MÉMOIRE DU STAGE DE L3 :

GÉNÉRALISATION DE LA STATISTIQUE DE KOLMOGOROV-SMIRNOV

Vernet Élodie
Maître de stage : Nicolas Vayatis

20/06/2010

Adresse électronique : elodie.vernet@ens-cachan.fr

Lieu : CMLA de l'ENS de Cachan

Résumé : Le but de ce stage est d'étudier théoriquement une probabilité dépendant de paramètres qui traduit la vitesse d'une convergence uniforme en probabilité, puis de comparer la théorie aux expériences numériques sur un exemple précis.

Remerciements : Je remercie mon maître de stage Nicolas Vayatis pour l'encadrement de ce stage ainsi que Nicolas Baskiotis pour son aide précieuse.

Table des matières

Introduction

On étudie, ici, la convergence uniforme presque sûre de la mesure empirique vers la mesure. On réalise cette étude afin de généraliser la statistique de Kolmogorov-Smirnov et notamment pour trouver un test d'adéquation pour des dimensions supérieures à deux. On essaye également, par cette étude, de justifier un méthode permettant de trouver un classifieur optimal.

Ce problème n'a été résolu à ce jour que pour certains exemples précis. Kolmogorov et Smirnov se sont intéressés aux fonctions de répartition sur $\mathbb{R}$. Il existe aussi quelques résultats en dimension supérieure (comme celui de Beran et Millar). Enfin la théorie de Vapnik et Chervonenkis traîte le pire des cas. Le problème est donc encore ouvert. On s'est inspiré de la thèse de Nicolas Vayatis ([10]), pour cette étude.

Dans ce rapport, on analysera le comportement asymptotique ou non de la probabilité que la distance entre la mesure empirique et la mesure soit supérieure à un certain réel. Puis, on comparera la théorie aux expériences numériques.

1 Motivations

On introduit μ une mesure de probabilité sur $\mathbb{R}^d$, $\mathcal{C}$ un ensemble de parties mesurables de $\mathbb{R}^d$. On peut estimer la mesure d'une partie $c \in \mathcal{C}$: $\mu(c) = \mathbb{P}\{X_1 \in c\} = \mathbb{E}\left[\mathbb{I}_{\{c\}}(X_1)\right]$, grâce à la mesure empirique de c associée à μ. Plus précisément, on tire des variables aléatoires $X_1, X_2, ..., X_n$ indépendantes, identiquement distribuées de loi μ et on calcule $\hat{\mu}_n(c) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbb{I}_{\{c\}}(X_i)$ qui est la mesure empirique de c associée à μ. Par la loi forte des grands nombres, on sait que $\hat{\mu}_n(c)$ tend presque sûrement vers $\mu(c)$. On s'intéresse alors à la convergence uniforme en probabilité de $\hat{\mu}_n$ vers μ (c'est-à-dire à la version uniforme de la loi des grands nombres), ce qui revient à étudier la probabilité suivante $K(n, \mu, t, \mathcal{C}) = \mathbb{P}\left\{\sup_{c\in\mathcal{C}} |\hat{\mu}_n(c) - \mu(c)| > t\right\}$ pour $n \in \mathbb{N}$, μ une mesure de probabilité, $t \in \mathbb{R}$ et $\mathcal{C}$ un ensemble de parties de $\mathbb{R}^d$.

Il y a deux principales motivations à l'étude de cette probabilité. La première vient du fait qu'elle généralise la statistique de Kolmogorov-Smirnov. La seconde nous vient de la théorie de la classification.

1.1 Statistique de Kolmogorov-Smirnov

La statistique de Kolmogorov-Smirnov repose sur $K\left(n, \mu, \frac{t}{\sqrt{n}}, \mathcal{C}_0\right)$ où $\mathcal{C}_0 = \{]-\infty, a], a \in \mathbb{R}\}$, c'est-à-dire l'ensemble des demi-droites à gauche. On introduit F la fonction de répartition associée à μ, $\hat{F}_n : \begin{cases} \mathbb{R} \to \mathbb{R} \\ x \to \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{]-\infty,x]\}}(X_i) \end{cases}$ la fonction de répartition empirique pour l'échantillon $X_1, X_2, ..., X_n$ (i.i.d. de loi μ). Alors on a $K\left(n, \mu, \frac{t}{\sqrt{n}}, \mathcal{C}_0\right) = \mathbb{P}\left\{\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \frac{t}{\sqrt{n}}\right\}$. Or Kolmogorov [8] nous donne le comportement asymptotique de cette probabilité.

Théorème 1.1.1. ***(Kolmogorov)*** $\forall t \in \mathbb{R}$,

$$\lim_{n\to+\infty} K\left(n, \mu, \frac{t}{\sqrt{n}}, \mathcal{C}_0\right) = 2\sum_{k=1}^{+\infty}(-1)^{k-1}e^{-2k^2t^2}.$$

Ce comportement asymptotique permet de construire un test d'adéquation, c'est-à-dire un test qui permet de déterminer avec une certaine probabilité si un échantillon suit ou non une certaine loi. Ainsi l'étude de $K(n, \mu, t, \mathcal{C})$ dans le cas où $d \geq 2$ pourrait permettre de trouver un test d'adéquation pour des dimensions supérieures à 1. Ce problème reste ouvert à ce jour.

1.2 Classification

Intéressons-nous désormais à la seconde motivation. Et en premier lieu, expliquons ce qu'est la classification [5] (on ne s'intéresse ici qu'à la classification binaire). Le but de la classification est de deviner à quelle classe appartient une observation. Par exemple, on connaît la température, la pression artérielle, le taux de globules blancs d'un patient et on veut déterminer si il est malade ou non.

On note $X \in \mathbb{R}^d$ l'observation (dans l'exemple précédent, c'est un vecteur à 3 composantes contenant la température, la pression et le taux de globules), $Y \in \{-1, 1\}$ la classe de l'observation (malade ou non). Et on représente la loi jointe de (X, Y) par le couple $(\mathbb{P}_X, \eta)$ où $\eta(x) = \mathbb{P}(Y = 1|X = x)$, $\forall x \in \mathbb{R}$.

On appelle classifieur toute fonction $g : \mathbb{R}^d \to \{-1, 1\}$. Ainsi un classifieur prédit à quelle classe appartient une observation (dans l'exemple, il a le rôle du médecin : à partir des 3 données physiologiques, il prédit si le patient est malade). On note $\mathcal{G}$ un ensemble de classifieurs (on peut, par exemple, considérer les classifieurs qui séparent les deux classes grâce à un hyperplan). Le but de la classification est de trouver le meilleur classifieur (le meilleur

médecin) de $\mathcal{G}$ c'est-à-dire celui qui minimisera un certain critère de performance. Ici, on prend comme critère de performance l'espérance du nombre d'erreurs commises : $L(g) = \mathbb{E}[\mathbb{I}\{Y \neq g(X)\}]$. On note $\bar{g} = \underset{g\in\mathcal{G}}{\operatorname{argmin}}\, L(g)$ le meilleur classifieur de la classe $\mathcal{G}$.

Proposition 1.2.1. *Soient* $\mathcal{G}_0 = \mathcal{F}(\mathbb{R}^d, \{-1, 1\})$ *et* $g^*(x) = \begin{cases} 1 \text{ si } \eta(x) > \frac{1}{2} \\ -1 \text{ si } \eta(x) \leq \frac{1}{2} \end{cases}$, *alors*

$$\forall g \in \mathcal{G}_0,\ L^* = L(g^*) \leq L(g).$$

Mais, en pratique, on ne connaît pas la loi jointe de (X,Y) et on ne connaît donc pas g^*. On a seulement accès à des observations de la loi jointe : (X_1, Y_1), (X_2, Y_2), ... ,(X_n, Y_n). Une stratégie possible est alors de minimiser le risque empirique : $\hat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{g(X_i) \neq Y_i\}$. On note $\hat{g}_n = \underset{g\in\mathcal{G}}{\operatorname{argmin}}\, \hat{L}_n(g)$. D'après la loi des grands nombres, $\forall g \in \mathcal{G}$, $\lim\limits_{n\to\infty} \hat{L}_n(g) = L(g)$ p.s..

Mais a-t-on $\lim\limits_{n\to\infty} L(\hat{g}_n) = L^*$ p.s., c'est-à-dire en prenant une suite de classifieurs minimisant le risque empirique, approche-t-on le risque du meilleur classifieur ?

Proposition 1.2.2. *On suppose maintenant que* $g^* \in \mathcal{G}$, *alors*

$$L(\hat{g}_n) - L^* \leq 2 \sup_{g\in\mathcal{G}} |\hat{L}_n(g) - L(g)|.$$

On considère alors la loi jointe de (X, Y) représentée par le couple $(\mathbb{P}_X, \eta)$, que l'on note ν. On note $c_g = \left\{(x, y) \in \mathbb{R}^d \times \{-1, 1\}, g(x) \neq y\right\}$ et la famille d'ensembles $\mathcal{C}_\mathcal{G} = \{c_g, g \in \mathcal{G}\}$. Par Borel-Cantelli et la proposition précédente, il suffirait d'avoir une majoration de $K(n, \nu, t, \mathcal{C}_\mathcal{G})$ à décroissance exponentielle en n pour que $\lim\limits_{n\to\infty} L(\hat{g}_n) = L^*$ p.s..

