Compléments : Processus empiriques et applications statistiques

Les processus empiriques constituent une classe particulière de processus stochastiques où le paramètre vit dans un espace fonctionnel. Dans ces notes, on donnera les notions de base qui interviennent dans l'étude de ces processus et on donnera une idée de leurs applications statistiques.

1 Motivations

1.1 M-estimation

Modèle 1: modèle logistique paramétrique

On observe des couples $Z_i = (X_i, Y_i)$ i.i.d. où les X_i sont des variables aléatoires sur $\mathbb{R}$ et les Y_i sont des variables aléatoires binaires sur $\{0, 1\}$.

Le modèle statistique est défini par la donnée de (μ, η) la loi de (X_1, Y_1) où

$$\begin{aligned} \forall A \in \mathcal{B}(\mathbb{R}), \quad \mu(A) &= \mathbb{P}(X \in A) \\ \forall x \in \mathbb{R}, \quad \eta(x) &= \mathbb{P}(Y = 1 \mid X = x) \end{aligned}$$

On suppose que la fonction de régression η est de la forme paramétrique suivante

$$\forall x \in \mathbb{R}, \quad \eta(x) = F(\theta x)$$

où F est connue, et $\theta \in \mathbb{R}$ est le paramètre à estimer. Dans la suite, on fixe $F(u) = \frac{e^u}{1+e^u}$.

On donne également la notation suivante pour la densité conditionnelle de Y sachant X

$$p_\theta(y \mid x) = (F(\theta x))^y (1 - F(\theta x))^{1-y} .$$

Comme exemple de M-estimateur, on considère l'estimateur du maximum de vraisemblance

$$\hat{\theta}_n = \arg\max_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \log p_\theta(Y_i \mid X_i)$$

dont on souhaite caractériser le comportement asymptotique.

On introduit

$$q_\theta(x, y) = \frac{\partial}{\partial \theta} \log p_\theta(y \mid x) = x(y - F(\theta x))$$

puisque $F'(u) = F(u)(1 - F(u))$ pour le choix particulier de la fonction F considéré ici.

Soit

$$g_\theta(x) = \begin{cases} \dfrac{q_\theta(x, y) - q_{\theta_0}(x, y)}{\theta - \theta_0} = -x \left(\dfrac{F(\theta x) - F(\theta_0 x)}{\theta - \theta_0} \right) & \text{si } \theta \neq \theta_0 \\ -x^2 F(\theta_0 x)(1 - F(\theta_0 x)) & \text{si } \theta = \theta_0 . \end{cases}$$

Lemme 1 *Supposons que*

$$\textbf{(H)} \qquad \frac{1}{n}\sum_{i=1}^{n} g_{\hat{\theta}_n}(X_i) \xrightarrow{P} \mathbb{E}(g_{\theta_0}(X_1)) := I_{\theta_0} > 0$$

Alors

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{I_{\theta_0}}\right) .$$

Preuve. On peut voir que $\hat{\theta}_n$ est un Z-estimateur vérifiant

$$\sum_{i=1}^{n} q_{\hat{\theta}_n}(X_i, Y_i) = 0 .$$

On a alors, par définition de g_θ :

$$\sum_{i=1}^{n} q_{\theta_0}(X_i, Y_i) + (\hat{\theta}_n - \theta_0)\sum_{i=1}^{n} g_{\hat{\theta}_n}(X_i) = 0$$

et donc :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_{\theta_0}(X_i, Y_i)}{\frac{1}{n}\sum_{i=1}^{n} g_{\hat{\theta}_n}(X_i)} .$$

Or, d'après le théorème central limite, on a :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_{\theta_0}(X_i, Y_i) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_{\theta_0})$$

puisque

$$\mathbb{E}_{\theta_0} q_{\theta_0}(X, Y) = 0$$
$$\mathbb{V}_{\theta_0} q_{\theta_0}(X, Y) = I_{\theta_0} .$$

On utilise alors l'hypothèse **(H)** du théorème et, en appliquant le théorème de Slutsky, le lemme est démontré. ■

Commentaire :

- Retour sur l'hypothèse **(H)** : il s'agit d'une généralisation de la Loi des Grands Nombres; on peut se contenter de démontrer (exercice) que

$$\sup_{\theta}\left(\frac{1}{n}\sum_{i=1}^{n} g_\theta(X_i) - \mathbb{E}g_\theta(X)\right) \xrightarrow{P} 0 .$$

- Une autre méthode pour démontrer la normalité asymptotique est d'utiliser l'équicontinuité asymptotique du processus

$$\theta \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(q_\theta(X_i, Y_i) - \mathbb{E}q_\theta(X, Y)\right) .$$

Modèle 2: modèle non-paramétrique

On considère le même modèle que précédemment sauf qu'on abandonne la structure paramétrique de la fonction de régression. On suppose ici que $\eta(x) = F(x)$, $\forall x \in \mathbb{R}$, avec F croissante, à valeurs dans $[0,1]$ et le "paramètre" à estimer est à rechercher dans l'espace fonctionnel suivant:

$$\Theta = \{F \ : \mathbb{R} \to [0,1], F \text{ croissante}\}\,.$$

L'estimateur du maximum de vraisemblance est défini par

$$\hat{F}_n = \arg\max_{F\in\Theta} \sum_{i=1}^{n} (Y_i \log F(X_i) + (1-Y_i)\log(1-F(X_i)))\,.$$

Maintenant, plusieurs questions apparaissent.

1. comment dériver dans l'espace des paramètres ?
2. comment mesurer la distance entre $\hat{F}_n$ et la *vraie* F_0 ?

On ne donnera pas de réponse précise à ces questions mais seulement la nature des résultats dans le cas non-paramétrique. Auparavant, on rappelle la notation O_P.

Définition 2 *Soit (Z_n) une suite aléatoire et k_n une suite déterministe. On dira que $Z_n = O_P(k_n)$ si et seulement si*

$$\lim_{T\to+\infty} \limsup_{n\to+\infty} \mathbb{P}(|Z_n| > Tk_n) = 0\,.$$

Si $k_n \to 0$, on dira que k_n est la vitesse de convergence de Z_n (vers 0).

Si on mesure avec la distance L_2, on obtient des vitesses de convergence :

$$\|\hat{F}_n - F_0\| = O_P(n^{-1/3})\,,$$

à comparer avec $\hat{\theta}_n - \theta_0 = O_P(n^{-1/2})$ dans le cas paramétrique.

En conclusion : il y a un prix à payer pour assouplir la structure du modèle et passer du modèle paramétrique (Modèle 1) au modèle non paramétrique (Modèle 2).

Remarque 3 *Il existe aussi des modèles intermédiaires dits semi-paramétriques où on impose des contraintes supplémentaires sur F (par exemple, sur ses dérivées). On obtient alors typiquement des vitesses en $O_P(n^{-2/5})$.*

1.2 Classification

On considère un modèle statistique similaire au précédent mais un peu plus général. Les observations sont des copies i.i.d. du couple (X, Y) où X est une quantité aléatoire à valeurs sur un espace mesurable $(\mathcal{X}, \mathcal{A})$ et Y est une variable aléatoire binaire sur $\{0, 1\}$. Le modèle est décrit comme précédemment par le couple (μ, η) constitué par μ la loi marginale de X et η la fonction de régression.

Ce qui distingue la classification de l'approche statistique habituelle, c'est qu'on s'intéresse avant tout au problème de la prédiction du label Y sur la base de l'observation X (et pas tant à identifier la loi générant les observations qui est un problème plus difficile).

Définition 4 *On introduit la terminologie de la classification :*

- *Un* ***classifieur*** *est une fonction mesurable*

$$g : \mathcal{X} \to \{0, 1\} .$$

- *L'****erreur du classifieur*** g *est donnée par*

$$L(g) = \mathbb{P}(g(X) \neq Y) .$$

- *Le* ***classifieur de Bayes*** *est défini par*

$$g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}$$

*et l'****erreur de Bayes*** *correspond à* $L^* = L(g^*)$.

La notion de classifieur de Bayes correspond à la situation optimale comme le montre la proposition suivante.

Proposition 5 *Pour tout classifieur* g*, on a :* $L^* \leq L(g)$.

Preuve. On écrit d'abord l'erreur en conditionnant par $X = x$.

$$\begin{aligned}\mathbb{P}(g(X) \neq Y \mid X = x) &= \mathbb{P}(g(X) = 0, Y = 1 \mid X = x) + \mathbb{P}(g(X) = 1, Y = 0 \mid X = x) \\ &= \eta(x)\, \mathbb{I}_{\{g(X)=0\}} + (1 - \eta(x))\, \mathbb{I}_{\{g(X)=1\}} \\ &= (1 - \eta(x)) + (2\eta(x) - 1)\, \mathbb{I}_{\{g(X)=0\}} .\end{aligned}$$

On évalue alors la différence :

$$\mathbb{P}(g(X) \neq Y \mid X = x) - \mathbb{P}(g^*(X) \neq Y \mid X = x) = (2\eta(x) - 1)(\mathbb{I}_{\{g(X)=0\}} - \mathbb{I}_{\{g^*(X)=0\}}) \geq 0$$

par définition de g^*. Pour finir, on intègre en x. ■

Proposition 6 *On a, pour tout* g *:*

$$L(g) - L^* = \mathbb{E}(|2\eta(X) - 1| \cdot \mathbb{I}_{\{g(X) \neq g^*(X)\}}) .$$

De plus :

$$L^* = \mathbb{E}(\min\{\eta(X), 1 - \eta(X)\}) .$$

Idéalement, on souhaiterait, sans faire d'hypothèses sur la loi sous-jacente (μ, η) et à partir des observations i.i.d. $(X_1, Y_1), ..., (X_n, Y_n)$, construire un estimateur $\hat{g}_n$ qui s'approche du classifieur optimal g^* au sens de l'erreur de classification L. En fait, on va se fixer un objectif moins ambitieux. A cet effet, on se donne une classe $\mathcal{C}$ de classifieurs et on cherche un estimateur approchant le meilleur dans la classe.

On considère maintenant le critère empirique suivant.

