ANNALES DE L'I. H. P., SECTION B

LUCIEN BIRGE

Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées

Annales de l'I. H. P., section B, tome 20, nº 3 (1984), p. 201-223
<http://www.numdam.org/item?id=AIHPB_1984__20_3_201_0>

Ann. Inst. Henri Poincaré,
Vol. 20, n° 3, 1984, p. 201-223. *Probabilités et Statistiques*

Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées

par

Lucien BIRGE
E. R. A. C. N. R. S. 532
U. E. R. de Sciences Économiques,
Université Paris-X-Nanterre, 200, avenue de la République,
F-92001, Nanterre Cedex, France

RÉSUMÉ. — Soit à estimer un paramètre inconnu $\theta \in \Theta$ à partir de n observations indépendantes de loi P_θ. Le risque minimax correspondant à cette expérience statistique est défini par

$$R_n(\Theta) = \inf_{T} \sup_{\theta \in \Theta} \mathbb{E}[nh^2(P_T, P_\theta)].$$

où T est un estimateur de θ et h la distance de Hellinger. Supposons que l'on perturbe Θ de ε_n, on obtient un nouvel espace de paramètres $\overline{\Theta}_n$ qui est un ε_n-voisinage de Θ. Le nouveau risque $R_n(\overline{\Theta}_n)$ sera-t-il très différent ? Plus précisément nous allons chercher des suites ε_n telles que

$$R_n(\overline{\Theta}_n) \leqq KR_n(\Theta)$$

où K ne dépend ni de n, ni de Θ. Les suites du type $\varepsilon_n = cn^{-1/2}$ conviennent (K ne dépendant alors que de c). Pour un choix de la forme $\varepsilon_n = cn^{-1/2}R_n(\Theta)$, on peut trouver des contre-exemples mais le résultat subsiste pour une large famille de modèles, réguliers en ce sens que leur risque est approximativement déterminé par leur structure métrique. De nombreux problèmes non paramétriques classiques sont de ce type.

ABSTRACT. — Consider the problem of estimating the unknown parameter θ, which belongs to the parameter space Θ, from n i. i. d. variables

of law P_θ. We shall define the minimax risk corresponding to this experiment by

$$R_n(\Theta) = \inf_{T} \sup_{\theta\in\Theta} \mathbb{E}_\theta[nh^2(P_T, P_\theta)],$$

where T denotes an arbitrary estimator of θ and h is the Hellinger distance. Suppose that we enlarge the parameter space or make some perturbations of it of size ε_n; we shall get a new parameter space $\overline{\Theta}_n$ with the property that any point of $\overline{\Theta}_n$ will be within a distance ε_n from some point in Θ. The question is: how much will $R_n(\overline{\Theta}_n)$ differ from $R_n(\Theta)$? More precisely we shall try to see for which choice of the sequence $\{\varepsilon_n\}_{n>0}$ there exists some constant K independent of n and Θ such that $R_n(\overline{\Theta}_n) < KR_n(\Theta)$.

For $\varepsilon_n = cn^{-1/2}$ we shall prove such an inequality (with K depending only on c).

For larger perturbations of the form $\varepsilon_n = cn^{-1/2}R_n(\Theta)$ the result does not hold in the general case but it is true for a large class of models which are regular in the sense that their risk is determined by the dimensional properties of the parameter space (Θ, h) considered as a metric space.

I. INTRODUCTION

Considérons le problème suivant : estimer un paramètre θ, appartenant à un espace Θ, à partir de n observations aléatoires $X_1, \ldots, X_n$ indépendantes à valeur dans un espace mesuré $(\Omega, \mathscr{A})$ et de même loi inconnue P_θ. Si l'on veut appliquer les résultats théoriques obtenus dans ce cadre à un problème pratique d'estimation, on va modéliser l'ensemble des lois possibles par $\{P_\theta^n, \theta\in\Theta\} = \Theta_n$; mais toute modélisation de ce type n'est qu'approximative et pose le problème suivant : que devient la qualité de notre estimation si le véritable modèle est en fait régi par une loi P′ un peu différente de P_θ^n. Le principal objectif de la théorie de la robustesse est d'obtenir des procédures d'estimation « robustes », c'est-à-dire dont les performances ne sont pas trop affectées par certains types de perturbations ce qui s'exprime en général sous forme de stabilité des caractéristiques (biais, variance asymptotique, etc.) des estimateurs en question. Cependant, dans de nombreux problèmes, Θ est trop gros (infini-dimensionnel) pour que l'on puisse adopter ce point de vue. Dans les problèmes classiques de robustesse, on estime un paramètre de dimension finie avec une vitesse de l'ordre de $n^{-1/2}$ et les propriétés de stabilité se situent au niveau des constantes, les perturbations envisagées n'affectant pas la vitesse d'estimation. Mais si Θ représente par exemple un ensemble de densités assez large, on sait que la vitesse d'estimation peut être sensiblement plus lente, en $\frac{1}{n^\alpha}$, $\alpha < \frac{1}{2}$ par exemple et il n'est pas du tout évident que des perturbations arbitraires de taille $n^{-1/2}$ ne vont pas entraîner des variations de cette vitesse.

Par ailleurs on peut se demander aussi quel serait l'effet de perturbations de taille plus importante et jusqu'où l'on peut aller ? Les perturbations que nous allons ici envisager vont représenter des « grossissements » en ce sens que l'on va remplacer Θ par un de ses voisinages et que celui-ci contiendra des boules. Une perturbation de taille ε entraînera l'apparition dans le nouvel espace des paramètres de boules de rayon ε pour la distance de Hellinger. Or si la perte est également mesurée par cette distance, on peut voir que (si Ω est infini) le meilleur estimateur pour une boule pleine est son centre, indépendamment des observations (cf. [*4*]). Donc, pour de telles fonctions de perte, autoriser des perturbations de taille ε implique d'obtenir un risque au moins égal à ε. Il s'ensuit que pour un problème d'estimation simple dont la vitesse est en $n^{-1/2}$, il est sans intérêt de considérer des perturbations d'amplitude supérieure à $cn^{-1/2}$. Par contre lorsque la vitesse initiale du problème est n^α, $0 > \alpha > -\frac{1}{2}$, on peut se demander quel sera l'effet de perturbations de taille comparable (c'est-à-dire cn^α). C'est de ce type de problème que nous allons nous préoccuper, et dans un cadre assez général qui nous amène à mettre l'accent sur la stabilité de la vitesse d'estimation, plutôt que sur la recherche d'estimateurs robustes, même si les démonstrations fournissent des méthodes générales pour construire des estimateurs dont la vitesse de convergence est celle que l'on attend (les constantes étant ici négligées, ce qui implique que les estimateurs obtenus seront généralement loin de l'optimal mais les hypothèses que nous ferons ne permettent guère d'investigations précises !).

Le modèle théorique est celui de l'estimation d'un paramètre à partir de n observations indépendantes équidistribuées de sorte que l'expérience initiale est décrite par l'ensemble des probabilités produits $\Theta_n = \{ P_\theta^n, \theta \in \Theta \}$. Dans la suite nous identifierons souvent θ à P_θ ou P_θ^n. La vitesse d'estimation est mesurée par le risque minimax associé à la distance de Hellinger (cf. [*1*] ou [*7*]):

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

de la manière suivante : sur Θ_n on considère la fonction de perte

$$H^2(P_\theta^n, P_{\theta'}^n) = nh^2(P_\theta, P_{\theta'})$$

que nous noterons le plus souvent $H^2(\theta, \theta')$ lorsque il n'y a pas d'ambiguïté sur la valeur de n. Le risque minimax s'écrit alors

$$R(\Theta_n) = \inf_{T_n} \sup_{\theta \in \Theta} E_\theta [nh^2(T_n, \theta)]$$

où T_n est un estimateur quelconque à valeurs dans l'ensemble $\mathscr{M}_1$ des probabilités sur Ω.