Ainsi, l'étude de $K(n, \mu, t, \mathcal{C})$ permettrait de jusitifier la méthode décrite précédemment (minimiser le risque empirique pour chercher un classifieur optimal).

2 Résultats connus - Inégalités et théorèmes limites

Nous avons donné les deux principales motivations à l'étude de la probabilité $K(n, \mu, t, \mathcal{C})$. Intéressons-nous désormais au comportement de cette quantité dans un cadre asymptotique ou non et dans certains cas particuliers.

Nous séparerons en deux cas ce travail, le cas où $\mathcal{C}$ est fini et le cas où $\mathcal{C}$ est infini.

2.1 Cas fini

Nous nous intéressons ici au cas où $\mathcal{C}$ est fini. La quantité étudiée ne dépend alors plus d'un supremum mais d'un maximum. Et nous pouvons nous contenter d'étudier le cas où $\mathcal{C}$ n'a qu'un élément. En effet,

$$\mathbb{P}\left(\max_{c_j\in\{c_1,...,c_k\}} |\hat{\mu}_n(c_j)-\mu(c_j)|>t\right) = \mathbb{P}\left(\exists j\in\{1,...,k\}, |\hat{\mu}_n(c_j)-\mu(c_j)|>t\right) =$$

$$= \mathbb{P}\left(\cup_{j=1}^{k}\{|\hat{\mu}_n(c_j)-\mu(c_j)|>t\}\right) \leq \sum_{j=1}^{k}\mathbb{P}\left(|\hat{\mu}_n(c_j)-\mu(c_j)|>t\right)$$

$$\leq k\max_j \mathbb{P}\left(|\hat{\mu}_n(c_j)-\mu(c_j)|>t\right)$$

Dans cette partie, nous supposons donc que $\mathcal{C}=\{c\}$ et la probabilité étudiée devient donc :

$$K(n,\mu,t,\{c\}) = \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{\{X_i\in c\}}-\mathbb{E}[\mathbb{I}_{X_1\in c}]\right|>t\right\} = \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}Y_i-\mathbb{E}[Y_1]\right|>t\right\}$$

où $Y_1, Y_2, ..., Y_n \sim \mathcal{B}(\mathbb{P}[X_1\in c])$ *i.i.d.*. On considère, par la suite, $Y_1, Y_2, ..., Y_n \sim \mathcal{B}(p)$ *i.i.d.*.

2.1.1 Inégalités

Nous donnerons ici trois inégalités simples (pour plus d'inégalités de concentration voir l 'article [6]).

Commençons par utiliser l'inégalité bien connue de Chebychev.

Proposition 2.1.1. ***(Chebychev)*** $\forall t\in\mathbb{R}$, $\forall n\in\mathbb{N}$

$$K(n,\mu,t,\{c\}) = \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}Y_i-\mathbb{E}[Y_1]\right|>t\right\}$$

$$\leq \frac{var(\frac{1}{n}\sum_{i=1}^{n}Y_i)}{t^2} = \frac{p(1-p)}{nt^2} \leq \frac{1}{4nt^2}$$

On a alors une majoration de $K(n,\mu,t,\{c\})$ qui ne permet d'utiliser Borel-Cantelli et d'avoir la loi uniforme des grands nombres puisque $\sum_{n\in\mathbb{N}}\frac{1}{4nt^2}$ diverge.

Quant à l'inégalité de Hoeffding, elle donne une majoration à décroissance exponentielle qui permet d'utiliser Borel-Cantelli.

Proposition 2.1.2. ***(Hoeffding)*** *Les variables aléatoires Y_i étant comprises entre 0 et 1 avec une probabilité égale à 1, $\forall t \in \mathbb{R}$, $\forall n \in \mathbb{N}$,*

$$K(n, \mu, t, \{c\}) = \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} Y_i - \mathbb{E}[Y_1] \right| > t \right\} \leq 2 \exp\left(-2nt^2\right).$$

À partir d'un certain rang, l'inégalité d'Hoeffding est bien meilleur que celle de Chebychev du fait de leur comportement quand n tend vers $+\infty$.

Corollaire 2.1.2.1. *$\forall c \in \mathcal{C}$, $\hat{\mu}_n(c)$ converge uniformément presque sûrement vers $\mu(c)$.*

Preuve 2.1.2.1. *Comme la somme* $\sum_{n \in \mathbb{N}} \exp\left(-2nt^2\right)$ *converge, par Borel-Cantelli*

$$\mathbb{P}\left\{ \cap_{N \in \mathbb{N}} \cup_{n \geq N} \left\{ |\hat{\mu}_n(c) - \mu(c)| > t \right\} \right\} = 0, \ \forall t \in \mathbb{R}.$$

Or,

$$\begin{aligned} &\mathbb{P}\left\{ \cap_{N \in \mathbb{N}} \cup_{n \geq N} \left\{ |\hat{\mu}_n(c) - \mu(c)| > t \right\} \right\} \\ &= \mathbb{P}\left\{ \exists N \in \mathbb{N}, \forall n \geq N, |\hat{\mu}_n(c) - \mu(c)| > t \right\}. \end{aligned}$$

D'où

$$\begin{aligned} &\mathbb{P}\left\{ \lim_{n \to +\infty} |\hat{\mu}_n(c) - \mu(c)| = 0 \right\} \\ &= \mathbb{P}\left\{ \forall i \in \mathbb{N}, \forall N \in \mathbb{N}, \exists n \geq N, |\hat{\mu}_n(c) - \mu(c)| < \frac{1}{i} \right\} \\ &= 1 - \mathbb{P}\left\{ \exists i \in \mathbb{N}, \exists N \in \mathbb{N}, \forall n \geq N, |\hat{\mu}_n(c) - \mu(c)| > \frac{1}{i} \right\} \\ &= 1 - \mathbb{P}\left\{ \cup_{i \in \mathbb{N}} \left\{ \exists N \in \mathbb{N}, \forall n \geq N, |\hat{\mu}_n(c) - \mu(c)| > \frac{1}{i} \right\} \right\} \\ &\geq 1 - \sum_{i \in \mathbb{N}} \mathbb{P}\left\{ \exists N \in \mathbb{N}, \forall n \geq N, |\hat{\mu}_n(c) - \mu(c)| > \frac{1}{i} \right\} = 1 \ \square \end{aligned}$$

Utilisons maintenant la méthode de Chernoff (cette technique est aussi utilisée dans la démonstration de l'inégalité d'Hoeffding) : pour une variable aléatoire X et un réel t, on écrit la majoration suivante grâce à Markov $\mathbb{P}[X \geq t] = \mathbb{P}[e^{sX} \geq e^{st}] \leq \frac{\mathbb{E}[e^{sX}]}{e^{st}}$ et on minimise ensuite le membre de droite en s.

Proposition 2.1.3. ***(Chernoff)*** *On note $\forall t \in \mathbb{R}$, $\forall p \in [0, 1]$*

$$\Gamma(t, p) = (1 - p - t) \log\left(\frac{1 - p - t}{1 - p}\right) + (t + p) \log\left(\frac{p + t}{p}\right),$$

$$\Lambda(t, p) = (1 + t - p) \log\left(\frac{1 + t - p}{1 - p}\right) + (p - t) \log\left(\frac{p - t}{p}\right),$$

alors

$$K(n,\mu,t,\{c\}) = \mathbb{P}\Big[\big|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}[Y_1]\big| > t\Big]$$

$$\leq e^{-n\Gamma(t,p)} + e^{-n\Lambda(t,p)} \leq 2e^{-n.\min(\Gamma(t,p),\Lambda(t,p))}$$

Preuve 2.1.3.1. *Utilisons la méthode de Chernoff dans le cas où* $Y = \frac{1}{n}\sum_{i=1}^{n} Y_i$ *avec* $Y_i \sim \mathcal{B}(p = \mathbb{P}[X_1 \in c] = \lambda(c))$. *Et* nS *suit alors une Bernouilli de paramètres* n *et* p.