Définition 7 *On définit l'erreur empirique associée au classifieur g sur l'échantillon $D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ par*

$$\hat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{g(X_i) \neq Y_i\}} .$$

On note

$$\hat{g}_n = \arg\min_{g \in \mathcal{C}} \hat{L}_n(g)$$

l'estimateur résultant de l'optimisation du critère. On souhaite contrôler le risque suivant :

$$\Delta_n = L(\hat{g}_n) - \inf_{g \in \mathcal{C}} L(g) .$$

Remarque 8 *Il faut noter que $L(\hat{g}_n)$ est une quantité aléatoire. En effet, on peut l'écrire*

$$L(\hat{g}_n) = \mathbb{P}\left(\hat{g}_n(X) \neq Y \mid (X_1, Y_1)...(X_n, Y_n)\right) .$$

Remarque 9 *Pour avoir un contrôle sur l'erreur globale $L(\hat{g}_n) - L^*$, on doit traiter le problème d'approximation du classifieur de Bayes par les éléments de $\mathcal{C}$. En effet, l'erreur globale peut se décomposer en deux termes quantifiant chacun l'erreur d'estimation et l'erreur d'approximation :*

$$L(\hat{g}_n) - L^* = \underbrace{(L(\hat{g}_n) - \inf_{g \in \mathcal{C}} L(g))}_{estimation} + \underbrace{(\inf_{g \in \mathcal{C}} L(g) - L^*)}_{approximation} .$$

En utilisant le fait que $\hat{g}_n$ minimise l'erreur empirique, on obtient le contrôle suivant suivant sur l'erreur d'estimation Δ_n :

$$\begin{aligned} \Delta_n &= L(\hat{g}_n) - \hat{L}_n(\hat{g}_n) + \hat{L}_n(\hat{g}_n) - \inf_{g \in \mathcal{C}} L(g) \\ &\leq 2 \sup_{g \in \mathcal{C}} |\hat{L}_n(g) - L(g)| . \end{aligned}$$

Or, pour tout g fixé, la variable aléatoire $n\hat{L}_n(g)$ se comporte comme une loi binomiale $\mathrm{Bin}(n, L(g))$. La convergence de l'erreur d'estimation vers zéro repose donc sur le contrôle des déviations uniformes de la loi binomiale.

Conclusion: Dans les deux cas (M-estimation et classification), on s'intéresse à certains processus indexés par des paramétres vivant éventuellement dans des espaces fonctionnels. Les résultats de consistance d'estimateurs dans chaque contexte reposent essentiellement sur des lois des grands nombres **uniformes**.

2 Lois limites uniformes

2.1 Processus empirique et propriété de Glivenko-Cantelli

On supposera tout au long de cette présentation que les $X_1, ..., X_n$ sont des quantités aléatoires i.i.d. de loi P. On note

$$\forall f \in \mathcal{F}, \quad Pf = \mathbb{E}f(X) = \int f \, dP$$

et aussi

$$\forall f \in \mathcal{F}, \quad P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \int f \, dP_n$$

où $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ désigne la mesure empirique associée à l'échantillon $X_1, ..., X_n$.

Définition 10 *Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé et $\mathcal{F}$ une classe de fonctions. On note $X, X_1, ..., X_n$ des quantités aléatoires i.i.d. de loi P. Le processus stochastique ayant la forme*

$$\begin{aligned} &\mathcal{F} \times \Omega \to \mathbb{R} \\ &(f, \omega) \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i(\omega)) - \mathbb{E}f(X)) = \sqrt{n}(P_n(\omega) - P)(f) \end{aligned}$$

s'appelle le ***processus empirique centré normalisé*** *indexé par $\mathcal{F}$.*

On s'intéresse alors aux conditions sur $\mathcal{F}$ pour avoir un contrôle sur la quantité

$$\sup_{f \in \mathcal{F}} |(P_n - P)(f)| .$$

Une telle classe $\mathcal{F}$ est dite classe de (ou "ayant la propriété de") Glivenko-Cantelli.

Définition 11 *Une classe $\mathcal{F}$ de fonctions mesurables $f : \mathcal{X} \to \mathbb{R}$ est dite P-**Glivenko-Cantelli** si*

$$\sup_{f \in \mathcal{F}} |(P_n - P)(f)| \to 0 \quad P\text{-p.s.}$$

Remarque 12 *Il faut noter que dans le cas où la classe $\mathcal{F}$ est finie et tous ses éléments satisfont $P|f| < \infty$, la propriété est triviale puisqu'elle découle de la loi forte des grands nombres. On suppose désormais que $|\mathcal{F}| = \infty$ (dénombrable).*

Pour démontrer la propriété, on doit d'abord caractériser la taille de la famille $\mathcal{F}$. Une caractérisation possible passe par la notion d'entropie métrique qu'on va introduire progressivement. On verra plus tard une caractérisation combinatoire (*cf.* **Section 4**).

On donne à présent un premier résultat. On consière $\mathcal{F}$ une famille de fonctions mesurables sur $\mathbb{R}^d$. On suppose $P|f| < \infty$, $\forall f \in \mathcal{F}$.

Théorème 13 *On suppose que pour tout $\epsilon > 0$, il existe une classe finie $\mathcal{F}_\epsilon$ telle que pour tout $f \in \mathcal{F}$, on peut trouver des éléments $f_{\epsilon,L}$ et $f_{\epsilon,U}$ dans $\mathcal{F}_\epsilon$ tels que*

$$f_{\epsilon,L} \leq f \leq f_{\epsilon,U} \quad \text{et} \quad P(f_{\epsilon,U} - f_{\epsilon,L}) < \epsilon .$$

Alors

$$\sup_{f \in \mathcal{F}} |(P_n - P)(f)| \to 0 \quad p.s.$$

Ce théorème nous dit que si on peut discrétiser la famille infinie $\mathcal{F}$ en une famille finie (on parle aussi de ϵ-réseau) $\mathcal{F}_\epsilon$ arbitrairement fine (*i.e.* pour tout $\epsilon > 0$), alors on a une loi uniforme des grands nombres.

Preuve du théorème. On note que, pour tout $f \in \mathcal{F}$ et tout $\epsilon > 0$:

$$\begin{aligned}(P_n - P)(f) &\leq P_n f_{\epsilon,U} - Pf \\ &= (P_n - P)f_{\epsilon,U} + P(f_{\epsilon,U} - f) \\ &\leq (P_n - P)f_{\epsilon,U} + \epsilon\end{aligned}$$

et aussi

$$\begin{aligned}(P - P_n)(f) &\leq Pf - P_n f_{\epsilon,L} \\ &= P(f - f_{\epsilon,L}) + (P - P_n)f_{\epsilon,L} + \\ &\leq \epsilon + (P - P_n)f_{\epsilon,L}\ .\end{aligned}$$

Il s'ensuit que, pour tout $f \in \mathcal{F}$ et tout $\epsilon > 0$:

$$|(P_n - P)(f)| \leq \max\left\{|(P - P_n)f_{\epsilon,L}|, |(P_n - P)f_{\epsilon,U}|\right\} + \epsilon$$

ou encore:

$$\sup_{f\in\mathcal{F}} |(P_n - P)(f)| \leq \max_{g\in\mathcal{F}_\epsilon} |(P - P_n)g| + \epsilon$$

Or d'après la loi forte des grands nombres, on a, pour tout g fixé:

$$|(P - P_n)g| \xrightarrow{p.s.} 0, \quad \text{quand} \quad n \to \infty$$

et donc

$$\max_{g\in\mathcal{F}_\epsilon} |(P - P_n)g| \xrightarrow{p.s.} 0, \quad \text{quand} \quad n \to \infty\ .$$

On en déduit, pour tout $\epsilon > 0$:

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{F}} |(P_n - P)(f)| \leq \epsilon$$

ce qui termine la preuve. ■

2.2 Cas des fonctions de répartition

On s'intéresse maintenant au cas particulier du processus empirique défini par les fonctions de répartition empiriques d'une variable aléatoire réelle.

On considère des variables $X, X_1, ..., X_n$ i.i.d. de loi F. On note

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[X_i \leq t]} \quad \text{et} \quad F(t) = \mathbb{P}\{X \leq t\}\,.$$

Le processus $\hat{F}_n$ indexé par $t \in \mathbb{R}$ peut s'interpréter comme un processus empirique indexé par la classe $\mathcal{F} = \{\mathbb{I}_{]-\infty,t]} : t \in \mathbb{R}\}$.

Pour t fixé, on peut appliquer les théorèmes limites classiques et obtenir :

$$\hat{F}_n(t) \xrightarrow{\text{p.s.}} F(t)$$
$$\sqrt{n}(\hat{F}_n(t) - F(t)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, F(t)(1-F(t)))\,.$$

On rappelle que la quantité aléatoire

$$\|\hat{F}_n - F\|_\infty = \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)|$$

est connue sous le nom de statistique de Kolmogorov-Smirnov. On démontre facilement (voir les cours fondamentaux de Probabilités/Statistiques) le théorème de Glivenko-Cantelli :

$$\|\hat{F}_n - F\|_\infty \xrightarrow{\text{p.s.}} 0\,.$$

Un résultat plus raffiné nous donne la vitesse de cette convergence.

Théorème 14 (Kolmogorov (1936), Smirnov (1936)) *Soit $X_1, ..., X_n$ i.i.d. de f.d.r. F. On a, $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\{\sqrt{n}\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| > \epsilon\} = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\epsilon^2}$$
$$\lim_{n\to\infty} \mathbb{P}\{\sqrt{n}\sup_{t\in\mathbb{R}} (\hat{F}_n(t) - F(t)) > \epsilon\} = e^{-2\epsilon^2}$$
$$\lim_{n\to\infty} \mathbb{P}\{\sqrt{n}\sup_{t\in\mathbb{R}} (F(t) - \hat{F}_n(t)) > \epsilon\} = e^{-2\epsilon^2}$$

Ce résultat remarquable par le fait que la vitesse est universelle (*i.e.* indépendante de la loi F) a donné lieu au test non-paramétrique du même nom.

