13] à propos de ce phénomène.

Dans tous les exemples considérés aux chapitres 4 et 5, variables indépendantes, chaînes de Markov, processus gaussiens, lorsque l'on peut construire les d-estimateurs, le risque pourra être approché (à des constantes multiplicatives près) par la fonction $L\tilde{d}$. On pourrait donc supposer que les seuls critères de dimension métrique suffisent à expliquer la vitesse d'estimation des problèmes statistiques et que les bornes qu'ils fournissent sont toujours adéquates. Il n'en est rien. Le chapitre 6 sera donc consacré à l'étude des mérites et des limites de cette théorie métrique. Nous nous placerons ici dans le cas de suites de n variables indépendantes équidistribuées et de la distance de Hellinger (mais les résultats sont les mêmes avec la distance $\mathbb{L}^2$). Nous construirons toute une famille de contre-exemples, pour lesquels

$$\sup_\theta \mathbb{E}_\theta[nh^2(\theta, \hat{\theta}_n)] < 1,$$

où $\hat{\theta}_n$ est l'estimateur du maximum de vraisemblance (et il en serait de même pour certains d-estimateurs), alors que la fonction $\tilde{h}(\varepsilon)$ est totalement arbitraire et que les informations de Kullback entre les points de Θ sont infinies. Ceci montre que ni les majorations du risque (même pour les d-estimateurs) obtenues par la théorie métrique, ni les minorations que donnent les informations de Kullback ne sont adéquates dans certains cas.

D'autre part, pour toutes les vitesses d'estimation «raisonnables», c'est-à-dire celles qui peuvent s'écrire $r(n)$ avec $r(n)$ décroissante et $n^{\frac{1}{2}} r(n)$ croissante (ceci en raison de 1.1) on peut construire un modèle d'estimation, en variables indépendantes toujours, pour lequel i) la vitesse de $R_n(d^q)$ est $r^q(n)$, ii) les d-estimateurs forment une suite adéquate, iii) les bornes supérieures et inférieures données par la théorie métrique sont adéquates. On peut donc en conclure que faute d'autres informations sur le modèle que celles du type métrique, l'utilisation des d-estimateurs semble justifiée et constitue le moins mauvais choix possible. Ceci d'autant plus que dans les cas classiques d'estimation de densités ils fournissent les bonnes vitesses, aussi bien que sur d'autres exemples qui, à notre connaissance, n'avaient pas été étudiés précédemment: entre autre l'estimation dans les classes Λ_ω considérées comme familles de densités de variables indépendantes, ou de densités spectrales d'un processus gaussien et dans certains types de chaînes de Markov.

En conclusion, nous évoquerons quelques problèmes posés par les limitations de cette méthode. Certaines difficultés sont liées à l'utilisation de la distance de Hellinger pour laquelle les calculs de dimension métrique ou d'entropie ne sont pas très aisés. On peut lui préférer les distances $\mathbb{L}^p$. Malheureusement, pour $p \leqq 2$, elles ne sont pas toujours comparables à h, et ne le sont pas pour $p > 2$. Il s'ensuit que l'on ne parvient plus à relier facilement les performances des tests de s contre t à la distance $d(s,t)$. Par contre nous verrons que les d-estimateurs présentent de très bonnes propriétés de robustesse ce qui rend plausible leur utilisation pratique lorsque aucun estimateur classique robuste n'est disponible.

Enfin le problème reste ouvert de trouver une quantité que ne serait ni h, ni K, et avec laquelle on puisse effectuer à la fois les calculs de majoration et de minoration. Toutefois les démonstrations concernant les minorations laissent à penser qu'il ne suffit pas d'utiliser les couples de mesures, mais plutôt des ensembles plus gros, ce qui risque de rendre les quantités en question très complexes et difficilement calculables et leur ôterait donc beaucoup d'intérêt. En l'état actuel des choses, l'utilisation des d-estimateurs semble convenir à de nombreux cas pour lesquels la théorie métrique rend correctement compte des vitesses d'estimation et donne le moyen de les calculer à partir de la fonction $L\tilde{d}$.

2. Etude théorique dans le cas d'une seule observation

La construction de nos estimateurs tout comme celle de Le Cam [24] et [26], est fondée sur une famille de tests. L'estimateur est défini par une procédure de minimax relatif à une certaine fonction de perte et contrairement à celui de Le Cam, le processus ne nécessite qu'une seule étape. En fait, la construction des d-estimateurs est un cas particulier d'une construction plus générale que nous allons décrire d'abord.

Nous supposons donnés un recouvrement de Θ par des ensembles B_s, s dans S, une famille de tests non randomisés $\{\varphi_{s,t}\}_{(s,t)\in S\times S}$ de B_s contre B_t avec $\varphi_{s,t} = 1 - \varphi_{t,s}$ et une fonction de perte ℓ sur S^2. Posons

$$L_s(X) = \sup_{t \neq s} \ell(s,t)\varphi_{s,t}(X), \qquad L(X) = \inf_{s\in S} L_s(X).$$

Nous supposerons que si θ est un point de B_s, $P_\theta[L_s(X) > n] \xrightarrow[n\to+\infty]{} 0$. Il s'ensuit que $L(X)$ est fini p.s. et l'on peut choisir comme estimateur $\hat{\theta}(X)$ un point arbitraire de B_s où s est tel que $L_s(X) \leqq L(X) + \varepsilon$, ε étant choisi arbitrairement petit, éventuellement nul lorsque cela est possible.

Lorsqu'il existe un η-réseau S, un recouvrement $\{B_s\}_{s\in S}$ de Θ centré sur S et une famille de tests $\varphi_{s,t}$ satisfaisant aux hypothèses H1 et H2, on fera la construction précédente en prenant $\ell(s,t) = d(s,t)$ et en choisissant pour $\hat{\theta}(X)$ non pas un point quelconque d'un B_s pour lequel $L_s(X) \leqq L(X) + \varepsilon$, mais un point t de S tel que $L_t(X) = L(X)$ ce qui est possible vu le

Lemme 2.1. *Les hypothèses* $\mathrm{H1}(\eta, k_0, \delta)$ *et* $\mathrm{H2}(k_0\eta, A_1, A_2)$ *étant satisfaites, il existe* $\hat{\theta}(X)$ *dans* S *tel que* $L_{\hat{\theta}(X)}(X) = L(X) < +\infty$ p.s. *De plus, si* s *est un point de* S *et* θ *est dans* B_s, *on a pour* $p \geqq k_0$

$$P_\theta[d(s,\hat{\theta}) \geqq 2^p\eta] \leqq A_1 \sum_{n=p}^{+\infty} \exp[-A_2 2^{2n}\eta^2] 2^{\delta(n+1)} < +\infty. \tag{2.1}$$

Preuve. Supposons que θ soit la vraie valeur du paramètre et que $L_s(X) \geqq 2^p\eta$. Alors il existe t dans S et $n \geqq p$ tels que $d(s,t) \geqq 2^n\eta$ et $\varphi_{s,t}(X) = 1$. Mais d'après (1.4), puisque θ est dans B_s, la probabilité d'un tel évènement est inférieure à $A_1 \exp[-A_2 2^{2n}\eta^2]$. Considérons alors les sphères de centre s et rayons $2^n\eta$, $n \geqq p$. Elles engendrent une suite de couronnes recouvrant le complémentaire de $\mathscr{B}(s, 2^p\eta)$. La couronne $\mathscr{B}(s, 2^{n+1}\eta) - \mathscr{B}(s, 2^n\eta)$ contient au plus $2^{(n+1)\delta}$ points de S, tous à distance de s supérieure à $2^n\eta$. Il s'ensuit aisément que l'on a:

$$P_\theta[L_s(X) \geqq 2^p\eta] \leqq A_1 \sum_{n=p}^{+\infty} \exp[-A_2 2^{2n}\eta^2] 2^{\delta(n+1)}.$$

Il est clair que cette série converge et donc $L(X)$ est fini p.s. Soient alors t tel que $L_t(X) < L(X) + \eta$ et u tel que $d(u,t) \geqq L(X) + \eta$. D'après la définition de L_t, $\varphi_{t,u}(X) = 1$ et $L_u(X) \geqq d(u,t) \geqq L(X) + \eta > L_t(X)$. Ceci montre que pour calculer $L(X) = \inf_{u \in S} L_u(X)$, il suffit de se limiter à calculer l'inf lorsque u appartient à $S \cap \mathscr{B}(t, L(X) + \eta)$. D'après (1.3), cet ensemble est fini et il s'ensuit que $L(X) = L_{\hat{\theta}(X)}(X)$ pour un certain élément $\hat{\theta}(X)$ de S.

Supposons donc que $\hat{\theta}(X) = t$ et $d(s,t) \geqq 2^p\eta$. Ou bien $\varphi_{s,t}(X) = 1$ et $L_s(X) \geqq 2^p\eta$, où bien $\varphi_{t,s}(X) = 1$ et $L_s(X) \geqq L_t(X) \geqq 2^p\eta$ puisque $L_t(X) = L(X)$. On aura donc

$$P_\theta[d(s, \hat{\theta}(X)) \geqq 2^p\eta] \leqq P_\theta[L_s(X) \geqq 2^p\eta].$$

Le lemme est ainsi démontré. □

Corollaire 2.2. *Supposons que les tests* $\varphi_{s,t}$ *sont des tests de rapport de vraisemblance au seuil* 0 *entre* P_s *et* P_t, *c'est-à-dire que* $\varphi_{s,t} = 0$ (*accepte* B_s) *si* $\log \frac{dP_s}{dP_t} > 0$, $\varphi_{s,t} = 1$ (*rejette* B_s) *si* $\log \frac{dP_s}{dP_t} < 0$ *et que ces tests vérifient* H2 *quelle que soit la valeur de* $\varphi_{s,t}$(0 *ou* 1) *sur l'ensemble* $\left\{\log \frac{dP_s}{dP_t} = 0\right\}$. *Alors le maximum de vraisemblance sur* S *est atteint en au moins un point et tout estimateur du maximum de vraisemblance* (*avec règle de décision non randomisée dans le cas d'un maximum multiple*) *sur* S *est un cas particulier de la construction précédente et vérifie* (2.1).

Preuve. Reprenons la démonstration du lemme précédent. Si L est tel que $L_t(X) = L(X)$ alors $\log \frac{dP_t}{dP_u}(X) \geqq 0$ dès que $d(t,u) > L(X)$. Si t' maximise la vraisemblance sur $S \cap \{\bar{\mathscr{B}}(t, L(X))\}$ alors $\log \frac{dP_{t'}}{dP_t}(X) \geqq 0$ et t' maximise la vraisemblance sur S. Soit $\tilde{\theta}(X)$ un maximum de vraisemblance égal à t' pour une certaine valeur de X. Il coïncide avec l'estimateur $\hat{\theta}(X)$ si pour $t' \neq t$ on a choisi $\varphi_{t',t}=0$ lorsque $\log\frac{dP_{t'}}{dP_t}=0$. Les diverses valeurs du maximum de vraisemblance lorsque celui-ci n'est pas unique correspondent en fait à divers choix des $\varphi_{t',t}$ sur les ensembles $\left\{\log\frac{dP_t}{dP_{t'}}=0\right\}$. Ils vérifient donc certainement (2.1). □

Définition 2.3. Tout estimateur construit selon la méthode précédente (en particulier les estimateurs du maximum de vraisemblance sur un réseau) seront dits d-estimateurs.

Théorème 2.4. *Supposons* $\mathrm{H1}(\eta, k_0, \delta)$ *et* $\mathrm{H2}(k_0\eta, A_1, A_2)$ *satisfaites. Si* ℓ *est une fonction de perte telle que* $\ell(2x)-\ell(x)\leqq e^{\alpha x^2}$, $0<\alpha<A_2$, *il existe une constante* $C_1(\eta, \delta, k_0, A_1, A_2, \ell)$ *telle que pour tout d-estimateur* $\hat{\theta}$ *on ait*

$$\sup_{s\in S}\{\sup_{\theta\in B_s}\mathbb{E}_\theta[\ell(d(s,\hat{\theta}))]\}\leqq C_1<+\infty. \tag{2.2}$$

Si de plus $\{B_s\}_{s\in S}$ *est un* (η, d)*-recouvrement on a*

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta[\ell(d(\theta,\hat{\theta}))]\leqq C_1. \tag{2.3}$$

Preuve. Nous utiliserons (2.1) et l'inégalité suivante qui est vraie pour toute fonction croissante ℓ et toute suite croissante $\{x_n\}$ tendant vers $+\infty$:

$$\mathbb{E}_\theta[\ell(X)]\leqq\ell(x_p)+\sum_{n=p}^{+\infty}P_\theta[X\geqq x_n][\ell(x_{n+1})-\ell(x_n)].$$

Si l'on pose $x_p=(2^p+1)\eta$ on a d'après (2.1) si $B_s\subset\bar{\mathscr{B}}(s,\eta)$:

$$P_\theta[d(\theta,\hat{\theta})\geqq x_p]\leqq A_1\sum_{n=p}^{+\infty}\exp[-A_2 2^{2n}\eta^2+\delta(n+1)\log 2]$$

pour tout θ dans Θ et $2^p\geqq k_0$.

Posons $a_n=-2^{2n}A_2\eta^2+\delta(n+1)\log 2$. La suite $\{a_n\}$ tend vers $-\infty$ lorsque n croît. Si l'on a $a_p\leqq -0{,}96$ et $p\geqq 2$, alors il est immédiat que l'on obtient $a_{n+1}<4a_n$ pour $n\geqq p$ et donc $\exp(a_{n+1})\leqq\exp(a_n)\exp(-2{,}88)$. Finalement

$$\sum_{n=p}^{+\infty}\exp(a_n)\leqq\exp(a_p)\sum_{n=0}^{+\infty}\exp(-2{,}88n)=\exp(a_p)\frac{1}{1-e^{-2{,}88}}<1{,}06\exp(a_p).$$

Donc si p est le plus petit entier tel que $2^p\geqq k_0$, $p\geqq 2$, $a_p\leqq -0{,}96$, on obtient

$$\mathbb{E}_\theta[\ell(d(\theta,\hat{\theta}))]\leqq\ell(x_p)+A_1\times 1{,}06\sum_{n=p}^{+\infty}\exp(a_n)[\ell(x_{n+1})-\ell(x_n)]. \tag{2.4}$$

Mais $\frac{1}{2}x_{n+1}\leqq x_n$, donc

$$\ell(x_{n+1})-\ell(x_n)\leqq\ell(x_{n+1})-\ell\left(\frac{x_{n+1}}{2}\right)\leqq\exp\left[\alpha\frac{(x_{n+1})^2}{4}\right],$$

et finalement

$$\begin{aligned}&\exp(a_n)[\ell(x_{n+1})-\ell(x_n)]\\&\leqq \exp[-2^{2n}\eta^2[A_2-\alpha(1+2^{-n-1})^2-\delta\eta^{-2}2^{-2n}(n+1)\log 2]].\end{aligned}$$

Comme $\alpha<A_2$, la série au second membre de (2.4) converge et (2.3) s'ensuit. (2.2) se démontre d'une manière analogue avec $x_p=2^p\eta$. □

Désormais nous supposerons que B_s est un η-recouvrement centré sur S.

Corollaire 2.5. *Sous les hypothèses* H1 *et* H2 *avec* $\eta^2\geqq\delta\geqq\frac{1}{2}$, $k_0\geqq 4$, *si* $\ell(2x)\leqq 2^q\ell(x)$ *on peut écrire pour tout d-estimateur* $\hat\theta$, *et* $q>0$,

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta[\ell(\theta,\hat\theta)]\leqq\ell(\eta)C_2(k_0,q,A_1,A_2) \tag{2.5}$$

Preuve. Il est facile de voir que l'on a $a_n\leqq-\eta^2[2^{2n}A_2-(n+1)\log 2]$ et avec $\eta^2\geqq\frac{1}{2}$, $a_n\leqq-\frac{1}{2}[2^{2n}A_2-(n+1)\log 2]$. Donc si p est le plus petit entier tel que $2^p\geqq k_0$ et $\frac{1}{2}[2^{2p}A_2-(p+1)\log 2]\geqq 0{,}96$, on peut appliquer (2.4), p ne dépendant que de k_0 et A_2. De plus $\ell(x_n)<\ell(\eta)2^{(n+1)q}$. On obtient donc

$$\mathbb{E}_\theta[d^q(\theta,\hat\theta)]\leqq\ell(\eta)2^{(p+1)q}+(1{,}06)A_1\ell(\eta)\sum_{n=p}^{+\infty}\exp[a_n+q(n+2)\log 2],$$

ce qu'on peut clairement majorer (très brutalement) par

$$\ell(\eta)\left\{2^{(p+1)q}+1{,}06A_1\sum_{n=p}^{+\infty}\exp\left[-2^{2n-1}A_2+\log 2\left(\frac{n+1}{2}+(n+2)q\right)\right]\right\}.$$

D'où le résultat. □

L'inégalité (2.5) est générale. Dans les applications, il conviendra de faire un calcul plus précis adapté à chaque cas selon les valeurs des diverses constantes et de ℓ pour obtenir de meilleures majorations. En particulier il est facile de voir grâce à (1.11) ou à [6] dans le cas de n variables indépendantes équidistribuées et de la distance $d(s,t)=n^{\frac{1}{2}}h(P_s,P_t)$ que l'on pourra obtenir une amélioration de (1.4) sous la forme

$$\sup_{s'\in B_s}\mathbb{E}_{s'}[\varphi_{s,t}]\leqq\exp[-nh^2(B_s,B_t)]. \tag{2.6}$$

Si l'on suppose que $\tilde h(\varepsilon)<n\varepsilon^2$, on obtiendra $\tilde d(\eta)<\eta^2$ avec $\eta=n^{\frac{1}{2}}\varepsilon$ et H1 pourra être satisfaite pour un réseau bien choisi. Une démonstration analogue à la précédente utilisant une version modifiée de H2 grâce à (2.6) conduit par des calculs un peu plus précis au

Corollaire 2.6. *Si l'on dispose de n variables indépendantes équidistribuées et si* $n\varepsilon^2>\tilde h(\varepsilon)\geqq 1$, *il existe un* $(n^{\frac{1}{2}}h)$*-estimateur* $\hat\theta_n$ *qui vérifie:*

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta[h^2(\theta,\hat\theta_n)]\leqq 25\varepsilon^2+57\varepsilon^2\exp(-1{,}9n\varepsilon^2). \tag{2.7}$$

Mais ces résultats ne prétendent en aucun cas fournir des majorations optimales, seulement des bornes grossières pouvant donner un ordre de grandeur du risque vu que la méthode de majoration qui conduit à (2.1) est elle-même très grossière.

Nous allons maintenant nous intéresser à la recherche de bornes inférieures pour le risque minimax; ces bornes reposent sur le lemme fondamental suivant, très légère variante du lemme VII.1.1. de Ibragimov et Khas'minskii [16]:

Lemme 2.7. *Supposons que sur l'espace $\{\mathscr{X}, \mathscr{A}\}$ soient données n mesures de probabilité $P_1, \ldots, P_n$ et une fonction mesurable $\psi: \mathscr{X} \to \{1; \ldots; n\}$, $n \geqq 2$. Alors on a*

$$\sup_{1 \leqq i \leqq n} P_i\{\psi(x) \neq i\} \geqq 1 - \frac{n^{-2} \sum_{i,j} K(P_i, P_j) + \log 2}{\log(n-1)}. \tag{2.8}$$

Preuve. Définissons sur l'espace produit $\mathscr{X}' = \mathscr{X} \times \{1; \ldots; n\}$ la probabilité P par $dP(x, i) = n^{-1} dP_i(x)$ pour tout i, $1 \leqq i \leqq n$. Appelons p la projection $(x, i) \mapsto i$ sur le second facteur de $\mathscr{X}'$. Toutes les espérances sont prises par rapport à P. Nous avons (cf. [16] p. 323-325):

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}(p=i \mid x) \log \mathbb{E}(p=i \mid x) \\
&\quad = \mathbb{E}(p=\psi(x) \mid x) \log \mathbb{E}(p=\psi(x) \mid x) + \mathbb{E}(p \neq \psi(x) \mid x) \log \mathbb{E}(p \neq \psi(x) \mid x) \\
&\qquad + \mathbb{E}(p \neq \psi(x) \mid x) \sum_{i \neq \psi(x)} \frac{\mathbb{E}(p=i \mid x)}{\mathbb{E}(p \neq \psi(x) \mid x)} \log \frac{\mathbb{E}(p=i \mid x)}{\mathbb{E}(p \neq \psi(x) \mid x)} \\
&\quad \geqq -\log 2 - \mathbb{E}(p \neq \psi(x) \mid x) \log(n-1).
\end{aligned}$$

D'autre part $\mathbb{E}[p=i \mid x] = \dfrac{dP(x,i)}{\sum_{j=1}^{n} dP(x,j)}$ et donc

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}[p=i \mid x] \log \mathbb{E}[p=i \mid x]\right] &= n^{-1} \sum_{i=1}^{n} \int \log \frac{dP_i(x)}{\sum_{j=1}^{n} dP_j(x)} dP_i(x) \\
&\leqq n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int \log \frac{dP_i(x)}{dP_j(x)} dP_i(x) - \log n = n^{-2} \sum_{i,j} K(P_i, P_j) - \log n
\end{aligned}$$

en utilisant la convexité. En réunissant les deux inégalités, on obtient finalement

$$n^{-2} \sum_{i,j} K(P_i, P_j) - \log n \geqq -\log 2 - \log(n-1) P(p \neq \psi(x)).$$

Or

$$P[p \neq \psi(x)] = n^{-1} \sum_{i=1}^{n} P_i[\psi(x) \neq i] \leqq \sup_{1 \leqq i \leqq n} P_i[\psi(x) \neq i].$$

La conclusion s'ensuit immédiatement. □

La proposition suivante s'en déduit alors aisément:

Proposition 2.8. *Si* (Θ, d) *vérifie* $H3(\eta', \gamma)$ *avec* $\gamma < 1$, *pour tout estimateur* T, *on a*

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta[\ell(d(\theta, T))] \geqq (1-\gamma)\ell\left(\frac{\eta'}{2}\right). \tag{2.9}$$

Preuve. Choisissons θ dans S' et remplaçons l'estimateur T à valeur dans Θ par un estimateur T^* à valeur dans S' de façon que $d(T, T^*) = d(T, S')$. Alors si $T^* \neq \theta$, $d(T, \theta) \geqq \frac{\eta'}{2}$, à cause de (1.12) et il s'ensuit que pour tout θ dans S'

$$\mathbb{E}_\theta[\ell(d(\theta, T))] \geqq \ell\left(\frac{\eta'}{2}\right) P_\theta[T^* \neq \theta].$$

Il suffit alors d'appliquer le lemme 2.7 et d'utiliser (1.13) pour obtenir le résultat. □

Dans le cas où $\ell(x) = x^q$ qui nous intéresse tout particulièrement, on peut relier cette conclusion à celle du corollaire 2.5. On obtient ainsi le corollaire suivant qui constitue le résultat fondamental pour l'approximation du risque. Sa démonstration est immédiate à partir du corollaire 2.5 et de la proposition 2.8.