Alors $\forall s \in \mathbb{R}$, $\mathbb{P}[Y - \mathbb{E}[Y] > t] \leq \dfrac{\mathbb{E}[e^{sn(Y-\mathbb{E}[Y])}]}{e^{nst}} = e^{-sn(\mathbb{E}[Y]+t)}\mathbb{E}[e^{nsY}] = e^{-sn(p+t)}(pe^s + 1 - p)^n$. *On note* $\Gamma_1(s,t,p) = sp + st - \log(1 - p + pe^s)$ *et on a donc démontré que*

$$\forall s \in \mathbb{R},\ \mathbb{P}[\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}[Y_1] > t] \leq e^{-n\Gamma_1(s,t,p)}.$$

Minimisons maitenant le terme de droite en s, *c'est-à-dire maximisons* $\Gamma_1(s,t,p)$ *en* s. *Par le calcul des dérivées première et seconde, on trouve que* $\Gamma_1(s,t,p)$ *est minimale pour* $s^* = \log\big(\frac{(1-p)(p+t)}{p(1-p-t)}\big)$ *et alors* $\Gamma_1(s^*,t,p) = \Gamma(t,p) = (1-p-t)\log\big(\frac{1-p-t}{1-p}\big) + (t+p)\log\big(\frac{p+t}{p}\big)$. *On a donc*

$$\mathbb{P}[\hat{\mu}_n(c) - \mu(c) > t] \leq e^{-n\Gamma(t,p)} = e^{-n\left[(1-p-t)\log\left(\frac{1-p-t}{1-p}\right)+(t+p)\log\left(\frac{p+t}{p}\right)\right]}$$

En procédant de même pour $\mathbb{P}[\mathbb{E}[Y] - Y > t]$, *on obtient* $\forall s \in \mathbb{R}$, $\mathbb{P}[\mathbb{E}[Y_1] - \frac{1}{n}\sum_{i=1}^{n} Y_i > t] \leq e^{-n\Lambda_1(s,t,p)}$ *où* $\Lambda_1(s,t,p) = st - sp - \log(pe^s + 1 - p)$. *Et on trouve une valeur maximale pour* Λ_1 *à* t *et* p *fixés en* $\bar{s} = \log\big(\frac{p(1-p+t)}{(1-p)(p-t)}\big)$ *et on a alors* $\Lambda(t,p) = \Lambda_1(\bar{s},t,p) = (1+t-p)\log\big(\frac{1+t-p}{1-p}\big) + (p-t)\log\big(\frac{p-t}{p}\big)$. *On a donc*

$$\mathbb{P}[\mu(c) - \mu_n(c) > t] \leq e^{-n\Lambda(t,p)} = e^{-n\left[(1+t-p)\log\left(\frac{1+t-p}{1-p}\right)+(p-t)\log\left(\frac{p-t}{p}\right)\right]}$$

Finalement, comme les événements $\{\mu(c) - \mu_n(c) > t\}$ *et* $\{\mu_n(c) - \mu(c) > t\}$ *sont disjoints, on a*

$$K(n,\mu,t,\{c\}) = \mathbb{P}\Big[\big|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}[Y_1]\big| > t\Big]$$

$$\leq e^{-n\Gamma(t,p)} + e^{-n\Lambda(t,p)} \leq 2e^{-n.\min(\Gamma(t,p),\Lambda(t,p))}\ \square$$

Et cette borne, comme celle d'Hoeffding, décroit exponentiellement. Elle permet donc d'utiliser Borel-Cantelli.

On pourra comparer ces comportements sur la figure 1.

Nous avons ainsi tenter de décrire le compotement à distance finie (non asymptotique) de $K(n,\mu,t,\mathcal{C})$ quand $\mathcal{C} = \{c\}$. Étudions désormais le comportement asymptotique de $K(n,\mu,t,\mathcal{C})$, toujours dans le cas fini.

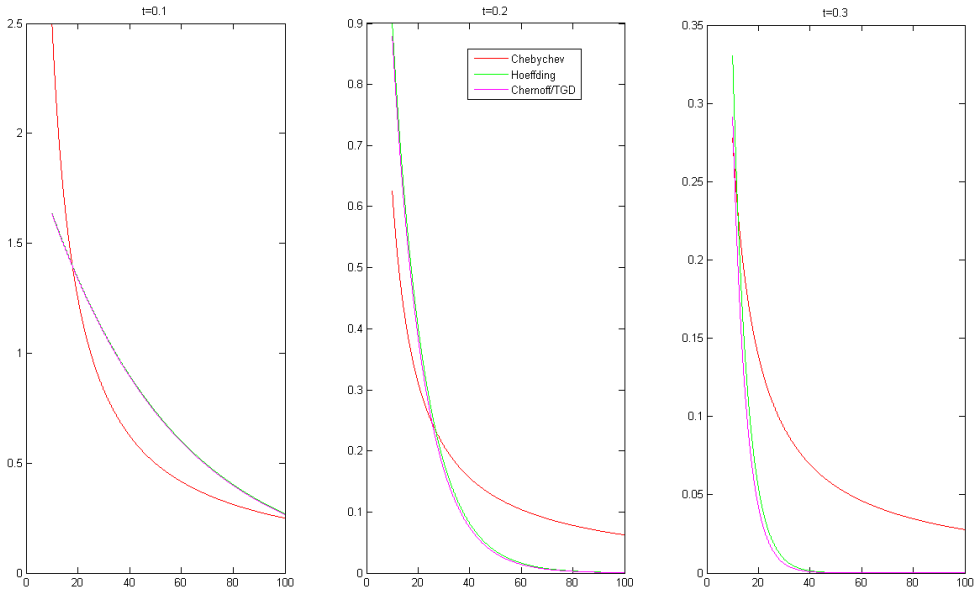

FIGURE 1: Le comportement des majorants en fonction de n pour différentes valeurs de t.

2.1.2 Techniques de grandes déviations

Les techniques de grandes déviations permettent de comprendre le comportement des événements A dont la probabilité $(\nu_n(A))$ décroit exponentiellement, avec (ν_n) une suite de mesures de probabilité. Les résultats sont donnés sous la forme $\lim_{n\to\infty} \frac{1}{n} \log(\nu_n(A))$. Pour plus de précisions, on pourra lire [9] et [3].

Commençons par donner quelques définitions générales.

Définition 2.1.1. ***(semi-continuité)*** *On dit qu'une fonction f est semi continue en x_0 si $\limsup_{x\to x_0} f(x) \leq f(x_0)$.*

Définition 2.1.2. ***(fonction de taux)*** *On dit qu'une fonction I définie sur un espace métrisable E à valeur dans $[0,\infty]$ est une fonction de taux si*

$$i)\ I \text{ n'est pas indentiquement égale à } \infty,$$

$$ii)\ I \text{ est semi-continue inférieurement,}$$

$$iii)\ \forall L \geq 0,\ \text{l'ensemble } \{x, I(x) \leq L\} \text{ est compact.}$$

Définition 2.1.3. *Une famille de mesures de probabilité $(\nu_n)_{n\in\mathbb{N}}$ vérifie le principe de grandes déviations avec une fonction de taux I si*

$$i)\ \forall F \subset E \text{ fermé},\ \limsup_{n\to\infty} \frac{1}{n} \log(\nu_n(F)) \leq -\inf_{x\in F} I(x)$$

$$ii)\ \forall O \subset E \text{ ouvert, } \liminf_{n\to\infty} \frac{1}{n}\log(\nu_n(O)) \geq -\inf_{x\in O} I(x).$$

Nous venons de donner le cadre général des techniques de grandes déviations, revenons à l'étude asymptotique de notre probabilité avec le théorème de Cramér.

Théorème 2.1.1. ***(Cramér)*** *On note Y_1, Y_2, ..., $Y_n \sim \tau$ i.i.d. à valeur dans $\mathbb{R}$. On définit $S_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ et on appelle ν_n la loi de cette somme. On introduit $\Phi(\theta) = \mathbb{E}[e^{\theta Y_1}]$ et $I(x) = \sup_\theta[\theta x - \log\Phi(\theta)]$. Alors I est une fonction de taux ; de plus, (ν_n) vérifie le principe de grandes déviations avec cette fonction de taux I.*

Donnons également quelques propriétés utiles de I.

Proposition 2.1.4. *I est convexe, atteint son minimum en $m = \mathbb{E}[Y_1]$ et est donc décroissante sur $]-\infty, m]$ et croissante sur $[m, +\infty[$.*

Utilisons ce résultat pour notre problème, c'est-à-dire Y_1, Y_2, ..., $Y_n \sim \mathcal{B}(p)$ *i.i.d.*.

Proposition 2.1.5.

$$\lim_{n\to\infty} \frac{1}{n}\log(K(n,\mu,t,\{c\})) = -\min(\Lambda(t,p), \Gamma(t,p)).$$

Preuve 2.1.5.1. *Calculons alors la fonction de taux associée à notre problème. On a $\Phi(\theta) = pe^\theta + (1-p)$ et*

$$I(x) = \sup_\theta[\theta x - \log(\Phi(\theta))] = x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right)$$

.