On a en fait une borne supérieure non-asymptotique, appelée inégalité de Dvoretsky, Kiefer et Wolfowitz,

$$\mathbb{P}\{\sqrt{n}\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| > \epsilon\} \leq 2e^{-2\epsilon^2}$$

pour tout $n\epsilon^2$ assez grand. Il est également remarquable que le facteur multiplicatif 2 devant l'exponentiel n'ait été obtenu que relativement récemment par Massart (1990) au prix de complications techniques significatives.

On s'oriente à présent vers la formulation d'un théorème central limite uniforme (on parle aussi de TCL fonctionnel). Pour cela, on se place dans l'espace de Skorohod, noté $D[-\infty, +\infty]$, des fonctions *càdlàg* (continues à droite, limite à gauche) muni de la norme du supremum. De plus, on fera appel à la notion de *tension* d'un élément aléatoire.

Définition 15 *On dira qu'un élément aléatoire mesurable X à valeurs dans un espace métrique est* **tendu** *si, pour tout $\epsilon > 0$, il existe un compact K tel que $\mathbb{P}\{X \notin K\} < \epsilon$.*

Théorème 16 (Donsker, 1952) *Soit $X_1, ..., X_n$ i.i.d. de loi F, alors la suite de processus empiriques $\sqrt{n}(\hat{F}_n - F)$ converge en loi dans $D[-\infty, +\infty]$ vers un élément aléatoire tendu G_F dont les lois marginales sont gaussiennes, de moyenne nulle et de fonction de covariance $\mathbb{E}(G_F(t)G_F(s)) = F(t\wedge s) - F(t)F(s)$, pour tout $s, t \in \mathbb{R}$.*

Remarque 17 *L'élément aléatoire G_F est connu sous le nom de pont brownien. En effet, les trajectoires de ce processus sont "attachées" à 0 en $+\infty$ et $-\infty$. De plus, les trajectoires sont continues mais nulle part différentiables. (Exercice: pour se faire une idée du pont brownien, simuler le processus empirique associé à la loi uniforme pour n=50, 100, 1000, 10000).*

2.3 Classes de Donsker

Le théorème de Donsker vu précédemment pour les fonctions de répartition relève en fait d'une propriété générale qu'on peut associer aux processus empiriques.

On note

$$G_n(f) = \sqrt{n}(P_n(f) - P(f))\,.$$

Par le TCL multivarié, on a, pour tout ensemble fini de fonctions mesurables $f_1, ..., f_k$ telles que $P(f_i^2) < +\infty$,

$$(G_n(f_1), ..., G_n(f_k)) \xrightarrow{\mathcal{D}} (G_P(f_1), ..., G_P(f_k))$$

où $(G_P(f_1), ..., G_P(f_k))$ est un vecteur gaussien sur $\mathbb{R}^k$ d'espérance nulle et de fonction de covariance

$$\forall f, g \in \mathcal{F}, \quad \mathbb{E}(G_P(f)G_P(g)) = P(fg) - P(f)P(g)\,.$$

La version uniforme du fait précédent est la propriété de Donsker qui s'opère en général dans l'espace $\ell^\infty(\mathcal{F})$ des fonctionnelles bornées sur $\mathcal{F}$ muni de la norme infinie.

Définition 18 *Une classe $\mathcal{F}$ de fonctions mesurables est dite **classe P-Donsker** si la suite de processus $\{G_n(f) : f \in \mathcal{F}\}$ converge en loi vers un processus limite G_P tendu dans l'espace $\ell^\infty(\mathcal{F})$. Le processus G_P est un processus gaussien de moyenne nulle et de fonction de covariance la fonction ci-dessus et G_P s'appelle un P-pont brownien.*

Remarque 19 *Il faut noter que si l'ensemble paramétrique est totalement borné (*i.e. *pour tout $\epsilon > 0$, on peut le recouvrir avec un nombre fini de boules de rayon ϵ) alors il existe une semi-métrique rendant presque toutes les trajectoires du processus limite uniformémement continues. Pour plus de détails sur ce point, voir, par exemple, p.262-263 du livre de Van der Vaart "Asymptotic Statistics".*

2.4 Exemples de calculs d'entropies

2.4.1 Définitions

Définition 20 (Entropie avec crochet) *Quelques définitions :*

- *un **crochet** $[l, u]$ est l'ensemble des éléments $f \in \mathcal{F}$ tels que $l \leq f \leq u$.*
- *un ϵ-**crochet** pour la norme $\|\cdot\|$ vérifie en plus $\|u - l\| < \epsilon$.*
- *le **nombre de crochets à l'échelle** ϵ, noté $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$, est le nombre minimum de ϵ-crochets nécessaires pour couvrir $\mathcal{F}$.*
- *l'**entropie avec crochets à l'échelle** ϵ est la quantité $\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$.*

2.4.2 Classe paramétrique

On considère

$$\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathbb{R}\,,\, \theta \in \Theta\}$$

où f_θ mesurable, et l'ensemble Θ est un sous-ensemble borné de $(\mathbb{R}^d, \|\cdot\|)$.

On commence par démontrer un résultat sur le recouvrement des boules de $\mathbb{R}^d$ par des petites boules.

Lemme 21 *Soit $\|\cdot\|$, métrique euclidienne sur $\mathbb{R}^d$. La boule ouverte $\mathcal{B}(R)$ de rayon R dans $\mathbb{R}^d$ peut être couverte par $\left(\frac{4R+\epsilon}{\epsilon}\right)^d$ boules de rayon ϵ.*

Preuve. Soit $\{c_j\}_{j=1,\dots,N} \subset \mathcal{B}(R)$ le plus grand ensemble tel que deux points c_{j_1}, c_{j_2} soients séparés d'une distance d'au moins ϵ. Alors les boules de rayon ϵ, de centre c_j couvrent $\mathcal{B}(R)$. On note $B_j = \mathcal{B}(c_j, \frac{\epsilon}{4})$, ces boules sont disjointes et

$$\bigcup_{j=1}^{N} B_j \subset \mathcal{B}\left(R + \frac{\epsilon}{4}\right)$$

or, le volume d'une boule de rayon ρ dans $\mathbb{R}^d$ est donné par $C_d\rho^d$ où $C_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$. D'où :

$$NC_d\left(\frac{\epsilon}{4}\right)^d \leq C_d\left(R + \frac{\epsilon}{4}\right)^d$$

et donc

$$N \leq \left(\frac{R + \frac{\epsilon}{4}}{\frac{\epsilon}{4}}\right)^d .$$

■

On requiert une certaine régularité de la famille paramétrique pour pouvoir transporter le dénombrement des boules dans $\mathbb{R}^d$ à celui des crochets dans l'espace fonctionnel $\mathcal{F}$.

Théorème 22 *Soit $\mathcal{F}$ une classe paramétrique indexée par Θ un sous-ensemble borné de $\mathbb{R}^d$. On suppose qu'il existe une fonction mesurable m telle que*

$$\textbf{(HL)} \qquad \forall \theta_1, \theta_2 \in \Theta,\, \forall x \in \mathcal{X}, \quad |f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x)\|\theta_1 - \theta_2\|\,.$$

Si on a $\|m\|_{P,r} = P|m(X)|^r < \infty$, alors il existe une constante $K = K(\Theta, d)$ telle que

$$\forall \epsilon \in]0, \mathrm{diam}(\Theta)[, \quad N_{[]}(\epsilon\|m\|_{P,r}, \mathcal{F}, L_r(P)) \leq K\left(\frac{\mathrm{diam}(\Theta)}{\epsilon}\right)^d .$$

Preuve du théorème. On considère des crochets de la forme $[f_\theta - \epsilon m, f_\theta + \epsilon m]$. La taille des crochets en norme $L_r(P)$ est $2\epsilon\|m\|_{P,r}$. Donc, si θ varie sur une grille de pas ϵ sur Θ, les crochets couvrent $\mathcal{F}$ car d'après la condition de Lipschitz **(HL)**, on a :

$$f_{\theta_1} - \epsilon m \leq f_{\theta_2} \leq f_{\theta_1} + \epsilon m$$

dès que $\|\theta_1 - \theta_2\| < \epsilon$. Donc, on a besoin d'autant de crochets pour couvrir $\mathcal{F}$ que de boules de rayon $\frac{\epsilon}{2}$ pour couvrir Θ. On applique le **Lemme 21** pour obtenir le résultat. ■

2.4.3 Classe de Sobolev

Définition 23 *On dit qu'une fonction $g : I \subset \mathbb{R} \longrightarrow \mathbb{R}$ est **absolument continue** si c'est la primitive d'une fonction intégrable h. En d'autres termes, la fonction g est presque partout différentiable de dérivée h.*

Pour $k \in \mathbb{N}$ fixé, on considère la classe de fonctions suivante :

$$\mathcal{F}_k = \left\{ f : [0,1] \to \mathbb{R},\ \|f\|_\infty \leq 1,\ f^{(k-1)} \text{ absolument continue },\ \int (f^{(k)})^2(x)\, dx \leq 1 \right\} .$$

Théorème 24 *Pour tout $k \in \mathbb{N}$, on a la borne suivante sur l'entropie à crochets de la classe $\mathcal{F}_k$:*

$$\exists K \ \ t.q.\ \forall \epsilon > 0, \quad \log N_{[]}(\epsilon, \mathcal{F}_k, \| \cdot \|_\infty) \leq K \left(\frac{1}{\epsilon}\right)^{1/k} .$$

2.4.4 Classe de fonctions monotones

Soit $\mathcal{F}$ la classe des fonctions monotones à valeurs dans $[0,1]$

$$\mathcal{F} = \{ f : \mathbb{R} \to [0,1], f \text{ monotone } \} .$$

Théorème 25 *Il existe une constante $K = K(r)$ telle que, pour toute mesure de probabilité Q et tout entier $r \geq 1$, on a :*

$$\forall \epsilon > 0, \quad \log N_{[]}(\epsilon, \mathcal{F}, L_r(Q)) \leq K \left(\frac{1}{\epsilon}\right) .$$

2.5 Conditions suffisantes pour être Glivenko-Cantelli ou Donsker

On a déjà vu une caractérisation des classes de Glivenko-Cantelli (**Théorème 13**) mais on va la reformuler à nouveau en utilisant les concepts introduits précédemment.