Corollaire 2.9. *Supposons que soient vérifiés* $H1(\eta, k_0, \delta)$, $H2(k_0\eta, A_1, A_2)$ *et* $H3(c\eta, \gamma)$ *avec* $\eta^2 \geqq \delta \geqq \frac{1}{2}$, $k_0 \geqq 4$, $\gamma < 1$, *alors si* $\hat{\theta}$ *est un d-estimateur, on peut écrire:*

$$(1-\gamma)\left(\frac{c}{2}\right)^q \eta^q \leqq R(d^q) \leqq \sup_{\theta \in \Theta} \mathbb{E}_\theta[d^q(\theta, \hat{\theta})] \leqq C_2(k_0, q, A_1, A_2)\eta^q \tag{2.10}$$

où C_2 *est défini comme en* (2.5) *et donc indépendant de* η, δ.

Remarque. Il est nécessaire, pour cette démonstration, de se restreindre à $\ell(x) = x^q$ ou plus généralement à des fonctions telles que $\ell(2x) \leqq C\ell(x)$, c'est-à-dire à croissance polynomiale. En effet les majorations sont fonction de $\ell(\eta)$ et les minorations de $\ell\left(\frac{c\eta}{2}\right)$. Lorsque ℓ est à croissance trop rapide, on ne sait plus comparer ces deux nombres indépendamment de η, le rapport entre majorations et minorations dépend de η et n'est plus borné lorsque η tend vers l'infini.

3. Suites d'observations – Liens entre vitesse et dimension

Nous venons de voir comment les hypothèses H1, H2 et H3 nous permettent d'encadrer le risque. Si nous avons une suite d'expériences $\mathscr{E}_n = (\mathscr{X}^n, \mathscr{A}^n, E_n, d_n, \Theta_n)$, nous pouvons appliquer le corollaire 2.9 aux espaces (Θ_n, d_n) s'ils vérifient ces hypothèses et obtenir ainsi des encadrements successifs des $R_n(d_n^q)$. Mais pour que l'on puisse définir une vitesse, il faudra que les constantes γ, c, C_2 qui interviennent dans (2.10) soient indépendantes de n, ce qui suppose dans les hypothèses H1, H2, H3 une certaine uniformité en n. D'autre part, il serait bien agréable de pouvoir définir l'η qui intervient dans (2.10) de façon essentielle, en fonction uniquement de n et des propriétés de d. C'est dans ce but que nous avons, dans l'introduction, défini la fonction $\tilde{d}$ ainsi que la transformation L. Nous allons donc introduire de nouvelles hypothèses.

Supposons que δ est une fonction de $\mathbb{R}^+$ dans $\mathbb{R}^+ \cup \{+\infty\}$, nous poserons les hypothèses suivantes relatives à (Θ, d):

H4$(\delta, \varepsilon_0, n_0, k_0, A_1, A_2)$: Pour tout $\varepsilon \leqq \varepsilon_0$, il existe un ε-réseau S et un ε-recouvrement de Θ centré sur S: $\{B_s\}_{s\in S}$ tels que

i) S satisfait à H1$(\varepsilon, k_0, \delta(\varepsilon))$ si $\delta(\varepsilon) < +\infty$

ii) si $d(s,t) \geqq k_0 \varepsilon$ et $n \geqq n_0$, il existe un test non randomisé de B_s contre B_t fondé sur n observations dont chacune des erreurs est majorée par $A_1 \exp[-nA_2 d^2(s,t)]$.

Dans le cas de n observations indépendantes équidistribuées, si $d=h$, d'après (1.11) ii) sera vérifié sans peine en prenant $B_s = \mathscr{B}(s, \varepsilon)$. Il suffira donc de choisir $\delta(\varepsilon)$ de façon adéquate, et toute fonction strictement supérieure à $\tilde{d}$ conviendra. Il en est de même si $d \leqq Mh$.

L'utilisation de ce nouveau jeu d'hypothèses permet d'obtenir immédiatement une version uniforme en n du corollaire 2.5.

Proposition 3.1. *Supposons que les hypothèses* H4 *soient vérifiées et que l'on ait* $k_0 \geqq 4$, $\varepsilon \leqq \varepsilon_0$, $n \geqq n_0$, $n\varepsilon^2 \geqq \delta(\varepsilon) \geqq \frac{1}{2}$. *Alors si* $\hat{\theta}_n$ *est un* d_n*-estimateur fonction de* n *observations et construit à partir des éléments donnés par* H4, *il vérifie:*

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta[d^q(\theta, \hat{\theta}_n)] \leqq \varepsilon^q C_2[k_0, q, A_1, A_2]. \tag{3.1}$$

Preuve. Si l'on considère les éléments donnés par H4 et qu'on les transporte sur (Θ_n, d_n), S devient un η-réseau avec $\eta = n^{\frac{1}{2}}\varepsilon$ et vérifie H1$(\eta, k_0, \delta(\varepsilon))$. De plus, en utilisant ii), on voit que H2$(k_0\eta, A_1, A_2)$ est vérifiée par les B_s qui forment un η-recouvrement. On a de plus $\eta^2 \geqq \delta(\varepsilon) \geqq \frac{1}{2}$, ce qui permet d'appliquer (2.5) qui devient

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta[d_n^q(\theta, \hat{\theta})] \leqq \eta^q C_2.$$

On obtient (3.1) en simplifiant par $n^{\frac{q}{2}}$. □

Chaque fois que cela ne prête pas à confusion, nous parlerons encore de d-estimateur pour $\hat{\theta}_n$ (plutôt que de d_n-estimateur).

Remarque. Lorsque $\tilde{d}(\varepsilon)$ est bornée on pourra choisir la foction $\delta(\varepsilon)$ constante égale à D. On prendra alors $\eta^2 = D$ soit $\varepsilon^2 = n^{-1}D$. Ceci nous donnera finalement si $q=2$

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta[nd^2(\theta, \hat{\theta}_n)] \leqq C_2 D.$$

Nous allons de même modifier l'hypothèse H3 pour l'adapter au cas de n observations. Considérons une fonction J, positive, décroissante et posons

H5(J, n_0, γ): Pour tout $n \geqq n_0$, il existe un sous-ensemble fini S'_n de Θ de cardinal $m+1$, tel que l'on ait

$$d(s,t) \geqq J(n) \quad \text{pour } s \text{ et } t \text{ dans } S'_n,\ s \neq t \tag{3.2}$$

$$\sup_{s,t \in S'_n} K(P_{s,n}, P_{t,n}) \leqq \gamma \log m - \log 2 \tag{3.3}$$

On peut en déduire immédiatement une version uniforme de la proposition 2.8:

Corollaire 3.2. *Supposons que* (Θ, d) *vérifie* $\mathrm{H5}(J, n_0, \gamma)$ *avec* $\gamma < 1$, *alors on a pour* $n \geqq n_0$

$$R_n(d^q) \geqq (1-\gamma) 2^{-q} (J(n))^q.$$

Pour démontrer le théorème qui donne la vitesse d'estimation nous aurons besoin de quelques propriétés de la transformation L décrite dans l'introduction et dont nous rappellerons qu'elle est définie par $L\delta(n) = \inf\{x \mid nx^2 > \delta(x)\}$. Nous pouvons en déduire:

Proposition 3.3. *La transformation* L *est croissante et la fonction* $L\delta$ *décroissante. De plus, si* δ *est décroissante, on a*

$$\delta[\beta L\delta(n)] < n\beta^2 [L\delta(n)]^2 \quad \textit{si } \beta > 1 \tag{3.4}$$

$$(n+1)^{\frac{1}{2}} L\delta(n+1) \geqq n^{\frac{1}{2}} L\delta(n) \geqq L\delta(1) \tag{3.5}$$

Preuve. Ces résultats sont des conséquences immédiates de la définition, sauf (3.5). Choisissons x tel que $\left(\frac{n+1}{n}\right)^{\frac{1}{2}} x < L\delta(n)$. Alors $(n+1)x^2 \leqq \delta\left[\left(\frac{n+1}{n}\right)^{\frac{1}{2}} x\right] \leqq \delta(x)$ et donc $L\delta(n+1) \geqq x$. D'où le résultat. □

Dans ce qui suit, nous utiliserons toujours une fonction δ décroissante. Ce n'est pas indispensable, mais permet d'éviter d'inutiles complications dans les démonstrations, d'autant plus que dans les applications, nous aurons toujours affaire à de telles fonctions.

Théorème 3.4. *Soient* δ *une fonction décroissante, minorée par* $\frac{1}{2}$, $k_0 \geqq 4$, n_0 et $\varepsilon_0 > L\delta(n_0)$. *Supposons que* $\mathrm{H4}(\delta, \varepsilon_0, n_0, k_0, A_1, A_2)$ *soit vérifiée ainsi que* $\mathrm{H5}(cL\delta, n_0, \gamma)$ *avec* γ *et* $c < 1$. *Alors, pour tout* $q > 0$, *il existe une suite adéquate* $\{\hat{\theta}_n\}$ *de* d*-estimateurs de* θ *et l'on a*

$$R_n(d^q) \asymp [L\delta(n)]^q \quad \textit{pour } n \geqq n_0.$$

Preuve. D'après (3.4), si $\beta > 1$, on aura $n\beta^2 [L\delta(n)]^2 > \delta[\beta L\delta(n)] \geqq \frac{1}{2}$. Choisissons $\beta < 2$ (éventuellement très proche de 1). Posons $\varepsilon_n = \beta L\delta(n)$. Comme $\varepsilon_0 > L\delta(n_0)$, on choisit β de façon que $\varepsilon_n \leqq \varepsilon_0$. On peut alors appliquer la proposition 3.1 avec ε_n à la place de ε et obtenir un d-estimateur $\hat{\theta}_n$ qui vérifie pour $n \geqq n_0$

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta [d^q(\theta, \hat{\theta}_n)] \leqq \beta^q C_2(k_0, q, A_1, A_2) [L\delta(n)]^q.$$

La première moitié du théorème est démontrée. Pour ce qui est de la seconde, on applique le corollaire 3.2 qui implique

$$R_n(d^q) \geqq (1-\gamma) \left(\frac{c}{2}\right)^q [L\delta(n)]^q, \quad \text{si } n \geqq n_0.$$

D'où la conclusion. $\square$

Nous obtenons aussi la proposition essentielle suivante qui montre qu'en général la vitesse obtenue est celle déduite de la dimension $\tilde{d}$ et que si une fonction δ satisfait aux hypothèses du théorème, elle ressemble à la fonction $\tilde{d}$ en ce sens que leurs transformées par L sont équivalentes. Pour le démontrer, nous utiliserons une propriété de $\tilde{d}$ qui est une conséquence immédiate de (1.5):

$$\tilde{d}(\varepsilon') \leqq \tfrac{4}{3}\tilde{d}(\varepsilon) \qquad \text{si } 2\varepsilon \geqq \varepsilon' > \varepsilon. \tag{3.6}$$

Proposition 3.5. *Supposons que soient vérifiées les hypothèses du théorème* (3.4) *et qu'il existe* $\beta > 0$, *indépendant de* n, *tel que*

$$\beta n d^2(s,t) \leqq K(P_{s,n}, P_{t,n}) \tag{3.7}$$

pour tous s, t *dans l'ensemble* S'_n *qui satisfait à* H5, *alors pour* $n \geqq n_0$, *il existe une constante* $\beta' < 1$ *telle que*

$$\beta' L\delta(n) \leqq L\tilde{d}(n) \leqq L\delta(n).$$

Preuve. L'hypothèse H4 implique que pour $\varepsilon \leqq \varepsilon_0$, $\tilde{d}(\varepsilon) \leqq \delta(\varepsilon)$ et donc $L\tilde{d} \leqq L\delta$. D'autre part, H5 entraîne l'existence d'un ensemble S'_n de cardinal $m+1$ dont tous les points sont à des distances mutuelles supérieures ou égales à $\eta = cL\delta(n)$. Supposons que S'_n soit inclus dans une boule de rayon $(2^j - 1)\frac{\eta}{3}$ et considérons un $\frac{\eta}{3}$-réseau S. Tout point de S'_n est à distance inférieure à $\frac{\eta}{3}$ d'un point de S et un point de S est associé de cette manière à au plus un point de S'_n. Donc le nombre de points de S dans une boule de rayon $2^j\frac{\eta}{3}$ est supérieur à $m+1$. Il s'ensuit que l'on a:

$$\tilde{d}\left(\frac{\eta}{3}\right) \geqq \frac{\log(m+1)}{j\log 2}, \qquad \text{si } j \geqq 3. \tag{3.8}$$

L'inégalité (3.7) implique d'autre part

$$\eta^2 \leqq \sup_{s,t \in S'_n} d^2(s,t) \leqq (\gamma \log m - \log 2) n^{-1} \beta^{-1},$$

ce qui fait que l'on peut choisir $j \geqq 3$ de façon que

$$n\beta(2^{j-1} - 1)^2 \frac{\eta^2}{9} \leqq \gamma \log m - \log 2 \leqq n\beta \frac{\eta^2}{9}(2^j - 1)^2. \tag{3.9}$$

D'autre part

$$\log(m+1) \geqq \gamma \log m - \log 2 \qquad \text{si } \gamma < 1. \tag{3.10}$$

En réunissant (3.8), (3.9) et (3.10) on obtient:

$$\tilde{d}\left(\frac{\eta}{3}\right) \geqq n\eta^2 \frac{\beta}{9} \frac{(2^{j-1} - 1)^2}{j\log 2} \geqq \beta_1 n\eta^2 \tag{3.11}$$

pour une certaine constante $\beta_1 = \frac{\beta}{3\log 2}$, puisque $j \geqq 3$. De plus, on a par définition $n\eta'^2 > \delta(\eta')$ dès que $\eta' > L\delta(n)$ et donc si $p > c^{-1}$, $np^2\eta^2 > \delta(p\eta)$, ce qui grâce à (3.11) nous donne avec une constante $\beta_2 = p^{-2}\beta_1$

$$\tilde{d}\left(\frac{\eta}{3}\right) \geqq \beta_2 \delta(p\eta),$$

ou si l'on pose $x_n = \frac{c}{3} L\delta(n)$

$$\tilde{d}(x_n) \geqq \beta_2 \delta(3px_n), \qquad n \geqq n_0.$$

Alors si $\frac{x_n}{2} < x \leqq x_n$, il vient en utilisant (3.6) et la décroissance de δ

$$\tilde{d}(x) \geqq \tfrac{3}{4}\beta_2 \delta(3px_n) \geqq \tfrac{3}{4}\beta_2 \delta(6px),$$

et finalement en posant $\beta_3 = \sup(\frac{4}{3}\beta_2^{-1}, 6p) > 6c^{-1}$,

$$\beta_3^2 \tilde{d}(x) \geqq \delta(\beta_3 x). \tag{3.12}$$

Comme d'après (3.5), $\frac{x_n}{2} < x_{n+1}$, l'inégalité (3.12) sera vérifiée pour tous les x de l'intervalle $\left]\frac{c}{6} L\delta(+\infty), \frac{c}{3} L\delta(n_0)\right]$ en convenant que $L\delta(+\infty) = \lim_{n \to +\infty} L\delta(n)$. Si $L\delta(+\infty) > 0$, $\delta(x) = +\infty$ pour $x < L\delta(+\infty)$. Or (3.11) s'écrit

$$\tilde{d}(x_n) \geqq 9\beta_1 n x_n^2.$$

Comme x_n tend vers $\frac{c}{3} L\delta(+\infty) > 0$ lorsque n tend vers l'infini, il s'ensuit en utilisant (3.6) que $\tilde{d}(x) = +\infty$ pour $x < \frac{c}{3} L\delta(+\infty)$, ce qui entraîne que (3.12) est encore vérifié pour $x \leqq \frac{c}{6} L\delta(+\infty)$. Donc (3.12) est satisfaite dans tous les cas sur $\left]0, \frac{c}{3} L\delta(n_0)\right]$. Alors, soit $n \geqq n_0$ et $x < \beta_3^{-1} L\delta(n)$; il s'ensuit que $n\beta_3^2 x^2 \leqq \delta(\beta_3 x)$ et $x < \frac{c}{6} L\delta(n_0)$. Donc $nx^2 \leqq \tilde{d}(x)$ et $x \leqq L\tilde{d}(n)$. Finalement $L\tilde{d}(n) \geqq \beta_3^{-1} L\delta(n)$, ce qui est la conclusion cherchée. □

Remarque 1. On voit qu'on a en fait des inégalités entre $\tilde{d}$ et δ elles-mêmes, et non seulement entre leurs transformées. Ceci permettrait, dans les applications, d'obtenir un véritable encadrement de $\tilde{d}$, mais en fait cette fonction n'a pas grand intérêt en elle-même puisque l'on dispose par ailleurs d'autres quantités telles que l'entropie pour mesurer la taille des ensembles.

Remarque 2. L'hypothèse (3.7), qui peut sembler peu naturelle, est en fait absolument immédiate lorsque l'on travaille sur des variables indépendantes équidistribuées et avec la distance de Hellinger, comme nous le verrons au chapitre suivant. C'est là ce qui donne tout son intérêt à ce résultat.

Ainsi que nous l'avons dit, dans le cas des variables indépendantes équidistribuées, l'hypothèse H4 se trouve vérifiée au moins pour $d=h$, pourvu que l'on choisisse $\delta>\tilde{d}$. Pour ce qui est de H5, on verra que l'on dispose généralement sur S' de l'inverse de (3.7) c'est-à-dire d'une inégalité du type $K(P_{s,n}, P_{t,n}) \leqq \beta n d^2(s,t)$ ce qui permet de majorer le premier membre de (3.3) connaissant les valeurs de $d(s,t)$ pour s et t dans S'. Le problème le plus délicat sera celui de l'évaluation de m et pour ce faire nous aurons très souvent recours à un résultat relatif aux «systèmes de petites perturbations», méthode très utilisée déjà, notamment par Farrell [12, 13], Bretagnolle-Huber [8] et Ibragimov-Khas'minkii [17] pour obtenir des minorations de la vitesse d'estimation. Nous aurons ici besoin des définitions suivantes:

Définition 3.6. Soit d une distance sur un sous-ensemble $\mathscr{L}$ d'un certain espace $\mathbb{L}^1(\mu)$, μ étant une mesure sur $(\mathscr{X}, \mathscr{A})$. Nous dirons que $\Phi(d)$ est suradditive (Φ étant une fonction de $\mathbb{R}^+$ dans $\mathbb{R}$), si pour toute partition finie $\{A_i\}_{1\leqq i\leqq p}$ de l'espace $\mathscr{X}$, on a pour f et g dans $\mathscr{L}$.

$$\Phi[d(f,g)] = \sum_{i=1}^{p} \Phi[d(f 1_{A_i}, g 1_{A_i})]. \tag{3.13}$$

Définition 3.7. Dans l'espace métrique (Θ, d), un ensemble S' est dit η-discernable si l'on a

$$d(s,t)>\eta \qquad s,t\in S', \ s\neq t.$$

Nous remarquerons que la propriété (3.13) introduite par Bretagnolle et Huber dans [8] implique que $f 1_{A_i}$ est un élément de $\mathscr{L}$. Dans la suite, nous n'utiliserons que des fonctions $\Phi(x)=x^r$, $r\geqq 1$. En particulier $\|f-g\|_r^r$ est suradditive sur $\mathbb{L}^r(\mu)$ et h^2 l'est également sur la partie positive de $\mathbb{L}^1(\mu)$, ce qui fait que nous pourrons leur appliquer la proposition suivante:

Proposition 3.8. *Supposons que soient donnés une partition $\{A_i\}_{1\leqq i\leqq p}$ de $\mathscr{X}$ et des éléments f, g_i, g_i' de $\mathbb{L}^1(\mu)$ tels que g_i, g_i' aient leur support dans A_i et que l'ensemble $\Theta'=\left\{f+\sum_{i=1}^{p}\lambda_i \mid \lambda_i=g_i \text{ ou } g_i'\right\}$ soit inclus dans Θ. Si l'on a pour tout i $d(f+g_i, f+g_i')\geqq\eta$, et si d^r est suradditive, pour un $r\geqq 1$, alors il existe un sous-ensemble S' de Θ', η'-discernable avec $\eta'=\left(\frac{p}{8}\right)^{1/r}\eta$ et dont le cardinal satisfait à*

$$\log[\operatorname{Card} S'-1] > \begin{cases} 0{,}316\,p, & \textit{pour } p\geqq 8 \\ 0{,}648\,p, & \textit{si } 3\leqq p\leqq 7. \end{cases} \tag{3.14}$$

Preuve. Considérons l'ensemble $J=\{-1;1\}^p$ et munissons-le de la distance d' définie par

$$d'(x,y)=\tfrac{1}{2}\sum_{i=1}^{p}|x_i-y_i| \tag{3.15}$$

où x_i est la i-ème coordonnée de x; choisissons un sous-ensemble J' de J, de cardinal $m+1$, qui soit $(k+\alpha)$-discernable maximal, avec k entier, $0\leqq\alpha<1$ et $k+\alpha=\frac{p}{8}$. Il existe une bijection π de J dans Θ' consistant à choisir $\lambda_i=g_i$ si x_i

$=1$ et g_i' si $x_i=-1$. En utilisant les hypothèses et la suradditivité de d^r, il est clair que par cette bijection on obtient

$$d^r(\pi(x), \pi(y)) \geqq \eta^r d'(x, y).$$

Il s'ensuit que si $S'=\pi(J')$, S' est η'-discernable avec $\eta'=(k+\alpha)^{\frac{1}{r}}\eta$ et que Card S' $-1=m$. Reste à vérifier (3.14). Si $3 \leqq p \leqq 7$, $k=0$ et $m+1=2^p$, c'est immédiat. Sinon, comme J' est maximal, c'est un $(k+\alpha)$-réseau et donc un k-réseau de J. Or une boule de rayon k dans J contient $\sum_{i=0}^{k} C_p^i$ points de J; comme les boules de rayon k centrées sur J' recouvrent J, on aura

$$m+1 \geqq 2^p \left[\sum_{i=0}^{k} C_p^i\right]^{-1}.$$

On peut majorer $2^{-p} \sum_{i=0}^{k} C_p^i$ en utilisant des bornes exponentielles pour les grandes déviations de la loi binomiale (cf. [29]) ou plus simplement la formule de Stirling: $n!=(2\pi n)^{1/2} n^n e^{-n} \exp[\varphi(n)]$; $\frac{1}{12n+1}<\varphi(n)<\frac{1}{12n}$. Pour $i \leqq \frac{p}{8}$ $\cdot\, C_p^{i-1}<\frac{1}{7} C_p^i$ donc $\sum_{i=0}^{k} C_p^i \leqq \frac{7}{6} C_p^k$ de sorte que

$$m+1 \geqq \tfrac{6}{7} 2^p \frac{k!(p-k)!}{p!} \geqq 2^p \frac{3}{4\sqrt{7}} (2\pi p)^{1/2} (7^{7/8}/8)^p$$

en utilisant la formule de Stirling. Finalement on vérifie que pour $p \geqq 8$

$$\log m > p[\tfrac{7}{8} \log 7-\log 4]>0{,}316 p. \quad \square$$

Remarque. En pratique, nous utiliserons cette proposition lorsque les A_i sont des intervalles égaux, les g_i des translatées d'une même fonction et $g_i'=-g_i$. Ce qui fait que $d(f+g_i, f-g_i)$ sera indépendant de i et égal à η. Dans ce cas, il est clair que Θ' et S' sont inclus dans une boule de rayon $p^{\frac{1}{r}}\eta$, ce qui permet d'obtenir une minoration de $\tilde{d}(\eta')$ puisque la démonstration précédente donne une minoration du nombre de points d'un η'-réseau dans une boule de rayon $8^{\frac{1}{r}}\eta'$. Comme $r \geqq 1$, il vient

$$\tilde{d}\left(\left(\frac{p}{8}\right)^{\frac{1}{r}} \eta\right)>0{,}15 p, \quad \text{si } p \geqq 3.$$

Du résultat précédent on pourra déduire H5 de la manière suivante qui s'appliquera à de nombreux exemples.