Comme nous l'avons dit, par perturbations nous entendons des grossissements et nous supposerons qu'ils affectent non seulement l'ensemble Θ mais aussi l'équidistribution (jamais l'indépendance !). Donc à tout $\varepsilon > 0$, nous associerons l'ε-voisinage Θ_n^ε de Θ_n ainsi défini : Θ_n^ε est l'ensemble des probabilités produits $P = \bigotimes_{i=1}^{n} P_i$ telles qu'il existe un point θ dans Θ pour lequel tous les P_i $(i = 1, \ldots, n)$ appartiennent à la boule fermée $\mathscr{B}(P_\theta, \varepsilon)$ de centre P_θ et rayon ε. On prolonge alors H à Θ_n^ε en posant

$$H^2\left(\bigotimes_{i=1}^{n} P_i, \bigotimes_{i=1}^{n} Q_i\right) = \sum_{i=1}^{n} h^2(P_i, Q_i)$$

de sorte que sur Θ_n^ε le risque minimax devient

$$R(\Theta_n^\varepsilon) = \inf_{T_n} \sup_{\theta' \in \Theta_n^\varepsilon} \mathbb{E}_{\theta'} [H^2(T_n, \theta')]$$

où T_n désigne maintenant un estimateur à valeurs dans l'ensemble $\{ P^n, P \in \mathscr{M}_1 \}$. De manière générale nous appellerons ε-perturbé de Θ_n tout ensemble Θ_n' tel que $\Theta_n \subset \Theta_n' \subset \Theta_n^\varepsilon$. Pour étudier la stabilité du risque, nous nous efforcerons de comparer $R(\Theta_n')$ et $R(\Theta_n)$ en fonction de ε. En fait, étant donné les propriétés du risque minimax, il suffira bien évidemment de comparer $R(\Theta_n^\varepsilon)$ et $R(\Theta_n)$.

Par ailleurs comme le risque minimax n'est pratiquement jamais accessible exactement nous utiliserons dans tout ce qui suit des majorants D_n de $R(\Theta_n)$, et nous étudierons le problème suivant : comment choisir ε pour que l'on ait $R(\Theta_n^\varepsilon) \leqslant KD_n$, où K est une constante universelle. Nous pourrons apporter à la question plusieurs réponses, suivant les propriétés de l'espace Θ.

Dans un premier temps, nous étudierons l'effet de petites perturbations, de taille $cn^{-1/2}$, ce qui correspond à la vitesse de séparation de deux probabilités puisque l'on sait d'après [7] que l'on ne peut séparer, à partir de n observations, deux probabilités dont la distance est beaucoup plus petite que $n^{-1/2}$. Nous verrons dans ce cas, et sous des hypothèses assez faibles (à savoir que la suite D_n est non décroissante), que toute expérience est stable (au sens précédent) à des perturbations de taille $cn^{-1/2}$, K ne dépendant que de c.

Par ailleurs, comme nous l'avons indiqué, le risque $R(\Theta_n^\varepsilon)$ est au moins de l'ordre de $n\varepsilon^2$ de sorte qu'il est sans intérêt de considérer des perturbations de taille supérieure à $c(D_n/n)^{1/2}$. Le problème est donc entièrement résolu lorsque la suite D_n est bornée. Mais lorsque D_n tend vers l'infini avec n, comme c'est le cas dans de nombreux problèmes statistiques (cf. [1], [5] et [6]), $(D_n/n)^{1/2}$ qui correspond en un certain sens à la distance moyenne entre la valeur du paramètre et son estimée est infiniment grand devant $cn^{-1/2}$ et il est naturel de se demander si des résultats identiques ne pourraient pas être obtenus avec $\varepsilon = c(D_n/n)^{1/2}$. Nous verrons qu'il n'en est rien dans le cas général (même si la suite D_n est croissante) mais que la stabilité existe dans certains cas réguliers qui ont été étudiés dans [0] et [1] pour lesquels on sait relier le risque à la structure métrique de Θ. Pour expliciter cela nous aurons besoin de quelques définitions.

DÉFINITION 1. — Pour toute fonction f de $\mathbb{R}^+$ dans $\mathbb{R}^+ \cup \{+\infty\}$ nous définirons la fonction Lf de $\mathbb{N} - \{0\}$ dans $\mathbb{R}^+ \cup \{+\infty\}$ par

$$(1.1) \qquad Lf(n) = \inf \{ x \mid nx^2 > f(x) \}.$$

DÉFINITION 2. — Étant donné un espace métrique (E, d), S est dit x-réseau de E si tout point de E est à distance inférieure ou égale à x d'un point de S.

DÉFINITION 3. — Dans un espace métrique (E, d) la fonction $\tilde{d}$ de $\mathbb{R}^+ - \{0\}$ dans $\mathbb{R}^+ \cup \{+\infty\}$ est définie par

$$(1.2) \qquad \tilde{d}(x) = \inf_{S_x} \sup_{\substack{j \geq 3 \\ j \in \mathbb{N}}} \sup_{y \in E} \frac{\operatorname{Log} \operatorname{Card} [\mathscr{B}(y, 2^j x) \cap S_x]}{j \operatorname{Log} 2}$$

où S_x parcourt l'ensemble des x-réseaux de (E, d) et $\mathscr{B}(y, r)$ désigne la boule fermée de centre y et de rayon r.

Si d est une distance bornée $\tilde{d}(x)$ est borné lorsque x est assez grand de sorte que $L\tilde{d}$ est une fonction finie et il est démontré dans [1] que l'on a toujours

$$(1.3) \qquad R(\Theta_n) \leq Cn[L\tilde{h}(n)]^2, \quad \text{dès que} \quad nL\tilde{h}(n) \geq \frac{1}{2}.$$

Pour certains problèmes réguliers on obtient également une inégalité analogue en sens inverse ; ce sont les problèmes de ce type que nous allons considérer. Pour cela nous utiliserons la définition suivante :

DÉFINITION 4. — Une expérience Θ_n sera dite a-dimensionnelle si

$$(1.4) \qquad R(\Theta_n) \geqslant an[L\tilde{h}(n)]^2 .$$

L'appellation dimensionnelle correspond au fait que $\tilde{h}$ décrit la dimension métrique de (Θ, h). Évidemment, tout problème est a-dimensionnel pour une certaine constante a, mais il l'est d'autant moins que a est petit et cette notion n'a plus de sens pour des valeurs trop élevées de $\frac{C}{a}$. L'intérêt de cette définition est de permettre d'obtenir des résultats uniformes en n et Θ lorsque a est fixé. En particulier les problèmes a-dimensionnels sont stables pour des perturbations de taille $\varepsilon = c\left(\frac{D_n}{n}\right)^{1/2}$ avec une constante K ne dépendant que de a et c. Malheureusement, ce résultat ne peut se généraliser aux problèmes non-dimensionnels car le rapport $\frac{R(\Theta_n^\varepsilon)}{R(\Theta_n)}$ peut devenir arbitrairement grand lorsque a tend vers 0 ce que nous montrerons en construisant des modèles convenables et qui laisse à penser que les problèmes non-dimensionnels sont en un sens pathologiques.

Les trois chapitres qui suivent vont développer successivement ces trois aspects du problème, à savoir :

i) l'étude de l'effet de perturbations de taille $cn^{-1/2}$,

ii) la stabilité aux perturbations de taille $c(D_n/n)^{-1/2}$ des modèles dimensionnels,

iii) le comportement très instable des vitesses pour certains modèles non-dimensionnels.

Ceci dit, il ne faut pas voir dans les constructions qui suivent une tentative pour exhiber des estimateurs robustes utilisables en pratique. Ces constructions n'ont d'intérêt que théorique mais permettent de savoir *a priori* à quel degré de stabilité on peut s'attendre dans un problème donné et donc quelles performances maximales on peut espérer d'estimateurs robustes.

II. UN RÉSULTAT GÉNÉRAL DE STABILITÉ

Étant donné le problème d'estimation associé à l'expérience Θ_n précédemment définie, nous allons étudier le comportement du risque lorsque l'on remplace Θ_n par $\Theta'_n \subset \Theta_n^\varepsilon$, ε étant de la forme $cn^{-1/2}$ avec c fixé. L'objet de ce chapitre est de montrer que pour de telles perturbations, le résultat suivant est vérifié :

THÉORÈME 1. — Étant données une constante $c > 0$ et une expérience Θ_n telle que $R(\Theta_m) \leqslant D$ pour $m \leqslant n$, $D \geqslant 1/4$, tout problème Θ'_n, ε-perturbé de Θ_n avec $\varepsilon = cn^{-1/2} < 1$, aura un risque $R(\Theta'_n)$ majoré par

$$R(\Theta'_n) \leqslant K(c)D \tag{2.1}$$

où K est indépendant de Θ, n et D.

Remarque. — La restriction $D \geqslant 1/4$ est arbitraire. Il importe seulement de se fixer une borne inférieure pour D, ce qui signifie que l'ensemble Θ n'est pas trop lacunaire. En effet, il est bien clair que si Θ'_n contient des boules de rayon $cn^{-1/2}$ on a nécessairement $R(\Theta'_n) \geqslant c^2$ ce qui est incompatible avec (2.1) si D peut être arbitrairement petit !

Construction d'un estimateur $\tilde{\theta}$.