Théorème 26 *Toute classe $\mathcal{F}$ de fonctions mesurables telles que*

$$\forall \epsilon > 0, \quad N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < +\infty$$

est P-Glivenko-Cantelli.

Il existe bien d'autres conditions suffisantes pour qu'une classe soit Glivenko-Cantelli. Elles reposent sur d'autres notions de complexité ou sur d'autres variantes de l'entropie.

On passe maintenant à la propriété de Donsker.

Définition 27 *On définit l'**intégrale entropique à crochets à l'échelle** δ comme étant la quantité suivante*

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))}\, d\epsilon .$$

Remarque 28 *La convergence de l'intégrale entropique dépend du comportement de l'entropie au voisinage de 0.*

Théorème 29 *Toute classe $\mathcal{F}$ de fonctions mesurables telles que*

$$\forall \epsilon > 0, \quad J_{[]}(\epsilon, \mathcal{F}, L_2(P)) < +\infty$$

est P-Donsker.

Pour démontrer ce résultat, on doit vérifier l'équicontinuité asymptotique du processus à savoir : pour tout $\epsilon, \eta > 0$, il existe une partition $\mathcal{F}_1, ..., \mathcal{F}_k$ de $\mathcal{F}$ telle que :

$$\limsup_{n\to\infty} \mathbb{P}\{\sup_i \sup_{f,g\in\mathcal{F}_i} \mid G_n(f-g) \mid > \epsilon\} < \eta\,.$$

En guise d'application, on peut se rapporter aux exemples de la section précédente.

En conclusion : pour aller au-delà des résultats asymptotiques et obtenir des vitesses de convergence, on doit élaborer des inégalités dites maximales (bornes sur les moments ou les queues de distribution d'un supremum de variables aléatoires). Ces inégalités reposent sur deux arguments :

- inégalités exponentielles (*cf.* **Section 3**),
- discrétisation de la classe $\mathcal{F}$ (*cf.* **Section 4**).

3 Inégalités exponentielles

3.1 Inégalités de déviation

On rappelle le principe de la méthode de Chernoff qui est la clé de plusieurs inégalités exponentielles classiques.

Soit Z une variable aléatoire telle que $\forall s > 0, \mathbb{E}e^{sZ} < +\infty$. On a, pour tout $t, s > 0$,

$$\begin{aligned}\mathbb{P}(Z \geq t) &= \mathbb{P}(e^{sZ} \geq e^{st}) \\ &\leq e^{-st}\mathbb{E}e^{sZ} \\ &\leq e^{-st+\log \mathbb{E}e^{sZ}}\end{aligned}$$

Cette inégalité étant satisfaite pour tout $s > 0$, elle est satisfaite en particulier pour la valeur optimisant la borne. On a donc

$$\mathbb{P}(Z \geq t) \leq e^{\phi(t)}$$

où $\phi(t) = \inf_{s>0} \left\{-st + \log \mathbb{E}e^{sZ}\right\}$.

Le calcul de bornes exponentielles explicites dépend donc du contrôle de la transformée de Laplace de la variable Z. Une situation agréable correspond à celui des variables aléatoires dites sous-gaussiennes. En voilà un exemple dans le lemme suivant :

Lemme 30 *Soit Z une variable aléatoire centrée et bornée,* i.e. $\mathbb{E}Z = 0$, $Z \in [a, b]$ *p.s.. On a alors*

$$\mathbb{E}e^{sZ} \leq e^{s^2(b-a)^2/8}\,.$$

Preuve. On utilise la convexité de la fonction exponentielle. Ainsi, pour tout $z \in [a, b]$, on a :

$$e^{sz} \leq \frac{z-a}{b-a}e^{sb} + \frac{b-z}{b-a}e^{sa}\,.$$

En intégrant et en utilisant le fait que la variable Z est centrée, on obtient :

$$\begin{aligned}\mathbb{E}e^{sZ} &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= e^{sa}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{s(b-a)}\right) \\ &= e^{-pu}(1-p+pe^u) = e^{g(u)}\end{aligned}$$

où

$$\begin{aligned}p &= -\frac{a}{b-a} \\ u &= s(b-a)\end{aligned}$$

et $g(u) = -pu + \log(1-p+pe^u)$. On note que $g(0) = g'(0) = 0$ donc un développement de Taylor à l'ordre 2 autour de 0 donne

$$g(u) = \frac{u^2}{2}g''(\theta), \qquad \text{pour un certain } \theta\,.$$

On note enfin que

$$g''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leq \frac{1}{4}\,,$$

car $4xy \leq (x+y)^2$. D'où le résultat. ■

La méthode de Chernoff s'illustre particulièrement dans le cas où Z est une moyenne de variables aléatoires indépendantes et bornées.

Une application classique et immédiate de ce qui précède est l'inégalité de Hoeffding.

Théorème 31 *Soient $X_1, ..., X_n$ des variables aléatoires indépendantes bornées telles que $\forall i, X_i \in [a_i, b_i]$ p.s. . Soit $Z = \sum_{i=1}^n X_i$. Alors, on a :*

$$\mathbb{P}(Z - \mathbb{E}(Z) > \epsilon) \leq \exp\left\{-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}\,,$$

et aussi

$$\mathbb{P}(\mathbb{E}(Z) - Z > \epsilon) \leq \exp\left\{-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}\,.$$

Remarque 32 *Dans le cas de variables de Bernoulli de paramètre p, l'inégalité de Hoeffding donne :*

$$\mathbb{P}\left(\frac{S_n}{n} - p > \epsilon\right) \leq e^{-2n\epsilon^2}\,.$$

*Il faut noter que cette inégalité donne une borne supérieure du pire-des-cas puisqu'elle est obtenue pour $p = 1/2$. La méthode de Chernoff fournit en réalité un résultat plus précis pour un $p \neq 1/2$ fixé, à condition de ne pas utiliser le **Lemme 30** et de faire un calcul exact de la transformée de Laplace.*

L'inégalité précédente permet de vérifier la propriété de Glivenko-Cantelli dans le cas d'une classe finie de fonctions. Par exemple, dans le cas de la classification (**Section 1.2**), on a :

Corollaire 33 *Soit une famille $\mathcal{C}$ de classifieurs de cardinal fini N. On a alors, pour tout $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{g \in \mathcal{C}} |\hat{L}_n(g) - L(g)| > \epsilon\right) \leq 2N \exp\{-2n\epsilon^2\}\,.$$

Par ailleurs, il est utile aussi d'établir un contrôle sur la déviation uniforme moyenne. Pour y parvenir, on peut suivre deux approches:

- appliquer l'identité suivante :

$$\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}(X > t)\, dt$$

  vraie pour toute variable aléatoire positive X satisfaisant $\lim_{x \to +\infty}[x(1 - F(x))] = 0$.

- exploiter le caractère sous-gaussien des variables individuelles.

On peut montrer (exercice) que la borne obtenue par la première approche est $O\left(\frac{N}{\sqrt{n}}\right)$.

On se concentre à présent sur la deuxième approche qui donne une meilleure borne. Pour cela, on établit une inégalité maximale pour les variables aléatoires sous-gaussiennes.

Lemme 34 *Soit $\sigma > 0$, $N \geq 2$ et $Y_1, ..., Y_N$ des variables aléatoires réelles telles que*

$$\forall s > 0, \forall i, \quad \mathbb{E}e^{sY_i} \leq e^{s^2\sigma^2/2}\,.$$

Alors, on a :

$$\mathbb{E}\left\{\max_{1\leq i\leq N} Y_i\right\} \leq \sigma\sqrt{2\ln N}\,.$$

Si, de plus, on a : $\mathbb{E}e^{-sY_i} \leq e^{s^2\sigma^2/2}$, alors :

$$\mathbb{E}\left\{\max_{1\leq i\leq N} |Y_i|\right\} \leq \sigma\sqrt{2\ln(2N)}\,.$$

Preuve. On applique d'abord l'inégalité de Jensen :

$$\begin{aligned}\forall s > 0, \quad \exp\left\{s\mathbb{E}\left\{\max_{1\leq i\leq N} Y_i\right\}\right\} &\leq \mathbb{E}\left\{\exp\left\{s\max_{1\leq i\leq N} Y_i\right\}\right\}\\ &= \mathbb{E}\left\{\max_{1\leq i\leq N}\exp\{sY_i\}\right\}\\ &\leq \sum_{i=1}^{N}\mathbb{E}\{\exp\{sY_i\}\}\\ &\leq Ne^{s^2\sigma^2/2}\,.\end{aligned}$$

D'où :

$$\forall s > 0, \quad \mathbb{E}\left\{\max_{1\leq i\leq N} Y_i\right\} \leq \frac{\ln N}{s} + \frac{s\sigma^2}{2}\,.$$

On pose alors : $s = \sqrt{\frac{2\ln N}{\sigma^2}}$, ce qui prouve l'inégalité. ■

Corollaire 35 *Soit une famille $\mathcal{C}$ de classifieurs de cardinal fini N. On a alors*

$$\mathbb{E}\sup_{g\in\mathcal{C}}|\hat{L}_n(g) - L(g)| \leq \sqrt{\frac{\ln(2N)}{2n}}\,.$$

Ces premiers résultats uniformes, même s'ils donnent une première idée, sont très restrictifs puisque d'une part, ils ne portent que sur des classes finies, et d'autre part, ils sont très conservateurs car ils s'appuient sur l'inégalité de Hoeffding qui fournit une borne dans le cas le moins favorable qui soit.

On verra plus loin (**Section 4**) que pour pallier au premier défaut, il faut développer des notions fines de complexité afin de mettre en oeuvre des techniques de discrétisation du supremum pour des classes fonctionnelles générales.

En ce qui concerne le deuxième défaut, on peut formuler quelques inégalités de déviation plus fines que l'inégalité de Hoeffding mais tout aussi classiques, telles que les inégalités de Bennett et Bernstein. Dans ces résultats, on voit apparaître la variance de la somme des variables.