Corollaire 3.9. *Supposons que la distance d soit telle que d^r soit suradditive et que δ soit une fonction positive décroissante. Supposons que pour tout $n \geqq n_0$, il* *existe* p_n et η_n et une constante c indépendante de n telle que l'on ait

$$\left(\frac{p_n}{8}\right)^{\frac{2}{r}} \eta_n^2 \geqq c^2 (L\delta(n))^2, \qquad p_n \geqq 3. \tag{3.16}$$

Si pour ces n, il existe un ensemble Θ'_n, *inclus dans* Θ, *répondant aux hypothèses de la proposition* 3.8, *et*

$$\sup_{\substack{s,t \in \Theta'_n \\ d(s,t) \geqq cL\delta(n)}} K(P_{s,n}, P_{t,n}) \leqq \frac{p_n}{10}, \tag{3.17}$$

alors il existe $\gamma \leqq 0{,}6$ *tel que l'hypothèse* H5 $(cL\delta, \gamma, n_0)$ *soit satisfaite sur* (Θ, d).

Preuve. C'est une conséquence facile de la proposition 3.8. En effet Θ'_n contient un ensemble $cL\delta(n)$-discernable S'_n de cardinal $m+1$ avec $\log m > 0{,}316 p_n$ ou $0{,}648 p_n$ suivant la valeur de p_n et sur cet ensemble $K(P_{s,n}, P_{t,n})$ est majoré par $\frac{p_n}{10}$. On vérifie alors aisément que $\frac{p_n}{10} \leqq 0{,}316 \gamma p_n - \log 2$ si $p_n \geqq 8$ et que $\frac{p_n}{10} \leqq 0{,}648 \gamma p_n - \log 2$ si $p_n < 8$. □

Remarque. Dans le cas de variables indépendantes équidistribuées, la proposition 3.8 permet d'obtenir une inégalité analogue à celle de la proposition 2.1 de Bretagnolle et Huber [8]. Nous nous placerons dans le cadre de la proposition 3.8 où d^r est une distance suradditive, en choisissant $g'_i = -g_i$. Dans ce cas, et à la condition que $\left\|\frac{g_i}{f}\right\|_\infty < 1$ et $\int g_i = 0$ pour tout i, il est montré dans [8] que

$$R_n(d^r) \geqq \tfrac{1}{2} \sum_{i=1}^{p} \left(\frac{\eta}{2}\right)^r \exp[-nK_i], \tag{3.18}$$

avec $K_i = K(f+g_i, f-g_i)$ et $d(f,g) = \|f-g\|_r$. Si l'on applique la proposition 3.8, la proposition 2.8 et la surradditivité de K, on obtient que pour

$$n \sum_{i=1}^{p} K_i \leqq \gamma \times 0{,}316 p - \log 2, \qquad R_n(d^r) \geqq (1-\gamma) \left(\frac{\eta}{2}\right)^r \times \frac{p}{8}. \tag{3.19}$$

L'utilisation de ces inégalités se fait évidemment pour n et p grands et η petit. En particulier dans (3.18) on s'arrangera pour avoir K_i de l'ordre de n^{-1}. Plus particulièrement, choisissons η assez petit pour que $K_i \leqq \frac{1}{10n}$. Alors (3.19) sera satisfaite avec $\gamma = 0{,}4$ si p est grand et l'on obtiendra par (3.18) $R_n(d^r) \geqq \frac{1}{2} p \left(\frac{\eta}{2}\right)^r \exp[-\frac{1}{10}]$ et par (3.19) $R_n(d^r) \geqq 0{,}6 \frac{p}{8} \left(\frac{\eta}{2}\right)^r$. Ces minorations seront donc analogues, à des constantes multiplicatives près, bien que (3.18) soit un peu meilleure. Il s'ensuit que dans le cas des petites perturbations, (3.18) serait plus avantageuse. Néanmoins, nous continuerons d'utiliser la proposition 3.8 qui permet de revenir aux minorations plus générales qu'exprime la proposition 2.8 ou de se ramener directement à l'hypothèse H5 grâce au corollaire 3.9.

4. Variables indépendantes équidistribuées

Dans le cas des variables indépendantes équidistribuées la distance d est la distance de Hellinger entre les lois d'une variable ou une distance déduite de la norme $\mathbb{L}^p$, $1 \leqq p \leqq 2$. En ce qui concerne les majorations, les hypothèses nécessaires seront toujours vérifiées si $d = h$. En effet, on choisira une fonction $\delta > \tilde{d}$ lorsque $\tilde{d}$ est finie, égale sinon, et par définition on pourra trouver un ε-réseau S satisfaisant à $H_1(\varepsilon, 4, \delta(\varepsilon))$. On choisit pour B_s la boule $\bar{\mathscr{B}}(s, \varepsilon)$. On sait que l'on peut appliquer (1.9) et obtenir (1.11) ce qui nous conduit finalement à écrire:

$$\pi(B_s^n, B_t^n) \leqq \exp\left[-n(h(s,t) - 2\varepsilon)^2\right] \quad \text{si } h(s,t) > 2\varepsilon. \tag{4.1}$$

Il est alors évident que $\mathrm{H}4(\delta, \frac{1}{3}, 1, 4, A_1, \frac{1}{4})$ est satisfaite avec $A_1 = 2 + \eta$ et η arbitrairement petit, étant donnée la définition de π. Si l'on utilise les tests définis dans [5] chaque erreur est majorée par le second membre de (4.1) et $A_1 = 1$. Il est à noter que l'inégalité (4.1) confère aux d-estimateurs des propriétés de robustesse à la non-équidistribution. En effet on teste entre B_s^n et B_t^n ce qui permet de supposer que l'espace des paramètres ne contient pas seulement les lois $P_{s'}^n$ mais aussi les lois $\bigotimes_{i=1}^{n} P_i$ où $P_i \in B_s$ pour tout i.

Pour utiliser une distance de type $\mathbb{L}^p$, des précautions supplémentaires seront nécessaires parce que l'on a alors besoin que cette distance soit majorée par un multiple de h et aussi à boules convexes pour appliquer (1.9), mais ce point est clair. Nous utiliserons pour vérifier cela le lemme immédiat suivant:

Lemme 4.1. *Soit f_θ une famille de densités dans $\mathbb{L}^2(\mu)$, si les f_θ sont uniformément majorées par un nombre γ (ce que nous noterons $BS(\gamma)$) on a*

$$\|f_\theta - f_{\theta'}\|_2^2 \leqq 8\gamma h^2(f_\theta, f_{\theta'}); \tag{4.2}$$

si elles sont uniformément minorées par $\gamma' > 0$ (noté $BI(\gamma')$)

$$h^2(f_\theta, f_{\theta'}) \leqq \frac{1}{8\gamma'} \|f_\theta - f_{\theta'}\|_2^2. \tag{4.3}$$

En particulier sous $LB(\gamma_1, \gamma_2)$, h et $\| \ \|_2$ sont équivalentes.

Il s'ensuit, en utilisant l'inégalité sur les normes

$$\|f_\theta - f_{\theta'}\|_p \leqq \|f_\theta - f_{\theta'}\|_q, \quad \text{si } 1 \leqq p \leqq q, \tag{4.4}$$

vraie lorsque μ est une probabilité, que si Θ satisfait en outre à $BS(\gamma)$:

$$\|f_\theta - f_{\theta'}\|_p \leqq \sqrt{8\gamma}\, h(\theta, \theta') \qquad 1 < p \leqq 2.$$

Pour $p = 1$, on peut appliquer (1.8) même sans condition BS, ce qui fait que tout ce qui suit pourrait être fait en utilisant la distance $\mathbb{L}^1$ plutôt que h. Ceci permet donc, sous l'hypothèse $BS(\gamma)$, de voir que $A_1 \exp\left[-\frac{n}{4} h^2(s,t)\right] \leqq A_1$

$\cdot \exp[-nA_2 d^2(s,t)]$ lorsque d est une distance du type précédent, avec $A_2 = \frac{1}{32\gamma}$. Donc sous les hypothèses $BS(\gamma)$, nous pourrons encore utiliser H4, ce qui permettra d'appliquer la proposition 3.1. Dans le cas de la distance de Hellinger, on peut comme nous l'avons vu au corollaire 2.6 améliorer les constantes en utilisant (4.1) afin d'obtenir (2.7).

Remarque 1. Malgré des définitions un peu différentes de la dimension, l'inégalité (2.7) est comparable à celle obtenue par Le Cam pour ses estimateurs $\hat{\theta}'_n$ dans [26]

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta[nh^2(\theta, \hat{\theta}'_n)] \leqq 6{,}5D + 5{,}5$$

où 2^D est le nombre minimal d'ensembles de diamètre ε nécessaires à recouvrir un ensemble de diamètre 2ε si $\varepsilon \geqq \varepsilon_0$ avec $\varepsilon_0 = 0{,}1 \left(\frac{D}{n}\right)^{1/2}$.

Remarque 2. Le Cam a également étudié le cas d'observations non-équidistribuées, introduisant entre les mesures une distance H telle que H^2 soit la somme des carrés des distances de Hellinger entre les composantes c'est à dire que $P_\theta = \bigotimes_{i \in I} P_{\theta,i}$ et

$$H^2\left(\bigotimes_{i \in I} P_{\theta,i}, \bigotimes_{i \in I} P_{\theta',i}\right) = \sum_{i \in I} h^2(P_{\theta,i}, P_{\theta',i}).$$

Il obtient ainsi pour le risque (cf. [24, 25]) des bornes du type

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta[H^2(P_\theta, P_{\hat{\theta}})] \leqq CD \log D.$$

Les méthodes précédentes permettent d'obtenir des résultats analogues mais des travaux récents de Birgé [7] concernant les tests de deux boules en distance H permettent de construire des H-estimateurs pour lesquels le risque présente la même forme que dans le cas équidistribué c'est à dire

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta[H^2(P_\theta, P_{\hat{\theta}})] \leqq C'D.$$

La construction de h-estimateurs à partir de la famille de tests construits dans [6], si elle est effectivement réalisable, peut conduire à des calculs pénibles. Pour une application pratique on peut souhaiter les simplifier. Or le corollaire 2.2 implique que si les tests $\varphi_{s,t}$ sont des tests de vraisemblance entre les points du η-réseau S, on peut choisir comme d-estimateur un maximum de vraisemblance sur S, ce qui conduira à des calculs plus agréables dans la pratique. Il est donc intéressant de voir à quelles conditions les tests entre les points de S pourront vérifier l'hypothèse H4. Nous supposerons que Θ est une partie de $\mathbb{L}^1(\mu)$ et nous utiliserons la distance uniforme entre les éléments de $\mathbb{L}^1(\mu)$ définie par $\|f-g\|_\infty = \operatorname{ess\,sup}_\mu |f-g|$. Les résultats reposent sur le lemme fondamental suivant:

Lemme 4.2. *Supposons que l'on ait trois densités f, f' et g par rapport à une probabilité μ qui vérifient*

$$\|f'-f\|_\infty \leqq \varepsilon, \quad f \geqq \gamma, \quad \rho(f,g) = 1 - k^2\varepsilon^2, \quad k \geqq 5\gamma^{-1}.$$

Alors si $P_{f'}^n$ désigne la puissance n-ième de la mesure $f' \cdot \mu$ on a l'inégalité

$$P_{f'}^n \left[\sum_{i=1}^n \log \frac{g(x_i)}{f(x_i)} \geqq 0 \right] \leqq \exp \left[-\frac{n}{2} h^2(f, g) \right].$$

Preuve. Posons $f'(x) = f(x)[1 + \eta(x)]$, alors $|\eta(x)| \leqq \frac{\varepsilon}{\gamma}$ et aussi

$$\int \left(\frac{g}{f}\right)^{\frac{1}{2}} f' d\mu = \int (fg)^{\frac{1}{2}} (1 + \eta) d\mu.$$

Puisque f, f' et g sont des densités, $\int f\eta d\mu = 0$ et

$$\int (fg)^{\frac{1}{2}}(1+\eta) d\mu = 1 - \tfrac{1}{2} \int (\sqrt{f} - \sqrt{g})^2 \, (1+\eta) d\mu + \tfrac{1}{2} \int f\eta d\mu + \tfrac{1}{2} \int g\eta d\mu,$$

d'où

$$\int \left(\frac{g}{f}\right)^{\frac{1}{2}} f' d\mu \leqq 1 - \left(1 - \frac{\varepsilon}{\gamma}\right) h^2(f, g) + \tfrac{1}{2} \int |f - g| \, |\eta| \, d\mu$$

$$\leqq 1 - \left(1 - \frac{\varepsilon}{\gamma}\right) h^2(f, g) + \tfrac{1}{2} \frac{\varepsilon}{\gamma} \, \|f - g\|_1.$$

En utilisant (1.8), on trouve finalement:

$$\int \left(\frac{g}{f}\right)^{\frac{1}{2}} f' d\mu \leqq 1 - \left(1 - \frac{\varepsilon}{\gamma}\right) k^2 \varepsilon^2 + \sqrt{2} \frac{\varepsilon}{\gamma} k\varepsilon.$$

Comme $k\varepsilon \leqq 1$ ceci peut encore s'écrire

$$\int \left(\frac{g}{f}\right)^{\frac{1}{2}} f' d\mu \leqq 1 - k^2 \varepsilon^2 \left[1 - \left(\frac{1 + \sqrt{2}}{k\gamma}\right)\right] \leqq 1 - \frac{k^2 \varepsilon^2}{2} = 1 - \tfrac{1}{2} h^2(f, g),$$

puisque $k\gamma \geqq 5$. Pour conclure, il suffit d'utiliser l'inégalité exponentielle

$$P_{f'}^n \left[\sum_{i=1}^n \log \frac{g(x_i)}{f(x_i)} \geqq 0 \right] \leqq \left[\int \exp \left(\tfrac{1}{2} \log \frac{g}{f}\right) f' d\mu \right]^n. \quad \square$$

Supposons alors que Θ soit une partie de $\mathbb{L}^1(\mu)$ satisfaisant à $BI(\gamma)$. Notons pour simplifier $d'(f, g) = \|f - g\|_\infty$. Si d est la distance de Hellinger ou majorée par un de ses multiples, on peut supposer d'après $BI(\gamma)$ que $d \leqq d'$ (quitte à remplacer d par un multiple) en utilisant (4.3) et (4.4). Supposons alors que δ majore $\tilde{d}'$, on va pouvoir trouver un (ε, d')-réseau S qui vérifiera $\mathrm{H1}(\varepsilon, 4, \delta(\varepsilon))$ (et sera aussi un (ε, d)-réseau) et lui associer un recouvrement par des boules de rayon ε (pour d'). D'après le lemme 4.2 les tests entre les centres de ces boules vérifieront H4, ii) avec $A_1 = 1$, $A_2 = \frac{1}{2}$ si $d = h$ $\left(\text{ou } \frac{1}{2M} \text{ si } d^2 \leqq Mh^2\right)$ et $k_0 = \frac{5}{\gamma}$ $\left(\text{ou } \frac{5\sqrt{M}}{\gamma} \text{ si } d^2 \leqq Mh^2\right)$. H4 sera satisfaite et tout maximum de vraisemblance sur S sera un d-estimateur d'après le corollaire 2.2. En résumé nous pouvons énoncer le résultat suivant:

Proposition 4.3. *Soit Θ un ensemble de probabilités muni d'une distance d majorée par un multiple de la distance de Hellinger et supposons que $P_{\theta,n}=P_\theta^n$.*

i) *Si $\delta(\varepsilon)>\tilde{d}(\varepsilon)$ pour $\varepsilon\leqq\varepsilon_0$, l'hypothèse H4$(\delta,\varepsilon_0,1,4,A_1,A_2)$ sera satisfaite avec A_1 arbitrairement proche de 2 et $A_2=\frac{1}{4M}$ si $d^2\leqq Mh^2$.*

ii) *Si Θ est inclus dans $\mathbb{L}^1(\mu)$ où μ est une probabilité, et vérifie $BI(\gamma)$, si $d'(f,g)=\|f-g\|_\infty$ et $\delta(\varepsilon)>\tilde{d}'(\varepsilon)$ pour $\varepsilon\leqq\varepsilon_0\leqq k_0^{-1}$, l'hypothèse H4$(\delta,\varepsilon_0,1,k_0,A_1,A_2)$ sera satisfaite avec $A_1=1$ et $A_2=\frac{1}{2M}$ si $d^2\leqq Mh^2$, $d\leqq d'$ et pour d-estimateurs on pourra choisir des maximums de vraisemblance sur des (ε,d')-réseaux.*

Dans le cas ii), on pourrait s'attendre à ce que le fait de remplacer $\tilde{d}$ par $\tilde{d}'$ nous fasse perdre beaucoup, puisque les majorations que l'on obtient seront fonction de $L\delta$ et donc de $L\tilde{d}'$ pour un bon choix de δ. Mais il se trouve que pour les exemples où l'on souhaitera utiliser ii), $L\tilde{d}'\asymp L\tilde{d}$ et les majorations obtenues sont du même ordre de grandeur que si l'on travaillait avec un (ε,d)-réseau. Ceci permettra de les faire coïncider (aux constantes multiplicatives près) avec les minorations.

Pour ce qui est des minorations, nous remarquons avant tout que $K(P_s^n,P_t^n)=nK(P_s,P_t)$. Contrairement aux hypothèses H4, les hypothèses H5 ne seront pas automatiquement vérifiées. Elles ont simplement de «bonnes chances» de l'être comme le montre le raisonnement heuristique suivant. Supposons que $\tilde{d}(\varepsilon)$ soit à peu près égal à $n\varepsilon^2$, d'après ce qui précède, le risque obtenu par les d-estimateurs sera de l'ordre de ε^2 (si la perte est quadratique) avec n observations. Dans les cas ordinaires $\tilde{d}$ est approximativement décroissante, ce qui fait que pour une structure métrique de Θ régulière et un bon choix de $\eta,\eta<\frac{\varepsilon}{4A}$, $A>1$, on pourra d'après la définition de $\tilde{d}$ trouver un ensemble S' n-discernable de cardinal supérieur à $2^{\tilde{d}(\varepsilon)}$ contenu dans une boule de diamètre $\frac{\varepsilon}{2A}$. Si cet ensemble S' est choisi de manière que sur S' K et d^2 soient comparables, en particulier si $K\leqq A^2d^2$, on aura $K(P_s,P_t)\leqq\frac{\varepsilon^2}{4}$ pour s et t dans S' et en passant aux mesures produits $K(P_s^n,P_t^n)\leqq n\frac{\varepsilon^2}{4}$. Il est facile alors de voir que H3(η,γ) sera satisfaite avec $\gamma>\frac{1}{2}$ et donc que le risque minimax sera de l'ordre de η^2. Il suffit alors d'écrire η sous la forme $\frac{\varepsilon}{A'}$ avec $A'>4A$ pour pouvoir comparer les majorations et les minorations. Si on peut trouver un A' indépendant de n, la vitesse obtenue par les d-estimateurs sera adéquate. Le point crucial du «raisonnement» précédent est la possibilité de trouver un ensemble S' assez gros sur lequel $K\leqq A^2d^2$. Pour ce faire, le lemme suivant sera très utile:

Lemme 4.4. *Soit μ une mesure, $P=f\cdot\mu$, $Q=g\cdot\mu$. Supposons que $1-\alpha\leqq\frac{g}{f}\leqq1+\beta$, $\alpha<1$, $\beta>0$. Il existe alors deux fonctions monotones $C(\alpha)$ et $C'(\beta)$ telles que*

$$2h^2(P,Q) \leqq C'(\beta) h^2(P,Q) \leqq K(P,Q) \leqq C(\alpha) h^2(P,Q); \tag{4.5}$$

C et C' tendant vers 4 lorsque α et β tendent vers 0. On peut vérifier que l'on a $C'(1) > 3{,}7$, $C(\frac{1}{2}) < 5$, $C(\frac{1}{4}) < 4{,}3$.

