

Faculté des sciences

Inégalités de concentration

Pour évènements rares

Auteur : Stéphane Lhaut

Promoteur : Johan Segers

Lecteurs : Augusto Ponce et Rainer von Sachs

Année académique 2020-2021

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
FACULTÉ DES SCIENCES
ÉCOLE DE MATHÉMATIQUES

Inégalités de concentration pour évènements rares

Stéphane Lhaut
Promoteur : Johan Segers
Lecteurs : Augusto Ponce et Rainer von Sachs

Mémoire de Master
Juin 2021

Remerciements

Quand je pense à mon mémoire, la première personne qui me vient en tête est certainement mon promoteur, le Professeur Johan Segers. Je voudrais vivement le remercier pour sa grande disponibilité ainsi que pour ses précieux conseils, tant au niveau du fond que de la forme. Malgré un contexte sanitaire difficile, c'est toujours avec grand plaisir que j'ai rejoint les nombreuses réunions Teams qui ont permis d'aboutir au mémoire qui suit.

Je remercie également les Professeurs Augusto Ponce et Rainer von Sachs qui ont accepté d'être lecteurs de mon travail. Les différentes matières que vous m'avez enseignées durant mon parcours académique furent indispensables à la réalisation d'un tel travail. J'espère que sa lecture vous sera agréable.

Je pense aussi à mes parents qui ont eu la lourde tâche de relire toutes ces pages sans pouvoir en comprendre le sens : merci. Si des erreurs orthographiques subsistent, que j'en sois désigné comme l'unique responsable.

Introduction

Quoi de plus frustrant pour toute personne disposant d'une boîte aux lettres électronique de voir cette dernière envahie de « spams » provenant d'internautes malveillants ou de personnes peu scrupuleuses s'adonnant à des pratiques telles que le « fishing » ? Heureusement, la majorité des fournisseurs de boîtes mails disposent aujourd'hui d'algorithmes de *classification* qui effectuent un premier tri des courriers électroniques qui nous sont adressés. Néanmoins, ce tri n'est pas toujours correct et des erreurs sont parfois commises. Un des rôles de la *théorie de l'apprentissage statistique* est de quantifier la probabilité de survenance de telles erreurs [15] [7].

Par exemple, supposons que l'algorithme de tri se base sur d quantités liées au mail (nombres de caractères, nombres d'espaces, etc.), notées $X \in \mathbb{R}^d$, pour décider si celui-ci est un spam ou non et qu'il associe à ces données un nombre $Y \in \{0, 1\}$ (où 1 = « le mail est un spam » et 0 = « le mail n'est pas un spam ») via une *fonction de classification* $g : \mathbb{R}^d \rightarrow \{0, 1\}$. Nous dirons qu'il commet une erreur si $g(X) \neq Y$ et, la probabilité d'une telle erreur est donnée par le *risque*

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

Le problème est que cette dernière quantité n'est pas explicitement calculable car elle requiert la connaissance de la distribution jointe du couple (X, Y) , et, dans les cas pratiques, nous disposons seulement d'exemples provenant du passé contenus dans un échantillon $(X_i, Y_i)_{i=1}^n$. Chercher une fonction de classification g qui minimiserait donc le risque $L(g)$ semble impossible... Une solution est de travailler avec le *risque empirique* associé à l'échantillon donné par

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

Supposons que la fonction de classification g soit choisie au sein d'une classe (potentiellement très large) $\mathcal{G}$. Un algorithme pourrait alors simplement choisir une fonction g^* (elle n'est pas nécessairement unique) qui minimise $\hat{L}_n$ parmi toutes les fonctions de $\mathcal{G}$. Si nous voulons quantifier la probabilité qu'une erreur soit commise par un tel algorithme, nous nous apercevons rapidement (voir [15]) qu'une quantité d'intérêt est la *déviatio*

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

En particulier, nous cherchons à établir des résultats de la forme

$$\mathbb{P} \left(\sup_{g \in \mathcal{G}} |\hat{L}_n(g) - L(g)| \geq t \right) \leq B(n, t),$$

pour n'importe quel $t > 0$, où $B(n, t)$ est une borne indépendante de la distribution des données. C'est un premier exemple de ce que nous appellerons *inégalité de concentration* dans la suite.

De manière générale, le présent texte s'inscrit dans le cadre similaire de la théorie des *processus empiriques* [28] indexés par une classe $\mathcal{A}$ (souvent infinie) de boréliens de l'espace euclidien $\mathbb{R}^d$. Nous étudions la mesure empirique associée à un échantillon indépendant et identiquement distribué de données $X_1, \dots, X_n \sim \mu$ (où μ est une mesure de probabilité inconnue sur $\mathbb{R}^d$) donnée par

$$\mu_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

Plus spécifiquement, nous établissons des bornes sur

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right), \quad (0.1)$$

avec, comme objectif, de développer des résultats adaptés au cadre des *événements rares*, où il sera supposé que la probabilité $\mu(A)$ de n'importe quel borélien de la classe $\mathcal{A}$ est très petite (typiquement, nous avons en tête une probabilité de l'ordre 10^{-2} ou 10^{-3}). Notons que dans le cadre de la classification introduit ci-dessus, les données se présentent sous la forme de couples $(X_i, Y_i)_{i=1}^n$ dans l'espace $\mathbb{R}^d \times \{0, 1\}$ et les ensembles de la classe $\mathcal{A}$ correspondent aux ensembles $\{(x, y) \in \mathbb{R}^d \times \{0, 1\} : g(x) \neq y\}$, avec $g \in \mathcal{G}$ (ces ensembles sont aussi appelés « errors sets » par Lugosi dans [15, p. 22]). L'hypothèse de rareté des événements A rencontrés se traduit également de manière directe en termes du risque pour le problème de la classification : pour n'importe quel $g \in \mathcal{G}$, le risque d'erreur $\mathbb{P}(g(X) \neq Y)$ est supposé très faible et nous cherchons, dans cette classe $\mathcal{G}$, la meilleure fonction de classification, en nous basant sur le risque empirique.

Afin de pouvoir développer cette théorie, un long chemin est nécessaire et nous en expliquons maintenant les principales étapes au travers d'une description des différents chapitres.

Le chapitre 1 s'intéresse au phénomène de concentration des sommes de variables aléatoires indépendantes bornées. Cela se translate directement au cas de la mesure empirique associée à un échantillon indépendant car $n\mu_n(A)$ (pour un borélien $A \in \mathcal{A}$ fixé) est, dans ce cas, une somme de variables aléatoires indépendantes à valeurs dans $[0, 1]$. Nous développons trois grandes inégalités classiques dans le domaine [6], toutes provenant d'une même méthode connue sous le nom de *méthode de Chernoff*. Ce phénomène de concentration s'observe également pour des fonctions de variables aléatoires indépen-

dantes qui ne sont pas forcément des sommes. Les *fonctions à différences finies* satisfont des propriétés similaires. Nous démontrons l'inégalité standard sur le sujet, connue sous le nom d'*inégalité des différences finies*. Nous terminons le chapitre par quelques extensions des résultats précédents, nécessaires pour notre étude du cas des événements rares.

Le chapitre 2 montre que, si $|\mathcal{A}| = +\infty$, l'étude des probabilités liées à la déviation maximale, comme dans (0.1), ne peut se faire sans une théorie additionnelle. Cette composante manquante est précisément la théorie combinatoire de Vladimir Naumovitch Vapnik (né en 1936) et d'Alexey Chervonenkis (1938–2014) [29]. Nous développons les concepts liés à cette théorie de manière très détaillée, notamment au travers d'exemples bien plus complets que ceux habituellement rencontrés dans la littérature (voir exemple 2.25). Nous montrons ensuite comment une telle théorie permet d'obtenir des inégalités intéressantes sur (0.1), même si la classe $\mathcal{A}$ est infinie. Cela nous a notamment permis, avec Johan Segers, d'améliorer l'inégalité classique du domaine (voir théorème 2.17). La question de la mesurabilité des objets rencontrés y est traitée avec grand soin, contrairement à la littérature classique (voir remarque 2.1). Des comparaisons graphiques sont également présentées afin d'aider le lecteur à percevoir l'impact des différentes techniques utilisées.

Le chapitre 3 est consacré à l'adaptation des résultats obtenus dans le deuxième chapitre lorsque nous nous plaçons dans le cas des événements rares, un cadre motivé par la *théorie des valeurs extrêmes*. De nombreuses inégalités sont démontrées et quasiment toutes sont basées sur un lemme de conditionnement, initialement proposé dans [13], sans démonstration. Nous en avons effectué la démonstration en détails et avons observé que, malgré une apparente simplicité, une grande richesse se cache dans celle-ci. Tous les résultats du chapitre sont principalement basés sur les travaux d'Anne Sabourin et de Johan Segers dans [25]. Néanmoins, nous avons proposé diverses améliorations et variations dans nos inégalités et espérons qu'à l'avenir, un article pourra naître d'une future collaboration entre les différents auteurs. Des comparaisons graphiques sont également proposées.

Dans chacun des chapitres, une grande attention est portée aux détails au sein des démonstrations. L'auteur est fermement convaincu que, par souci de longueur, trop de textes mathématiques sont imprécis sur de nombreux points, au détriment parfois de la compréhension du lecteur ou de sa capacité d'auto-apprentissage. Pour cette raison, tous les résultats présentés dans ce texte ont été rendus aussi explicites que possible et l'auteur espère que le lecteur ne se sentira alors pas découragé par la longueur plus conséquente de l'écrit.

Table des matières

Remerciements	5
Introduction	7
1 Inégalités classiques	13
1.1 Somme de variables aléatoires indépendantes	15
1.1.1 Inégalité de Hoeffding	17
1.1.2 Inégalité de Bennett	20
1.1.3 Inégalité de Bernstein	21
1.2 Fonctions à différences finies	23
1.2.1 Inégalité d'Efron-Stein	23
1.2.2 Inégalités des différences finies	27
1.2.3 Extensions	29
2 Mesures empiriques	33
2.1 Quelques quantités combinatoires	36
2.1.1 Coefficient de désintégration et dimension VC	36
2.1.2 Nombre couvrant	42
2.2 Concentration de mesures empiriques	43
2.2.1 Bornes provenant de la distribution	45
2.2.2 Bornes provenant de l'espérance	54
2.3 Comparaison des bornes	69
3 Mesures empiriques : le cas des évènements rares	73
3.1 Un argument de conditionnement	74
3.2 Bornes provenant de la distribution	77
3.3 Bornes provenant de l'espérance	82
3.4 Comparaison des bornes	86
Conclusion	89
Bibliographie	91

Chapitre 1

Inégalités classiques

Ce chapitre se base principalement sur [6, Chapitre 2, 3] et [15, Chapitre 1].

Partant du constat que l'inégalité de Bienaymé–Tchebychev (théorème 1.2) est trop faible pour décrire avec précision le phénomène de concentration de variables aléatoires autour de leur moyenne, nous montrons comment l'améliorer dans diverses situations. La section 1.1 traite des inégalités de concentration pour des sommes de variables aléatoires indépendantes, un cas qui recouvre de nombreuses applications (ainsi en témoigneront les chapitres suivants). La section 1.2 se veut un peu plus générale et présente certains résultats quand la fonction n'est plus une somme, mais n'est pas trop sensible à la modification d'un de ses arguments.

Une inégalité bien connue de la théorie des probabilités est l'inégalité de Markov (voir par exemple [26, Exercice 3.6.1]) que nous rappelons ici.

Théorème 1.1 (Inégalité de Markov). *Soit X une variable aléatoire positive définie sur un espace probabilisé $(\Omega, \mathcal{F}, P)$. Alors, pour tout $t > 0$,*

$$P(X \geq t) \leq \frac{E[X]}{t}.$$

Démonstration. Observons que pour tout $t > 0$, nous avons

$$t\mathbb{1}_{\{X \geq t\}} \leq X.$$

Prendre l'espérance de chaque coté permet de conclure. ■

Il est alors aisé de déduire la célèbre inégalité de Bienaymé–Tchebychev.

Théorème 1.2 (Inégalité de Bienaymé–Tchebychev). *Soit X une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathcal{F}, P)$, de variance finie. Alors, pour tout $t > 0$,*

$$P(|X - E[X]| \geq t) \leq \frac{\text{Var}[X]}{t^2}.$$

Démonstration. Il suffit d'utiliser l'inégalité de Markov comme suit :

$$P(|X - E[X]| \geq t) = P((X - E[X])^2 \geq t^2) \leq \frac{E[(X - E[X])^2]}{t^2} = \frac{\text{Var}[X]}{t^2}. \quad \blacksquare$$

Cette dernière inégalité fournit de l'information sur le comportement de X : elle assure que la probabilité de trouver des valeurs de X écartées de sa moyenne de plus de t est bornée par une certaine quantité en $O(t^{-2})$. Ce résultat a l'avantage d'être très général et fournit une borne simple à calculer (bien que dépendante de la distribution de X). Il représente souvent un moyen facile pour montrer certaines convergences en probabilité. Cependant, l'inégalité de Bienaymé–Tchebychev est souvent un résultat très faible comme le suggère l'exemple suivant.

Exemple 1.3 (Cas binomial). Soit $X := n^{-1} \sum_{i=1}^n X_i$, où les X_i sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) suivant une loi de Bernoulli de paramètre p , notée $Be(p)$, définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$. Alors, en utilisant l'indépendance, nous déduisons facilement la variance de X :

$$\text{Var}[X] = \frac{1}{n^2} n \text{Var}[X_1] = \frac{p(1-p)}{n}.$$

En appliquant directement l'inégalité de Bienaymé–Tchebychev, nous trouvons que pour tout $t > 0$,

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - p\right| \geq t\right) \leq \frac{p(1-p)}{nt^2}.$$

Or, nous savons que le théorème de la limite centrale garantit la convergence

$$P\left(\frac{\frac{1}{n} \sum_{i=1}^n X_i - p}{\sqrt{p(1-p)/n}} \geq t\right) \longrightarrow 1 - \Phi(t),$$

quand $n \rightarrow +\infty$, où Φ désigne la fonction de répartition d'une loi normale standard, i.e.

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-y^2/2} dy.$$

Observons que pour tout $x \in \mathbb{R}$, le théorème fondamental du calcul différentiel et intégral implique

$$\begin{aligned} \int_x^{+\infty} \frac{x}{\sqrt{2\pi}} e^{-y^2/2} dy &\leq \int_x^{+\infty} \frac{y}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &= \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \frac{d}{dy} (-e^{-y^2/2}) dy = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \end{aligned}$$

Ainsi,

$$1 - \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-y^2/2} dy \leq \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}. \quad (1.1)$$

Donc pour n suffisamment large, nous nous attendons à trouver une borne de type exponentielle, clairement bien plus forte que celle fournie par l'inégalité de Bienaymé–Tchebychev.

Remarque 1.4 (Espaces probabilisés). Jusque-là, nous avons toujours spécifié l'espace probabilisé sous-jacent aux variables aléatoires rencontrées. En particulier, il était sous-entendu que l'espérance d'une variable aléatoire X définie sur $(\Omega, \mathcal{F}, P)$ correspond simplement à l'intégrale de X par rapport à la mesure de probabilité P , i.e. $E[X] := \int_{\Omega} X(\omega) dP(\omega)$. Dans la suite, nous n'explicitons plus forcément cet espace et il sera toujours supposé que les variables aléatoires sont définies sur un même espace probabilisé, noté par défaut $(\Omega, \mathcal{F}, P)$, à moins que nous ne spécifions explicitement le contraire.

1.1 Somme de variables aléatoires indépendantes

L'idée principale pour arriver à obtenir des inégalités de concentration assez fines repose sur la *méthode de Chernoff* : nous remarquons que pour toute variable aléatoire réelle X et pour tout $\lambda > 0$, l'inégalité de Markov implique

$$P(X \geq t) = P(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{E[e^{\lambda X}]}{e^{\lambda t}} = \frac{m_X(\lambda)}{e^{\lambda t}},$$

où l'on a défini $m_X(\lambda) := E[e^{\lambda X}]$ la fonction génératrice des moments de X . Il est alors possible de reformuler cette quantité en termes de la fonction génératrice des cumulants, définie par $\psi_X(\lambda) := \log(m_X(\lambda))$, où $\log$ désigne, et désignera toujours dans la suite, le logarithme népérien :

$$\frac{m_X(\lambda)}{e^{\lambda t}} = \frac{e^{\psi_X(\lambda)}}{e^{\lambda t}} = e^{-(\lambda t - \psi_X(\lambda))}.$$

Comme $\lambda > 0$ n'est pas fixé, si l'on introduit la *transformée de Cramer* de X , définie par

$$\psi_X^*(t) = \sup_{\lambda > 0} (\lambda t - \psi_X(\lambda)),$$

nous obtenons la borne

$$P(X \geq t) \leq e^{-\psi_X^*(t)}. \quad (1.2)$$

Cette borne constitue la base des inégalités de concentration présentées dans cette section. Elle est initialement développée par Herman Chernoff dans [9, Théorème 1] pour étudier les queues de la loi binomiale. Avant de continuer, quelques remarques s'imposent :

- Les fonctions génératrices des moments et des cumulants n'existent pas toujours. Cependant, dans les cas qui nous intéresseront, nous traiterons souvent de variables

aléatoires bornées et l'existence est alors vérifiée. L'inégalité (1.2) assure alors que la queue de leur distribution décroît exponentiellement vers zéro.

- En appliquant l'inégalité de Hölder, nous observons facilement que la fonction génératrice des cumulants est une fonction convexe. Ainsi, la transformée de Cramer se compare immédiatement à la transformée de Fenchel–Legendre d'une fonction convexe (excepté que le supremum est pris sur $(0, +\infty)$ et non $\mathbb{R}$ tout entier).

Pour se familiariser avec la méthode, considérons un premier exemple simple.

Exemple 1.5 (Cas normal). Soit X une variable aléatoire suivant une loi normale de moyenne 0 et de variance σ^2 , i.e. $X \sim N(0, \sigma^2)$. Alors, il est bien connu que sa fonction génératrice des moments est donnée par

$$m_X(\lambda) = e^{\lambda^2 \sigma^2 / 2}.$$

Nous obtenons donc la fonction génératrice des cumulants $\psi(\lambda) = \lambda^2 \sigma^2 / 2$. Il s'agit clairement d'une fonction convexe et différentiable. Ainsi, la transformée de Cramer se calcule très simplement par différentiation et vaut

$$\psi_X^*(t) = t^2 / 2\sigma^2.$$

En conséquence, la méthode de Chernoff fournit la borne

$$P(X \geq t) \leq \exp(-t^2 / 2\sigma^2).$$

Comme attendu, la loi normale possède des queues exponentielles.

Nous souhaitons maintenant appliquer la méthode de Chernoff à des sommes de variables aléatoires prenant des valeurs dans des intervalles bornés. Cela permet un contrôle de la variance, comme montré dans le lemme suivant.

Lemme 1.6. *Soit X une variable aléatoire réelle à valeurs dans l'intervalle $[a, b]$ presque sûrement (p.s.). Alors*

$$\text{Var}[X] \leq \frac{(b-a)^2}{4}.$$

Démonstration. Rappelons que la variance est solution du problème de minimisation suivant

$$\text{Var}[X] = \min_{s \in \mathbb{R}} E[(X - s)^2],$$

où le minimum est atteint pour $s = E[X]$ (voir par exemple [26], p. 41). Si nous choisissons $s = (a + b)/2$, le milieu de l'intervalle, nous obtenons donc

$$\text{Var}[X] \leq E \left[\left(X - \frac{a+b}{2} \right)^2 \right].$$

Comme $X \in [a, b]$ p.s., avec probabilité 1 il ne peut être écarté du milieu de l'intervalle de plus d'une demi-longueur, ainsi nous avons p.s.

$$\left(X - \frac{a+b}{2}\right)^2 \leq \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{4}.$$

En combinant cette inégalité à la précédente, nous obtenons le résultat. ■

Comme nous le verrons plus loin, ce contrôle de la variance peut se transposer en un contrôle de la fonction génératrice des cumulants, qui, à son tour, permet de minorer la transformée de Cramer.

Nous énonçons et démontrons trois grandes inégalités de concentration pour des sommes de variables aléatoires bornées.

1.1.1 Inégalité de Hoeffding

Nous aurons besoin d'un lemme démontré par Wassily Hoeffding dans [20] (en particulier il s'agit de l'inégalité (4.16) dans son papier). Nous reformulons son résultat en termes de la fonction génératrice des cumulants afin de pouvoir directement fournir une borne inférieure sur la transformée de Cramer.

Lemme 1.7 (Lemme de Hoeffding). *Soit X une variable aléatoire réelle de moyenne nulle, à valeurs dans $[a, b]$ p.s. Alors, pour tout $\lambda > 0$,*

$$\psi_X(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}.$$

Démonstration. Nous utilisons sans démonstration les faits élémentaires suivants

$$\begin{cases} \psi_X(0) &= 0 \\ (\psi_X)'(0) &= E[X] = 0. \end{cases}$$

Par le théorème de Taylor et la formule du reste de Lagrange, pour tout $\lambda > 0$, il existe $\theta \in (0, \lambda)$ tel que

$$\psi_X(\lambda) = \psi_X(0) + (\psi_X)'(0)\lambda + \frac{1}{2}(\psi_X)''(\theta)\lambda^2 = \frac{1}{2}(\psi_X)''(\theta)\lambda^2.$$

Pour conclure, il suffit d'exprimer $(\psi_X)''(\theta)$ comme la variance d'une variable aléatoire Z réelle gardant p.s. ses valeurs dans $[a, b]$ et d'utiliser le lemme 1.6. Un calcul simple montre que (il faut échanger dérivée et intégrale mais cela se justifie dans ce cadre)

$$(\psi_X)''(\theta) = \frac{E[X^2 e^{\theta X}] E[e^{\theta X}] - (E[X e^{\theta X}])^2}{(E[e^{\theta X}])^2} = \frac{E[X^2 e^{\theta X}]}{E[e^{\theta X}]} - \left(\frac{E[X e^{\theta X}]}{E[e^{\theta X}]} \right)^2.$$

Nous prenons Z tel que pour toute fonction mesurable $f : \mathbb{R} \rightarrow \mathbb{R}$

$$\mathbb{E}[f(Z)] = \frac{\mathbb{E}[f(X)e^{\theta X}]}{\mathbb{E}[e^{\theta X}]}.$$

Cette construction est connue sous le nom de *changement de mesure exponentiel* : l'idée est de modifier la distribution de X par un facteur exponentiel et de normaliser. Cela se voit en prenant $f = \mathbb{1}_A$ pour A un borélien de $\mathbb{R}$:

$$\mathbb{P}(Z \in A) = \mathbb{E}[\mathbb{1}\{Z \in A\}] = \frac{\mathbb{E}[\mathbb{1}\{X \in A\}e^{\theta X}]}{\mathbb{E}[e^{\theta X}]}.$$

Remarquons que la variable aléatoire Z garde ses valeurs dans $[a, b]$ p.s.

$$\mathbb{P}(a \leq Z \leq b) = \mathbb{E}[\mathbb{1}\{a \leq Z \leq b\}] = \frac{\mathbb{E}[\mathbb{1}\{a \leq X \leq b\}e^{\theta X}]}{\mathbb{E}[e^{\theta X}]} = 1,$$

car $X \in [a, b]$ p.s. Le lemme 1.6 permet alors de conclure car

$$(\psi_X)''(\theta) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 = \text{Var}[Z] \leq \frac{(b-a)^2}{4}. \quad \blacksquare$$

Théorème 1.8 (Inégalité de Hoeffding). *Soient $X_1, \dots, X_n$ des variables aléatoires réelles indépendantes telles que $X_i \in [a_i, b_i]$ p.s. pour chaque $i \in \{1, \dots, n\}$. Si nous notons $S_n := \sum_{i=1}^n X_i$ leur somme, nous avons les inégalités suivantes pour tout $t > 0$:*

$$\begin{cases} \mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) & \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\ \mathbb{P}(S_n - \mathbb{E}[S_n] \leq -t) & \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \end{cases}$$

Une conséquence immédiate du théorème 1.8 est l'inégalité

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2 \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \quad (1.3)$$

qui est elle-même souvent désignée par l'inégalité de Hoeffding.

Observons que la borne fournie par le théorème 1.8 est identique à celle du cas d'une normale centrée (voir exemple 1.5) où la variance est donnée par $\sigma^2 = (1/4) \sum_{i=1}^n (b_i - a_i)^2$. Cela va dans le sens attendu car nous utiliserons le lemme de Hoeffding (lemme 1.7) dans la démonstration de ce théorème, qui lui-même est basé sur le lemme 1.6 qui garantit que

$$\text{Var}[S_n] = \sum_{i=1}^n \text{Var}[X_i] \leq \frac{1}{4} \sum_{i=1}^n (b_i - a_i)^2.$$

Avant de commencer la démonstration, rappelons que l'intérêt de travailler avec les fonctions génératrices de moments et de cumulants pour des sommes de variables aléatoires indépendantes est le suivant : en utilisant les mêmes notations que dans le théorème 1.8, pour tout $\lambda > 0$

$$\begin{cases} m_{S_n}(\lambda) &= \prod_{i=1}^n m_{X_i}(\lambda), \\ \psi_{S_n}(\lambda) &= \sum_{i=1}^n \psi_{X_i}(\lambda). \end{cases} \quad (1.4)$$

Démonstration du théorème 1.8. Pour la démonstration, nous supposons que les X_i sont tous de moyenne nulle. Cela évite de devoir définir de nouvelles variables $Y_i = X_i - \mathbb{E}[X_i]$ et de changer les intervalles $[a_i, b_i]$. Comme la borne obtenue ne fait intervenir que la longueur des intervalles, cela ne constitue pas une perte de généralité. Par (1.4) et le lemme de Hoeffding,

$$\psi_{S_n}(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda) \leq \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2.$$

Comme expliqué auparavant, cela fournit une borne inférieure sur la transformée de Cramer

$$\psi_{S_n}^*(t) = \sup_{\lambda > 0} (\lambda t - \psi_{S_n}(\lambda)) \geq \sup_{\lambda > 0} \left\{ \lambda t - \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2 \right\}.$$

Par différenciation, nous trouvons que le supremum est atteint pour $\lambda = 4t / \sum_{i=1}^n (b_i - a_i)^2$. Ainsi,

$$\sup_{\lambda > 0} \left\{ \lambda t - \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2 \right\} = \frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2},$$

et la méthode de Chernoff fournit la borne

$$\mathbb{P}(S_n \geq t) \leq e^{-\psi_{S_n}^*(t)} \leq \exp \left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right),$$

qui correspond à la première inégalité du théorème 1.8 dans le cas où les X_i sont centrés. L'autre inégalité est immédiate en appliquant la première sur les variables aléatoires $Y_i := -X_i$. ■

Si nous appliquons l'inégalité de Hoeffding dans le cas binomial de l'exemple 1.3, il faut prendre $[a_i, b_i] = [0, 1]$ et $\mathbb{E}[X_i] = p$, pour tout $i \in \{1, \dots, n\}$. Il vient alors

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| \geq t \right) = \mathbb{P} \left(\left| \sum_{i=1}^n X_i - np \right| \geq nt \right) \leq 2e^{-2n^2 t^2 / n} = 2e^{-2nt^2},$$

qui est un résultat bien meilleur que celui fournit par l'inégalité de Bienaymé–Tchebychev. Chernoff fournissait déjà cette borne pour la binomiale dans [9], et en ce sens, le théorème 1.8 étend son résultat.

La borne apparaissant dans l'inégalité de Hoeffding ne dépend pas de la distribution des X_i et est donc particulièrement simple à calculer ; cependant, cela peut représenter un désavantage. L'estimation de la variance utilisée (lemme 1.6) peut être trop grossière, et nous aimerions obtenir également des bornes exponentielles faisant intervenir directement la variance des X_i .

1.1.2 Inégalité de Bennett

Nous commençons par démontrer une inégalité due à George Bennett [4].

Lemme 1.9. *Soit X une variable aléatoire réelle telle que $|X| \leq c$ p.s. pour $c > 0$. Alors pour tout $\lambda > 0$,*

$$\begin{aligned}\psi_X(\lambda) &\leq \log \left(1 + \lambda \mathbb{E}[X] + \mathbb{E}[X^2] \frac{\phi(c\lambda)}{c^2} \right) \\ &\leq \lambda \mathbb{E}[X] + \mathbb{E}[X^2] \frac{\phi(c\lambda)}{c^2},\end{aligned}$$

où ϕ est définie sur $\mathbb{R}$ par $\phi(u) = e^u - u - 1$.