Théorème 36 (Bennett) *Soient $X_1, ..., X_n$ des variables aléatoires indépendantes telles que $\mathbb{E}X_i = 0$ et $|X_i| \leq c$ p.s.. On pose*

$$Z = \sum_{i=1}^{n} X_i$$
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}(X_i)\,.$$

Alors,

$$\forall t > 0, \quad \mathbb{P}(Z > t) \leq \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right)\right)$$

où $h(u) = (1+u)\ln(1+u) - u$ pour $u \geq 0$.

Preuve. La preuve s'appuie également sur la méthode de Chernoff, simplement elle requiert une majoration plus fine de la transformé de Laplace. On a

$$\forall s > 0, \quad \mathbb{E}\exp(sX_i) = 1 + \sum_{r=2}^{\infty} \frac{s^r \mathbb{E}(X_i^r)}{r!} = 1 + s^2 \mathbb{V}(X_i) F_i$$

où on a posé

$$F_i = \sum_{r=2}^{\infty} \frac{s^{r-2}\mathbb{E}(X_i^r)}{r!\mathbb{V}(X_i)}\,.$$

Or, on a : $\mathbb{E}(X_i^r) \leq c^{r-2}\mathbb{V}(X_i)$ d'où :

$$F_i \leq \frac{e^{sc} - 1 - sc}{(sc)^2}\,.$$

Il reste alors à optimiser en s pour obtenir la majoration désirée. ■

En utilisant une simple minoration de la fonction h ($h(u) \geq \frac{u^2}{2+2u/3}$), on obtient le résultat suivant.

Corollaire 37 (Bernstein) *Sous les mêmes hypothèses que précédemment, on a*

$$\forall t > 0, \quad \mathbb{P}(Z > t) \leq \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}ct}\right)\,.$$

On donne en complément une preuve alternative de l'inégalité de Bennett qui se révèle très instructive.

Preuve alternative de l'inégalité de Bennett :

On introduit la log-Laplace associée à la variable aléatoire Z :

$$M(s) = \ln \mathbb{E}\exp((sZ))\,.$$

On va démontrer l'inégalité suivante :

$$M(s) \leq \frac{\sigma^2}{c^2}(e^{sc} - 1 - sc)\,.$$

Par le théorème de convergence dominée, on montre qu'on peut intervertir dérivation et espérance. Ainsi, on obtient :

$$M'(s) = \sum_{i=1}^{n} \mathbb{E}\left(\frac{X_i \exp(sX_i)}{\mathbb{E}\exp(sX_i)}\right)$$
$$M''(s) = \sum_{i=1}^{n} \left(\mathbb{E}\left(\frac{X_i^2 \exp(sX_i)}{\mathbb{E}\exp(sX_i)}\right) - \left(\mathbb{E}\left(\frac{X_i \exp(sX_i)}{\mathbb{E}\exp(sX_i)}\right)\right)^2\right).$$

On introduit une nouvelle mesure de probabilité Q_s^i indexée par s et i de densité par rapport à P^i la loi de X_i donnée par

$$\frac{dQ_s^i}{dP^i} = \frac{\exp(sX_i)}{\mathbb{E}\exp(sX_i)}.$$

On en déduit

$$M'(s) = \sum_{i=1}^{n} \mathbb{E}_{Q_s^i}(X_i)$$
$$M''(s) = \sum_{i=1}^{n} \mathbb{V}_{Q_s^i}(X_i).$$

On établit maintenant une majoration pour $\mathbb{V}_{Q_s^i}(X_i)$:

$$\begin{aligned}\mathbb{V}_{Q_s^i}(X_i) &\leq \mathbb{E}_{Q_s^i}(X_i^2) \\ &\leq \mathbb{E}(X_i^2 \exp(sX_i)) \\ &\leq e^{sc}\mathbb{E}(X_i^2)\end{aligned}$$

où on a utilisé successivement l'inégalité $\mathbb{E}(\exp(sX_i)) \geq 1$ (d'après Jensen), puis l'hypothèse $X_i \leq c$.

Ceci implique

$$M''(s) \leq \sigma^2 e^{sc}$$

et par intégrations successives, on démontre l'inégalité voulue. ■

3.2 Inégalités de concentration

Dans beaucoup de situations, on souhaiterait établir des inégalités analogues à celles présentées précédemment pour des variables aléatoires Z qui ne sont pas des sommes mais des fonctions plus compliquées des variables initiales $X_1, ..., X_n$, supposées indépendantes.

Dans ce paragraphe, on considère

$$Z = g(X_1, ..., X_n).$$

Il s'avère que sous des hypothèses de régularité sur g, de tels résultats existent. On présente le premier résultat de ce type dû à McDiarmid (1989).

Définition 38 *On dit que $g : \mathbb{R}^n \to \mathbb{R}$ est une* ***fonction aux différences bornées*** *s'il existe des constantes c_i telles que, pour tout i,*

$$\sup_{x_1,...,x_n,x_i'} |g(x_1,...,x_n) - g(x_1,...,x_{i-1},x_i',x_{i+1},...,x_n)| \leq c_i\,.$$

Théorème 39 *Soient $X_1,...,X_n$ des variables aléatoires indépendantes et $Z = g(X_1,...,X_n)$ où g est une fonction aux différences bornées. On a alors :*

$$\forall t > 0, \quad \mathbb{P}(Z - \mathbb{E}(Z) > t) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right\}$$

et

$$\forall t > 0, \quad \mathbb{P}(\mathbb{E}(Z) - Z > t) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right\}\,.$$

Pour la preuve du théorème, on a besoin d'un lemme préliminaire qui est simplement une reformulation du **Lemme 30**.

Lemme 40 *Soient V et Z deux variables aléatoires telles que $\mathbb{E}(V \mid Z) = 0$ p.s. et soient h une fonction et c une constante positive ou nulle telles que*

$$h(Z) \leq V \leq h(Z) + c\,.$$

Alors, on a, pour tout $s > 0$,

$$\mathbb{E}(e^{sV} \mid Z) \leq e^{s^2c^2/8}\,.$$

Preuve du théorème. Soit

$$\forall i = 1,...,n, \quad H_i(X_1,...,X_i) = \mathbb{E}(g(X_1,...,X_n) \mid X_1,...,X_i)$$
$$H_0 = \mathbb{E}(g(X_1,...,X_n))\,.$$

On pose également :

$$V_i = H_i(X_1,...,X_i) - H_{i-1}(X_1,...,X_{i-1})$$
$$V = g - \mathbb{E}(g)\,.$$

On a alors : $V = \sum_{i=1}^n V_i$.
Soit maintenant

$$W_i = \sup_u H_i(X_1,...,X_{i-1},u) - H_{i-1}(X_1,...,X_{i-1})$$
$$Z_i = \inf_v H_i(X_1,...,X_{i-1},v) - H_{i-1}(X_1,...,X_{i-1})\,.$$

On a alors : $W_i \leq V_i \leq Z_i$ p.s., et aussi, par hypothèse,

$$W_i - Z_i \leq \sup_u \sup_v (H_i(X_1,...,X_{i-1},u) - H_i(X_1,...,X_{i-1},v)) \leq c_i\,.$$

On en déduit d'après le lemme :

$$\forall s > 0, \quad \mathbb{E}(e^{sV_i} \mid X_1,...,X_{i-1}) \leq e^{s^2c_i^2/8}\,.$$

On applique enfin la méthode de Chernoff

$$\begin{aligned}\forall t>0, \quad \mathbb{P}(g-\mathbb{E}(g)>t) &\leq e^{-st}\mathbb{E}(e^{s\sum_{i=1}^{n}V_i}) \\ &= e^{-st}\mathbb{E}(e^{s\sum_{i=1}^{n-1}V_i}\mathbb{E}(e^{sV_n} \mid X_1,...,X_{n-1})) \\ &\leq e^{-st+\frac{s^2}{8}c_n^2}\mathbb{E}(e^{s\sum_{i=1}^{n-1}V_i}) \\ &\leq ... \leq e^{-st+\frac{s^2}{8}(\sum_{i=1}^{n}c_i^2)}\,.\end{aligned}$$

On pose $s=\frac{4t}{\sum c_i^2}$ ce qui donne le résultat. ■

On peut appliquer le résultat précédent dans le modèle de classification. On pose

$$Z=\sup_{f\in\mathcal{C}}|\hat{L}_n(f)-L(f)|\,.$$

Il est facile de vérifier que la fonction des X_i est une fonction aux différences bornées avec $c_i=\frac{1}{n}$. On en déduit alors de façon directe à partir du **Théorème 39** :

Proposition 41 *Pour toute classe $\mathcal{C}$ de classifieurs, on a :*

$$\forall t>0, \quad \mathbb{P}\left(\sup_{f\in\mathcal{C}}|\hat{L}_n(f)-L(f)|-\mathbb{E}\left(\sup_{f\in\mathcal{C}}|\hat{L}_n(f)-L(f)|\right)>t\right)\leq 2e^{-2nt^2}\,.$$

Dans l'étude des vitesses de convergence, on peut souvent se ramener à l'étude de l'espérance de la déviation uniforme au lieu de la queue de sa distribution.

4 Mesures de complexité

4.1 Complexités combinatoires

L'approche combinatoire dans l'étude des processus empiriques est connue sous le nom de théorie de Vapnik-Chervonenkis. On illustre cette approche en formulant un résultat dans un cadre proche du modèle de classification.

Notations : Soient des variables aléatoires $X_1, ..., X_n$ i.i.d. dans $\mathbb{R}^d$ de loi μ. On note :

$$\forall A \in \mathcal{B}(\mathbb{R}^d), \quad \mu(A) = \mathbb{P}(X_1 \in A)$$
$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_A(X_i)$$

On introduit une classe $\mathcal{A}$ d'ensembles mesurables de $\mathbb{R}^d$ dont on décrit la complexité combinatoire à travers sa capacité à discriminer sur un ensemble de points.