On obtient donc par le théorème de Cramér avec $k > t$, $F = [p+t, p+k]$ et $O =]p+t, p+k[$,

$$-\inf_{x\in]p+t,p+k[} I(x) \leq \liminf_{n\to\infty} \frac{1}{n}\log(\nu_n(]p+t,p+k[))$$

$$\leq \limsup_{n\to\infty} \frac{1}{n}\log(\nu_n([p+t,p+k])) \leq -\inf_{x\in[p+t,p+k]} I(x).$$

Or $\inf_{x\in]p+t,p+k[} I(x) = \inf_{x\in[p+t,p+k]} I(x) = I(p+t)$
et $\liminf_{n\to\infty} \frac{1}{n}\log(\nu_n(]p+t,p+k[)) = \liminf_{n\to\infty} \frac{1}{n}\log(\nu_n([p+t,p+k]))$ car $\nu_n(]p+t,p+k[) = \nu_n([p+t,p+k])$ et de même pour la limite supérieure.
Finalement,

$$\forall k > t,\ \lim_{n\to\infty} \frac{1}{n}\log(\nu_n(]p+t,p+k[))$$

$$= \lim_{n\to\infty} \frac{1}{n} \log(\nu_n([p+t, p+k])) = -I(p+t).$$

Et donc, $\lim_{n\to\infty} \frac{1}{n} \log(\nu_n(]p+t, +\infty[)) = -I(p+t)$.

En procédant de même avec $k > t$, $F = [p-k, p-t]$ *et* $O =]p-k, p-t[$, *on obtient* $\lim_{n\to\infty} \frac{1}{n} \log(\nu_n(]-\infty, p-t[)) = -I(p-t)$.

Et avec $k > t$, $F = [p-k, p-t] \cup [p+t, p+k]$ *et* $O =]p+t, p+k[\cup]p-k, p-t[$, *on a*

$$\lim_{n\to\infty} \frac{1}{n} \log(K(n, \mu, t, \{c\})) = \lim_{n\to\infty} \frac{1}{n} \log(\mathbb{P}[S_n \in]-\infty, p-t[\cup]p+t, +\infty[)$$

$$= \lim_{n\to\infty} \frac{1}{n} \log(\nu_n(]-\infty, p-t[\cup]p+t, +\infty[)) = -\min(I(p-t), I(p+t)).$$

On remarquera que $I(p-t) = \Lambda(t,p)$ *et* $I(p+t) = \Gamma(t,p)$. *D'où le résultat.*□

2.2 Cas infini

Intéressons-nous désormais au cas infini et dans un premier temps revenons à l'exemple des fonctions de répartition sur $\mathbb{R}$ dont on a déjà parlé dans les motivations.

2.2.1 Fonctions de répartition sur $\mathbb{R}$

Nous avons déjà vu que $K(n, \mu, \frac{t}{\sqrt{n}}, \mathcal{C}_0) = \mathbb{P}\left\{\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \frac{t}{\sqrt{n}}\right\}$ avec F la fonction de répartition de la loi μ, $\hat{F}_n$ la fonction de répartition empirique et $\mathcal{C}_0 = \{]-\infty, a], a \in \mathbb{R}\}$ l'ensemble des demi-droites à gauche. De nombreux résultats (tant asymptotiques que non asymptotiques) ont déjà été trouvés dans ce cas.

Donnons tout d'abord les résultats asymptotiques.

Le théorème de Glivenko-Cantelli nous donne la convergence uniforme en probabilité de la fonction de répartition empirique vers la fonction de répartition.

Théorème 2.2.1. ***(Glivenko-Cantelli)***

$$\lim_{n\to\infty} K(n, \mu, t, \mathcal{C}_0) = \lim_{n\to+\infty} \mathbb{P}\Big(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > t\Big) = 0.$$

On s'intéresse maintenant à la vitesse de convergence, Kolmogorov [8] a décrit cette vitesse :

Théorème 2.2.2. ***(Kolmogorov)***

$$\lim_{n\to\infty} K(n,\mu,\frac{t}{\sqrt{n}},\mathcal{C}_0) = \lim_{n\to+\infty} \mathbb{P}\Big(\sup_{\{x\in\mathbb{R}\}} |\hat{F}_n(x) - F(x)| > \frac{t}{\sqrt{n}}\Big)$$

$$= 2\sum_{k=1}^{+\infty} (-1)^{k-1} \exp\left(-2k^2t^2\right)$$

Ce théorème implique notamment le théorème de Glivenko-Cantelli.

Smirnov s'est quant à lui intéressé à la même probabilité sans la valeur absolue.

Théorème 2.2.3. ***(Smirnov)***

$$\lim_{n\to+\infty} \mathbb{P}\Big(\sup_{x\in\mathbb{R}} \hat{F}_n(x) - F(x) > \frac{t}{\sqrt{n}}\Big) = \exp\left(-2t^2\right)$$

en sachant que $\mathbb{P}\Big(\sup_{x\in\mathbb{R}} \hat{F}_n(x) - F(x) > \frac{t}{\sqrt{n}}\Big) \leq \mathbb{P}\Big(\sup_{\{x\in\mathbb{R}\}} |\hat{F}_n(x) - F(x)| > \frac{t}{\sqrt{n}}\Big)$.

Ces deux derniers résultats décrivent la loi de la limite de $\sqrt{n}\sup_{\{x\in\mathbb{R}\}} |\hat{F}_n(x) - F(x)|$ et de $\sqrt{n}\sup_{x\in\mathbb{R}} \left(\hat{F}_n(x) - F(x)\right)$.

Le principal résultat non asymptotique est celui de Dvoretzky-Wolfowitz-Kiefer amélioré par Massart [7] :

Théorème 2.2.4. ***(Dvoretzky-Wolfowitz-Kiefer,Massart)***

$$K(n,\mu,t,\mathcal{C}_0) = \mathbb{P}\Big(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > t\Big) \leq 2\exp\left(-2nt^2\right).$$

Et on voit qu'ici aussi, on trouve un majorant de décroissance exponentielle (comme dans le cas fini).

On remarquera que les résultats ne dépendent pas de la probabilité μ.

2.2.2 Un résultat pour les demi-espaces en dimension quelconque

Dans leur article [4], Beran et Millar nous donne le comportement en $+\infty$ de $K(n,\mu,t,\mathcal{C})$ dans $\mathbb{R}^q,\ q \geq 2$ pour $\mathcal{C}_6$ l'ensemble des demi-espaces.

Définition 2.2.1. ***(Sphère unité)*** *On note $S_q = \{s \in \mathbb{R}^q, ||s||_2 = 1\}$ la sphère unité.*

Définition 2.2.2. ***(Demi-espace)*** *Un demi-espace $A(s,u)$ est défini par deux paramètre $s \in S^q$ et $u \in \mathbb{R}$: $A(s,u) = \{x \in \mathbb{R}^q, \langle s,x\rangle \leq u\}$.*

Définition 2.2.3. *L'ensemble des demi-espaces est* $\mathcal{C}_6 = \{A(s,u), (s,u) \in S^q \times \mathbb{R}\}$.

Définition 2.2.4. ***(Processus gaussien)*** *Un processus* $\{X_t, t \in \mathbb{R}\}$ *est un processus gaussien si pour tout entier* n *et pour tout* n*-uplet* $-\infty < t_1 < t_2 < ... < t_n < +\infty$*, le vecteur* $(X_{t_1}, ...X_{t_n})$ *est un vecteur gaussien.*

Théorème 2.2.5. ***(Beran et Millar)*** *Si on note* $W = \{W(s,u), (s,u) \in S_q \times \mathbb{R}\}$ *le processus gaussien de moyenne nulle et de covariances :* $\mathbb{E}[W(s,u)W(s',u')] = \mathbb{P}[A(s,u) \cap A(s',u')] - \mathbb{P}[A(s,u)]\mathbb{P}[A(s',u')]$*, alors*

$$\lim_{n\to\infty} K\left(n, \mu, \frac{t}{\sqrt{n}}, \mathcal{C}_6\right) = \lim_{n\to\infty} \mathbb{P}\left\{\sqrt{n}\sup_{c\in\mathcal{C}_6} |\hat{\mu}_n(c) - \mu(c)| > t\right\}$$

$$= \mathbb{P}\left\{\sup_{(s,u)\in S_q\times\mathbb{R}} |W(s,u)| > t\right\}.$$

Dans ce cas particulier, Beran et Millar ont pu décrire de façon précise la loi asymptotique de $\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_6} |\hat{\mu}_n(c) - \mu(c)|$.

2.2.3 Théorie de Vapnik-Chervonenkis

Introduisons quelques concepts de la théorie de Vapnik-Chervonenkis ([10]) afin de donner des résultats sur $K(n, \mu, t, \mathcal{C})$, les plus généraux possibles et non asymptotiques.

Tout d'abord, donnons quelques définitions afin de décrire la complexité combinatoire de $\mathcal{C}$.

Définition 2.2.5. ***(trace)****La trace de* $\mathcal{C}$ *sur un ensemble de points de* $\mathbb{R}^d$ $(x^n = \{x_1, x_2, ...x_n\}, x_1, x_2, ...x_n \in \mathbb{R}^d)$ *est* $tr(\mathcal{C}, x^n) = \{x^n \cap c, c \in \mathcal{C}\}$.

Exemple 2.2.5.1. *Dans* $\mathbb{R}$*, pour* $x^3 = \{-1, 2, 3\}$ *et* $\mathcal{C}_2 = \{]-\infty, a], [a, +\infty[, a \in \mathbb{R}\}$ *l'ensemble des demi-droites,* $tr(\mathcal{C}, x^n) = \{\oslash, \{-1\}, \{-1, 2\}, \{-1, 2, 3\}, \{2, 3\}, \{3\}\}$.