Cette construction n'utilisera que les 2^m premières observations, où m est choisi de sorte que

$$\frac{n}{2} < 2^m \leqslant n . \tag{2.2}$$

Donnons-nous trois constantes : l entier non-négatif et a, $\lambda \geqslant 3$ de telle sorte que

$$2^l = (\lambda c)^2 \qquad \text{et} \qquad \beta = a^{-1} + \sqrt{2}\lambda^{-1} < 1/2 . \tag{2.3}$$

Nous pouvons supposer que $Da2^{l-m} < 1/4$ sinon $2^m \leqslant 4Da(\lambda c)^2$ et puisque $R(\Theta'_n) \leqslant n$,

$$R(\Theta'_n) \leqslant n2^{-m}4Da\lambda^2c^2$$

ce qui démontre le théorème grâce à (2.2). Fixons donc $D' > D$ et tel que $D'a2^{l-m} < 1/4$ et pour $l \leqslant i \leqslant k$ avec k défini par

$$1/8 \leqslant D'a2^{k-m} < 1/4 \tag{2.4}$$

posons

$$b_i^2 = D'a2^{i-m} . \tag{2.5}$$

Clairement $k < m$ vu que $D'a \geqslant 3/4$ et la suite $\{ b_i \}_{l \leqslant i \leqslant k}$ est croissante.

Pour tout $i, l \leqslant i \leqslant k$, posons $I = 2^{m-i}$ et définissons 2^i variables I-dimensionnelles Y_j^I à partir des variables observées $X_1, \ldots, X_n$ par

$$(2.6) \qquad Y_j^I = (X_{(j-1)I+1}, \ldots, X_{jI}) \qquad 1 \leqslant j \leqslant 2^i .$$

Par hypothèse, pour tout I et tout j il existe un estimateur $\hat{\theta}_I(Y_j^I)$ tel que

$$(2.7) \qquad \sup_{\theta \in \Theta} \mathbb{E}_\theta [I h^2(\hat{\theta}_I, \theta)] \leqslant D' .$$

Notons $\mathscr{P}_I$ l'ensemble des $\hat{\theta}_I(Y_j^I)$, $1 \leqslant j \leqslant 2^i$, et considérons dans Θ les sous-ensembles T_i $(l \leqslant i \leqslant k)$ des P_θ tels que

$$\operatorname{Card} [\mathscr{B}(P_\theta, b_i) \cap \mathscr{P}_I] > 2^{i-1} .$$

Nous définirons alors une suite d'ensembles de confiance S_i pour $k + 1 \geqslant i \geqslant l - 1$ de la manière suivante :

$$S_{k+1} = \Theta ; \quad S_k = S_{k+1} \cap T_k ; \ldots ; S_i = S_{i+1} \cap T_i ; \ldots ; S_{l-1} = \varnothing .$$

Il est immédiat que pour $l \leqslant i \leqslant k$ le diamètre de T_i et donc celui de S_i sont inférieurs à $2b_i$. Nous pouvons alors construire l'estimateur $\tilde{\theta}$: si $\tau = \inf \{ i \mid S_i \neq \varnothing \}$, $\tilde{\theta}$ sera un point arbitraire dans S_τ.

Risque associé à l'estimateur $\tilde{\theta}$.

Par hypothèse, il existe un point θ de Θ tel que la loi Q de $X_1, \ldots, X_n$ s'écrive $Q = \bigotimes_{i=1}^{n} Q_i$ avec $h(P_\theta, Q_i) \leqslant \varepsilon$ pour $1 \leqslant i \leqslant n$, de sorte que sur Θ'_n on peut écrire

$$\mathbb{E}_Q[H^2(P_{\tilde{\theta}}^n, Q)] = \mathbb{E}_Q\left[\sum_{i=1}^{n} h^2(P_{\tilde{\theta}}, Q_i)\right]$$
$$\leqslant n \mathbb{E}_Q [(h(\theta, \tilde{\theta}) + \varepsilon)^2] .$$

Si θ est dans S_i, $\tau \leqslant i$ et $\tilde{\theta} \in S_\tau \subset S_i$ si bien que θ et $\tilde{\theta}$ appartiennent à T_i et $h(\theta, \tilde{\theta}) \leqslant 2b_i$, d'où la majoration

$$(2.8) \quad n^{-1} \mathbb{E}_Q [H^2(\tilde{\theta}, Q)] \leqslant Q [\theta \in S_l](2b_l + \varepsilon)^2$$
$$+ \sum_{i=l}^{k-1} Q [\theta \in S_{i+1} - S_i](2b_{i+1} + \varepsilon)^2 + Q [\theta \notin S_k] .$$

Afin de majorer cette dernière quantité nous démontrerons d'abord quelques lemmes.

LEMME 1. — Soient $P = \bigotimes_{i=1}^{s} P_i$ et $Q = \bigotimes_{i=1}^{s} Q_i$ avec $h(P_i, Q_i) \leqslant \varepsilon$ pour $1 \leqslant i \leqslant s$; alors

$$h^2(P, Q) \leqslant s\varepsilon^2 . \tag{2.9}$$

Preuve. — Il suffit d'utiliser les propriétés de $\rho(P, Q) = \int \sqrt{dPdQ}$ (cf. [1] ou [7]) et l'inégalité $1 - s\varepsilon^2 \leqslant (1 - \varepsilon^2)^s$ pour $s \geqslant 1$ ce qui donne :

$$\begin{aligned} h^2(P, Q) &= 1 - \prod_{i=1}^{s} \int \sqrt{dP_i dQ_i} = 1 - \prod_{i=1}^{s} (1 - h^2(P_i, Q_i)) \\ &\leqq 1 - (1 - \varepsilon^2)^s \leqq \min(1, s\varepsilon^2) . \end{aligned}$$ ■

LEMME 2. — Pour $l \leqslant i \leqslant k$ et $I = 2^{m-i}$

$$Q[h(\hat{\theta}_I(Y_j^I), \theta) > b_i] \leqslant \beta_i = \frac{1}{a} + (2I\varepsilon^2)^{1/2} . \tag{2.10}$$

Preuve. — D'après (2.7) et (2.5) on a

$$P_\theta^I[h(\hat{\theta}_I(Y_j^I), \theta) > b_i] \leqslant \frac{D'}{2^{m-i}b_i^2} = \frac{1}{a} .$$

Notons $Q'_{I,j}$ la loi I-dimensionnelle de Y_j^I, on peut lui appliquer le Lemme 1 de sorte que $h^2(P_\theta^I, Q'_{I,j}) \leqslant I\varepsilon^2$. (2.10) s'ensuit d'après une inégalité classique entre la distance de Hellinger et la distance en variation (cf. [7]). ■

LEMME 3 (d'après Le Cam). — Supposons donnés $b > 0$ et $\beta \leqslant \frac{1}{2}$ et un estimateur $\hat{\theta}(X)$ de θ tel que pour toute probabilité P appartenant à un certain voisinage $\mathscr{V}_\theta$ de θ on ait

$$P[h(\theta, \hat{\theta}) > b] \leqslant \beta .$$

Soient s observations indépendantes $X_1, \ldots, X_s$ de lois respectives Q_i appartenant à $\mathscr{V}_\theta$ alors si $Q = \bigotimes_{i=1}^{s} Q_i$,

$$Q\left[\operatorname{Card}\left[\mathscr{B}(\theta, b) \cap \{ \hat{\theta}(X_1), \ldots, \hat{\theta}(X_s) \} \leqslant \frac{s}{2}\right]\right] \leqslant [4\beta(1-\beta)]^{s/2} . \tag{2.11}$$

Preuve. — Soient $B_\theta = \mathscr{B}(\theta, b)$ et $N_\theta = \sum_{i=1}^{s} 1_{\{\hat{\theta}(X_i) \in B_\theta\}}$. Il s'agit de majorer $Q\left[N_\theta \leqslant \frac{s}{2}\right]$. Or d'après l'inégalité exponentielle

$$Q\left[N_\theta \leqslant \frac{s}{2}\right] \leqslant \prod_{i=1}^{s} \mathbb{E}_Q\left[\exp t\left(\frac{1}{2} - 1_{\{\hat{\theta}(X_i)\in B_\theta\}}\right)\right] \quad \text{si} \quad t > 0,$$

$$= \prod_{i=1}^{s}\left[p_i \exp\left(\frac{t}{2}\right) + (1 - p_i)\exp\left(-\frac{t}{2}\right)\right],$$

si $p_i = P_i[h(\theta, \hat{\theta}) > b]$. Si $\exp\left(\frac{t}{2}\right) = \left(\frac{1-\beta}{\beta}\right)^{1/2}$, comme $p_i \leqslant \beta \leqslant \frac{1}{2}$ on obtient

$$Q\left[N_\theta \leqslant \frac{s}{2}\right] \leqslant (4\beta(1-\beta))^{s/2}. \quad \blacksquare$$