Définition 42 *Soit $x_1^n = \{x_1, ..., x_n\}$, un jeu de n points fixés de $\mathbb{R}^d$.*

- *On définit la* **trace** *de $\mathcal{A}$ sur x_1^n par :*
$$\mathrm{Tr}(\mathcal{A}, x_1^n) = \{A \cap x_1^n \ : \ A \in \mathcal{A}\}\,.$$
- *On appelle* **coefficient d'éclatement** *de $\mathcal{A}$ la fonction suivante :*
$$\mathcal{S}_{\mathcal{A}}(n) = \max_{x_1^n} |\,\mathrm{Tr}(\mathcal{A}, x_1^n)|$$
où la notation $|T|$ désigne le cardinal de l'ensemble T.
- *On dit que $\mathcal{A}$* **pulvérise** *ou* **éclate** *l'ensemble x_1^n de points si $|\,\mathrm{Tr}(\mathcal{A}, x_1^n)| = 2^n$.*

Exercice: déterminer le coefficient d'éclatement pour les demi-droites et les intervalles de $\mathbb{R}$.

Théorème 43

$$\mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{2 \ln(2\mathcal{S}_{\mathcal{A}}(2n))}{n}}\,.$$

Preuve. La preuve s'effectue en quatre mouvements :

1. **Première symétrisation :** On introduit un échantillon dit "fantôme" $X_1', ..., X_n'$ indépendant de et de même loi que $X_1, ..., X_n$ et on pose
$$\mu_n'(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_A(X_i')\,.$$
On a alors :
$$\begin{aligned}\mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| &= \mathbb{E} \sup_{A \in \mathcal{A}} |\mathbb{E}\,(\mu_n(A) - \mu_n'(A) \mid X_1, ..., X_n)| \\ &\leq \mathbb{E} \sup_{A \in \mathcal{A}} \mathbb{E}\,(|\mu_n(A) - \mu_n'(A)| \mid X_1, ..., X_n) \\ &\leq \mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu_n'(A)|\end{aligned}$$
où on a utilisé successivement l'inégalité de Jensen (convexité de la valeur absolue), ainsi que le fait que $\sup \mathbb{E}(\,\cdot\,) \leq \mathbb{E} \sup(\,\cdot\,)$

2. **Deuxième symétrisation :** On introduit à présent des variables aléatoires de signe $\sigma_1, ..., \sigma_n$ i.i.d. telles que $\mathbb{P}(\sigma_i = \pm 1) = \frac{1}{2}$, et supposées indépendantes de $X_1, ..., X_n, X'_1, ..., X'_n$. On note alors que les variables

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbb{I}_A(X_i) - \mathbb{I}_A(X'_i)) \quad \text{et} \quad \frac{1}{n}\sum_{i=1}^{n}\sigma_i(\mathbb{I}_A(X_i) - \mathbb{I}_A(X'_i))$$

ont la même loi (on vérifie qu'elles ont la même fonction caractéristique).

Par conséquent :

$$\begin{aligned}\mathbb{E}\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu'_n(A)| &= \frac{1}{n}\mathbb{E}\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\sigma_i(\mathbb{I}_A(X_i) - \mathbb{I}_A(X'_i))\right| \\ &= \frac{1}{n}\mathbb{E}\mathbb{E}\left(\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\sigma_i(\mathbb{I}_A(X_i) - \mathbb{I}_A(X'_i))\right| \mid X_1, ..., X_n, X'_1, ..., X'_n\right).\end{aligned}$$

3. **Dénombrement :** On est ramené finalement à l'étude de la quantité suivante :

$$\mathbb{E}\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\sigma_i(\mathbb{I}_A(x_i) - \mathbb{I}_A(x'_i))\right|$$

où les $x_1, ..., x_n, x'_1, ..., x'_n$ sont des points fixés.

Soit $\hat{\mathcal{A}} \subset \mathcal{A}$ une sous-famille finie de $\mathcal{A}$ permettant de reconstituer la trace de $\mathcal{A}$ sur les $2n$ points $x_1, ..., x_n, x'_1, ..., x'_n$. On a donc : $|\hat{\mathcal{A}}| \leq \mathcal{S}_{\mathcal{A}}(2n)$.

On peut donc écrire à présent

$$\mathbb{E}\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\sigma_i(\mathbb{I}_A(x_i) - \mathbb{I}_A(x'_i))\right| = \mathbb{E}\max_{A\in\hat{\mathcal{A}}}\left|\sum_{i=1}^{n}\sigma_i(\mathbb{I}_A(x_i) - \mathbb{I}_A(x'_i))\right|$$

A travers les étapes de symmétrisation, on a pu convertir le supremum sur une classe infinie en un maximum sur une classe finie. Cette dernière est aléatoire (elle dépend des réalisations des variables X_i, X'_i) mais son cardinal est contrôlé par une grandeur déterministe.

4. **Cas fini :** On peut maintenant appliquer le **Lemme 34** sur l'espérance du maximum d'une collection finie de variables aléatoires sous-gaussiennes. On pose $Y_i(A) = \sigma_i(\mathbb{I}_A(x_i) - \mathbb{I}_A(x'_i))$. Il est facile de voir que $\mathbb{E}(Y_i(A)) = 0$ et que $Y_i(A) \in [-1, 1]$. On en déduit, par le **Lemme 30**, que :

$$\forall s > 0, \quad \mathbb{E}e^{sY_i(A)} \leq e^{s^2/2},$$

et par indépendance des $Y_i(A)$, on a:

$$\forall s > 0, \quad \mathbb{E}e^{s\sum_{i=1}^{n}Y_i(A)} \leq e^{ns^2/2}.$$

Le caractère sous-gaussien de la collection de variables $\left(\sum_{i=1}^n Y_i(A)\right)_{A\in\hat{\mathcal{A}}}$ étant acquis, on peut lui appliquer le **Lemme 34**. On obtient donc :

$$\mathbb{E}\max_{A\in\hat{\mathcal{A}}}|\sum_{i=1}^n \sigma_i(\mathbb{I}_A(x_i)-\mathbb{I}_A(x_i'))| \leq \sqrt{2n\ln(2|\hat{\mathcal{A}}|)}$$
$$\leq \sqrt{2n\ln(2\mathcal{S}_{\mathcal{A}}(2n))}\,.$$

Cette dernière majoration est uniforme en ce sens qu'elle ne dépend plus des X_i, X_i'. On intègre et le résultat est obtenu.

■

Remarque 44 *L'inégalité originelle établie par Vapnik et Chervonenkis (1971) portait sur la queue de la distribution de la déviation uniforme :*

$$\mathbb{P}(\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|>t) \leq 4\mathcal{S}_{\mathcal{A}}(2n)\,e^{-nt^2/8}\,.$$

Remarque 45 *Il est possible d'obtenir une borne plus raffinée sous réserve d'un contrôle de la variance du processus. En effet, si on pose* $\Sigma = \sup_{A\in\mathcal{A}}\sqrt{\mu(A)(1-\mu(A))}$*, on peut démontrer :*

$$\mathbb{E}\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)| \leq \frac{16\ln(2\mathcal{S}_{\mathcal{A}}(2n))}{n} + \sqrt{\frac{32\Sigma^2\ln(2\mathcal{S}_{\mathcal{A}}(2n))}{n}}\,.$$

Si on met bout à bout le théorème précédent avec l'inégalité de concentration du **Théorème 39**, on en déduit le corollaire suivant.

Corollaire 46 *Soit une classe* $\mathcal{A}$ *d'ensembles mesurables de* $\mathbb{R}^d$ *telle que*

$$\frac{\ln(\mathcal{S}_{\mathcal{A}}(n))}{n} \to 0$$

quand $n\to+\infty$*. Alors, la classe de fonctions indicatrices*

$$\mathcal{F} = \{\mathbb{I}_A\,:\,A\in\mathcal{A}\}$$

est une classe de Glivenko-Cantelli.

Au vu de ce corollaire, une question s'impose : comment le coefficient d'éclatement dépend-il de n ?

En fait, dans beaucoup d'exemples, la dépendance est polynomiale en n.

Exemples

- $\mathcal{A} = \{$ demi-droites de $\mathbb{R}\}$: $\mathcal{S}_{\mathcal{A}}(n) = n+1$.
- $\mathcal{A} = \{$ intervalles de $\mathbb{R}\}$: $\mathcal{S}_{\mathcal{A}}(n) = \frac{n(n+1)}{2}$.
- $\mathcal{A} = \{$ demi-espaces de $\mathbb{R}^d\}$: $\mathcal{S}_{\mathcal{A}}(n) = O(n^{d+1})$.
- $\mathcal{A} = \{$ hyper-rectangles de $\mathbb{R}^d\}$: $\mathcal{S}_{\mathcal{A}}(n) = O(n^{2d})$.

Dans les exemples précédents, il existe une valeur critique de n à partir de laquelle, on ne peut plus obtenir toutes les parties d'un ensemble à n points en les interceptant par des éléments de $\mathcal{A}$. Cette valeur critique s'appelle **dimension de Vapnik-Chervonenkis**. Il faut noter qu'il existe des classes d'ensembles qui sont trop riches de ce point de vue et pour lesquelles, la dimension de Vapnik-Chervonenkis est infinie (par exemple, la classe des polygones convexes de $\mathbb{R}^2$).

4.2 Entropies

La méthode d'approximation est une approche analytique plutôt que combinatoire. Elle consiste à discrétiser l'espace fonctionnel sous-jacent en considérant une approximation de cardinal fini. La mesure de complexité qui intervient dans ce contexte est l'entropie métrique. On donne un résultat établi grâce à la méthode de chaînage, due à Dudley.