Preuve. Elle est purement calculatoire. On écrit $\log(1+h) = h - t(h)h^2$ et $\sqrt{1+h} = 1 + \frac{h}{2} - s(h)h^2$. Le problème se ramène à étudier le comportement du rapport $\frac{t(h)}{s(h)}$ sur $]-1; +\infty[$. L'inégalité $K \geqq 2h^2$ est classique (voir [10]). □

Ce lemme s'appliquera évidement toujours dans le cas $LB(\gamma_1, \gamma_2)$. Grâce au lemme 4.1, on constate qu'alors h^2, $\| \|_2^2$ et K sont équivalents. C'est le cas le plus favorable pour l'application du théorème 3.4.

Quelques exemples

Les ellipsoïdes de $\mathbb{L}^2$ (cf. [30] pour des résultats plus précis)

Le premier exemple que nous examinerons est celui des ellipsoïdes. Considérons l'ensemble des fonctions f positives, continues sur $[0;1]$ et telles que $f(0) = f(1)$. Ces fonctions admettent des développements en séries trigonométriques de la forme

$$f(x) = x_0 + \sum_{k=1}^{+\infty} [x_k \cos(2\pi k x) + y_k \sin(2\pi k x)].$$

Puisque seules les densités par rapport à la mesure de Lebesgue sur $[0; 1]$ nous intéressent, nous supposerons que $x_0 = 1$ et considérerons l'ensemble $E(a)$ associé à la suite strictement positive décroissante $\{a_k\}_{k \geqq 1}$ par

$$\sum_{k=1}^{+\infty} \frac{x_k^2 + y_k^2}{a_k^2} \leqq 1. \tag{4.6}$$

Nous supposerons en outre que la série $\sum a_k^2$ converge. Posons $x_k^2 + y_k^2 = \alpha_k^2$, $\alpha_k \geqq 0$; alors l'inégalité de Cauchy-Schwartz et (4.6) entraînent que $|x_k \cos(2\pi k x) + y_k \sin(2\pi k x)| \leqq \alpha_k$ et $(\sum \alpha_k)^2 \leqq \sum a_k^2$, donc que $E(a)$ satisfait à la condition BS; celle-ci comme nous l'avons vu permet de remplacer la distance de Hellinger par la distance $\mathbb{L}^2$, adaptée à ce problème hilbertien. Pour vérifier H4 il suffira donc de majorer $\tilde{d}$. Pour cela nous aurons recours au lemme suivant de théorie de l'approximation (voir Lorentz [27]).

Lemme 4.5. *Soit une boule $B = \bar{\mathscr{B}}(s, p\varepsilon)$, $p \geqq 1$ dans l'espace euclidien $\mathbb{R}^k$ (ou $\mathbb{R}^k$ muni d'une autre norme). Si S est un sous-ensemble de B, ε-discernable et maximal, c'est un ε-réseau de B tel que*

$$p^k \leqq \operatorname{Card} S \leqq (2p+1)^k.$$

Considérons dans $E(a)$ la partie W_k obtenue en faisant $x_j = y_j = 0$ pour $j > k$. W_k approxime $E(a)$ à a_{k+1} près. Alors si $\varepsilon \geqq a_{k+1}$ et S est un ensemble ε-

discernable maximal dans W_k, c'est un 2ε-réseau de $E(a)$. En utilisant le lemme 4.5, il est aisé de voir que $\tilde{d}(2\varepsilon)<4k$ vu que W_k est une partie d'espace affine de dimension inférieure à $2k$. Ceci nous amène à poser $\delta(2\varepsilon)=4k$ pour $a_k>\varepsilon\geqq a_{k+1}$. Considérons maintenant l'ensemble S_k des f de $E(a)$ pour lesquels $\sum_{j=1}^{k} x_k^2 \leqq a_k^2$, les autres coefficients étant nuls. Cet ensemble est une boule de rayon a_k. D'autre part comme $\sum a_k^2<+\infty$, $\sqrt{k}\,a_k \xrightarrow[k\to+\infty]{} 0$. Donc pour k assez grand, tous les éléments de S_k seront positifs (dès que $\sqrt{k}\,a_k<1$) et même log-bornés uniformément. Ceci entraîne que pour k grand on aura sur S_k $K(f,g)\leqq A\|f-g\|_2^2$. De plus, si $\varepsilon\leqq a_k$, on pourra trouver dans S_k un ensemble $\frac{\varepsilon}{2}$-discernable ayant plus de 2^k éléments et sur lequel K sera majoré par $4A\varepsilon^2$. On voit que

$$L\delta(n)=\begin{cases} 2\sqrt{\dfrac{k}{n}} & \text{si } n\in\left]\dfrac{k}{a_k^2}, \dfrac{k}{a_{k+1}^2}\right] \\ 2a_{k+1} & \text{si } n\in\left]\dfrac{k}{a_{k+1}^2}, \dfrac{k+1}{a_{k+1}^2}\right] \end{cases} \tag{4.7}$$

Nous voulons vérifier H5$(cL\delta, n_0, \gamma)$. Pour n_0 grand, nous pouvons supposer que le k qui intervient dans $L\delta(n)$ est grand et donc que l'inégalité $K(f,g)\leqq A\|f-g\|_2^2$ est vérifiée sur S_k. Soit $A'>1$, on a visiblement $\frac{L\delta(n)}{2A'}\leqq a_k$ si $\frac{k+1}{a_{k+1}^2}>n\geqq\frac{k}{a_k^2}$. Donc on peut trouver un ensemble S', $\frac{L\delta(n)}{4A'}$-discernable avec $\operatorname{Card} S'\geqq 2^k$ sur lequel K sera majoré par $[L\delta(n)]^2\,\frac{A}{A'^2}$ et pour n observations $K(f^n,g^n)$ sera majoré par $n[L\delta(n)]^2\,\frac{A}{A'^2}$ et donc d'après (4.7) par $4(k+1)\,\frac{A}{A'^2}$. H5 sera donc satisfaite dès que

$$4(k+1)\,\frac{A}{A'^2}\leqq\gamma\log[\operatorname{Card} S'-1]-\log 2.$$

Comme k est grand et $\operatorname{Card} S'\geqq 2^k$, ceci sera vérifié pour $A'=3\sqrt{A}$ par exemple. On peut donc appliquer le théorème 3.4 d'où l'on conclut que $R_n(d^q)\asymp[L\delta(n)]^q$. Vu la forme de $L\delta(n)$ ceci n'est guère explicite, mais on peut regarder des cas particuliers:

– si $a_k=k^\lambda$ avec $\lambda<-\frac{1}{2}$, on trouve que $L\delta(n)\asymp n^{\frac{\lambda}{1-2\lambda}}$ et $R_n(d^q)\asymp n^{\frac{q\lambda}{1-2\lambda}}$

– si $a_k=e^{-\lambda k}$, $\lambda>0$ on trouve $L\delta(n)\asymp\left(\frac{\log n}{n}\right)^{\frac{1}{2}}$ et $R_n(d^q)\asymp\left(\frac{\log n}{n}\right)^{\frac{q}{2}}$.

Les familles uniformément équicontinues

Nous considérons ici la famille A_ω des fonctions sur $[0;1]$ ayant un module de continuité majoré par une fonction concave croissante ω, c'est-à-dire telles que

$$|f(x+h)-f(x)|\leqq\omega(h) \qquad \text{pour tous } x, h;\ x \text{ et } x+h \text{ dans } [0;1].$$

En fait on considère d'ordinaire des fonctions continues, ω étant continu avec $\omega(0)=0$. Mais nous permettrons ici aux fonctions d'être discontinues et supposerons que $0 \leqq \omega(0) \leqq \frac{1}{2}$. Bien évidemment, nous restreindrons Θ aux éléments de Λ_ω qui sont des densités positives par rapport à la mesure de Lebesgue. Nous ferons ici encore les calculs en utilisant la distance $\mathbb{L}^2$. Comme il est clair que toutes ces densités sont majorées par $1+\omega(1)$, la condition BS est satisfaite et H4 sera vérifiée aisément dès que l'on sait majorer δ. Pour cela nous utiliserons un calcul analogue à celui effectué par Kolmogorov-Tihomirov dans [20].

Soit $t>0$ donné et $\omega(t)=\varepsilon$. Dans $[0;1]\times[0;1+\omega(1)]$, nous considérerons un réseau de points de la forme $(kt, k'\varepsilon)$ avec k et k' entiers positifs et l'ensemble des fonctions qui joignent ces points c'est-à-dire telles que $\psi(kt)=k'\varepsilon$ pour toutes les valeurs possibles, et qui sont linéaires entre kt et $(k+1)t$. En fait nous imposerons à ψ la contrainte supplémentaire de ne pas varier plus que ω, c'est-à-dire que pour tous les k:

$$\psi[(k+1)t]=\psi(kt)+j\varepsilon \qquad \text{avec } j=0;\ 1 \text{ ou} -1.$$

Alors, pour toute f de Θ, il existe ψ de ce type telle que $\|f-\psi\|_\infty \leqq 2\varepsilon$. On le voit facilement par récurrence. Si $k'\varepsilon \leqq f(kt) < (k'+1)\varepsilon$, nous prendrons $\psi(kt)=k'\varepsilon$. On a alors $k'\varepsilon+j\varepsilon \leqq f((k+1)t) < (k'+j+1)\varepsilon$ avec $j=0$; 1 ou -1. On choisit $\psi[(k+1)t]=(k'+j)\varepsilon$. Il est facile de voir que dans chacun des 3 cas, $|\psi-f| \leqq 2\varepsilon$ sur l'intervalle $[kt,(k+1)t[$. Il est alors aisé d'obtenir ainsi un 2ε-réseau S dans Θ, pour la distance uniforme. Pour cela on n'a besoin que des fonctions ψ pour lesquelles $\psi(kt)=1$ ou $1-\varepsilon$ pour au moins une valeur de k vu que f est une densité et qu'on aura donc $|f(kt)-1|<\varepsilon$ pour un certain k. Il s'ensuit que l'on peut prendre $\operatorname{Card} S < 2t'3^{t'}$ où $t'=[\![t^{-1}]\!]+1$ ($[\![u]\!]$ désigne la partie entière de u). Mais comme on a évidemment sur Θ $\|f-\psi\|_2 \leqq \|f-\psi\|_\infty$, S est aussi un $(2\varepsilon, d)$-réseau ce qui entraîne puisque $\varepsilon=\omega(t)$:

$$\tilde{d}(2\omega(t)) \leqq \tfrac{1}{2}[\log 2t' + t'\log 3] \leqq \tfrac{3}{2}t^{-1} \qquad \text{pour } t \leqq \tfrac{1}{4}.$$

Nous poserons donc $\delta(2\omega(t))=\dfrac{3}{2t}$ si $t \leqq \frac{1}{4}$.

La fonction $t\omega^2(t)$ est continue croissante sur $[0;\frac{1}{4}]$ et donc pour tout $n \geqq n_0$ avec n_0 bien choisi, il existe $t_n \leqq \frac{1}{4}$ avec $t_n\omega^2(t_n)=\dfrac{3}{8n}$ et on en déduit

$$L\delta(n)=2\omega(t_n) \qquad \text{si } n \geqq n_0 \text{ et } t_n\omega^2(t_n)=\frac{3}{8n}. \tag{4.8}$$

Pour vérifier H5, nous considérerons le système de perturbations suivant: on a choisi une partition de $[0;1]$ en $t'=[\![(2t)^{-1}]\!]$ intervalles de longueur $2t$ et un résidu de longueur $<2t$ qu'on laisse de côté. Posons $a_j=2jt$ et g_j est la fonction sur $[a_j;a_{j+1}]$, linéaire par morceaux avec $g_j(a_j)=0$,

$$g_j\left(\frac{3a_j+a_{j+1}}{4}\right)=\frac{\varepsilon}{2}, \quad g_j\left(\frac{a_j+a_{j+1}}{2}\right)=0, \quad g_j\left(\frac{a_j+3a_{j+1}}{4}\right)=-\frac{\varepsilon}{2},$$

$$g_j(a_{j+1})=0 \quad \text{et} \quad \varepsilon=\omega(t)<1.$$

Nous considérons le sous-ensemble Θ' défini par $\Theta'=\left\{1+\sum_{j=0}^{t'-1}\lambda_j g_j\right\}$, $\lambda_j=\pm 1$. Θ' est une partie de Θ, vu le choix de ε et la concavité de ω. De plus on voit aisément que l'on a

$$\|1+g_j,\ 1-g_j\|_2^2=4\int_{a_i}^{a_i+1} g_j^2\,d\lambda=\tfrac{2}{3}t\varepsilon^2,$$

d'où $K(1+g_j,\ 1-g_j)\leqq\frac{2}{3}t\varepsilon^2$, grâce à l'inégalité donnée dans [8]

$$\left|(1+z)\log\left(\frac{1-z}{1+z}\right)-2z\right|\leqq(1-|z|)^{-1}2z^2,$$

utilisée avec $z=g_j$ et $|g_j|=\dfrac{\varepsilon}{2}<\frac{1}{2}$.

Nous appliquerons alors le corollaire 3.9 avec $r=2$ et δ comme en (4.8). Soit $k>1$. Pour n assez grand $\geqq n_0$, $L\delta(n)<2$, $t_n\leqq\frac{1}{6}$ et donc dans (4.8) $\omega(t_n)<1$. Par continuité, il existe $u_n<t_n$ tel que

$$u_n\omega^2(u_n)=\frac{t_n\omega^2(t_n)}{k}\quad\text{et}\quad 1>\omega^2(u_n)>\frac{\omega^2(t_n)}{k}.$$

Choisissons un ensemble Θ'_n comme il vient d'être montré avec $t=u_n$, $\varepsilon=\omega(u_n)$. Pour que cet ensemble satisfasse aux hypothèses du corollaire 3.9, il faut que l'on ait:

$$\left(\frac{[\![(2u_n)^{-1}]\!]}{8}\right)\tfrac{2}{3}u_n\omega^2(u_n)\geqq 4c^2\omega^2(t_n),\qquad [\![(2u_n)^{-1}]\!]\geqq 3$$

et

$$\tfrac{2}{3}nu_n\omega^2(u_n)[\![(2u_n)^{-1}]\!]\leqq\tfrac{1}{10}[\![(2u_n)^{-1}]\!]. \tag{4.9}$$

La première inégalité est satisfaite avec $c^2\leqq\dfrac{1}{108k}$. La deuxième l'est parce que $t_n\leqq\frac{1}{6}$. La dernière inégalité nécessite seulement $k\geqq 2{,}5$ puisque $t_n\omega^2(t_n)=\dfrac{3}{8n}$. Ainsi l'hypothèse $H5(cL\delta,\gamma,n_0)$ est satisfaite si n_0 est assez grand et le théorème 3.4 s'applique ici. Dans le cas particulier des fonctions Lipschitziennes $\omega(x)=Cx$ où C est une constante et donc

$$L\delta(n)\asymp n^{-\frac{1}{3}},\qquad R_n(d^q)\asymp n^{-\frac{q}{3}}.$$

Pour des fonctions α-Hölderiennes $\omega(x)=Cx^\alpha$ d'où

$$L\delta(n)\asymp n^{-\frac{\alpha}{2\alpha+1}},\qquad R_n(d^q)\asymp n^{-\frac{q\alpha}{2\alpha+1}}.$$

Enfin lorsque $\omega(0)=a>0$, $L\delta(n)\xrightarrow[n\to+\infty]{}2a$ et $R_n(d^q)\xrightarrow[n\to+\infty]{}a'>0$.

Remarque. On s'est placé ici sur $[0;1]$, il est clair que la démonstration peut se transposer sur tout intervalle fermé de $\mathbb{R}$.

Autres problèmes d'estimation de densités

Ce dernier exemple se trouve être tout à fait typique en ce sens que d'autres problèmes d'estimation de densités pourraient se traiter de façon absolument analogue. En effet les majorations de $\tilde{d}$, tout comme l'obtention du système de petites perturbations g_i, sont calquées sur les méthodes utilisées en théorie de l'approximation pour calculer les exposants d'entropie des ensembles Λ_ω. Donc chaque fois que l'ensemble Θ consiste en une famille de densités appartenant à une classe bien connue en théorie de l'approximation, on pourra s'inspirer de cette théorie pour calculer δ et vérifier H5. Les références importantes sont Kolmogorov-Tihomirov [20], Lorentz [27] et [28] et Vitushkin [32], et nous conduisent aux familles déjà étudiées dans la littérature (cf. [8, 12, 17] et [33]) c'est à dire les classes $\Lambda^s_{\alpha,r}$ et $\Lambda^s_{\alpha,r}(p)$ décrites dans l'introduction (mais le plus souvent avec $s=1$). Ces classes satisfont à des conditions BS et on peut donc utiliser diverses distances. Mais pour $\Lambda^s_{\alpha,r}(p)$, les calculs d'approximation étant connus en distance $\mathbb{L}^p$, c'est celle-ci qu'il convient d'utiliser. Dans ces deux derniers cas, les vitesses sont celles mentionnées dans l'introduction c'est-à-dire $R_n(d^q) \asymp n^{\frac{-q(r+\alpha)}{s+2(r+\alpha)}}$. L'exemple de $\Lambda^s_{\alpha,r}$ et celui de Λ_ω (qui ne semble pas avoir été déjà traité dans la littérature) appellent un certain nombre de remarques qui peuvent se généraliser à tout problème du même type.

Remarque 1. Les calculs de réseaux, se font en distance uniforme et donc sont encore valables pour les distances $\mathbb{L}^p(1 \leqq p \leqq 2)$ et pour celle de Hellinger également si l'on a en plus une condition BI permettant d'utiliser (4.3). On obtient ainsi une même fonction δ pour toutes ces distances, les minorations ne posant pas de problème puisque pour les systèmes de petites perturbations grâce auxquels on les obtient, toutes ces distances sont équivalentes.

Remarque 2. Modulo une condition BI si $d=h$, la proposition (4.3) ii) sera appliquable puisque $\delta \geqq \tilde{d}'$ où d' est la distance uniforme et on pourra utiliser comme estimateurs des maximums de vraisemblance sur des ε-réseaux, lesquels sont fournis par la théorie de l'approximation.

Remarque 3. Il est facile de voir que la distance h entre f et g peut être considérée comme une distance $\mathbb{L}^2$ entre $\sqrt{f}$ et $\sqrt{g}$. Il s'ensuit que si l'on connaît la structure métrique de l'ensemble Λ en distance $\mathbb{L}^2$, on connaîtra aussi celle de $\Lambda' = \{f \mid \sqrt{f} \in \Lambda\}$ en distance de Hellinger. Bien évidemment, si f est une densité, $\sqrt{f}$ n'en est pas une en général et vice-versa. Toutefois ceci n'entraîne que de légères modifications dans le choix des systèmes de perturbations utilisés. En particulier, on pourrait traiter la classe $\Lambda'_\omega = \{$densités $f \mid \sqrt{f} \in \Lambda_\omega\}$ en distance de Hellinger par cette méthode.

Remarque 4. Il serait facile de vérifier, dans les deux exemples précédemment étudiés ainsi que sur $\Lambda^s_{\alpha,r}$ que pour les sous-ensembles Θ' de Θ utilisés dans les minorations $K \asymp d^2$ et donc que (3.7) est satisfaite. Par la proposition 3.5 on a donc $L\delta \asymp L\tilde{d}$.

D'autres exemples, moins classiques, et permettant d'obtenir des vitesses différentes seront abordés au chapitre 6, mais avant de passer aux questions de dépendance, nous dirons quelques mots des problèmes paramétriques et de dimension finie, bien que ceux-ci relèvent en principe de méthodes classiques plus fines adaptées à chaque cas particulier.

Les problèmes paramétriques sont ceux où l'ensemble Θ est «naturellement» isomorphe à un sous-ensemble de $\mathbb{R}^k$. Ce sont les familles exponentielles, les problèmes de paramètres de translation ou d'échelle, etc. D'une manière générale les méthodes que nous avons développées ici sont peu adaptées à ces problèmes, et ceci pour deux raisons. D'abord les calculs précédents sont grossiers, alors que l'on connaît généralement de «bons» estimateurs donnant des résultats précis et qu'il existe un très abondante littérature sur ces problèmes, chaque cas étant étudié en particulier et de manière spécifique. Les outils très généraux que nous utilisons ne sauraient faire aussi bien. Une seconde raison est que si l'on peut obtenir par nos méthodes des majorations, il sera souvent plus délicat de vérifier les hypothèses H5, beaucoup de problèmes paramétriques se prêtant mal à l'utilisation de l'information de Kullback. Il n'y a qu'à penser aux familles de translation sur $\mathbb{R}$ avec des densités à support compact.

De plus, une bonne partie des problèmes paramétriques sont de dimension finie et vérifient donc

$$\sup_{n\geqq 1}\sup_{\theta\in\theta} \mathbb{E}_\theta[nh^2(\theta,\hat\theta_n)] \leqq C.$$

En fait ces problèmes sont toujours de dimension finie au sens de la métrique euclidienne sur Θ, mais pas nécessairement au sens de celle de Hellinger (ou $\mathbb{L}^p$). Il faut pour cela que $\|\theta-\theta'\|$ et $h(f_\theta, f_{\theta'})$ soient correctement reliés. Le problème plus général des relations de dimension pour deux distances a été traité par Assouad [1]. Dans le cas d'un paramètre de translation sur $\mathbb{R}$, une condition nécessaire et suffisante pour que (Θ, h) soit de dimension finie est que $h(f_0, f_\theta)=\theta^p\varphi(\theta)$ où φ est une fonction équivalente à une fonction croissante (voir [1]). Divers cas de ce type sont étudiés dans [10], en particulier les cas habituels très réguliers sont de dimension finie (pour h).

5. Chaînes de Markov et processus gaussiens stationnaires

Lorsque l'on tente d'étendre les résultats précédents aux variables dépendantes, on se heurte à divers problèmes. La théorie métrique est en effet conçue pour les variables indépendantes, pour lesquelles la vérification des hypothèses H4 est immédiate. Ce n'est plus le cas. On rencontre également d'autres difficultés concernant les minorations. Le fait capital, en variables indépendantes est la vitesse de séparation des mesures produits qui s'exprime par les deux égalités

$$\rho(P^n, Q^n)=\rho^n(P,Q), \qquad K(P^n, Q^n)=nK(P,Q).$$

Ceci justifie l'utilisation de $d_n=n^{1/2}d$ et facilite la vérification de H5 puisque si $d^2(\theta,\theta')$ et $K(P_{\theta,1}, P_{\theta',1})$ sont comparables, il en sera de même de $d_n^2(\theta,\theta')$ et $K(P_{\theta,n}, P_{\theta',n})$. Nous aurons donc besoin de résultats analogues, donnant le comportement de $\rho(P_{\theta,n}, P_{\theta',n})$ et $K(P_{\theta,n}, P_{\theta',n})$ en fonction de n.