Démonstration. Commençons par observer que la fonction $u \in \mathbb{R} \rightarrow u^{-2}\phi(u)$ est croissante. Cela se remarque directement en regardant son graphe, ou en calculant sa dérivée. Cette fonction est bien définie en zéro par prolongement continu car $\phi(u)$ correspond à l'exponentielle moins les deux premiers termes de son développement de Taylor en zéro et se comporte donc en $O(u^2)$ lorsque u est proche de zéro.

Dès lors, pour $\lambda > 0$

$$\frac{e^{\lambda X} - \lambda X - 1}{X^2} \leq \frac{e^{\lambda c} - \lambda c - 1}{c^2}.$$

Ainsi, en réarrangeant les termes et en prenant l'espérance, il vient

$$m_X(\lambda) \leq 1 + \lambda \mathbb{E}[X] + \frac{\phi(\lambda c)}{c^2} \mathbb{E}[X^2].$$

En passant au logarithme, nous trouvons la première inégalité. La seconde vient du fait que $\log(1 + u) \leq u$. ■

Théorème 1.10 (Inégalité de Bennett). *Soient $X_1, \dots, X_n$ des variables aléatoires réelles indépendantes telles que $|X_i| \leq c$ p.s. pour chaque $i \in \{1, \dots, n\}$, où $c > 0$. Si nous notons $S_n := \sum_{i=1}^n X_i$ leur somme et $\sigma^2 := (1/n) \sum_{i=1}^n \text{Var}[X_i]$ la somme normalisée des variances, nous avons l'inégalité suivante pour tout $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp \left(-\frac{n\sigma^2}{c^2} h \left(\frac{ct}{n\sigma^2} \right) \right),$$

où h est définie pour $u \geq 0$ par $h(u) = (1 + u) \log(1 + u) - u$.

Démonstration. Nous effectuons la démonstration dans le cas de variables centrées, $E[X_i] = 0$, pour réduire les notations. Cela ne constitue pas une perte de généralité car on peut toujours centrer les X_i sans changer la constante c .

Par (1.4) et par le lemme précédent, pour tout $\lambda > 0$,

$$\psi_{S_n}(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda) \leq \frac{\phi(c\lambda)}{c^2} \sum_{i=1}^n E[X_i^2] = \frac{\phi(c\lambda)}{c^2} n\sigma^2,$$

où ϕ est définie comme dans le lemme 1.9. Ayant une borne supérieure sur la fonction génératrice des cumulants, nous déduisons une borne inférieure pour la transformée de Cramer

$$\psi_{S_n}^*(t) = \sup_{\lambda > 0} (\lambda t - \psi_{S_n}(\lambda)) \geq \sup_{\lambda > 0} \left\{ \lambda t - \frac{\phi(c\lambda)}{c^2} n\sigma^2 \right\} = \frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right),$$

où le supremum est atteint pour $\lambda = \log(1 + ct/n\sigma^2)/c$. La méthode de Chernoff fournit alors la borne

$$P(S_n \geq t) \leq e^{-\psi_{S_n}^*(t)} \leq \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right)\right),$$

ce qui correspond au résultat pour des variables X_i centrées. ■

Malgré le fait que la borne obtenue soit plus précise que l'inégalité de Hoeffding, elle est assez compliquée à utiliser en pratique et nous lui préférons souvent l'inégalité de Bernstein présentée ci-après.

1.1.3 Inégalité de Bernstein

Le résultat suivant est attribué à Sergei Natanovich Bernstein dans [5]. Comme l'explique George Bennett dans [4], l'article original n'a pas été traduit et est quasiment impossible à consulter. Nous le citons uniquement par souci de complétude.

Théorème 1.11 (Inégalité de Bernstein). *Soient $X_1, \dots, X_n$ des variables aléatoires réelles indépendantes telles que $|X_i| \leq c$ p.s. pour chaque $i \in \{1, \dots, n\}$, où $c > 0$. Si nous notons $S_n := \sum_{i=1}^n X_i$ leur somme et $\sigma^2 := (1/n) \sum_{i=1}^n \text{Var}[X_i]$ la somme normalisée des variances, nous avons l'inégalité suivante pour tout $t > 0$,*

$$P(S_n - E[S_n] \geq t) \leq \exp\left(-\frac{t^2}{2(n\sigma^2 + ct/3)}\right).$$

Démonstration. Le résultat découle immédiatement de l'inégalité de Bennett et de l'inégalité

$$h(u) \geq \frac{u^2}{2(1+u/3)}, \quad (1.5)$$

où h est définie comme au théorème 1.10 et $u \geq 0$.

Pour montrer (1.5), nous remarquons pour tout $u \geq 0$,

$$h'(u) = \log(1+u) = \int_0^u \frac{1}{1+t} dt.$$

Dès lors, en changeant l'ordre d'intégration, il vient

$$h(u) = \int_0^u h'(s) ds = \int_0^u \int_0^s \frac{1}{1+t} dt ds = \int_0^u \int_t^u \frac{1}{1+t} ds dt = \int_0^u \frac{(u-t)}{1+t} dt.$$

Supposons $u > 0$ (le cas $u = 0$ est immédiat). Il résulte du précédent calcul que nous devons prouver l'inégalité

$$\int_0^u \frac{1}{1+t} \frac{2(u-t)}{u^2} dt \geq \frac{1}{1+u/3}. \quad (1.6)$$

Un calcul simple montre que, peu importe $u > 0$, la fonction

$$t \in \mathbb{R} \mapsto f(t) := \frac{2(u-t)}{u^2} \mathbb{1}_{[0,u]}(t) \geq 0$$

est une densité. Dès lors, l'inégalité (1.6) peut se réécrire

$$\mathbb{E} \left[\frac{1}{1+T} \right] \geq \frac{1}{1+\mathbb{E}[T]},$$

pour T une variable aléatoire de densité f (en effet, $\mathbb{E}[T] = u/3$ pour une telle variable aléatoire). La fonction $x \in [0, +\infty) \mapsto 1/(1+x)$ étant convexe, le résultat suit de l'inégalité de Jensen. ■

Historiquement, il est important de remarquer que ce résultat, ici présenté comme une simple conséquence de l'inégalité de Bennett, a été démontré bien avant cette dernière.

L'inégalité de Bernstein est évidemment un peu moins forte que la précédente, cependant, la borne obtenue est plus simple à manipuler et fournit un contrôle souvent suffisant. De plus, si nous supposons que tous les X_i proviennent de la même distribution et que $t = o(n)$, la borne est de l'ordre

$$e^{-(1+o(1)) \frac{t^2}{2\sigma^2}}, \quad \text{quand } n \rightarrow +\infty.$$

Cela correspond à l'ordre attendu au regard du théorème de la limite centrale qui garantit qu'asymptotiquement, la variable aléatoire $S_n - \mathbb{E}[S_n]$ se comporte comme une normale $N(0, \sigma^2)$.

Observons que dans le cas de l'exemple 1.3, un contexte qui reviendra souvent dans la suite, l'inégalité de Bernstein devient

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - p \geq t\right) = \mathbb{P}\left(\sum_{i=1}^n X_i - np \geq nt\right) \leq \exp\left(-\frac{nt^2}{2(p(1-p) + t/3)}\right),$$

avec $c = 1$ et $\sigma^2 = p(1-p)$.

1.2 Fonctions à différences finies

Dans la section précédente, nous avons vu comment, grâce à la méthode de Chernoff et à des bornes sur la fonction génératrice des cumulents, il est possible de fournir des bornes exponentielles sur la probabilité qu'une somme de variables aléatoires indépendantes s'écarte de sa moyenne. L'idée de cette deuxième section est de généraliser ces résultats, sous certaines hypothèses, pour des fonctions mesurables de variables aléatoires indépendantes qui ne sont pas forcément des sommes.

1.2.1 Inégalité d'Efron-Stein

Nous considérons une fonction mesurable $f : \mathcal{X}^n \rightarrow \mathbb{R}$, où $\mathcal{X}$ est un espace mesurable quelconque, et $X_1, \dots, X_n$ des variables aléatoires indépendantes à valeurs dans $\mathcal{X}$. Nous pouvons alors définir la variable aléatoire réelle

$$Z := f(X_1, \dots, X_n).$$

Clairement, si les X_i étaient des variables aléatoires réelles et $f(x_1, \dots, x_n) := \sum_{i=1}^n x_i$, nous aurions que

$$\text{Var}[Z] = \sum_{i=1}^n \text{Var}[X_i],$$

et nous retombons dans les cas traités dans la première section. Afin de généraliser ces résultats, nous observons que même si Z n'est pas une somme de variables aléatoires réelles, nous pouvons borner sa variance de la manière suivante. L'origine de ce résultat remonte au papier [12] d'Efron et Stein.

Théorème 1.12 (Inégalité d'Efron–Stein). *Soit Z définie comme précédemment. Si*

$$\mathbb{E}^{(i)}[Z] := \mathbb{E}[Z | X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n]$$

désigne l'espérance conditionnelle de Z sachant $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, alors

$$\text{Var}[Z] \leq \sum_{i=1}^n \mathbb{E} \left[(Z - \mathbb{E}^{(i)}[Z])^2 \right].$$

De plus, si $X'_1, \dots, X'_n$ est une copie indépendante de $X_1, \dots, X_n$, et si

$$Z'_i := f(X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n),$$

alors

$$\sum_{i=1}^n \mathbb{E} \left[(Z - \mathbb{E}^{(i)}[Z])^2 \right] = \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[(Z - Z'_i)^2 \right].$$

Ce résultat suggère de s'intéresser à une classe particulière de fonctions si l'on souhaite obtenir une borne uniforme sur la variance de Z .

Définition 1.13 (Fonction à différences finies). Soit $f : \mathcal{X}^n \rightarrow \mathbb{R}$. Nous disons que f est une fonction à différences finies s'il existe des constantes $c_1, \dots, c_n \geq 0$ telles que pour tout $i \in \{1, \dots, n\}$,

$$\sup_{\substack{x_1, \dots, x_n \in \mathcal{X} \\ x'_i \in \mathcal{X}}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i.$$

Dès lors, si $Z := f(X_1, \dots, X_n)$, où f est une fonction à différences finies, le théorème 1.12 assure que

$$\text{Var}[Z] \leq \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[(Z - Z'_i)^2 \right] \leq \frac{1}{2} \sum_{i=1}^n c_i^2.$$

La démonstration de l'inégalité d'Efron–Stein repose sur un lemme qui garantit que la variance de Z peut être réécrite comme une somme de variances d'incrément d'une martingale de Doob. (Si ce vocabulaire n'est pas familier au lecteur, il pourra consulter [18], bien que cela ne soit pas nécessaire pour la suite.)

Lemme 1.14. Soit Z définie comme précédemment. Si nous notons

$$\mathbb{E}_i[Z] := \mathbb{E}[Z | X_1, \dots, X_i],$$

l'espérance conditionnelle de Z sachant $X_1, \dots, X_i$ (où $\mathbb{E}_0[Z] := \mathbb{E}[Z]$) et si

$$\Delta_i := \mathbb{E}_i[Z] - \mathbb{E}_{i-1}[Z],$$

désigne les incréments de la martingale $(\mathbb{E}_i[Z])_{i=0, \dots, n}$ associée à Z , alors

$$\text{Var}[Z] = \sum_{i=1}^n \mathbb{E}[\Delta_i^2] = \sum_{i=1}^n \text{Var}[\Delta_i].$$

La suite de variables aléatoires $(Y_i := E_i[Z])_{i=0,\dots,n}$ est effectivement une martingale de Doob par rapport à la filtration $\mathcal{F}_i := \sigma(X_1, \dots, X_i)$, où $\sigma(X_1, \dots, X_i)$ désigne la tribu engendrée par les variables aléatoires $X_1, \dots, X_i$, comme définie dans [26], p.60. En effet, pour tout $i \in \{1, \dots, n-1\}$,

$$\begin{aligned} E[Y_{i+1}|\mathcal{F}_i] &= E[E_{i+1}[Z]|X_1, \dots, X_i] = E[E[Z|X_1, \dots, X_{i+1}]|X_1, \dots, X_i] \\ &= E[Z|X_1, \dots, X_i] = E_i[Z] = Y_i, \end{aligned}$$

où, pour le passage à la deuxième ligne, nous avons utilisé [18, Propriété 3], p.16.

Bien que ce formalisme ne soit pas essentiel dans notre cadre, il est important à mentionner car les résultats que nous présenterons dans cette section sont également disponibles en théorie des martingales avec des incréments bornés (voir l'*inégalité d'Azuma*, [17, Théorème 3.10], nommée selon le mathématicien japonais Kazuoki Azuma (né en 1939), qui généralise largement l'inégalité de concentration que nous exposerons ici).

Démonstration du lemme 1.14. Observons que $E_n[Z] = E[Z|X_1, \dots, X_n] = Z$ et rappelons que nous avons défini $E_0[Z] = E[Z]$. Ainsi, nous pouvons réécrire l'écart entre Z et sa moyenne de la manière suivante :

$$Z - E[Z] = \sum_{i=1}^n (E_i[Z] - E_{i-1}[Z]) = \sum_{i=1}^n \Delta_i.$$

Dès lors,

$$\text{Var}[Z] = E[(Z - E[Z])^2] = E[(\Delta_1 + \dots + \Delta_n)^2] = \sum_{i=1}^n E[\Delta_i^2] + 2 \sum_{i < j} E[\Delta_i \Delta_j].$$

Pour montrer le résultat annoncé, il suffit donc de montrer que l'espérance des termes croisés de la somme est nulle. Cela se fait aisément de la manière suivante : remarquons que pour tout $i < j$,

$$E_i[\Delta_j] = E_i[E_j[Z] - E_{j-1}[Z]] = E_i[Z] - E_i[Z] = 0.$$

Donc, pour tout $i < j$,

$$E[\Delta_i \Delta_j] = E[E_i[\Delta_i \Delta_j]] = E[\Delta_i E_i[\Delta_j]] = 0,$$

ce qui conclut la démonstration de la première égalité. La seconde suit du fait que pour tout $i \in \{1, \dots, n\}$, nous avons $E[\Delta_i] = 0$. ■

Démonstration du théorème 1.12. Par le lemme 1.14,

$$\text{Var}[Z] = \sum_{i=1}^n E[\Delta_i^2] = \sum_{i=1}^n E[(E_i[Z] - E_{i-1}[Z])^2].$$

Si pour chaque $i \in \{1, \dots, n\}$, nous notons μ_i la distribution sur $\mathcal{X}$ associée à X_i , alors le théorème de Fubini assure que

$$\begin{aligned} \mathbb{E}_i \left[\mathbb{E}^{(i)}[Z] \right] &= \int_{\mathcal{X}^{n-i}} \left(\int_{\mathcal{X}} f(X_1, \dots, X_{i-1}, x_i, \dots, x_n) d\mu_i(x_i) \right) d\mu_{i+1}(x_{i+1}) \dots d\mu_n(x_n) \\ &= \int_{\mathcal{X}^{n-(i-1)}} f(X_1, \dots, X_{i-1}, x_i, \dots, x_n) d\mu_i(x_i) \dots d\mu_n(x_n) = \mathbb{E}_{i-1}[Z]. \end{aligned}$$

Ainsi, par l'inégalité de Jensen pour les espérances conditionnelles,

$$\begin{aligned} \text{Var}[Z] &= \sum_{i=1}^n \mathbb{E} \left[(\mathbb{E}_i[Z] - \mathbb{E}_{i-1}[Z])^2 \right] = \sum_{i=1}^n \mathbb{E} \left[\left(\mathbb{E}_i \left[Z - \mathbb{E}^{(i)}[Z] \right] \right)^2 \right] \\ &\leq \sum_{i=1}^n \mathbb{E} \left[\mathbb{E}_i \left[\left(Z - \mathbb{E}^{(i)}[Z] \right)^2 \right] \right] = \sum_{i=1}^n \mathbb{E} \left[\left(Z - \mathbb{E}^{(i)}[Z] \right)^2 \right]. \end{aligned}$$

Pour montrer l'autre partie du théorème, nous remarquons que, conditionnellement à $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, les variables aléatoires Z et Z'_i sont indépendantes et identiquement distribuées. Nous observons également que si X et Y sont deux variables aléatoires i.i.d. à valeurs dans $\mathcal{X}$, alors

$$\text{Var}[X] = \frac{1}{2} \mathbb{E} \left[(X - Y)^2 \right].$$

En effet, partant du membre de droite, comme $\mathbb{E}[X - Y] = 0$,

$$\frac{1}{2} \mathbb{E} \left[(X - Y)^2 \right] = \frac{1}{2} \text{Var}[X - Y] = \frac{1}{2} (\text{Var}[X] + \text{Var}[Y] - 2 \text{Cov}[X, Y]) = \text{Var}[X],$$

où la dernière inégalité vient directement du fait que X et Y sont i.i.d. Par conséquent, en travaillant conditionnellement à $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, nous obtenons, par le raisonnement qui précède,

$$\mathbb{E}^{(i)} \left[(Z - \mathbb{E}^{(i)}[Z])^2 \right] =: \text{Var}^{(i)}[Z] = \frac{1}{2} \mathbb{E}^{(i)} \left[(Z - Z'_i)^2 \right].$$

Et finalement, il vient

$$\begin{aligned} \sum_{i=1}^n \mathbb{E} \left[(Z - \mathbb{E}^{(i)}[Z])^2 \right] &= \sum_{i=1}^n \mathbb{E} \left[\mathbb{E}^{(i)} \left[(Z - \mathbb{E}^{(i)}[Z])^2 \right] \right] = \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[\mathbb{E}^{(i)} \left[(Z - Z'_i)^2 \right] \right] \\ &= \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[(Z - Z'_i)^2 \right]. \quad \blacksquare \end{aligned}$$

Il est intéressant de noter que si les X_i sont toutes des variables aléatoires réelles et si $Z = f(X_1, \dots, X_n) := \sum_{i=1}^n X_i$, alors la borne sur la variance proposée par l'inégalité d'Efron–Stein devient une égalité et est donc optimal dans ce sens.

1.2.2 Inégalités des différences finies

Si f est une fonction à différences finies, nous pouvons fournir, au-delà d'un contrôle sur la variance, des inégalités de concentration exponentielles pour Z . Commençons par une extension de l'inégalité de Hoeffding.

Théorème 1.15 (Inégalité des différences finies). *Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes à valeurs dans $\mathcal{X}$ et $Z := f(X_1, \dots, X_n)$, où $f : \mathcal{X}^n \rightarrow \mathbb{R}$ est une fonction à différences finies. Alors pour tout $t > 0$,*

$$\begin{cases} \mathbb{P}(Z - \mathbb{E}[Z] \geq t) & \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right), \\ \mathbb{P}(Z - \mathbb{E}[Z] \leq -t) & \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right). \end{cases}$$

L'inégalité des différences finies généralise l'inégalité de Hoeffding car, sous les mêmes hypothèses que dans le théorème 1.8, nous pouvons prendre $c_i = b_i - a_i$ et nous retrouvons la même borne. Dès lors, il n'est pas étonnant que la démonstration du théorème 1.12 soit très similaire à celle de l'inégalité de Hoeffding. Nous commençons par un lemme qui correspond au lemme de Hoeffding (lemme 1.7) pour les espérances conditionnelles.

Lemme 1.16. *Soient X et Y deux variables aléatoires, où X est supposée à valeurs dans $\mathbb{R}$. Si $\mathbb{E}[X|Y] = 0$ p.s. et s'il existe une fonction h et une constante $c \geq 0$ telles que*

$$h(Y) \leq X \leq h(Y) + c,$$

alors pour tout $\lambda > 0$,

$$\mathbb{E}\left[e^{\lambda X} | Y\right] \leq \exp\left(\frac{\lambda^2 c^2}{8}\right).$$

L'idée est d'appliquer ce lemme avec $X := \Delta_i$ et $Y := (X_1, \dots, X_{i-1})$ pour chaque $i \in \{1, \dots, n\}$. La démonstration du lemme 1.14 a en effet révélé que pour tout $i < j$, nous avons $\mathbb{E}_i[\Delta_j] = 0$. Ces choix pour X et Y vérifient donc bien $\mathbb{E}[X|Y] = \mathbb{E}_{i-1}[\Delta_i] = 0$. En utilisant la décomposition $Z - \mathbb{E}[Z] = \sum_{i=1}^n \Delta_i$, cela permettra de fournir une borne sur la fonction génératrice des cumulants de $Z - \mathbb{E}[Z]$.

Démonstration du lemme 1.16. Remarquons que nous pouvons réécrire le lemme de Hoeffding (lemme 1.7) de la manière suivante : si X est une variable aléatoire de moyenne nulle à valeurs dans $[a, b]$, alors

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Or, il suit directement des hypothèses choisies ici que $X|Y$ est de moyenne nulle p.s. et est à valeurs dans un intervalle de longueur c . La conclusion suit donc du lemme de Hoeffding. ■

Démonstration du théorème 1.15. Rappelons que

$$\Delta_i = E_i[Z] - E_{i-1}[Z] =: g(X_1, \dots, X_i).$$

Si nous définissons $h(X_1, \dots, X_{i-1}) := \inf_{x \in \mathcal{X}} g(X_1, \dots, X_{i-1}, x)$, il est alors clair que

$$h(X_1, \dots, X_{i-1}) \leq \Delta_i.$$

Comme f est supposée à différences finies, nous avons également que

$$\sup_{x \in \mathcal{X}} g(X_1, \dots, X_{i-1}, x) \leq \inf_{x \in \mathcal{X}} g(X_1, \dots, X_{i-1}, x) + c_i,$$

pour une certaine constante $c_i \geq 0$. En effet, g ne dépend de X_i que via $E_i[Z]$ qui, clairement, est également une fonction à différences finies de $X_1, \dots, X_i$ avec les mêmes constantes c_i que f . Dès lors,

$$\Delta_i = g(X_1, \dots, X_i) \leq \sup_{x \in \mathcal{X}} g(X_1, \dots, X_{i-1}, x) \leq h(X_1, \dots, X_{i-1}) + c_i.$$

Par le lemme précédent, avec X et Y choisis comme annoncé, il vient

$$E_{i-1} [e^{\lambda \Delta_i}] \leq \exp \left(\frac{\lambda^2 c_i^2}{8} \right),$$

pour chaque $i \in \{1, \dots, n\}$ et pour tout $\lambda > 0$. En utilisant la décomposition $Z - E[Z] = \sum_{i=1}^n \Delta_i$ et en conditionnant, il est alors simple de montrer par récurrence que pour tout $\lambda > 0$,

$$\psi_{Z-E[Z]}(\lambda) \leq \frac{\lambda^2 \sum_{i=1}^n c_i^2}{8}.$$

En effet, pour $n = 1$, nous avons

$$\psi_{Z-E[Z]}(\lambda) = \log \left(E \left[e^{\lambda(Z-E[Z])} \right] \right) = \log \left(E \left[e^{\lambda \Delta_1} \right] \right) = \log \left(E_0 \left[e^{\lambda \Delta_1} \right] \right) \leq \frac{\lambda^2 c_1^2}{8}$$

Pour $n > 1$, si nous supposons que le cas $n - 1$ vérifié, il suffit de conditionner comme suit :

$$\begin{aligned} \psi_{Z-E[Z]}(\lambda) &= \log \left(E \left[e^{\lambda(Z-E[Z])} \right] \right) = \log \left(E \left[e^{\lambda \sum_{i=1}^n \Delta_i} \right] \right) = \log \left(E \left[\prod_{i=1}^n e^{\lambda \Delta_i} \right] \right) \\ &= \log \left(E \left[E_{n-1} \left[\prod_{i=1}^n e^{\lambda \Delta_i} \right] \right] \right) = \log \left(E \left[\prod_{i=1}^{n-1} e^{\lambda \Delta_i} E_{n-1} \left[e^{\lambda \Delta_n} \right] \right] \right) \end{aligned}$$

$$\begin{aligned}
&\leq \log \left(\exp \left(\frac{\lambda^2 c_n^2}{8} \right) \mathbb{E} \left[\prod_{i=1}^{n-1} e^{\lambda \Delta_i} \right] \right) = \frac{\lambda^2 c_n^2}{8} + \log \left(\mathbb{E} \left[\prod_{i=1}^{n-1} e^{\lambda \Delta_i} \right] \right) \\
&\leq \frac{\lambda^2 c_n^2}{8} + \frac{\lambda^2 \sum_{i=1}^{n-1} c_i^2}{8} = \frac{\lambda^2 \sum_{i=1}^n c_i^2}{8}.
\end{aligned}$$

Dès lors, comme dans le cas des sommes de variables indépendantes, nous obtenons une borne inférieure sur la transformée de Cramer,

$$\psi_{Z-\mathbb{E}[Z]}^*(t) = \sup_{\lambda>0} (\lambda t - \psi_{Z-\mathbb{E}[Z]}(\lambda)) \geq \sup_{\lambda>0} \left(\lambda t - \frac{\lambda^2 \sum_{i=1}^n c_i^2}{8} \right).$$

Nous trouvons que le supremum est atteint pour $\lambda = 4t / \sum_{i=1}^n c_i^2$. Ainsi,

$$\sup_{\lambda>0} \left(\lambda t - \frac{\lambda^2 \sum_{i=1}^n c_i^2}{8} \right) = \frac{2t^2}{\sum_{i=1}^n c_i^2},$$

et la méthode de Chernoff fournit la borne attendue

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq e^{-\psi_{Z-\mathbb{E}[Z]}^*(t)} \leq \exp \left(\frac{-2t^2}{\sum_{i=1}^n c_i^2} \right).$$

Pour l'autre inégalité, il suffit d'appliquer le raisonnement précédent à la variable aléatoire $\tilde{Z} = -f(X_1, \dots, X_n)$ car $-f$ est également une fonction à différences finies avec les mêmes constantes c_i . ■

Comme nous le verrons dans la suite, l'inégalité des différences finies possède de nombreuses applications. Néanmoins, cette hypothèse peut sembler restrictive et le lecteur pourrait apprécier la lecture de [6, Chapitre 6] qui présente comment l'hypothèse peut être relaxée. Par exemple, les écarts de la fonction en changeant la valeur de la variable x_i peuvent être bornés par des constantes $c_i(x^{(i)})$, dépendant de $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$, à condition que la somme de ces constantes ne soit pas trop grande. Les démonstrations se basent sur la *sous-additivité de l'entropie* (la version continue de l'entropie de Shannon utilisée en théorie de l'information). Cette méthode de démonstration nécessite l'introduction de certaines quantités et résultats qui ne seront plus utilisés dans la suite de ce texte. Nous lui avons donc préféré l'alternative proposée par [15].

De nombreuses possibilités existent pour démontrer le théorème 1.15. Nous en mentionnons une que nous avons particulièrement appréciée, qui utilise un argument de transport optimal [24, Théorème 5.1].

1.2.3 Extensions

Dans le chapitre 3, nous aurons également besoin d'une inégalité similaire au théorème 1.15 faisant intervenir directement la variance dans la borne, comme les inégalités

de Bennett (théorème 1.10) et de Bernstein (théorème 1.11). Nous présentons les résultats dans un cadre ne nécessitant pas l'hypothèse des différences finies, bien que cette hypothèse puisse être utile si nous souhaitons borner les quantités apparaissant dans les inégalités.

Comme pour l'inégalité des différences finies, l'extension de l'inégalité de Bennett repose sur un lemme correspondant au lemme 1.9 pour les espérances conditionnelles.

Lemme 1.17. *Soient X et Y deux variables aléatoires, où X est supposée réelle. Si $E[X|Y] = 0$ et si $|X| \leq c$ p.s. pour un certain $c > 0$, alors pour tout $\lambda > 0$,*

$$E[e^{\lambda X}|Y] \leq \exp\left(\text{Var}[X|Y] \frac{\phi(c\lambda)}{c^2}\right),$$

où $\text{Var}[X|Y] := E[X^2|Y] - (E[X|Y])^2$ désigne la variance conditionnelle et ϕ est définie comme dans le lemme 1.9.

La démonstration, de la même manière que celle du lemme 1.16, est identique à celle du lemme 1.9 appliqué à $X|Y$.

En prenant $X := \Delta_i = E_i[Z] - E_{i-1}[Z]$ et $Y := (X_1, \dots, X_{i-1})$ et en utilisant la décomposition $Z - E[Z] = \sum_{i=1}^n \Delta_i$ comme pour la démonstration de l'inégalité des différences finies, nous obtenons le résultat suivant.