Définition 47 (Nombre de couverture) *Soit B un ensemble munie d'une métrique ρ.*

- *Une **couverture** à l'échelle $r > 0$ est un ensemble B_r tel que, pour tout $b \in B$, il existe $c \in B_r$ tel que $\rho(b, c) \leq r$.*
- *Le **nombre de couverture** $N(r, B)$ est le cardinal de la plus petite couverture à l'échelle r.*

Définition 48 (Empreinte binaire) *Soit $\mathcal{A}$ une classe d'ensembles mesurables de $\mathbb{R}^d$ et $x_1^n = \{x_1, ..., x_n\}$ un jeu de n points fixés. On appelle **empreinte binaire** de x_1^n l'ensemble des vecteurs binaires suivant :*

$$\mathcal{A}(x_1^n) = \{b = (b_1, ..., b_n) \in \{0, 1\}^n \, : \, b_i = \mathbb{I}_{[x_i \in A]}, \, A \in \mathcal{A}\} \, .$$

*On munit cet ensemble de la **distance de Hamming** définie par :*

$$\rho(b, c) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{[b_i \neq c_i]}} \, .$$

Théorème 49

$$\mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \frac{24}{\sqrt{n}} \max_{x_1^n} \int_0^1 \sqrt{\ln(2N(\epsilon, \mathcal{A}(x_1^n)))} \, d\epsilon \, .$$

Preuve. La preuve s'effectue par étapes. L'étape de symétrisation préalable est également indispensable ici.

1. **Symétrisation :** on se réfère aux deux premières étapes de la méthode combinatoire.

$$\begin{aligned} \mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| &\leq \frac{1}{n} \mathbb{E}\mathbb{E} \left(\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i (\mathbb{I}_A(X_i) - \mathbb{I}_A(X_i')) \right| \mid X_1, ..., X_n, X_1', ..., X_n' \right) \\ &\leq \frac{2}{n} \mathbb{E}\mathbb{E} \left(\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i (\mathbb{I}_A(X_i) \right| \mid X_1, ..., X_n \right) \, . \end{aligned}$$

2. **Discrétisation :** on utilise alors l'empreinte binaire de la classe $\mathcal{A}$ pour $x_1^n = (x_1, ..., x_n)$ fixé.

$$\mathbb{E} \left(\sup_{A \in \mathcal{A}} | \sum_{i=1}^{n} \sigma_i (\mathbb{I}_A(x_i)| \right) = \mathbb{E} \left(\sup_{b \in \mathcal{A}(x_1^n)} | \sum_{i=1}^{n} \sigma_i b_i| \right)$$

On construit une suite $(B_k)_{1 \leq k \leq M}$ de couvertures minimales d'échelle $r = 2^{-k}$ de $\mathcal{A}(x_1^n)$ pour la distance de Hamming où on prend :

- à l'échelle la plus grossière ($r = 1$) : $B_0 = \{b_0 = (0, ..., 0)\}$.
- à l'échelle la plus fine : $B_M = \mathcal{A}(x_1^n)$ avec $M = \lfloor \log_2 \sqrt{n} + 1 \rfloor$.

En effet, si $b, c \in \{0,1\}^n$ diffèrent seulement pour une coordonnée, alors $\rho(b,c) = \frac{1}{\sqrt{n}}$. Alors, on peut choisir pour B_M l'ensemble des points de $\mathcal{A}(x_1^n)$ comme centres et $r = \frac{1}{2\sqrt{n}}$ comme échelle.

3. **Chaînage :** on note $b^* = (b_1^*, ..., b_n^*)$ le vecteur où le maximum est atteint et $b^{(k)} = (b_1^{(k)}, .., b_n^{(k)})$ le plus proche voisin de b^* dans B_k. On a donc :

$$\rho(b^{(k)}, b^*) \leq 2^{-k}$$

et

$$\begin{aligned}\rho(b^{(k)}, b^{(k-1)}) &\leq \rho(b^{(k)}, b^*) + \rho(b^{(k-1)}, b^*) \\ &\leq 2^{-k} + 2^{-k+1} \leq 3 \cdot 2^{-k}\,.\end{aligned}$$

Le chaînage consiste à décomposer le processus relativement aux projections de plus en plus grossières. On a:

$$b_i^* = b_i^{(0)} + \sum_{k=1}^{M} (b_i^{(k)} - b_i^{(k-1)})$$

donc :

$$\sum_{i=1}^{n} \sigma_i b_i^* = \sum_{k=1}^{M} \sum_{i=1}^{n} \sigma_i (b_i^{(k)} - b_i^{(k-1)})\,.$$

On en déduit :

$$\mathbb{E}\left(\max_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i b_i \right| \right) \leq \sum_{k=1}^{M} \mathbb{E}\left(\max_{b \in B_k, c \in B_{k-1}, \rho(b,c) \leq 3 \cdot 2^{-k}} \left| \sum_{i=1}^{n} \sigma_i (b_i - c_i) \right| \right)\,.$$

4. **Majoration :** On applique maintenant le **Lemme 30** et on obtient : $\forall b \in B_k$, $\forall c \in B_{k-1}$ avec $\rho(b,c) \leq 3 \cdot 2^{-k}$,

$$\forall s > 0, \quad e^{s \sum_{i=1}^{n} \sigma_i (b_i - c_i)} \leq e^{n \frac{s^2}{2} (3 \cdot 2^{-k})^2}\,.$$

Le nombre de couple (b, c) est borné par

$$|B_k| \cdot |B_{k-1}| \leq |B_k|^2 = N^2(2^{-k}, \mathcal{A}(x_1^n))\,.$$

On en déduit par le **Lemme 34** :

$$\mathbb{E}\left(\max_{b \in B_k, c \in B_{k-1}, \rho(b,c) \leq 3 \cdot 2^{-k}} \left| \sum_{i=1}^{n} \sigma_i (b_i - c_i) \right| \right) \leq 3\sqrt{n}\, 2^{-k} \sqrt{2 \ln(2N^2(2^{-k}, \mathcal{A}(x_1^n)))}\,.$$

Donc, en utilisant le fait que la fonction $x \mapsto N(x, \mathcal{A}(x_1^n))$ est décroissante, on obtient finalement

$$\begin{aligned}\mathbb{E}\left(\max_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\sigma_i b_i\right|\right) &\leq 12\sqrt{n}\sum_{k=1}^{+\infty}2^{-(k+1)}\sqrt{2\ln(2N(2^{-k},\mathcal{A}(x_1^n)))}\\ &\leq 12\sqrt{n}\int_0^1\sqrt{\ln(2N(\epsilon,\mathcal{A}(x_1^n)))}\,d\epsilon\,,\end{aligned}$$

qui nous mène directement au résultat.

■

5 Application statistique

Il existe de nombreuses applications des processus empiriques à des problèmes statistiques. On propose ici d'en regarder une en particulier qui est la question de la normalité asymptotique des M- et Z-estimateurs.

On souhaite couvrir deux types de situations à l'aide de théorèmes généraux :

1. **les estimateurs du maximum de (log-)vraisemblance :** étant données $X_1, ..., X_n$ des variables aléatoires i.i.d. de densité p_θ, on considère
$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1} \ln p_\theta(X_i)\,.$$

2. **les estimateurs de localisation :** étant données $X_1, ..., X_n$ des variables aléatoires i.i.d. dont le paramètre de localisation est θ, on propose un estimateur $\hat{\theta}_n$ défini comme racine de l'équation en θ :
$$\sum_{i=1} \psi(X_i - \theta) = 0\,.$$
Dans le cas où $\psi(x) = x$, on retrouve l'estimateur de la moyenne, alors que si $\psi(x) = |x|$, c'est l'estimateur de la médiane.

Remarque 50 *La première situation (1.) peut conduire à une caractérisation comme Z-estimateur de ce qui est en apparence un M-estimateur dès lors que l'application $\theta \mapsto \ln p_\theta(x)$ est différentiable. D'autre part, la situation (2.) peut également se ramener assez fréquemment à une formulation de type M-estimateur dès que la fonction ψ est intégrable. En particulier, l'estimateur de la médiane peut être obtenu comme le maximum de la fonction $\theta \mapsto -|x - \theta|$.*

Or, les théorèmes généraux classiques pour le maximum de vraisemblance établis systématiquement à partir des années '40 exigent trop de régularité pour pouvoir s'étendre au cas de l'estimateur de la médiane.

On va donner un résultat de normalité asymptotique avec des hypothèses minimales.

Argument heuristique :

Soient $X_1, ..., X_n$ des quantités aléatoires i.i.d. de loi P_θ où $\theta \in \Theta \subset \mathbb{R}^d$. On considère le critère empirique à maximiser suivant :
$$P_n m_\theta = \frac{1}{n} \sum_{i=1} m_\theta(X_i)\,.$$

On suppose que le vrai critère $\theta \mapsto Pm_\theta$ est deux fois différentiable et admet un maximum unique en $\theta_0 \in \dot{\Theta}$. Par conséquent, son gradient s'annule en θ_0 et la matrice hessienne V_{θ_0} en θ_0 est définie négative. On rappelle la notation pour le processus empirique centré normalisé : $G_n = \sqrt{n}(P_n - P)$. En admettant qu'on a suffisamment de régularité (on note $\dot{m}_\theta = \nabla_\theta m_\theta$), on peut alors écrire, le développement suivant, pour tout θ fixé :

$$\begin{aligned} P_n(m_\theta - m_{\theta_0}) &= P(m_\theta - m_{\theta_0}) + \frac{1}{\sqrt{n}} G_n(m_\theta - m_{\theta_0}) \\ &= \frac{1}{2}(\theta - \theta_0)^T V_{\theta_0}(\theta - \theta_0) + \frac{1}{\sqrt{n}}(\theta - \theta_0)^T G_n \dot{m}_{\theta_0} + o(\|\theta - \theta_0\|^2) + o_P\left(\frac{\|\theta - \theta_0\|}{\sqrt{n}}\right)\,. \end{aligned}$$

Si on néglige les restes, on trouve que le maximum du terme de droite est réalisé pour θ vérifiant :

$$\sqrt{n}(\theta - \theta_0) = -V_{\theta_0}^{-1} G_n \dot{m}_{\theta_0} .$$

On peut s'attendre donc au comportement limite suivant pour le M-estimateur $\hat{\theta}_n$ maximisant le terme de gauche :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} G_n \dot{m}_{\theta_0} + o_P(1) .$$

On en déduit alors la normalité asymptotique de l'estimateur $\hat{\theta}_n$ comme conséquence du théorème central limite appliqué à la suite de variables $\dot{m}_{\theta_0}(X_i)$. On obtient donc :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}(0, \Sigma)$$

où

$$\Sigma = V_{\theta_0}^{-1} P(\dot{m}_{\theta_0} \dot{m}_{\theta_0}^T) V_{\theta_0}^{-1} .$$

On note que V_{θ_0} étant symétrique, c'est aussi le cas pour son inverse.