LEMME 4. — Si $l \leqslant i \leqslant k$ et $\beta_l = a^{-1} + (\varepsilon^2 2^{m-l+1})^{1/2} \leqslant 1/2$,

$$(2.12) \quad n^{-1}\mathbb{E}_Q[H^2(Q, \tilde{\theta})] \leqslant (2b_l + \varepsilon)^2\left[1 - \sum_{i=l}^{k} \gamma_i\right] + \sum_{i=l}^{k-1} \gamma_i(2b_{i+1} + \varepsilon)^2 + \gamma_k$$

avec

$$(2.13) \quad \begin{cases} \gamma_i = \{4[a^{-1} + (2\varepsilon^2 2^{m-i})^{1/2}][1 - a^{-1} - (2\varepsilon^2 2^{m-i})^{1/2}]\}^{2^{i-1}} \text{ si } i \geqslant 1 \\ \gamma_0 = a^{-1} + (\varepsilon^2 2^{m+1})^{1/2}). \end{cases}$$

Preuve. — Soit γ_i un majorant de $Q[\theta \in S_{i+1} - S_i]$. Comme

$$Q[\theta \in S_l] = 1 - \sum_{i=l}^{k} Q[\theta \in S_{i+1} - S_i]$$

et que la suite b_i est croissante, (2.8) entraîne facilement (2.12). D'autre part si $\theta \in S_{i+1} - S_i$, $\theta \notin T_i$ et donc Card $[\mathscr{B}(\theta, b_i) \cap \mathscr{P}_I] \leqslant 2^{i-1}$ ce qui d'après le Lemme 3 implique

$$Q[\theta \in S_{i+1} - S_i] \leqslant [4\beta_i(1-\beta_i)]^{2^{i-1}} \quad \text{si} \quad i \geqslant 1, \qquad Q[\theta \notin S_0] \leqslant \beta_0,$$

où β_i est donné par le Lemme 2 ; d'où l'expression de γ_i. $\blacksquare$

Preuve du Théorème 1.

En partant de (2.12) et en remarquant que

$$(2b_{i+1} + \varepsilon)^2 \leqslant 2(2b_i + \varepsilon)^2 \qquad \text{et} \qquad 1 < 2(2b_k + \varepsilon)^2$$

on obtient

$$(2.14)\quad n^{-1}\mathbb{E}_Q[H^2] < (2b_l + \varepsilon)^2\left[1 - \sum_{i=l}^{k}\gamma_i + \sum_{i=l}^{k-1}\gamma_i 2^{i+1-l} + 2\gamma_k 2^{k-l}\right]$$

$$= (2b_l + \varepsilon)^2\left[1 + \sum_{i=l}^{k}\gamma_i(2^{i+1-l} - 1)\right].$$

D'autre part d'après (2.2) et (2.3)

$$\beta_l = a^{-1} + c(2^{m-l+1}n^{-1})^{1/2} \leqslant \beta < 1/2 .$$

Il s'ensuit que si $l = 0\gamma_0 \leqslant \beta$ et si $i \geqslant 1 \vee l$, $\gamma_i \leqslant [4\beta(1-\beta)]^{2^{i-1}}$ étant donné que la fonction $g(x) = x(1-x)$ est croissante pour $0 \leqslant x \leqslant 1/2$. En posant $\gamma = [4\beta(1-\beta)]^{1/2}$ on en déduit que si $i \geqslant 1 \vee l$, $\gamma_i \leqslant \gamma^{2^i}$. (2.14) conduit alors aux majorations suivantes du risque :

$$(2.15)\quad \begin{aligned} &\mathbb{E}_Q(H^2) < n(2b_l + \varepsilon)^2\left[1 + \beta + \sum_{i=1}^{+\infty}\gamma^{2^i}(2^{i+1} - 1)\right] && \text{si}\quad l = 0,\\ &\mathbb{E}_Q(H^2) < n(2b_l + \varepsilon)^2\left[1 + \sum_{j=0}^{+\infty}\gamma^{2^{l+j}}(2^{j+1} - 1)\right] && \text{si}\quad l \geqslant 1. \end{aligned}$$

Les séries sont clairement convergentes puisque $\gamma < 1$. Par ailleurs

$$2b_l + \varepsilon = 2(D'a2^{l-m})^{1/2} + cn^{-1/2} \leqslant 2\lambda c(D'a2^{-m})^{1/2} + c2^{-m/2},$$

donc

$$n(2b_l+\varepsilon)^2 \leqslant 4n2^{-m}(\lambda c)^2 D'a[1+(4\lambda^2 D'a)^{-1/2}]^2 .$$

Étant donné que le choix des constantes a et λ détermine β et γ et implique que $4\lambda^2 D'a > 27$, le théorème est démontré puisque $n2^{-m} < 2$ et D' est arbitrairement proche de D. Si $n = 2^m$, (2.15) prend la forme suivante :

$$R(\Theta'_n) \leqslant 4 \times 2^l Da\left[1 + \frac{c}{2}(Da2^l)^{-1/2}\right]^2\left[s + \sum_{j>1}(2^{j+1} - 1)\Gamma^{2^j}\right]$$

avec

$$\beta = a^{-1} + c2^{(1-l)/2},\quad \gamma^2 = 4\beta(1-\beta),\quad \Gamma = (\gamma^2)^{2^{l-1}}\quad \text{et}\quad s = \begin{cases} 1+\beta \text{ si } l=0 \\ 1+\Gamma \text{ sinon}. \end{cases}$$

Sous cette forme, il n'est pas difficile d'obtenir numériquement des majorations de $K(c)$. Avec $D = 1$ on a trouvé $K(0,1) < 368$, $K(0,01) < 156$, $K(0^+) < 134$. Ces valeurs ne constituent bien entendu que des majorations grossières. ■

REMARQUE 1. — Les constantes obtenues ici paraissent très grandes. Cela tient à la fois au principe même de la méthode et à la façon dont on a conduit les calculs. Dans le principe, la perte essentielle est due à l'utilisation de l'inégalité de Tchebychev dans la démonstration du Lemme 2. En pratique, on pourrait espérer obtenir mieux que a^{-1} dans (2.10) et (2.13).

REMARQUE 2. — La construction précédente est loin d'être optimale et malgré l'utilisation de l'inégalité de Tchebychev, il est possible de l'améliorer sensiblement. Elle a été choisie ici pour sa simplicité qui permet de faire une démonstration courte du Théorème 1. Cependant il est possible de faire mieux en utilisant pour l'estimateur une construction un peu différente et plus complexe. Par ailleurs le Lemme 3 et l'inégalité (2.11) peuvent également être améliorés. On peut ainsi, pour les très petites valeurs de c obtenir une constante K de l'ordre de 60. Nous ne présenterons pas ici cette autre démonstration qui est nettement plus longue et technique que celle que nous avons adoptée sans rien apporter de plus du point de vue théorique. Simplement, on peut imaginer que pour un problème précis, on puisse, en remplaçant également l'inégalité de Tchebychev par une inégalité plus fine, construire un estimateur ayant des propriétés plus « raisonnables » que celles que nous avons obtenues ici.

Il est bien connu que dans la majorité des problèmes statistiques classiques, tant paramétriques que non-paramétriques, le risque converge à des vitesses de l'ordre de $n^{-1/2}$ ou plus lentes. A tous ces problèmes, on peut donc appliquer la théorie précédente et il serait donc raisonnable de rechercher des procédures robustes à des perturbations de taille $n^{-1/2}$ puisque l'on sait que de telles procédures doivent exister. En particulier tous les problèmes paramétriques classiques (à vitesse $n^{-1/2}$) doivent pouvoir être traités de manière robuste.

III. ROBUSTESSE DES ESTIMATEURS DIMENSIONNELS

Dans [*1*] nous avons étudié une classe de problèmes qui se caractérisent par le fait que leur structure de risque est extrêmement liée à leur structure métrique. Nous qualifierons ces problèmes de « dimensionnels » puisque c'est la dimension métrique de Θ qui donne l'allure de la fonction de risque. En particulier nous avons montré que si Θ possédait une structure dimensionnelle donnée par la fonction $L\tilde{h}(n)$, on pouvait construire des estimateurs $\tilde{\theta}_n$ dont le risque est majoré par $CnL\tilde{h}(n)$ et qui possèdent en outre certaines propriétés de robustesse. Nous allons montrer ici que l'on peut modifier ces estimateurs pour les rendre robustes à de plus grosses perturbations, au prix d'une augmentation de la constante C.