Définition 2.2.6. *Le cardinal de la trace est noté* $N(\mathcal{C}, x^n) = |tr(\mathcal{C}, x^n)|$*. On remarque que* $N(\mathcal{C}, x^n) \leq 2^n$ *(nompre de parties de* x^n*).*

Exemple 2.2.6.1. *On reprend l'exemple précédent et* $N(\mathcal{C}, x^n) = 6 \leq 2^3$.

Définition 2.2.7. ***(n-iéme coefficient de pulvérisation)****On appelle le* n*-iéme coefficient de pulvérisation de* $\mathcal{C}$ *le plus grand cardinal de la trace de* $\mathcal{C}$ *et de tous les ensembles de* n *points de* $\mathbb{R}^d$ *:* $s(\mathcal{C}, n) = \sup_{x^n} N(\mathcal{C}, x^n)$.

Exemple 2.2.7.1. *Pour l'exemple précédent,* $s(\mathcal{C}_2, 1) = 2 = 2^1$*,* $s(\mathcal{C}_2, 2) = 4 = 2^2$*,* $s(\mathcal{C}_2, 3) = 6 < 2^3$.

Définition 2.2.8. ***(dimension VC)*** *La dimension VC de $\mathcal{C}$ est $V(\mathcal{C}) = \max\{k \in \mathbb{N} : s(\mathcal{C}, k) = 2^k\}$. C'est la plus grande valeur de k telle qu'il existe un ensemble x^k de k points de $\mathbb{R}^d$ tel qu'en l'interceptant avec les ensembles c de $\mathcal{C}$, on obtient toutes les parties de x^k.*

Exemple 2.2.8.1. *Toujours pour l'exemple précédent, on a $V(\mathcal{C}_2) = 2$.*

Exemple 2.2.8.2. *Dans $\mathbb{R}$, pour $\mathcal{C}_0 = \{]-\infty, a], a \in \mathbb{R}\}$, c'est-à-dire l'ensemble des demi-droites à gauche, $V(\mathcal{C}_0) = 1$.*

Exemple 2.2.8.3. *Dans $\mathbb{R}$, pour $\mathcal{C}_3 = [a, b], a, b \in \mathbb{R}\}$, c'est-à-dire l'ensemble des segments, $V(\mathcal{C}_3) = 2$.*

Exemple 2.2.8.4. *Dans $\mathbb{R}^2$, pour $\mathcal{C}_4 = \{]-\infty, a]\times]-\infty, b], a, b \in \mathbb{R}\}$ l'ensemble des orthants, $V(\mathcal{C}_4) = 2$.*

Exemple 2.2.8.5. *Dans $\mathbb{R}^2$, pour $\mathcal{C}_5$ l'ensemble des polygones, $V(\mathcal{C}_5) = +\infty$.*

Exemple 2.2.8.6. *En dimension quelconque d, pour l'ensemble des demi-espaces $\mathcal{C}_6$, $V(\mathcal{C}_6) = d + 1$.*

Définition 2.2.9. ***(entropie VC)*** *Soit X^n un échantillon i.i.d. $X_1, X_2, ...X_n$ de loi μ, l'entropie VC de $\mathcal{C}$ et μ est : $H(\mathcal{C}, \mu, n) = \mathbb{E}[ln(N(\mathcal{C}, X^n))]$.*

Maintenant que nous avons défini la dimension combinatoire, énonçons quelques résultats.

Sous certaines hypothèse, on peut généraliser le théorème de Glivenko-Cantelli : (théorème dû à Vapnik et Chevonenkis)

Théorème 2.2.6. ***(Vapnik et Chevonenkis)*** *Pour une probabilité donnée μ et une famille d'ensembles $\mathcal{C}$:*

$$\forall t \in \mathbb{R} \lim_{n\to+\infty} K(n, \mu, t, \mathcal{C}) = \lim_{n\to\infty} \mathbb{P}\left\{\sup_{c\in\mathcal{C}} |\hat{\mu}_n(c) - \mu(c)| > t\right\} = 0$$

$$ssi \lim_{n\to+\infty} \frac{H(\mathcal{C}, \mu, n)}{n} = 0.$$

Remarque 2.2.6.1. *Si la dimension VC de $\mathcal{C}$ est finie, alors quelque soit la probabilité μ, l'hypothèse sera vérifiée et on aura donc une convergence uniforme en probabilité de $\hat{\mu}_n$ vers μ.*

Et si la dimension VC de $\mathcal{C}$ est finie, Vapnik et Chervonenkis nous donnent une majoration de $K(n, \mu, t, \mathcal{C})$.

Théorème 2.2.7. ***(Vapnik et Chevonenkis)*** *Si $V(\mathcal{C}) < \infty$, $\forall t \in \mathbb{R}$, $\forall n \in \mathbb{N}$*

$$\sup_{\mu} K(n, \mu, t, \mathcal{C}) \leq 4\Big(\frac{2en}{V(\mathcal{C})}\Big)^{V(\mathcal{C})} \exp\left(-\frac{nt^2}{8}\right).$$

Cette majoration est uniforme en μ, le majorant est donc valable dans le pire des cas. On se doute donc que dans des cas particuliers, cette majoration n'est pas optimale.

3 Expériences numériques

3.1 Cadre des simulations et objectifs

Nous allons maintenant estimer la probabilité étudiée $K(n, \mu, t, \mathcal{C})$ afin de pouvoir comparer la théorie à la réalité.

Donnons tout d'abord le cadre des simulations. Nous travaillerons dans $\mathbb{R}$ avec μ et n fixés. On prendra $\mu \sim \mathcal{U}([0,1])$ et on étudiera plusieurs cas de familles $\mathcal{C}$ d'ensembles.

Théoriquement, nous avons trouvé plusieurs majorations de $K(n, \mu, t, \mathcal{C})$ (les inégalités dues à Chebychev, Hoeffding, Chernoff et Vapnik-Chervonenkis) et un comportement asymptotique dû aux techniques de grandes déviations.
Chebychev :

$$K(n, \mu, t, \{c\}) \leq \frac{1}{4nt^2}$$

Hoeffding :

$$K(n, \mu, t, \{c\}) \leq 2\exp\left(-2nt^2\right)$$

Chernoff :

$$K(n, \mu, t, \{c\}) \leq \exp\left(-n\Gamma(t,p)\right) + \exp\left(-n\Lambda(t,p)\right)$$

$$\leq 2\exp\left(-n\min(\Gamma(t,p), \Lambda(t,p))\right)$$

Vapnik-Chervonenkis :

$$K(n, \mu, t, \mathcal{C}) \leq 4\Big(\frac{2en}{V(\mathcal{C})}\Big)^{V(\mathcal{C})} \exp\left(-\frac{nt^2}{8}\right)$$

Techniques de grandes déviations :

$$\lim_{n\to\infty} \frac{1}{n} \log(K(n, \mu, t, \{c\})) = -\min(\Lambda(t,p), \Gamma(t,p))$$

Théoriquement, nous avonc des résultats dans le cas où $\mathcal{C}$ est fini et un résultat dans le cas où $\mathcal{C}$ est infini (Vapnik-Chervonenkis). Mais ce dernier

résultat est une inégalité qui est vraie dans le pire des cas, elle ne doit donc pas être optimale dans les cas précis que nous allons étudiés. Quant aux résultats dans le cas fini, ils ne tiennent pas en compte du suprémum. Néanmoins, vu l'ensemble des résultats, on s'attend à trouver un comportement de $K(n, \mu, t, \mathcal{C})$ à décroissance exponentielle en n.

3.2 Difficultés

Les difficultés pour simuler numériquement cette probabilité sont dues à deux facteurs. Tout d'abord on se confronte à la simulation d'évènements rares, de plus on doit calculer un suprémum sur un ensemble infini.

3.2.1 Simulation d'évènements rares

Les évènements rares sont difficiles à simuler, car il faut un nombre très important de simulations avant de voir l'évènement se produire. Prenons un exemple : nous voulons estimer la probabilité qu'un évènement rare A^j, dépendant de $X_1^j, X_2^j, ..., X_n^j$, se produise. Les n-uplets $(X_1^1, X_2^1, ..., X_n^1), (X_1^2, ..., X_n^2)$, $..., (X_1^N, ..., X_n^N)$ sont i.i.d. On estime cette probabilité par $\hat{p}_N = \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}(A^j)$. On cherche à trouver le nombre (N) de n-uplets qu'il faut simuler afin que l'erreur relative $(\frac{\hat{p}_N - p}{p})$ soit inférieure à 1% avec une probabilité supérieure à $1 - \alpha$.