Approche rigoureuse :

Il s'agit maintenant de préciser le résultat, mais il y a diverses façons de le faire. On peut, par exemple, faire l'hypothèse que la fonction $\theta \mapsto m_\theta(x)$ est deux fois différentiable en tout point x, mais c'est trop contraignant par rapport à notre objectif. En effet, pour l'estimation de la médiane, on a la fonction $m_\theta(x) = -|x - \theta|$ qui n'est pas dérivable en tout x.

En réalité, on a surtout besoin d'un développement de Taylor à l'ordre 2 pour la fonction $\theta \mapsto P m_\theta(X)$ autour du maximum et aussi d'une notion faible de dérivabilité pour $\theta \mapsto m_\theta(x)$.

Définition 51 *Un M-estimateur consistant est une suite $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ vérifiant :*

- $P_n m_{\hat{\theta}_n} \geq \sup_\theta P_n m_\theta - o_P(n^{-1})$.
- $\hat{\theta}_n = \theta_0 + o_P(1)$

Un M-estimateur $\hat{\theta}_n$ sera dit $\sqrt{n}$-consistant s'il vérifie en plus $\hat{\theta}_n = \theta_0 + O_P\left(n^{-1/2}\right)$.

Théorème 52 *Soit $\theta \in \Theta \subset \mathbb{R}^d$ et $x \mapsto m_\theta(x)$ une fonction mesurable. On suppose :*

- **(H1)** *la fonction $\theta \mapsto P m_\theta(X)$ est de classe C^2 et admet un maximum en θ_0. On note V_{θ_0} la matrice hessienne en θ_0.*
- **(H2)** *il existe une fonction $\dot{m}_{\theta_0}$ telle que :*

$$\mathbb{E}(m_\theta - m_{\theta_0} - (\theta - \theta_0)^T \dot{m}_{\theta_0})^2 = o(\|\theta - \theta_0\|^2) .$$

- **(H3)** *La classe $\mathcal{F}_\delta$ de fonctions définie, pour un certain δ, par*

$$\left\{ \frac{m_\theta - m_{\theta_0} - (\theta - \theta_0)^T \dot{m}_{\theta_0}}{\|\theta - \theta_0\|} \ : \|\theta - \theta_0\| < \delta \right\} \text{ est } P\text{-Donsker.}$$

Alors pour tout M-estimateur $\sqrt{n}$-consistant $\hat{\theta}_n$, on a :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}(0, V_{\theta_0}^{-1} P(\dot{m}_{\theta_0} \dot{m}_{\theta_0}^T) V_{\theta_0}^{-1}) .$$

Lemme 53 *On suppose que les hypothèses* **(H2)** *et* **(H3)** *sont satisfaites. Alors, pour toute suite $\tilde{\theta}_n$ telle que $\tilde{\theta}_n = \theta_0 + o_P(1)$, on a*

$$G_n(m_{\tilde{\theta}_n} - m_{\theta_0}) = (\tilde{\theta}_n - \theta_0)^T G_n \dot{m}_{\theta_0} + o_P(\|\tilde{\theta}_n - \theta_0\|) .$$

Preuve. On introduit la notation : $\Theta_\delta = \{\theta : \|\theta - \theta_0\| < \delta\}$. Soit la fonction

$$\begin{aligned} f : \ell^\infty(\Theta_\delta) \times \Theta_\delta &\to \mathbb{R}^d \\ (z, \theta) &\mapsto z(\theta) \end{aligned}$$

Cette fonction est continue en tout point (z, θ_0) dès que la fonction $\theta \mapsto z(\theta)$ est continue en θ_0. En effet, si on a $(z_n, \theta_n) \to (z, \theta_0)$ dans $\ell^\infty(\Theta_\delta) \times \Theta_\delta$, alors $z_n \to z$ uniformément et donc $z_n(\theta_n) = z(\theta_n) + o(1) \to z(\theta_0)$ si z continue en θ_0.

Soit maintenant le processus stochastique Z_n indexé par Θ_δ défini par

$$Z_n(\theta) = G_n \left(\frac{m_\theta - m_{\theta_0} - (\theta - \theta_0)^T \dot{m}_{\theta_0}}{\|\theta - \theta_0\|} \right) .$$

Par hypothèse, ce processus converge en loi dans $\ell^\infty(\Theta_\delta)$ vers un processus gaussien tendu Z. Or, le processus limite Z possède des trajectoires continues sur Θ_δ relativement à la pseudo-distance définie par $\rho^2(\theta_1, \theta_2) = \mathbb{E}(Z(\theta_1) - Z(\theta_2))^2$. Or, l'hypothèse **(H2)** nous indique précisément que $\rho(\theta, \theta_0) \to 0$ lorsque $\theta \to \theta_0$ (car $Z(\theta_0) = 0$ p.s. et la fonction de covariance caractérisant le processus limite correspond à la covariance des fonctions de $\mathcal{F}_\delta$). Ainsi, la fonction f est continue en presque tout point (Z, θ_0). On a alors, par le théorème de Slutsky, que $(Z_n, \tilde{\theta}_n) \xrightarrow{\mathcal{D}} (Z, \theta_0)$ et par le théorème de continuité, on peut transporter la convergence par l'application f continue en (Z, θ_0). Finalement, on a : $Z_n(\tilde{\theta}_n) \xrightarrow{\mathcal{D}} Z(\theta_0) = 0$. On conclut en utilisant le fait que la convergence en loi vers une constante équivaut à la convergence en probabilité. ■

Preuve du théorème. En utilisant le développement de Taylor de $\theta \mapsto Pm_\theta(X)$ ainsi que le premier lemme, on obtient :

$$P_n(m_{\hat{\theta}_n} - m_{\theta_0}) = \frac{1}{2}(\hat{\theta}_n - \theta_0)^T V_{\theta_0}(\hat{\theta}_n - \theta_0) + \frac{1}{\sqrt{n}}(\hat{\theta}_n - \theta_0)^T G_n \dot{m}_{\theta_0} + o_P\left(\|\hat{\theta}_n - \theta_0\|^2 + \frac{\|\hat{\theta}_n - \theta_0\|}{\sqrt{n}} \right) .$$

On utilise le caractère $\sqrt{n}$-consistant de $\hat{\theta}_n$ pour simplifier le reste :

$$P_n(m_{\hat{\theta}_n} - m_{\theta_0}) = \frac{1}{2}(\hat{\theta}_n - \theta_0)^T V_{\theta_0}(\hat{\theta}_n - \theta_0) + \frac{1}{\sqrt{n}}(\hat{\theta}_n - \theta_0)^T G_n \dot{m}_{\theta_0} + o_P\left(\frac{1}{n} \right) .$$

On applique la même relation à la suite $\tilde{\theta}_n = \theta_0 - \frac{1}{\sqrt{n}} V_{\theta_0}^{-1} G_n \dot{m}_{\theta_0}$ qui présente également le caractère $\sqrt{n}$-consistant et satisfait le lemme. On obtient :

$$P_n(m_{\tilde{\theta}_n} - m_{\theta_0}) = -\frac{1}{2n}(G_n \dot{m}_{\theta_0})^T V_{\theta_0}^{-1}(G_n \dot{m}_{\theta_0}) + o_P\left(\frac{1}{n} \right) .$$

On prend la différence entre les deux relations pour en déduire que :

$$P_n m_{\hat{\theta}_n} - P_n m_{\tilde{\theta}_n} = \frac{1}{2}(\hat{\theta}_n - \tilde{\theta}_n)^T V_{\theta_0}(\hat{\theta}_n - \tilde{\theta}_n) + o_P\left(\frac{1}{n}\right) .$$

Or, le membre de gauche est positif à un terme d'ordre $o_P\left(\frac{1}{n}\right)$ près. Comme V_{θ_0} est définie négative, on a donc

$$\hat{\theta}_n - \theta_0 + \frac{1}{\sqrt{n}} V_{\theta_0}^{-1} G_n \dot{m}_{\theta_0} = o_P\left(\frac{1}{\sqrt{n}}\right)$$

et on a le résultat. ■

Remarque 54 *La $\sqrt{n}$-consistance est satisfaite sous les hypothèses du théorème dès lors que l'intégrale entropique à crochets associée à la classe $\mathcal{F}_\delta$ est proportionnelle au diamètre δ de l'ensemble paramétrique Θ_δ. On note que c'est le cas si l'entropie avec crochets est en $\epsilon^{-\alpha}$.*

Exemples :

1. Cas de la médiane

   $$m_\theta(x) = -|x - \theta|$$

   avec $\theta \in [-1, 1]$. On note f la densité des X_i et on suppose $f(\theta_0) > 0$.

   On peut vérifier que cette fonction satisfait toutes les hypothèses du théorème en posant $\dot{m}_\theta(x) \equiv -\operatorname{sgn}(x - \theta)$. On note que $\theta \mapsto Pm_\theta$ est deux fois dérivable avec $V_{\theta_0} = -2f(\theta_0)$. On en déduit

   $$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{(2f(\theta_0))^2}\right) .$$

2. Estimateur robuste de la moyenne :

   $$m_\theta(x) = (x - \theta)^2 \mathbb{I}_{[|x-\theta| \leq a]} + (2a|x - \theta| - a^2)\mathbb{I}_{[|x-\theta|>a]}, ,$$

   avec $\theta \in \mathbb{R}$ et $a > 0$ fixé. On suppose qu'il existe un unique maximum en θ_0. On note F la f.d.r des X_i.

   On obtient en appliquant le théorème :

   $$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2\right) ,$$

   où

   $$\sigma^2 = \frac{a^2 F(-a) + \int_{-a}^{a} x^2 dF(x) + a^2(1 - F(a))}{(F(a) - F(-a))^2} .$$