PROPOSITION 1 (*). — Étant donnés une expérience Θ_n à laquelle est associée une dimension $\tilde{h}$ et un réel $k \geqslant 0$, considérons l'η-voisinage Θ_n^η de Θ_n défini par $\eta = k\mathrm{L}\tilde{h}(n)$. Alors il existe une constante C' telle que si $n[\mathrm{L}\tilde{h}(n)]^2 \geqslant \frac{1}{7}$ on ait

$$\mathrm{R}(\Theta_n^\eta) \leqslant \mathrm{C}' n[(k+1)\mathrm{L}\tilde{h}(n)]^2 . \tag{3.1}$$

Preuve. — La construction de l'estimateur est strictement identique à celle qui a été donnée dans [1]. On fixe n et $\varepsilon > \mathrm{L}\tilde{h}(n)$ tel que $n\varepsilon^2 > \tilde{h}(\varepsilon)$. Alors d'après la définition de $\tilde{h}$, il existe un ε-réseau S tel que pour t dans Θ et tout entier $j \geqslant 3$

$$\mathrm{Card}\ [\mathscr{B}(t, 2^j\varepsilon) \cap \mathrm{S}] \leqslant \exp\ [jn\varepsilon^2 \operatorname{Log} 2] . \tag{3.2}$$

De plus ε peut être choisi arbitrairement proche de $\mathrm{L}\tilde{h}(n)$. Comme l'ensemble $\overline{\Theta}_n = \bigcup_{s \in \mathrm{S}} \mathscr{B}^{\otimes n}(\mathrm{P}_s, (k+1)\varepsilon)$ contient $\Theta_n^{k\varepsilon}$, il nous suffira de majorer $\mathrm{R}(\overline{\Theta}_n)$.

Étant donnés deux points s et t de Θ tels que $h(s,t) > 2(k+1)\varepsilon$, on sait d'après [3] qu'il existe un test non randomisé de $\mathscr{B}^{\otimes n}(\mathrm{P}_s, (k+1)\varepsilon)$ contre $\mathscr{B}^{\otimes n}(\mathrm{P}_t, (k+1)\varepsilon)$ dont les erreurs décroissent en fonction exponentielle de n et plus précisément qu'il existe une fonction $\varphi_{s,t}$ à valeurs 0 ou 1 telle que

$$\left\{ \begin{array}{l} \mathbb{E}_{\mathrm{P}}[\varphi_{s,t}] \\ \mathbb{E}_{\mathrm{Q}}[\varphi_{t,s}] \end{array} \right. \leqslant \exp\ [-n[h(s,\ t) - 2(k+1)\varepsilon]^2]; \qquad \varphi_{t,s} = 1 - \varphi_{s,t} , \tag{3.3}$$

pour tout P dans $\mathscr{B}^{\otimes n}(\mathrm{P}_s, (k+1)\varepsilon)$ et tout Q dans $\mathscr{B}^{\otimes n}(\mathrm{P}_t, (k+1)\varepsilon)$.

Définissons $\mathrm{T}(s) = \{ t \in \mathrm{S} \mid h(s,t) > 2(k+1)\varepsilon \}$ et

$$\mathrm{L}_s = \sup_{t \in \mathrm{T}(s)} \varphi_{s,t} h(s,t), \qquad \mathrm{L} = \inf_{s \in \mathrm{S}} \mathrm{L}_s .$$

Nous avons montré dans [1] que L_s était fini p. s. et que le minimum L est atteint sur S. Choisissons pour estimateur $\tilde{\theta}$ un quelconque des points t tels que $\mathrm{L} = \mathrm{L}_t$. Il s'ensuit selon [1] que si $h(s, \tilde{\theta}) > r$, $\mathrm{L}_s > r$ de sorte que

$$\mathrm{P}[h(s, \tilde{\theta}) > r] \leqslant \mathrm{P}[\mathrm{L}_s > r] . \tag{3.4}$$

(*) Une autre construction de d-estimateurs robustes est indiquée dans [9], utilisant les résultats de [2]. Elle pourrait être adaptée à la situation présente pour donner une version améliorée de la Proposition 1.

Il est facile de voir en découpant S en couronnes de centre s, en comptant le nombre de points d'une telle couronne, noté

$$a_j = \text{Card}\,[S \cap [\mathscr{B}(s, 2^{j+1}\varepsilon) - \mathscr{B}(s, 2^j\varepsilon)]]$$

et en utilisant (3.3) que si $P = \bigotimes_{i=1}^{n} P_i$ où P_i est dans $\mathscr{B}(s, (k+1)\varepsilon)$, on a

$$\text{(3.5)} \qquad P[L_s > 2^p\varepsilon] \leqslant \sum_{j=p}^{+\infty} \exp\{-n[2^j - 2(k+1)]^2\varepsilon^2\}\, a_j,$$

de sorte que si l'on pose $\operatorname{Log} A_j = -n[2^j - 2(k+1)]^2\varepsilon^2$ il vient selon (3.4)

$$P[h(s, \tilde{\theta}) > 2^p\varepsilon] \leqslant \sum_{j=p}^{+\infty} A_j a_j .$$

Comme $h(P_s, P_i) \leqslant (k+1)\varepsilon$, on a

$$h(P_{\tilde{\theta}}, P_i) \leqslant (k+1)\varepsilon + h(s, \tilde{\theta}) \qquad \text{et} \qquad H^2(P_{\tilde{\theta}}^n, P) \leqslant n[(k+1)\varepsilon + h(s, \tilde{\theta})]^2$$

d'où

$$\mathbb{E}_P[H^2(P, \tilde{\theta})] \leqslant n\varepsilon^2\left[x_p^2 + \sum_{j=p}^{+\infty}\left(\sum_{l=j}^{+\infty} A_l a_l\right)(x_{j+1}^2 - x_j^2)\right], \qquad x_j = 2^j + k + 1$$

puisque $\mathbb{E}_P[H^2(P, \tilde{\theta})] = \int_0^{+\infty} P[H^2(P, \tilde{\theta}) > t]\,dt$. Ceci donne finalement

$$\frac{1}{n\varepsilon^2}\,\mathbb{E}_P[H^2(P, \tilde{\theta})] \leqslant x_p^2 + \sum_{j=p}^{+\infty} A_j a_j (x_{j+1}^2 - x_p^2).$$

Mais d'après (3.2)

$$\text{(3.6)} \qquad b_q = \sum_{j=p}^{q} a_j \leqslant \exp\,[(q+1)n\varepsilon^2 \operatorname{Log} 2], \qquad \text{si} \quad p \geqslant 2,$$

de sorte que

$$\text{(3.7)} \quad \frac{1}{n\varepsilon^2}\,\mathbb{E}_P[H^2(P, \tilde{\theta})] \leqslant x_p^2 + A_p b_p (x_{p+1}^2 - x_p^2) + \sum_{j=p+1}^{+\infty} A_j (b_j - b_{j-1})(x_{j+1}^2 - x_p^2)$$

$$= x_p^2 + \sum_{j=p}^{+\infty} b_j [A_j (x_{j+1}^2 - x_p^2) - A_{j+1}(x_{j+2}^2 - x_p^2)] .$$

Supposons que p soit fixé de sorte que

$$8(k+1) > 2^p \geqslant 4(k+1) \tag{3.8}$$

alors on peut montrer que l'on a $\dfrac{x_{j+2}^2 - x_p^2}{x_{j+1}^2 - x_p^2} < 5$ et $\dfrac{[2^{j+1}-2(k+1)]^2}{[2^j-2(k+1)]^2} > 4$ pour $j \geqslant p$. Comme pour $x \geqslant \dfrac{\text{Log } 5}{3}$ on a $\exp[-4x] \leqslant \dfrac{1}{5}\exp[-x]$ et que $n\varepsilon^2[2^j-2(k+1)]^2 \geqslant \dfrac{\text{Log } 5}{3}$ d'après (3.8) et $n\varepsilon^2 > n[\mathrm{L}\tilde{h}(n)]^2 \geqslant \dfrac{1}{7}$ il s'ensuit que $\mathrm{A}_{j+1} \leqslant \dfrac{\mathrm{A}_j}{5}$ et que la suite $\mathrm{A}_j(x_{j+1}^2 - x_p^2)$ est décroissante pour $j \geqslant p$. Alors en utilisant (3.6) et (3.7) on obtient

$$\begin{aligned}\frac{1}{n\varepsilon^2}\mathbb{E}_{\mathrm{P}}[\mathrm{H}^2(\mathrm{P}, \tilde{\theta})] &\leqslant x_p^2 + \sum_{j=p}^{+\infty} \exp[(j+1)n\varepsilon^2 \text{ Log } 2][\mathrm{A}_j(x_{j+1}^2 - x_p^2) \\ &- \mathrm{A}_{j+1}(x_{j+2}^2 - x_p^2)] \leqslant x_p^2 + \mathrm{A}_p \exp[(p+1)n\varepsilon^2 \text{ Log } 2](x_{p+1}^2 - x_p^2) \\ &+ \sum_{j=p+1}^{+\infty} \mathrm{A}_j\,[\exp[(j+1)n\varepsilon^2 \text{ Log } 2] - \exp[jn\varepsilon^2 \text{ Log } 2]][x_{j+1}^2 - x_p^2].\end{aligned} \tag{3.9}$$