Plus délicat encore est le problème des tests entre les boules qui se résolvait assez simplement dans le cas des variables indépendantes grâce à des résultats de convexité et principalement (1.9) ou [6]. Il n'existe aucun analogue ici. On devra avoir recours à un artifice qui consiste à montrer des variantes du lemme 4.2. Malheureusement, ceci suppose l'utilisation de la distance uniforme et nécessite des hypothèses beaucoup plus restrictives sur les familles de densités que nous considérerons. Malgré cela, de nombreux exemples pourront être traités par ces méthodes.

Nous rappellerons d'abord le cadre de travail. Pour les processus de Markov, les données sont une probabilité μ, des mesures initiales ν_θ et une famille de fonctions $p_\theta(x,y)$ auxquelles on associe les transitions $p_\theta(x,y)\mu(dy)$. Un point de Θ est un couple (ν_θ, p_θ) mais on ne s'intéresse qu'à l'estimation de p_θ puisque l'information sur ν_θ n'augmente pas avec le nombre des observations. Nous abandonnerons donc la première observation X_0 de loi ν_θ, et considérerons les mesures $P_{\theta,n}$, lois de $(X_1, \ldots, X_n)$ sous (ν_θ, p_θ). La distance d est définie par

$$d^2(s,t)=\tfrac{1}{2}\int [\sqrt{p_s(x,y)}-\sqrt{p_t(x,y)}]^2\,\mu(dx)\mu(dy).$$

Pour les processus gaussiens, nous considérerons n observations d'un processus stationnaire sur $\mathbb{R}$, ayant une densité spectrale f_θ par rapport à la probabilité de Lebesgue λ sur $[0;2\pi[$. La distance est la distance $\mathbb{L}^2$ entre ces densités spectrales.

Dans les deux cas, nous ferons la même hypothèse $LB(\gamma_1,\gamma_2)$ sur la famille $\{p_\theta\}$ ou $\{f_\theta\}$, nécessaire pour montrer ce qui va suivre. Nous noterons toujours d' la distance uniforme.

Les chaînes de Markov

Proposition 5.1. *Supposons que* $n\geqq 30$, $d'(s,s')\leqq\varepsilon$ *et* $d(s,t)\geqq k_0\varepsilon$ *avec* $k_0=7\gamma_1^{-2}\gamma_2^{1/2}$.

$$P_{s',n}\left[\log\frac{dP_{t,n}}{dP_{s,n}}\geqq 0\right]\leqq 1{,}03\left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{1}{2}}\exp\left[-n\frac{\gamma_1}{5}d^2(s,t)\right]$$

si la famille $\{p_\theta\}_{\theta\in\Theta}$ *vérifie* $LB(\gamma_1,\gamma_2)$, $\gamma_1<1<\gamma_2$.

Preuve. Supposons s', s, t fixés et définissons le noyau R' par

$$R'(x,dy)=\left(\frac{p_t(x,y)}{p_s(x,y)}\right)^{\frac{1}{2}}p_{s'}(x,y)\,\mu(dy),$$

avec $\|p_s(x,y)-p_{s'}(x,y)\|_\infty\leqq\varepsilon$. Nous poserons également $R(x,dy)=(p_t(x,y)\,p_s(x,y))^{\frac{1}{2}}\mu(dy)$. $R(x,dy)$ est un noyau sous-markovien et si l'on écrit

$$\int_{\mathscr{X}} R(x,dy)=R(x,\mathscr{X})=1-k^2(x)\varepsilon^2,\qquad k(x)\geqq 0$$

alors $1-d^2(s,t)=\int R(x,\mathscr{X})\mu(dx)$ et nous nous trouvons dans un cas identique à celui du lemme 4.2. En suivant la même démonstration, nous arrivons aisément à l'inégalité analogue:

$$R'(x,\mathscr{X})\leqq 1-k^2(x)\varepsilon^2\left[1-\frac{(1+\sqrt{2})}{\gamma_1 k(x)}\right]. \tag{5.1}$$

De plus nous avons aussi

$$\left(\frac{p_t}{p_s}\right)^{\frac{1}{2}} p_{s'}=(p_t p_s)^{\frac{1}{2}}\frac{p_{s'}}{p_s}\geqq\gamma_1\left[1-\frac{\varepsilon}{\gamma_1}\right]=\gamma_1-\varepsilon\geqq\tfrac{6}{7}\gamma_1$$

puisque $p_s\geqq\gamma_1$ et $\|p_s-p_{s'}\|_\infty\leqq\varepsilon\leqq\frac{\gamma_1}{7}$ ainsi que

$$\frac{p_{s'}}{p_s}\leqq\tfrac{8}{7}\quad\text{et}\quad R'(x,\mathscr{X})\leqq 1+\left(\frac{1+\sqrt{2}}{2\gamma_1}\right)^2\varepsilon^2\leqq 1{,}03$$

en maximisant en k dans (5.1). En intégrant (5.1) il vient:

$$\begin{aligned}\int R'(x,\mathscr{X})R'(y,dx)\leqq R'(y,\mathscr{X})&-\frac{6\gamma_1}{7}\varepsilon^2\int k^2(x)\mu(dx)\\&+\left(\frac{1+\sqrt{2}}{\gamma_1}\right)\frac{8\varepsilon^2}{7}\int k(x)R(y,dx);\end{aligned} \tag{5.2}$$

comme

$$\begin{aligned}\int k(x)R(y,dx)&\leqq\left(\int p_t(y,x)p_s(y,x)\mu(dx)\int k^2(x)\mu(dx)\right)^{1/2}\\&\leqq\left(\gamma_2\int k^2(x)\mu(dx)\right)^{1/2},\end{aligned}$$

il vient en posant $\bar{k}^2=\int k^2(x)\mu(dx)$

$$\sup_{y\in\mathscr{X}}\int R'(x,\mathscr{X})R'(y,dx)\leqq 1+\varepsilon^2\left[\left(\frac{1+\sqrt{2}}{2\gamma_1}\right)^2-\frac{6\gamma_1\bar{k}^2}{7}+\left(\frac{1+\sqrt{2}}{\gamma_1}\right)\frac{8\gamma_2^{1/2}\bar{k}}{7}\right].$$

En utilisant le fait que $\bar{k}\geqq 7\gamma_1^{-2}\gamma_2^{1/2}$ on obtient finalement

$$\sup_{y\in\mathscr{X}}\int R'(x,\mathscr{X})\,R'(y,dx)\leqq 1-\frac{3\gamma_1\bar{k}^2\varepsilon^2}{7}.$$

Par ailleurs, il est clair que la loi de X_1 a une densité $g_s(x)$ par rapport à μ, comprise entre γ_1 et γ_2 lorsque s est la vraie valeur du paramètre. On peut alors en utilisant l'inégalité exponentielle écrire

$$P_{s',n}\left[\log\frac{dP_{t,n}}{dP_{s,n}}>0\right]\leqq\int\left(\frac{g_t(x_1)}{g_s(x_1)}\right)^{\frac{1}{2}}g_{s'}(x_1)\mu(dx_1)R'(x_1,dx_2)\dots R'(x_{n-1},dx_n).$$

Finalement on obtient à partir de (5.3) si $n=2p+1$

$$P_{s',n}\left[\log\frac{dP_{t,n}}{dP_{s,n}}>0\right]\leqq\left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{1}{2}}\left(1-\frac{3\bar{k}^2\varepsilon^2\gamma_1}{7}\right)^p<\left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{1}{2}}\exp\left[-\frac{3p\gamma_1}{7}\bar{k}^2\varepsilon^2\right],$$

et si $n=2p+2$

$$P_{s',n}\left[\log\frac{dP_{t,n}}{dP_{s,n}}>0\right]\leqq\left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{1}{2}}\left(1+\left(\frac{1+\sqrt{2}}{2\gamma_1}\right)^2\varepsilon^2\right)\left(1-\frac{3\bar{k}^2\varepsilon^2\gamma_1}{7}\right)^p$$
$$<1{,}03\left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{1}{2}}\exp\left[-\frac{3p\gamma_1}{7}\bar{k}^2\varepsilon^2\right].$$

Finalement puisque $d^2(s,t)=\bar{k}^2\varepsilon^2$ et $n\geqq 30$ on a le résultat voulu. □

Si S est un (ε,d')-réseau, ce n'est pas nécessairement un (ε,d)-réseau, mais d'après (4.3) et (4.4), ce sera un ε-réseau pour un multiple de d. Comme tous les résultats sont donnés à des constantes multiplicatives près, remplacer d par un multiple ne change rien. Il s'ensuit que l'on pourra vérifier H4, avec des constantes adéquates dès que $\delta>\tilde{d}'$ en utilisant des (ε,d')-réseaux et les recouvrements associés. Pour les minorations, nous utiliserons le lemme suivant:

Lemme 5.2. *Supposons que l'on ait l'hypothèse $LB(\gamma_1,\gamma_2)$ et*

$$\int\log\frac{p_s(x,y)}{p_t(x,y)}\,p_s(x,y)\mu(dx)\mu(dy)\leqq A, \tag{5.4}$$

alors pour $n\geqq A^{-1}\gamma_2^{-1}\log\frac{\gamma_2}{\gamma_1}$ nous aurons $K(P_{s,n},P_{t,n})\leqq 2n\gamma_2 A$.

Preuve. Nous noterons comme précédemment $g_s(x)=\int p_s(y,x)v_s(dy)$ la densité de la loi de X_1. Nous pouvons écrire

$$K(P_{s,n},P_{t,n})=\sum_{i=1}^{n-1}\int\log\frac{p_s(x_i,x_{i+1})}{p_t(x_i,x_{i+1})}\prod_{j=1}^{n-1}p_s(x_j,x_{j+1})g_s(x_1)\mu(dx_1)\dots\mu(dx_n)$$
$$+\int\log\frac{g_s(x_1)}{g_t(x_1)}\,g_s(x_1)\mu(dx_1).$$

Si l'on considère le i-ème terme I_i et que l'on intègre en x_j, $j\geqq i+2$ on trouve pour $i\geqq 2$

$$I_i=\int\log\frac{p_s(x_i,x_{i+1})}{p_t(x_i,x_{i+1})}\,p_s(x_i,x_{i+1})p_s(x_{i-1},x_i)$$
$$\cdot\prod_{j=1}^{i-2}p_s(x_j,x_{j+1})g_s(x_1)\mu(dx_1)\dots\mu(dx_{i+1})$$
$$\leqq\gamma_2 A\int\prod_{j=1}^{i-2}p_s(x_j,x_{j+1})g_s(x_1)\mu(dx_1)\dots\mu(dx_{i-1})=\gamma_2 A.$$

Le calcul est analogue pour $i=1$ et finalement

$$K(P_{s,n},P_{t,n})\leqq(n-1)\gamma_2 A+\log\frac{\gamma_2}{\gamma_1}.$$

Le résultat s'en déduit immédiatement. □

Il est clair que le 1[er] membre de (5.4) est une information de Kullback intégrée en μ et que grâce aux hypothèses $LB(\gamma_1, \gamma_2)$ elle sera d'après le lemme 4.4 équivalente à $d^2(s,t)$ (qui est une distance de Hellinger intégrée). Ainsi pour vérifier H5, on utilisera un ensemble S' sur lequel $d^2(s,t)$ sera de l'ordre de $[L\delta(n)]^2$ donc A également et $K(P_{s,n}, P_{t,n})$ sera de l'ordre de $[nL\delta(n)]^2$ ce qui est ce que nous cherchons. L'inégalité sur n sera toujours vérifiée en pratique vu que $L(\delta(n))^2$ est grand devant n^{-1}.

La conséquence de ces deux lemmes est que tout se passera exactement comme dans le cas des variables indépendantes, pourvu que l'on puisse travailler avec la distance uniforme comme avec la distance d et que $\tilde{d}'$ et $\tilde{d}$ se ressemblent. C'était le cas en variables indépendantes pour les familles Λ_ω ou $\Lambda^s_{\alpha,r}$. Il en sera de même ici pour leurs analogues bi-dimensionnels. Les d-estimateurs obtenus seront des estimateurs du maximum de vraisemblance sur un réseau. Plutôt que de tenter de faire un panorama des cas possibles, nous traiterons en détail un exemple, pour bien marquer l'analogie avec le cas des variables indépendantes.

Exemple. $\mathscr{X}$ est l'ensemble $[0;1]$, μ la mesure de Lebesgue et Θ est l'ensemble des densités $p(x,y)$ C-lipschitziennes en x et C'-α-hölderiennes en y, satisfaisant en outre $LB(\gamma_1, \gamma_2)$. Nous aurons donc

$$\begin{cases} |p_s(x,y) - p_s(x',y)| \leqq C|x-x'| & \text{avec } C>0 \ \forall y, \\ |p_s(x,y) - p_s(x,y')| \leqq C'|y-y'|^\alpha & 0<\alpha\leqq 1 \ \forall x. \end{cases}$$

La démonstration est analogue à celle faite pour Λ_ω dans le cas indépendant. Nous utiliserons sur $[0;1]^2$ un quadrillage en pq rectangles de côtés p^{-1} et q^{-1}. Nous supposerons que $Cp^{-1} + C'q^{-\alpha} \leqq \varepsilon$. Alors la variation de $p_s(x,y)$ sur un rectangle du quadrillage est inférieure à ε. Si nous considérons les fonctions constantes sur ces rectangles, à valeurs $k\varepsilon$, k entier, nous obtiendrons ainsi un réseau qui approche Θ à ε près pour la distance uniforme. Mais nous n'utiliserons pas tous les points du réseau parce que $p_s(x,y)$ étant une densité en y prend la valeur 1 en au moins un point et que p_s varie d'au plus ε d'un rectangle à un rectangle contigu. Il s'ensuit que $[3^q 2q]^p$ est une majoration du nombre de points d'un ε-réseau. Nous supposerons que $p = \left[\!\left[\frac{2C}{\varepsilon}\right]\!\right] + 1$, $q = \left[\!\left[\left(\frac{2C'}{\varepsilon}\right)^{\frac{1}{\alpha}}\right]\!\right] + 1$. Il s'ensuit que pour ε assez petit $\leqq \varepsilon_0$, p et q sont $\geqq 2$ et que $\frac{\varepsilon}{2} < Cp^{-1} + C'q^{-\alpha} \leqq \varepsilon$. Le nombre de points d'un ε-réseau est donc majoré par $\exp[pq\log 3 + p\log 2q] \leqq \exp(3pq)$, d'où

$$\tilde{d}'(\varepsilon) \leqq \frac{pq}{\log 2} \leqq C_1 \varepsilon^{-\frac{1+\alpha}{\alpha}} \qquad \text{si } \varepsilon \leqq \varepsilon_0,$$

où C_1 est une constante indépendante de ε. Nous poserons donc $\delta(\varepsilon) = C_1 \varepsilon^{-\frac{1+\alpha}{\alpha}}$. Il s'ensuit que l'on a $L\delta(n) \asymp n^{-\frac{\alpha}{3\alpha+1}}$ qui tend vers 0 quand n tend vers l'infini, ce qui fait que pour $n \geqq n_0$, $L\delta(n) \leqq \varepsilon_0$ et le calcul précédent est justifié. Donc d'après la proposition 5.1, H4 sera satisfaite avec des constantes adéquates. Reste à vérifier H5. Pour cela nous utiliserons encore un système de perturbations et le corollaire 3.9 vu que d^2 est suradditive. Définissons les fonctions $\bar{g}$ et $\bar{\bar{g}}$ sur $\mathbb{R}^2$ par

$$\begin{aligned}\bar{g}(x,y)&=(1-2C|x|)(\tfrac{1}{2}-C'|y|^{\alpha}) \text{ pour } |x|\leqq(2C)^{-1},\ |y|\leqq(2C')^{-\frac{1}{\alpha}}\\ &=0 \text{ sinon,}\\ \bar{\bar{g}}(x,y)&=\bar{g}(x,y)-\bar{g}(x,y-2(2C')^{-\frac{1}{\alpha}}).\end{aligned}$$

Ainsi $\bar{\bar{g}}$ a un support dans l'ensemble $\{x,y\,|\,|x|\leqq(2C)^{-1};\ -(2C')^{-\frac{1}{\alpha}}\leqq y\leqq 3(2C')^{-\frac{1}{\alpha}}\}$, est C-lipschitzienne en x et C'-α-hölderienne en y. De plus son intégrale est nulle. Posons finalement $g(x,y)=\eta\bar{\bar{g}}(x\eta^{-1},y\eta^{-\frac{1}{\alpha}})$, avec $\eta>0$. g a les mêmes propriétés que $\bar{\bar{g}}$, mais son support est inclus dans le rectangle

$$|x|\leqq\frac{\eta}{2C},\qquad -\left(\frac{\eta}{2C'}\right)^{\frac{1}{\alpha}}\leqq y\leqq 3\left(\frac{\eta}{2C'}\right)^{\frac{1}{\alpha}}.$$

Ce sont des translatées de g qui nous serviront comme petites perturbations de la densité constante $p_s(x,y)=1$. Choisissons $p=\left[\!\left[\frac{2C}{\eta}\right]\!\right]$, $q=\left[\!\left[\left(\frac{2C'}{\eta}\right)^{\frac{1}{\alpha}}\right]\!\right]$ et le quadrillage associé. Si η est assez petit $\eta\leqq Cp^{-1}+C'q^{-\alpha}<2\eta$. De ce quadrillage, on extrait un quadrillage plus grossier en groupant 8 rectangles élémentaires pour obtenir des rectangles de côtés $\frac{2}{p}$ et $\frac{4}{q}$ en nombre $N=\left[\!\left[\frac{p}{2}\right]\!\right]\times\left[\!\left[\frac{q}{4}\right]\!\right]$. Si $\eta\leqq\eta_0$ assez petit, p et q sont assez grands pour que ce nombre soit supérieur à $\frac{pq}{9}$. On voit que chacun de ces rectangles est plus grand que le support de g. Donc pour chaque rectangle d'indice j nous considérons une translatée g_j de g qui a son support dans ce rectangle. Alors le système $\Theta'=\left\{1+\sum_{j=1}^{N}\lambda_j g_j\,|\,\lambda_j=\pm1\right\}$ est inclus dans Θ et $N\geqq\frac{pq}{9}$.

Nous avons

$$\begin{aligned}\tfrac{1}{2}\int(\sqrt{1+g_j}-\sqrt{1-g_j})^2\,dx\,dy&=\tfrac{1}{2}\int(\sqrt{1+g}-\sqrt{1-g})^2\,dx\,dy\\ &=8\int_0^{\frac{\eta}{2C}}dx\int_0^{\left(\frac{\eta}{2C'}\right)^{1/\alpha}}(\sqrt{1+g(x,y)}-\sqrt{1-g(x,y)})^2\,dy\\ &=8\int_0^{\frac{1}{2C}}\eta\,dz\int_0^{\left(\frac{1}{2C'}\right)^{\alpha^{-1}}}\eta^{\frac{1}{\alpha}}(\sqrt{1+\eta\bar{\bar{g}}(z,w)}-\sqrt{1-\eta\bar{\bar{g}}(z,w)})^2\,dw.\end{aligned}$$

Comme pour x petit $x\leqq\sqrt{1+x}-\sqrt{1-x}\leqq\frac{4}{3}x$, il vient finalement comme $|\bar{\bar{g}}|\leqq\frac{1}{2}$ et η_0 est petit

$$C_2\eta^{3+\frac{1}{\alpha}}\leqq\tfrac{1}{2}\int(\sqrt{1+g_j}-\sqrt{1-g_j})^2\,dx\,dy\leqq\tfrac{4}{3}C_2\eta^{3+\frac{1}{\alpha}}.$$

D'autre part comme $|g_j| \leqq \frac{\eta}{2}$, on peut utiliser l'inégalité (4.5). On obtiendra finalement, toujours avec η_0 petit:

$$\int \log \frac{1+g_j}{1-g_j}(1+g_j)\,dx\,dy \leqq \tfrac{20}{3} C_2 \eta^{3+\frac{1}{\alpha}}.$$

Choisissons alors $\eta = \frac{L\delta(n)}{k}$ avec k entier et n assez grand pour que η soit $\leqq \eta_0$ et que toutes les inégalités précédentes soient vérifiées. On a alors:

$$N \geqq \frac{pq}{9} \geqq C_3 \eta^{-\frac{\alpha+1}{\alpha}} \tag{5.5}$$

$$\sup_{s,t\in\Theta'} K(P_{s,n}, P_{t,n}) \leqq 2\gamma_2 \tfrac{20}{3} C_2 \eta^{3+\frac{1}{\alpha}} n N, \tag{5.6}$$

grâce au lemme 5.2 avec

$$A = \tfrac{20}{3} C_2 \eta^{3+\frac{1}{\alpha}} N \geqq \tfrac{20}{3} C_2 C_3 \eta^2 = \frac{20}{3k^2} C_2 C_3 [L\delta(n)]^2 = C_4 n^{-\frac{2\alpha}{3\alpha+1}}$$

qui est grand devant n^{-1}, si n est grand. Il reste à vérifier (3.16) qui s'écrit ici grâce à (5.5) avec $p_n = N$ (lequel est $\geqq 3$ pour η petit)

$$\frac{N}{8} C_2 \eta^{3+\frac{1}{\alpha}} \geqq \frac{C_2 C_3}{8} \eta^2 = \frac{C_2 C_3}{8} \left(\frac{L\delta(n)}{k}\right)^2,$$

et pour (3.17) il suffit de voir d'après (5.6) que l'on a

$$2\gamma_2 \tfrac{20}{3} C_2 \eta^{3+\frac{1}{\alpha}} n \leqq \tfrac{1}{10}.$$

Comme $\eta^{3+\frac{1}{\alpha}} = n^{-1} k^{-\left(3+\frac{1}{\alpha}\right)}$, ce sera vrai si k a été choisi convenablement. Le corollaire 3.9 implique alors que H5 est satisfaite. On en déduit que le théorème 3.4 s'applique donnant pour n grand la valeur du risque $R_n(d^q) \asymp n^{-\frac{\alpha q}{3\alpha+1}}$.