Théorème 1.18. *Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes à valeurs dans $\mathcal{X}$ et $Z := f(X_1, \dots, X_n)$, où $f : \mathcal{X}^n \rightarrow \mathbb{R}$ est une fonction mesurable. Si*

$$\sigma := \sum_{i=1}^n \sup_{x_1, \dots, x_{i-1} \in \mathcal{X}} \text{Var}[\Delta_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}],$$

et (par un léger abus de notation)

$$c := \max_{i=1, \dots, n} \sup_{x_1, \dots, x_{i-1} \in \mathcal{X}} \sup_{x \in \mathcal{X}} \Delta_i(x_1, \dots, x_{i-1}, x),$$

alors pour tout $t > 0$,

$$P(Z - E[Z] \geq t) \leq \exp\left(-\frac{\sigma}{c^2} h\left(\frac{ct}{\sigma}\right)\right),$$

où h est définie comme dans l'inégalité de Bennett.

Démonstration. Soit $i \in \{1, \dots, n\}$. Il suit du lemme précédent avec $X = \Delta_i$ et $Y := (X_1, \dots, X_{i-1})$ que pour tout $\lambda > 0$,

$$E_{i-1}[e^{\lambda \Delta_i}] \leq \exp\left(\text{Var}[\Delta_i | X_1, \dots, X_{i-1}] \frac{\phi(c\lambda)}{c^2}\right) \leq \exp\left(\sigma_i \frac{\phi(c\lambda)}{c^2}\right),$$

où $\sigma_i := \sup_{x_1, \dots, x_{i-1} \in \mathcal{X}} \text{Var}[\Delta_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}]$. En utilisant la décomposition $Z - \mathbb{E}[Z] = \sum_{i=1}^n \Delta_i$ et une démonstration par récurrence tout à fait similaire à celle présentée dans la preuve de l'inégalité des différences finies, nous obtenons

$$\psi_{Z-\mathbb{E}[Z]}(\lambda) \leq \frac{\phi(c\lambda)}{c^2} \sum_{i=1}^n \sigma_i = \frac{\phi(c\lambda)}{c^2} \sigma.$$

Comme précédemment, cette borne se traduit en une borne inférieure sur la transformée de Cramer,

$$\psi_{Z-\mathbb{E}[Z]}^*(t) = \sup_{\lambda > 0} (\lambda t - \psi_{Z-\mathbb{E}[Z]}(\lambda)) \geq \sup_{\lambda > 0} \left(\lambda t - \frac{\phi(c\lambda)}{c^2} \sigma \right) = \frac{\sigma}{c^2} h\left(\frac{ct}{\sigma}\right),$$

La méthode de Chernoff assure que pour tout $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{\sigma}{c^2} h\left(\frac{ct}{\sigma}\right)\right). \quad \blacksquare$$

En réutilisant l'inégalité (1.5), nous obtenons un résultat similaire à l'inégalité de Bernstein. Il est également présenté chez [13, Proposition 11].

Théorème 1.19. *Sous les mêmes hypothèses et en utilisant les mêmes notations que dans le théorème précédent, nous avons, pour tout $t > 0$,*

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{t^2}{2(\sigma + ct/3)}\right).$$

Comme pour l'inégalité des différences finies, ces résultats peuvent être obtenus via des méthodes de théorie de l'information. Si le lecteur est intéressé, il pourra par exemple consulter [8, Chapitre 9].

Chapitre 2

Mesures empiriques

Ce chapitre se base principalement sur [15, Chapitre 1], [7] et [10, Chapitre 12 et 13].

Dans le chapitre 1, nous avons vu que si $Y_1, \dots, Y_n$ sont des variables aléatoires i.i.d. de même loi $Be(p)$, alors l'inégalité de Hoeffding (théorème 1.8) permet d'obtenir l'inégalité de concentration

$$P\left(\frac{1}{n} \sum_{i=1}^n Y_i - p \geq t\right) \leq e^{-2nt^2},$$

pour tout $t > 0$. En particulier, si nous disposons d'un échantillon i.i.d. de données $X_1, \dots, X_n$ de même distribution μ , où μ est une mesure de probabilité sur $\mathbb{R}^d$, alors, pour A un borélien de $\mathbb{R}^d$, les variables aléatoires $Y_i := \mathbf{1}\{X_i \in A\}$ (qui correspondent simplement, pour $\omega \in \Omega$ fixé, à la fonction indicatrice de A appliquée à $X_i(\omega)$) sont i.i.d. de même loi $Be(\mu(A))$. Ainsi, en définissant $\mu_n(A)$, la *mesure empirique* de A , par

$$\mu_n(A) := \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \in A\},$$

nous avons les inégalités

$$P\left(\mu_n(A) - \mu(A) \geq t\right) \leq e^{-2nt^2},$$

et

$$P\left(\mu_n(A) - \mu(A) \leq -t\right) \leq e^{-2nt^2},$$

pour tout $t > 0$. Autrement dit, la mesure empirique $\mu_n(A)$ est fortement concentrée autour de sa moyenne, $\mu(A)$. Supposons maintenant que nous ayons une classe $\mathcal{A}$ de boréliens de $\mathbb{R}^d$ (notés $\mathcal{B}(\mathbb{R}^d)$), qui sera toujours supposée non vide dans la suite, et que nous souhaitons étudier la déviation maximale entre la mesure empirique et sa moyenne au sein de la classe, i.e. que nous soyons intéressés par la quantité

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|. \quad (2.1)$$

La seule chose que nous puissions dire grâce au chapitre 1 est que, si la classe $\mathcal{A}$ est de

cardinalité finie, disons N , alors

$$\begin{aligned}
\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) &= \mathbb{P} \left(\exists A \in \mathcal{A} : |\mu_n(A) - \mu(A)| \geq t \right) \\
&\leq \mathbb{P} \left(\bigcup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) \\
&\leq \sum_{A \in \mathcal{A}} \mathbb{P} (|\mu_n(A) - \mu(A)| \geq t) \\
&\leq 2Ne^{-2nt^2}.
\end{aligned}$$

Par conséquent, dès que la classe $\mathcal{A}$ contient une infinité de boréliens de $\mathbb{R}^d$, la théorie du chapitre 1 ne permet plus d'obtenir des bornes sur la probabilité d'intérêt.

Dans ce second chapitre, nous montrons comment, en utilisant certaines quantités combinatoires liées à la classe $\mathcal{A}$, il est possible de fournir des bornes intéressantes sur cette probabilité. Ces quantités proviennent de la *théorie de Vapnik–Chervonenkis* due à Vladimir Naumovitch Vapnik et Alexey Chervonenkis dans [29]. Elle est surtout développée pour l'apprentissage statistique, comme expliqué dans l'introduction (ou, par exemple, chez [7]). Néanmoins, l'approche que nous suivrons est purement probabiliste. La section 2.1 introduit les quantités combinatoires qui seront utilisées dans la suite, la section 2.2 présente les inégalités et les comparaisons graphiques se situent dans la section 2.3.

Remarque 2.1 (Mesurabilité). À ce stade, le lecteur averti peut se soucier de la mesurabilité des quantités qui seront étudiées dans ce chapitre (et dans le chapitre suivant) ; par exemple, (2.1) n'est pas une variable aléatoire si $\mathcal{A}$ n'est pas dénombrable. Une solution possible est de travailler avec des *probabilités extérieures* qui sont définies, même pour des ensembles non mesurables : si nous disposons d'un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, alors pour n'importe quel $E \subset \Omega$, nous définissons sa probabilité extérieure par

$$\mathbb{P}^*(E) := \inf \{ \mathbb{P}(F) : E \subset F, F \in \mathcal{F} \},$$

et il est alors, par exemple, possible de montrer ([28, Lemme 1.2.3]) que l'infimum est atteint par un ensemble $F^* \in \mathcal{F}$. De la même manière, l'*espérance extérieure* peut être définie pour une fonction arbitraire $T : \Omega \rightarrow \mathbb{R}$, pas nécessairement mesurable, par

$$\mathbb{E}^*[T] := \inf \{ \mathbb{E}[U] : U \geq T, U : \Omega \rightarrow \mathbb{R} \text{ mesurable et } \mathbb{E}[U] \text{ existe} \}.$$

Le premier chapitre de [28] traite en détails la convergence en distribution (et d'autres modes de convergence) dans ce cadre de non mesurabilité, en pointant les similitudes et les différences avec la théorie classique. Par exemple, le théorème de Fubini qui permet les arguments de conditionnement ultérieurement utilisés n'est plus valide en toute généralité ([28, Lemme 1.2.6 et Lemme 1.2.7]). Dès lors, cette approche n'est pas celle que nous suivrons dans la suite. Nous effectuerons l'hypothèse additionnelle sur $\mathcal{A}$ qu'il existe une

collection au plus dénombrable $\mathcal{A}_0 \subset \mathcal{A}$ telle que pour tout $A \in \mathcal{A}$, il existe une suite $(A_n)_{n \in \mathbb{N}} \subset \mathcal{A}_0$ qui satisfait

$$\lim_{n \rightarrow +\infty} \mathbb{1}_{A_n}(x) = \mathbb{1}_A(x), \quad (2.2)$$

pour n'importe quel $x \in \mathbb{R}^d$. Dans le langage de [28, Exemple 2.3.4], cela correspond à demander que la classe des indicatrices $\{\mathbb{1}_A : A \in \mathcal{A}\}$ soit *ponctuellement mesurable*. Si nous définissons la limite supérieure et la limite inférieure d'une suite d'ensembles, comme dans [26] p. 61, par

$$\liminf_{n \rightarrow +\infty} A_n := \bigcup_{n \geq 1} \bigcap_{k \geq n} A_k \quad \text{et} \quad \limsup_{n \rightarrow +\infty} A_n := \bigcap_{n \geq 1} \bigcup_{k \geq n} A_k,$$

alors on peut montrer que la condition (2.2) est équivalente à demander que $A = \liminf_{n \rightarrow +\infty} A_n = \limsup_{n \rightarrow +\infty} A_n =: \lim_{n \rightarrow +\infty} A_n$. C'est une conséquence directe du fait que

$$\limsup_{n \rightarrow +\infty} \mathbb{1}_{A_n}(x) := \inf_{n \in \mathbb{N}} \sup_{k \geq n} \mathbb{1}_{A_k}(x) = \inf_{n \in \mathbb{N}} \mathbb{1}_{\bigcup_{k \geq n} A_k}(x) = \mathbb{1}_{\limsup_{n \rightarrow +\infty} A_n}(x)$$

et

$$\liminf_{n \rightarrow +\infty} \mathbb{1}_{A_n}(x) := \sup_{n \in \mathbb{N}} \inf_{k \geq n} \mathbb{1}_{A_k}(x) = \sup_{n \in \mathbb{N}} \mathbb{1}_{\bigcap_{k \geq n} A_k}(x) = \mathbb{1}_{\liminf_{n \rightarrow +\infty} A_n}(x).$$

Ainsi, $\mathcal{A}$ peut être remplacé partout par $\mathcal{A}_0$ et la mesurabilité du supremum est alors garantie car

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| = \sup_{A \in \mathcal{A}_0} |\mu_n(A) - \mu(A)|.$$

L'inégalité « $\geq$ » est évidente. Pour montrer que l'inégalité « $\leq$ » tient également, considérons la fonction $J : A \in \mathcal{A} \mapsto J(A) := |\mu_n(A) - \mu(A)| \in \mathbb{R}$. Soit une suite maximisante $(A_k)_{k \in \mathbb{N}} \subset \mathcal{A}$ telle que $\lim_{k \rightarrow +\infty} J(A_k) = \sup_{A \in \mathcal{A}} J(A)$. Par hypothèse de mesurabilité ponctuelle, pour tout $k \in \mathbb{N}$, il existe une suite $(A_l^k)_{l \in \mathbb{N}} \subset \mathcal{A}_0$ telle que $\lim_{l \rightarrow +\infty} A_l^k = A_k$. Soit $\epsilon > 0$. Par définition de limite, il existe $k(\epsilon) \in \mathbb{N}$ tel que pour tout $k \geq k(\epsilon)$,

$$\left| \sup_{\mathcal{A}} J - J(A_{k(\epsilon)}) \right| \leq \epsilon.$$

De plus, comme la mesure μ est une mesure de probabilité, le théorème de convergence dominée assure que la fonction J est continue par rapport à la convergence (2.2). Par conséquent, il existe $l[k(\epsilon)]$ tel que pour tout $l \geq l[k(\epsilon)]$,

$$\left| J(A_l^{k(\epsilon)}) - J(A_{k(\epsilon)}) \right| \leq \epsilon.$$

Par l'inégalité triangulaire, il suit que pour $l \geq l[k(\epsilon)]$,

$$\left| \sup_{\mathcal{A}} J - J(A_l^{k(\epsilon)}) \right| \leq \left| \sup_{\mathcal{A}} J - J(A_{k(\epsilon)}) \right| + \left| J(A_l^{k(\epsilon)}) - J(A_{k(\epsilon)}) \right| \leq 2\epsilon.$$

Comme $A_l^{k(\epsilon)} \in \mathcal{A}_0$ pour tout $l \in \mathbb{N}$ et comme $\epsilon > 0$ est arbitraire, nous concluons que le supremum de J sur $\mathcal{A}$ ne peut excéder le supremum de J sur $\mathcal{A}_0$.

2.1 Quelques quantités combinatoires

L'idée principale de la théorie de Vapnik et Chervonenkis (nous ferons référence à ces auteurs via l'abréviation VC) est que l'élément $A \in \mathcal{A}$ pouvant atteindre le supremum

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|,$$

se trouve dans une sous-classe de $\mathcal{A}$ dont la cardinalité, même si $\mathcal{A}$ contient un nombre infini de boréliens de $\mathbb{R}^d$, peut être bornée (en tout cas pour diverses classes qui nous intéressent).

2.1.1 Coefficient de désintégration et dimension VC

Une première quantité intéressante liée à la complexité de la classe $\mathcal{A}$ est son *coefficient de désintégration* (de l'anglais « *shatter coefficient* »).

Définition 2.2 (Coefficient de désintégration). Soient $n \in \mathbb{N}$ et $\mathcal{A}$ une classe de boréliens de $\mathbb{R}^d$. On définit le coefficient de désintégration de $\mathcal{A}$ de taille n par

$$\mathbb{S}_{\mathcal{A}}(n) := \max_{x_1, \dots, x_n \in \mathbb{R}^d} |\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A}\}|.$$

Autrement dit, $\mathbb{S}_{\mathcal{A}}(n)$ décrit le nombre maximal de sous-ensembles de n points que l'on peut former en intersectant ces points avec les éléments de la classe $\mathcal{A}$. Nous comprenons intuitivement que si la classe $\mathcal{A}$ est « riche », alors certainement ce nombre sera grand. Dans tous les cas, $\mathbb{S}_{\mathcal{A}}(n) \leq 2^n$ pour tout $n \in \mathbb{N}$ car nous ne pourrions jamais former plus de sous-ensembles de $\{x_1, \dots, x_n\}$ que l'ensemble de ses parties. Si, pour un ensemble de points $\{x_1, \dots, x_n\}$, nous avons que

$$|\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A}\}| = 2^n,$$

alors nous dirons que $\mathcal{A}$ *désintègre* l'ensemble $\{x_1, \dots, x_n\}$. Cela conduit naturellement à la définition suivante.

Définition 2.3 (Dimension VC). La *dimension VC* d'une classe $\mathcal{A}$ de boréliens de $\mathbb{R}^d$, notée V (ou $V_{\mathcal{A}}$ si une confusion est possible), est définie par

$$V := \sup\{n \in \mathbb{N} : \mathbb{S}_{\mathcal{A}}(n) = 2^n\}.$$

Clairement, V est bien définie car si pour un certain $n \in \mathbb{N}$, la classe $\mathcal{A}$ ne sait désintégrer aucun ensemble de points de cette taille, i.e. $\mathbb{S}_{\mathcal{A}}(n) < 2^n$, alors certainement elle n'est pas capable de désintégrer un ensemble de taille plus grande, i.e. pour tout naturel $m > n$, $\mathbb{S}_{\mathcal{A}}(m) < 2^m$.

Il est important de remarquer que ces notions sont définies pour un cadre bien plus général que celui des boréliens de $\mathbb{R}^d$. Dès que nous disposons d'un espace E , d'une classe d'ensembles de cet espace, disons $\mathcal{H}$, et d'un ensemble C dans E , toutes les notions précédemment introduites font sens. Par exemple, en définissant,

$$\mathcal{H} \cap C := \{H \cap C : H \in \mathcal{H}\},$$

nous disons que $\mathcal{H}$ désintègre C si

$$|\mathcal{H} \cap C| = 2^{|C|}.$$

De manière similaire, la dimension VC de $\mathcal{H}$ est la plus grande cardinalité telle qu'il existe un ensemble de cette taille désintégré par $\mathcal{H}$. Cependant, comme nous nous intéressons uniquement au cas de mesures empiriques sur $\mathbb{R}^d$, nous avons préféré présenter ces quantités combinatoires dans le même cadre.

La dimension VC se calcule assez facilement pour des classes intéressantes.

Exemple 2.4 (Fonction de répartition empirique, calcul de la dimension VC). Supposons que $d = 1$ et

$$\mathcal{A} := \{(-\infty, t] : t \in \mathbb{R}\}. \quad (2.3)$$

Cette classe présente un réel intérêt pratique car, dans ce cas, la mesure empirique associée à un échantillon i.i.d. de variables aléatoires réelles $X_1, \dots, X_n$ devient

$$\mu_n((-\infty, t]) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \leq t\} = F_n(t),$$

où $F_n(t)$ est la *fonction de répartition empirique* associée aux observations $X_1, \dots, X_n$. Notons que la classe $\mathcal{A}$ définie dans (2.3) satisfait clairement l'hypothèse de mesurabilité ponctuelle évoquée à la remarque 2.1, par densité des nombres rationnels dans les nombres réels. Si F désigne la répartition commune des X_i , la statistique

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

est très utilisée en pratique, notamment dans des tests non paramétriques célèbres comme celui de Kolmogorov–Smirnov, voir par exemple [21, Exemple 11.2.12].

Pour cette classe, nous sommes capables de calculer sa dimension VC et montrons que $V = 1$. En effet, supposons dans un premier temps que nous disposons d'un seul point $x_1 \in \mathbb{R}$. Alors en prenant $t < x_1$ et choisissant $I_t := (-\infty, t] \in \mathcal{A}$, nous avons $I_t \cap \{x_1\} = \emptyset$. De plus, en prenant $t > x_1$, nous avons $I_t \cap \{x_1\} = \{x_1\}$. Donc $\mathcal{A}$ désintègre un singleton. Par contre, si nous avons deux points distincts $\{x_1, x_2\}$ (que l'on suppose ordonnés par ordre croissant), la classe $\mathcal{A}$ ne sait plus désintégrer cet ensemble. Nous pourrions obtenir les ensembles $\emptyset, \{x_1\}, \{x_1, x_2\}$ mais jamais $\{x_2\}$. Cela montre que $V = 1$. Il est en fait simple de voir que pour tout $n \in \mathbb{N}$, nous avons $\mathbb{S}_{\mathcal{A}}(n) = n + 1$.

Exemple 2.5 (Rectangles). Prenons maintenant pour $\mathcal{A}$ l'ensemble de tous les rectangles de $\mathbb{R}^d$ alignés avec les éléments de la base canonique de $\mathbb{R}^d$. Alors $V = 2d$. Nous le démontrerons dans le cas $d = 2$ pour simplifier la visualisation, mais cela fonctionne pour le cas général de la même manière.

Pour montrer le résultat, il faut d'une part montrer que la classe $\mathcal{A}$ permet de désintégrer un ensemble de points jusqu'à une cardinalité de 4 (où l'ensemble peut être choisi comme on le souhaite). D'autre part, il faut montrer que pour tout ensemble de points de cardinalité 5, il n'est pas possible de récupérer toutes les parties de cet ensemble en intersectant l'ensemble avec les éléments de $\mathcal{A}$.

La première partie est évidente et les détails sont laissés au lecteur. Nous présentons simplement dans la figure 2.1 une suggestion pour la construction d'un ensemble de 4 points pouvant être désintégré par $\mathcal{A}$.

Pour la seconde partie, considérons un ensemble arbitraire $\{x_1, \dots, x_5\}$ de points de $\mathbb{R}^d$. Quitte à renommer les points, nous pouvons supposer que x_1 possède la plus grande première composante, x_2 possède la plus petite première composante, x_3 possède la plus grande seconde composante et x_4 possède la plus petite seconde composante. Alors tout rectangle qui contient les points $\{x_1, \dots, x_4\}$ contient forcément $\{x_5\}$ et nous ne pouvons pas obtenir l'ensemble $\{x_1, \dots, x_4\}$ par intersection avec un rectangle aligné avec les axes. La figure 2.2 illustre ce propos.

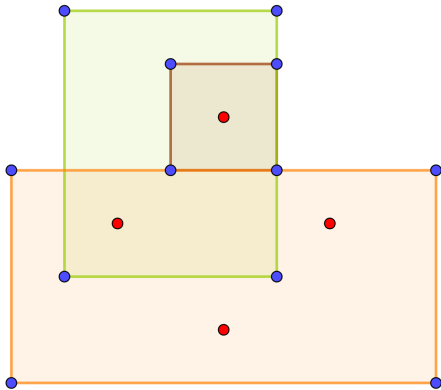


FIGURE 2.1 – La classe $\mathcal{A}$ désintègre les points rouges.

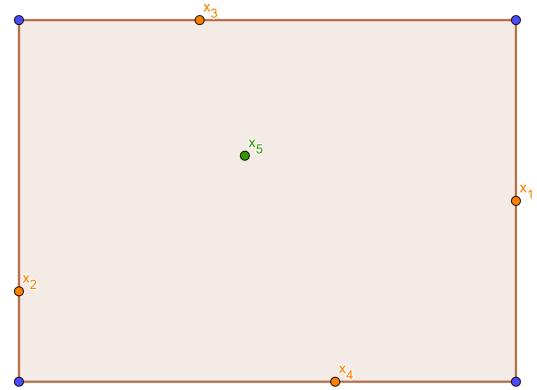


FIGURE 2.2 – Tout rectangle contenant $x_1, \dots, x_4$ contient aussi x_5 .

Nous avons donc montré le résultat pour $\mathbb{R}^2$ et le résultat général suit d'une construction similaire. Dès que l'on dépasse $2d$ points, l'ajout d'un nouveau point empêche la désintégration via $\mathcal{A}$.

Il est assez clair de l'exemple précédent que la capacité de désintégration d'une classe de convexes est liée à la rigidité des éléments de la classe. S'ils sont trop rigides, leur

intérieur contient, à partir d'un certain nombre de points, des éléments que l'on souhaiterait exclure et nous ne sommes plus capables de désintégrer les ensembles trop grands. Si, par contre, nous disposons d'une classe d'éléments convexes assez flexibles, la dimension peut être infinie, comme l'assure le résultat suivant provenant de [10, Théorème 13.10].

Proposition 2.6. *La classe de tous les polygones convexes de $\mathbb{R}^2$ a une dimension VC infinie.*

Esquisse de démonstration. Pour le voir, considérons un nombre de points n fixés sur le cercle unité. Il est simple de voir que l'ensemble des polygones convexes peut désintégrer l'ensemble de ces points. Comme cela fonctionne pour tout $n \in \mathbb{N}$, nous avons forcément que $V = +\infty$. ■

L'utilité de la dimension VC peut sembler limitée dans un premier temps : en effet, si la dimension VC d'une classe $\mathcal{A}$ est finie, tout ce que l'on sait pour l'instant est que

$$\begin{cases} \mathbb{S}_{\mathcal{A}}(n) = 2^n & \text{pour } n \leq V, \\ \mathbb{S}_{\mathcal{A}}(n) < 2^n & \text{pour } n > V. \end{cases}$$

Il s'avère que pour $n > V$, la vitesse de croissance du coefficient de désintégration est en réalité polynomiale. Ce résultat est connu sous le nom de *lemme de Sauer* (plus particulièrement le corollaire 2.8).

Théorème 2.7 (Lemme de Sauer). *Soit $\mathcal{A}$ une classe de boréliens de $\mathbb{R}^d$. Si la dimension VC de $\mathcal{A}$ est finie, alors, pour tout $n \in \mathbb{N}$,*

$$\mathbb{S}_{\mathcal{A}}(n) \leq \sum_{i=0}^V \binom{n}{i},$$

où les coefficients binomiaux $\binom{n}{i}$ sont définis par

$$\binom{n}{i} := \begin{cases} \frac{n!}{i!(n-i)!} & \text{si } 0 \leq i \leq n, \\ 0 & \text{sinon.} \end{cases}$$

Nous suivons la démonstration de [2, Lemme 1.8].

Démonstration. La démonstration se fait par induction sur n .

Commençons par vérifier le cas initial. Si $n = 0$, alors peu importe la classe $\mathcal{A}$ (qui est toujours supposée non vide, comme annoncé en début de chapitre)

$$\mathbb{S}_{\mathcal{A}}(0) = |\{\emptyset \cap A : A \in \mathcal{A}\}| = |\emptyset| = 1 = \sum_{i=0}^{V_{\mathcal{A}}} \binom{0}{i}.$$

Supposons maintenant le résultat vrai jusque $n - 1$. Fixons $\{x_1, \dots, x_n\}$ des points de $\mathbb{R}^d$ et définissons la collection (que nous supposons également non vide dans un premier temps)

$$\tilde{\mathcal{A}} := \{A \in \mathcal{A} \mid \exists B \in \mathcal{A} : \{x_1, \dots, x_{n-1}\} \cap A = \{x_1, \dots, x_{n-1}\} \cap B \\ \text{et } \{x_1, \dots, x_n\} \cap A \neq \{x_1, \dots, x_n\} \cap B\}.$$

Si $\tilde{\mathcal{A}}$ désintègre $\{x_1, \dots, x_{n-1}\}$, alors certainement $\mathcal{A}$ désintègre $\{x_1, \dots, x_{n-1}\} \cup \{x_n\}$ car pour chaque sous-ensemble de points de $\{x_1, \dots, x_n\}$ sélectionné par un élément de $\tilde{\mathcal{A}}$, un différent sous-ensemble est sélectionné par un autre élément de $\mathcal{A}$. Comme la cardinalité de l'ensemble des parties de $\{x_1, \dots, x_n\}$ est exactement le double de celui de $\{x_1, \dots, x_{n-1}\}$, cela implique forcément que $\{x_1, \dots, x_n\}$ est désintégré par $\mathcal{A}$. Dès lors, nous avons

$$V_{\tilde{\mathcal{A}}} + 1 \leq V_{\mathcal{A}},$$

ou, de manière équivalente,

$$V_{\tilde{\mathcal{A}}} \leq V_{\mathcal{A}} - 1.$$

De plus, montrons également que

$$|\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A}\}| = |\{\{x_1, \dots, x_{n-1}\} \cap \tilde{A} : \tilde{A} \in \tilde{\mathcal{A}}\}| \\ + |\{\{x_1, \dots, x_{n-1}\} \cap A : A \in \mathcal{A}\}|.$$

Pour comprendre la provenance de cette égalité, il faut remarquer que l'on peut couper la classe $\mathcal{A}$ en deux parties disjointes $\tilde{\mathcal{A}}$ et $\mathcal{A} \setminus \tilde{\mathcal{A}}$, de telle sorte que

$$|\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A}\}| = |\{\{x_1, \dots, x_n\} \cap A : A \in \tilde{\mathcal{A}}\}| \\ + |\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A} \setminus \tilde{\mathcal{A}}\}|.$$

Une intersection entre $\{x_1, \dots, x_n\}$ et un élément de $\mathcal{A} \setminus \tilde{\mathcal{A}}$ est uniquement représentée par l'intersection de cet élément avec $\{x_1, \dots, x_{n-1}\}$ car il n'existe pas d'autre élément de $\mathcal{A}$ qui puisse avoir la même intersection $\{x_1, \dots, x_{n-1}\}$ et une intersection différente avec $\{x_1, \dots, x_n\}$. Dès lors,

$$|\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A} \setminus \tilde{\mathcal{A}}\}| = |\{\{x_1, \dots, x_{n-1}\} \cap A : A \in \mathcal{A} \setminus \tilde{\mathcal{A}}\}|.$$

En outre, il suit de la définition de $\tilde{\mathcal{A}}$ que

$$|\{\{x_1, \dots, x_n\} \cap A : A \in \tilde{\mathcal{A}}\}| = 2|\{\{x_1, \dots, x_{n-1}\} \cap A : A \in \tilde{\mathcal{A}}\}|.$$

En combinant ces relations, nous trouvons l'identité attendue :

$$|\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A}\}| \\ = |\{\{x_1, \dots, x_n\} \cap A : A \in \tilde{\mathcal{A}}\}| + |\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A} \setminus \tilde{\mathcal{A}}\}| \\ = 2|\{\{x_1, \dots, x_{n-1}\} \cap A : A \in \tilde{\mathcal{A}}\}| + |\{\{x_1, \dots, x_{n-1}\} \cap A : A \in \mathcal{A} \setminus \tilde{\mathcal{A}}\}|$$

$$= |\{\{x_1, \dots, x_{n-1}\} \cap A : A \in \tilde{\mathcal{A}}\}| + |\{\{x_1, \dots, x_{n-1}\} \cap A : A \in \mathcal{A}\}|.$$

Par hypothèse de récurrence et en utilisant la relation $\binom{n-1}{i-1} + \binom{n-1}{i} = \binom{n}{i}$, nous avons donc

$$\begin{aligned} |\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A}\}| &\leq \mathbb{S}_{\tilde{\mathcal{A}}}(n-1) + \mathbb{S}_{\mathcal{A}}(n-1) \\ &\leq \sum_{i=0}^{V_{\mathcal{A}}-1} \binom{n-1}{i} + \sum_{i=0}^{V_{\mathcal{A}}} \binom{n-1}{i} \\ &= \sum_{i=0}^{V_{\mathcal{A}}} \binom{n-1}{i-1} + \sum_{i=0}^{V_{\mathcal{A}}} \binom{n-1}{i} \\ &= \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}. \end{aligned}$$

En passant au maximum sur $\{x_1, \dots, x_n\}$, nous obtenons le résultat.