Il est alors facile de voir que pour $n\varepsilon^2 \geqslant \dfrac{1}{7}$, cette expression est majorée par $\mathrm{C}_1 x_p^2$ où C_1 est une constante universelle. Vu la définition (3.8) de p on en déduit

$$\mathbb{E}_{\mathrm{P}}[\mathrm{H}^2(\mathrm{P}, \tilde{\theta})] \leqslant n\varepsilon^2 \mathrm{C}_1 x_p^2 < 81\mathrm{C}_1(k+1)^2 n\varepsilon^2 ;$$

la conclusion s'ensuit puisque ε est arbitrairement proche de $\mathrm{L}\tilde{h}(n)$. ■

Remarques. — 1) Dans un cas précis où $n\varepsilon^2$ est donné ainsi que k, on a intérêt pour évaluer le risque à se reporter à (3.9), un calcul numérique fournissant aisément le résultat y compris pour $k = 0$.

2) On remarquera que, aux constantes près, (3.1) est optimal puisque selon [*4*] le risque minimax pour une boule de Hellinger de rayon r sur un espace infini est nr^2 ; comme ici Θ_n' contient des boules de rayon $k\mathrm{L}\tilde{h}(n)$, on ne peut faire mieux que $n[k\mathrm{L}\tilde{h}(n)]^2$.

COROLLAIRE 1. — Soient Θ_n une expérience supposée a-dimensionnelle et D une borne supérieure du risque : $\mathrm{D} \geqslant \mathrm{R}(\Theta_n)$. Alors cette expérience est stable aux perturbations de taille $c(\mathrm{D}/n)^{1/2}$ et le risque pour le problème perturbé Θ'_n est majoré par

$$\text{(3.10)} \qquad R(\Theta'_n) \leqslant 4C'D \sup\left[c^2, \frac{1}{a}\right],$$

si $nL\tilde{h}(n) \geqslant \frac{1}{7}$.

Preuve. — Par hypothèse $D \geqslant R(\Theta_n) \geqslant an[L\tilde{h}(n)]^2$. Fixons

$$k = [L\tilde{h}(n)]^{-1}c\left(\frac{D}{n}\right)^{1/2}$$

dans la Proposition 1. Le risque pour Θ'_n sera donc majoré par (3.1). Si $k \geqslant 1$

$$R(\Theta'_n) \leqslant C'n[(k+1)L\tilde{h}(n)]^2 \leqslant C'n\left[\frac{k+1}{k}c\left(\frac{D}{n}\right)^{1/2}\right]^2 \leqslant 4C'c^2D.$$

Si $k < 1$

$$R(\Theta'_n) \leqslant 4C'n[L\tilde{h}(n)]^2 \leqslant \frac{4C'D}{a},$$

ce qui achève la démonstration. ■

Remarque 1. — Dans tout ce qui précède, la restriction $n[L\tilde{h}(n)]^2 \geqslant \frac{1}{7}$ n'en est pas réellement une puisqu'elle équivaut au fait que $\tilde{h}(x) \geqslant \frac{1}{7}$ lorsque x est petit, ce qui revient à dire que l'espace Θ n'est pas trop lacunaire, cas pathologique que l'on souhaite bien évidemment exclure.

Remarque 2. — Les résultats précédents s'étendent sans aucune modification au cas où Θ''_n est défini comme l'ensemble des probabilités produits Q sur Ω^n telles qu'il existe θ dans Θ et $H(P^n_\theta, Q) \leqslant n^{1/2}\eta$. Ceci est dû aux propriétés des tests entre deux boules pour la distance H et démontre une propriété de robustesse à des perturbations plus larges. On trouvera les détails dans [*2*]; voir aussi [*9*].

Les exemples relevant des techniques précédentes sont très nombreux. En particulier les problèmes classiques d'estimation des densités en font partie. Dans ces problèmes, la vitesse d'estimation est en général de l'ordre de $n^{-\alpha}$ avec $\alpha < 1/2$ et elle est liée à la dimension de l'espace des paramètres ainsi qu'il est expliqué dans [*0*] et [*1*]. Il s'ensuit que tous ces problèmes dont on trouvera des exemples dans [*5*] ou [*6*], sont stables aux perturbations de taille $cn^{-\alpha}$ pour un α correspondant à la vitesse initiale de convergence du risque. A titre d'illustration on peut considérer le cas particulier suivant : soit à estimer une densité unimodale sur $\mathbb{R}$, la perte étant calculée à partir de la distance déduite de la norme $\mathbb{L}^1$. Dans ce cas la vitesse de convergence sera $n^{-1/3}$. On pourra donc, avec un estimateur convenable, estimer de la même manière une densité « presque unimodale » en ce sens qu'elle est éloignée d'au plus $cn^{-1/3}$ d'une densité unimodale. Ici le remplacement de la distance de Hellinger par la distance en variation ne change rien à la stabilité, tout ce chapitre pouvant être traité aussi bien avec l'une ou l'autre distance étant donné leurs relations (cf. [*1*]).

IV. INSTABILITÉ DES PROBLÈMES NON DIMENSIONNELS

D'après ce qu'on vient de voir, les problèmes dimensionnels sont stables à des perturbations de taille $c\left(\frac{R(\Theta_n)}{n}\right)^{1/2}$ et l'on peut se demander ce qu'il en est dans le cas général. Toutefois on remarque que si l'on s'éloigne de la dimensionnalité, (lorsque a devient petit) la constante au second membre de (3.10) augmente. On pourrait croire que cette tendance à l'instabilité de la borne si a tend vers 0 est un effet des calculs mais nous allons voir en exhibant une famille de contre-exemples qu'il n'en est rien et que le Corollaire 1 est faux dans un cadre général. Il s'agit de problèmes pathologiques, perturbations de problèmes dimensionnels, mais n'ayant plus cette propriété. En fait on peut construire des contre-exemples à partir de tout problème dimensionnel dont le risque tend vers l'infini avec le nombre des observations. On peut même renforcer les conditions afin de se replacer dans le cadre du Théorème 1 et voir que le risque pour le problème perturbé n'est pas seulement très grand devant $R(\Theta_n)$ mais aussi devant $\sup_{m>0} R(\Theta_m)$.

Afin de décrire ces expériences, nous étudierons d'abord un type particulier de perturbation d'un problème statistique. Considérons un ensemble $\overline{\Theta}_1 = \{ P_\theta, \theta \in \overline{\Theta} \}$ formé de probabilités sur $(\Omega, \mathscr{A})$, ne chargeant pas les points et supposons que l'espace Ω est infini et qu'il existe un ε-réseau S dénombrable dans $\overline{\Theta}$. A S, associons une partition $\{ A_s \}_{s\in S}$ de $\overline{\Theta}$ telle que pour tout s, $s \in A_s \subset \mathscr{B}(s, \varepsilon)$, ainsi qu'une collection dénombrable $\{ x_s \}_{s\in S}$ de points distincts de Ω. Nous définirons la perturbation $\Theta_1(\lambda)$ de $\overline{\Theta}_1$ comme l'ensemble des probabilités P_{θ_λ}, $\theta_\lambda \in \Theta(\lambda)$, de la forme

$$P_{\theta_\lambda} = (1 - \lambda)P_\theta + \lambda\delta_{x_s} \quad \text{si} \quad \theta \in A_s, \tag{4.1}$$

δ_x désignant la probabilité portée par le point x. (4.1) permet de définir une bijection j de $\overline{\Theta}$ sur $\Theta(\lambda)$ par $j(\theta) = \theta_\lambda$. Les propriétés suivantes sont claires pour tout estimateur T à valeurs dans $\mathscr{M}_1$.

$$h^2(\theta, j(\theta)) = 1 - (1 - \lambda)^{1/2} \leqslant \lambda, \tag{4.2}$$

$$h^2(j(\theta), \mathrm{T}) \leqslant 2[h^2(\theta, \mathrm{T}) + \lambda], \tag{4.3}$$

$$h^2(j(\theta), \mathrm{T}) \geqslant \frac{1}{2} h^2(\mathrm{T}, \theta) - \lambda. \tag{4.4}$$

(4.2) implique que $\Theta_n(\lambda)$ est une $\lambda^{1/2}$-perturbation de $\overline{\Theta}_n$ et (4.3), (4.4) entraînent les relations suivantes entre les risques minimax des deux expériences.