Sachant que $\mathbb{P}(A^1), \mathbb{P}(A^2), ..., \mathbb{P}(A^N)$ sont i.i.d. de loi $\mathcal{B}(p)$, par le théorème central limite, on obtient

$$\sqrt{N}\frac{\hat{p}_N - p}{\sqrt{p(1-p)}} \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} Z \sim \mathcal{N}(0, 1).$$

On note $t_{\frac{\alpha}{2}}$ le réel (quantile de la loi normale) tel que $\mathbb{P}\left\{|Z| < t_{\frac{\alpha}{2}}\right\} = 1 - \alpha$. Il faut alors que $N \geq \frac{1-p}{p}\left(\frac{t_{\frac{\alpha}{2}}}{0.01}\right)^2 \sim \frac{1}{p}\left(\frac{t_{\frac{\alpha}{2}}}{0.01}\right)^2$ (pour p petit) pour que $\mathbb{P}\left\{\frac{|\hat{p}_N - p|}{p} < 1\%\right\} = 1 - \alpha$. On s'apperçoit que plus l'événement est rare, plus il faut simuler d'échantillons. Avec $p = 10^{-10}$ et $\alpha = 5\%$, il faut simuler plus de 3.10^{14} n-uplets $(X_1^j, X_2^j, ..., X_n^j)$ pour avoir la précision voulue (erreur relative de 1% avec une probabilité de 95%) ! ! !

3.2.2 Le problème dû au suprémum

Le second problème rencontré est dû au suprémum (qui posait déjà un problème théoriquement, notamment pour passer du cas fini au cas infini). En effet, pour estimer $K(n, \mu, t, \mathcal{C})$, on simule des variables $X_1^1, X_2^1, ..., X_n^1, X_1^2, X_2^2$ $..., X_n^2, ..., X_1^N, X_2^N, ..., X_n^N$ i.i.d.de loi μ. On note $\hat{\mu}_n^j(c) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i^j \in c\}$ la

mesure empiriques de μ associée à l'échantillon $X_1^j, X_2^j, ..., X_n^j$. Puis on calcule $\hat{p}_N = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}\left\{\sup_{c\in\mathcal{C}} |\hat{\mu}_n^j(c) - \mu(c)| > t\right\}$ qui est l'estimateur par la méthode des moments de $K(n, \mu, t, \mathcal{C})$. On remarque alors que l'estimation de $K(n, \mu, t, \mathcal{C})$ nécessite le calcul de $\sup_{c\in\mathcal{C}} |\hat{\mu}_n^j(c) - \mu(c)|, \forall\ 1 \leq j \leq N$ qui est un suprémum sur un ensemble infini en général. Les calculs sont difficiles car on cherche le suprémum d'une fonction à valeur cotinue et on ne peut donc pas énumérer ses valeurs.

3.3 Résultats

3.3.1 Résolution du problème dû au suprémum

Nous résolvons tout d'abord le problème du suprémum (on s'arrangera pour prendre des valeurs de t et n petites afin de ne pas recontrer de problème du fait de la simulation d'évènements rares). Dans $\mathbb{R}$, ce problème se résout assez facilement car le suprémum est en fait un maximum. En premier lieu, on considère $\mathcal{C}_0 = \{[0, a], a \in \mathbb{R}\}$ alors $F_n^j(x) = \mu_n^j([0, x]), x \in \mathbb{R}$ ne change de valeur qu'en n points : $X_1^j, X_2^j, ..., X_n^j$ (Figure 2). Donc $\sup_{c\in\mathcal{C}_0} |\mu_n^j(c) - \mu(c)| = \max_{x\in\{X_1^j,..,X_n^j\}} |\mu_n^j([0, x[) - \mu([0, x[)|, |\mu_n^j([0, x]) - \mu([0, x])|$.

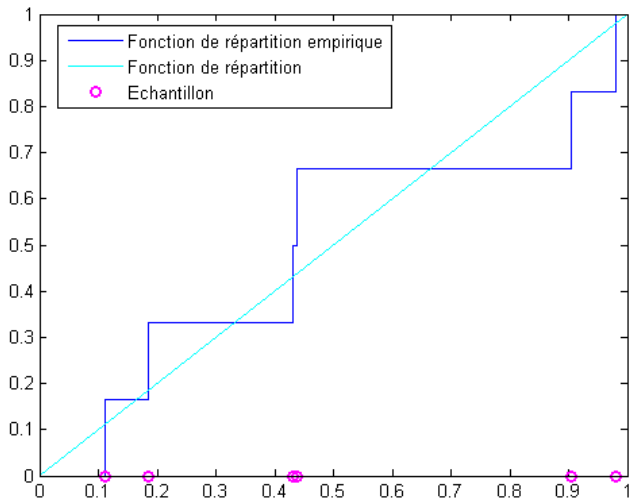

FIGURE 2: Fonction de répartition empirique pour un échantillon donné

On remarque que dès que la dimension de l'espace considéré ou que la complexité de la famille d'ensembles augmente, il est plus difficile de trouver une méthode pour calculer le suprémum. Par exemple, lorsque nous considérons les intervales $\mathcal{C}_3 = \{[a, b], a, b \in \mathbb{R}\}$ sur $\mathbb{R}$ alors le calcul du suprémum a une

complexité qui est égale au carré de celle du calcul du suprémum pour les demi-droites à gauche, en effet : $\sup_{c\in\mathcal{C}_3}|\mu_n^j(c)-\mu(c)| = \max_{x\leq y\in\{X_1^j,..,X_n^j\}}\Big(|\mu_n^j([x,y[)-\mu([x,y[)|, |\mu_n^j([x,y])-\mu([x,y])|, |\mu_n^j(]x,y])-\mu(]x,y])|, |\mu_n^j(]x,y[)-\mu(]x,y[)|\Big)$. On adaptera ces exemples dans les autres cas.

3.3.2 Ensembles critiques

Introduisons désormais la notion d'ensemble critique. Par la méthode de Chernoff, nous avons trouvé un résultat non asymptotique très similaire au résultat asymptotique obtenu par les techniques de grandes déviations :

$$K(n,\mu,t,\{c\}) \leq e^{-n\Gamma(t,p)} + e^{-n\Lambda(t,p)} \leq 2e^{-n.\min(\Gamma(t,p),\Lambda(t,p))}$$

et

$$\lim_{n\to\infty}\frac{1}{n}\log(K(n,\mu,t,\{c\})) = -\min(I(p-t),I(p+t)),$$

où $I(p-t)=\Lambda(t,p)=(1+t-p)\log\left(\frac{1+t-p}{1-p}\right)+(p-t)\log\left(\frac{p-t}{p}\right)$ et $I(p+t)=\Gamma(t,p)=(1-p-t)\log\left(\frac{1-p-t}{1-p}\right)+(t+p)\log\left(\frac{p+t}{p}\right)$. On pense que le suprémum est principalement atteint pour les valeurs de p (dites valeurs critiques) qui minimisent $\min(I(p-t),I(p+t)) = \min(\Gamma(t,p),\Lambda(t,p))$ à t fixé (c'est-à-dire celles qui maximisent le majorant dans le cas fini et la valeur asymptotique). Á défaut de pouvoir faire les calculs exacts, on trace $\Gamma(t,p)$ et $\Lambda(t,p)$ à t fixé et on cherche leur minimum (on pourra voir ce tracé sur la Figure 3, les courbes les plus hautes correspondent aux t les plus grands).

Par un développement asymptotique, on obtient que $\Gamma(t,p) \underset{t\to 0}{=} \frac{t^2}{p(1-p)}+o(t^2)$ et de même $\Lambda(t,p) \underset{t\to 0}{=} \frac{t^2}{p(1-p)}+o(t^2)$. Ainsi le minimum de $\Gamma(t,p)$ et $\Lambda(t,p)$ est atteint pour p proche de $\frac{1}{2}$ pour t fixé petit (car $\frac{t^2}{p(1-p)}$ est minimum pour $p=\frac{1}{2}$), ceci se voit aussi sur la Figure 3.

Par un calcul simple, on voit que $\Lambda(t,p)=\Gamma(t,1-p)$ (ce qui se remarque aussi sur la figure 3). Ainsi pour un t fixé, si $\Gamma(t,p)$ atteint son minimum en p^* alors $\Lambda(t,p)$ atteint son minimum en $1-p^*$.

Nous allons noté $p_\Gamma^*(t) = \underset{p,\exists c\in\mathcal{C},\mu(c)=p}{\operatorname{argmin}}\Gamma(t,p)$, $p_\Lambda^*(t) = \underset{p,\exists c\in\mathcal{C},\mu(c)=p}{\operatorname{argmin}}\Lambda(t,p)$ (n'est pas forcément égale à $1-p_\Gamma^*(t)$) et $p^*(t) = \underset{p,\exists c\in\mathcal{C},\mu(c)=p}{\operatorname{argmin}}\min[\Gamma(t,p),\Lambda(t,p)]$. On a alors $\sup_{\{c\},c\in\mathcal{C}} K(n,\mu,t,\{c\}) \leq 2e^{-n.\min(\Gamma(t,p^*),\Lambda(t,p^*))}$. Attention, ces arguments minimum ne sont pas forcément uniques ! !