Si la classe $\tilde{\mathcal{A}}$ est vide, il suit du raisonnement précédent que

$$|\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A}\}| = |\{\{x_1, \dots, x_{n-1}\} \cap A : A \in \mathcal{A}\}|.$$

Ainsi,

$$|\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A}\}| \leq \mathbb{S}_{\mathcal{A}}(n-1) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n-1}{i} \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i},$$

et comme précédemment, nous trouvons le résultat en passant au maximum sur $\{x_1, \dots, x_n\}$. ■

Ce résultat possède un important corollaire qui sera utilisé de manière extensive dans la suite.

Corollaire 2.8. *Soit $\mathcal{A}$ une classe de boréliens de $\mathbb{R}^d$. Si la dimension VC de $\mathcal{A}$ est finie, alors, pour tout $n \in \mathbb{N}$,*

$$\mathbb{S}_{\mathcal{A}}(n) \leq (n+1)^V,$$

et pour tout $n \geq V$,

$$\mathbb{S}_{\mathcal{A}}(n) \leq \left(\frac{ne}{V}\right)^V.$$

Démonstration. Pour démontrer la première inégalité, rappelons la formule du binôme de Newton : pour tout $x, y \in \mathbb{R}$ et $k \in \mathbb{N}$, nous avons

$$(x+y)^k = \sum_{i=0}^k \binom{k}{i} x^i y^{k-i}.$$

Appliquons ce résultat avec $x = n$, $y = 1$ et $k = V$

$$(n+1)^V = \sum_{i=0}^V \binom{V}{i} n^i = \sum_{i=0}^V \frac{V! n^i}{i! (V-i)!} = \sum_{i=0}^V \frac{n^i}{i!} \frac{V!}{(V-i)!} \geq \sum_{i=0}^V \frac{n^i}{i!} \geq \sum_{i=0}^V \binom{n}{i}.$$

En vertu du lemme de Sauer, cela montre la première inégalité.

Pour la seconde, nous utilisons encore la formule du binôme de Newton et le fait que $V/n \leq 1$. Il suit de la définition de e que

$$e^V \geq \left(1 + \frac{V}{n}\right)^n = \sum_{i=0}^n \binom{n}{i} \left(\frac{V}{n}\right)^i \geq \sum_{i=0}^V \binom{n}{i} \left(\frac{V}{n}\right)^i \geq \left(\frac{V}{n}\right)^V \sum_{i=0}^V \binom{n}{i}.$$

Dès lors, pour $n \geq V$, par le lemme de Sauer,

$$\mathbb{S}_{\mathcal{A}}(n) \leq \sum_{i=0}^V \binom{n}{i} \leq \left(\frac{ne}{V}\right)^V. \quad \blacksquare$$

2.1.2 Nombre couvrant

Rappelons que le coefficient de désintégration de taille n d'une classe $\mathcal{A}$ est défini par

$$\mathbb{S}_{\mathcal{A}}(n) = \max_{x_1, \dots, x_n \in \mathbb{R}^d} |\{\{x_1, \dots, x_n\} \cap A : A \in \mathcal{A}\}|.$$

Une manière alternative de représenter cette quantité est de définir

$$\mathcal{A}(x_1^n) := \{b = (b_1, \dots, b_n) \in \{0, 1\}^n : b_i = \mathbb{1}\{x_i \in A\} \text{ pour } A \in \mathcal{A}\},$$

et

$$\mathbb{S}_{\mathcal{A}}(n) := \max_{x_1, \dots, x_n \in \mathbb{R}^d} |\mathcal{A}(x_1^n)|.$$

Autrement dit, chaque élément de l'ensemble des parties de $x_1^n := \{x_1, \dots, x_n\}$ est représenté par un sommet de l'hypercube $[0, 1]^n$. Comme l'hypercube de $\mathbb{R}^n$ admet 2^n sommets, il permet cette représentation.

Cette manière alternative de définir le coefficient de désintégration mène à une autre mesure de la complexité de la classe $\mathcal{A}$, le *nombre couvrant*. Avant d'introduire la définition précise de cette notion, nous devons munir les sommets de l'hypercube d'une distance.

Définition 2.9 (Distance de Hamming). Soient $b = (b_1, \dots, b_n)$ et $c = (c_1, \dots, c_n)$ dans $\{0, 1\}^n$. La distance de Hamming entre b et c est définie par

$$\rho(b, c) := \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{b_i \neq c_i\}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (b_i - c_i)^2}.$$

Autrement dit, la distance de Hamming correspond à la distance euclidienne sur l'hypercube, normalisée par un facteur $n^{-1/2}$ (en particulier, elle satisfait donc bien les propriétés d'une distance).

Nous pouvons maintenant définir le *recouvrement* d'un ensemble de sommets de l'hypercube (par des boules).

Définition 2.10 (Recouvrement). Soit $B \subseteq \{0, 1\}^n$. Un recouvrement de rayon r de B est un ensemble $B_r \subseteq \{0, 1\}^n$ tel que pour tout $b \in B$, il existe $c \in B_r$ tel que $\rho(b, c) \leq r$.

Cela conduit à la définition du *nombre couvrant* d'un ensemble de $\{0, 1\}^n$.

Définition 2.11 (Nombre couvrant). Soient $B \subseteq \{0, 1\}^n$ et $r > 0$. Le nombre couvrant $\mathcal{N}(r, B)$ est la cardinalité du plus petit recouvrement de B de rayon r .

Intuitivement, si une classe $\mathcal{A}$ de boréliens de $\mathbb{R}^d$ est « riche », le nombre couvrant $\mathcal{N}(r, \mathcal{A}(x_1^n))$ est grand. Bien qu'il soit compliqué à calculer en pratique, nous avons le résultat suivant, dû à Haussler ([19, Théorème 1]), qui assure que si $\mathcal{A}$ est de dimension VC finie, nous avons un contrôle sur le nombre couvrant $\mathcal{N}(r, \mathcal{A}(x_1^n))$.

Théorème 2.12. Soient $\mathcal{A}$ une classe de boréliens de $\mathbb{R}^d$, $\{x_1, \dots, x_n\}$ des points de $\mathbb{R}^d$ et $0 < r \leq 1$. Si $V = V_{\mathcal{A}} < +\infty$, nous avons

$$\mathcal{N}(r, \mathcal{A}(x_1^n)) \leq e(V + 1) \left(\frac{2e}{r^2} \right)^V.$$

La démonstration est longue et n'apporte pas d'idées utiles pour la suite. Si le lecteur souhaite consulter la preuve, nous l'invitons à parcourir directement le papier de Haussler [19].

2.2 Concentration de mesures empiriques

Nous souhaitons fournir des bornes sur la probabilité

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right),$$

où $t > 0$ et $\mathcal{A}$ désigne une classe non vide de boréliens de $\mathbb{R}^d$. Afin de présenter les résultats sous une forme plus explicite, nous suivrons une approche totalement équivalente qui consiste à fournir une borne qui tient avec grande probabilité sur $Z = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$. Ceci est formalisé dans le lemme suivant.

Lemme 2.13 (Inversion). *Soit $\delta \in (0, 1)$. Si pour tout $t > 0$ nous avons une inégalité de la forme*

$$P(Z \geq t) \leq S(t),$$

où $S : [0, +\infty) \rightarrow [0, \infty)$ est une fonction inversible, alors avec probabilité au moins $1 - \delta$,

$$Z \leq S^{-1}(\delta),$$

et l'inégalité tient encore si elle est remplacée par une inégalité stricte.

Démonstration. Soit $t = t(\delta) := S^{-1}(\delta)$. Alors,

$$P(Z \geq t(\delta)) \leq S(t(\delta)) = \delta.$$

Autrement dit, la probabilité que Z dépasse $S^{-1}(\delta)$ est bornée par δ . Cela implique que Z est borné (strictement) par $S^{-1}(\delta)$ avec probabilité au moins $1 - \delta$. ■

Deux méthodes seront principalement étudiées.

Dans un premier temps, nous obtenons des bornes sur Z en étudiant directement sa distribution et en appliquant le lemme d'inversion. Les inégalités reposent principalement sur des méthodes de symétrisation, de randomisation et de conditionnement (des termes qui seront expliqués dans la suite).

Dans un second temps, nous utilisons le fait que $Z = Z(X_1, \dots, X_n)$ est une fonction à différences finies (définition 1.13) avec des constantes $c_i = 1/n$. L'inégalité des différences finies (théorème 1.15) assure alors que

$$P(Z - E[Z] \geq t) \leq e^{-2nt^2}. \quad (2.4)$$

Autrement dit, Z est fortement concentrée autour de sa moyenne. Par conséquent, il suffit de borner $E[Z]$ pour obtenir des bornes sur Z . Cela suit directement du lemme d'inversion appliqué à l'équation (2.4). En effet dans ce cas, $S(t) = e^{-2nt^2}$, donc avec probabilité au moins $1 - \delta$, nous avons $Z - E[Z] \leq S^{-1}(\delta) = \sqrt{\frac{1}{2n} \log(1/\delta)}$. De manière équivalente, avec probabilité au moins $1 - \delta$,

$$Z \leq \sqrt{\frac{1}{2n} \log(1/\delta)} + E[Z], \quad (2.5)$$

et, fournir une borne sur la moyenne de Z se traduit immédiatement en une borne sur Z . Encore une fois, les raisonnements seront principalement basés sur des méthodes de symétrisation, de randomisation et de conditionnement. Toutefois, une méthode additionnelle, dite de *chaînage* (de l'anglais « chaining »), due à Richard Dudley (1938–2020) dans [11], sera également discutée afin d'obtenir des bornes asymptotiquement meilleures sur $E[Z]$.

Une synthèse et une comparaison détaillée des différents résultats seront proposées en fin de chapitre, basées sur des calculs explicites dans le cas de la fonction de répartition empirique.

2.2.1 Bornes provenant de la distribution

La première borne que nous présentons est la célèbre *inégalité de Vapnik–Chervonenkis* ([29, Théorème 2]), reformulée en utilisant le lemme d'inversion.

Théorème 2.14 (Inégalité de Vapnik–Chervonenkis). *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Alors, pour $\delta \in (0, 1)$, nous avons avec probabilité $1 - \delta$,*

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2}{n} [\log(4/\delta) + \log(\mathbb{S}_{\mathcal{A}}(2n))]}.$$

Ce résultat est particulièrement intéressant quand nous avons à faire à une classe $\mathcal{A}$ de dimension VC finie. Dans ce cas, par le corollaire 2.8, nous avons $\log(\mathbb{S}_{\mathcal{A}}(2n)) \leq V \log(2n + 1)$, et avec probabilité $1 - \delta$, il suit que

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2}{n} [\log(4/\delta) + V \log(2n + 1)]},$$

une quantité qui tend rapidement vers zéro avec n .

Cette inégalité est souvent présentée dans sa forme équivalente, comme l'inégalité de concentration suivante (voir par exemple [7], p.186) qui assure que pour tout $t > 0$,

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-nt^2/8}. \quad (2.6)$$

La démonstration de cette inégalité repose sur un *lemme de symétrisation* que nous présentons dans une forme plus générale que la formulation classique présentée chez [7, Lemme 2]. Cela nous permettra ultérieurement (théorème 2.17) d'améliorer l'inégalité de Vapnik et Chervonenkis.

Lemme 2.15 (Lemme de symétrisation). *Soit $X_1, \dots, X_n, X'_1, \dots, X'_n$ un échantillon i.i.d. de taille $2n$ de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$. Notons μ_n la distribution empirique associée à $X_1, \dots, X_n$ et μ'_n la distribution empirique associée à*

$X'_1, \dots, X'_n$. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Soient $t > 0$ et $a \in (0, 1)$ tels que $4(1-a)^2 nt^2 \geq 2$, alors

$$\left. \begin{aligned} & \mathbb{P} \left(\sup_{A \in \mathcal{A}} (\mu_n(A) - \mu(A)) \geq t \right) \\ & \mathbb{P} \left(\sup_{A \in \mathcal{A}} (\mu(A) - \mu_n(A)) \geq t \right) \end{aligned} \right\} \leq 2 \mathbb{P} \left(\sup_{A \in \mathcal{A}} (\mu_n(A) - \mu'_n(A)) \geq at \right).$$

Le choix $a = 1/2$ représente le choix classique. L'échantillon $X'_1, \dots, X'_n$ est souvent appelé *échantillon fantôme* (de l'anglais « ghost sample »).

Démonstration du lemme 2.15. Soient $A \in \mathcal{A}$ et $t > 0$. Observons que

$$\mathbb{1}\{\mu_n(A) - \mu(A) \geq t\} \mathbb{1}\{\mu'_n(A) - \mu(A) \leq (1-a)t\} \leq \mathbb{1}\{\mu_n(A) - \mu'_n(A) \geq at\}. \quad (2.7)$$

En effet, pour que le membre de gauche soit 1, il faut que les deux événements associés aux indicatrices se produisent, i.e.

$$\begin{cases} \mu_n(A) - \mu(A) \geq t, \\ \mu'_n(A) - \mu(A) \leq (1-a)t. \end{cases}$$

Cela implique forcément que l'évènement associé à l'indicatrice du membre de droite se réalise car dans ce cas,

$$\mu_n(A) - \mu'_n(A) = (\mu_n(A) - \mu(A)) - (\mu'_n(A) - \mu(A)) \geq t - (1-a)t = at.$$

En prenant l'espérance de chaque côté de (2.7) par rapport à l'échantillon fantôme, il vient

$$\mathbb{1}\{\mu_n(A) - \mu(A) \geq t\} \mathbb{P}'(\mu'_n(A) - \mu(A) \leq (1-a)t) \leq \mathbb{P}'(\mu_n(A) - \mu'_n(A) \geq at),$$

où nous avons utilisé la notation $\mathbb{P}'$ pour mettre en évidence le fait que nous avons intégré uniquement par rapport à l'échantillon fantôme. Par l'inégalité de Bienaymé–Tchebychev (théorème 1.2), nous avons

$$\begin{aligned} \mathbb{P}'(\mu'_n(A) - \mu(A) \leq (1-a)t) &= 1 - \mathbb{P}'(\mu'_n(A) - \mu(A) \geq (1-a)t) \\ &\geq 1 - \frac{\text{Var}[\mu'_n(A)]}{(1-a)^2 t^2} \\ &= 1 - \frac{\text{Var}[\mathbb{1}\{X'_1 \in A\}]}{n(1-a)^2 t^2} \\ &\geq 1 - \frac{1}{4n(1-a)^2 t^2} \\ &\geq 1/2 \end{aligned}$$

où l'avant-dernière inégalité vient du fait que la variance d'une indicatrice est toujours bornée supérieurement par 1/4 (car $\max_{p \in [0,1]} \{p(1-p)\} = 1/4$) et la dernière inégalité suit par hypothèse. En regroupant les inégalités, nous trouvons

$$\mathbb{1}\{\mu_n(A) - \mu(A) \geq t\} \leq 2P'(\mu_n(A) - \mu'_n(A) \geq at).$$

Comme cela tient pour tout $A \in \mathcal{A}$, nous avons également que

$$\begin{aligned} \mathbb{1}\{\sup_{A \in \mathcal{A}} \mu_n(A) - \mu(A) \geq t\} &= \sup_{A \in \mathcal{A}} \mathbb{1}\{\mu_n(A) - \mu(A) \geq t\} \leq 2 \sup_{A \in \mathcal{A}} P'(\mu_n(A) - \mu'_n(A) \geq at) \\ &\leq 2P'(\sup_{A \in \mathcal{A}} (\mu_n(A) - \mu'_n(A)) \geq at). \end{aligned}$$

En prenant l'espérance par rapport à l'échantillon $X_1, \dots, X_n$, il vient alors

$$P\left(\sup_{A \in \mathcal{A}} (\mu_n(A) - \mu(A)) \geq t\right) \leq 2P\left(\sup_{A \in \mathcal{A}} (\mu_n(A) - \mu'_n(A)) \geq at\right).$$

L'autre inégalité se démontre de manière similaire. ■

Un autre ingrédient utilisé pour démontrer l'inégalité VC consiste en une égalité en distribution, notée $=_d$, dont la démonstration est laissée au lecteur. En utilisant les mêmes notations que dans le lemme de symétrisation,

$$\sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}) =_d \sum_{i=1}^n \sigma_i (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}), \quad (2.8)$$

où nous avons introduit des *variables signes* (aussi appelées *variables de Rademacher*), i.e. une suite $\sigma_1, \dots, \sigma_n$ de variables aléatoires i.i.d. et indépendantes des échantillons $X_1, \dots, X_n$ et $X'_1, \dots, X'_n$ telles que

$$\sigma_i = \begin{cases} +1 & \text{avec probabilité } 1/2, \\ -1 & \text{avec probabilité } 1/2. \end{cases}$$

Ce procédé correspond à ce que nous appelons *randomisation*.

Démonstration du théorème 2.14. Nous supposons $nt^2 \geq 2$, car dans l'autre cas, la borne proposée dans l'inégalité VC excède 1 et est triviale. Ainsi, par le lemme de symétrisation appliqué avec $a = 1/2$ et par randomisation, nous avons

$$\begin{aligned} P\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t\right) &\leq 4P\left(\sup_{A \in \mathcal{A}} (\mu_n(A) - \mu'_n(A)) \geq t/2\right) \\ &= 4P\left(\sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}) \geq t/2\right) \end{aligned}$$

$$= 4 \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}) \geq t/2 \right).$$

Fixons $X_1 = x_1, \dots, X_n = x_n, X'_1 = x'_1, \dots, X'_n = x'_n$. Nous remarquons que le supremum est forcément atteint dans une sous-classe $\tilde{\mathcal{A}}$ de $\mathcal{A}$ qui reprend l'ensemble des éléments $A \in \mathcal{A}$ tels que toutes les intersections possibles entre $\{x_1, \dots, x_n, x'_1, \dots, x'_n\}$ et un élément de $\mathcal{A}$ soient représentées. Forcément, $|\tilde{\mathcal{A}}| \leq \mathbb{S}_{\mathcal{A}}(2n)$. Ainsi,

$$\begin{aligned} & \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\}) \geq t/2 \right) \\ &= \mathbb{P} \left(\max_{A \in \tilde{\mathcal{A}}} \frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\}) \geq t/2 \right) \\ &\leq \sum_{A \in \tilde{\mathcal{A}}} \mathbb{P} \left(\sum_{i=1}^n \sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\}) \geq nt/2 \right). \end{aligned}$$

Les variables aléatoires $\sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\})$ étant de moyenne nulle, à valeurs dans $[-1, 1]$ et indépendantes, nous pouvons appliquer l'inégalité de Hoeffding (théorème 1.8) pour obtenir,

$$\mathbb{P} \left(\sum_{i=1}^n \sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\}) \geq nt/2 \right) \leq e^{-nt^2/8}, \text{ pour tout } A \in \tilde{\mathcal{A}}.$$

En regroupant ces résultats, nous trouvons l'inégalité VC par un argument de conditionnement

$$\begin{aligned} & \mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) \\ &\leq 4 \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}) \geq t/2 \right) \\ &= 4 \mathbb{E} \left[\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}) \geq t/2 \mid X_1, \dots, X_n, X'_1, \dots, X'_n \right) \right] \\ &\leq 4 \mathbb{E} \left[\mathbb{S}_{\mathcal{A}}(2n) e^{-nt^2/8} \right] = 4 \mathbb{S}_{\mathcal{A}}(2n) e^{-nt^2/8}. \end{aligned}$$

Une application simple du lemme d'inversion sur (2.6) fournit le résultat. ■

Exemple 2.16 (Fonction de répartition empirique, théorème de Glivenko–Cantelli). Dans le cas de la fonction de répartition empirique (nous rappelons de l'exemple 2.4 que $V = 1$), nous obtenons gratuitement le théorème de Glivenko–Cantelli [10, théorème 12.4]

pour la statistique du test de Kolmogorov–Smirnov. Celui-ci assure que

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \rightarrow 0 \quad \text{p.s.}$$

L'inégalité VC implique directement ce résultat via le lemme de Borel–Cantelli (voir [26, Lemme 5.3.1]) car, pour tout $t > 0$, la série

$$\sum_{n=1}^{+\infty} \mathbb{P} \left(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \geq t \right) \leq \sum_{n=1}^{+\infty} 4(2n+1)e^{-nt^2/8}$$

est convergente, et ainsi

$$\mathbb{P} \left(\limsup_{n \rightarrow +\infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \geq t \right) = 0,$$

pour tout $t > 0$. Autrement dit,

$$\mathbb{P} \left(\limsup_{n \rightarrow +\infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0 \right) = 1.$$

La convergence de la série est en fait rendue encore bien plus aisée par le résultat suivant dû à Massart [22] qui affirme que

$$\mathbb{P} \left(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \geq t \right) \leq 2e^{-2nt^2},$$

pour tout $t > 0$. Cette inégalité est bien meilleure que celle induite par l'inégalité VC et nous ne discutons pas sa démonstration.

Comme annoncé précédemment, l'avantage d'avoir démontré un lemme de symétrisation plus général que celui classiquement rencontré est qu'il est maintenant possible de renforcer l'inégalité VC (théorème 2.14). Ceci est formalisé dans le théorème suivant, apparemment nouveau, dû à l'auteur et à Johan Segers.

Théorème 2.17. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Alors, pour tout $\delta \in (0, 1)$, nous avons avec probabilité $1 - \delta$,*

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{1}{2n}} + \sqrt{\frac{2}{n} [\log(4/\delta) + \log(\mathbb{S}_{\mathcal{A}}(2n))]}, \quad (2.9)$$

ou, de manière équivalente, nous avons pour tout $t > 0$,

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) \leq 4\mathbb{S}_{\mathcal{A}}(2n) \exp \left\{ -\frac{nt^2}{2} \left(1 - \sqrt{\frac{2}{nt^2}} \right) - \frac{1}{4} \right\}.$$

Le résultat obtenu traduit le fait que l'exponentielle se comporte comme $e^{-nt^2/2} = (e^{-nt^2/8})^4$ quand nt^2 est suffisamment grand et non comme $e^{-nt^2/8}$. Cela représente une grande amélioration en pratique (voir section 2.3).

Démonstration du théorème 2.17. Nous commençons par démontrer la formulation du résultat comme une inégalité de concentration. Nous avons démontré que pour tout $t > 0$,

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

En particulier, nous avons remarqué que si $nt^2 < 2$, la borne était triviale et nous avons donc travaillé sous l'hypothèse $nt^2 \geq 2$. Grâce à notre lemme de symétrisation, si pour $a \in (0, 1)$ nous avons $4(1-a)^2nt^2 \geq 2$, nous pouvons montrer de la même manière que

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-a^2nt^2/2}. \quad (2.10)$$

Si $a > 1/2$, cette borne est meilleure car nous améliorons la vitesse de convergence de l'exponentielle vers zéro. Toutefois, le prix à payer est que la condition $4(1-a)^2nt^2 \geq 2$ devient plus difficile à satisfaire. Étant donnés n et t , nous pouvons choisir le meilleur (i.e. le plus grand) $a \in (0, 1)$ tel que la condition $4(1-a)^2nt^2 \geq 2$ est respectée. Il est donné par

$$a = a(n, t) = 1 - \sqrt{\frac{1}{2nt^2}}, \quad (2.11)$$

qui est positif à condition que $nt^2 \geq 1/2$. En utilisant ce choix de a dans la borne (2.10) et le fait que la borne est triviale si $nt^2 < 1/2$, nous trouvons la borne annoncée.

La formulation alternative s'obtient maintenant aisément. En effet, il suffit d'appliquer le lemme d'inversion (lemme 2.13) qui garantit qu'avec probabilité $1 - \delta$,

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq t(\delta),$$

où $t(\delta) > 0$ est solution de

$$4\mathbb{S}_{\mathcal{A}}(2n) \exp \left\{ -\frac{nt(\delta)^2}{2} \left(1 - \sqrt{\frac{2}{nt(\delta)^2}} \right) - \frac{1}{4} \right\} = \delta.$$

Cette dernière équation se réécrit simplement comme une équation quadratique en $t(\delta)$

$$\frac{n}{2}t(\delta)^2 - \sqrt{\frac{n}{2}}t(\delta) + \left[\frac{1}{4} - \log(4/\delta) - \log(\mathbb{S}_{\mathcal{A}}(2n)) \right] = 0.$$

La solution positive de cette équation correspond à la borne supérieure proposée dans (2.9). ■

Nous présentons maintenant une alternative pour les déviations relatives, dont la preuve est due à Anthony et Shawe–Taylor [1, Théorème 2.1] (bien que cette dernière n’était en réalité pas complète puisque l’inégalité provenant de [14, Théorème 1] était utilisée sans preuve de sa validité).

Théorème 2.18. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Alors, pour $\delta \in (0, 1)$, nous avons avec probabilité $1 - \delta$,*

$$\left. \begin{aligned} \sup_{A \in \mathcal{A}} \frac{\mu_n(A) - \mu(A)}{\sqrt{\mu_n(A)}} \\ \sup_{A \in \mathcal{A}} \frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \end{aligned} \right\} \leq 2\sqrt{\frac{1}{n} [\log(4/\delta) + \log(\mathbb{S}_{\mathcal{A}}(2n))]}.$$

Ce résultat est fort utile dans le cas où nous disposons d’informations à priori sur la distribution μ et la classe $\mathcal{A}$. Par exemple, supposons que nous sachions d’avance que pour tout $A \in \mathcal{A}$ nous avons $\mu(A) \leq p$, pour un certain $p \in (0, 1)$ fixé. Alors la deuxième inégalité du théorème 2.18 assure qu’avec probabilité $1 - \delta$,

$$\sup_{A \in \mathcal{A}} (\mu(A) - \mu_n(A)) \leq 2\sqrt{\frac{p}{n} [\log(4/\delta) + \log(\mathbb{S}_{\mathcal{A}}(2n))]} \quad (2.12)$$

Si p est très petit, cela peut largement améliorer la borne fournie par l’inégalité VC. Ce type de problème sera traité en détails dans le chapitre suivant.

Le théorème 2.18 est souvent présenté sous sa forme (équivalente) d’inégalités de concentration : pour tout $t > 0$,

$$\left. \begin{aligned} \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{\mu_n(A) - \mu(A)}{\sqrt{\mu_n(A)}} \geq t \right) \\ \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \geq t \right) \end{aligned} \right\} \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-nt^2/4}. \quad (2.13)$$

La preuve utilise le même raisonnement que celui développé dans la démonstration de l’inégalité VC, en se basant sur des variantes du lemme de symétrisation, dont nous effectuons la démonstration en détails dans le cas de la seconde inégalité.