LEMME 5. — Pour tout $n \geqslant 1$, les inégalités suivantes sont satisfaites :

$$(1-\lambda)^n\left[\frac{1}{2}\mathrm{R}(\overline{\Theta}_n) - n\lambda\right] \leqslant \mathrm{R}(\Theta_n(\lambda)) \leqslant 2(1-\lambda)^n[n\lambda + \mathrm{R}(\overline{\Theta}_n)] + (1-\lambda)\varepsilon^2 n. \tag{4.5}$$

Preuve. — Soit $\mathrm{F} = \{ x_s \}_{s \in \mathrm{S}}$, S étant dénombrable et P_θ ne chargeant pas les points, on a $\mathrm{P}_\theta^n(\mathrm{F}^{\otimes n}) = 0$ pour θ dans $\overline{\Theta}$. Posons $\mathrm{F}_n = (\mathrm{F}^c)^{\otimes n}$ et désignons par $\hat{\theta}_n$ un estimateur quelconque de θ_λ à valeurs dans $\mathscr{M}_1$ construit à partir de n observations. Alors $\mathrm{R}(\Theta_n(\lambda)) = n[\mathrm{R}_1 + \mathrm{R}_2]$ où R_1 et R_2 sont définis par

$$\left\{ \begin{aligned} \mathrm{R}_1 &= \inf_{\hat{\theta}_n} \sup_{\theta_\lambda \in \Theta(\lambda)} \mathrm{E}_{\theta_\lambda}[h^2(\hat{\theta}_n, \theta_\lambda) 1_{\mathrm{F}_n}], \\ \mathrm{R}_2 &= \inf_{\hat{\theta}_n} \sup_{\theta_\lambda \in \Theta(\lambda)} \mathrm{E}_{\theta_\lambda}[h^2(\hat{\theta}_n, \theta_\lambda) 1_{\mathrm{F}_n^c}]. \end{aligned} \right. \tag{4.6}$$

Si $\theta_\lambda = j(\theta)$ est l'image d'un point θ de A_s, p. s. les observations qui appartiennent à F sont égales à x_s, de sorte que si l'on choisit $\hat{\theta}_n = j(s)$ comme estimateur lorsque toutes les observations qui tombent dans F valent x_s, on aura p. s. sur F_n^c

$$h^2(\hat{\theta}_n, \theta_\lambda) = h^2(j(s), j(\theta_\lambda)) = (1 - \lambda) h^2(s, \theta_\lambda) \leqslant (1 - \lambda)\varepsilon^2.$$

On en déduit

$$0 \leqslant \mathrm{R}_2 \leqslant (1 - \lambda)\varepsilon^2. \tag{4.7}$$

D'autre part si X a pour loi $\mathrm{P}_{\theta_\lambda}$, la loi conditionnelle de X sachant que X est dans F^c est P_θ avec $\theta = j^{-1}(\theta_\lambda)$, ce qui entraîne

$$\mathbb{E}_{\theta_\lambda}[h^2(\hat{\theta}_n, \theta_\lambda) 1_{\mathrm{F}_n}] = \mathbb{E}_\theta[h^2(\hat{\theta}_n, \theta_\lambda)](1 - \lambda)^n. \tag{4.8}$$

Il s'ensuit en utilisant (4.3) et (4.4) que

$$(1-\lambda)^n\left[\frac{1}{2}\,\mathrm{R}(\overline{\Theta}_n) - n\lambda\right] \leqslant n\mathrm{R}_1 \leqslant 2(1-\lambda)^n[\mathrm{R}(\overline{\Theta}_n) + n\lambda]$$

ce qui joint à (4.7) permet de conclure. ■

On peut alors démontrer :

THÉORÈME 3. — Soit Θ un espace de paramètres séparable tel que les risques $r(n) = \mathrm{R}(\overline{\Theta}_n)$ associés tendent vers l'infini. Choisissons une suite $\{ k_n \}_{n\geqslant 1}$, $k_n < n$, tendant vers l'infini assez lentement pour que

$$k_n \exp(k_n) = o(r(n)). \tag{4.9}$$

Alors il existe une suite d'expériences $\{ \Theta_n(\lambda_n) \}$ avec $\lambda_n = n^{-1}k_n$ telles que $\Theta_n(\lambda_n)$ soit une $\lambda_n^{1/2}$-perturbée de $\overline{\Theta}_n$ dont le risque vérifie

$$k_n = o(\mathrm{R}(\Theta_n(\lambda_n)))\,;\ \mathrm{R}(\Theta_n(\lambda_n)) = o(r(n)). \tag{4.10}$$

S'il existe en outre des constantes α et $\delta > 0$ telles que pour tous les n suffisamment grands on ait pour $m < n$

$$\frac{r(m)}{r(n)} \leqslant \alpha\left(\frac{m}{n}\right)^\delta \tag{4.11}$$

on obtiendra également

$$\sup_{m>0} \mathrm{R}(\Theta_m(\lambda_n)) = o(r(n)). \tag{4.12}$$

Remarque. — L'existence d'expériences $\overline{\Theta}_n$ ayant des propriétés de risque du type précédent ne pose pas de problème. On trouvera en particulier dans [*1*] [*5*] et [*6*] des exemples nombreux de fonctions de risque à croissance polynomiale qui vérifient (4.11). On peut par exemple prendre pour $\overline{\Theta}$ l'ensemble des densités L-lipschitziennes sur $[0,1]$ comprises entre $1/2$ et $3/2$.

Preuve du théorème. — Définissons $\Theta(\lambda_n)$ comme perturbé de $\overline{\Theta}$ ainsi qu'il a été indiqué en (4.1) en remplaçant ε par $n^{-1/2}$ et λ par λ_n. Il s'ensuit que pour tout $m > 0$ $\Theta_m(\lambda_n)$ est une $\lambda_n^{1/2}$-perturbation de $\overline{\Theta}_m$ et d'après (4.5) que l'on a

$$(1-\lambda_n)^m\left[\frac{r(m)}{2} - m\lambda_n\right] \leqslant \mathrm{R}(\Theta_m(\lambda_n)) \leqslant 2(1-\lambda_n)^m[r(m)+m\lambda_n]+(1-\lambda_n) \tag{4.13}$$

Étant donné que $xe^{-x} \leqslant e^{-1}$ si $x > 0$, on a

$$(1 - \lambda_n)^m m\lambda_n < m\lambda_n \exp(-m\lambda_n) \leqslant e^{-1},$$

donc

$$\frac{1}{2}r(m)(1-\lambda_n)^m - e^{-1} \leqslant \mathrm{R}(\Theta_m(\lambda_n)) \leqslant 2r(m)(1-\lambda_n)^m + 2e^{-1} + 1. \tag{4.14}$$

D'autre part comme $r(n) \leqslant n$, (4.9) implique que $k_n = \mathcal{O}(\operatorname{Log} n)$ et que $\lambda_n = \mathcal{O}\left(\frac{\operatorname{Log} n}{n}\right)$. Il s'ensuit que

$$(1-\lambda_n)^n r(n) \sim r(n) \exp(-n\lambda_n) = r(n) \exp(-k_n) = o(r(n))$$

ce qui d'après (4.9) et (4.14) entraîne (4.10). On a également

$$\mathrm{R}(\Theta_m(\lambda_n))r^{-1}(n) \leqslant 2\frac{r(m)}{r(n)}(1-\lambda_n)^m + (2e^{-1}+1)r^{-1}(n),$$

si bien que pour démontrer (4.12) il suffit de vérifier que

$$\sup_{m>0}\left[\operatorname{Log}\left(\frac{r(m)}{r(n)}\right) - \frac{m}{n}k_n\right] \xrightarrow[n\to+\infty]{} -\infty. \tag{4.15}$$