Pour un $p^*(t)$ donné, un ensemble critique est un ensemble $c\in\mathcal{C}$ tel que $\mu(c)=p^*(t)$. Par exemple dans le cas de $\mathcal{C}_0=\{[0,a],a\in[0,1]\}$ et $\mu\sim\mathcal{U}([0,1])$, $c=[0,p^*(t)]$ est le seul ensemble critique associé à $p^*(t)$ alors que pour $\mathcal{C}_3=\{[a,b],a,b\in\mathbb{R}\}$, les ensembles critiques sont les ensembles de la forme $c=[a,a+p^*(t)],a\in[0,1-p^*(t)]$. Introduisons une nouvelle famille d'ensembles $\mathcal{C}_7(maxa)=\{[0,a],0\leq a<maxa\leq 1\}$. On remarque que sa dimension VC est la même que celle de $\mathcal{C}_0$, alors que cette nouvelle famille semble moins

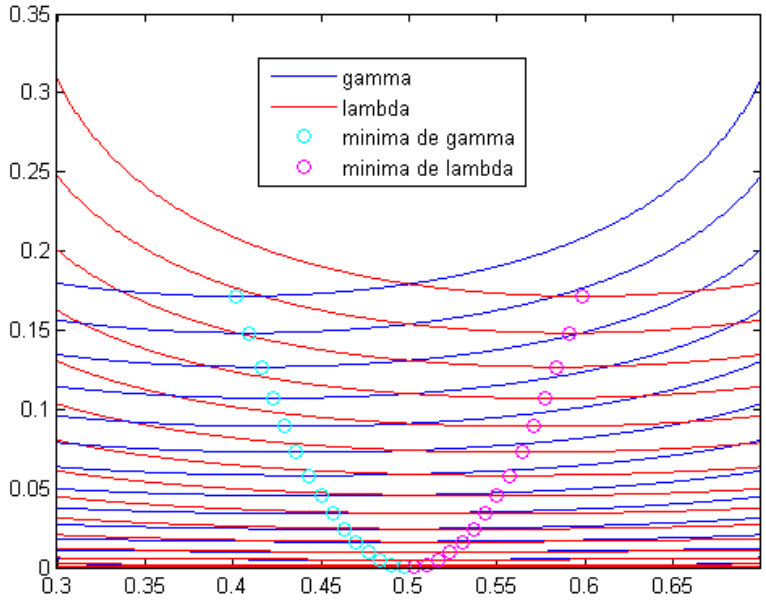

FIGURE 3: Γ et Λ en fonction de p pour différentes valeurs de $t \in [0.01, 0.3]$ et leur minimum

riche!! Soit $t \in \mathbb{R}$ tel que $p^*(t)(\mathcal{C}_0) > maxa$, alors $p^*(t)(\mathcal{C}_7) = maxa$. Cette dernière propriété est visible sur la Figure 4 où nous voyons la répartition des $a \leq maxa$ ($c = [0, a]$) pour lesquels le $\sup_{c \in \mathcal{C}_7(maxa)} |\hat{\mu}_n(c) - \mu(c)|$ est atteint.

3.3.3 L'échantillonnage préférentiel

Comme nous l'avons remarqué dans la partie précédente, une des difficultés dans ce type de simulation est due à la simulation d'évènements rares. Ce problème peut être résolu grâce à l'échantillonnage préférentiel. Donnons le principe de cette méthode.

On veut estimer $p = \mathbb{E}[\eta(X)]$ où X est une variable aléatoire de loi μ et η une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$. Habituellement, on utilise l'estimateur des moments qui converge presque sûrement vers p, par la loi forte des grands nombres. Soient $X_1, X_2, ..., X_N$ i.i.d. de loi μ, l'estimateur des moments associé à cet échantillon est $\hat{p}_N = \frac{1}{N} \sum_{j=1}^{N} \eta(X_j)$. Utilisons une autre méthode pour estimer p. Soit ν une mesure de probabilité telle que μ est absolument continue par rapport à ν (ainsi la dérivée de Radon-Nikodym $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ existe) et $Y_1, Y_2, ..., Y_N$ i.i.d. de loi ν. On estime p par $\bar{p}_N = \frac{1}{N} \sum_{j=1}^{N} \eta(Y_j) \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(Y_j)$ qui converge également presque sûrement vers p puisque $\mathbb{E}[\bar{p}_N] = p$.

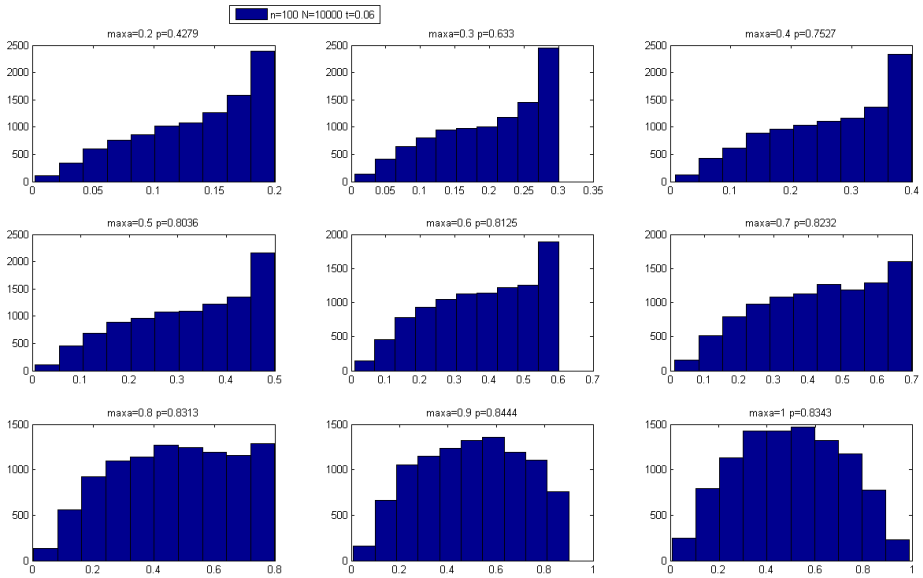

FIGURE 4: Histogramme des valeurs de a pour lesquels le sup est atteint sachant que numériquement : $p^*_\Gamma(0.06)(\mathcal{C}_0) \sim 0.48$ et $p^*_\Lambda(0.06)(\mathcal{C}_0) \sim 0.52$

Il faut désormais trouver le bon changement de mesure (celui qui minimisera la variance), c'est-à-dire celui qui augmente la probabilité de l'évènement rare ($\{\sup_{c\in\mathcal{C}} |\hat{\mu}_n(c)-\mu(c)|\}$) que l'on veut mesurer. On fait alors de l'échantillonnage préférentiel. On a vu que le suprémum étudié était plus fréquemment atteint sur les ensembles critiques, on note c^* un ensemble critique assocoé à t, μ et $\mathcal{C}$. On veut donc augmenter la probabilité de l'évènement rare $\{|\hat{\mu}_n(c^*) - \mu(c^*)| > t\}$ qui est la réunion des deux évènements disjoints suivants $\{\hat{\mu}_n(c^*) > \mu(c^*) + t\}$ et $\{\hat{\mu}_n(c^*) < \mu(c^*) - t\}$. On choisit d'augmenter la probabilité du premier évènement $\{\hat{\mu}_n(c^*) > \mu(c^*) + t\}$. On choisit donc une mesure de proba μ qui donne un poids $\mu(c^*) + t$ à c^* (i.e. $\nu(c^*) = \mu(c^*) + t$) et $\nu((c^*)^c) = 1 - (\mu(c^*) + t)$.

On applique cette méthode au cas que nous étudions ($\mu \sim \mathcal{U}([0,1])$), pour $\mathcal{C}_7(maxa) = \{[0,a], a \in [0, maxa[\}$. Soit $t \in \mathbb{R}$ assez petit (ainsi $p^*(t)(\mathcal{C}_0) \sim \frac{1}{2}$), alors

$$p^*(t)(\mathcal{C}_7(maxa)) = \begin{cases} \frac{1}{2} \text{ si } maxa > \frac{1}{2} \\ maxa \text{ si } maxa \leq \frac{1}{2} \end{cases} = ma.$$

On prends alors ν de densité $f_\nu = (1+\frac{t}{ma})\mathbb{I}_{[0,ma]} + \frac{1-ma-t}{1-ma}\mathbb{I}_{[ma,1]}$. C'est également la probabilité obtenue par le changement exponentiel de mesure. Dans les Figure 5 et 6, on compare les variances des estimateurs de la probabilité étudiée avec et sans échantillonnage préférentiel. Et on voit que pour des probabilités inférieures à 0.02, l'estimateur avec échantillonnage préférentiel est plus fiable (de variance faible) que celui sans échantillonnage préférentiel.