Démonstration. Considérons la deuxième inégalité de (2.13). Nous supposons $nt^2 \geq 2$ car la borne est triviale dans les autres cas. Commençons par montrer la symétrisation associée, que nous énonçons en réutilisant les mêmes notations que dans le lemme 2.15 : pour tout $t > 0$,

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \geq t \right) \leq 4 \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{\mu'_n(A) - \mu_n(A)}{\sqrt{(1/2)(\mu'_n(A) + \mu_n(A))}} \geq t \right). \quad (2.14)$$

L'inégalité suit du fait que pour tout $t > 0$ et $A \in \mathcal{A}$,

$$\mathbb{1}\left\{\frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \geq t\right\} \mathbb{1}\{\mu'_n(A) \geq \mu(A)\} \leq \mathbb{1}\left\{\frac{\mu'_n(A) - \mu_n(A)}{\sqrt{(1/2)(\mu'_n(A) + \mu_n(A))}} \geq t\right\}. \quad (2.15)$$

En effet, pour que le produit des indicatrices dans le membre de gauche soit 1, il faut que les deux conditions des indicatrices soient satisfaites simultanément,

$$\begin{cases} \frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \geq t, \\ \mu'_n(A) \geq \mu(A). \end{cases}$$

La première condition implique forcément que $\mu_n(A) \leq \mu(A)$ car $t > 0$. Ainsi, en utilisant le fait que la fonction $x \geq 0 \mapsto (x - a)/\sqrt{x + a}$ est croissante, pour n'importe quel $a \geq 0$, cela implique forcément que

$$\frac{\mu'_n(A) - \mu_n(A)}{\sqrt{(1/2)(\mu'_n(A) + \mu_n(A))}} \geq \frac{\mu(A) - \mu_n(A)}{\sqrt{(1/2)(\mu(A) + \mu_n(A))}} \geq \frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \geq t,$$

autrement dit, il faut que l'indicatrice du membre de droite dans (2.15) soit 1 également. En prenant l'espérance par rapport à l'échantillon fantôme, il vient

$$\mathbb{1}\left\{\frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \geq t\right\} P'(\mu'_n(A) \geq \mu(A)) \leq P'\left(\frac{\mu'_n(A) - \mu_n(A)}{\sqrt{(1/2)(\mu'_n(A) + \mu_n(A))}} \geq t\right).$$

Remarquons que, pour que l'indicatrice dans le membre de gauche vaille 1, il faut $\frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \geq t$, ce qui implique que $\mu(A) \geq t^2$. Dès lors,

$$n \geq \frac{1}{t^2} \geq \frac{1}{\mu(A)}.$$

Comme la variable aléatoire $n\mu'_n(A)$ est de loi binomiale de paramètre $(n, \mu(A))$, ce que nous notons $n\mu'_n(A) \sim \text{Bin}(n, \mu(A))$, nous pouvons appliquer le [14, Théorème 1] qui assure que

$$P'(\mu'_n(A) \geq \mu(A)) = P'(n\mu'_n(A) \geq n\mu(A)) \geq \frac{1}{4}.$$

Dès lors, en réarrangeant les termes, en passant au supremum sur $A \in \mathcal{A}$ et en prenant l'espérance par rapport à $X_1, \dots, X_n$ comme dans la démonstration du lemme de symétrisation (lemme 2.15) nous obtenons (2.14).

En randomisant comme dans la démonstration de l'inégalité VC nous avons

$$P\left(\sup_{A \in \mathcal{A}} \frac{\mu'_n(A) - \mu_n(A)}{\sqrt{(1/2)(\mu'_n(A) + \mu_n(A))}} \geq t\right)$$

$$\begin{aligned}
&= \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{\frac{1}{n} \sum_{i=1}^n (\mathbb{1}\{X'_i \in A\} - \mathbb{1}\{X_i \in A\})}{\sqrt{(1/2)(\mu'_n(A) + \mu_n(A))}} \geq t \right) \\
&= \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbb{1}\{X'_i \in A\} - \mathbb{1}\{X_i \in A\})}{\sqrt{(1/2)(\mu'_n(A) + \mu_n(A))}} \geq t \right).
\end{aligned}$$

Fixons $X_1 = x_1, \dots, X_n = x_n, X'_1 = x'_1, \dots, X'_n = x'_n$. Comme précédemment, le supremum est forcément atteint dans une sous-classe $\tilde{\mathcal{A}}$ de $\mathcal{A}$ telle que $|\tilde{\mathcal{A}}| \leq \mathbb{S}_{\mathcal{A}}(2n)$. Par conséquent, en utilisant l'inégalité de Hoeffding, nous obtenons

$$\begin{aligned}
&\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbb{1}\{x'_i \in A\} - \mathbb{1}\{x_i \in A\})}{\sqrt{\frac{1}{2n} \sum_{j=1}^n (\mathbb{1}\{x'_j \in A\} + \mathbb{1}\{x_j \in A\})}} \geq t \right) \\
&\leq \sum_{A \in \tilde{\mathcal{A}}} \mathbb{P} \left(\sum_{i=1}^n \sigma_i (\mathbb{1}\{x'_i \in A\} - \mathbb{1}\{x_i \in A\}) \geq t \left[\frac{n}{2} \sum_{j=1}^n (\mathbb{1}\{x'_j \in A\} + \mathbb{1}\{x_j \in A\}) \right]^{1/2} \right) \\
&\leq \sum_{A \in \tilde{\mathcal{A}}} \exp \left\{ - \frac{nt^2 \sum_{j=1}^n (\mathbb{1}\{x'_j \in A\} + \mathbb{1}\{x_j \in A\})}{4 \sum_{j=1}^n (\mathbb{1}\{x'_j \in A\} - \mathbb{1}\{x_j \in A\})^2} \right\} \leq \mathbb{S}_{\mathcal{A}}(2n) e^{-nt^2/4}.
\end{aligned}$$

En conditionnant comme dans la preuve du théorème 2.14 et en appliquant le lemme d'inversion, nous arrivons au résultat.

La première inégalité se base sur une symétrisation légèrement différente, dont nous laissons la démonstration aux soins du lecteur. ■

Nous ne pourrions clore cette partie sans mentionner une inégalité particulièrement forte due à Michel Talagrand [27, Théorème 1.1] qui assure que, si $\mathcal{A}$ est de dimension VC finie, il existe une constante $C > 0$ telle que

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) \leq \frac{C}{t\sqrt{n}} \left(\frac{Cnt^2}{V} \right)^V e^{-2nt^2}.$$

Néanmoins, il n'existe, à la connaissance de l'auteur, pas d'estimation uniforme indépendante de t pour la constante C . Dès lors, dans un souhait de rester explicite, nous ne ferons pas usage de cette inégalité dans la suite.

2.2.2 Bornes provenant de l'espérance

Rappelons qu'au début de la section, nous avons vu (2.5) qu'avec probabilité $1 - \delta$,

$$Z = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{1}{2n} \log(1/\delta)} + \mathbb{E}[Z],$$

où $\delta \in (0, 1)$. Par conséquent, il est naturel de vouloir chercher des bornes sur $\mathbb{E}[Z]$. Bien évidemment, toutes les inégalités obtenues précédemment peuvent fournir de telles bornes via la formule bien connue

$$\mathbb{E}[Z] = \int_0^{+\infty} \mathbb{P}(Z \geq t) dt,$$

toutefois, les inégalités ainsi obtenues ne sont pas aussi bonnes que celles que nous proposons dans la suite.

Commençons par un lemme simple ([15, Lemme 1.3]) qui permet de borner l'espérance d'un maximum entre plusieurs variables aléatoires, potentiellement dépendantes, quand nous disposons d'un contrôle sur la fonction génératrice des moments.

Lemme 2.19. *Soient $\sigma > 0$, $m \geq 2$ et $Y_1, \dots, Y_m$ des variables aléatoires réelles. Si pour tout $1 \leq i \leq m$ et pour tout $\lambda > 0$,*

$$m_{Y_i}(\lambda) = \mathbb{E}[e^{\lambda Y_i}] \leq e^{\lambda^2 \sigma^2 / 2},$$

alors,

$$\mathbb{E}\left[\max_{1 \leq i \leq m} Y_i\right] \leq \sigma \sqrt{2 \log(m)}.$$

Si de plus, pour tout $\lambda > 0$ et $1 \leq i \leq m$, $m_{-Y_i}(\lambda) \leq e^{\lambda^2 \sigma^2 / 2}$, nous avons également que pour tout $m \geq 1$,

$$\mathbb{E}\left[\max_{1 \leq i \leq m} |Y_i|\right] \leq \sigma \sqrt{2 \log(2m)}.$$

Démonstration. Il suit de l'inégalité de Jensen que pour tout $\lambda > 0$,

$$e^{\lambda \mathbb{E}[\max_{1 \leq i \leq m} Y_i]} \leq \mathbb{E}[e^{\lambda \max_{1 \leq i \leq m} Y_i}] = \mathbb{E}\left[\max_{1 \leq i \leq m} e^{\lambda Y_i}\right] \leq \sum_{i=1}^m \mathbb{E}[e^{\lambda Y_i}] \leq m e^{\lambda^2 \sigma^2 / 2}.$$

Dès lors,

$$\mathbb{E}\left[\max_{1 \leq i \leq m} Y_i\right] \leq \frac{\log(m)}{\lambda} + \frac{\lambda \sigma^2}{2}.$$

Minimisant la borne sur $\lambda > 0$, nous trouvons la première inégalité (le minimum est atteint en $\lambda = \sqrt{2 \log(m)}/\sigma$).

La deuxième inégalité suit du fait que

$$\max_{1 \leq i \leq m} |Y_i| = \max\{Y_1, -Y_1, \dots, Y_m, -Y_m\}$$

et de la première inégalité. ■

Nous proposons une première borne sur l'espérance $E[Z]$. Ce résultat est parfois également appelé « inégalité de Vapnik–Chervonenkis » (voir e.g. [15, Théorème 1.9]).

Théorème 2.20 (Inégalité VC pour la moyenne). *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Alors,*

$$E[Z] = E \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}}.$$

La démonstration est en fait elle aussi assez similaire à celle de l'inégalité VC. L'avantage est que, pour l'espérance, la symétrisation est plus simple. De plus, l'inégalité obtenue en combinaison avec (2.5) est meilleure que l'inégalité VC pour des échantillons de taille raisonnable (voir section 2.3).

Démonstration. Nous commençons par introduire un échantillon fantôme $X'_1, \dots, X'_n$ (toujours supposé indépendant de l'échantillon de base) et des variables de Rademacher $\sigma_1, \dots, \sigma_n$ (toujours supposées indépendantes des deux échantillons) qui nous permettront d'effectuer la symétrisation et la randomisation (comme expliqué auparavant). Si μ'_n désigne la mesure empirique associée à l'échantillon fantôme, nous avons

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

$$= \frac{1}{n} \mathbb{E} \left[\mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}) \right| \middle| X_1, \dots, X_n, X'_1, \dots, X'_n \right] \right].$$

Fixons $X_1 = x_1, \dots, X_n = x_n, X'_1 = x'_1, \dots, X'_n = x'_n$ et définissons la classe $\tilde{\mathcal{A}}$ comme dans la démonstration du théorème 2.14. Le supremum est forcément atteint par un élément de $\tilde{\mathcal{A}}$, dès lors

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

Pour tout $A \in \tilde{\mathcal{A}}$, les variables aléatoires $\sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\})$ étant de moyenne nulle, à valeurs dans $[-1, 1]$ et indépendantes, nous avons par lemme de Hoeffding (lemme 1.7) que pour tout $\lambda > 0$

$$\mathbb{E} \left[e^{\lambda \sum_{i=1}^n \sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\})} \right] = \prod_{i=1}^n \mathbb{E} \left[e^{\lambda \sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\})} \right] \leq e^{n\lambda^2/2}.$$

Comme la distribution de $\sum_{i=1}^n \sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\})$ est symétrique et $|\tilde{\mathcal{A}}| \leq \mathbb{S}_{\mathcal{A}}(2n)$, il suit du lemme 2.19 que

$$\mathbb{E} \left[\max_{A \in \tilde{\mathcal{A}}} \left| \sum_{i=1}^n \sigma_i (\mathbb{1}\{x_i \in A\} - \mathbb{1}\{x'_i \in A\}) \right| \right] \leq \sqrt{2n \log(2\mathbb{S}_{\mathcal{A}}(2n))}.$$

Ainsi, par le raisonnement précédent,

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \frac{1}{n} \sqrt{2n \log(2\mathbb{S}_{\mathcal{A}}(2n))} = \sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}}. \quad \blacksquare$$

De la même manière que le théorème 2.18 permettait de fournir une borne plus fine si nous disposions à priori d'information sur la classe $\mathcal{A}$, nous pouvons également obtenir un résultat similaire pour l'espérance (nous exposons une inégalité légèrement meilleure que [15, Théorème 1.10]).

Théorème 2.21. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Notons $\Sigma = \sup_{A \in \mathcal{A}} \sqrt{\mu(A)(1 - \mu(A))}$. Alors,*

$$\mathbb{E}[Z] = \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \frac{8 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n} + \Sigma \sqrt{\frac{32 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}}.$$

Cette inégalité est intéressante si nous savons à l'avance que $\mu(A) \leq p$ pour tout $A \in \mathcal{A}$, où p est très petit (ce qui sera typiquement le cas dans le chapitre suivant). Dans ce cas Σ sera petit également, et le théorème précédent peut être vu comme une interpolation

entre un « bon cas » (décroissance en $1/n$) et un « mauvais cas » (décroissance en $1/\sqrt{n}$) dont l'influence sera atténuée par la présence du Σ .

Gabor Lugosi dit de ce théorème [15, page 15] qu'il représente une amélioration significative du théorème 2.20 quand Σ est très petit. Nous verrons que les techniques développées dans le chapitre 3 permettent d'obtenir une amélioration encore bien plus significative (voir section 3.4).

Démonstration. Le début de la démonstration est identique à celle de l'inégalité VC pour la moyenne et n'est pas répété. Nous partons du fait que

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

Au lieu d'appliquer le lemme de Hoeffding en invoquant que, pour $X_1, \dots, X_n, X'_1, \dots, X'_n$ et $A \in \mathcal{A}$ supposés fixés, les variables aléatoires $\sigma_i (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\})$ sont i.i.d. à valeurs dans $[-1, 1]$, nous pouvons être plus fins et directement utiliser le fait que

$$\sigma_i (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}) \in \left[-(\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}), \mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\} \right].$$

Ainsi, pour $A \in \mathcal{A}$ fixé et pour tout $\lambda > 0$, le lemme de Hoeffding implique

$$\mathbb{E} \left[e^{\lambda \sum_{i=1}^n \sigma_i (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\})} \middle| X_1, \dots, X_n, X'_1, \dots, X'_n \right] \leq e^{\lambda^2 \sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\})^2 / 2}.$$

Comme le supremum est atteint sur la classe $\tilde{\mathcal{A}}$ de cardinalité $|\tilde{\mathcal{A}}| \leq \mathbb{S}_{\mathcal{A}}(2n)$, nous obtenons par le lemme 2.19,

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \frac{1}{n} \mathbb{E} \left[\sup_{A \in \mathcal{A}} \sqrt{\sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\})^2} \right] \sqrt{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}. \quad (2.16)$$

En utilisant l'inégalité de Jensen, nous pouvons borner l'espérance dans le membre de droite de la manière suivante,

$$\begin{aligned} & \mathbb{E} \left[\sup_{A \in \mathcal{A}} \sqrt{\sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\})^2} \right] \\ & \leq \sqrt{\mathbb{E} \left[\sup_{A \in \mathcal{A}} \sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\})^2 \right]} \\ & = \sqrt{\mathbb{E} \left[\sup_{A \in \mathcal{A}} \sum_{i=1}^n \left\{ (\mathbb{1}\{X_i \in A\} - \mu(A)) + (\mu(A) - \mathbb{1}\{X'_i \in A\}) \right\}^2 \right]} \end{aligned}$$

$$\begin{aligned}
&\leq \sqrt{4 \mathbb{E} \left[\sup_{A \in \mathcal{A}} \sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mu(A))^2 \right]} \\
&\leq 2\Sigma\sqrt{n} + 2\sqrt{\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mu(A))| \right]} \\
&= 2\Sigma\sqrt{n} + 2\sqrt{n \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right]},
\end{aligned}$$

En combinant cela avec (2.16), nous obtenons

$$\mathbb{E}[Z] \leq 2 \left(\Sigma + \sqrt{\mathbb{E}[Z]} \right) \sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}}.$$

En posant $m = \sqrt{\mathbb{E}[Z]}$, cette dernière inégalité peut se réécrire

$$m^2 - 2\sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}}m - 2\Sigma\sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}} \leq 0. \quad (2.17)$$

Cela implique en particulier que

$$m = \sqrt{\mathbb{E}[Z]} \leq m_+,$$

où m_+ est la solution positive de l'équation du second degré associée à l'inégalité (2.17). Nous calculons

$$m_+ = \sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}} + \sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n} + 2\Sigma\sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}}}.$$

Dès lors,

$$\begin{aligned}
\mathbb{E}[Z] &\leq m_+^2 \\
&\leq 2 \left(\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n} + \frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n} + 2\Sigma\sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}} \right) \\
&= \frac{8 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n} + \Sigma\sqrt{\frac{32 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}},
\end{aligned}$$

où nous avons utilisé le fait que pour tout $a, b \in \mathbb{R}$, nous avons $(a + b)^2 \leq 2(a^2 + b^2)$. ■

Les théorèmes que nous venons de démontrer assurent que

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] = O \left(\sqrt{\frac{\log(\mathbb{S}_{\mathcal{A}}(2n))}{n}} \right), \quad \text{quand } n \rightarrow +\infty.$$

En particulier, si $\mathcal{A}$ est de dimension VC finie, le corollaire 2.8 garantit

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] = O \left(\sqrt{\frac{\log(n)}{n}} \right), \quad \text{quand } n \rightarrow +\infty.$$

Nous allons montrer comment, en utilisant le nombre couvrant (définition 2.11) et la représentation du coefficient de désintégration dans l'hypercube, il est possible d'obtenir également

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] = O \left(\frac{1}{\sqrt{n}} \right),$$

un résultat qui, asymptotiquement, est certainement meilleur que ceux présentés jusque-là.

Nous commençons par démontrer, en utilisant le chaînage, une borne sur $\mathbb{E}[Z]$ qui fait intervenir le nombre couvrant (une version améliorée de l'inégalité classique [15, Théorème 1.16] due à Johan Segers). Ensuite, nous utilisons l'inégalité de Haussler (théorème 2.12).

Théorème 2.22. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Pour tout $a > 1$, nous avons*

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \frac{4}{\sqrt{n}} \frac{a^2}{a-1} \max_{x_1, \dots, x_n \in \mathbb{R}^d} \int_0^{a^{-1}} \sqrt{\log \left(\sqrt{2} \mathcal{N}(r, \mathcal{A}(x_1^n)) \right)} dr.$$

Cela représente une amélioration de [15, Théorème 1.16]

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \frac{24}{\sqrt{n}} \max_{x_1, \dots, x_n \in \mathbb{R}^d} \int_0^1 \sqrt{\log \left(2 \mathcal{N}(r, \mathcal{A}(x_1^n)) \right)} dr$$

car le facteur 24 devant l'intégrale a été réduit à

$$4 \inf_{a>1} \frac{a^2}{a-1} = 16,$$

la borne supérieure de l'intégrale a été réduite à $a^{-1} < 1$ et le facteur 2 dans l'intégrale est devenu $\sqrt{2}$.

Démonstration. Le début de la preuve commence comme l'inégalité VC pour la moyenne. Nous supposons donc connu que

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \frac{1}{n} \mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i (\mathbf{1}\{X_i \in A\} - \mathbf{1}\{X'_i \in A\}) \right| \right].$$

Il suit que

$$\begin{aligned}
\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] &\leq \frac{1}{n} \mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i (\mathbb{1}\{X_i \in A\}) \right| \right] + \frac{1}{n} \mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i (\mathbb{1}\{X'_i \in A\}) \right| \right] \\
&= \frac{2}{n} \mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i (\mathbb{1}\{X_i \in A\}) \right| \right] \\
&= \frac{2}{n} \mathbb{E} \left[\mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i (\mathbb{1}\{X_i \in A\}) \right| \middle| X_1, \dots, X_n \right] \right],
\end{aligned}$$

où la deuxième ligne vient du fait que $X_1, \dots, X_n, X'_1, \dots, X'_n$ est un échantillon i.i.d. de taille $2n$ de distribution commune μ .

Fixons $X_1 = x_1, \dots, X_n = x_n, X'_1 = x'_1, \dots, X'_n = x'_n$. Par définition de $\mathcal{A}(x_1^n)$, nous avons

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \sigma_i (\mathbb{1}\{x_i \in A\}) \right| \right] = \mathbb{E} \left[\max_{b \in \mathcal{A}(x_1^n)} \left| \sum_{i=1}^n \sigma_i b_i \right| \right]. \quad (2.18)$$

Fixons $a > 1$. Soit $M = \lfloor \log_a(\sqrt{n}) \rfloor + 1$, où $\lfloor x \rfloor$ désigne le plus grand entier n tel que $n \leq x$. Nous définissons les sous-ensembles de $\{0, 1\}^n$ suivants : $B_0 = \{(0, \dots, 0)\}$ et B_k un recouvrement (définition 2.10) minimal de rayon a^{-k} de $\mathcal{A}(x_1^n)$. Observons que, comme la distance de Hamming (définition 2.9) est toujours plus petite que 1, B_0 est un recouvrement minimal de $\mathcal{A}(x_1^n)$ de rayon $a^0 = 1$. De plus, comme la distance de Hamming est toujours plus grande que $1/\sqrt{n}$ pour deux points distincts, nous avons également que $B_M = \mathcal{A}(x_1^n)$.

Soit $\sigma = (\sigma_1, \dots, \sigma_n)$ une réalisation du vecteur aléatoire de variables de Rademacher et soit $b^* = b^*(\sigma)$ le vecteur de $\{0, 1\}^n$ associé tel que le maximum dans (2.18) est atteint. Définissons $b^{(M)}, b^{(M-1)}, \dots, b^{(0)}$ de manière récursive, comme suit (notons que ces quantités dépendent également de σ) :

$$\begin{cases} b^{(M)} &= b^* \\ b^{(k-1)} &\in \operatorname{argmin}\{\rho(b, b^{(k)}) : b \in B_{k-1}\}, \quad k = M, \dots, 1. \end{cases}$$

Notons que $b^{(0)} = (0, \dots, 0)$. Par définition de B_{k-1} , nous avons également

$$\rho(b^{(k)}, b^{(k-1)}) \leq a^{-(k-1)}, \quad \forall k = 1, \dots, M.$$

Nous pouvons alors procéder au chaînage, de la manière suivante :

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

Par conséquent,

$$\begin{aligned}
\mathbb{E} \left[\max_{b \in \mathcal{A}(x_1^n)} \left| \sum_{i=1}^n \sigma_i b_i \right| \right] &= \mathbb{E} \left[\left| \sum_{k=1}^M \sum_{i=1}^n \sigma_i (b_i^{(k)} - b_i^{(k-1)}) \right| \right] \\
&\leq \sum_{k=1}^M \mathbb{E} \left[\left| \sum_{i=1}^n \sigma_i (b_i^{(k)} - b_i^{(k-1)}) \right| \right] \\
&\leq \sum_{k=1}^M \mathbb{E} \left[\max_{\substack{b \in B_k, \ c \in B_{k-1} \\ \rho(b,c) \leq a^{-k+1}}} \left| \sum_{i=1}^n \sigma_i (b_i - c_i) \right| \right].
\end{aligned}$$

Par le lemme de Hoeffding, nous avons que pour tout $\lambda > 0$ et pour tout $(b, c) \in B_k \times B_{k-1}$ tels que $\rho(b, c) \leq a^{-k+1}$,

$$\mathbb{E} \left[e^{\lambda \sum_{i=1}^n \sigma_i (b_i - c_i)} \right] = \prod_{i=1}^n \mathbb{E} \left[e^{\lambda \sigma_i (b_i - c_i)} \right] \leq \prod_{i=1}^n e^{\lambda^2 (b_i - c_i)^2 / 2} = e^{\lambda^2 n \rho(b,c)^2 / 2} \leq e^{\lambda^2 n a^{-2k+2} / 2}.$$

De plus, il est évident que $|B_k| |B_{k-1}| \leq |B_k|^2 = \mathcal{N}(a^{-k}, \mathcal{A}(x_1^n))^2$. Il suit alors du lemme 2.19 que

$$\begin{aligned}
\mathbb{E} \left[\max_{\substack{b \in B_k, \ c \in B_{k-1} \\ \rho(b,c) \leq a^{-k+1}}} \left| \sum_{i=1}^n \sigma_i (b_i - c_i) \right| \right] &\leq \sqrt{n} a^{-k+1} \sqrt{2 \log(2 \mathcal{N}(a^{-k}, \mathcal{A}(x_1^n))^2)} \\
&= 2\sqrt{n} a^{-k+1} \sqrt{\log(\sqrt{2} \mathcal{N}(a^{-k}, \mathcal{A}(x_1^n)))}.
\end{aligned}$$

Si nous définissons

$$f : r \in (0, 1] \rightarrow f(r) = \sqrt{\log(\sqrt{2} \mathcal{N}(r, \mathcal{A}(x_1^n)))} \in [0, +\infty),$$

nous obtenons

$$\mathbb{E} \left[\max_{b \in \mathcal{A}(x_1^n)} \left| \sum_{i=1}^n \sigma_i b_i \right| \right] \leq \sum_{k=1}^M 2\sqrt{n} a^{-k+1} f(a^{-k}) \leq 2a\sqrt{n} \sum_{k=1}^{+\infty} a^{-k} f(a^{-k}). \quad (2.19)$$

Comme la fonction f est décroissante, il suit que

$$\begin{aligned}
\sum_{k=1}^{+\infty} a^{-k} f(a^{-k}) &= \sum_{k=1}^{+\infty} \frac{a^{-k} - a^{-k-1}}{1 - a^{-1}} f(a^{-k}) \\
&= \frac{1}{1 - a^{-1}} \sum_{k=1}^{+\infty} \int_{a^{-k-1}}^{a^{-k}} f(a^{-k}) \, dr \\
&\leq \frac{1}{1 - a^{-1}} \sum_{k=1}^{+\infty} \int_{a^{-k-1}}^{a^{-k}} f(r) \, dr
\end{aligned}$$

$$= \frac{a}{a-1} \int_0^{a^{-1}} f(r) dr.$$

Par les calculs précédents, il vient alors

$$\mathbb{E} \left[\max_{b \in \mathcal{A}(x_1^n)} \left| \sum_{i=1}^n \sigma_i b_i \right| \right] \leq 2\sqrt{n} \frac{a^2}{a-1} \int_0^{a^{-1}} f(r) dr.$$

En multipliant ce résultat par $2/n$ et en passant au maximum sur $x_1, \dots, x_n \in \mathbb{R}^d$, nous obtenons la conclusion. $\blacksquare$

Si la dimension VC de la classe $\mathcal{A}$ est finie, nous avons le corollaire suivant.

Corollaire 2.23. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Si $\mathcal{A}$ est de dimension VC finie, nous avons*

$$\mathbb{E}[Z] = \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq 15 \sqrt{\frac{V}{n}}.$$

Bien que ce résultat semble meilleur que tout ce que nous avons présenté jusque-là, il ne donne pas forcément la plus petite borne en pratique. Si nous ne disposons que d'un « petit échantillon », il est tout à fait possible que les résultats précédents soient plus performants (et nous le constaterons d'ailleurs sur un exemple dans la section 2.3).

Intuitivement, ce corollaire n'est pas surprenant : le théorème 2.22 a permis d'obtenir une borne sur l'espérance faisant intervenir le nombre couvrant (et plus le coefficient de désintégration, comme dans l'inégalité VC pour la moyenne). De plus, quand $V < +\infty$, la borne de Haussler (théorème 2.12) sur ce nombre couvrant est indépendante de n (à l'inverse de la borne obtenue sur le coefficient de désintégration dans le corollaire 2.8). Il n'est alors pas étonnant d'obtenir une inégalité en $O(1/\sqrt{n})$.