Dans cet exemple, l'hypothèse $C>0$ est fondamentale. $C=0$ signifie que $p_\theta(x,y)$ est indépendant de x. C'est le cas des variables indépendantes pour lequel on trouverait comme on l'a vu un risque en $n^{-\frac{q\alpha}{1+2\alpha}}$. De même, il est essentiel de supposer $\alpha>0$.

On pourrait traiter de la même façon d'autres exemples où $p_\theta(x,y)$ appartient à une classe Λ_ω ou $\Lambda^1_{\alpha,r}$ si une variable est fixée et une autre classe de ce type si l'autre variable est fixée, l'exemple précédent n'étant qu'une illustration. La seule restriction est toujours la condition $LB(\gamma_1,\gamma_2)$ dont il est malheureusement actuellement impossible de se débarasser.

Remarque 1. Le lemme 5.2 montre que $K(P_{s,n}, P_{t,n}) \geqq C' n A$. Mais la démonstration implique également que $K(P_{s,n}, P_{t,n}) \geqq C' n A$ avec une autre constante C'. De plus sur les systèmes de petites perturbations, il est facile de voir en utilisant le lemme 4.4 que le A de (5.4) est de l'ordre de $4d^2(s,t)$. Donc finalement, on aura certainement sur les systèmes de petites perturbations (3.7) et on pourra appliquer la proposition 3.5 soit $L\delta(n) \asymp L\tilde{d}(n)$.

Remarque 2. Cet exemple nous a fait étudier un cas de fonctions de 2 variables, lipschitziennes en x, α-hölderiennes en y. Pour cette classe de fonctions on a majoré $\tilde{d}$ par δ, calculé $L\delta$, et exhibé des systèmes de petites perturbations de taille $cL\delta$. Il est bien clair que le fait que le cadre soit markovien n'a joué là aucun rôle puisque grâce aux lemmes 5.1 et 5.2 tout s'est passé comme si les variables étaient indépendantes. En fait on aurait aussi bien pu traiter l'exemple en variables indépendantes, avec une distance h ou $\mathbb{L}^p (1 \leqq p \leqq 2)$ et des densités sur le carré $[0;1]^2$. Les résultats seraient strictement les mêmes, (et aussi en remplaçant $[0;1]^2$ par un rectangle quelconque). On pourrait de même traiter des exemples multi-dimensionnels avec par exemple un certain module de continuité en chaque variable.

Remarque 3. Des travaux récents de Birgé ont amélioré sensiblement le lemme 5.1 permettant de travailler avec les chaînes de Markov comme avec les variables indépendantes sans utiliser la distance uniforme. On pourra trouver les détails dans [7].

Les processus gaussiens stationnaires

Nous voulons à nouveau démontrer les analogues des lemmes 5.1 et 5.2 mais ici le problème est plus difficile. Dans le cas des chaînes de Markov, les démonstrations consistaient, grâce à quelques artifices, à se ramener au cas indépendant. Ici, cela n'est plus possible et il faudra travailler directement sur les rapports de vraisemblances et surtout sur les matrices de covariances.

Si f est une densité spectrale élément de Θ, $P_{f,n}$ est la loi gaussienne des n premières observations, de densité $p_{f,n}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ et la matrice de covariance associée à ces n observations s'écrit $T_n f$. Ses coefficients sont des coefficients de Fourier de f et l'on a

$$p_{f,n}(X) = (2\pi)^{-\frac{n}{2}} |T_n f|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2} X^t (T_n f)^{-1} X\right].$$

Les démonstrations vont nécessiter des calculs sur les matrices symétriques, ce qui nous amène à faire quelques rappels: $\|A\|$ désigne la norme de l'opérateur linéaire de $\mathbb{R}^n$ associé, soit le sup des valeurs absolues des valeurs propres. On définit un ordre sur les matrices symétriques en posant $A \leqq B$ lorsque $B-A$ est positive. Si f est positive, $T_n f$ est une matrice positive. Il s'ensuit que T_n est un opérateur linéaire et positif. Nous utiliserons sans rappel les propriétés de cet ordre sur les matrices (cf. [4]) ainsi que les lemmes faciles suivants:

Lemme 5.3. *Soient A et $B \geqq 0$, $\gamma_2 I \geqq B \geqq \gamma_1 I$, $\gamma_1 > 0$. Si $\{a_i\}_{i \leqq n}$ et $\{c_i\}_{i \leqq n}$ sont les valeurs propres (en ordre croissant) de A et de $C = \sqrt{B}\, A \sqrt{B}$, alors $\gamma_2 a_i \geqq c_i \geqq \gamma_1 a_i$ pour $1 \leqq i \leqq n$.*

Lemme 5.4. *Soient A et B symétriques, $\{a_i\}_{i\leqq n}$ les valeurs propres de A, et $\|B\|\leqq\gamma$ alors*

$$\operatorname{Tr}(AB)\leqq\gamma\sum_{i=1}^{n}|a_i|.$$

Lemme 5.5. *Soit A une matrice symétrique de valeurs propres $\{a_i\}_{i\leqq n}$ ordonnées dans l'ordre croissant des valeurs absolues, $B\geqq 0$ avec $\gamma_1 I\leqq B\leqq\gamma_2 I$, $\gamma_1>0$. Posons $C=\sqrt{B}A\sqrt{B}$. Si $\{c_i'\}_{i\leqq n}$ sont les valeurs propres de C^2, alors*

$$\gamma_1^2 a_i^2\leqq c_i'\leqq\gamma_2^2 a_i^2, \quad et \quad \gamma_1^2\operatorname{Tr}A^2\leqq\operatorname{Tr}C^2\leqq\gamma_2^2\operatorname{Tr}A^2.$$

Nous pouvons maintenant démontrer les deux résultats principaux permettant de vérifier les hypothèses H4 et H5:

Proposition 5.5. *Si l'ensemble Θ vérifie $LB(\gamma_1,\gamma_2)$ et s' il existe n_0, δ, k_1 tels que pour tout f dans Θ on a*

$$\sum_{j=-\infty}^{-[\![\frac{n}{2}]\!]-1} a_j^2+\sum_{j=[\![\frac{n}{2}]\!]+1}^{+\infty} a_j^2\leqq k_1^2\varepsilon^2, \tag{5.7}$$

avec $a_j=\frac{1}{2\pi}\int_0^{2\pi}e^{-ijx}f_\theta(x)dx$ et $\varepsilon=L\delta(n)\leqq\frac{\sqrt{5}\gamma_1}{3}$, $n\geqq n_0$, alors il existe deux constantes k_0 et A_2 indépendantes de n et ε telles que si $\|g-g'\|_\infty\leqq\varepsilon$ et $\|f-g\|_2\geqq k_0\varepsilon$ avec f, g, g' dans Θ

$$P_{g',n}\left[\log\frac{P_{f,n}}{P_{g,n}}\geqq 0\right]\leqq\exp\left[-nA_2\|f-g\|_2^2\right].$$

Preuve. De l'inégalité exponentielle on obtient

$$P_{g',n}\left[\log\frac{P_{f,n}}{P_{g,n}}\geqq 0\right]\leqq\int\left(\frac{p_{f,n}}{p_{g,n}}\right)^{\frac{1}{2}}p_{g',n}dx_1\ldots dx_n.$$

Un calcul facile montre que ce second membre peut s'écrire

$$\rho'=\frac{|T_n g|^{\frac{1}{4}}\left|\frac{(T_n f)^{-1}+2(T_n g')^{-1}-(T_n g)^{-1}}{2}\right|^{-\frac{1}{2}}}{|T_n f|^{\frac{1}{4}}|T_n g'|^{\frac{1}{2}}}.$$

Nous verrons que le numérateur est bien défini. Comme $T_n g\geqq\gamma_1 I$, on peut poser

$$\Delta'=(T_n g)^{-\frac{1}{2}}\Delta(T_n g)^{-\frac{1}{2}}, \qquad \Omega'=(T_n g)^{-\frac{1}{2}}\Omega(T_n g)^{-\frac{1}{2}},$$

$$T_n f=T_n g+\Delta=(T_n g)^{\frac{1}{2}}[I+\Delta'](T_n g)^{\frac{1}{2}}, \qquad T_n g'=T_n g+\Omega=(T_n g)^{\frac{1}{2}}[I+\Omega'](T_n g)^{\frac{1}{2}},$$

$$\Gamma=(I+\Delta')^{-\frac{1}{2}}+(I+\Delta')^{\frac{1}{2}}, \qquad \Lambda=\tfrac{1}{2}[\Gamma+\Omega'(I+\Delta')^{-\frac{1}{2}}-\Omega'(I+\Delta')^{\frac{1}{2}}].$$

On peut alors voir que $\rho' = |A|^{-\frac{1}{2}}$. Notons $\{\delta'_i\}_{i \leqq n}$ les valeurs propres ordonnées de Δ' $I + \Delta' = (T_n g)^{-\frac{1}{2}} T_n f (T_n g)^{-\frac{1}{2}}$ et donc d'après le lemme 5.3

$$\gamma_1 \gamma_2^{-1} \leqq 1 + \delta'_i \leqq \gamma_1^{-1} \gamma_2, \qquad 1 \leqq i \leqq n. \tag{5.8}$$

Soit alors x_i un vecteur propre de Δ' associé à δ'_i, de norme 1,

$$A(x_i) = \tfrac{1}{2}[((1+\delta'_i)^{\frac{1}{2}} + (1+\delta'_i)^{-\frac{1}{2}}) x_i - ((1+\delta'_i)^{\frac{1}{2}} - (1+\delta'_i)^{-\frac{1}{2}}) \Omega'(x_i)].$$

Comme $\Omega = T_n(g' - g)$, on a $-\varepsilon I \leqq \Omega \leqq \varepsilon I$ et donc $\|\Omega'\| \leqq \dfrac{\varepsilon}{\gamma_1} = \varepsilon'$; il s'ensuit que si l'on pose $\gamma_1^{-1} \gamma_2 = C_1^2$,

$$\inf_{t \geqq 1} \{(t + t^{-1}) - (t - t^{-1})\} \varepsilon' \leqq \|2A(x_i)\| \leqq (C_1 + C_1^{-1}) + \varepsilon'(C_1 - C_1^{-1})$$

et finalement vu que le minimum du premier membre est obtenu pour $t = \left(\dfrac{1+\varepsilon'}{1-\varepsilon'}\right)^{\frac{1}{2}}$ et que $\varepsilon' \leqq \dfrac{\sqrt{5}}{3}$, il existe $C_2 > 0$ tel que $(1 - \varepsilon'^2)^{\frac{1}{2}} \leqq \|A(x_i)\| \leqq 1 + C_2$. Comme les $\{x_i\}_{i \leqq n}$ forment une base orthonormée, il s'ensuit que les valeurs propres $\{\lambda_i\}_{i \leqq n}$ de A satisfont à

$$\tfrac{2}{3} \leqq (1 - \varepsilon'^2)^{\frac{1}{2}} \leqq \lambda_i \leqq 1 + C_2. \tag{5.9}$$

Ceci implique que ρ' est bien défini ainsi qu'il était annoncé. D'autre part, l'application du lemme 5.4 entraîne

$$\begin{aligned} \operatorname{Tr}(A) &= \tfrac{1}{2}[\operatorname{Tr}(\Gamma) + \operatorname{Tr}[\Omega'((I+\Delta')^{-\frac{1}{2}} - (I+\Delta')^{\frac{1}{2}})]] \\ &\geqq \tfrac{1}{2}\operatorname{Tr}(\Gamma) - \frac{\varepsilon'}{2} \sum_{i=1}^{n} |(1+\delta'_i)^{-\frac{1}{2}} - (1+\delta'_i)^{\frac{1}{2}}|. \end{aligned}$$

Comme on a les bornes (5.8) sur $1 + \delta'_i$, il s'ensuit qu'il existe une constante C_3 telle que:

$$\operatorname{Tr}(A) \geqq \tfrac{1}{2}\operatorname{Tr}(\Gamma) - C_3 \varepsilon' \sum_{i=1}^{n} |\delta'_i|.$$

Les mêmes bornes (5.8) sur $1 + \delta'_i$ entraînent que l'on a pour une constante C_4

$$\tfrac{1}{2}[(1+\delta'_i)^{\frac{1}{2}} + (1+\delta'_i)^{-\frac{1}{2}}] \geqq 1 + C_4 \delta_i'^2$$

et donc que $\frac{1}{2}\operatorname{Tr}(\Gamma) \geqq n + C_4 \sum_{i=1}^{n} \delta_i'^2$, d'où

$$\operatorname{Tr}(A) \geqq n + C_4 \sum_{i=1}^{n} \delta_i'^2 - C_3 \varepsilon' \sum_{i=1}^{n} |\delta'_i|.$$

Si l'on note $\{\delta_i\}_{i \leqq n}$ les valeurs propres de Δ, on sait que l'on a d'après le lemme 5.5

$$\sum_{i=1}^{n} \delta_i'^2 \geqq \gamma_2^{-2} \sum_{i=1}^{n} \delta_i^2 = \gamma_2^{-2} \operatorname{Tr}(\Delta^2).$$

Soit c_{ii} le i-ème coefficient diagonal de Δ^2, on pourra écrire puisque $\Delta = T_n(f-g)$,

$$c_{ii} = \sum_{k=1}^{n} b_{i-k}^2 \quad \text{avec } b_j = b_{-j} = \frac{1}{2\pi} \int_0^{2\pi} e^{-ijx}[f(x)-g(x)]\,dx.$$

D'où

$$\operatorname{Tr}(\Delta^2) = \sum_{i=1}^{n} c_{ii} = \sum_{i=1}^{n} \sum_{j=i-1}^{i-n} b_j^2 \geqq \frac{n}{2} \sum_{j=-[\![\frac{n}{2}]\!]}^{[\![\frac{n}{2}]\!]} b_j^2. \tag{5.10}$$

Comme d'après l'hypothèse $\sum_{j=-\infty}^{-[\![\frac{n}{2}]\!]-1} b_j^2 + \sum_{j=[\![\frac{n}{2}]\!]+1}^{+\infty} b_j^2 \leqq 4k_1^2\varepsilon^2$, on obtient

$$\operatorname{Tr}(\Delta^2) \geqq \frac{n}{2}[\|f-g\|_2^2 - 4k_1^2\varepsilon^2];$$

si $k_0 \geqq 3k_1$, on trouve finalement,

$$\sum_{i=1}^{n} \delta_i'^2 \geqq \gamma_2^{-2} \operatorname{Tr}\Delta^2 \geqq (2\gamma_2)^{-2} n \|f-g\|_2^2.$$

Nous poserons $\|f-g\|_2 = k\varepsilon$ et $k' = \dfrac{\gamma_1 k}{2\gamma_2}$. Dans ce cas $\sum_{i=1}^{n} \delta_i'^2 \geqq k'^2 \varepsilon'^2 n$. D'autre part on a $\left(\sum_{i=1}^{n} |\delta_i'|\right)^2 \leqq n \sum_{i=1}^{n} \delta_i'^2$, ce qui entraîne que pour $k' \geqq \dfrac{2C_3}{C_4}$

$$C_4 \sum_{i=1}^{n} \delta_i'^2 - C_3 \varepsilon' \sum_{i=1}^{n} |\delta_i'| \geqq \frac{C_4}{2} \sum_{i=1}^{n} \delta_i'^2.$$

Ceci sera assuré si l'on choisit $k_0 \geqq \dfrac{4C_3\gamma_2}{C_4\gamma_1}$ et alors

$$\operatorname{Tr}(\Lambda) \geqq n + \frac{C_4}{2} k'^2 \varepsilon'^2 n.$$

D'autre part $\operatorname{Tr}(\Lambda) = \sum_{i=1}^{n} \lambda_i$, lesquels λ_i sont bornés par (5.9). Si $\frac{2}{3} \leqq u \leqq 1$, $\log u \geqq \frac{5}{4}(u-1)$ et il existe une constante C_5 telle que pour $1 \leqq u \leqq 1 + C_2$, $\log u \geqq C_5(u-1)$. On a donc

$$\begin{aligned}\log|\Lambda| = \sum_{i=1}^{n} \log(\lambda_i) &\geqq C_5 \sum_{\lambda_i \geqq 1} (\lambda_i - 1) + \tfrac{5}{4} \sum_{\lambda_i < 1} (\lambda_i - 1) \\ &\geqq C_5 \sum_{i=1}^{n} (\lambda_i - 1) + (\tfrac{5}{4} - C_5) \sum_{\lambda_i < 1} (\lambda_i - 1).\end{aligned}$$

Comme $\sum_{\lambda_i < 1} (\lambda_i - 1) \geqq -n[1-(1-\varepsilon'^2)^{\frac{1}{2}}] \geqq -n\varepsilon'^2$, on obtient

$$\log|\Lambda| \geqq C_5[\operatorname{Tr}\Lambda - n] - (\tfrac{5}{4} - C_5)n\varepsilon'^2 \geqq C_5 \frac{C_4}{2} k'^2 \varepsilon'^2 n - (\tfrac{5}{4} - C_5)n\varepsilon'^2.$$

En supposant que $\frac{C_4 C_5}{2} k'^2 \geqq 3(\frac{5}{4} - C_5)$, ce qui sera vérifié si $\frac{C_4 C_5 \gamma_1^2 k_0^2}{4\gamma_2^2} \geqq 3(\frac{5}{4} - C_5)$ c'est-à-dire si k_0 est assez grand, on a

$$\log|A| \geqq \frac{C_4 C_5}{3} k'^2 \varepsilon'^2 n = \frac{C_4 C_5}{12\gamma_2^2} n k^2 \varepsilon^2.$$

En revenant à notre point de départ, nous trouvons finalement

$$P_{g',n}\left[\log \frac{P_{f,n}}{P_{g,n}} \geqq 0\right] \leqq \exp[-\tfrac{1}{2}\log|A|] \leqq \exp\left[-\frac{C_4 C_5}{24\gamma_2^2} n k^2 \varepsilon^2\right],$$

ce qui est le résultat voulu. □

Lemme 5.6. *Supposons que Θ vérifie $BI(\gamma_1)$ et que f et g soient tels que $\|f - g\|_\infty \leqq \frac{\gamma_1}{2}$. Alors on a*

$$K(P_{f,n}, P_{g,n}) \leqq \frac{3}{5\gamma_1^2} n \|f-g\|_2^2.$$

Preuve. Nous avons d'après l'inégalité de Jensen

$$\begin{aligned} K(P_{f,n}, P_{g,n}) &= \int_{\mathbb{R}^n} \log\left(\frac{p_{f,n}}{p_{g,n}}\right) p_{f,n} dx_1, \ldots, dx_n \\ &\leqq \log \int_{\mathbb{R}^n} \frac{p_{f,n}^2}{p_{g,n}} dx_1, \ldots, dx_n, \end{aligned}$$

et l'on trouve facilement

$$\begin{aligned} \int_{\mathbb{R}^n} \frac{p_{f,n}^2}{p_{g,n}} dx_1 \ldots dx_n &= |T_n g|^{\frac{1}{2}} |T_n f|^{-1} |2(T_n f)^{-1} - (T_n g)^{-1}|^{-\frac{1}{2}} \\ &= |(T_n g)^{-1}|^{-\frac{1}{2}} |2 T_n f - T_n f (T_n g)^{-1} T_n f|^{-\frac{1}{2}}. \end{aligned}$$

Posons $T_n f = T_n g + \Delta = (T_n g)^{\frac{1}{2}}[I + \Delta'](T_n g)^{\frac{1}{2}}$, $\Delta' = (T_n g)^{-\frac{1}{2}} \Delta (T_n g)^{-\frac{1}{2}}$. Alors

$$\begin{aligned} I - \Delta'^2 &= (T_n g)^{-\frac{1}{2}}[T_n g - \Delta (T_n g)^{-1} \Delta](T_n g)^{-\frac{1}{2}} \\ &= (T_n g)^{-\frac{1}{2}}[2 T_n f - T_n f (T_n g)^{-1} T_n f](T_n g)^{-\frac{1}{2}}. \end{aligned}$$

Donc $K(P_{f,n}, P_{g,n}) \leqq -\frac{1}{2}\log|I - \Delta'^2|$. Comme $-\frac{\gamma_1}{2} I \leqq T_n(f-g) \leqq \frac{\gamma_1}{2} I$, toutes les valeurs propres de Δ ont leurs valeurs absolues inférieures à $\frac{\gamma_1}{2}$. Il s'ensuit par le lemme 5.5 que les valeurs propres $\delta_i'^2$ de Δ'^2 sont dans l'intervalle $[0; \frac{1}{4}]$ et que l'on a

$$\sum_{i=1}^n \delta_i'^2 = \operatorname{Tr} \Delta'^2 \leqq \gamma_1^{-2} \operatorname{Tr} \Delta^2 = \gamma_1^{-2} \sum_{i=1}^n \delta_i^2,$$

puisque $(T_n g)^{-1} \leqq \gamma_1^{-1} I$. Donc

$$\log|I-\Delta'^2|=\sum_{i=1}^{n}\log(1-\delta_i'^2)\geqq -\tfrac{6}{5}\sum_{i=1}^{n}\delta_i'^2\geqq -\frac{6}{5\gamma_1^2}\operatorname{Tr}\Delta^2$$

et $K(P_{f,n},P_{g,n})\leqq\frac{3}{5\gamma_1^2}\operatorname{Tr}\Delta^2$. Mais le calcul déjà fait en (5.10) nous indique que $\operatorname{Tr}\Delta^2\leqq n\|f-g\|_2^2$ d'où le résultat. □

Pour obtenir H4 à partir de la proposition 5.5, nous procéderons exactement comme dans le cas des variables indépendantes en choisissant une fonction δ qui majore $\tilde{d}'$ où d' est la distance uniforme sur $[0;2\pi[$. Ici, il y a apparemment un problème supplémentaire en raison de (5.7). Mais sur les exemples classiques Λ_ω et $\Lambda^1_{\alpha,r}$ il ne se posera pas, du fait des propriétés d'approximation des fonctions de ces classes par les polynômes trigonométriques, qui font que si Θ est approché à ε près par les polynômes trigonométriques d'ordre $p, \tilde{d}'(\varepsilon)$ est (à des constantes multiplicatives près) p. Comme on s'est placé dans une situation où $\varepsilon=L\delta(n)<1$, $p\varepsilon<\tilde{d}'(\varepsilon)$ et $L\delta(p)>\varepsilon$ donc $p<n$ (souvent très inférieur si ε est petit) l'approximation de Θ par les polynômes trigonométriques de degré n est meilleure (et bien meilleure si ε est petit) que ε (à des constantes multiplicatives près) au sens de la distance uniforme mais bien sûr également en norme $\mathbb{L}^2$. Comme la meilleure approximation en distance $\mathbb{L}^2$ d'une fonction par un polynôme trigonométrique est obtenue en prenant le début de son développement de Fourier, l'inégalité (5.7) sera vérifiée car dans ces exemples, $\tilde{d}'(\varepsilon)$ a un comportement polynomial et l'approximation par les polynômes de degré $\frac{n}{2}$ est à une constante multiplicative près la même que celle obtenue avec des polynômes de degré n.