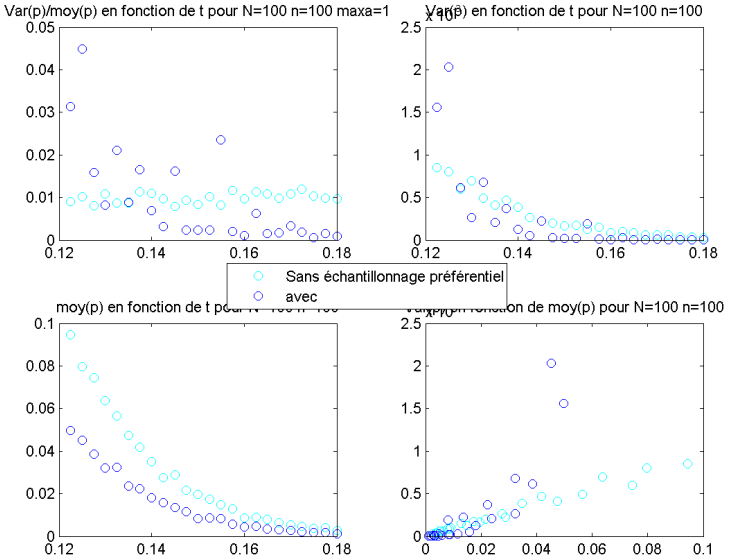

FIGURE 5: De gauche à droite et de haut en bas : $\frac{var(\hat{p})}{moy(\hat{p})}$, $var(\hat{p})$, $\hat{p}$ enfonction de t et $var(\hat{p})$ en fonction de $moy(\hat{p})$ pour les estimateurs avec et sans échantillonnage préférentiel. Nombre de simulation par estimation : $N = 100$, $n = 100$

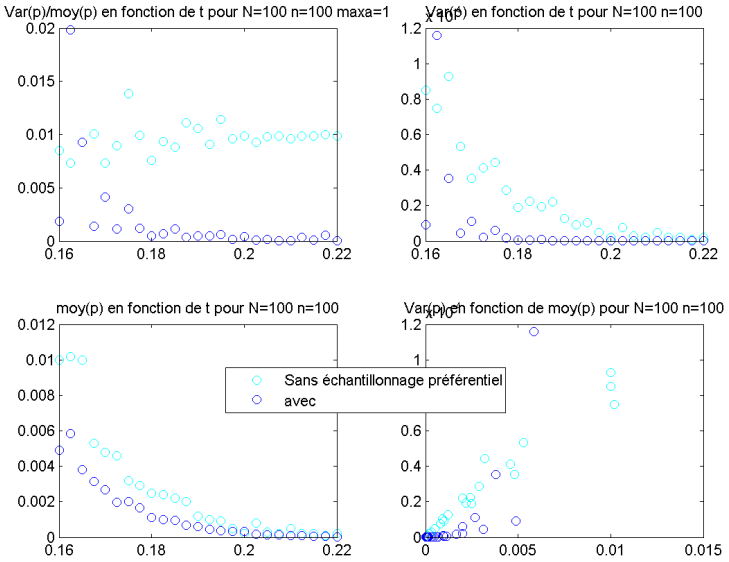

FIGURE 6: De gauche à droite et de haut en bas : $\frac{var(\hat{p})}{moy(\hat{p})}$, $var(\hat{p})$, $\hat{p}$ enfonction de t et $var(\hat{p})$ en fonction de $moy(\hat{p})$ pour les estimateurs avec et sans échantillonnage préférentiel. Nombre de simulation par estimation : $N = 100$, $n = 100$

3.3.4 Comparaison théorie/réalité

Nous comparons sur la Figure 7 les comportements des majorants obtenus par Chebychev, Hoeffding et Chernov au comportement des estimateurs avec et sans échantillonnage préférentiel. On remarquera que l'estimateur avec échantillonnage préférentiel n'est valable que pour ses petites valeurs (≤ 0.02) et donc pour les grandes valeurs de n (pour $n \leq 170$, la courbe de l'estimateur avec échantillonnage préférentiel admet des piques qui traduisent une variance élevée). On voit également que les majorants trouvés dans le cas fini majorent ici les estimateurs (si on ne tient pas en compte des piques dus à une variance trop élevée et donc à une mauvaise estimation). Le comportement à décroissance exponentiel est donc confirmé. Dans cet exemple, on a pris comme valeur critique $p^* = \frac{1}{2}$, le majorant trouvé par la méthode de Chernov est alors égal à celui d'Hoeffding. On pense que pour avoir une meilleure majoration, il faut plus tenir compte des ensembles critiques et de la complexité de $\mathcal{C}$. En effet, sur la figure 4, on peut voir que l'estimateur dépend fortement de cette famille d'ensembles.

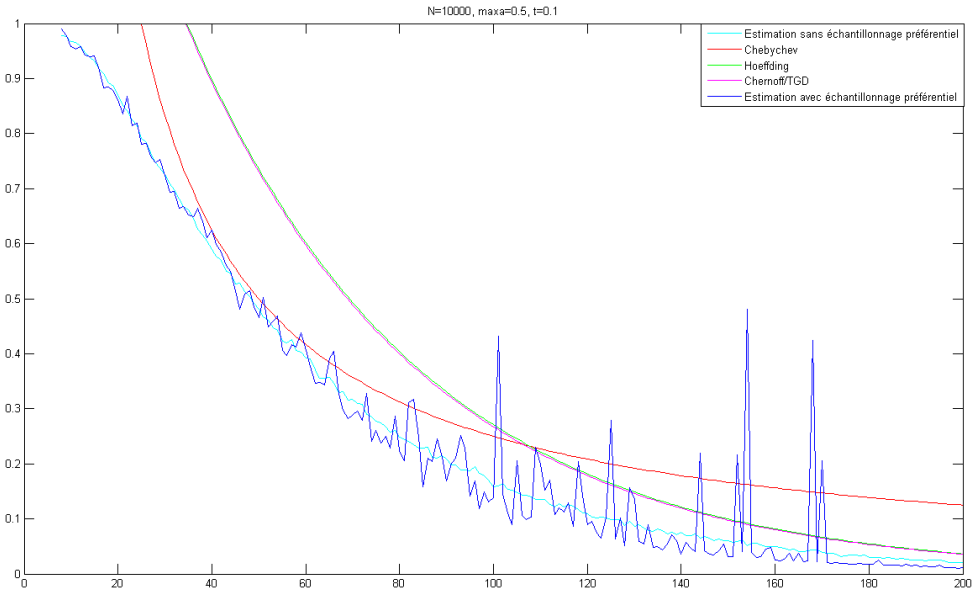

FIGURE 7: Comparaison du comportement des majorants trouvés et ses estimations avec et sans échantillonnage préférentiel en fonction de n. Nombre de simulation par estimation : 10000, $maxa = 0.5$, $t = 0.1$

Conclusion et perspectives

Dans ce stage, nous avons étudié le comportement théorique asymptotique ou non de la quantité $K(n, \mu, t, , \mathcal{C})$. Le comportement de cette quantité a été décrit précisément pour les fonctions de répartition avec Kolmogorov-Smirnov et son comportement général dans le cadre de la théorie de Vapnik-Chervonenkis (dans le pire des cas). Nous avons trouvé des résultats dans le cas fini grâce aux techniques de grandes déviations et la méthode de Chernov. Nous avons également étudié numériquement cette quantité dans un cas précis et nous l'avons comparée aux résultats théoriques. Nous avons alors confirmé le comportement à décroissance exponentielle de cette quantité. Et nous avons remarqué une forte dépendance de cette quantité avec les ensembles critiques de $\mathcal{C}$. Nous souhaiterions poursuivre cette étude en améliorant notre connaissance sur cette dépendance, en étudiant notemment d'aures familles d'ensembles.

Nous aimerions également étudié un cas particulier dans $\mathbb{R}^2$ avec une uniforme sur $[0, 1] \times [0, 1]$ et les orthants. La principale difficulté serait le calcul du suprémum (plus difficile à calculer que dans $\mathbb{R}^2$. On pourrait s'inspirer des travaux de Justel, Peña et Zamar ([2] [1]) sur les tests d'adéquation en dimension deux pour résoudre ce problème.

Bibliographie

[1] D. Peña et R. Zamar A. Justel. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, 35 :251–259, 1997.

[2] D. Peña et R. Zamar A. Justel. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics and Econonometrics*, 13 :1–16, Septembre 1994.

[3] James Antonio Bucklew. *Introduction to Rare Event Simulation*. Springer, 2004.

[4] R. Beran et P. W. Millar. Confidence sets for a multivariate distribution. *The annals of Statistics*, 14 :431–443, 1986.

[5] Gábor Lugosi. Pattern classification and learning theory, notes de CISM international centre for machanical sciences, 2002.

[6] Gábor Lugosi. Concentration of measure inequalities, notes de machine learning summer school 2003, 2005.

[7] P.Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The annals of probability*, 1990.

[8] C. Read B. Vidakovic et N. L. Johnson S. Kotz, N. Balakrishnan. *Encyclopedia of Statistical Sciences*, volume 2. John Wiley & Sons, 1988.

[9] D.W. Stroock. *An Introduction to the theory of Large Deviations*. Springer Verlag, 1984.

[10] Nicolas Vayatis. Inégalités de Vapnik-Chervonenkis et mesures de complexité, 2000.