Démonstration. Comme $\mathcal{A}$ est de dimension finie, nous pouvons appliquer le théorème 2.12, i.e.

$$\mathcal{N}(r, \mathcal{A}(x_1^n)) \leq e(V+1) \left(\frac{2e}{r^2} \right)^V.$$

Donc pour n'importe quels $x_1, \dots, x_n \in \mathbb{R}^d$ et $a > 1$,

$$\int_0^{a^{-1}} \sqrt{\log \left(\sqrt{2} \mathcal{N}(r, \mathcal{A}(x_1^n)) \right)} dr \leq \int_0^{a^{-1}} \sqrt{\log \left(\sqrt{2} e(V+1) \right) + V \log(2e/r^2)} dr.$$

Comme $V \geq 1$, nous avons forcément que $V^{-1} \log(1 + V) \leq \log(2)$. Dès lors, pour tout $x_1, \dots, x_n \in \mathbb{R}^d$,

$$\begin{aligned} \int_0^{a^{-1}} \sqrt{\log(\sqrt{2} \mathcal{N}(r, \mathcal{A}(x_1^n)))} \, dr &\leq \sqrt{V} \frac{1}{\sqrt{V}} \int_0^{a^{-1}} \sqrt{\log(\sqrt{2}e(V+1)) + V \log(2e/r^2)} \, dr \\ &\leq \sqrt{V} \int_0^{a^{-1}} \sqrt{\frac{5}{2} \log(2) + 2 - 2 \log(r)} \, dr \\ &= \sqrt{V} a^{-1} \int_0^1 \sqrt{\frac{5}{2} \log(2) + 2 - 2 \log(a^{-1}z)} \, dz \\ &=: \sqrt{V} a^{-1} I(a). \end{aligned}$$

Il suit alors du théorème précédent que pour tout $a > 1$

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \sqrt{\frac{V}{n}} \frac{4a}{a-1} I(a).$$

Il suffit donc de montrer

$$\inf_{a>1} \frac{4a}{a-1} I(a) \leq 15$$

pour obtenir le résultat.

Définissons $b = b(a) := (5/2) \log(2) + 2 + 2 \log(a)$. Montrons d'abord que

$$I(a) = \int_0^1 \sqrt{b - 2 \log(z)} \, dz = \sqrt{b} + \sqrt{2\pi} e^{b/2} (1 - \Phi(\sqrt{b})),$$

où Φ désigne la fonction de répartition d'une loi normale standard (voir exemple 1.3). En effet, nous avons

$$\begin{aligned} I(a) &= \int_0^1 \sqrt{b - 2 \log(z)} \, dz = \int_{-\infty}^0 \sqrt{b - 2x} \, e^x \, dx \\ &= \int_0^{+\infty} \sqrt{b + 2x} \, e^{-x} \, dx = \sqrt{2} \int_0^{+\infty} \sqrt{b/2 + x} \, e^{-x} \, dx \\ &= \sqrt{2} e^{b/2} \int_{b/2}^{+\infty} \sqrt{t} \, e^{-t} \, dt = \sqrt{2} e^{b/2} \left(\sqrt{\frac{b}{2}} e^{-b/2} + \frac{1}{2} \int_{b/2}^{+\infty} \frac{1}{\sqrt{t}} \, e^{-t} \, dt \right) \\ &= \sqrt{b} + e^{b/2} \int_{\sqrt{b}}^{+\infty} e^{-s^2/2} \, ds = \sqrt{b} + \sqrt{2\pi} e^{b/2} (1 - \Phi(\sqrt{b})). \end{aligned}$$

En utilisant l'inégalité (1.1) démontrée dans l'exemple 1.3, il vient alors

$$I(a) \leq \sqrt{b} + \sqrt{2\pi} e^{b/2} \left(\frac{1}{\sqrt{2\pi b}} e^{-b/2} \right) = \sqrt{b} + \frac{1}{\sqrt{b}}.$$

Dès lors,

$$\inf_{a>1} \frac{4a}{a-1} I(a) \leq \inf_{a>1} \frac{4a}{a-1} \left(\sqrt{b(a)} + \frac{1}{\sqrt{b(a)}} \right),$$

avec $b(a) := (5/2) \log(2) + 2 + 2 \log(a)$. Bien que minimiser cette fonction « manuellement » soit sans doute possible, nous avons, dans le but d'éviter de longues lignes de calculs, effectué la minimisation via le logiciel libre R et nous avons obtenu

$$\inf_{a>1} \frac{4a}{a-1} \left(\sqrt{b(a)} + \frac{1}{\sqrt{b(a)}} \right) \approx 14.352 < 15,$$

où le minimum est atteint en $a \approx 3.588$. Cela termine donc la démonstration. ■

Il est en fait possible de réduire la constante 15 du corollaire 2.23 à 11 si nous ne bornons pas la série obtenue en (2.19) par une intégrale en utilisant la monotonie, comme dans la démonstration classique. À la place, nous utilisons directement l'inégalité de Haussler dans la série et ensuite nous appliquons l'inégalité de Jensen. Les techniques de démonstration étant assez différentes, nous avons présenté ces deux résultats séparément.

Le prochain résultat est dû à Johan Segers.

Théorème 2.24. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Si $\mathcal{A}$ est de dimension VC finie, nous avons*

$$\mathbb{E}[Z] = \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq 11 \sqrt{\frac{V}{n}}.$$

Démonstration. Nous partons de l'inégalité (2.19) dans la démonstration du théorème 2.22 qui assure que pour tout $a > 1$ et pour tout $x_1, \dots, x_n \in \mathbb{R}^d$,

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \frac{4a}{\sqrt{n}} \sum_{k=1}^{+\infty} a^{-k} f(a^{-k}),$$

où

$$f : r \in (0, 1] \rightarrow f(r) = \sqrt{\log \left(\sqrt{2} \mathcal{N}(r, \mathcal{A}(x_1^n)) \right)} \in [0, +\infty).$$

Observons que, comme dans la démonstration du corollaire précédent, en utilisant la borne de Haussler pour le nombre couvrant, nous avons pour tout $r \in (0, 1]$,

$$\frac{1}{\sqrt{V}} \sqrt{\log \left(\sqrt{2} \mathcal{N}(r, \mathcal{A}(x_1^n)) \right)} \leq \sqrt{\frac{5}{2} \log(2) + 2 - 2 \log(r)} := g(r).$$

La fonction g ainsi définie est convexe et décroissante sur $(0, 1]$. Introduisons les quantités

$$c_k = a^{-k} \frac{2a}{1+a},$$

qui satisfont

$$a^{-k} < c_k < a^{-k+1}, \quad \frac{1}{2}(c_k + c_{k+1}) = a^{-k} \quad \text{et} \quad c_k - c_{k+1} = \frac{2(a-1)}{a+1} a^{-k}.$$

Si U désigne une variable aléatoire de loi uniforme sur $[c_{k+1}, c_k]$, ce que nous noterons dorénavant $U \sim \text{Uni}[c_{k+1}, c_k]$, il suit alors de l'inégalité de Jensen que

$$\begin{aligned} \sum_{k=1}^{+\infty} a^{-k} g(a^{-k}) &= \frac{a+1}{2(a-1)} \sum_{k=1}^{+\infty} (c_k - c_{k+1}) g\left(\frac{1}{2}(c_k + c_{k+1})\right) \\ &= \frac{a+1}{2(a-1)} \sum_{k=1}^{+\infty} (c_k - c_{k+1}) g(\mathbb{E}[U]) \\ &\leq \frac{a+1}{2(a-1)} \sum_{k=1}^{+\infty} (c_k - c_{k+1}) \mathbb{E}[g(U)] \\ &= \frac{a+1}{2(a-1)} \sum_{k=1}^{+\infty} (c_k - c_{k+1}) \int_{c_{k+1}}^{c_k} g(r) \frac{1}{c_k - c_{k+1}} dr \\ &= \frac{a+1}{2(a-1)} \sum_{k=1}^{+\infty} \int_{c_{k+1}}^{c_k} g(r) dr \\ &= \frac{a+1}{2(a-1)} \int_0^{\frac{2}{1+a}} g(r) dr, \end{aligned}$$

où la dernière inégalité vient du fait que $c_1 = 2/(1+a)$.

En minimisant encore une fois via $\mathbf{R}$, nous obtenons

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \sqrt{\frac{V}{n}} \inf_{a > 1} \frac{2a(a+1)}{a-1} \int_0^{\frac{2}{1+a}} g(r) dr \leq 11 \sqrt{\frac{V}{n}}. \quad \blacksquare$$

Avant de passer à la synthèse et à la comparaison de toutes les inégalités exposées dans ce chapitre, nous considérons un cas particulier dans lequel nous sommes capables de calculer de manière exacte le nombre couvrant (évitant ainsi l'utilisation de la borne de Haussler) et effectuons une comparaison des résultats obtenus sur l'espérance.

Exemple 2.25 (Fonction de répartition empirique, comparaison des bornes sur l'espérance). Comme nous l'avons vu dans l'exemple 2.4, dans le cas de la fonction de répartition empirique nous pouvons calculer le coefficient de désintégration directement. Nous avons obtenu

$$\mathbb{S}_{\mathcal{A}}(n) = n + 1.$$

L'inégalité VC pour la moyenne (théorème 2.20) fournit donc la borne

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \sqrt{\frac{2 \log(4n+2)}{n}}$$

Nous avons également vu que $V = 1$, donc le corollaire 2.23 et le théorème 2.24 impliquent respectivement

$$\begin{aligned} \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] &\leq \frac{15}{\sqrt{n}} \text{ (majoration par l'intégrale et Haussler),} \\ \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] &\leq \frac{11}{\sqrt{n}} \text{ (majoration par la série et Haussler).} \end{aligned}$$

Il est aussi possible de calculer exactement le nombre couvrant $\mathcal{N}(r, \mathcal{A}(x_1^n))$ pour n'importe quels $r \in (0, 1]$ et $x_1, \dots, x_n \in \mathbb{R}$. Cela nous évite de devoir utiliser la borne de Haussler pour cette quantité et nous obtiendrons une borne de $\mathbb{E}[Z]$ plus fine en utilisant directement le théorème 2.22 au lieu du corollaire 2.23.

Soient $x_1, \dots, x_n \in \mathbb{R}$. Quitte à renommer les points, nous pouvons supposer que ceux-ci sont ordonnés de manière croissante. Par un raisonnement similaire à celui effectué dans l'exemple 2.4, il est aisé de voir que pour

$$\mathcal{A} := \{(-\infty, t] : t \in \mathbb{R}\},$$

nous avons

$$\mathcal{A} \cap \{x_1, \dots, x_n\} = \{\emptyset, \{x_1\}, \{x_1, x_2\}, \{x_1, x_2, x_3\}, \dots, \{x_1, \dots, x_{n-1}\}, \{x_1, \dots, x_n\}\}.$$

Dès lors,

$$\mathcal{A}(x_1^n) = \{(0, \dots, 0), (1, 0, \dots, 0), (1, 1, 0, \dots, 0), \dots, (1, \dots, 1, 0), (1, \dots, 1)\}.$$

Soit $r \in (0, 1]$. Pour calculer le nombre couvrant, il faut distinguer différents cas. Pour $k = 1, \dots, n$, si

$$\sqrt{\frac{k-1}{n}} \leq r < \sqrt{\frac{k}{n}},$$

alors on vérifie que

$$\mathcal{N}(r, \mathcal{A}(x_1^n)) = \left\lceil \frac{n+1}{2k-1} \right\rceil, \quad (2.20)$$

où $\lceil x \rceil$ désigne le plus petit entier z tel que $x \leq z$. Notons également que si $r = 1$ (un cas qui n'est pas traité par la formule précédente), alors, comme la distance de Hamming est toujours plus petite que 1,

$$\mathcal{N}(1, \mathcal{A}(x_1^n)) = 1.$$

Afin d'aider le lecteur à comprendre l'origine de la formule (2.20), nous illustrons sa construction sur la figure 2.3 avec $n = 6$ (donc $|\mathcal{A}(x_1^n)| = 7$) et $k = 2$.

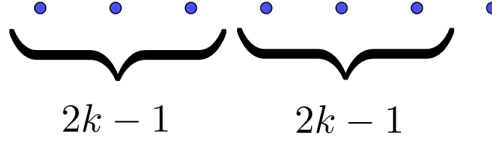


FIGURE 2.3 – Construction de la formule exacte pour le nombre couvrant dans le cas de la fonction de répartition empirique : les points représentent les sommets « ordonnés » de l'hypercube (par rapport à leur distance de Hamming avec l'origine) et les accolades représentent les points qui seront couverts par une même boule.

Remarquons que, quand $r \rightarrow 0$, la borne de Haussler assure que $\mathcal{N}(r, \mathcal{A}(x_1^n)) = O(1/r^2)$. Notre résultat n'est pas différent en ce sens que la formule (2.20) donne le même taux de convergence (car $k \approx nr^2 + 1$ quand $r \rightarrow 0$). La différence se situe au niveau des constantes, bien plus élevées dans l'inégalité de Haussler.

Maintenant que nous disposons d'une formule pour le nombre couvrant, nous pouvons calculer explicitement l'intégrale apparaissant dans la borne du théorème 2.22. Soient $a > 1$ et $x_1, \dots, x_n \in \mathbb{R}^d$ fixés. Alors,

$$\begin{aligned}
& \int_0^{a^{-1}} \sqrt{\log \left(\sqrt{2} \mathcal{N}(r, \mathcal{A}(x_1^n)) \right)} dr \\
&= \sum_{k=1}^{\lfloor \frac{n}{a^2} \rfloor} \int_{\sqrt{\frac{k-1}{n}}}^{\sqrt{\frac{k}{n}}} \sqrt{\log \left(\sqrt{2} \left\lceil \frac{n+1}{2k-1} \right\rceil \right)} dr + \int_{\sqrt{\frac{\lfloor \frac{n}{a^2} \rfloor}{n}}}^{a^{-1}} \sqrt{\log \left(\sqrt{2} \mathcal{N}(r, \mathcal{A}(x_1^n)) \right)} dr \\
&= \sum_{k=1}^{\lfloor \frac{n}{a^2} \rfloor} \int_{\sqrt{\frac{k-1}{n}}}^{\sqrt{\frac{k}{n}}} \sqrt{\log \left(\sqrt{2} \left\lceil \frac{n+1}{2k-1} \right\rceil \right)} dr + \int_{\sqrt{\frac{\lfloor \frac{n}{a^2} \rfloor}{n}}}^{a^{-1}} \sqrt{\log \left(\sqrt{2} \left\lceil \frac{n+1}{2\lfloor n/a^2 \rfloor + 1} \right\rceil \right)} dr \\
&= \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor \frac{n}{a^2} \rfloor} \left\{ \left(\sqrt{k} - \sqrt{k-1} \right) \sqrt{\log \left(\sqrt{2} \left\lceil \frac{n+1}{2k-1} \right\rceil \right)} \right\} \\
&\quad + \left(a^{-1} - \sqrt{\frac{\lfloor \frac{n}{a^2} \rfloor}{n}} \right) \sqrt{\log \left(\sqrt{2} \left\lceil \frac{n+1}{2\lfloor n/a^2 \rfloor + 1} \right\rceil \right)}.
\end{aligned}$$

Cela se traduit directement en une inégalité sur $E[Z]$. L'expression obtenue est évidemment plus difficile à manipuler mais a l'avantage d'être plus fine car l'intégrale est calculée de manière exacte.

La figure 2.4 compare (les deux axes sont en échelle logarithmique) les différentes bornes pour $E[Z]$, dans le cas de la fonction de répartition empirique, pour des échantillons

de taille raisonnable. Dans la borne provenant du calcul exact du nombre couvrant, le paramètre $a > 1$ a été choisi de façon à minimiser (numériquement, via la fonction `optimize` de R) l'expression obtenue.

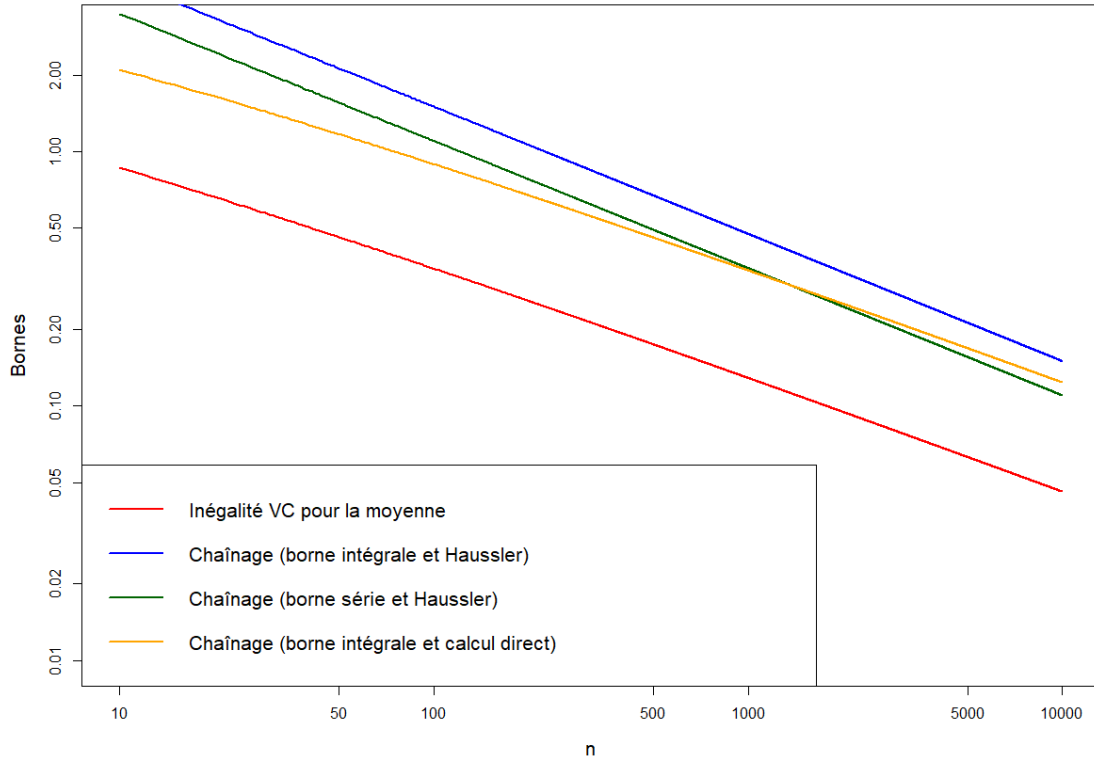


FIGURE 2.4 – Comparaison, en échelle logarithmique, des bornes sur $E[Z]$ pour la fonction de répartition empirique.

D'abord, nous observons que l'inégalité VC (pour la moyenne) est largement meilleure que le chaînage pour des échantillons de taille raisonnable. Cela s'explique par les constantes trop élevées qui apparaissent dans les bornes, inhérentes à la méthode ; même si celles-ci ne contiennent plus de terme dépendant de n au numérateur, comme le numérateur de l'inégalité VC croît extrêmement lentement avec n , cette dernière demeure meilleure dans les applications pratiques.

Ensuite, nous remarquons un autre phénomène très intéressant. Parmi les inégalité utilisant le chaînage, deux catégories sont à distinguer : celles qui utilisent une borne de la série (2.19) par une intégrale, un argument qui utilise implicitement la monotonie des expressions rencontrées (théorème 2.22 et corollaire 2.23) et celle qui travaille directement sur la série avec l'inégalité de Jensen (théorème 2.24). Au sein de celles qui utilisent une borne intégrale, il est tout à fait normal et rassurant de constater que le calcul exact du nombre couvrant améliore l'inégalité obtenue avec le résultat de Haussler (même si

les courbes ont l'air de se rapprocher, le lecteur peut vérifier graphiquement que l'ordre demeure valide pour n largement supérieur). Par contre, pour n excédant 1000, nous constatons que la courbe verte qui utilise directement la série et l'inégalité de Jensen performe mieux que la borne obtenue via le calcul exact du nombre couvrant (courbe orange), malgré que le résultat de Haussler ait été utilisé pour démontrer la borne représentée par la courbe verte. Ceci est en réalité une manifestation assez générale d'un principe empirique que nous avons constaté et que nous retrouverons au chapitre suivant : les inégalités basées sur Jensen sont souvent largement meilleures que celles obtenues par un argument de monotonie.

2.3 Comparaison des bornes

Nous avons, au travers de différentes méthodes, obtenu de nombreuses bornes pour

$$Z = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|,$$

qui tiennent avec probabilité $1 - \delta$ pour quelque $\delta \in (0, 1)$. Nous proposons de synthétiser ici les résultats principaux obtenus (les résultats qui nécessitent de l'information supplémentaire sur la classe $\mathcal{A}$, tels que le théorème 2.18 ou le théorème 2.21, seront traités en détails dans le chapitre suivant et ne sont donc pas rappelés ici) :

- L'inégalité VC classique (théorème 2.14) :

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2}{n} [\log(4/\delta) + \log(\mathbb{S}_{\mathcal{A}}(2n))]}.$$

- L'inégalité VC améliorée que nous avons présentée dans le théorème 2.17 :

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{1}{2n}} + \sqrt{\frac{2}{n} [\log(4/\delta) + \log(\mathbb{S}_{\mathcal{A}}(2n))]}.$$

- L'inégalité VC pour la moyenne (théorème 2.20) combinée à (2.5) :

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{1}{2n} \log(1/\delta)} + \sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n}}.$$

- La méthode de chaînage (théorème 2.22) combinée à (2.5) :

$$\begin{aligned} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| &\leq \sqrt{\frac{1}{2n} \log(1/\delta)} \\ &+ \frac{4}{\sqrt{n}} \frac{a^2}{a-1} \max_{x_1, \dots, x_n \in \mathbb{R}^d} \int_0^{a^{-1}} \sqrt{\log(\sqrt{2} \mathcal{N}(r, \mathcal{A}(x_1^n)))} \, dr, \end{aligned}$$

où $a > 1$.

- Si la dimension VC de $\mathcal{A}$ est finie, le théorème 2.24 combiné à (2.5) :

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{1}{2n} \log(1/\delta)} + 11\sqrt{\frac{V}{n}}.$$

Nous n'avons pas inclus la borne dérivée du corollaire 2.23 car, si la dimension VC est finie, nous utiliserons toujours le théorème 2.24.

Dans le cas de la fonction de répartition empirique, nous avons calculé toutes les quantités intervenant dans ces bornes (voir exemples 2.4 et 2.25). Nous présentons sur la figure 2.5 une comparaison de ces bornes (les deux axes sont en échelle logarithmique) avec $\delta = 1\%$.

Premièrement, nous observons que la borne obtenue via l'inégalité VC pour la moyenne semble la meilleure option pour des échantillons de taille raisonnable. Cette constatation va dans le sens de l'exemple 2.25. Intuitivement, cela peut sembler contre-intuitif : les bornes provenant d'une inégalité de concentration (en l'occurrence, l'inégalité des différences finies) et d'un contrôle sur l'espérance sont toujours composées de l'addition des deux résultats, tandis que les bornes provenant directement de la distribution de Z sont obtenues sans avoir recours à une inégalité de concentration préalable. De plus, les techniques de démonstration sont assez similaires. Il semblerait que la différence provienne de la symétrisation, qui, même optimisée dans le cas des distributions (au sens du théorème 2.17) parvient seulement à concurrencer la borne dérivée de l'inégalité VC pour la moyenne pour n très large (courbe bleue). La symétrisation dans le cas de l'espérance est beaucoup plus simple à traiter, et, visiblement, plus performante également. Néanmoins, observons tout de même l'amélioration significative engendrée par l'optimisation du processus de symétrisation par rapport à l'inégalité VC classique (courbe noire).

Deuxièmement, nous constatons, sans étonnement, que la méthode de chaînage gagne en performance avec n car elle ne contient aucun terme dépendant de n au numérateur (courbe verte), mais qu'elle n'est pas très utile en pratique à cause de ses constantes (cf. exemple 2.25). En effet, si nous comparons la borne sur l'espérance obtenue par chaînage avec l'inégalité VC pour la moyenne, nous constatons que le chaînage ne devient meilleur que si

$$n \geq \frac{\exp(121/2) - 4}{4} \approx 4.7 \times 10^{25},$$

une quantité dépassant largement la taille d'un échantillon réaliste.

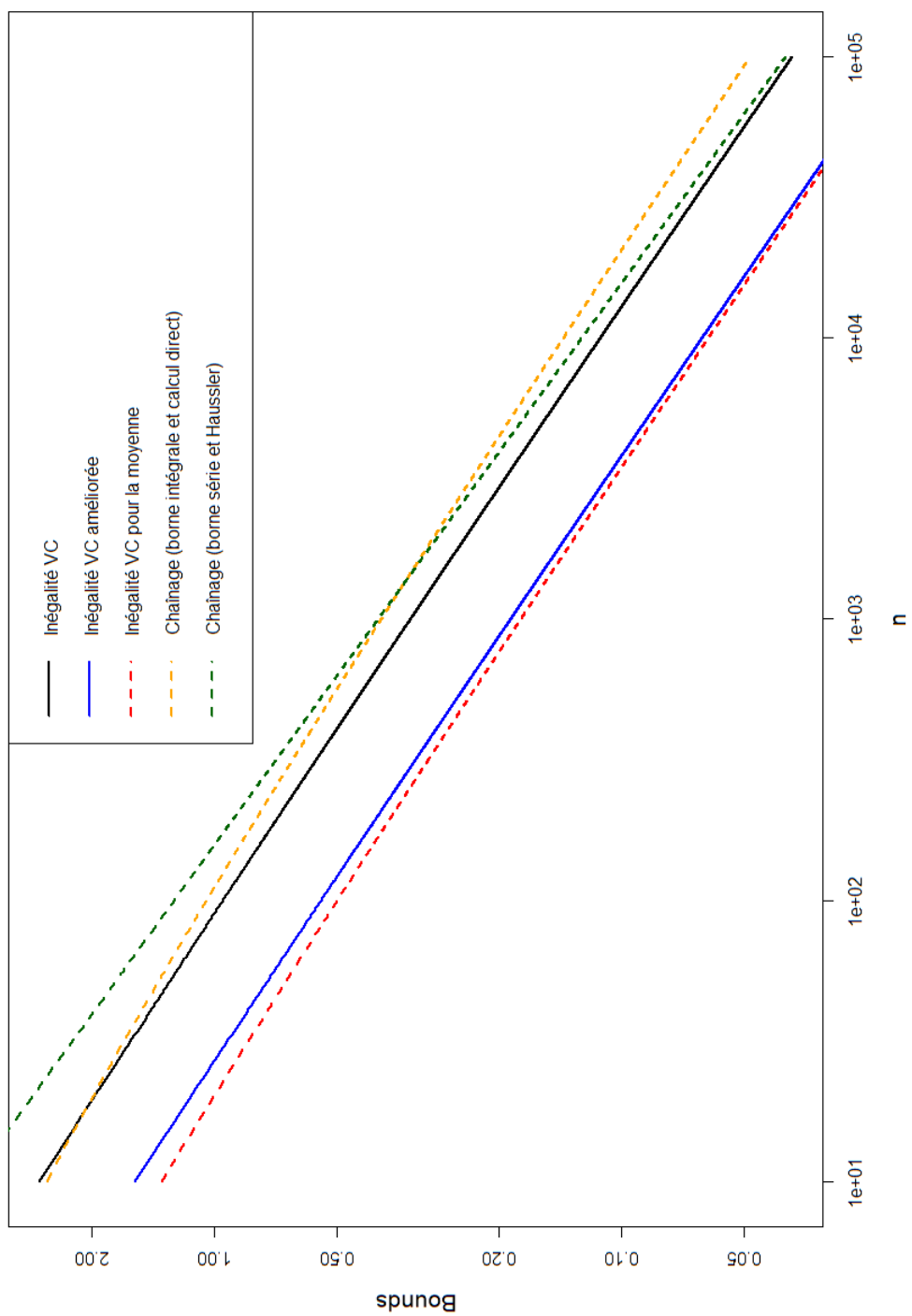


FIGURE 2.5 – Comparaison des bornes pour la fonction de répartition empirique avec $\delta = 1\%$. Les inégalités obtenues à partir de la distribution sont en traits continus et les inégalités obtenues à partir d'une borne sur l'espérance sont en traits discontinus.

Chapitre 3

Mesures empiriques : le cas des évènements rares

Ce chapitre se base principalement sur [25].

Comme dans le chapitre précédent, nous souhaitons obtenir des bornes sur la probabilité

$$P \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right),$$

où μ est une mesure de probabilité sur $\mathbb{R}^d$, μ_n est la mesure empirique associée à un échantillon indépendant $X_1, \dots, X_n$ de distribution commune μ et $\mathcal{A}$ est une classe de boréliens de $\mathbb{R}^d$. Cependant, nous voulons maintenant adapter ces bornes pour les *évènements rares*, i.e. si nous notons

$$\mathbb{A} := \bigcup_{A \in \mathcal{A}} A \quad \text{et} \quad p := \mu(\mathbb{A}), \quad (3.1)$$

alors p est supposé très petit (nous imaginons, par exemple, que p est de l'ordre de 10^{-2} ou 10^{-3}). La mesurabilité de l'union $\mathbb{A}$ est garantie par l'hypothèse faite sur $\mathcal{A}$ dans la remarque 2.1 car elle assure que $\bigcup_{A \in \mathcal{A}} A = \bigcup_{A \in \mathcal{A}_0} A$, et la mesurabilité d'une union dénombrable est évidente.