Nous traiterons ici plus particulièrement la classe Λ_ω sur $[0;2\pi[$ identifié au cercle trigonométrique (c'est-à-dire que nous considérons les fonctions sur $[0;2\pi[$ telles que $f(0)=f(2\pi)$). Nous supposerons en outre que $\omega(0)=0$. Nous avons calculé $L\delta(n)$ en (4.8). D'autre part, nous savons que les polynômes trigonométriques de degré $(p-1)$ approchent les fonctions de Λ_ω à $\omega\left(\frac{\pi}{p}\right)$ près (cf. [28]). Soit donc $L\delta(n)=2\omega(t_n)$ avec $t_n\omega^2(t_n)=\frac{3}{8n}$. Comme $\omega(0)=0$, nous pouvons supposer, lorsque n est assez grand que $\omega(t_n)\leqq\frac{1}{4}$, $t_n\leqq\pi$. Alors il existe $p\geqq 1$ tel que $\frac{\pi}{p+1}<t_n\leqq\frac{\pi}{p}$. Il s'ensuit vu la croissance de ω que $\frac{\pi}{p}\,\omega^2\left(\frac{\pi}{p}\right)\geqq\frac{3}{8n}$ et donc que $\frac{n}{2}\geqq p$. Finalement l'approximation par les polynômes trigonométriques d'ordre $\left[\!\left[\frac{n}{2}\right]\!\right]$ est meilleure que $\omega\left(\frac{\pi}{p}\right)\leqq\omega(2t_n)\leqq 2\omega(t_n)$ d'après la concavité de ω et l'inégalité (5.7) sera vérifiée avec $k_1=1$.

Pour ce qui est de la minoration, il n'y a aucune difficulté en utilisant le lemme 5.6 la démonstration est strictement la même que dans le cas des variables indépendantes à ceci près que la formule (4.9) doit être modifiée d'un facteur $\frac{3}{5\gamma_1^2}$, ce qui amène à choisir non plus $k\geqq 2{,}5$ mais $k\geqq\frac{3}{2\gamma_1^2}$. Ainsi tout se passe comme dans le cas des variables indépendantes, et il en serait de même pour la classe $\Lambda^1_{\alpha,r}$.

6. Possibilités de la théorie métrique; exemples et contre-exemples

En considérant les nombreux exemples précédemment traités nous pourrions penser que cette théorie métrique est sans défaut et que nous avons découvert une famille d'estimateurs «universels» qui, au moins dans le cas de variables indépendantes équidistribuées permettrait de traiter la plupart des problèmes ayant une structure métrique régulière et aussi sous des hypothèses LB bien des problèmes markoviens ou gaussiens stationnaires. En fait, il n'en est rien et les raisons en sont bien simples. Il n'y a qu'à regarder la démonstration du théorème 2.4 ou le lemme 2.7 pour s'en rendre compte. Pour obtenir les majorations, on a simplement ajouté les erreurs $P(A_i)$ des divers tests en remplaçant $P(\bigcup_i A_i)$ par $\sum_i P(A_i)$, faisant comme si les ensembles A_i étaient disjoints. Cela est faux, mais comme nous l'avons vu, cette approximation est souvent suffisante, ce qui veut dire que ces ensembles ne se recouvrent pas trop. Toutefois, dans certains cas, l'inverse se produira et de nombreux A_i seront identiques. Les majorations obtenues n'auront alors plus rien à voir avec la vitesse réelle du problème. Ceci n'implique déailleurs rien quant à la qualité des estimateurs, c'est simplement la méthode de calcul qui est en cause. Nous en verrons un exemple un peu plus loin.

Pour ce qui est des minorations, il est facile aussi de voir que l'introduction de l'information de Kullback est assez artificielle (mais indispensable) et liée aux bonnes propriétés de la fonction logarithme; le fait que l'information soit éventuellement infinie ôte tout caractère général à l'inégalité (2.8) qui n'est utilisable que lorsque l'information de Kullback ressemble au carré de la distance de Hellinger. Il serait bien sûr préférable de travailler toujours avec h, mais il est impossible avec la seule distance de Hellinger d'obtenir un analogue du lemme 2.7.

Le contre-exemple qui suit illustrera ces faits, en ce sens qu'il s'agit là d'un cas où la fonction $\tilde{d}$ est absolument arbitraire (éventuellement égale à $+\infty$) et où pourtant l'estimation a lieu à la vitesse optimale de $n^{-1/2}$. Dans ce cas, ni les majorations, ni les minorations précédemment obtenues ne sont d'aucune utilité, les majorations ne correspondent plus à rien et les informations de Kullback sont infinies. Pourtant, il existe des suites adéquates de d-estimateurs, ce qui prouve que seuls les calculs sont ici inadaptés.

Nous noterons $\mathbb{N}^* = \mathbb{N} - \{0\}$ et $\bar{\mathbb{N}} = \mathbb{N}^* \cup \{+\infty\}$. Soit f une fonction de $\mathbb{N}^*$ dans $\bar{\mathbb{N}}$ et I_n l'ensemble $\{1; 2 \ldots; f(n)\}$ lorsque $f(n) < +\infty$, sinon $I_n = \mathbb{N}^*$.

Nous choisissons pour espace Θ le produit des I_n et pour $\mathscr{X}$ la somme directe des sections commençantes de Θ, c'est-à-dire

$$\Theta = \prod_{n=1}^{+\infty} I_n, \qquad \mathscr{X} = \bigoplus_{n=1}^{+\infty} \left[\prod_{j=1}^{n} I_j\right].$$

$\mathscr{X}$ est dénombrable et $\mathscr{A}$ est la tribu des parties de $\mathscr{X}$. Notons p_n la projection de Θ sur $\prod_{j=1}^{n} I_j$, Si θ est un point de Θ, nous noterons θ_n pour $p_n(\theta)$. Pour tout n, θ_n est un point de $\mathscr{X}$ et nous pouvons considérer les probabilités P_θ associées à Θ: $P_\theta = \sum_{j=1}^{+\infty} 2^{-j} \delta_{\theta_j}$ où δ_x désigne la masse unité au point x.

Proposition 6.1. *Pour un choix convenable de f, $\tilde{h}$ peut prendre des valeurs arbitrairement élevées ($+\infty$ inclus). Pourtant si l'on cherche à estimer θ à partir de n variables indépendantes de loi P_θ, on obtient*

$$an^{-1} \leqq R_n(h^2) < n^{-1}, \qquad a > 0. \tag{6.1}$$

De plus, il existe une suite adéquate de h-estimateurs $\hat{\theta}_n$, tels que

$$\sup_{n \geqq 1} \sup_{\theta \in \Theta} \mathbb{E}_\theta[nh^2(\theta, \hat{\theta}_n)] \leqq 1.$$

Preuve. Sur l'espace $\Theta \times \Theta$ on définit la fonction v à valeurs dans $\bar{\mathbb{N}}$ par

$$v(\theta, \theta') = \inf\{j \mid \theta_j \neq \theta'_j\} \qquad (= +\infty \text{ si } \theta = \theta').$$

Elle vérifie

$$v(\theta, \theta'') \geqq \sup[v(\theta, \theta'), v(\theta', \theta'')].$$

Sur Θ, les distances $\mathbb{L}^1$ et de Hellinger sont reliées à v par

$$\|P_\theta - P_{\theta'}\|_1 = 2h^2(\theta, \theta') = 2 \sum_{j=v(\theta,\theta')}^{+\infty} 2^{-j} = 2^{2-v(\theta,\theta')}.$$

Il est immédiat de construire des 2^{-q}-réseaux minimaux pour la distance $\mathbb{L}^1$, comme suit. Soit r_q un inverse de p_q, c'est-à-dire que $p_q \circ r_q$ est l'identité sur $J_q = \prod_{j=1}^{q} I_j$. Alors l'image de r_q est un 2^{-q+1}-réseau S_q et deux points distincts θ, θ' de S_q sont à distance supérieure à 2^{-q+2}. Il s'ensuit que $\tilde{h}(2^{-\frac{q}{2}}) \geqq \dfrac{f(q)}{3 \log 2}$. Donc $\tilde{h}$ est arbitrairement grand.

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi P_θ. Pour tout i, X_i est l'une des projections θ_q de θ sur un certain J_q. Posons $q(i) = q$ si X_i appartient à J_q et $k(X_1, \ldots, X_n) = \sup_{1 \leqq i \leqq n} q(i)$. Alors $\hat{\theta}(X_1, \ldots, X_n)$ est un estimateur du maximum de vraisemblance si $p_k(\hat{\theta}) = X_i$ et $q(i) = k(X_1, \ldots, X_n)$. Dans ce cas

$$\|P_{\hat{\theta}} - P_\theta\|_1 \leqq 2^{1-k(X_1,\ldots,X_n)} \quad P_\theta \text{ p.s.}$$

Il s'ensuit que

$$\mathbb{E}_\theta[nh^2(\theta, \hat{\theta})] \leqq n \sum_{j=1}^{+\infty} P_\theta[k(X_1, \ldots, X_n) = j] 2^{-j}.$$

Mais

$$P_\theta[k(X_1, \ldots, X_n) \leqq j] = \left[P_\theta\left(X_1 \in \sum_{i=1}^{j} J_i\right)\right]^n = (1 - 2^{-j})^n.$$

Finalement

$$\mathbb{E}_\theta[nh^2(\theta, \hat{\theta})] \leqq n \sum_{j=1}^{+\infty} (1 - 2^{-j})^n 2^{-j-1} \leqq n \sum_{j=1}^{+\infty} 2^{-j-1} \exp(-n 2^{-j}).$$

En utilisant le fait que

$$2^{-j-1} \exp(-n 2^{-j}) < \log 2 \int_j^{j+1} 2^{-t} \exp(-n 2^{-t}) dt$$

on obtient

$$\mathbb{E}_\theta[nh^2(\theta, \hat{\theta})] < n \log 2 \int_1^{+\infty} 2^{-t} \exp[-n 2^{-t}] dt = 1 - e^{-\frac{n}{2}}.$$

Par ailleurs, si S_q est un $(2^{-\frac{q}{2}}, h)$-réseau construit comme précédemment, il est facile de voir que tout test du maximum de vraisemblance au seuil zéro entre deux points θ, θ' de S_q à distance $2^{-\frac{r}{2}}$, $r<q$, est un test entre les boules de centres θ et θ' et rayons $2^{-\frac{q}{2}}$ dont les erreurs sont majorées par $P_\theta[k(X_1, \dots, X_n) \leqq r]$ c'est-à-dire $(1-2^{-r})^n$. Ces erreurs sont inférieures à exp($-nh^2(\theta, \theta')$) et l'hypothèse H2 est vérifiée par cette famille de tests. Il s'ensuit qu'un estimateur du maximum de vraisemblance sur un $(2^{-\frac{q}{2}}, h)$-réseau est un h-estimateur. Il est facile de voir que si $q=10n$, par exemple, le calcul précédent reste approximativement valable pour le h-estimateur vu que $n2^{-q-1}P_\theta[k(X_1, \dots, X_n)>q]$ est environ $n^2 2^{-20n-1}$ ce qui est négligeable et donc que le h-estimateur est aussi bon que le maximum de vraisemblance. La minoration dans (6.1) provient de (1.1), les conditions nécessaires étant ici bien évidemment vérifiées. Ceci achève la démonstration. □

Remarque. Bien que les conditions de dimension soient ici arbitraires, on arrive à trouver de bons h-estimateurs. Mais cela suppose qu'on ne tient aucun compte de la fonction $L\tilde{h}$. Ici on a arbitrairement décidé que pour n observations on construisait l'estimateur à partir d'un 2^{-5n}-réseau. En fait plus le réseau est petit, meilleur sera le résultat. C'est tout à fait contraire à ce qui se passe dans d'autres cas où l'on a intérêt quand la dimension est grande à considérer des réseaux moins fins, de l'ordre de $L\tilde{d}(n)$. Ici la dimension n'a aucune importance, tout se passe comme si elle était 1 parce que c'est uniquement k qui détermine les erreurs de tous les tests et celles-ci ne s'ajoutent absolument pas.

Un autre problème se pose, c'est celui des vitesses d'estimation possibles. Les nombreux exemples précédents montrent que, outre la vitesse classique $n^{-\frac{1}{2}}$ qui est d'après (1.1) la meilleure possible, on peut rencontrer des vitesses en $n^{-\alpha}$ avec $\alpha<\frac{1}{2}$. Un certain nombre de questions se posent à ce propos: toutes les fonctions à décroissance plus lente que $n^{-\frac{1}{2}}$ sont-elles des vitesses possibles, pour ces vitesses quelles sont alors les performances des d-estimateurs, enfin existe-t-il des vitesses pour lesquelles les d-estimateurs ne peuvent pas être adéquats?

Nous pourrions d'abord supposer que les classes Λ_ω, pour diverses fonctions ω, nous donnent toutes les vitesses possibles, mais il n'en est rien en raison de la concavité de ω. Par exemple si $\omega(0)=0$, elle implique que dans (4.8) on obtient pour $p>n$ $t_p \leqq \left(\frac{n}{p}\right)^{\frac{1}{3}} t_n$ et donc que $L\delta(n) \geqq \left(\frac{3}{2t_1}\right)^{\frac{1}{2}} n^{-\frac{1}{3}}$ et finalement que $R_n(d^q) \geqq Cn^{-\frac{q}{3}}$. Afin d'obtenir toutes les vitesses raisonnables et de répondre aux questions précédentes, nous allons construire sur $\mathbb{R}^+$ des modèles de perturbations d'un type particulier qui constitueront des modèles universels pour la vitesse d'estimation.

$\mathscr{X}$ sera $\mathbb{R}^+$ muni de ses boréliens et λ la mesure de Lebesgue. Nous considérons des suites de variables indépendantes équidistribuées de densités f_θ par rapport à λ. La famille $\{f_\theta\}_{\theta \in \Theta}$ se construit de la manière suivante. Donnons-nous une suite $\{a_i\}_{i \geqq 1}$ strictement décroissante et positive avec

$$a_1=1, \quad a_i \leqq 2(a_{i+1}), \quad i \geqq 1 \tag{6.2}$$

et une suite non-décroissante d'entiers $\{p_n\}_{n\geqq 1}$ avec la condition

$$p_n \geqq \frac{3\log(n-q_n)}{2\log 2}, \quad n\geqq 1, \ p_1 \geqq 3 \tag{6.3}$$

où $q_0=0$, $q_n=\sup\{i \mid a_i \geqq 4a_n\}$ ou $q_n=0$ quand cet ensemble est vide. Une conséquence de (6.2) est que si $n\geqq 3$, $q_n \leqq n-2$. Si a_i est de la forme $a2^{-i}$, alors toute suite $\{p_n\}$ minorée par 3 vérifie (6.3).

Pour $i\geqq 1$ posons $\Theta_{(i)}=\{-1;+1\}^{p_i}$ et $\Theta=\bigoplus_{i=1}^{+\infty}\Theta_{(i)}$. Pour repérer un élément de $\Theta_{(i)}$ nous l'écrirons sous la forme $\theta=\{i;\theta_1;\ \ldots;\ \theta_{p_i}\}$ et noterons θ_0 la première coordonnée qui indique que θ appartient à $\Theta_{(i)}$. Alors $\theta_j=\pm 1$ si $j\geqq 1$. Comme $0<a_i\leqq 1$ les intervalles $[i;i+a_i[$ sont disjoints. Nous diviserons $[i;i+a_i[$ en p_i intervalles égaux, chacun d'eux étant lui-même divisé en deux parties égales A'_{ij} et A''_{ij} avec

$$A'_{ij}=\left[i+\frac{(j-1)a_i}{p_i};\ i+\frac{(j-\frac{1}{2})a_i}{p_i}\right[, \qquad A''_{ij}=\left[i+\frac{(j-\frac{1}{2})a_i}{p_i};\ i+\frac{j a_i}{p_i}\right[.$$

Sur $A'_{ij}+A''_{ij}$ nous construirons une petite perturbation f_{ij} en posant $f_{ij}=\frac{1}{4}(1_{A'_{ij}}-1_{A''_{ij}})$. Alors si $\theta=\{i,\theta_1,\ldots,\theta_{p_i}\}$ la mesure P_θ est constituée par une masse $1-a_i$ uniformément répartie sur $[0;1-a_i]$ et une masse a_i sur $[i;i+a_i[$ donnée par le système de perturbations $1+\sum_{j=1}^{p_i}\theta_j f_{ij}$, soit si $\theta_0=i$

$$\frac{dP_\theta}{d\lambda}=f_\theta=1_{[0;1-a_i]}+1_{[i;i+a_i]}+\sum_{j=1}^{p_i}\theta_j f_{ij}.$$

On peut alors montrer:

Proposition 6.2. *Pour la classe de modèles ainsi construits, les h-estimateurs sont adéquats et l'on a*

$$R_n(h^q)\asymp[L\tilde{h}(n)]^q.$$

Preuve. Remarquons d'abord que si $\theta'_0\neq\theta_0$, $h^2(P_\theta,P_{\theta'})=a_q$ où $q=\inf(\theta_0,\theta'_0)$ et si $\theta_0=\theta'_0=i$ on a avec $C_1=1-\frac{\sqrt{15}}{4}\simeq 0{,}03$

$$h^2[1+f_{ij},1-f_{ij}]=\frac{C_1 a_i}{p_i}\leqq h^2(P_\theta,P_{\theta'})\leqq C_1 a_i<a_{i+1}, \qquad 1\leqq j\leqq p_i.$$

Il s'ensuit qu'on peut calculer une borne supérieure de $\tilde{h}(\varepsilon)$. Supposons que $a_{i+1}\leqq\varepsilon^2<a_i$. Nous construisons un $\sqrt{a_{i+1}}$-réseau S de la manière suivante: nous prenons un point de $\Theta_{(i)}$, un point de $\Theta_{(i+1)}$ et tous les points de $\Theta_{(k)}$ pour $k\leqq i-1$. Définissons la suite $\{r_n\}$ par récurrence avec $r_0=i$ et $r_n=q_{r_{n-1}}$. D'après la définition de q_m, il s'ensuit que $a_{r_m}\geqq 4^m a_i$, si $r_m>0$. Considérons la boule ouverte $\mathscr{B}(\theta,\sqrt{4^m a_i})$. Elle est incluse dans $\mathscr{B}(\theta,\sqrt{a_{r_m}})$ si $r_m\neq 0$. Supposons que θ soit dans $\Theta_{(k)}$ et $r_m \neq 0$. Si $k \leqq r_m$, cette boule contient au plus 2^{p_k} points de S. Si $k > r_m$ elle contient au plus $(2+2^{p_i-1}+\ldots+2^{p_{r_m}+1})$ points de S, ce qui est inférieur à $2^{p_i}(i-r_m)$. Comme $r_j - r_{j+1} \geqq 2$, on a

$$\log(i-r_m) = \log(r_0 - r_m) \leqq \sum_{j=0}^{m-1} \log(r_j - r_{j-1})$$

$$\leqq \tfrac{2}{3}\log 2 \sum_{j=0}^{m-1} p_{r_j} \leqq \left(\frac{2m}{3}\log 2\right) p_i$$

en utilisant (6.3). Il s'ensuit que

$$\log \operatorname{Card}[S \cap \mathscr{B}(\theta, 2^m \sqrt{a_i})] \leqq \left(\frac{2m}{3}+1\right) p_i \log 2.$$

On peut vérifier que la formule subsiste si $r_m = 0$, donc

$$\tilde{h}(\varepsilon) \leqq p_i \qquad \text{si } a_{i+1} \leqq \varepsilon^2 < a_i$$

et il est immédiat que $\tilde{h}(\varepsilon) = 0$ si $\varepsilon^2 \geqq a_1$. On choisira donc pour fonction δ

$$\delta(\varepsilon) = p_i \quad \text{si } a_{i+1} \leqq \varepsilon^2 < a_i, \quad \delta(\varepsilon) = 0 \quad \text{si } \varepsilon^2 \geqq a_1.$$

Si la suite (a_i) ne tend pas vers 0, on prendra $\delta(\varepsilon) = +\infty$ pour $\varepsilon \leqq \lim_i a_i$. Il s'ensuit que l'on a

$$\begin{aligned} (L\delta(n))^2 &= n^{-1} p_i && \text{si } n a_{i+1} \leqq p_i < n a_i, \\ &= a_{i+1} && \text{si } p_i < n a_{i+1} \leqq p_{i+1}, \\ &= 1 && \text{si } p_1 \geqq n, \end{aligned} \tag{6.4}$$

ou plus simplement

$$a_i > (L\delta(n))^2 \geqq a_{i+1} \qquad \text{si } p_i a_i^{-1} < n \leqq p_{i+1} a_{i+1}^{-1}. \tag{6.5}$$

Pour vérifier H5, nous prendrons comme ensemble de petites perturbations un des $\Theta_{(k)}$ auquel nous appliquerons le corollaire 3.9. Nous remarquerons que, vu la forme des f_{ij}, on a suivant le lemme 4.4

$$K(1+f_{ij}, 1-f_{ij}) \leqq 5h^2(1+f_{ij}, 1-f_{ij}) = 5\,C_1 \frac{a_i}{p_i}.$$

Il s'ensuit que si θ et θ' sont deux points de $\Theta_{(k)}$ il vient

$$K(P_\theta^n, P_{\theta'}^n) \leqq n p_k 5\, C_1 \frac{a_k}{p_k} = 5n\, C_1 a_k.$$

Posons $C_2 = 50\, C_1$, d'où $C_2 \simeq 1{,}5$. Alors pour tout i, on peut trouver un $k > i$ tel que

$$C_2 p_{i+1} a_{i+1}^{-1} \leqq p_k a_k^{-1} \qquad \text{et } a_k \geqq \frac{a_i}{2\,C_2} \tag{6.6}$$

d'après (6.2) et (6.3). Si la suite a_i converge vers 0 on choisit a_k approximativement égal à $\frac{a_i}{2C_2}$ et si cela est impossible parce que $a_i \to a > 0$, alors p_n tend vers l'infini et (6.6) peut être réalisé pour k grand. Il suffit dès lors d'appliquer le corollaire 3.9 à l'ensemble $\Theta_{(k)}$, k étant défini par (6.6) et i par (6.5) (ou $i=1$ si $n \leqq p_1$). (3.16) et (3.17) s'écrivent dans ce cas

$$\left(\frac{p_k}{8}\right) C_1 \frac{a_k}{p_k} \geqq c^2 a_i \quad \text{et} \quad 5n C_1 a_k \leqq \frac{p_k}{10}.$$

La première inégalité est vérifiée avec $c^2 = \frac{C_1}{16 C_2} = \frac{1}{800}$ et la seconde grâce à (6.6). Donc H5 est satisfaite et on peut appliquer le théorème 3.4 ainsi que la proposition 3.5 grâce au lemme 4.4 qui implique que $K(P_\theta^n, P_{\theta'}^n) \geqq 2nh^2(\theta, \theta')$. D'où le résultat. □

Nous allons montrer maintenant que grâce à cette famille de modèles, nous pouvons obtenir toutes les vitesses «raisonnables». En effet, si $r(n)$ représente une vitesse, il est nécessaire d'imposer que $r(n)$ soit décroissante (l'estimation s'améliorant avec le nombre des observations), mais aussi qu'elle ne décroisse pas trop vite puisque l'on a vu que l'on ne peut pas estimer plus vite que $n^{-\frac{1}{2}}$ dès que l'espace des paramètres contient des points suffisamment proches. Comme les vitesses sont définies à équivalence près, le choix de $r(1)$ est arbitraire. Pour la commodité des calculs nous normaliserons $r(n)$ par $r(1)=2$.

Théorème 6.3. *Soit une suite $r(n)$ de $\mathbb{N}^*$ dans $]0; 2]$ telle que*

$$r(1)=2, \quad r(n+1) \leqq r(n), \quad nr^2(n) \leqq (n+1) r^2(n+1), \quad \lim_{n\to\infty} r(n) \leqq 1. \tag{6.7}$$

Alors il existe des suites $\{a_i\}_{i \geqq 1}$ et $\{p_i\}_{i \geqq 1}$ satisfaisant (6.2) *et* (6.3) *telles que pour le modèle du type précédent associé, on ait $R_n(h^q) \asymp r^q(n)$.*

Preuve. 1^er cas: supposons que $r(n) \xrightarrow[n\to\infty]{} 0$. Posons alors $a_i = 2^{-i+1}$ et

$$j_i = \inf\{n \mid r^2(n) < a_i\}, \qquad p_i = \max\{[\![j_i 2^{-i} + 1]\!], p_{i-1}\}.$$

Il s'ensuit que la suite $\{p_i\}$ est croissante avec $p_1 \geqq 3$ puisque $j_1 \geqq 4$. En effet, si $n < 4$, $nr^2(n) \geqq r^2(1) = 4$ et donc $r^2(n) > 1 = a_1$. (6.3) est vérifiée trivialement et l'on a toujours $p_i > 2^{-i} j_i$. Comme $nr^2(n)$ est croissante, on a

$$r^2(n) \geqq \frac{4}{n}; \qquad r^2(kn) \geqq k^{-1} r^2(n), \; k \geqq 1. \tag{6.8}$$

On en déduit, comme $r^2(j_i - 1) \geqq a_i$, que $r^2[2(j_i - 1)] \geqq a_{i+1}$ et donc que $j_{i+1} \geqq 2j_i - 1$. En multipliant cette inégalité par 2^{-i-1}, on obtient par récurrence

$$2^{-(i+1)} j_{i+1} \geqq 2^{-k}(j_k - 1), \qquad 1 \leqq k \leqq i.$$

Il s'ensuit que $p_i \leqq \sup_{1 \leqq k \leqq i} [\![j_k 2^{-k} + 1]\!] \leqq [\![j_i 2^{-i} + \frac{3}{2}]\!]$ et

$$[\![j_i 2^{-i} + 1]\!] \leqq p_i \leqq [\![j_i 2^{-i} + \tfrac{3}{2}]\!]. \tag{6.9}$$

D'après la proposition (6.2), $R_n(h^q) \asymp [L\delta(n)]^q$ où $L\delta(n)$ est défini par (6.4) ou (6.5)

– si $n \leqq p_1 a_1^{-1} = p_1$, alors $1 \geqq L\delta(n) \geqq \frac{1}{2}$ et $4 \geqq r^2(n) \geqq 1$.

– si $p_i a_i^{-1} < n \leqq p_{i+1} a_{i+1}^{-1}$, $\sqrt{a_{i+1}} \leqq L\delta(n) < \sqrt{a_i}$ et d'après (6.9)

$$\frac{j_i}{2} < n \leqq \frac{j_{i+1}}{2} + \tfrac{3}{2} a_{i+1}^{-1}.$$

Mais aussi $2n > j_i$, donc $r^2(2n) < a_i$ et $r^2(n) \leqq 2r^2(2n) < 2a_i$ et de même $n < j_{i+1}$ (car $r^2(j_{i+1}) < a_{i+1}$ entraîne par (6.8) que $j_{i+1} > 4a_{i+1}^{-1}$) et donc $r^2(n) \geqq a_{i+1}$. Finalement $a_{i+1} \leqq r^2(n) < 2a_i$. Il s'ensuit que $L\delta(n) \asymp r(n)$.

$2^{ème}$ *cas:* si $r(n)$ tend vers une limite $a > 0$ lorsque $n \to +\infty$, on reprend la construction précédente, mais il existe un q et un n_0 tels que

$$a_q > r^2(n) \geqq \frac{a_q}{2} = a_{q+1} \qquad \text{pour } n \geqq n_0,$$

ce qui entraîne que $j_q < +\infty$ et $j_{q+1} = +\infty$. Pour $i \leqq q$, on prendra les a_i et p_i comme précédemment. Pour $i > q$, $\{a_i\}$ est n'importe quelle suite décroissante telle que $a_q > a_i \geqq \frac{a_q}{2}$ et $\{p_i\}$ n'importe quelle suite croissante tendant vers l'infini et satisfaisant (6.3). On remarquera que q_n tend vers une limite qui est a_{q-2} et qu'il suffit de choisir $p_n \geqq 3 \log n$ pour obtenir (6.3). Le calcul précédent reste donc valable pour $i \leqq q-1$, c'est-à-dire pour $n \leqq p_q a_q^{-1}$. Pour $n > p_q a_q^{-1}$, on a certainement $a_q > (L\delta(n))^2 \geqq \frac{a_q}{2}$ et aussi $n > \frac{j_q}{2}$. Il s'ensuit qu'on a aussi $2a_q > r^2(n) \geqq \frac{a_q}{2}$. Donc l'équivalence $r(n) \asymp L\delta(n)$ demeure vérifiée. □

7. Conclusion

Nous pouvons déduire du chapitre précédent que, si les d-estimateurs ne fournissent pas une méthode universelle d'estimation, dans le cas de variables indépendantes équidistribuées ils permettent de retrouver les résultats classiques sur l'estimation des densités, ainsi que de traiter des exemples nouveaux tels que les classes Λ_ω, aussi bien dans le cas indépendant que pour des densités spectrales et des problèmes variés d'estimation de transitions pour les chaînes de Markov. D'autre part, on donne ici une construction explicite de ces estimateurs, soit comme estimateurs du maximum de vraisemblance sur des réseaux, soit à partir de tests entre des boules de Hellinger (voir [6]).

Le théorème 6.3 montre en outre que pour toute vitesse $r(n)$ satisfaisant à (6.7) il existe un modèle pour lequel la théorie précédente s'applique, $R_n(h^q)$ étant équivalent à $r^q(n)$ et les suites de h-estimateurs étant adéquates. Il s'ensuit que lorsque l'on a peu d'informations sur le modèle, en particulier sur les conditions de régularité qui permettraient d'utiliser des méthodes classiques, mais seulement la connaissance de la structure métrique de Θ par $\tilde{d}$, on sera naturellement amené (au moins dans le cas des variables indépendantes équidistribuées et de la distance de Hellinger) à utiliser les d-estimateurs, pour lesquels on aura une majoration du risque dont on sait qu'elle sera adéquate pour certains types de modèles au moins, ayant une structure $\tilde{d}$ analogue.

Par ailleurs, sur tous les modèles auxquels elle s'applique, la théorie précédente permet de calculer a priori la vitesse d'estimation et de l'expliquer par la structure métrique de l'espace des paramètres, ce qui n'était pas le cas jusqu'à présent pour les problèmes d'estimation de densités où l'on se contentait de faire les calculs dans chaque cas, sans disposer d'un outil général permettant de lier la vitesse à une structure particulière de l'espace des paramètres. En outre, il ressort de ce qui précède que la vitesse d'estimation est davantage un phénomène métrique que vectoriel, même si la théorie classique a avant tout utilisé les propriétés vectorielles de l'espace des paramètres et surtout les propriétés d'orthogonalité dans les espaces euclidiens et hilbertiens. Les propriétés vectorielles impliquent bien évidemment des propriétés métriques, mais l'inverse n'est pas vrai. Ce fait est souligné par les résultats récents d'Assouad [2] lequel a obtenu une réciproque des résultats de Le Cam [24] disant que si la dimension métrique de Θ est finie, l'estimation se fait à la vitesse $n^{-1/2}$. Assouad montre qu'inversement, si pour toute expérience métriquement équivalente à celle ayant Θ pour espace de paramètres, la vitesse d'estimation est en $n^{-1/2}$, alors Θ est de dimension métrique finie.

Enfin une propriété fort importante des d-estimateurs dans le cas des variables indépendantes équidistribuées est leur robustesse. Ce fait sera développé dans un prochain article mais on peut le voir aisément car il est clair que ce sont des estimateurs sur $\Theta' = \bigcup_{s\in S} \mathscr{B}(s,\eta)$ où S est un η-réseau de Θ, c'est-à-dire sur un voisinage Θ' de Θ. Si dans cette construction on remplace $\mathscr{B}(s,\eta)$ par $\mathscr{B}(s,k\eta)$ où k est une constante fixée, on ne multiplie le risque que par une constante. Ainsi, au prix d'une légère modification dans la construction des d-estimateurs, on disposera d'une estimation sur l'espace de paramètres élargi $\Theta_k = \bigcup_{s\in S} \mathscr{B}(s,k\eta)$ ce qui permet des erreurs égales à $(k-1)\eta$. Si l'on se place dans le cadre d'une suite d'observations, cela signifie que l'on autorise des déviations de taille $(k-1)L\delta(n)$ pour la distance d, si $L\delta(n)$ représente la vitesse d'estimation sur Θ. En fait le procédé ci-dessus revient à construire des d'-estimateurs (avec $d'=kd$) en utilisant des d-réseaux, c'est-à-dire en choisissant $\delta \geqq \tilde{d}$ (et non $\tilde{d}'$). De plus, dans le cas des variables indépendantes, les déviations peuvent être différentes pour chaque observation.

Pour pouvoir prendre en compte des erreurs plus importantes, en particulier des déviations pour une distance $\bar{d}$ inférieure à d, il est nécessaire de faire une hypothèse supplémentaire qui assure que le modèle reste identifiable et que sa structure métrique n'est pas modifiée. Si la construction d'un d-estimateur sur Θ requiert un η-réseau S, l'espace des paramètres élargi Θ' que l'on considérera devra satisfaire aux hypothèses suivantes:

Si Q_s (resp. Q_t) dans Θ' est une contaminée d'un point P_s (resp. P_t) de Θ on a

$$d(Q_s, Q_t) \geqq k'\, d(P_s, P_t) - k\eta \tag{7.1}$$

où k et k' sont deux constantes fixées. En fait Θ' sera de la forme $\Theta' = \bigcup_{s \in S} \mathscr{V}_s$ où $\mathscr{V}_s$ est un voisinage de P_s. Toute la théorie pourra fonctionner avec Θ' à la place de Θ si les ensembles $\mathscr{V}_s$ satisfont à l'hypothèse H2. Dans le cas des variables indépendantes équidistribuées, il sera facile de le montrer à partir de [6] ou [23] pourvu que les $\mathscr{V}_s$ soient convexes et satisfassent à (7.1). En utilisant (2.2) on peut montrer que l'estimateur obtenu sera tel que si Q est un point de $\mathscr{V}_s$, $\mathbb{E}_Q[d^q(\hat{\theta}, s)]$ sera majoré par un multiple de η^q. Ainsi, dans le cas de suites d'observations, on obtiendra la vitesse donnée par la structure métrique de Θ (seules les constantes seront modifiées). Il serait en particulier facile d'appliquer ces résultats à des modèles paramètriques de dimension finie. Il n'est pas possible par contre d'obtenir une borne pour $\mathbb{E}_Q[d^q(P_{\hat{\theta}}, Q)]$ car il se peut fort bien que même si Q appartient à $\mathscr{V}_s$, $d(Q, P_s)$ soit très grand.

Nous terminerons ce chapitre par un certain nombre de remarques.

Remarque 1. Il est clair, vu la façon dont sont définis $\tilde{d}$ d'une part, les réseaux servant aux minorations d'autre part, que les bornes supérieures du risque sont obtenues en utilisant les propriétés globales de Θ et les bornes inférieures des propriétés locales. Le fait que l'on puisse les faire coïncider est dû, de façon essentielle, à l'utilisation du risque minimax. Toutefois il est possible que si l'on se restreint à une partie $\mathscr{V}$ de Θ, un petit voisinage d'un point donné en particulier, on obtienne une valeur de $\tilde{d}$ qui dépendra de $\mathscr{V}$ et qu'elle puisse être notablement inférieure à la valeur de $\tilde{d}$ correspondant à Θ tout entier. Il se peut donc, dans certains cas, qu'on ait intérêt à localiser grossièrement le paramètre dans un premier temps en utilisant un estimateur préliminaire et à utiliser ensuite des d-estimateurs sur une partie seulement de Θ.

Remarque 2 [4]. On peut se demander pourquoi utiliser les d-estimateurs qui ne sont jamais que des avatars robustes du maximum de vraisemblance plutôt que ce dernier. La réponse est qu'il semble impossible de développer une telle théorie métrique en utilisant le maximum de vraisemblance (ou alors en imposant des hypothèses très strictes et difficiles à vérifier) en particulier en raison de la non-robustesse de ce dernier. Il existe des exemples (cf. [3] et [22]) pour lesquels le maximum de vraisemblance n'est même pas consistant alors que la fonction $\tilde{h}(\varepsilon)$ est finie ce qui implique que les h-estimateurs se comportent de façon satisfaisante. Si l'on reprend l'exemple de Bahadur [3], on peut vérifier que $\tilde{h}(\varepsilon) < -C \log \varepsilon$; les h-estimateurs convergeront donc à une vitesse meilleure que $\left(\frac{\log n}{n}\right)^{1/2}$.

Remarque 3. Il est bien clair que la construction des d-estimateurs repose essentiellement sur l'hypothèse H2, laquelle, dans le cas de n variables indépendantes vient du fait que

$$\pi(P^n, Q^n) \leqq \rho(P^n, Q^n) \leqq \rho^n(P, Q) = \exp[n \log(1 - h^2(P, Q))]$$

et du lien entre h et d. Dans les autres cas, on utilise des inégalités exponentielles analogues. Il s'ensuit que dans ces inégalités, on pourrait remplacer h^2 par

[4] En réponse à une question de G. Tusnády

$-\log\rho$. C'est en fait cette quantité qui mesure les erreurs des tests et avec laquelle on devrait travailler. Malheureusement, $\sqrt{-\log\rho}$ n'est pas une distance, tout au plus localement une quasi-distance. On pourrait étendre la théorie à ce cas, mais cela ne présente guère d'intérêt et complique la présentation des résultats ainsi que les calculs (cf. [5]). Le fait que les erreurs des tests s'expriment aisément en fonction de ρ confère à h son importance et entraîne que les distances que l'on utilise doivent être convenablement reliées à h. En particulier les distances $\mathbb{L}^p$ pour $p>2$ ne conviennent pas parce qu'il n'est plus possible de les majorer par un multiple de h, en général. De toutes façons une telle distance n'est pas intrinsèque, dépendant du choix de la mesure dominante.

Remarque 4. C'est le lemme 2.7 qui montre l'importance de l'information de Kullback dans la recherche des minorations. Cette information n'est pas une distance, et pas davantage directement liée à ρ ou h. Elle s'introduit en raison des bonnes propriétés de la fonction logarithme, mais il serait souhaitable de la remplacer par autre chose dans la mesure où les minorations qu'elle fournit peuvent être, comme on l'a vu au chapitre précédent, très mauvaises. Malheureusement il paraît difficile d'obtenir des minorations satisfaisantes autrement que par des quantités faisant intervenir un grand nombre de probabilitités à la fois ou leurs mélanges, comme le risque bayésien. Ces quantités deviennent alors très difficiles à calculer et n'ont pas la bonne propriété de multiplicativité de l'information: $K(P^n, Q^n) = nK(P, Q)$. Le problème se pose donc de trouver une fonction de Θ qui soit calculable et permette d'établir à la fois des majorations et des minorations du risque. Jusqu'à présent cela n'a pas été possible et l'utilisation de h apparaît donc comme un bon compromis, h^2 étant un intermédiaire satisfaisant dans bien des cas entre $-\log\rho$ et K, comme nous l'avons vu. Ces deux fonctions ont d'ailleurs la même propriété d'être multipliées par n lorsque l'on passe de une à n observations indépendantes, ce qui est fondamental dans la théorie précédente.

Je remercie tout particulièrement Patrice Assouad, Robert Azencott, Didier Dacunha-Castelle et Lucien Le Cam pour l'aide qu'ils ont pu, sous des formes diverses, m'apporter durant la préparation et la rédaction de ce travail.

Bibliographie

1. Assouad, P.: Espaces métriques, plongements, facteurs. Thèse d'Etat, Orsay (1976)
2. Assouad, P.: Classes de Vapnik-Červonenkis et vitesse d'estimation. Prépublication: Orsay 1982
3. Bahadur, R.R.: Examples of inconsistency of maximum likelihood estimates. Sankhya **20**, 207–210 (1958)
4. Beckenbach, E.F., Bellman, R.: Inequalities. Berlin-Heidelberg-New York: Springer 1961
5. Birgé, L.: Vitesses optimales de convergence des estimateurs. Astérisque **68**, 171–185 (1979)
6. Birgé, L.: Sur un théorème de minimax et son application aux tests. Probab. Math. Statist. III, **2** [To appear 1983]
7. Birgé, L.: Robust testing for independent non-identically distributed variables and Markov chains. Dans: Specifying statistical models, pp. 134–162. New York-Heidelberg-Berlin: Springer 1983

8. Bretagnolle, J., Huber, C.: Estimation des densités: risque minimax. Z. Wahrscheinlichkeitstheorie verw. Gebiete **47**, 119–137 (1979)
9. Čencov, N.N.: Statistical decision rule and optimal inference. En russe. Moscou: Nauka 1972
10. Dacunha-Castelle, D.: Ecole d'Eté de Probabilités de Saint-Flour VII. Berlin-Heidelberg-New York: Springer 1977
11. Donsker, M.D., Varadhan, S.R.S.: Asymptotic evaluation of certain Markov process expectations for large time, I. Comm. Pure Appl. Math. **28**, 1–47 (1975)
12. Farrell, R.H.: On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. Ann. Math. Statist. **43**, 170–180 (1972)
13. Farrell, R.H.: Asymptotic lower bounds for the risk of estimators of the value of a spectral density function. Z. Wahrscheinlichkeitstheorie verw. Gebiete **49**, 221–234 (1979)
14. Huber, C.: Thèse d'Etat: Orsay 1979
15. Huber, P.J.: A robust version of the probability ratio test. Ann. Math. Statist. **36**, 1753–1758 (1965)
16. Ibragimov, I.A., Khas'minskii, R.Z.: Statistical estimation, Asymptotic theory. New York-Heidelberg-Berlin: Springer 1981
17. Ibragimov, I.A., Khas'minskii, R.Z.: On estimate of the density function. En russe. Zap. Nauchn. Semin. LOMI **98**, 61–85 (1980)
18. Ibragimov, I.A., Khas'minskii, R.Z.: On the non-parametric density estimates. En russe. Zap. Nauchn. Semin. LOMI **108**, 73–89 (1981)
19. Khas'minskii, R.Z.: A lower bound on the risks of non-parametric estimates of densities in the uniform metric. Theory Probability Appl. **23**, 794–796 (1978)
20. Kolmogorov, A.N., Tihomirov, V.M.: ε-entropy and ε-capacity of sets in function spaces. Amer. Math. Soc. Transl. (2) **17**, 277–364 (1961)
21. Kraft, C.: Some conditions for consistency and uniform consistency of statistical procedures. Univ. of Calif. Publ. in Statistics **1**, 125–142 (1955)
22. Le Cam, L.: Théorie asymptotique de la décision statistique. Presses de l'Université de Montréal: 1968
23. Le Cam, L.: Convergence of estimates under dimensionality restrictions. Ann. Statist. **1**, 38–53 (1973)
24. Le Cam, L.: On local and global properties in the theory of asymptotic normality of experiments. Dans: Stochastic processes and related topics **1**, 13–54. New York-San Francisco-London: Academic Press 1975
25. Le Cam, L.: An inequality concerning Bayes estimates. Prépublication (1979)
26. Le Cam, L.: Asymptotic methods in statistical decision theory. Livre à paraître
27. Lorentz, G.G.: Metric entropy and approximation. Bull. Amer. Math. Society **72**, 903–937 (1966)
28. Lorentz, G.G.: Approximation of functions. New York: Holt, Rinehart, Winston 1966
29. Okamoto, M.: Some inequalities relating to the partial sum of binomial probabilities. Ann. Inst. Statis. Math. **10**, 29–35 (1958)
30. Pinsker, M.S.: Optimal filtration of square-integrable signals in gaussian noise. Problems of Inform. Transmission **16**, 120–133 (1980)
31. Strasser, H.: Convergence of estimates: parts I and II. J. Multivariate Anal. **11**, 127–172 (1981)
32. Vitushkin, A.G.: Theory of transmission and processing of information. New York: Pergamon Press 1961
33. Wahba, G.: Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. Ann. Statist. **3**, 15–29 (1975)

Reçu le 28 Mars 1981; en forme révisée le 14 Avril 1983