La décroissance en p de $p^{-1}r(p)$ implique que si $m \geqslant n$ et $k_n \geqslant 1$,

$$\operatorname{Log}\left(\frac{r(m)}{r(n)}\right) - \frac{m}{n}k_n \leqslant \operatorname{Log}\left(\frac{m}{n}\right) - \frac{m}{n}k_n \leqslant -k_n$$

ce qui prouve (4.15) pour $m \geqslant n$. Si $m < n$, d'après (4.11) pour n assez grand

$$\operatorname{Log}\left(\frac{r(m)}{r(n)}\right) - \frac{m}{n}k_n \leqslant \operatorname{Log}\alpha + \delta\operatorname{Log}\left(\frac{m}{n}\right) - \frac{m}{n}k_n.$$

En supposant que $k_n \geqslant \delta$ cette quantité est majorée par

$$\operatorname{Log}\alpha + \delta\,[\operatorname{Log}\delta - \operatorname{Log}k_n - 1],$$

ce qui démontre (4.15) pour $m < n$. ■

On peut alors déduire aisément les résultats d'instabilité que nous avons annoncés du corollaire suivant :

Corollaire 2. — Quelles que soient les constantes c, $\mathrm{K} > 0$, il existe des entiers n et des expériences Θ_m avec $m = 1; 2; \ldots; n$ telles que

i) $\mathrm{R}(\Theta_m) \leqslant \mathrm{D}_n$ pour $1 \leqslant m \leqslant n$,

ii) si $\varepsilon_n = c\left[\frac{\mathrm{R}(\Theta_n)}{n}\right]^{1/2}$, $\mathrm{R}(\Theta_n^{\varepsilon_n}) > \mathrm{KD}_n$.

Preuve. — Il suffit de considérer un ensemble de paramètres $\overline{\Theta}$ satisfaisant aux hypothèses du Théorème, y compris (4.11) et de poser $\Theta_m = \Theta_m(\lambda_n)$, $D_n = \sup_{m>0} R(\Theta_m(\lambda_n))$. D'après (4.10) si n est assez grand $\lambda_n = \frac{k_n}{n} < \varepsilon_n^2$ de sorte que $\Theta_n^{\varepsilon_n}$ contient $\overline{\Theta}_n$ et donc $R(\Theta_n^{\varepsilon_n}) \geqslant r(n)$. Comme d'après (4.12) $D_n = o(r(n))$, on obtiendra le Corollaire si n est convenablement choisi. ■

Le résultat précédent appelle un certain nombre de remarques :

i) Si les problèmes $\overline{\Theta}_n$ sont dimensionnels, $r(n)$ est de l'ordre de $n\mathrm{L}\tilde{h}^2(n)$, donc $\lambda_n = \frac{k_n}{n} = o(\mathrm{L}\tilde{h}^2(n))$.

Il s'ensuit sans difficulté que j est presque isométrique en ce sens que $\frac{h(j(\theta), j(\theta'))}{h(\theta, \theta')}$ est voisin de 1 si $h(\theta, \theta') \geqslant \mathrm{L}\tilde{h}(n)$, donc que les fonctions $\tilde{h}(x)$ correspondant à Θ et $\Theta(\lambda_n)$ seront semblables pour $x \geqslant \mathrm{L}\tilde{h}(n)$ ce qui implique que les fonctions $\mathrm{L}\tilde{h}$ correspondant aux deux problèmes seront comparables. Il s'ensuit que $\Theta(\lambda_n)$ a un risque beaucoup plus petit que celui donné par la dimension.

ii) A partir de (4.14) et des évaluations précédentes on voit que $r^{-1}(n)R(\Theta_m(\lambda_n))$ a le même comportement asymptotique que $\frac{r(m)}{r(n)} \exp\left(-\frac{m}{n}k_n\right)$. Ceci implique que $r'(m) = R(\Theta_m(\lambda_n))$ aura des variations inhabituelles, tendant vers 0 lorsque m tend vers l'infini et passant par un maximum qui sera grand devant $r'(n)$ et atteint pour une valeur m inférieure à n si r a la croissance polynomiale décrite par (4.11).

iii) Le phénomène d'instabilité précédent n'est pas uniquement dû au fait que les probabilités considérées ne sont pas absolument continues les unes par rapport aux autres. On se convaincra facilement qu'il serait possible de construire des exemples analogues à partir de mesures mutuellement absolument continues de la manière suivante. Partons d'un $\overline{\Theta}$ formé de densités par rapport à la mesure de Lebesgue sur [0, 1], uniformément comprises entre $\frac{1}{2}$ et $\frac{3}{2}$ et donnant lieu à une fonction $r(n)$ telle que $\alpha n^\gamma \leqslant r(n) \leqslant \beta n^\gamma$, $\alpha, \beta, \gamma > 0$ et considérons un ε-réseau fini S de $\overline{\Theta}$ avec $\varepsilon = \mathrm{L}\tilde{h}(n)$. Dans les cas classiques d'estimation de densité, de tels réseaux existent. Aux points de S on n'associera plus des masses de Dirac δ_{x_s} mais des densités de la forme $a1_{B_s} + a^{-1}1_{B_s^c}$, les B_s étant des intervalles disjoints de même longueur $(a+1)^{-1}$, ce qui est possible si a est assez grand. Le cas que nous avons considéré est en quelque sorte le cas limite lorsque $a = +\infty$, pour lequel les calculs se simplifient, mais il est bien clair qu'en utilisant de telles perturbations construites à partir d'une suite $\{ a_n \}_{n \geq 1}$ tendant vers l'infini suffisamment vite, on obtiendra des résultats analogues aux précédents.

V. CONCLUSION

Il ressort de ce qui précède que si tous les problèmes considérés possèdent une certaine forme de stabilité à de petites perturbations (de l'ordre de $n^{-1/2}$), seuls des modèles relativement réguliers seront stables à des perturbations plus importantes $\left(\text{de l'ordre de } \left[\frac{R(\Theta_n)}{n}\right]^{1/2} \right)$ de la taille de l'erreur moyenne d'estimation. C'est la notion de problème dimensionnel qui traduit cette forme de régularité en l'absence de laquelle l'instabilité se manifeste sur certains exemples. Il est tentant de penser que cette instabilité est caractéristique des problèmes non-dimensionnels et se manifestera dans tous les cas, mais un tel résultat englobant tous les problèmes non-dimensionnels semble difficile à démontrer étant donné qu'il n'existe aucune théorie générale concernant la structure du risque minimax pour de telles expériences. Néanmoins, cette instabilité laisse à penser que ces problèmes sont pathologiques et ne peuvent donner lieu à une théorie générale de l'estimation satisfaisante et que seuls les problèmes dimensionnels, par leur relative stabilité aux perturbations, constituent de « bons modèles ».

Par ailleurs les constructions que nous avons utilisées ici pour mettre en évidence certains phénomènes d'instabilité sont évidemment « ad hoc » et il n'est guère plausible que de tels phénomènes se présentent en pratique au statisticien. Donc, d'un point de vue heuristique, on peut admettre qu'un problème pratique d'estimation sera stable à des perturbations du même ordre de grandeur que son risque minimax et qu'il convient de rechercher des méthodes robustes aux perturbations de cette amplitude.

BIBLIOGRAPHIE

[*0*] L. BIRGÉ, *Thèse*, Université Paris VII, 1980.

[*1*] L. BIRGÉ, Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, t. **65**, 1983, p. 181-237.

[*2*] L. BIRGÉ, Robust testing for independent non-identically distributed variables and Markov chains. Dans : *Specifying statistical models*, p. 134-162. New York-Heidelberg-Berlin: Springer, 1983.

[*3*] L. BIRGÉ, Sur un thèorème de minimax et son application aux tests. *Probab. Math. Statist. III*, t. **2**, 1984.

[*4*] L. BIRGÉ, Non-asymptotic minimax risk for Hellinger balls (1982). *A paraître dans Probab. Math. Statist. V.*

[5] J. BRETAGNOLLE, C. HUBER, Estimation des densités : risque minimax. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, t. **47**, 1979, p. 119-137.

[6] I. A. IBRAGIMOV, R. Z. KHAS'MINSKII, On estimate of the density function. En russe. *Zap. Nauchn. Semin. LOMI*, t. **98**, 1980, p. 61-85.

[7] L. LE CAM, Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, t. **1**, 1973, p. 38-53.

[8] L. LE CAM, On local and global properties in the theory of asymptotic normality of experiments. *Stochastic processes and related topics*, t. **1**, p. 13-54. New York-San Francisco-London: Academic Press, 1975.

[9] L. LE CAM, On the risk of Bayes estimates. *Statistical Decision Theory and Related Topics III*, t. **2**, 1982, p. 121-137.

(Manuscrit reçu le 12 novembre 1982)

(modifié le 2 avril 1984)