Ce cadre s'inscrit dans le contexte de la *théorie des valeurs extrêmes*, la branche de la statistique qui s'intéresse aux valeurs extrêmes d'un échantillon, telles que son maximum ou ses quantiles extrémaux. Celle-ci possède de nombreuses applications très diverses telles que la prévision d'évènements extrêmes en météorologie, la gestion de gros risques en réassurance ou encore la détection de nouveautés en apprentissage machine (voir [3, Section 1.3] pour d'autres applications).

Comme expliqué auparavant, cf. (2.12), le théorème 2.18 permet d'obtenir un premier résultat intéressant dans ce cadre car il fournit l'inégalité

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{p}{n} \left[\log(8/\delta) + \log(\mathbb{S}_{\mathcal{A}}(2n)) \right]},$$

avec probabilité $1 - \delta$, où $\delta \in (0, 1)$. Le facteur p permet de raffiner l'inégalité VC par la connaissance à priori d'information sur la classe $\mathcal{A}$. Néanmoins, ce résultat ne semble pas suffisant dans le sens où la *taille de l'échantillon effectif*, i.e. le nombre moyen d'observations X_i prenant leur valeur dans $\mathbb{A}$, est donnée par np , une quantité qui peut être largement inférieure à la taille de l'échantillon total n . Ainsi, nous nous attendons à ce que la mesure de complexité de la classe, donnée par $\mathbb{S}_{\mathcal{A}}$, fasse directement intervenir cette quantité et non n .

Dans la section 3.1, nous présentons le résultat fondamental du chapitre qui permet de développer des inégalités prenant la rareté des événements rencontrés en considération. Dans les sections 3.2 et 3.3, nous démontrons différentes inégalités en réutilisant les techniques introduites au chapitre précédent. La section 3.4 est consacrée à une comparaison graphique des bornes obtenues.

3.1 Un argument de conditionnement

Le résultat suivant est essentiel et sera utilisé dans toutes les inégalités de concentration qui suivront. Il est proposé dans [25] sous le nom de « conditioning trick » et provient originellement de [13, p. 17].

Lemme 3.1. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Soit $Y_1, \dots, Y_n$ un échantillon indépendant de distribution commune*

$$\mu_{\mathbb{A}}(\cdot) = \mu(\cdot | \mathbb{A}) = \frac{\mu(\cdot \cap \mathbb{A})}{\mu(\mathbb{A})} = \frac{\mu(\cdot \cap \mathbb{A})}{p}$$

et indépendant de $X_1, \dots, X_n$, avec $\mathbb{A}$ et p comme dans (3.1). Si pour tout $k = 1, \dots, n$ nous notons μ_k^Y la mesure empirique associée à $Y_1, \dots, Y_k$ et si $K \sim \text{Bin}(n, p)$ désigne le nombre d'observations X_i dans $\mathbb{A}$, alors nous avons l'égalité en distribution suivante pour tout $k \in \{0, \dots, n\}$:

$$\left[(\mu_n(A))_{A \in \mathcal{A}} \mid K = k \right] =_d \left(\frac{k}{n} \mu_k^Y(A) \right)_{A \in \mathcal{A}},$$

où l'égalité en distribution est à considérer au sens d'une égalité en distribution entre deux processus stochastiques, i.e. pour tout $A_1, \dots, A_N \in \mathcal{A}$, les vecteurs aléatoires

$$\left(\mu_n(A_1), \dots, \mu_n(A_N) \right) \mid K = k \quad \text{et} \quad \frac{k}{n} \left(\mu_k^Y(A_1), \dots, \mu_k^Y(A_N) \right) \quad (3.2)$$

ont même distribution (si $k = 0$ les deux côtés de l'égalité sont supposés nuls).

Dans la suite, nous pourrons donc travailler conditionnellement à $K = k$ pour un certain $k = 0, \dots, n$, fournir des inégalités de concentration pour μ_k^Y et intégrer par rapport à K à la fin de notre développement. Cela permettra d'obtenir des bornes ne prenant en compte que la taille de l'échantillon effectif. En particulier, nous ferons usage de l'inégalité de Bernstein (théorème 1.11) qui garantit que K est fortement concentré autour de sa moyenne $E[K] = np$, i.e. pour $t \geq 0$,

$$P(|K - np| \geq t) \leq 2 \exp\left(-\frac{t^2}{2(np(1-p) + t/3)}\right). \quad (3.3)$$

Notons que l'égalité en distribution du lemme 3.1 s'étend au supremum sur $\mathcal{A}$ en utilisant la remarque 2.1. En effet, en combinaison avec la continuité des probabilités par rapport aux suites monotones d'évènements [26, Proposition 1.2.9], elle assure que pour tout $t > 0$ et pour tout $k \in \{0, \dots, n\}$,

$$\begin{aligned} P\left(\sup_{A \in \mathcal{A}} \mu_n(A) \geq t \mid K = k\right) &= P\left(\sup_{A \in \mathcal{A}_0} \mu_n(A) \geq t \mid K = k\right) \\ &= P\left(\bigcap_{l \in \mathbb{N}} \left\{ \sup_{A \in \mathcal{A}_0} \mu_n(A) > t - 1/l \right\} \mid K = k\right) \\ &= P\left(\bigcap_{l \in \mathbb{N}} \bigcup_{i \in \mathbb{N}} \{\mu_n(A_i) > t - 1/l\} \mid K = k\right) \\ &= \lim_{l \rightarrow +\infty} P\left(\bigcup_{i \in \mathbb{N}} \{\mu_n(A_i) > t - 1/l\} \mid K = k\right) \\ &= \lim_{l \rightarrow +\infty} \lim_{N \rightarrow +\infty} P\left(\bigcup_{i=1}^N \{\mu_n(A_i) > t - 1/l\} \mid K = k\right) \\ &= \lim_{l \rightarrow +\infty} \lim_{N \rightarrow +\infty} P\left(\bigcup_{i=1}^N \{\mu_k^Y(A_i) > t - 1/l\}\right) \\ &= P\left(\sup_{A \in \mathcal{A}_0} \frac{k}{n} \mu_k^Y(A) \geq t\right) = P\left(\sup_{A \in \mathcal{A}} \frac{k}{n} \mu_k^Y(A) \geq t\right), \end{aligned}$$

où $A_1, A_2, A_3, \dots$ est une énumération de la classe dénombrable $\mathcal{A}_0$.

Démonstration du lemme 3.1. Soient $k \in \{1, \dots, n\}$ (si $k = 0$ le résultat est trivial) et $A_1, \dots, A_N \in \mathcal{A}$.

Commençons par supposer les A_i mutuellement disjoints. Quitte à ajouter l'ensemble $\mathbb{A} \setminus (A_1 \cup \dots \cup A_N) \in \mathcal{A}$ à la suite $(A_l)_{l=1}^N$, nous pouvons supposer, sans perte de généralité, que celle-ci forme une partition de $\mathbb{A}$. Prenons $j_1, \dots, j_N \in \{0, \dots, k\}$ tels que $\sum_{l=1}^N j_l = k$ (sinon les probabilités calculées ultérieurement sont nulles). Comme, dans ce cas, le vecteur aléatoire $(n\mu_n(A_1), \dots, n\mu_n(A_N)) \mid K = k$ est de loi multinomiale

$Mult(k; \mu(A_1)/p, \dots, \mu(A_N)/p)$, nous avons

$$P\left(\left(\mu_n(A_1), \dots, \mu_n(A_N)\right) = (j_1, \dots, j_N)/n \mid K = k\right) = \binom{k}{j_1, \dots, j_N} \prod_{l=1}^N \left[\frac{\mu(A_l)}{p}\right]^{j_l},$$

où nous avons utilisé le coefficient multinomial

$$\binom{k}{j_1, \dots, j_N} := \frac{k!}{(j_1)! \cdots (j_N)!}.$$

De même, nous calculons

$$\begin{aligned} P\left(\frac{k}{n}(\mu_k^Y(A_1), \dots, \mu_k^Y(A_N)) = (j_1, \dots, j_N)/n\right) \\ = P\left(\bigcap_{l=1}^N \left\{\sum_{i=1}^k \mathbb{1}\{Y_i \in A_l\} = j_l\right\}\right) = \frac{k!}{(j_1)! \cdots (j_N)!} \prod_{l=1}^N \left[\frac{\mu(A_l)}{p}\right]^{j_l}, \end{aligned}$$

vérifiant ainsi l'égalité en distribution des vecteurs de (3.2), dans le cas où les A_l sont tous disjoints.

Si les A_l ne sont pas mutuellement disjoints, l'égalité en distribution demeure valide. En effet, considérons d'abord le cas $N = 2$: clairement,

$$(\mu_n(A_1), \mu_n(A_2)) = F(\mu_n(A_1 \setminus A_2), \mu_n(A_1 \cap A_2), \mu_n(A_2 \setminus A_1)),$$

où $F : (x, y, z) \in \mathbb{R}^3 \mapsto (x + y, y + z) \in \mathbb{R}^2$. Or, il est clair que les ensembles $A_1 \setminus A_2, A_1 \cap A_2, A_2 \setminus A_1$ sont mutuellement disjoints et que la fonction F est mesurable. Dès lors, l'égalité en distribution dans le cas disjoint se « transporte » par F mesurable au cas non disjoint. Le cas $N > 2$ se traite en considérant une fonction mesurable $F : \mathbb{R}^{2^N-1} \rightarrow \mathbb{R}^N$ et une décomposition disjointe de $A_1 \cup \dots \cup A_N$ similaire à celle effectuée dans le cas $N = 2$. ■

Nous proposons maintenant diverses inégalités de concentration, adaptées au cadre des évènements rares. Comme dans le chapitre précédent, nous distinguons deux méthodes : les inégalités basées sur la distribution de $Z = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$ et les inégalités basées sur des bornes de $E[Z]$.

Nous serons dans la suite amenés à définir $\mathbb{S}_{\mathcal{A}}(x)$, pour $x \in \mathbb{R}$ un réel qui ne sera pas nécessairement dans $\mathbb{N}$. Nous procédons pour cela à une simple interpolation linéaire : il existe $\lambda \in [0, 1]$ tel que $x = \lambda \lfloor x \rfloor + (1 - \lambda) \lceil x \rceil$, nous définissons donc

$$\mathbb{S}_{\mathcal{A}}(x) := \lambda \mathbb{S}_{\mathcal{A}}(\lfloor x \rfloor) + (1 - \lambda) \mathbb{S}_{\mathcal{A}}(\lceil x \rceil).$$

3.2 Bornes provenant de la distribution

Le résultat suivant est une version améliorée de [25, Théorème 1].

Théorème 3.2. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Si $np \geq 2 \log(8/\delta)$, avec p comme dans (3.1), alors pour $\delta \in (0, 1)$, nous avons avec probabilité au moins $1 - \delta$,*

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2 \sqrt{\frac{2p}{n} [\log(8/\delta) + \log(\mathbb{S}_{\mathcal{A}}(8np))]} + \sqrt{\frac{2p}{n}}.$$

Cette inégalité peut être vue comme une version « événements rares » de l'inégalité VC (théorème 2.14). En particulier, la démonstration se base également sur le lemme de symétrisation (lemme 2.15) et un argument de randomisation.

Démonstration. Soient $a \in (0, 1)$ et $t > 0$. Par le lemme de symétrisation, si $4(1-a)^2 nt^2 \geq 2$, alors

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) \leq 4 \mathbb{P} \left(\sup_{A \in \mathcal{A}} (\mu_n(A) - \mu'_n(A)) \geq at \right),$$

où μ'_n est la mesure empirique associée à $X'_1, \dots, X'_n$, un échantillon fantôme indépendant de distribution commune μ et indépendant de $X_1, \dots, X_n$. En fait, nous vérifions facilement que, comme $\mu(A) \leq p$ pour tout $A \in \mathcal{A}$, la symétrisation reste valide sous l'hypothèse plus faible que $(1-a)^2 nt^2 \geq 2p$ (il suffit d'adapter la borne sur la variance $\text{Var}[\mathbb{1}\{X_i \in A\}]$ dans la démonstration).

Soit $\tilde{K} = \sum_{i=1}^n \mathbb{1}\{X_i \in \mathbb{A} \text{ ou } X'_i \in \mathbb{A}\} = \sum_{i=1}^n \mathbb{1}\{(X_i, X'_i) \in \tilde{\mathbb{A}}\}$, avec $\tilde{\mathbb{A}} = (\mathbb{A} \times \mathbb{R}^d) \cup (\mathbb{R}^d \times \mathbb{A})$. Alors, si $(Y_i, Y'_i)_{i=1}^n$ est un échantillon indépendant de $(X_i, X'_i)_{i=1}^n$ de distribution commune donnée par la distribution conditionnelle de (X_1, X'_1) sachant que ce dernier est dans $\tilde{\mathbb{A}}$, nous avons par le lemme 3.1

$$\begin{aligned} & \mathbb{P} \left(\sup_{A \in \mathcal{A}} (\mu_n(A) - \mu'_n(A)) \geq at \right) \\ &= \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}) \geq at \right) \\ &= \mathbb{E} \left[\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n (\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}) \geq at \mid \tilde{K} \right) \right] \\ &= \mathbb{E} \left[\mathbb{P} \left(\frac{\tilde{K}}{n} \sup_{A \in \mathcal{A}} \frac{1}{\tilde{K}} \sum_{i=1}^{\tilde{K}} (\mathbb{1}\{Y_i \in A\} - \mathbb{1}\{Y'_i \in A\}) \geq at \mid \tilde{K} \right) \right] \end{aligned}$$

$$= \mathbb{E} \left[\mathbb{P} \left(\sup_{A \in \mathcal{A}} \sum_{i=1}^{\tilde{K}} (\mathbb{1}\{Y_i \in A\} - \mathbb{1}\{Y'_i \in A\}) \geq nat \mid \tilde{K} \right) \right].$$

Par randomisation nous avons,

$$\sum_{i=1}^{\tilde{K}} (\mathbb{1}\{Y_i \in A\} - \mathbb{1}\{Y'_i \in A\}) =_d \sum_{i=1}^{\tilde{K}} \sigma_i (\mathbb{1}\{Y_i \in A\} - \mathbb{1}\{Y'_i \in A\}),$$

où $\tilde{K}$ est supposé fixé et $\sigma_1, \dots, \sigma_{\tilde{K}}$ sont des variables aléatoires de Rademacher supposées indépendantes de $(Y_1, Y'_1), \dots, (Y_n, Y'_n)$. Comme pour l'inégalité VC, nous avons par l'inégalité de Hoeffding (théorème 1.8),

$$\begin{aligned} & \mathbb{P} \left(\sup_{A \in \mathcal{A}} \sum_{i=1}^{\tilde{K}} (\mathbb{1}\{Y_i \in A\} - \mathbb{1}\{Y'_i \in A\}) \geq nat \mid \tilde{K} \right) \\ &= \mathbb{E}_Y \left[\mathbb{P}_\sigma \left(\sup_{A \in \mathcal{A}} \sum_{i=1}^{\tilde{K}} \sigma_i (\mathbb{1}\{Y_i \in A\} - \mathbb{1}\{Y'_i \in A\}) \geq nat \mid \tilde{K}, Y_1, Y'_1, \dots, Y_n, Y'_n \right) \right] \\ &\leq \min \left\{ 1, \mathbb{S}_{\mathcal{A}}(2\tilde{K}) \exp \left(-\frac{a^2(nt)^2}{2\tilde{K}} \right) \right\}, \end{aligned}$$

où la dernière inégalité suit du fait que l'espérance $\mathbb{E}_Y$ n'est prise que par rapport aux variables aléatoires $Y_1, Y'_1, \dots, Y_n, Y'_n$ et que toute probabilité est trivialement bornée par 1. Nous trouvons donc

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) \leq 4 \mathbb{E} \left[\min \left\{ 1, \mathbb{S}_{\mathcal{A}}(2\tilde{K}) \exp \left(-\frac{a^2(nt)^2}{2\tilde{K}} \right) \right\} \right]. \quad (3.4)$$

Afin de borner cette dernière quantité, nous distinguons selon que $\tilde{K} \leq 2np(1+s)$ ou non, avec $s > 0$ qui sera déterminé dans la suite. Par monotonie de l'intégrant en $\tilde{K}$, nous avons

$$\begin{aligned} & \mathbb{E} \left[\min \left\{ 1, \mathbb{S}_{\mathcal{A}}(2\tilde{K}) \exp \left(-\frac{a^2(nt)^2}{2\tilde{K}} \right) \right\} \right] \\ &\leq \mathbb{P}(\tilde{K} \geq 2np(1+s)) + \mathbb{S}_{\mathcal{A}}(4np(1+s)) \exp \left(-\frac{na^2t^2}{4p(1+s)} \right). \end{aligned}$$

Comme $a \in (0, 1)$ n'est pas fixé, nous pouvons, étant donnés n et t , choisir le plus grand $a < 1$ tel que la condition $(1-a)^2nt^2 \geq 2p$ soit respectée. Il est donné par

$a = a(n, t) = 1 - \sqrt{\frac{2p}{nt^2}}$. Pour un tel choix de a , nous trouvons donc

$$\begin{aligned} \mathbb{E} \left[\min \left\{ 1, \mathbb{S}_{\mathcal{A}}(2\tilde{K}) \exp \left(-\frac{a^2(nt)^2}{2\tilde{K}} \right) \right\} \right] \\ \leq \mathbb{P} \left(\tilde{K} \geq 2np(1+s) \right) + \mathbb{S}_{\mathcal{A}}(4np(1+s)) \exp \left(-\frac{nt^2 - 2\sqrt{2npt} + 2p}{4p(1+s)} \right). \end{aligned} \quad (3.5)$$

Soit $\delta \in (0, 1)$. Nous choisissons $s = s(\delta)$ tel que le premier terme dans (3.5) soit borné par $\delta/8$. Ensuite, nous choisissons $t = t(\delta)$ tel que le second terme soit égal à $\delta/8$. Ainsi, par (3.4), nous aurons que $\mathbb{P}(Z \geq t(\delta)) \leq \delta$ et par conséquent, nous obtiendrons $Z \leq t(\delta)$ avec probabilité au moins $1 - \delta$.

Par la version de l'inégalité de Bernstein rappelée dans (3.3), comme nous vérifions facilement que $\tilde{K} \sim \text{Bin}(n, \tilde{p})$ avec $\tilde{p} = p(2-p)$, nous avons

$$\begin{aligned} \mathbb{P} \left(\tilde{K} \geq 2np(1+s) \right) &= \mathbb{P} \left(\tilde{K} - n\tilde{p} \geq 2np(1+s) - n\tilde{p} \right) = \mathbb{P} \left(\tilde{K} - n\tilde{p} \geq np(2s+p) \right) \\ &\leq \exp \left(-\frac{(np)^2(2s+p)^2}{2(n\tilde{p}(1-\tilde{p}) + np(2s+p)/3)} \right) \\ &\leq \exp \left(-\frac{np(2s+p)^2}{2((2-p) + (2s+p)/3)} \right) \\ &= \exp \left(-\frac{np(s+p/2)^2}{1+s/3-p/3} \right) \\ &\leq \exp \left(-\frac{nps^2}{1+s/3} \right). \end{aligned}$$

En égalant cette expression à $\delta/8$, nous obtenons une équation quadratique en s dont la solution positive est donnée par

$$s_+ = s(\delta) := \frac{\log(8/\delta)}{3np} + \sqrt{\frac{\log(8/\delta)}{np}}.$$

Comme nous avons supposé $np \geq 2\log(8/\delta)$, il vient $s(\delta) \leq \frac{1}{6} + \frac{1}{\sqrt{2}} \leq 1$. Donc en prenant $s = 1$, le premier terme dans (3.5) est borné par $\delta/8$.

Pour un tel choix de s , le second terme de (3.5) devient

$$\mathbb{S}_{\mathcal{A}}(8np) \exp \left(-\frac{nt^2 - 2\sqrt{2npt} + 2p}{8p} \right).$$

Cette dernière expression est égale à $\delta/8$ à condition que

$$t = t(\delta) := 2\sqrt{\frac{2p}{n} \left[\log(8/\delta) + \log(\mathbb{S}_{\mathcal{A}}(8np)) \right]} + \sqrt{\frac{2p}{n}}. \quad \blacksquare$$

Nous pouvons obtenir une borne similaire en utilisant le lemme 3.1 sur la version non symétrisée de Z et en appliquant le théorème 2.17 directement sur μ_K^Y . Ceci est formalisé dans le résultat suivant, qui correspond à une version améliorée de [25, Théorème 2].

Théorème 3.3. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Si $np \geq 4 \log(4/\delta)$, avec p comme dans (3.1), alors pour $\delta \in (0, 1)$, nous avons avec probabilité au moins $1 - \delta$,*

$$\begin{aligned} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \\ \leq \frac{2}{3n} \log(4/\delta) + \sqrt{\frac{p}{n}} \left\{ \sqrt{2 \log(4/\delta)} + 2\sqrt{\log(8/\delta) + \log(\mathbb{S}_{\mathcal{A}}(4np))} + 1 \right\}. \end{aligned}$$

Démonstration. Soit $t > 0$. Si $K := \sum_{i=1}^n \mathbf{1}\{X_i \in \mathbb{A}\} \sim \text{Bin}(n, p)$, alors, en reprenant les notations du lemme 3.1,

$$\begin{aligned} \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{|\mu_n(A) - \mu(A)|}{p} \geq t \right) &= \mathbb{E} \left[\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{|\mu_n(A) - \mu(A)|}{p} \geq t \mid K \right) \right] \\ &= \mathbb{E} \left[\mathbb{P} \left(\sup_{A \in \mathcal{A}} \left| \frac{K}{np} \mu_K^Y(A) - \mu_{\mathbb{A}}(A) \right| \geq t \mid K \right) \right] \\ &\leq \mathbb{E} \left[\mathbb{P} \left(\frac{K}{np} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| + \left| \frac{K}{np} - 1 \right| \sup_{A \in \mathcal{A}} \mu_{\mathbb{A}}(A) \geq t \mid K \right) \right] \\ &\leq \mathbb{E} \left[\mathbb{P} \left(\frac{K}{np} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| + \left| \frac{K}{np} - 1 \right| \geq t \mid K \right) \right], \end{aligned}$$

où la dernière ligne suit du fait que $\mu_{\mathbb{A}}$ est encore une mesure de probabilité. Notons $t = u + s$, avec $u, s > 0$. Alors, cette dernière quantité est bornée par

$$\mathbb{P} \left(\left| \frac{K}{np} - 1 \right| \geq s \right) + \mathbb{E} \left[\mathbb{P} \left(\frac{K}{np} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \geq u \mid K \right) \mathbf{1} \left\{ \left| \frac{K}{np} - 1 \right| \leq s \right\} \right]. \quad (3.6)$$

Soit $\delta \in (0, 1)$. Comme dans la démonstration du théorème précédent, nous fixons $s(\delta)$ tel que le premier terme soit borné par $\delta/2$ et nous fixons ensuite $u(\delta)$ tel que le second terme soit également borné par $\delta/2$.

Le premier terme de (3.6) se borne directement en utilisant l'inégalité de Bernstein (3.3)

$$\begin{aligned} \mathbb{P} \left(\left| \frac{K}{np} - 1 \right| \geq s \right) &\leq \mathbb{P} (|K - np| \geq nps) \\ &\leq 2 \exp \left(-\frac{n^2 p^2 s^2}{2(np(1-p) + nps/3)} \right) \leq 2 \exp \left(-\frac{nps^2}{2(1+s/3)} \right). \end{aligned}$$

La solution positive s_+ associée à l'équation quadratique en s obtenue en égalant ce dernier terme à $\delta/2$ satisfait

$$s_+ \leq \frac{2}{3np} \log(4/\delta) + \sqrt{\frac{2}{np} \log(4/\delta)} =: s(\delta).$$

Pour traiter le second terme de (3.6), nous utilisons le théorème 2.17

$$\begin{aligned} \mathbb{P} \left(\frac{K}{np} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \geq u \mid K \right) &= \mathbb{P} \left(\sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \geq \frac{np u}{K} \mid K \right) \\ &\leq 4\mathbb{S}_{\mathcal{A}}(2K) \exp \left(-\frac{(np)^2 u^2}{2K} \left(1 - \sqrt{\frac{2K}{(np)^2 u^2}} \right) - \frac{1}{4} \right). \end{aligned}$$

Dès lors, par monotonie en K ,

$$\begin{aligned} \mathbb{E} \left[\mathbb{P} \left(\frac{K}{np} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \geq u \mid K \right) \mathbb{1} \left\{ \left| \frac{K}{np} - 1 \right| \leq s \right\} \right] \\ \leq 4\mathbb{S}_{\mathcal{A}}(2np(1+s)) \exp \left(-\frac{np u^2}{2(1+s)} \left(1 - \sqrt{\frac{2(1+s)}{np u^2}} \right) - \frac{1}{4} \right). \end{aligned}$$

Si $np \geq 4 \log(4/\delta)$, nous obtenons $s(\delta) \leq \frac{1}{6} + \frac{1}{\sqrt{2}} \leq 1$. Donc, pour un tel choix de s , le second terme est borné par

$$4\mathbb{S}_{\mathcal{A}}(4np) \exp \left(-\frac{np}{4} u^2 + \frac{\sqrt{np}}{2} u - \frac{1}{4} \right).$$

Cette dernière expression vaut $\delta/2$ si

$$u = u(\delta) = \frac{1}{\sqrt{np}} + 2\sqrt{\frac{1}{np} [\log(8/\delta) + \log(\mathbb{S}_{\mathcal{A}}(4np))]}.$$

En synthétisant les inégalités obtenues, nous avons montré qu'avec probabilité au moins $1 - \delta$,

$$\begin{aligned} \sup_{A \in \mathcal{A}} \frac{|\mu_n(A) - \mu(A)|}{p} \\ \leq \frac{2}{3np} \log(4/\delta) + \sqrt{\frac{1}{np}} \left\{ \sqrt{2 \log(4/\delta)} + 2\sqrt{\log(8/\delta) + \log(\mathbb{S}_{\mathcal{A}}(4np))} + 1 \right\}. \end{aligned}$$

Multipliant chaque coté de cette inégalité par p , nous obtenons le résultat annoncé. ■

3.3 Bornes provenant de l'espérance

Une approche alternative consiste à utiliser une inégalité de concentration pour $Z = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$ autour de sa moyenne $E[Z]$ et ensuite à borner cette dernière quantité (comme nous l'avons fait dans la section 2.2.2, mais avec des inégalités adaptées au cadre des événements rares). L'inégalité des différences finies dans sa version développée au théorème 1.19 combinée avec [13, Lemme 12], qui assure que, dans les mêmes notations que le théorème 1.19 et (3.1), $c \leq 1/n$ et $\sigma \leq 2p/n$, permet de montrer que pour tout $t > 0$,

$$P(Z - E[Z] \geq t) \leq \exp\left(-\frac{nt^2}{4p + 2t/3}\right).$$

Une application simple du lemme d'inversion (lemme 2.13) permet alors d'obtenir le résultat suivant [SS, Théorème 3].

Théorème 3.4 (Concentration du supremum dans le cas des événements rares). *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Alors, pour $\delta \in (0, 1)$, avec p comme dans (3.1), nous avons avec probabilité au moins $1 - \delta$,*

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \frac{2}{3n} \log(1/\delta) + 2\sqrt{\frac{p}{n} \log(1/\delta)} + E\left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right].$$

Le reste de la section est consacré à la recherche de bornes sur l'espérance.

Une première idée consiste simplement à appliquer le théorème 2.21 en utilisant le fait que $\Sigma = \sup_{A \in \mathcal{A}} \sqrt{\mu(A)(1 - \mu(A))} \leq \sqrt{p}$ dans notre cadre (cf. (3.1)).

Théorème 3.5. *Soient $X_1, \dots, X_n$ un échantillon indépendant de distribution commune μ , une mesure de probabilité sur $\mathbb{R}^d$, et μ_n la mesure empirique associée. Soit $\mathcal{A}$ une classe non vide de boréliens de $\mathbb{R}^d$. Alors, avec p comme dans (3.1), nous avons*

$$E\left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right] \leq \frac{8 \log(2\mathbb{S}_{\mathcal{A}}(2n))}{n} + \sqrt{\frac{p}{n} [32 \log(2\mathbb{S}_{\mathcal{A}}(2n))]}.$$

L'idée alternative est d'utiliser le lemme 3.1 et, dans les notations de ce lemme, d'appliquer les méthodes du chapitre précédent sur μ_K^Y pour faire apparaître la taille de l'échantillon effectif dans les bornes afin de tenir compte de la rareté des événements rencontrés.

Nous commençons par un lemme.

Lemme 3.6. *Dans les mêmes notations qu'au lemme 3.1, nous avons*

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] \leq \mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right] + \sqrt{\frac{p}{n}}.$$

Démonstration. Par le lemme 3.1, nous avons que

$$\begin{aligned} \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right] &= \mathbb{E} \left[\mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \mid K \right] \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^K \mathbb{1}\{Y_i \in A\} - \mu(A) \right| \mid K \right] \right] \\ &= \mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^K \mathbb{1}\{Y_i \in A\} - \mu(A) \right| \right]. \end{aligned}$$

Par l'inégalité triangulaire,

$$\begin{aligned} \mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^K \mathbb{1}\{Y_i \in A\} - \mu(A) \right| \right] &= \mathbb{E} \left[\sup_{A \in \mathcal{A}} \left| \frac{K}{n} (\mu_K^Y(A) - \mu_{\mathbb{A}}(A)) + \left(\frac{K}{n} - p \right) \mu_{\mathbb{A}}(A) \right| \right] \\ &\leq \mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right] + \mathbb{E} \left[\left| \frac{K}{n} - p \right| \right], \end{aligned}$$

où nous avons utilisé le fait que $\mu_{\mathbb{A}}$ est une distribution de probabilité sur $\mathbb{R}^d$. Le second terme se borne alors aisément en utilisant l'inégalité de Jensen et le fait que $K \sim \text{Bin}(n, p)$ car

$$\left(\mathbb{E} \left[\left| \frac{K}{n} - p \right| \right] \right)^2 \leq \mathbb{E} \left[\left(\frac{K}{n} - p \right)^2 \right] = \text{Var} \left[\frac{K}{n} - p \right] = \frac{\text{Var}[K]}{n^2} = \frac{np(1-p)}{n^2} \leq \frac{p}{n},$$

et le résultat suit directement. ■

Afin d'obtenir une borne sur l'espérance $\mathbb{E}[Z]$, il suffit donc d'obtenir une borne sur $\mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right]$. L'inégalité VC pour la moyenne (théorème 2.20) permet de fournir un premier résultat dans ce cadre.

Théorème 3.7. *Dans les mêmes notations qu'au lemme 3.1 et pour tout $\alpha > 1$, nous avons que*

$$\mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right] \leq \exp \left(-\frac{(\alpha - 1)^2 np}{2(1 + (\alpha - 1)/3)} \right) + \sqrt{\frac{2\alpha p}{n} \log(2\mathbb{S}_{\mathcal{A}}(2\alpha np))}.$$

Démonstration. Soit $\alpha > 1$. En distinguant selon que $K \leq \alpha np$ ou non et en observant que $\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \leq 1$, nous avons

$$\begin{aligned} & \mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right] \\ & \leq \mathbb{P}(K \geq \alpha np) + \mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \mathbf{1}_{\{K \leq \alpha np\}} \right]. \end{aligned} \quad (3.7)$$

Le premier terme de (3.7) se borne directement en utilisant l'inégalité de Bernstein (3.3)

$$\begin{aligned} \mathbb{P}(K \geq \alpha np) &= \mathbb{P}(K - np \geq (\alpha - 1)np) \\ &\leq \exp \left(-\frac{(\alpha - 1)^2(np)^2}{2(np + (\alpha - 1)np/3)} \right) = \exp \left(-\frac{(\alpha - 1)^2 np}{2(1 + (\alpha - 1)/3)} \right). \end{aligned}$$

Le second terme de (3.7) se borne en utilisant un argument de conditionnement et l'inégalité VC pour la moyenne (théorème 2.20)

$$\begin{aligned} & \mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \mathbf{1}_{\{K \leq \alpha np\}} \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \mathbf{1}_{\{K \leq \alpha np\}} \mid K \right] \right] \\ &\leq \mathbb{E} \left[\frac{K}{n} \sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2K))}{K}} \mathbf{1}_{\{K \leq \alpha np\}} \right] \\ &\leq \sqrt{\frac{2\alpha p}{n}} \log(2\mathbb{S}_{\mathcal{A}}(2\alpha np)), \end{aligned}$$

où la dernière ligne suit par monotonie. ■

Le paramètre α apparaissant dans la borne du théorème 3.7 est choisi par l'utilisateur en pratique et dépend de la taille de son échantillon effectif np . Si cette dernière est trop petite, alors certainement α devrait être choisi assez grand pour rendre le terme exponentielle très petit, au prix d'une légère croissance dans le second terme (si le coefficient de désintégration de sa classe $\mathcal{A}$ n'explose pas trop rapidement). Si, à l'inverse, np est déjà très grand, alors α peut être choisi plus petit car l'exponentielle sera déjà très proche de zéro.

Si la dimension VC de la classe $\mathcal{A}$ est finie, alors il n'est pas nécessaire d'utiliser la monotonie. Nous pouvons directement utiliser l'inégalité de Jensen et obtenir une borne plus fine que celle obtenue en appliquant le corollaire 2.8 à la borne du théorème 3.7 [25, Théorème 5].

Théorème 3.8. *Dans les mêmes notations qu'au lemme 3.1 et si la classe $\mathcal{A}$ a une dimension VC finie $V < +\infty$, alors*

$$\mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right] \leq \sqrt{\frac{2p}{n} [\log(2) + V \log(2np + 1)]}.$$

Démonstration. Par l'inégalité VC pour la moyenne (théorème 2.20) et par le corollaire 2.8, nous avons

$$\begin{aligned} \mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right] &= \mathbb{E} \left[\mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \mid K \right] \right] \\ &\leq \mathbb{E} \left[\frac{K}{n} \sqrt{\frac{2 \log(2\mathbb{S}_{\mathcal{A}}(2K))}{K}} \right] \\ &\leq \frac{1}{n} \mathbb{E} \left[\sqrt{2K [\log(2) + V \log(2K + 1)]} \right] \\ &\leq \frac{1}{n} \sqrt{2 \mathbb{E}[K] [\log(2) + V \log(2 \mathbb{E}[K] + 1)]} \\ &= \sqrt{\frac{2p}{n} [\log(2) + V \log(2np + 1)]}, \end{aligned}$$

où l'avant-dernière ligne suit de la concavité de $K \mapsto \sqrt{2K [\log(2) + V \log(2K + 1)]}$ et la dernière ligne suit du fait que $K \sim \text{Bin}(n, p)$. ■

La méthode de chaînage décrite dans le chapitre 2 représente une manière alternative d'obtenir une borne sur l'espérance $\mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right]$ apparaissant dans le lemme 3.6. Comme nous l'avons vu (théorème 2.24), cette méthode est particulièrement efficace quand la dimension VC de la classe $\mathcal{A}$ est finie. Ainsi, nous nous plaçons directement dans ce cadre pour énoncer le résultat suivant.

Théorème 3.9. *Dans les mêmes notations qu'au lemme 3.1 et si la classe $\mathcal{A}$ a une dimension VC finie $V < +\infty$, alors*

$$\mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right] \leq 11\sqrt{V} \sqrt{\frac{p}{n}}.$$

Démonstration. Le théorème 2.24 et l'inégalité de Jensen impliquent directement que

$$\begin{aligned} \mathbb{E} \left[\frac{K}{n} \sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \right] &= \mathbb{E} \left[\frac{K}{n} \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)| \mid K \right] \right] \\ &\leq \frac{11}{n} \sqrt{V} \mathbb{E} [\sqrt{K}] \leq \frac{11}{n} \sqrt{V} \sqrt{\mathbb{E}[K]} = 11\sqrt{V} \sqrt{\frac{p}{n}}, \end{aligned}$$

où la dernière égalité suit du fait que $K \sim \text{Bin}(n, p)$. ■

Encore une fois, comme dans le cadre classique du chapitre 2, ce résultat est meilleur que tous les autres en termes de convergence ; néanmoins, pour des tailles d'échantillon réalistes, la constante 11 apparaissant dans la borne peut limiter son utilité en pratique. Dans la section suivante, nous synthétisons les différents résultats obtenus dans ce chapitre et procédons à une comparaison graphique de ceux-ci pour la fonction de répartition empirique.

3.4 Comparaison des bornes

Nous avons obtenu différentes bornes pour $Z = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$, adaptées au cadre (3.1), qui tiennent toutes avec probabilité $1 - \delta$, pour $\delta \in (0, 1)$:

- Le théorème 3.2 (étude de la distribution symétrisée) :

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2p}{n} [\log(8/\delta) + \log(\mathbb{S}_{\mathcal{A}}(8np))]} + \sqrt{\frac{2p}{n}}.$$

- Le théorème 3.3 (étude de la distribution non symétrisée) :

$$\begin{aligned} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \\ \leq \frac{2}{3n} \log(4/\delta) + \sqrt{\frac{p}{n}} \left\{ \sqrt{2 \log(4/\delta)} + 2\sqrt{\log(8/\delta) + \log(\mathbb{S}_{\mathcal{A}}(4np))} + 1 \right\}. \end{aligned}$$

- Le théorème 3.4 combiné au théorème 3.5 (application du résultat de Massart [23]) :

$$\begin{aligned} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \\ \leq \frac{1}{n} \left\{ \frac{2}{3} \log(1/\delta) + 8 \log(2\mathbb{S}_{\mathcal{A}}(2n)) \right\} + \sqrt{\frac{p}{n}} \left\{ 2\sqrt{\log(1/\delta)} + \sqrt{32 \log(2\mathbb{S}_{\mathcal{A}}(2n))} \right\}. \end{aligned}$$

- Le théorème 3.4 combiné au lemme 3.6 et au théorème 3.7 (application de l'inégalité VC pour la moyenne et d'un argument de monotonie) : pour tout $\alpha > 1$,

$$\begin{aligned} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \\ \leq \frac{2}{3n} \log(1/\delta) + \exp\left(-\frac{(\alpha-1)^2 np}{2(1+(\alpha-1)/3)}\right) + \sqrt{\frac{p}{n}} \left\{ 2\sqrt{\log(1/\delta)} + \sqrt{2\alpha \log(2\mathbb{S}_{\mathcal{A}}(2\alpha np))} \right\}. \end{aligned}$$

- Si la dimension VC de $\mathcal{A}$ est finie, le théorème 3.4 combiné au lemme 3.6 et au théorème 3.8 (application de l'inégalité VC pour la moyenne et de l'inégalité de Jensen) :

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \frac{2}{3n} \log(1/\delta) + \sqrt{\frac{p}{n}} \left\{ 2\sqrt{\log(1/\delta)} + \sqrt{2[\log(2) + V \log(2np + 1)]} \right\}.$$

- Si la dimension VC de $\mathcal{A}$ est finie, le théorème 3.4 combiné au lemme 3.6 et au théorème 3.9 (application de la méthode de chaînage) :

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \frac{2}{3n} \log(1/\delta) + \sqrt{\frac{p}{n}} \left\{ 2\sqrt{\log(1/\delta)} + 11\sqrt{V} \right\}.$$

La figure 3.1 compare, pour $\delta = 1\%$, ces différentes bornes pour la fonction de répartition empirique, avec $p = 10^{-3}$ (les deux axes sont en échelle logarithmique). Le choix de α dans la borne obtenue via le théorème 3.7 a été choisi via la fonction `optimize` de R.

Commençons par comparer les deux bornes obtenues à travers l'étude de la distribution : la courbe noire (théorème 3.2) a été obtenue en profitant d'une meilleure symétrisation dans le cas des événements rares (comme expliqué dans la démonstration) et en appliquant le « conditioning trick » à une version symétrisée de Z (ce qui a rendu son application un peu plus délicate). La courbe verte (théorème 3.3) a, quant à elle, été obtenue sans symétrisation préalable à l'application du « conditioning trick », cette dernière intervenant seulement dans l'application de l'inégalité VC améliorée (théorème 2.17). Il ressort de la figure 3.1 que, pour des échantillons de taille raisonnable, commencer par symétriser Z est préférable, bien qu'asymptotiquement, les deux méthodes semblent tout à fait comparables.

Au niveau des bornes obtenues par une inégalité de concentration et d'un contrôle sur l'espérance $E[Z]$, nous remarquons directement que l'application du résultat de Massart décrit dans le théorème 2.21 (courbe mauve) fournit une borne largement moins performante que celles obtenues par les techniques de conditionnement décrites dans ce chapitre. Le reste de l'analyse est assez semblable à celle effectuée au chapitre 2 (section 2.3) : la borne obtenue par l'inégalité de Jensen (courbe rouge) semble meilleure que son homologue obtenue par un simple argument de monotonie (courbe bleue), et ce, peu importe la taille de l'échantillon. La méthode de chaînage (courbe orange) est, comme dans le cas classique, la meilleure technique du point de vue asymptotique et la moins bonne pour des échantillons réalistes (en raison des constantes, trop élevées, qui apparaissent dans les bornes) ; à titre d'exemple, toujours dans le cadre de la fonction de répartition empirique avec $p = 10^{-3}$, la borne déduite du théorème 3.9 ne devient plus petite que celle déduite du théorème 3.8 que pour

$$n \geq 500 \exp \left\{ \frac{121}{2} - \log(2) \right\} - 1 \approx 6.97 \times 10^{28}.$$

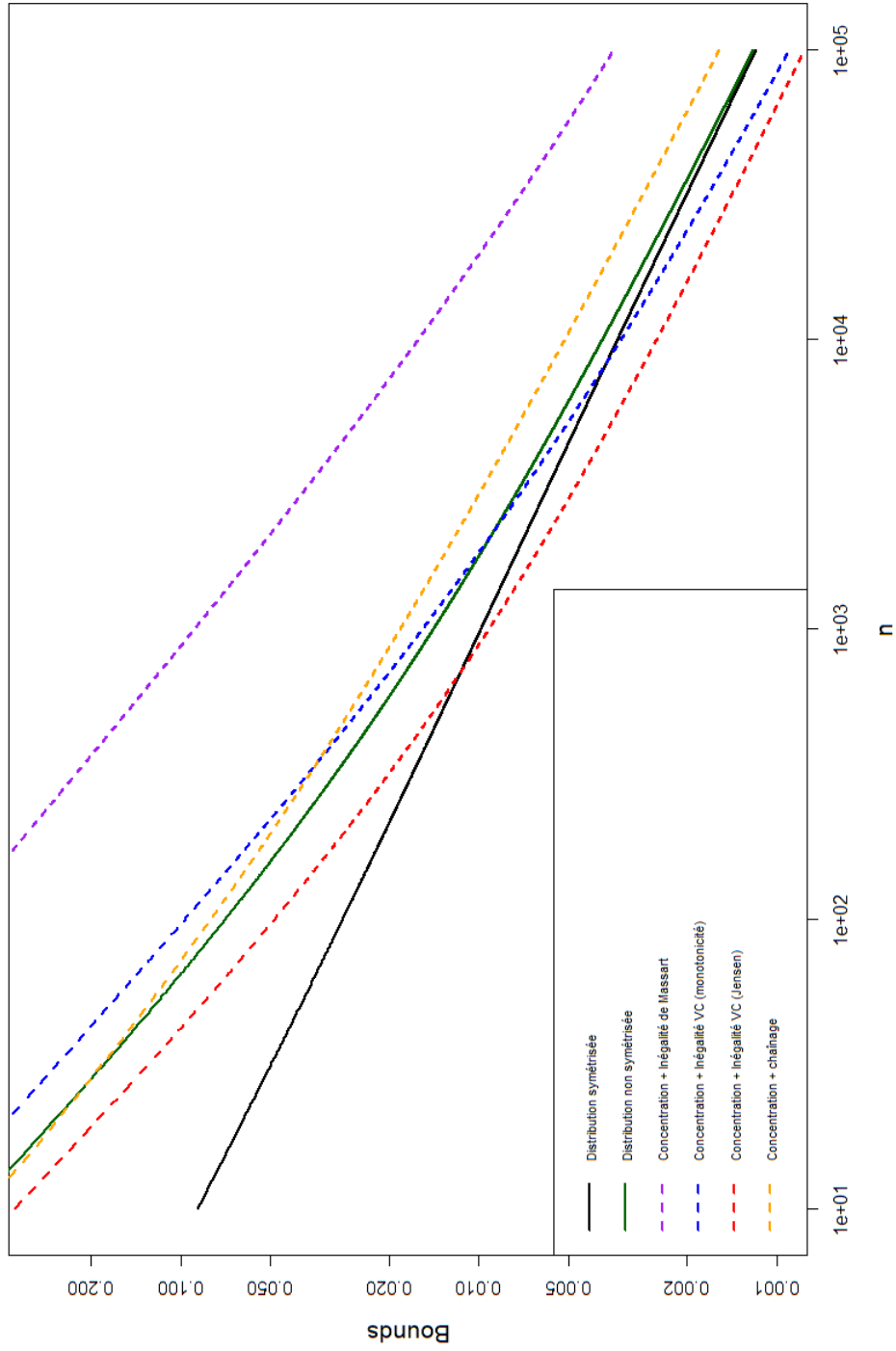


FIGURE 3.1 – Comparaison des bornes pour la fonction de répartition empirique avec $\delta = 1\%$. Les inégalités obtenues à partir de la distribution sont en traits continus et les inégalités obtenues à partir d'une borne sur l'espérance sont en traits discontinus.

Conclusion

Dans le labyrinthe d'inégalités proposées dans ce texte, nous espérons ne pas avoir perdu le lecteur. Au contraire, notre souhait est qu'il ait pu apprécier les différentes étapes menant au développement d'inégalités de concentration adaptées au cadre des événements rares. Au travers de ce chemin, de nombreuses techniques ont été développées et nous sommes convaincus que celles-ci pourraient être utiles à tout utilisateur de la théorie des probabilités. Tout d'abord, nous avons abordé les inégalités pour les sommes de variables aléatoires indépendantes (et plus généralement, les fonctions de variables aléatoires indépendantes) qui sont utilisées dans de très divers contextes et peuvent être démontrées de nombreuses manières [6]. Ensuite, nous avons exposé la théorie de Vapnik et Chervonenkis qui joue un rôle majeur dans toute théorie faisant intervenir des probabilités en dimension infinie. Enfin, nous avons proposé des inégalités adaptées aux événements rares qui seront utilisées à la place de la théorie classique dès que nous souhaitons étudier des cas extrêmes.

Notre approche fut différente de celle fréquemment rencontrée dans la littérature sur les processus empiriques (par exemple, [28]) : nos processus sont indexés par une classe de boréliens de $\mathbb{R}^d$ et non une classe de fonctions. Ce choix a principalement été motivé par une plus grande simplicité de visualisation des concepts combinatoires développés dans le second chapitre lorsqu'ils sont appliqués à des classes de boréliens. Il serait erroné de prétendre que cela ne représente pas une perte de généralité, cependant, notre motivation initiale étant le problème de la classification (comme décrit dans l'introduction), nous pouvions nous restreindre à ce cadre tout en restant général (voir [15, Chapitre 1]). Tous les outils développés pourraient être utilisés pour traiter des processus plus généraux et cela représente une piste certaine pour continuer ce travail.

De nombreuses autres directions sont également possibles. Les inégalités de concentration jouent un rôle important dans la statistique moderne. À l'inverse des résultats limites tels que la loi des grands nombres ou le théorème de la limite centrale, elles quantifient pour toute taille d'échantillon les déviations à la moyenne. Évidemment, la très large validité de ces inégalités est associée à une perte de précision (par exemple, si nous utilisons ces résultats afin de construire des régions de confiance, celles-ci seront souvent trop larges). Néanmoins, nous espérons avoir su convaincre le lecteur que des expressions assez fines peuvent tout de même être obtenues. La littérature sur ce sujet est sans cesse grandissante et nous supputons que notre bibliographie permettra au lecteur de trouver quelques premières idées de recherche s'il souhaite aller plus loin.

La théorie des valeurs extrêmes a, elle aussi, drastiquement gagné en popularité ces dernières années, notamment grâce à ses applications très diverses [3]. Même si notre travail ne fait qu’effleurer cette théorie au travers des hypothèses sous-jacentes aux inégalités développées dans le chapitre 3, il permet également d’introduire une piste de recherche dans cette direction. Par exemple, une mesure célèbre dans ce domaine est la *mesure angulaire* (ou mesure spectrale) [16, Section 6.1.4] qui permet de décrire la probabilité de survenance d’un « événement rare » dans une certaine direction asymptotique de l’espace. Il est alors possible de développer des inégalités de concentration pour la mesure angulaire empirique associée à un échantillon i.i.d. autour de la vraie mesure au lieu de travailler avec la mesure empirique classique comme nous l’avons fait.

La théorie du transport optimal donne un moyen de comparaison alternatif entre la mesure empirique et la vraie mesure, via la *distance de Wasserstein* $\mathcal{W}_p(\mu_n, \mu)$ [30, Chapitre 6]. Cette dernière quantité fait l’objet de récents papiers qui utilisent des techniques similaires à celles développées ici, notamment dans l’étude de la valeur moyenne $E[\mathcal{W}_p(\mu_n, \mu)]$ [31].

L’auteur espère que ce mémoire aura pu transmettre au lecteur, au moins en partie, le plaisir qu’a été son écriture et que les nombreuses invitations à continuer l’étude des sujets rencontrés seront une source de motivation à parcourir les très riches domaines que sont les inégalités de concentration et les processus empiriques.

Bibliographie

- [1] ANTHONY, Martin et SHAWE-TAYLOR, John. « A result of Vapnik with applications ». In : *Discrete Applied Mathematics* 47.3 (1993), p. 207-217.
- [2] ASSOUD, Patrick. « Densité et dimension ». fr. In : *Annales de l'institut Fourier* 33.3 (1983), p. 233-282. ISSN : 0373-0956.
- [3] BEIRLANT, Jan et al. *Statistics of extremes : theory and applications*. Wiley series in probability and statistics. Hoboken, NJ : Wiley, 2004. ISBN : 978-0-471-97647-9.
- [4] BENNETT, George. « Probability Inequalities for the Sum of Independent Random Variables ». en. In : *Journal of the American Statistical Association* 57.297 (mar. 1962), p. 33-45. ISSN : 0162-1459.
- [5] BERNSTEIN, Sergei. « Sur une modification de l'inégalité de Tchebychev ». In : *Annals Science Institute Sav. Ukraine. Math. I* (1924).
- [6] BOUCHERON, Stéphane, LUGOSI, Gábor et MASSART, Pascal. *Concentration inequalities : a nonasymptotic theory of independence*. en. 1st ed. OCLC : ocn818449985. Oxford : Oxford University Press, 2013. ISBN : 978-0-19-953525-5.
- [7] BOUSQUET, Olivier, BOUCHERON, Stéphane et LUGOSI, Gábor. « Introduction to Statistical Learning Theory ». In : *Advanced Lectures on Machine Learning*. Sous la dir. d'Olivier BOUSQUET, Ulrike von LUXBURG et Gunnar RÄTSCH. T. 3176. Series Title : Lecture Notes in Computer Science. Berlin, Heidelberg : Springer Berlin Heidelberg, 2004, p. 169-207. ISBN : 978-3-540-23122-6.
- [8] BOUSQUET, Olivier, LUXBURG, Ulrike von et RÄTSCH, Gunnar, éd. *Advanced lectures on machine learning : ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003 [and] Tübingen, Germany, August 4-16, 2003 : revised lectures*. en. Lecture notes in computer science, Lecture notes in artificial intelligence 3176. Berlin ; New York : Springer, 2004. ISBN : 978-3-540-23122-6.
- [9] CHERNOFF, Herman. « A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations ». en. In : *The Annals of Mathematical Statistics* 23.4 (déc. 1952), p. 493-507. ISSN : 0003-4851.
- [10] DEVROYE, Luc, GYÖRFI, László et LUGOSI, Gábor. *A Probabilistic Theory of Pattern Recognition*. en. T. 31. Stochastic Modelling and Applied Probability. New York, NY : Springer New York, 1996. ISBN : 978-1-4612-6877-2 978-1-4612-0711-5.
- [11] DUDLEY, R. M. « The sizes of compact subsets of Hilbert space and continuity of Gaussian processes ». In : *Journal of Functional Analysis* 1.3 (1967), p. 290-330. ISSN : 0022-1236.

- [12] EFRON, B. et STEIN, C. « The Jackknife Estimate of Variance ». In : *Annals of Statistics* 9.3 (1981), p. 586-596.
- [13] GOIX, Nicolas, SABOURIN, Anne et CLÉMENÇON, Stéphan. « Learning the dependence structure of rare events : a non-asymptotic study ». In : *arXiv :1505.06298 [math, stat]* (mai 2015).
- [14] GREENBERG, Spencer et MOHRI, Mehryar. « Tight lower bound on the probability of a binomial exceeding its expectation ». en. In : *Statistics & Probability Letters* 86 (mar. 2014), p. 91-98. ISSN : 01677152.
- [15] GYÖRFI, László, éd. *Principles of nonparametric learning : course material of the summer school, Udine, July 9 - 13, 2001*. eng. CISM courses and lectures 434. Meeting Name : Summer School Principles of Nonparametric Learning. Wien : Springer, 2002. ISBN : 978-3-211-83688-0.
- [16] HAAN, L. de et FERREIRA, Ana. *Extreme value theory : an introduction*. en. Springer series in operations research. New York ; London : Springer, 2006. ISBN : 978-0-387-23946-0.
- [17] HABIB, Michel et al., éd. *Probabilistic Methods for Algorithmic Discrete Mathematics*. English. Algorithms and Combinatorics 16. Berlin : Springer Berlin, 1998. ISBN : 978-3-642-08426-3.
- [18] HAINAUT, Donatien. *LMAT2470 - Processus stochastiques*. Français. Syllabus. Louvain-la-Neuve : UCLouvain, 2019.
- [19] HAUSSLER, David. « Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension ». en. In : *Journal of Combinatorial Theory, Series A* 69.2 (fév. 1995), p. 217-232.
- [20] Hoeffding, Wassily. « Probability Inequalities for Sums of Bounded Random Variables ». en. In : *Journal of the American Statistical Association* 58.301 (mar. 1963), p. 13-30. ISSN : 0162-1459.
- [21] LEHMANN, E. L. et ROMANO, Joseph P. *Testing statistical hypotheses*. en. 3rd ed. Springer texts in statistics. New York : Springer, 2005. ISBN : 978-0-387-98864-1.
- [22] MASSART, Pascal. « The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality ». In : *The Annals of Probability* 18.3 (1990), p. 1269-1283.
- [23] MASSART, Pascal. « About the constants in Talagrand's concentration inequalities for empirical processes ». en. In : *The Annals of Probability* 28.2 (avr. 2000). ISSN : 0091-1798.
- [24] MASSART, Pascal. *Concentration inequalities and model selection : Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Lecture notes in mathematics 1896. Berlin ; New York : Springer-Verlag, 2007. ISBN : 978-3-540-48497-4.
- [25] SABOURIN, Anne et SEGERS, Johan. *Concentration Inequalities for Rare Events*. (Non publié). 2020.

- [26] SEGERS, Johan. *LMAT1371 - Théorie des probabilités*. Français. Syllabus. Louvain-la-Neuve : UCLouvain, 2016.
- [27] TALAGRAND, M. « Sharper Bounds for Gaussian and Empirical Processes ». en. In : *The Annals of Probability* 22.1 (jan. 1994), p. 28-76. ISSN : 0091-1798.
- [28] VAART, Aad W. van der et WELLNER, Jon A. *Weak Convergence and Empirical Processes*. en. Springer Series in Statistics. New York, NY : Springer New York, 1996. ISBN : 978-1-4757-2547-6.
- [29] VAPNIK, V N et CHERVONENKIS, A Ya. « On the uniform convergence of relative frequencies of events to their probabilities (translated version) ». en. In : *Theory of probability and its applications* XVI.2 (1971), p. 264-280.
- [30] VILLANI, Cédric. *Optimal transport : old and new*. eng. Grundlehren der mathematischen Wissenschaften 338. Berlin : Springer, 2009. ISBN : 978-3-662-50180-1 978-3-540-71049-3 978-3-540-71050-9.
- [31] WEED, Jonathan et BACH, Francis. « Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance ». In : *Bernoulli* 25.4A (nov. 2019). ISSN : 1350-7265.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

Faculté des sciences

Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc