Quatrième partie

Annexes

CHAPITRE 22

Machines à sous

22.1 Corrigé possible du texte

22.1.1 Enoncé du problème et définition du modèle statistique associé

On étudie ici un modèle statistique avec observations incomplètes : on dispose d'observations parcellaires sur un certain phénomène aléatoire, dont on connaît la forme de la loi, mais pas les paramètres. Le phénomène est formé par des couples de variables aléatoires indépendantes $(E_t^A,\ E_t^B)$, à valeurs dans $\{0,\,1\}^2$, eux-mêmes indépendants et identiquement distribués lorsque t varie. La loi commune est donc un produit de lois de Bernoulli, de paramètres notés θ^A et θ^B. Pour éviter les trivialités, on exclut les cas extrêmes $\theta^A, \theta^B \in \{0,\,1\}$, de sorte que l'espace des paramètres est finalement $\Theta =]0,\,1[^2$. Les observations sont incomplètes (ou parcellaires) au sens où l'on n'observe à aucun tour t le vecteur $(E_t^A,\ E_t^B)$ en entier, mais seulement une de ses composantes, au libre choix du statisticien, celle d'index noté $C_t \in \{A,\,B\}$ ci-dessous (ou la première, ou la seconde composante). On note $E_t^{C_t} = X_t$ cette composante observée ; elle forme également le gain obtenu.

L'objectif ici n'est pas directement de bien estimer θ^A et θ^B, il ne s'agit donc pas vraiment d'un problème statistique au sens classique ; mais de faire en sorte que les sommes $E_1^{C_1} + \ldots + E_n^{C_n}$ soient grandes, en un sens précisé au paragraphe 22.1.2. On parle de problème d'apprentissage, et on verra que le joueur aura à arbitrer entre exploration (recueillir des informations sur les paramètres, et notamment, les estimer efficacement) et exploitation de ces informations. Contrairement aux problèmes statistiques habituels, on n'est pas tant intéressé par la bonne description du phénomène stochastique sous-jacent que par les conséquences de cette bonne modélisation en termes d'objectifs concrets (ici, gagner de l'argent !).

Explicitons maintenant le lien avec la réalité et développons la nature de ces objectifs concrets. Tout d'abord, il est bon de rappeler qu'un synonyme de machine à sous est bandit manchot ; une machine à sous a en effet un seul bras. Dans la suite, on va considérer que le joueur fait face à deux machines, ou, plus simplement, à une machine inhabituelle, à deux bras. C'est pourquoi on parle de bandits à deux bras.

Chacun des deux bras, à chaque fois qu'on l'actionne (avec une pièce de 1 €), tire un lot au hasard parmi la liste des lots disponibles (par exemple, 1 €, 2 €, 10 €, etc. jusqu'au gros lot). Les compteurs sont remis à zéro à chaque tirage, la machine est sans mémoire : cela garantit que les couples de paiements sont effectivement indépendants entre eux et identiquement distribués. Le fait que les bras sont sans influence l'un sur l'autre assure quant à lui que ces couples de paiements sont formés de variables aléatoires indépendantes. On note cependant

qu'*a priori*, les paiements ne sont pas donnés par une loi de Bernoulli, mais par une loi plus compliquée, discrète et admettant pour support les lots disponibles. La simplification proposée, dans le texte du chapitre 10, par le modèle de Bernoulli admet deux justifications (et demie).

Premièrement, on pourrait imaginer que l'on ne s'intéresse qu'au fait d'avoir un gain, peu importe son montant : on veut activer le bras qui obtient le plus souvent un gain, quel que soit le montant de ce dernier, ou alors, le plus souvent un gros gain (supérieur à un certain seuil psychologique, ce qui permet de considérer les petits gains comme peu intéressants pour le joueur). La mise de 1 € n'entre alors plus en ligne de compte. C'est une modélisation en termes d'ivresse ou de frisson du gain, pas tellement en termes d'efficacité des gains ; on pourrait notamment se retrouver à actionner presque exclusivement un bras qui gagne souvent mais peu, alors que l'autre bras a une meilleure espérance de gain.

Plus raisonnablement, on pourrait penser que le seul lot disponible est 2 € ; à chaque coup, on gagne ainsi, tous comptes faits et mise déduite, $2\,E_t^{C_t} - 1$ €, que l'on veut comparer au gain moyen du meilleur bras, $2\,\max\{\theta^A,\,\theta^B\} - 1$. La différence,

$$2\Big(E_t^{C_t} - \max\left\{\theta^A,\,\theta^B\right\}\Big)\,,$$

est, au facteur 2 près, la quantité d'intérêt dans le texte et dans ce qui suit. On pourra noter que les variables en jeu ici ne sont plus alors distribuées selon des lois de Bernoulli, mais selon des lois de Rademacher (*id est*, elles ont pour support $\{-1,\,1\}$) ; il est par ailleurs évident que les paramètres θ^A et θ^B doivent alors être plus petits que $1/2$, sans quoi, dans ce modèle, le casino perdrait de l'argent !

Enfin – et c'est la demi-justification –, les calculs ci-dessous, et notamment ceux mettant en jeu la loi des grands nombres pour les martingales, sont valables pour tout couple de lois sur les bras admettant toutes deux un moment d'ordre deux. Or, c'est le cas pour toutes les lois sur les gains des machines à sous, qui ont un support fini. Les lois de Bernoulli offrent une simplification encore suffisamment riche pour illustrer les méthodes de martingales sans trop alourdir ou compliquer le raisonnement ; les mêmes stratégies atteignent l'objectif fixé au paragraphe 22.1.4 sur des lois plus complexes. Seule la dérivation des intervalles de confiance par méthode de réinjection (« plug-in ») au paragraphe 22.1.7 utilise la forme particulière de la variance des lois de Bernoulli et leur estimation consistante à partir d'estimateurs de leur moyenne (de leur paramètre).

Définition mathématique d'une stratégie

On a parlé ci-dessus de stratégies, sans vraiment les définir formellement ; mais en faisant simplement appel à l'intuition que l'on peut s'en faire, à savoir que l'action choisie au tour t repose uniquement sur l'information disponible au début du tour t, qui est formée par la suite des actions et gains passés. Voici maintenant la même définition, mais formulée de manière rigoureuse mathématiquement.

Une stratégie (déterministe) est une suite de variables aléatoires $(C_n)_{n\geqslant 1}$ à valeurs dans $\{A,\,B\}$ telle que C_1 est déterministe, et pour $t \geqslant 2$, la variable aléatoire C_t est mesurable par

rapport à la tribu engendrée par

$$C_1, E_1^{C_1}, \ldots, C_{t-1}, E_{t-1}^{C_{t-1}} ;$$

en réalité, il est équivalent de requérir C_t soit mesurable par rapport à $E_1^{C_1}, \ldots, E_{t-1}^{C_{t-1}}$, comme on le montre par récurrence.

> Remarque au passage. Ce qui suit traite essentiellement des stratégies déterministes ; l'appel à l'aléa n'arrive que dans certains cas d'égalités entre estimées des paramètres des bras, où l'on veut alors tirer A ou B au hasard. Pour modéliser cela, et même des règles aléatoires plus générales, on suppose disposer d'une suite de variables aléatoires $U_1, U_2, \ldots$ indépendantes des gains des bras $(E_t^A,\ E_t^B)$, indépendantes entre elles, et identiquement distribuées selon une loi uniforme. On utilise le fait que toute probabilité sur $\{A, B\}$ peut être simulée grâce à une loi uniforme (voir le chapitre 9) : une stratégie aléatoire est alors une suite de variables aléatoires $(C_n)_{n\geqslant 1}$ telle que C_1 est U_1–mesurable, et pour $t \geqslant 2$, la variable aléatoire C_t est mesurable par rapport à
>
> $$U_1, C_1, E_1^{C_1}, \ldots, U_{t-1}, C_{t-1}, E_{t-1}^{C_{t-1}}, U_t .$$
>
> Ici encore, on peut montrer que les choix passés ne sont pas nécessaires dans la définition de la filtration.

22.1.2 Premiers exemples de stratégies et limitations de toute stratégie

Le but d'une stratégie est de recevoir un paiement moyen aussi proche que possible de celui du meilleur bras. On souhaite donc que

$$G_n = \frac{1}{n}\sum_{t=1}^{n} X_t = \frac{1}{n}\sum_{t=1}^{n} E_t^{C_t}$$

ne soit pas trop éloigné du gain du meilleur bras,

$$\max\left\{\frac{1}{n}\sum_{t=1}^{n} E_t^A,\ \frac{1}{n}\sum_{t=1}^{n} E_t^B\right\} \approx \max\left\{\theta^A, \theta^B\right\},$$

où l'approximation procède de la loi des grands nombres (et peut être quantifiée par les techniques du chapitre 3).

Certaines stratégies naïves ne conviennent pas, notamment celles qui jouent toujours le même bras, $C_t = A$ pour tout t ou $C_t = B$ pour tout t. En effet, selon les lois en question, l'une ou l'autre de ces deux stratégies, pas toujours la même, obtient un mauvais paiement moyen, égal à celui du plus mauvais bras. La stratégie qui alterne entre A et B et joue $C_{2t-1} = A$ puis $C_{2t} = B$ pour tout $t \geqslant 1$ n'obtient asymptotiquement qu'un paiement moyen égal à la moyenne de θ^A et θ^B.

On montre ci-dessous, au Corollaire 22.3, que le gain moyen G_n de toute stratégie vérifie

$$\min\left\{\theta^A,\ \theta^B\right\} \leqslant \liminf G_n \leqslant \limsup G_n \leqslant \max\left\{\theta^A,\ \theta^B\right\} \qquad \text{p.s.} \qquad (22.1)$$

Remarque au passage. Notons qu'ici, nous sommes en situation de prise de décisions séquentielles, contrairement aux situations de statistique plus classiques où toutes les observations sont disponibles avant l'estimation ; ce ne sont pas des assertions de convergence en probabilité que l'on désire obtenir alors, mais de convergence presque sûre, qui seules traitent du comportement individuel de toutes les trajectoires.

Au vu de ces inégalités asymptotiques, une stratégie pourra être considérée comme aussi bonne que possible lorsque son paiement moyen G_n tend effectivement vers la borne supérieure $\max\{\theta^A,\ \theta^B\}$; on verra qu'il existe de telles stratégies. Cette définition de bonne stratégie est un peu frustre au sens où il faudrait également s'intéresser à la vitesse de convergence du gain moyen vers la borne supérieure $\max\{\theta^A,\ \theta^B\}$. On n'a envie de considérer que des stratégies qui convergent suffisamment vite vers la limite. Le paragraphe 22.1.7 montre qu'une vitesse de convergence au moins en $1/\sqrt{n}$ est possible.

Pour que le reste du traitement du problème ne soit pas trivial, on suppose évidemment que $\theta^A \neq \theta^B$; pour le cas $\theta^A = \theta^B$, toutes les stratégies sont bonnes au sens défini ci-dessus et l'on n'a que des égalités dans (22.1). Pire, la loi de G_n est en fait indépendante de la stratégie employée, nG_n suit toujours une loi binomiale de paramètres n et $\theta^A = \theta^B$.

22.1.3 Preuve de la loi des grands nombres pour les martingales

Le Corollaire 22.3 découlera des résultats de martingale suivants. On commence par un bref retour sur les liens entre martingales et accroissements de martingale, et on démontre la loi des grands nombres pour les martingales.

Liens entre martingales et accroissements de martingale

On introduit un raccourci de notation efficace : étant donnée une filtration $(\mathcal{F}_n)_{n\geqslant 0}$, on note $\mathbb{E}_t$ l'espérance conditionnelle par rapport à $\mathcal{F}_{t-1}$, pour $t \geqslant 1$. (Notez bien le décalage d'indice.) On définit ci-dessous la notion d'accroissements de martingale à partir de la notion de martingale et on explique les liens entre deux tels processus.

On part d'une martingale $(M_n)_{n\geqslant 1}$ adaptée à une filtration $(\mathcal{F}_n)_{n\geqslant 1}$, que, sans perte de généralité, on étend à une martingale $(M_n)_{n\geqslant 0}$ adaptée à la filtration $(\mathcal{F}_n)_{n\geqslant 0}$, où $M_0 = \mathbb{E}\big[M_1\big]$ est l'espérance commune des éléments de la martingale et $\mathcal{F}_0$ est la tribu triviale (réduite à l'ensemble vide et à l'espace probabilisé tout entier). On définit alors la suite $(Y_n)_{n\geqslant 0}$ de ses accroissements par $Y_0 = M_0$ ($= y_0$, où y_0 est une constante), et $Y_t = M_t - M_{t-1}$ pour $t \geqslant 1$. La suite $(Y_n)_{n\geqslant 0}$ est également $(\mathcal{F}_n)_{n\geqslant 0}$–adaptée, et $\mathbb{E}_t\big[Y_t\big] = \mathbb{E}_t\big[M_t\big] - M_{t-1} = 0$ pour tout $t \geqslant 1$, car $(M_n)_{n\geqslant 0}$ est une martingale.

Réciproquement, une suite $(Y_n)_{n\geqslant 1}$ qui est $(\mathcal{F}_n)_{n\geqslant 1}$–adaptée et telle que $\mathbb{E}_t\big[Y_t\big] = 0$ pour tout $t \geqslant 1$ est appelée une suite d'accroissements de martingale. Le processus $(M_n)_{n\geqslant 0}$ défini par $M_0 = y_0$ (une constante librement choisie, éventuellement nulle) et $M_t = y_0 + Y_1 + \ldots + Y_t$ pour $t \geqslant 1$, est alors une martingale adaptée par rapport à $(\mathcal{F}_n)_{n\geqslant 0}$.

Enoncé de la loi des grands nombres pour les martingales

On part donc dans le texte d'une martingale $(M_n)_{n\geqslant 1}$ ou $(M_n)_{n\geqslant 0}$, que l'on décompose comme indiqué ci-dessus en la suite $(Y_t)_{t\geqslant 0}$ de ses accroissements. On définit alors le processus $(V_n)_{n\geqslant 0}$ comme la somme des variances conditionnelles des accroissements ; mais ceux-ci étant d'espérances conditionnelles nulles, $(V_n)_{n\geqslant 0}$ est simplement la somme des espérances conditionnelles des carrés des accroissements. Avec les notations ci-dessus, pour tout $n \geqslant 0$,

$$V_n = y_0^2 + \sum_{t=1}^{n} \mathbb{E}_t\Big[Y_t^2\Big],$$

où la sommation est évidemment nulle dans la définition lorsque $n = 0$, de sorte que $V_0 = y_0^2$.

$(V_n)_{n\geqslant 0}$ est une suite croissante de variables aléatoires, elle admet donc p.s. une limite $V_\infty = \lim V_n$. On remarque également, cela servira par la suite, qu'un élément V_t est mesurable par rapport à $\mathcal{F}_{t-1}$ (et pas seulement par rapport à $\mathcal{F}_t$). La suite $(V_n)_{n\geqslant 0}$ est donc $(\mathcal{F}_{n-1})_{n\geqslant 0}$– adaptée (où $\mathcal{F}_{-1} = \mathcal{F}_0$ est également la tribu triviale) ; ce qu'on appelle également le caractère prévisible du processus $(V_n)_{n\geqslant 0}$. La terminologie usuelle est que (V_n) est le compensateur prévisible du processus $\big(M_n^2\big)$. En effet, $\big(M_n^2\big)$ est une sous-martingale positive (par inégalité de Jensen), mais $\big(M_n^2 - V_n\big)$ est une martingale : pour tout $t \geqslant 1$,

$$\begin{aligned}\mathbb{E}\Big[M_t^2 - V_t \,\Big|\, \mathcal{F}_{t-1}\Big] &= \mathbb{E}\Big[\Big(Y_t^2 + 2M_{t-1}Y_t + M_{t-1}^2\Big) - \Big(V_{t-1} + \mathbb{E}\Big[Y_t^2 \,\Big|\, \mathcal{F}_{t-1}\Big]\Big) \,\Big|\, \mathcal{F}_{t-1}\Big] \\ &= 2M_{t-1}\,\mathbb{E}\Big[Y_t \,\Big|\, \mathcal{F}_{t-1}\Big] + M_{t-1}^2 - V_{t-1} = M_{t-1}^2 - V_{t-1}\,,\end{aligned}$$

où l'on a essentiellement utilisé que les processus étaient adaptés à la filtration et que (Y_n) est une suite d'accroissements de martingale. On peut maintenant énoncer l'extension de la loi forte des grands nombres aux martingales.

Théorème 22.1 (Loi des grands nombres pour les martingales) *Avec les notations précédentes,*

1. *sur* $\big\{(V_n) \text{ est bornée}\big\}$, *la martingale* (M_n) *converge p.s. dans* $\mathbb{R}$;
2. *si* $V_n \to \infty$ *p.s., alors* $M_n = o(V_n)$ *p.s.*

Preuve du point 1. du Théorème 22.1

Preuve Pour tout $c > 0$, on considère la variable aléatoire

$$\nu_c = \inf\big\{n \geqslant 0 : V_{n+1} > c\big\}$$

(avec la convention habituelle que $\inf \emptyset = \infty$). ν_c est un temps d'arrêt pour la filtration $(\mathcal{F}_n)$; en effet, pour tout entier k, au vu de la croissance de (V_n), on peut réécrire

$$\{\nu_c = k\} = \{V_{k+1} > c \text{ et } V_k \leqslant c\},$$

qui est $\mathcal{F}_k$–mesurable, puisque (V_n) est prévisible. Par théorème d'arrêt, en notant $n \wedge \nu_c = \min\{n,\ \nu_c\}$, on a que $(M_{n\wedge\nu_c})$ est encore une $(\mathcal{F}_n)$–martingale. On calcule maintenant le moment d'ordre deux de ses éléments ; on veut montrer qu'ils sont uniformément bornés (par c). Pour ce faire, on remarque que puisque $\left(M_n^2 - V_n\right)$ est également une martingale, le même théorème d'arrêt assure que $\left(M_{n\wedge\nu_c}^2 - V_{n\wedge\nu_c}\right)$ est elle aussi une $(\mathcal{F}_n)$–martingale. En particulier, le terme indexé par $n = 0$ étant nul, on a que l'espérance commune des éléments de la martingale est nulle : pour tout $n \geqslant 0$,

$$\mathbb{E}\left[M_{n\wedge\nu_c}^2\right] = \mathbb{E}\left[V_{n\wedge\nu_c}\right], \qquad \text{où} \quad \mathbb{E}\left[V_{n\wedge\nu_c}\right] \leqslant c$$

par définition même de ν_c.

Or, on sait qu'une martingale bornée dans $\mathbb{L}^2$ converge p.s. (et dans $\mathbb{L}^2$) : on définit légitimement $M_{\nu_c}^{\infty} = \lim M_{n\wedge\nu_c}$ pour tout $c > 0$ (par exemple pour tout $c \in \mathbb{N}^*$). On pose alors

$$M_\infty = \limsup_{c\to\infty} M_{\nu_c}^{\infty},$$

qui est bien une variable aléatoire (elle est bien mesurable), et on va montrer que (M_n) converge vers M_∞ sur $\{V_\infty < \infty\}$. Pour c fixé, sur l'événement $\{\nu_c = \infty\}$, on a $\nu_{c'} = \infty$ pour tout $c' \geqslant c$, de sorte que

$$M_\infty = \lim M_n \qquad \text{sur} \qquad \bigcup_{c\in\mathbb{N}^*} \{\nu_c = \infty\}\ .$$

Lorsque $V_\infty < \infty$, on a que $\nu_c = \infty$ pour tout $c > V_\infty$; réciproquement, lorsqu'il existe c tel que $\nu_c = \infty$, alors on sait que (V_n) est bornée par c, de sorte que $V_\infty \leqslant c < \infty$. On vient de prouver que

$$\bigcup_{c\in\mathbb{N}^*} \{\nu_c = \infty\} = \{V_\infty < \infty\} = \left\{(V_n) \text{ est bornée}\right\},$$

ce qui conclut la preuve du point 1. du Théorème 22.1. □

> Remarque au passage. Une des meilleures références pour avoir un aperçu rapide des résultats essentiels du cours de probabilités (martingales et chaînes de Markov) est [MPB98]. Au chapitre 3, on pourra relire par exemple l'énoncé du théorème d'arrêt (paragraphe 3.3(v)) et le passage sur la convergence des martingales bornées dans $\mathbb{L}^2$ (Théorème 3.7). Le résultat de convergence ci-dessus est prouvé de manière légèrement différente au Problème 3.7.

Preuve du point 2. du Théorème 22.1

On s'attache maintenant au point 2. du théorème, qui découle du point 1., comme on le verra. Il énonce que $M_n = o(V_n)$ p.s. sur $\{\liminf V_n = \infty\}$, ce qui est une généralisation de la loi forte des grands nombres aux martingales.

En effet, si on a une suite (Z_n) de variables aléatoires indépendantes et identiquement distribuées selon une loi admettant un moment d'ordre deux, d'espérance et de variance

notées respectivement μ et $\sigma^2 > 0$, alors $(Y_t) = (Z_t - \mu)$ est une suite d'accroissements de martingale ; le compensateur prévisible associé au carré de la martingale est $(V_n) = (n\sigma^2)$, il diverge p.s. vers ∞. Le point 2. du Théorème 22.1 assure alors que $Y_1 + \ldots + Y_n = o(n\sigma^2)$, soit exactement

$$\frac{1}{n}\sum_{t=1}^{n} Z_t \longrightarrow \mu \qquad \text{p.s.}$$

Preuve On note, pour tout $t \geqslant 1$,

$$Y'_t = \frac{Y_t}{V_t}\mathbb{I}_{\{t \geqslant \tau\}}$$

où $\tau = \inf\{n : V_n > 0\}$ est un temps d'arrêt prévisible, puisque (V_n) est prévisible. Ainsi, $\{t \geqslant \tau\} = \{V_t > 0\}$ est $\mathcal{F}_{t-1}$–mesurable. V_t étant également $\mathcal{F}_{t-1}$–mesurable et Y_t étant quant à elle $\mathcal{F}_t$–mesurable, on a bien que Y'_t est $\mathcal{F}_t$–mesurable ; de plus,

$$\mathbb{E}\left[\frac{Y_t}{V_t}\mathbb{I}_{\{t \geqslant \tau\}} \,\middle|\, \mathcal{F}_{t-1}\right] = \frac{1}{V_t}\,\mathbb{E}\big[Y_t \mid \mathcal{F}_{t-1}\big]\mathbb{I}_{\{t \geqslant \tau\}} = 0\,,$$

ce qui prouve que (Y'_t) est une suite d'accroissements de martingale. Par ailleurs,

$$\mathbb{E}\left[\left(\frac{Y_t}{V_t}\mathbb{I}_{\{t \geqslant \tau\}}\right)^2 \,\middle|\, \mathcal{F}_{t-1}\right] = \frac{1}{V_t^2}\,\mathbb{E}\big[Y_t^2 \,\big|\, \mathcal{F}_{t-1}\big]\mathbb{I}_{\{t \geqslant \tau\}} = \frac{1}{V_t^2}(V_t - V_{t-1})\mathbb{I}_{\{t \geqslant \tau\}}\,,$$

de sorte que le compensateur prévisible du carré de la martingale associée à (Y'_t) est (V'_n) défini, pour $n \geqslant 1$, par

$$V'_n = \sum_{t=1}^{n} \frac{1}{V_t^2}(V_t - V_{t-1})\mathbb{I}_{\{t \geqslant \tau\}}\,.$$

En utilisant pour l'inégalité le caractère croissant de (V_n), il vient que sur $\{\tau < \infty\}$,

$$\begin{aligned} V'_n &= \frac{1}{V_\tau}\mathbb{I}_{\{n \geqslant \tau\}} + \sum_{t=1}^{n} \mathbb{I}_{\{t \geqslant \tau+1\}} \int_{V_{t-1}}^{V_t} \frac{1}{V_t^2}\,\mathrm{d}x \\ &\leqslant \frac{1}{V_\tau} + \sum_{t=\tau+1}^{+\infty} \int_{V_{t-1}}^{V_t} \frac{1}{x^2}\,\mathrm{d}x = \frac{1}{V_\tau} + \int_{V_\tau}^{\infty} \frac{1}{x^2}\,\mathrm{d}x = \frac{2}{V_\tau} < \infty \end{aligned}$$

En fait, $\tau < \infty$ p.s., car $V_n \to \infty$ p.s. par hypothèse. Ainsi, la suite (V'_n) est p.s. bornée et par le point 1. du théorème,

$$\sum Y'_t = \sum \frac{Y_t}{V_t}\mathbb{I}_{\{t \geqslant \tau\}}$$

est p.s. une série convergente. On utilise encore une fois le fait que $\tau < \infty$ p.s. pour voir qu'il suffit de montrer que

$$\frac{1}{V_n}\sum_{t=\tau}^{n} Y_t \longrightarrow 0 \text{ p.s.} \qquad \text{pour avoir} \qquad M_n = o(V_n) \text{ p.s.}$$

Le résultat attendu procède de l'application du lemme de Kronecker (rappelé ci-dessous) aux suites de terme général respectivement défini, pour $k \geqslant 1$, par $u_k = Y_{k+\tau-1}/V_{k+\tau-1}$ et $b_k = V_{k+\tau-1}$:

$$\frac{1}{V_n}\sum_{t=\tau}^{n} Y_t = \frac{1}{b_{n-\tau+1}}\sum_{k=1}^{n-\tau+1} b_k u_k \longrightarrow 0 \qquad \text{p.s.}$$

□

Remarque au passage. On rappelle le lemme de Kronecker ; il est par exemple utilisé dans certaines preuves de la loi des grands nombres, voir [GS01, exercices du paragraphe 7.8].

Lemme 22.2 (Kronecker) Pour une série convergente (u_n) et pour tout suite croissante de réels strictement positifs (b_n) divergeant vers $+\infty$, on a

$$\lim_{n\to\infty} \frac{1}{b_n}\sum_{k=1}^{n} b_k u_k = 0\,.$$

Preuve On note $S_n = u_1 + \ldots + u_n$ pour $n \geqslant 1$ (et $S_0 = 0$) ; par hypothèse, il existe un réel L tel que $S_n \to L$. Une transformation d'Abel donne

$$\frac{1}{b_n}\sum_{k=1}^{n} b_k u_k = \frac{1}{b_n}\sum_{k=1}^{n} b_k\big(S_k - S_{k-1}\big) = S_n - \frac{1}{b_n}\sum_{k=1}^{n-1}\big(b_{k+1} - b_k\big)S_k \longrightarrow L - L = 0$$

par application d'une version généralisée du lemme de Cesaro au second terme (utilisant que $b_n \to \infty$ et que $b_{k+1} - b_k \geqslant 0$ pour tout k assez grand). □

22.1.4 Objectif optimal et définition de la notion de bonne stratégie

Nous sommes maintenant en mesure de prouver les inégalités donnant les possibilités et limitations de toute stratégie dans le cadre de jeux de bandits. La preuve du corollaire 22.3 montre d'ailleurs que même si l'on observait tous les paiements, et pas seulement celui du bras que l'on a utilisé, ces mêmes encadrements tiendraient encore.

Corollaire 22.3 La suite des gains moyens (G_n) de toute stratégie de jeu vérifie

$$\min\left\{\theta^A,\ \theta^B\right\} \leqslant \liminf G_n \leqslant \limsup G_n \leqslant \max\left\{\theta^A,\ \theta^B\right\} \qquad \text{p.s.}$$

On a donc désormais un critère nous permettant de dire quelles sont les stratégies que l'on apprécie : celles dont le gain moyen G_n atteint p.s., à la limite, la borne supérieure $\max\left\{\theta^A,\ \theta^B\right\}$ des inégalités du Corollaire 22.3. On a vu plus haut, au paragraphe 22.1.2, qu'il y a *a priori* un peu de travail à fournir, et l'étude des stratégies naïves qui suivent va confirmer ce sentiment. En attendant, nous prouvons le corollaire.

Preuve On rappelle qu'on note $X_t = E_t^{C_t}$ le gain obtenu au tour t. Les accroissements de martingale que l'on va considérer sont donnés par

$$Y_t = X_t - \theta^{C_t}$$

pour $t \geqslant 1$ (et $Y_0 = 0$). Ils sont $(\mathcal{G}_t)$–adaptés, où pour $t \geqslant 1$, on définit $\mathcal{G}_t$ comme la tribu associée à la définition d'une stratégie, celle engendrée par les variables aléatoires

$$\mathcal{G}_t = \sigma\big(U_1, C_1, X_1, \ldots, U_t, C_t, X_t, U_{t+1}\big)$$

(et $\mathcal{G}_0$ est la tribu engendrée par U_1) ; on pourra se reporter à la fin du paragraphe 22.1.1. On note en particulier que C_{t+1} est $\mathcal{G}_t$–mesurable, pour tout $t \geqslant 0$.

On montre maintenant que (Y_t) est une suite d'accroissements de martingale par rapport à $(\mathcal{G}_t)$: pour $t \geqslant 1$,

$$\begin{aligned}\mathbb{E}\big[Y_t \,|\, \mathcal{G}_{t-1}\big] &= \mathbb{E}\Big[X_t - \theta^{C_t} \,\Big|\, \mathcal{G}_{t-1}\Big] = \mathbb{E}\Big[E_t^{C_t} - \theta^{C_t} \,\Big|\, \mathcal{G}_{t-1}\Big] \\ &= \mathbb{I}_{\{C_t=A\}}\,\mathbb{E}\Big[E_t^A - \theta^A \,\Big|\, \mathcal{G}_{t-1}\Big] + \mathbb{I}_{\{C_t=B\}}\,\mathbb{E}\Big[E_t^B - \theta^B \,\Big|\, \mathcal{G}_{t-1}\Big] = 0\,,\end{aligned}$$

où l'on a utilisé, pour l'égalité finale, l'indépendance de E_t^A et E_t^B par rapport à $\mathcal{G}_{t-1}$ et le fait que leurs espérances respectives sont θ^A et θ^B.

On remarque, avec les notations du texte stipulant que N_n^A, N_n^B sont le nombre de fois où les bras A et B ont été joués durant les n premiers tours, que la martingale (M_n) associée à (Y_t) est donnée par

$$M_n = nG_n - \Big(N_n^A\theta^A + N_n^B\theta^B\Big)\,.$$

On calcule maintenant le compensateur prévisible de $\big(M_n^2\big)$: pour tout $n \geqslant 1$,

$$\begin{aligned}V_n &= \sum_{t=1}^{n} \mathbb{E}\Big[Y_t^2 \,\Big\|\, \mathcal{G}_{t-1}\Big] \\ &= \sum_{t=1}^{n} \mathrm{Var}\Big(\mathrm{Ber}\big(\theta^{C_t}\big) \,\Big|\, \mathcal{G}_{t-1}\Big) \\ &= N_n^A\,\theta^A\big(1-\theta^A\big) + N_n^B\,\theta^B\big(1-\theta^B\big) \geqslant n\sigma_*^2\,,\end{aligned}$$

où l'on a noté

$$\sigma_*^2 = \min\Big\{\theta^A\big(1-\theta^A\big),\ \theta^B\big(1-\theta^B\big)\Big\}$$

et où l'on a utilisé que Y_t suit, conditionnellement à C_t, la loi d'une variable de Bernoulli de paramètre θ^{C_t} recentrée par son espérance.

On note que $\sigma_* > 0$ puisque ni θ^A ni θ^B ne sont égaux à 0 ou 1. Ainsi, $V_n \to \infty$ et par loi des grands nombres pour les martingales, $M_n = o(V_n)$ p.s., de sorte que la majoration $V_n \leqslant n$ assure alors que $M_n = o(n)$. Cette comparaison asymptotique se reformule exactement comme

$$G_n - \frac{N_n^A\theta^A + N_n^B\theta^B}{n} \longrightarrow 0 \qquad \text{p.s.} \tag{22.2}$$

Les valeurs d'adhérence de (G_n) sont ainsi celles de la suite des

$$\frac{N_n^A \theta^A + N_n^B \theta^B}{n} ;$$

comme elles sont nécessairement comprises entre $\min\left\{\theta^A,\ \theta^B\right\}$ et $\max\left\{\theta^A,\ \theta^B\right\}$, cela conclut cette preuve. □

22.1.5 Une stratégie naïve : la décision simple selon les succès empiriques

Elle repose sur la considération d'estimateurs des paramètres θ^A et θ^B. Ceux construits au tour $t \geqslant 1$ sont notés $\widehat{\theta}_{t-1}^A$ et $\widehat{\theta}_{t-1}^B$ (parce qu'ils ne dépendent que des résultats des tours 1 à $t-1$, i.e., sont $\mathcal{G}_{t-1}$–mesurables). En particulier, on fixera des valeurs déterministes pour $\widehat{\theta}_0^A$ et $\widehat{\theta}_0^B$.

On considère la stratégie de jeu suivante, définie à partir d'eux et d'une suite $C_1', C_2', \ldots$ de variables aléatoires indépendantes et identiquement distribuées selon la loi uniforme sur $\{A,\ B\}$: pour $t \geqslant 1$,

$$C_t = \begin{cases} A & \text{si } \widehat{\theta}_{t-1}^A > \widehat{\theta}_{t-1}^B, \\ B & \text{si } \widehat{\theta}_{t-1}^A < \widehat{\theta}_{t-1}^B, \\ C_t' & \text{sinon.} \end{cases}$$

Tout est dans la définition des estimateurs : s'ils sont bons (on entend par exemple leur consistance), la stratégie le sera. S'ils sont mauvais, la stratégie ne remplira pas son contrat, qui est d'atteindre p.s. la borne supérieure énoncée au Corollaire 22.3. Les estimateurs les plus immédiats, qui sont définis par le choix d'une constante θ_0 selon

$$\widehat{\theta}_n^C = \frac{1}{N_n^C} \sum_{t=1}^{n} X_t \, \mathbb{I}_{\{C_t = C\}} \quad \text{si} \quad N_n^C = \sum_{t=1}^{n} \mathbb{I}_{\{C_t = C\}} \geqslant 1, \quad \text{et } \widehat{\theta}_n^C = \theta_0 \text{ sinon} \qquad (22.3)$$

pour $C \in \{A,\ B\}$ donnent lieu à la stratégie appelée de décision selon les succès empiriques. Malheureusement, ils ne sont pas automatiquement consistants ici.

On propose un exemple de mauvaise situation pour le cas où $\theta^A < \theta^B$ et $\theta_0 > 0$, les autres cas étant traités par des raisonnements tout à fait similaires.

On se place dans la réalisation où au premier tour, on joue le bras A, qui obtient un gain $E_1^A = 0$, et où au second tour on joue alors B, qui obtient également un gain $E_1^B = 0$. Cette réalisation arrive bien avec probabilité strictement positive. Comme alors

$$\widehat{\theta}_2^A = \widehat{\theta}_2^B = 0 \, ,$$

on a une probabilité $1/2$ de jouer A au tour $t = 3$ et A a une probabilité strictement positive d'obtenir le gain $E_3^A = 1$.

Arrivé dans cette situation, on montre par récurrence qu'on aura $\widehat{\theta}_t^A > 0$ pour tout $t \geqslant 3$ alors qu'on en restera à $\widehat{\theta}_t^B = 0$ et que B ne sera plus jamais tiré.

Cet exemple est un peu extrême au sens où les trois premiers tours de jeu peuvent tout déterminer. Cela dit, même si l'on s'offre un nombre fini $2N$, même arbitrairement grand, de tours d'exploration, où l'on tire chaque bras N fois (peu importe l'ordre) pour construire les estimateurs $\widehat{\theta}^A_{2N}$ et $\widehat{\theta}^B_{2N}$, le même problème survient, certes avec une probabilité plus faible (exponentielle en N). En effet, même avec $\theta^A < \theta^B$, il arrive avec probabilité strictement positive,

$$\left(1-\left(1-\theta^A\right)^N\right)\left(1-\theta^B\right)^N ,$$

que $\widehat{\theta}^B_{2N} = 0$ mais $\widehat{\theta}^A_{2N} > 0$, empêchant la stratégie de décision par les succès empiriques, si l'on utilise pour les tours postérieurs à $2N$, de choisir à nouveau le bon bras B.

> Remarque au passage. Ces arguments peuvent être généralisés à toute procédure statistique se fondant sur un nombre fini (même aléatoire) de tours pour calibrer des estimateurs avant de passer à la décision par succès empiriques : il existe alors toujours une probabilité strictement positive (même si on peut la rendre arbitrairement petite) que (G_n) ne converge pas vers $\max\left\{\theta^A,\ \theta^B\right\}$. Le paragraphe 22.1.8 expliquera pour quel type de critère une telle stratégie avec exploration préalable est intéressante.

Pour l'instant, il est donc clair qu'il faut continuer à affiner nos estimateurs, *id est*, à forcer au besoin la stratégie à jouer de temps en temps chacun des deux bras. On appelle cela l'exploration, par opposition aux tours de jeu où l'on utilise les estimateurs disponibles pour choisir un bras qui nous semble bon : ce sont les tours d'exploitation. Il s'agit de réaliser un compromis entre exploitation et exploration.

22.1.6 Une stratégie plus élaborée : la décision selon les succès empiriques avec excitation

Le manquement dans la stratégie précédente est qu'une exploration suffisante n'était pas garantie. Celle-ci est quantifiée mathématiquement par les conditions $N^A_n \to \infty$ p.s. et $N^B_n \to \infty$ p.s. Ces conditions assurent en retour la consistance forte des estimateurs $\widehat{\theta}^A_n$ et $\widehat{\theta}^B_n$ introduits en (22.3). L'énoncé et la preuve du résultat ci-dessous sont évidemment valables autant pour B que pour A.

Lemme 22.4 Lorsque $N^A_n \to \infty$ p.s., l'estimateur $\widehat{\theta}^A_n$ du paramètre θ^A est fortement consistant.

Preuve On considère la suite de variables aléatoires $\left(Y^A_t\right)$, définie, pour $t \geqslant 1$, par

$$Y^A_t = \left(X_t - \theta^{C_t}\right)\mathbb{I}_{\{C_t=A\}} ;$$

elle forme un accroissement de martingale par rapport à $(\mathcal{G}_t)$, comme on peut le voir par un argument similaire à celui du paragraphe 22.1.4. Il est immédiat que la martingale $\left(M^A_n\right)$ associée est définie par

$$M^A_n = N^A_n\left(\widehat{\theta}^A_n - \theta^A\right)$$

pour $n \geqslant 1$. Or, on a l'équivalence suivante pour la consistance forte désirée :

$$\widehat{\theta}_n^A \longrightarrow \theta^A \text{ p.s.} \qquad \Longleftrightarrow \qquad M_n^A = o\big(N_n^A\big) \text{ p.s.}$$

On veut appliquer la loi des grands nombres pour les martingales. Le compensateur prévisible de $\big((M_n^A)^2\big)$ est $\big(V_n^A\big)$ donné par $V_n^A = N_n^A\,\theta^A\big(1-\theta^A\big)$ pour $n \geqslant 1$, comme on peut le voir par des calculs similaires, encore une fois, à ceux du paragraphe 22.1.4. Or, par hypothèse d'excitation, $N_n^A \to \infty$ p.s., la loi des grands nombres s'applique donc ; $M_n^A = o(N_n^A)$ livre la consistance forte désirée. □

Il s'agit donc de faire en sorte
- que $N_n^A,\ N_n^B \to \infty$ p.s. d'une part, ce qu'on appelle l'exploration ;
- et d'utiliser à bon escient les estimateurs fortement consistants qui en résultent, ce qui forme l'exploitation.

Pour réaliser ce compromis entre exploration et exploitation, on recourt presque à la même stratégie que celle décrite plus haut (décision simple selon les succès empiriques), à ceci près qu'on la force de temps à temps à explorer, en lui faisant choisir un bras qu'elle n'avait pas décidé de jouer. Ces instants sont donnés par une suite (c_k) strictement croissante d'entiers dits instants d'exploration.

Avec les définitions (22.3), et on considérant toujours la même suite (C'_t) qu'au paragraphe précédent, on définit, pour $t \geqslant 1$,

$$C_t = \begin{cases} A & \text{si } t = c_{2k-1} \text{ pour } k \in \mathbb{N}^* \\ B & \text{si } t = c_{2k} \text{ pour } k \in \mathbb{N}^* \\ \left\{ \begin{array}{ll} A & \text{si } \widehat{\theta}_{t-1}^A > \widehat{\theta}_{t-1}^B \\ B & \text{si } \widehat{\theta}_{t-1}^A < \widehat{\theta}_{t-1}^B \\ C'_t & \text{si } \widehat{\theta}_{t-1}^A = \widehat{\theta}_{t-1}^B \end{array} \right\} & \text{aux autres tours } t. \end{cases}$$

Il est clair que le terme général c_k ne peut être proportionnel à k, sans quoi la stratégie résultante tire trop souvent le mauvais bras pour atteindre p.s. le gain moyen souhaité. Ainsi, on impose que $c_k \gg k$; le résultat ci-dessous montre que c'est également une condition suffisante pour une telle convergence. L'excitation corrige de la sorte les errements de la stratégie simple de décision selon les succès empiriques et on peut se sortir des pièges dans lesquels on était tombé dans les premiers tours.

Théorème 22.5 *Lorsque les instants d'exploration sont tels que $c_k \gg k$, la suite (G_n) des gains moyens de la stratégie de décision selon les succès empiriques avec excitation converge vers le gain maximal,*

$$G_n \longrightarrow \max\left\{\theta^A,\ \theta^B\right\} \qquad \text{p.s.}$$

Preuve Grâce à l'excitation, (c_k) étant strictement croissante, on a directement que $N_n^A,\ N_n^B \to \infty$, d'où la consistance des estimateurs, par application du Lemme 22.4. On suppose par exemple que $\theta^A > \theta^B$: on a indiqué au paragraphe 22.1.2 que le cas $\theta^A = \theta^B$ était trivial au sens où toute stratégie atteint l'objectif, et le cas $\theta^A < \theta^B$ est naturellement symétrique au cas que nous allons traiter.

On va prouver ici qu'à partir d'un certain rang, le bras B n'est plus joué que par suite de l'excitation. La consistance des estimateurs signifie exactement que pour tout $\varepsilon > 0$ (pris tel que $\varepsilon < \theta^A - \theta^B$), il existe une variable aléatoire T_0 finie p.s. telle que pour tout $t \geqslant T_0$,

$$\widehat{\theta}_t^B \leqslant \theta^B + \frac{\varepsilon}{2} \leqslant \theta^A - \frac{\varepsilon}{2} \leqslant \widehat{\theta}_t^A \ ; \tag{22.4}$$

ainsi, dans les tours de jeu où la stratégie emploie une décision selon les succès empiriques, B n'est pas jouée à partir du rang T_0. Or N_n^A est, aux temps d'exploration près, le nombre de succès de A, et donc

$$\frac{N_n^A}{n} \geqslant \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}_{\left\{\hat{\theta}_t^A > \hat{\theta}_t^B\right\}} - \frac{1}{n}\sum_{k=1}^{\infty} \mathbb{I}_{\{c_{2k} \leqslant n\}} \ .$$

Par (22.4),

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{I}_{\left\{\hat{\theta}_t^A > \hat{\theta}_t^B\right\}} \geqslant \frac{n - T_0 + 1}{n} \longrightarrow 1 \qquad \text{p.s.},$$

tandis que

$$\sum_{k=1}^{\infty} \mathbb{I}_{\{c_{2k} \leqslant n\}} \leqslant \sum_{k=1}^{\infty} \mathbb{I}_{\{c_k \leqslant n\}} = \max\left\{k \in \mathbb{N}^* :\ c_k \leqslant n\right\} \overset{\text{not.}}{=} c^{-1}(n) \ ,$$

où c^{-1} est appelée un inverse généralisé de l'application $c : \mathbb{N}^* \to \mathbb{N}^*$ définissant la suite (c_k). Puisque $c_k \gg k$ par hypothèse, il vient $c^{-1}(n) \ll n$. Au final,

$$1 \geqslant \frac{N_n^A}{n} \geqslant \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}_{\left\{\hat{\theta}_t^A > \hat{\theta}_t^B\right\}} - \frac{1}{n}\sum_{k=1}^{\infty} \mathbb{I}_{\{c_{2k} \leqslant n\}} \longrightarrow 1 \qquad \text{p.s.},$$

soit, par encadrement, $N_n^A/n \to 1$ p.s. Puis, $N_n^B/n = 1 - N_n^A/n \to 0$ p.s. Or, on a vu en (22.2) que

$$G_n - \frac{N_n^A \theta^A + N_n^B \theta^B}{n} \longrightarrow 0 \qquad \text{p.s.} \ ;$$

soit ici, $G_n \to \theta^A$ p.s. □

Remarque au passage. On a indiqué au paragraphe 22.1.2 pourquoi le cas $\theta^A = \theta^B$ était inintéressant : toutes les stratégies sont équivalentes. On peut malgré tout se demander si, pour la stratégie de décision selon les succès empiriques avec excitation

$$\frac{N_n^A}{n} \to \frac{1}{2} \qquad \text{et} \qquad \frac{N_n^B}{n} \to \frac{1}{2} \qquad \text{p.s.}$$

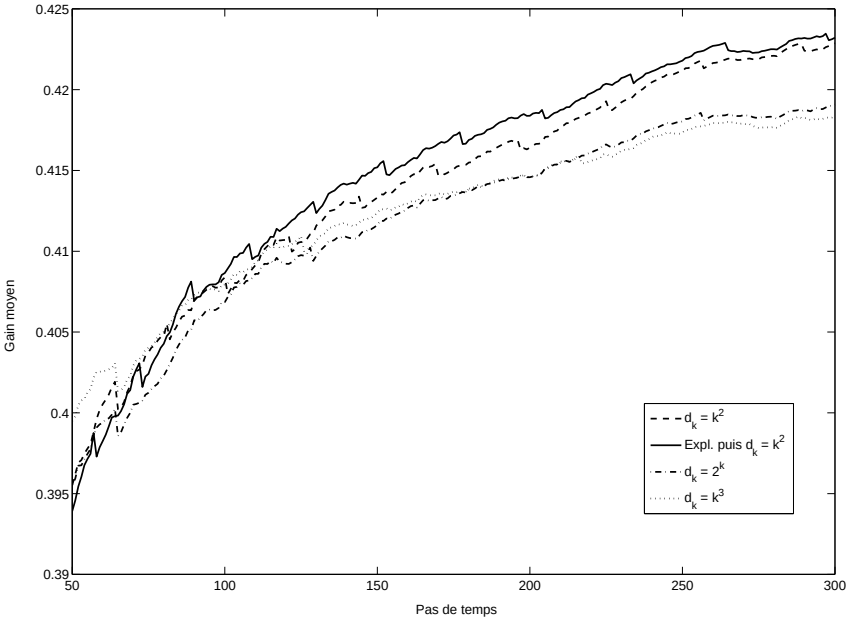

FIGURE 22.1. Performances moyennes (sur 1 000 simulations) de quatre suites d'excitation pour $\theta^A = 0.45$ et $\theta^B = 0.30$.

dans ce cas. C'est évidemment faux pour la stratégie simple, toujours à cause des mêmes pièges dans lesquels on peut tomber en temps fini. Il n'est pas évident, mais pas impossible non plus, que ce soit vrai en dépit de l'excitation, parce que l'échantillonnage des bras dépend de leurs performances passées.

Le choix précis de la suite des instants d'excitation (c_k) est assez libre, hormis la condition $c_k \gg k$. On verra au paragraphe 22.1.7 que des conditions supplémentaires du type $c_k \gg k^{2+\alpha}$, pour un $\alpha > 0$ librement choisi par le joueur, permettent d'obtenir des intervalles de confiance sur le gain. Cela ne nous donne donc pas vraiment la meilleure valeur (ni même la meilleure forme) de cette suite (c_k). On a essayé d'en comparer quelques-unes par simulations et on en reporte le résultat en figure 22.1. On a pris $\theta^A = 0.45$ et $\theta^B = 0.30$ (on explique le choix de ces valeurs au paragraphe 22.1.9 : elles conduisent aux figures 22.3, qui sont plus satisfaisantes que les figures 22.4).

Les quatre suites d'exploration sont de la forme

$$(c_k)_{k\geqslant 1} = (d_1,\ d_1 + 1,\ d_2,\ d_2 + 1,\ \ldots)\ ,$$

c'est-à-dire que dès que l'on explore le premier bras, on explore aussi le second au tour suivant. On considère des d_k donnés par k^2 (en pointillés) ; par $(d_1,\, d_2\, d_3,\, d_4) = (1,\, 3,\, 5,\, 7)$ et $d_k = 8 + (k-4)^2$ pour $k \geqslant 5$ (en trait plein), qui met en œuvre une exploration préliminaire de quatre coups sur chacun des bras avant de passer à une excitation moins fréquente ; par $d_k = 2^k$ (en traits mixtes pointillés–points) ; et enfin, $d_k = k^3$ (en points). Pour chaque suite, on relance 1 000 fois la stratégie par excitation et on calcule le gain moyen à chaque pas de temps sur ces 1 000 itérations ; c'est ce dernier que l'on a tracé à la figure 22.1. Là encore, on justifie ce lissage par répétition au paragraphe 22.1.9.

On retient essentiellement de cette étude rapide que les quatre choix sont très comparables (l'axe des ordonnées a une graduation resserrée). A cause de la condition supplémentaire sur (c_k) introduite au paragraphe 22.1.7 et parce que l'on étudiera d'autres stratégies par exploration préalable ci-dessous, on fixe pour les simulations suivantes le choix $d_k = 2^k$ et le choix résultant $(c_k) = (2,\, 3,\, 4,\, 5,\, 8,\, 9,\, \ldots)$. On note qu'ici $c_k \sim 2^{k/2}$.

22.1.7 Intervalles de confiance sur les paramètres

On veut établir la formule suivante de normalité asymptotique pour (G_n), afin d'en déduire des intervalles de confiance sur le paiement moyen maximal $\max\{\theta^A,\, \theta^B\}$.

Théorème 22.6 *Lorsque les instants d'exploration sont tels que $c_k \gg k^2$, la suite (G_n) des gains moyens de la stratégie de décision selon les succès empiriques avec excitation est asymptotiquement normale,*

$$\sqrt{n}\big(G_n - \theta^*\big) \rightsquigarrow \mathcal{N}\big(0,\, \theta^*(1-\theta^*)\big) \qquad \text{où} \quad \theta^* = \max\{\theta^A,\, \theta^B\}\,.$$

Ce résultat de normalité asymptotique est bien sûr à comparer au résultat de consistance procuré par le Théorème 22.5. La condition $c_k \gg k^2$ est équivalente à ce que $c^{-1}(n) \ll \sqrt{n}$, et est assurée par exemple lorsque $c_k \sim k^{2+\alpha}$ pour $\alpha > 0$.

Preuve Ici, comme au paragraphe précédent, le résultat est clair lorsque $\theta^A = \theta^B$ (par théorème de la limite centrale ordinaire, pour des observations indépendantes et identiquement distribuées) ; et par symétrie, il suffit de le prouver dans le cas où $\theta^A > \theta^B$. Le raisonnement est le suivant : à partir d'un certain rang, B n'est plus joué que du fait de l'exploration, dont on peut faire en sorte qu'elle soit suffisamment peu fréquente. G_n est ainsi à peu de choses près une moyenne de paiements du bras A, et une telle moyenne est asymptotiquement normale.

Plus précisément, le théorème de la limite centrale assure que

$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} E_t^A - \theta^A\right) \rightsquigarrow \mathcal{N}\Big(0,\, \theta^A\big(1-\theta^A\big)\Big)\,;$$

or,

$$\sqrt{n}\left|G_n - \frac{1}{n}\sum_{t=1}^{n} E_t^A\right| \leqslant \sqrt{n}\,\frac{N_n^B}{n} \leqslant \frac{T_0 + c^{-1}(n)}{\sqrt{n}} \;\longrightarrow\; 0 \qquad \text{p.s.},$$

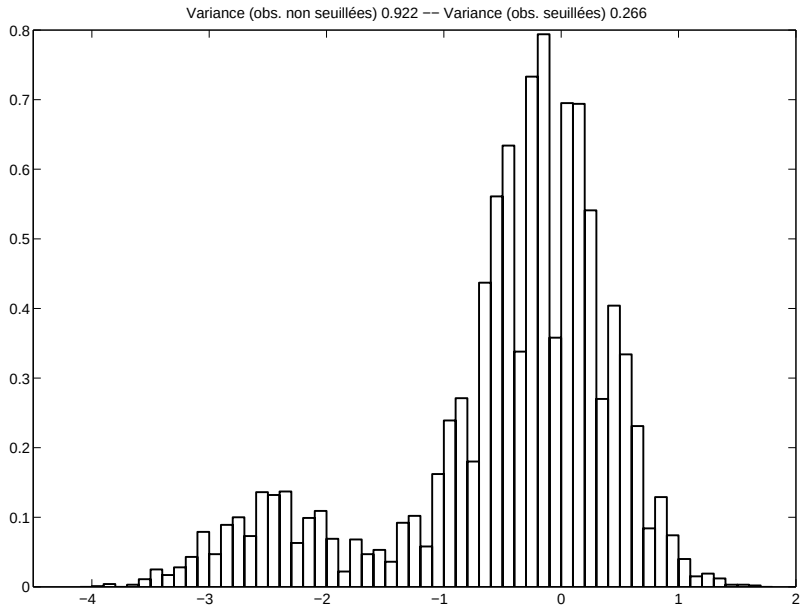

FIGURE 22.2. Histogramme des valeurs prises par $\sqrt{n}\left(G_n - \theta^A\right)$ sur 10 000 simulations pour $n = 300$, $\theta^A = 0.45$ et $\theta^B = 0.30$.

où l'on a majoré N_n^B en utilisant les résultats du paragraphe précédent, à savoir que B n'est plus joué après le rang T_0 (fini p.s.) que par exploration, et où la convergence vers 0 procède du choix, par hypothèse, d'une suite d'instants d'exploration telle que $c^{-1}(n) \ll \sqrt{n}$. Le lemme de Slutzky assure alors que

$$\sqrt{n}\big(G_n - \theta^A\big) \rightsquigarrow \mathcal{N}\Big(0,\ \theta^A\big(1 - \theta^A\big)\Big)\,,$$

ce qui conclut la preuve du cas où $\theta^A > \theta^B$. □

On illustre cette convergence par l'histogramme de la figure 22.2. On remarque que l'approximation de la distribution de la statistique d'intérêt, $\sqrt{n}\left(G_n - \theta^A\right)$, n'est pas encore tout à fait gaussienne. La distribution semble plutôt bimodale, avec deux modes gaussiens ; le mode dominant est centré en 0, et la variance des observations correspondantes (celles plus grandes que -1.5) est de 0.266, raisonnablement proche donc de la valeur théorique $\theta^A(1-\theta^A) = 0.2475$. L'autre mode concerne les chemins tels que le gain moyen G_n est encore loin de θ^A (qui est le meilleur paramètre). La masse de ce mode tend vers 0, mais lentement,

visiblement, puisqu'au rang $n = 300$, la simulation montre qu'un peu moins de 15 % de la masse de probabilité y est encore située.

On ne peut déduire directement du Théorème 22.6 un intervalle de confiance sur $\theta^* = \max\{\theta^A, \theta^B\}$, car la variance asymptotique $\sigma^2 = \theta^*(1-\theta^*)$ dépend du paramètre θ^* à estimer. On est exactement dans la situation du paragraphe 3.4. Une première méthode reposerait sur la majoration de σ^2 par $1/4$. Cependant, comme on dispose d'un estimateur (fortement) consistant de σ^2, en la personne de $G_n(1-G_n)$, il est plus efficace de procéder par réinjection (« plug-in »), comme au paragraphe 3.4.2 : une seconde application du lemme de Slutzky livre

$$\sqrt{\frac{n}{G_n(1-G_n)}}\left(G_n - \theta^*\right) \rightsquigarrow \mathcal{N}(0,1)\,,$$

d'où l'on tire l'un intervalle de confiance asymptotique de niveau $1-\alpha$ sur θ^*

$$I_{n,1-\alpha} = \left[G_n - z_{1-\alpha/2}\sqrt{\frac{G_n(1-G_n)}{n}},\ G_n + z_{1-\alpha/2}\sqrt{\frac{G_n(1-G_n)}{n}}\right],$$

où $z_{1-\alpha/2}$ désigne, ici encore, le quantile d'ordre $1-\alpha/2$ de la loi normale standard.

Arbitrage entre qualité d'estimation et obtention d'un bon paiement moyen

Pour avoir un intervalle de confiance de taille proportionnelle à $1/\sqrt{n}$ sur le paramètre $\theta_* = \min\left\{\theta^A, \theta^B\right\}$, il s'agirait que le mauvais bras soit actionné linéairement souvent ; mais cela empêcherait le gain moyen de tendre p.s. vers $\theta^* = \max\left\{\theta^A, \theta^B\right\}$. Il y a, là encore, une alternative dans la problématique : soit avoir un bon paiement moyen G_n et estimer efficacement $\max\left\{\theta^A, \theta^B\right\}$ mais pas $\min\left\{\theta^A, \theta^B\right\}$, soit arriver à estimer efficacement et simultanément θ^A et θ^B. C'est une question d'objectifs, tous ne sont pas conciliables.

Dans notre cadre, où l'on s'intéresse surtout au paiement moyen, les intervalles de confiance sur θ_* seront donc plus larges que ceux sur θ^*. Les théorèmes de la limite centrale existent pour les martingales, mais on voudrait éviter de recourir à eux ; aussi va-t-on se contenter de construire l'intervalle de confiance uniquement sur les instants d'exploration. Pour déterminer s'il faut utiliser l'exploration de A ou celle de B, on se fonde sur le classement courant entre $\widehat{\theta}_n^A$ et $\widehat{\theta}_n^B$: asymptotiquement, ils sont classés dans l'ordre correct parce que ce sont des estimateurs consistants,

$$\mathbb{I}_{\left\{\hat{\theta}_n^A > \hat{\theta}_n^B\right\}} \longrightarrow \mathbb{I}_{\{\theta^A > \theta^B\}} \qquad \text{p.s.}$$

On note

$$c_A^{-1}(n) = \max\{k \in \mathbb{N} :\ c_{2k+1} \leqslant n\} \qquad \text{et} \qquad c_B^{-1}(n) = \max\left\{k \in \mathbb{N}^* :\ c_{2k} \leqslant n\right\}$$

le nombre d'instants d'exploration pour A et B respectivement, jusqu'au tour n ; évidemment, $c^{-1}(n) = c_A^{-1}(n) + c_B^{-1}(n)$. Par une application de la loi forte des grands nombres ordinaire

(les instants d'exploration étant choisis de manière déterministe), on a l'estimation consistante

$$g_n = \left(\frac{1}{c_A^{-1}(n)} \sum_{t : c_{2t} \leqslant n} X_t \right) \mathbb{I}_{\left\{\hat{\theta}_n^A \leqslant \hat{\theta}_n^B\right\}} + \left(\frac{1}{c_B^{-1}(n)} \sum_{t : c_{2t+1} \leqslant n} X_t \right) \mathbb{I}_{\left\{\hat{\theta}_n^A > \hat{\theta}_n^B\right\}} \longrightarrow \theta_* \quad \text{p.s.}$$

De même, par application du théorème de la limite centrale ordinaire sur les temps d'exploration et du lemme de Slutzky,

$$\sqrt{c_A^{-1}(n)} \left(\frac{1}{c_A^{-1}(n)} \sum_{t\,:\,c_{2t+1} \leqslant n} X_t - \theta_* \right) \mathbb{I}_{\left\{\hat{\theta}_n^A \leqslant \hat{\theta}_n^B\right\}} + \sqrt{c_B^{-1}(n)} \left(\frac{1}{c_B^{-1}(n)} \sum_{t\,:\,c_{2t} \leqslant n} X_t - \theta_* \right) \mathbb{I}_{\left\{\hat{\theta}_n^A > \hat{\theta}_n^B\right\}} \rightsquigarrow \mathcal{N}\big(0,\ \theta_*(1-\theta_*)\big) .$$

Par la même méthode de réinjection que précédemment, en utilisant la consistance de (g_n) pour l'estimation de θ_*, on en tire un intervalle de confiance asymptotique sur θ_* de niveau $1-\alpha$:

$$\left[g_n - z_{1-\alpha/2} \sqrt{\frac{g_n(1-g_n)}{\min\left\{c_A^{-1}(n),\ c_B^{-1}(n)\right\}}},\ g_n + z_{1-\alpha/2} \sqrt{\frac{g_n(1-g_n)}{\min\left\{c_A^{-1}(n),\ c_B^{-1}(n)\right\}}} \right] .$$

Il est maintenant temps de comparer la largeur de cet intervalle de confiance à celle de $I_{n,1-\alpha}$, qui est proportionnelle à $1/\sqrt{n}$. Comme $\left| c_A^{-1}(n) - c_B^{-1}(n) \right| \leqslant 1$, on a

$$\frac{c^{-1}(n) - 1}{2} \leqslant \min\left\{c_A^{-1}(n),\ c_B^{-1}(n)\right\} \leqslant \frac{c^{-1}(n)}{2} ,$$

et la largeur est ici de l'ordre de $1/\sqrt{c^{-1}(n)}$ au lieu de $1/\sqrt{n}$. Dans les cas considérés plus haut, le choix de $d_k \sim k^{2+\alpha}$ mène à une vitesse de l'ordre de $n^{-1/(4+2\alpha)}$, et $d_k \sim 2^k$ à $1/\sqrt{\log n}$.

22.1.8 A propos des critères d'évaluation : uniformité, critère en espérance, critère d'$(1-\varepsilon)$–optimalité

Il n'existe pas de notion universelle de bonne stratégie ; le choix du critère pour déterminer la qualité d'une stratégie est crucial.

Intérêt d'un critère uniforme

Le texte introduit un critère de convergence presque sûre vers la meilleure moyenne possible : est convenable toute stratégie telle que, pour tous θ^A et θ^B dans $]0,1[$, le gain moyen G_n converge p.s. vers $\max\{\theta^A,\ \theta^B\}$, ce que l'on peut reformuler de manière compacte par

$$\inf_{\theta^A,\,\theta^B} \liminf_{n\to\infty} \left\{ G_n - \max\left\{\theta^A,\ \theta^B\right\} \right\} \geqslant 0 .$$

Voici une petite histoire, un paradoxe, pour illustrer, sur un cas simple (sur un jeu semblable aux jeux de bandits en un coup), la nécessité de raisonner pour tous les paramètres possibles θ^A et θ^B.

Remarque au passage. Un statisticien est l'heureux destinataire d'un lot tiré X uniformément au hasard dans $\{m, M\}$, où $m < M$ sont deux réels strictement positifs qui lui sont inconnus et qu'il n'a aucun moyen d'estimer. Au moment de recevoir X, on lui demande s'il veut échanger X contre l'autre valeur X' (*id est*, $X' = m$ si $X = M$ et $X' = M$ si $X = m$). Evidemment, la valeur X' ne lui est pas révélée et comme m et M lui sont inconnus, on ne voit pas très bien ce qu'il peut faire. Toujours est-il qu'on note Y son choix, qui dépend de X et éventuellement d'une randomisation auxiliaire indépendante de X et X'. Quelle est une bonne façon de faire ? Il est clair qu'en l'absence de toute information comme c'est le cas ici, on s'attendrait à ce que $\mathbb{E}[Y]$ soit égal à $\mathbb{E}[X] = (M + m)/2$ quelle que soit la stratégie Y. Cependant, on suppose disposer ici d'un aléa V de loi $\mathcal{E}(1)$, la loi exponentielle de paramètre 1 (ou, comme le montrera la preuve, toute loi à densité strictement positive sur $\mathbb{R}^+$). On conserve alors X si et seulement si $X \geqslant V$, et l'on obtient donc finalement

$$Y = X\,\mathbb{I}_{\{X \geqslant V\}} + X'\,\mathbb{I}_{\{X < V\}}\,.$$

Ce gain est tel que $\mathbb{P}\{Y = M\} > \mathbb{P}\{X = M\} = 1/2$; en effet, par disjonction d'événements puis indépendance de X et V,

$$\begin{aligned}
\mathbb{P}\{Y = M\} &= \mathbb{P}\{X = M \text{ et } V \leqslant M\} + \mathbb{P}\{X = m \text{ et } V > m\} \\
&= \mathbb{P}\{X = M\}\,\mathbb{P}\{V \leqslant M\} + \mathbb{P}\{X = m\}\,\mathbb{P}\{V > m\} \\
&= \frac{1}{2}\big(\mathbb{P}\{V \leqslant M\} + \mathbb{P}\{V > m\}\big) \\
&= \frac{1}{2}\big(1 + \mathbb{P}\{m < V \leqslant M\}\big) > \frac{1}{2} = \mathbb{P}\{X = M\}\,.
\end{aligned}$$

Notons que cette probabilité peut être arbitrairement proche de 1, selon la masse que la loi de V accorde à $]m, M]$; et par conséquent, l'espérance de gain $\mathbb{E}[Y]$, égale à $m\,\mathbb{P}\{Y = m\} + M\,\mathbb{P}\{Y = M\}$, peut être arbitrairement proche de M. Il semble donc qu'on gagne toujours à se fixer un seuil (aléatoire) et à changer de lot selon ce seuil ! Remarquons que c'est vrai même pour un seuil déterministe v : lorsque m et M sont soit tous deux plus petits, soit tous deux plus grands que v, alors l'espérance de gain de Y est celle de X ; et dans le cas où $m < v \leqslant M$, cette espérance est M. Cela nous met sur la voie de la résolution du paradoxe. C'est un paradoxe de super-efficacité, à rapprocher de l'exemple fameux de Hodge [vdV98, exemple 8.1]. S'il est vrai que $\mathbb{E}[Y] - (M + m)/2 > 0$ pour tous m et M avec le seuil aléatoire, il n'en demeure pas moins que dans un sens minimax, on ne peut pas mieux faire que la stratégie se contentant de X,

$$\sup_Y \inf_{0<m<M} \left\{\mathbb{E}\big[Y\big] - \frac{M+m}{2}\right\} = 0$$

où le supremum est sur toutes les stratégies Y possibles, éventuellement randomisées, et l'infimum, sur tous les choix possibles des couples m et M. On voit ici que requérir des comportements uniformes en m et M permet de ne pas considérer des améliorations trop locales, qui ne valent que par chance (elles dépendent de m et M, que l'on ne contrôle pas).

Critère en espérance

On définit donc le critère de qualité en espérance en prenant également en compte une certaine uniformité : est bonne, en espérance, toute stratégie telle que

$$\inf_{\theta^A,\,\theta^B} \liminf_{n\to\infty} \left\{ \mathbb{E}\big[G_n\big] - \max\left\{\theta^A,\,\theta^B\right\} \right\} \geqslant 0\,.$$

Ce critère est équivalent à la convergence en probabilité de (G_n) vers $\max\left\{\theta^A,\,\theta^B\right\}$, pour tout θ^A et θ^B.

> Remarque au passage. En effet, on procède de même que pour les équivalences discutées à la Définition–Théorème 20.2. Au paragraphe 22.1.4, on a exhibé une certaine martingale, d'espérance nulle : l'espérance du gain de toute stratégie vérifie
>
> $$\mathbb{E}\big[G_n\big] = \mathbb{E}\left[\frac{N_n^A\,\theta^A + N_n^B\,\theta^B}{n}\right].$$
>
> On a ainsi $\mathbb{E}\big[G_n\big] \to \max\left\{\theta^A,\,\theta^B\right\}$ si et seulement si
>
> $$\mathbb{E}\left[\max\left\{\theta^A,\,\theta^B\right\} - \left(\frac{N_n^A\,\theta^A + N_n^B\,\theta^B}{n}\right)\right] \longrightarrow 0\,,$$
>
> si et seulement si (s'agissant d'une variable aléatoire positive)
>
> $$\max\left\{\theta^A,\,\theta^B\right\} - \left(\frac{N_n^A\,\theta^A + N_n^B\,\theta^B}{n}\right) \xrightarrow{\mathbb{P}} 0\,,$$
>
> soit, en faisant appel à (22.2), si et seulement si
>
> $$G_n \xrightarrow{\mathbb{P}} \max\left\{\theta^A,\,\theta^B\right\}.$$

Convergences en probabilité et presque sûre étant très liées (par des arguments de sous-suites), ce nouveau critère en espérance est ainsi très loin d'être réellement plus facile à satisfaire que celui du texte. Par exemple, les stratégies naïves des paragraphes 22.1.5 et 22.1.2 ne conviennent pas non plus dans ce cadre.

Critère d'$(1-\varepsilon)$–optimalité

Une modification nous permettant en revanche de considérer et justifier la stratégie avec exploration préalable sur $2N$ tours décrite à la fin du paragraphe 22.1.5 est de fixer un seuil d'erreur $\varepsilon > 0$ et de requérir que pour tous θ^A et θ^B,

$$\liminf_{n\to\infty} \mathbb{E}\big[G_n\big] \geqslant (1-\varepsilon)\max\left\{\theta^A,\,\theta^B\right\} + \varepsilon \min\left\{\theta^A,\,\theta^B\right\}.$$

Une bonne stratégie pour ε fixé (choisi par le statisticien) repose alors par exemple sur l'identification en un temps fini (éventuellement aléatoire) et avec probabilité au moins $1-\varepsilon$ du bon bras.

On étudie dans un premier temps les capacités d'identification du bon bras par la stratégie qui explore uniformément les bras et utilise, pour $t \geqslant 1$, les choix $C_{2t-1} = A$ et $C_{2t} = B$. Pour tout $n \geqslant 1$, les estimateurs $\widehat{\theta}^A_{2n}$ et $\widehat{\theta}^B_{2n}$ sont alors des moyennes empiriques d'un nombre déterministe, n, de variables aléatoires indépendantes et identiquement distribuées. C'est le cas également de la différence $\widehat{\theta}^A_{2n} - \widehat{\theta}^B_{2n}$. On va construire à partir d'elles des intervalles de confiance $\widehat{\Delta}_n$ (de niveaux suffisamment grands) sur la différence entre les paramètres, $\Delta = \theta^A - \theta^B$.

Si pour un certain rang (aléatoire) T, la valeur 0 n'appartient plus à un de ces intervalles, alors pour tout $t \geqslant 2T + 1$, on jouera $C_t = A$ si $\widehat{\theta}^A_{2T} > \widehat{\theta}^B_{2T}$ et $C_t = B$ si $\widehat{\theta}^A_{2T} < \widehat{\theta}^B_{2T}$. On appellera cette stratégie la stratégie par exploration préalable.

Il reste à expliquer comment fixer T en fonction du paramètre ε que l'on se fixe. Il faut le prendre suffisamment grand pour qu'avec probabilité $1 - \varepsilon$, les estimateurs $\widehat{\theta}^A_T$ et $\widehat{\theta}^B_T$ soient classés dans le même ordre que θ^A et θ^B. Or, par inégalité de Hoeffding (pour variables aléatoires indépendantes et identiquement distribuées, Lemme 3.7), appliquée successivement aux Y_t et aux Y'_t, où pour $t \geqslant 1$,

$$Y_t = X_{2t-1} - X_{2t} - \theta^A + \theta^B \qquad \text{et} \qquad Y'_t = -Y_t \, ,$$

et une union de deux événements, on a que pour tout n, avec probabilité au moins $1 - \varepsilon'$ (avec $\varepsilon' > 0$),

$$\Delta = \theta^A - \theta^B \in \widehat{\Delta}_n \overset{\text{not.}}{=} \left[\widehat{\theta}^A_n - \widehat{\theta}^B_n - 2\sqrt{\frac{1}{2n} \log \frac{2}{\varepsilon'}},\ \widehat{\theta}^A_n - \widehat{\theta}^B_n + 2\sqrt{\frac{1}{2n} \log \frac{2}{\varepsilon'}} \right] .$$

Il ne suffit pas de prendre $\varepsilon' = \varepsilon$, il faut être un peu plus précautionneux ; cela vient essentiellement du fait que les $\widehat{\Delta}_n$ et T ne sont pas indépendants.

On choisit, pour tout $n \geqslant 1$, une probabilité $\varepsilon' = 6\varepsilon/(\pi^2 n^2)$: pour tout $n \geqslant 1$, avec probabilité au moins $1 - 6\varepsilon/(\pi^2 n^2)$,

$$\Delta = \theta^A - \theta^B \in \widehat{\Delta}_n = \left[\widehat{\theta}^A_n - \widehat{\theta}^B_n - 2\sqrt{\frac{1}{2n} \log \frac{\pi^2 n^2}{3\varepsilon}},\ \widehat{\theta}^A_n - \widehat{\theta}^B_n + 2\sqrt{\frac{1}{2n} \log \frac{\pi^2 n^2}{3\varepsilon}} \right] .$$

On définit alors T comme le temps d'arrêt construit sur ces nouveaux intervalles $\widehat{\Delta}_n$:

$$T = \inf \left\{ n \geqslant 1 : \ 0 \notin \widehat{\Delta}_n \right\} \, ;$$

T est fini presque sûrement dès que $\theta^A \neq \theta^B$, comme on peut le voir par loi des grands nombres. On a également

$$\mathbb{P}\left\{ \Delta \notin \widehat{\Delta}_T \right\} = \sum_{n \geqslant 1} \mathbb{P}\left\{ \Delta \notin \widehat{\Delta}_n \text{ et } T = n \right\} \leqslant \sum_{n \geqslant 1} \mathbb{P}\left\{ \Delta \notin \widehat{\Delta}_n \right\} \leqslant \sum_{n \geqslant 1} \frac{6\varepsilon}{\pi^2 n^2} = \varepsilon \, .$$

Par construction, $0 \notin \widehat{\Delta}_T$ (de manière déterministe), ce qui signifie que $\widehat{\Delta}_T$ est inclus soit dans $\mathbb{R}^*_+$ (auquel cas, l'algorithme jouera uniquement $C_t = A$ pour $t \geqslant 2T + 1$), soit dans

$\mathbb{R}^*_-$ (auquel cas, on joue B pour le reste des tours de jeu). On choisit le meilleur bras pour le reste du jeu dès qu'il est effectivement le cas que $\Delta \in \widehat{\Delta}_T$, ce qui arrive avec probabilité au moins $1 - \varepsilon$. Au final, la stratégie considérée obtient une espérance de gain moyen au moins égale, asymptotiquement, à

$$\liminf \mathbb{E}\big[G_n\big] \geqslant (1-\varepsilon) \max\left\{\theta^A,\, \theta^B\right\} + \varepsilon \min\left\{\theta^A,\, \theta^B\right\},$$

ce qui montre que le critère d'$(1 - \varepsilon)$–approximation est vérifié.

Remarquons que ce critère modifié, s'il mène à des stratégies de décision plus intuitives, est dangereux dans certaines situations. La phase préparatoire d'exploration peut faire jouer un grand nombre de fois le mauvais bras au début (ce qui ne pose pas de problème ici puisque ce nombre, même grand, est fini et que le critère est asymptotique). En revanche, dans les applications de la vie réelle comme, par exemple, la comparaison de deux traitements médicaux, on peut et doit, pour des raisons d'éthique et d'égalité entre les patients, utiliser autant que faire se peut et aussi rapidement que possible le bon traitement. (Au lieu de raisonner en termes de machines à sous et gains, on raisonne ici en termes de patients et de guérisons.) On n'a alors pas envie d'utiliser une stratégie en deux phases, exploration intensive puis exploitation des résultats de l'observation, mais d'exploiter immédiatement le peu d'information dont on dispose, comme on le faisait ci-dessus pour les stratégies dont le gain moyen convergeait presque sûrement vers celui du meilleur bras.

Le raisonnement précédent repose sur l'identification, avec grande probabilité, du meilleur bras. On pourrait penser que la théorie des tests serait utile ; il n'en est rien, car ceux-ci ne disent pas vraiment comment choisir entre deux hypothèses possibles, mais mettent seulement en exergue les situations où les observations contredisent gravement l'hypothèse testée (l'hypothèse nulle), auquel cas on passe alors à l'autre hypothèse, dite alternative. Ici, les deux hypothèses sont $\Delta > 0$ et $\Delta < 0$, et les tests considérés seraient des tests de conformité sur la moyenne, avec pour statistiques les $\widehat{\theta}^A_{2n} - \widehat{\theta}^B_{2n}$. Il se pourrait que des tests sur la moyenne choisissant chacun une des hypothèses comme hypothèse nulle la conservent tous deux (comme indiqué à la remarque 4.8). Evidemment, lorsque le nombre d'observations augmente, ce n'est plus le cas pour de bons tests (puissants) ; mais alors, cela revient à voir quand les intervalles de confiance de θ^A et θ^B sont disjoints, ce qui est exactement ce que l'on a fait (de manière non-asymptotique) ci-dessus, avec l'inégalité de Hoeffding. En résumé, le cadre des tests d'hypothèses n'apporte rien ici, à cause de la dissymétrie de traitement entre les hypothèses : pour les mettre en œuvre, il faudrait avoir une raison subjective de suspecter que l'un des deux paramètres θ^A et θ^B est plus grand que l'autre, ce qui n'est pas le cas face à deux machines à sous d'apparence identique (mais qui pourrait l'être dans l'exemple des traitements médicaux).

22.1.9 Performances pratiques de cinq stratégies et demie

On compare ici les performances des stratégies suivantes,

– la stratégie naïve N d'exploitation pure du paragraphe 22.1.5 ;

– cette même stratégie, précédée de 20 coups d'exploration pure, 10 pour chaque bras, non alternés : c'est la demi-stratégie, que l'on appelle EN ;
– la stratégie avec excitation X (utilisant, pour $k \geqslant 1$, les temps d'exploration $c_{2k-1} = 2^k$ et $c_{2k} = 2^k + 1$ comme indiqué au paragraphe 22.1.6) ;
– la stratégie E d'exploration préalable du paragraphe 22.1.8, pour $\varepsilon = 0.10$;
– la stratégie U dite UCB (« upper confidence bound ») dans la littérature de l'apprentissage ;
– une stratégie S du type « problème du scrutin ».

Il nous faut encore simplement préciser les définitions des stratégies U et S (et les heuristiques associées). La stratégie U joue chacun des deux bras aux deux premiers tours, puis choisit, pour $t \geqslant 3$, le bras $C \in \{A, B\}$ maximisant (avec les notations précédentes)

$$\widehat{\theta}^C_{t-1} + \sqrt{\frac{2\log(t-1)}{N^C_{t-1}}}$$

(et un bras au hasard en cas d'égalité). Les quantités en jeu peuvent être interprétées comme les bornes supérieures de certains intervalles de confiance sur les paramètres θ^A et θ^B. Cet algorithme est conçu pour minimiser une quantité appelée regret cumulé et on admet ici qu'il vérifie en particulier le critère de convergence en espérance du paragraphe 22.1.8.

La définition de la stratégie S ressemble quant à elle à celle de la stratégie naïve, excepté que lorsqu'un bras obtient un gain nul, on le pénalise pour cela, en retirant 1 à ses gains cumulés. Plus précisément, C_1 est tiré uniformément sur $\{A,\ B\}$, et pour $t \geqslant 2$,

$$C_t = \begin{cases} A & \text{si } \widehat{r}^A_{t-1} > \widehat{r}^B_{t-1}, \\ B & \text{si } \widehat{r}^A_{t-1} < \widehat{r}^B_{t-1}, \\ C'_t & \text{sinon, où } C'_t \text{ est tiré uniformément sur } \{A,\ B\}, \end{cases}$$

où $\widehat{r}^A_{t-1}$ et $\widehat{r}^B_{t-1}$ sont les estimateurs suivants de $r^A = 2\theta^A - 1$ et $r^B = 2\theta^B - 1$:

$$\widehat{r}^A_{t-1} = \frac{1}{N^A_{t-1}} \sum_{s=1}^{t-1} (2X_s - 1)\mathbb{I}_{\{C_s = A\}} \qquad \text{lorsque} \quad N^A_{t-1} \geqslant 1, \quad \text{et } \widehat{r}^A_{t-1} = 0 \text{ sinon}$$

(et de même pour B). On transforme ici les lois de Bernoulli (à support $\{0, 1\}$) en lois de Rademacher (à support $\{-1, 1\}$). C'est d'ailleurs comme cela que l'on peut implémenter en pratique l'algorithme S à partir de N : en le lançant sur les paiements transformés $2X_t - 1$. L'algorithme S évite certains écueils de N : il ne peut tomber dans certains pièges, par exemple celui où le mauvais bras est joué au premier tour, obtient un paiement de 1, et est joué à jamais ; avec une probabilité raisonnable, les pénalités -1 vont s'accumuler suffisamment ici pour que le bon bras soit de nouveau joué. Cela dit, il ne semble pas garanti que tous les pièges soient évités, leur probabilité est simplement moindre. On l'appelle algorithme du scrutin car sa formulation rappelle le problème du scrutin (voir par exemple [FF98, paragraphe 4.5] pour une description de ce dernier).

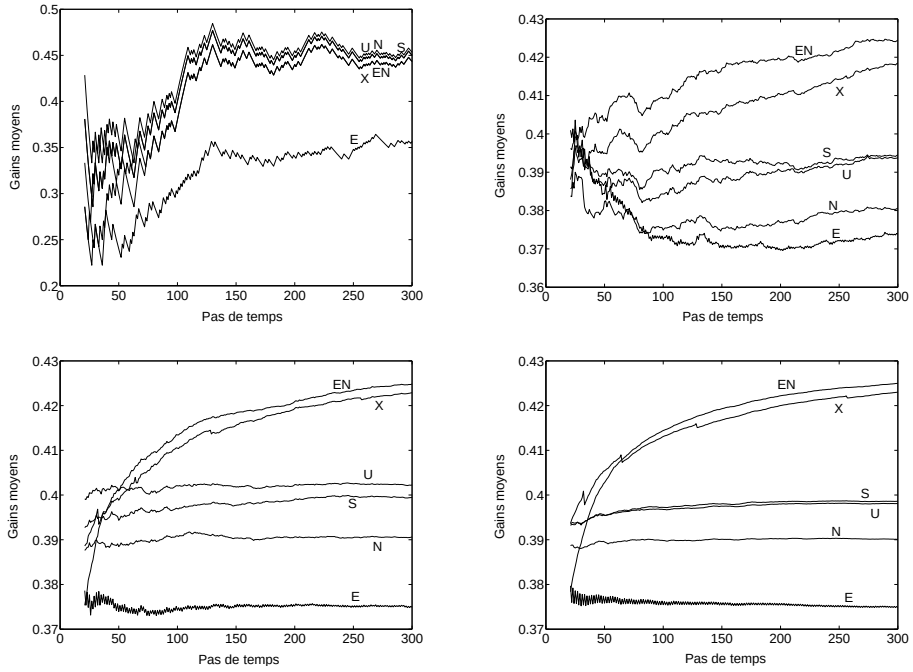

FIGURE 22.3. Performances moyennes des six algorithmes, en fonction du temps ; paramètres $\theta^A = 0.45$ et $\theta^B = 0.30$. En haut : sur 1 simulation (à gauche), 100 simulations (à droite) ; en bas : sur 1 000 simulations (à gauche), 10 000 simulations (à droite).

En résumé, on dispose de stratégies dont on sait ou pense qu'elles ont une probabilité strictement positive d'échouer (que leur gain moyen ne converge pas vers l'espérance du meilleur bras, ni p.s. ni même en espérance) : N, EN et S. Une assure une convergence en espérance, il s'agit de U ; et E assure une telle convergence à un facteur $1 - \varepsilon$ près. Quant à X, c'est celle qui dispose de la meilleure garantie théorique, son gain moyen convergeant presque sûrement vers l'espérance du meilleur bras. Il s'agit maintenant de voir si les performances pratiques sont liées ou non à ces garanties théoriques différentes. Bien que toutes ces stratégies s'étendent plus ou moins naturellement au cas de plusieurs bras, nous ne donnons de résultats expérimentaux que pour le cas de deux bras.

La figure 22.3 montre les performances moyennes des six algorithmes pour le couple de paramètres $\left(\theta^A,\ \theta^B\right) = (0.45,\ 0.30)$ et pour les tours de jeu 21 à 300. Pour chacun des graphiques de performances, on a tiré m réalisations des suites de paiements, pour des tailles $m \in \{1,\ 100,\ 1\,000,\ 10\,000\}$, lancé les algorithmes dessus, et représenté la moyenne (sur les m simulations) des moyennes mobiles des gains G_n ; c'est-à-dire, que l'on a réalisé une appro-

ximation des espérances $\mathbb{E}\big[G_n\big]$ selon la méthode de Monte-Carlo (voir le paragraphe 9.7.1). Evidemment, la qualité de l'approximation est d'autant meilleure que m est grand. Ce n'est pas surprenant : s'agissant de quantités toujours comprises dans [0, 1], il découle des résultats du chapitre 3 que la précision des moyennes par simulations par rapport aux vraies valeurs est de l'ordre de $1/\sqrt{m}$. Un comportement typique apparaît sur les graphiques de $m = 1\,000$ et $m = 10\,000$; en revanche, tel n'est pas le cas [1] pour les faibles valeurs (1 et 100) de m, l'allure des courbes y est fort différente. Pour les autres couples de paramètres, à cause du caractère typique qu'il révèle et vu le temps de calcul dix fois plus court, on effectuera nos simulations sur un nombre de $m = 1\,000$ simulations (voir figure 22.4).

Pour l'heure, on remarque sur la figure 22.3, une fois passé le comportement erratique du début, que les deux meilleures stratégies sont EN (pour laquelle on n'a pourtant aucune garantie !) et X (la meilleure théorique). Pour X, on lit facilement les temps d'exploration sur le graphique des performances, ces dernières marquant un creux. EN part d'une moyenne mobile plus basse que la plupart des algorithmes (à cause de la phase d'exploration pure), mais remonte le cap très rapidement ; il semble à privilégier lorsque l'on sait par avance que le nombre de tours de jeu sera grand (sans quoi, c'est X qui semble le plus adapté). Vient ensuite un groupe de trois formé par S, U et N. Bien que l'espérance des gains de U doive converger vers $\theta^A = 0.45$, cette convergence ne semble pas très rapide.

Enfin, E clôt la marche. Ses mauvaises performances sont expliquées par le fait que même sur les 10 000 simulations, aucune n'a permis à l'algorithme E de passer en mode d'exploitation (voir tableau 22.1) ; il a donc exploré tout le long et obtenu, comme il se doit, un paiement moyen dans cette phase proche de la moyenne des paramètres, ici, 0.375. On pouvait s'attendre à cela ; heuristiquement, le temps T à partir duquel l'exploration prend fin est proche du premier rang n tel que

$$2\sqrt{\frac{1}{2n}\log\frac{\pi^2 n^2}{3\varepsilon}} < \left|\theta^A - \theta^B\right| ,$$

qui ici, même pour ε grand (pris égal à 0.10 dans les simulations), requiert $n \geqslant 1\,600$, alors qu'ici l'étude se fait pour $n \leqslant 300$.

Pour comparer plus finement ces algorithmes, on a également calculé pour chaque algorithme (une approximation empirique sur $m = 10\,000$ simulations de) la probabilité que d'une part, il ait détecté que A était le bon bras, et que d'autre part, il soit celui des six qui obtienne le meilleur gain moyen. Les résultats sont reportés dans le tableau 22.1. La première ligne donne une estimation de la probabilité de tirer le bon bras au tour 300 ; c'est la mesure que l'on a choisi pour quantifier la bonne détection du meilleur bras. (Pour la stratégie E, on a en fait reporté la probabilité que la phase d'exploration soit finie : elle est nulle. Par conséquent, E alterne encore entre A et B, jouant A aux tours impairs.) Les résultats sont donnés avec deux chiffres significatifs seulement, car la précision de l'estimation est de l'ordre de 0.01 ($= 1/\sqrt{m}$), comme le montre l'exemple 3.9.

[1] Cela montre, en passant, l'intérêt de ne pas simuler uniquement des trajectoires individuelles, mais d'atteindre une représentation de comportements plus généraux. C'est une remarque valable pour l'ensemble des textes.

	N	EN	X	E	S	U
$\mathbb{P}\{C_{300} = A\}$	0.60	0.88	0.87	0.00	0.65	0.66
Prb. meilleur	0.38	0.23	0.21	0.02	0.33	0.34

TABLE 22.1. Différentes mesures numériques de performance des six stratégies, pour $\theta^A = 0.45$, $\theta^B = 0.30$ et $n = 300$. Estimées obtenues à partir de $m = 10\,000$ simulations.

La seconde ligne donne une estimation de la probabilité qu'un algorithme donné soit le meilleur (ou un des meilleurs, en cas d'égalités : à cause de cela, les probabilités se somment à 1.51, les cas d'égalité sont fréquents, notamment entre S et U). De tout cela, on retient essentiellement que la première ligne explique pourquoi X et EN ont une meilleure espérance de gain que les autres stratégies : c'est qu'ils ont détecté plus massivement le bon bras. Cette ligne permet aussi de voir avec quelle fréquence N tombe dans un piège ; pas si souvent que cela, mais suffisamment pour affecter ses performances. En faisant varier les instants où l'on évalue la probabilité de jouer le bon bras, on pourrait évaluer la vitesse de convergence (en probabilité) du gain moyen vers le gain moyen du meilleur bras. Il est cependant intéressant de noter que sur la seconde ligne, la meilleure performance est obtenue par N, celui-là même qui a la moins bonne espérance de gain moyen. C'est sans doute un artefact de sa tendance naturelle à tomber dans les pièges, en sa faveur ici (il peut être piégé par le bon bras), mais cela souligne l'importance du critère d'évaluation. Pour ce (mauvais, parce que relatif) critère d'évaluation, N serait le meilleur algorithme !

Pour être tout à fait complet, on propose pour conclure des simulations pour d'autres couples de paramètres, à la figure 22.4 (tous inférieurs ici encore à 1/2, sans quoi le casino serait perdant). Le comportement typique semble préservé pour tous les algorithmes (excepté E), que θ^B soit proche de θ^A (à droite) ou pas (à gauche) ; on peut toujours classer les algorithmes en les trois mêmes groupes de performance, X et EN, puis S, U et N, et enfin, E. Ce dernier semble sur le point de rattraper son retard sur le graphique de gauche, qui correspond à un écart important entre θ^A et θ^B, que E détecte cette fois-ci plus rapidement et même efficacement : à $n = 300$, il a une probabilité d'environ 98 % d'avoir détecté le bon bras, et n'est supplanté sur ce critère que par les algorithmes X et EN.

Cette étude empirique montre clairement que les meilleures stratégies sont X et EN (lorsque le nombre de tours n est petit), que les gains de E convergent très lentement vers le gain moyen du meilleur bras, et que N est effectivement mauvaise au sens des critères les plus naturels. Lorsque n est modérément grand, de l'ordre au moins de la centaine, on préfère EN à X ; lorsque n est petit, X est plus sûr. Les comportements sont sans doute différents pour n très grand, de l'ordre de plusieurs milliers, puisqu'alors le côté asymptotique des critères ressort plus nettement, en bien ou mal selon qu'ils sont vérifiés ou non par l'algorithme considéré.

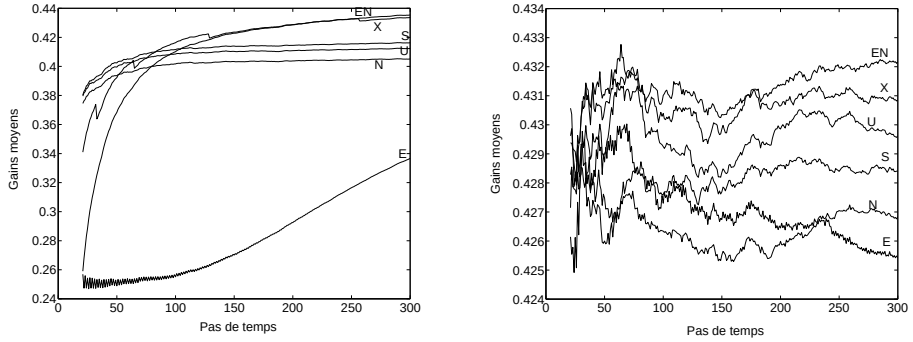

FIGURE 22.4. Performances moyennes des six algorithmes, en fonction du temps, sur 1 000 simulations. A gauche : $\theta^A = 0.40$ et $\theta^B = 0.05$; à droite : $\theta^A = 0.45$ et $\theta^B = 0.40$.

22.2 Questions possibles, conseils aux étudiants et bibliographie

22.2.1 Questions éventuelles d'un enseignant

Une question qui peut dérouter (à dessein, par sa simplicité) serait :

> Finalement, a-t-on vraiment besoin d'une loi des grands nombres pour les martingales ? Pourquoi la loi des grands nombres ordinaires ne s'applique-t-elle pas ? L'estimateur
>
> $$\widehat{\theta}_n^A = \frac{1}{N_n^A} \sum_{t:C_t=A} E_t^A \,,$$
>
> n'est-il pas une moyenne de variables aléatoires indépendantes et identiquement distribuées ?

Il s'agit de faire remarquer que la sommation est aléatoire : on ne somme pas toujours sur les mêmes éléments, les index de ceux-ci dépendent d'une manière complexe des performances du bras A, et même de celles de B. La valeur de E_t^A lorsque $C_t = A$ a un impact sur celle de C_{t+1} et des choix suivants. Dit autrement, il y a dépendance des $X_t = E_t^{C_t}$ lorsque t varie, alors même qu'il y avait indépendance des $\left(E_t^A,\ E_t^B\right)$ lorsque t varie (essentiellement, parce que l'on n'observe pas toutes les observations, mais qu'on en choisit une à chaque tour).

On pourrait également tester les connaissances sur les théorèmes de convergence de martingales (qui mettent en jeu soit des hypothèses de signe, soit des hypothèses de bornitude dans $\mathbb{L}^p$ pour $p \geqslant 1$) et/ou sur les théorèmes d'arrêt. Encore une fois, on ne peut que recommander la lecture des rappels de cours de [MPB98].

Enfin, pour susciter l'imagination et l'intuition d'un étudiant, on pourrait lui demander ce qu'il ferait si la machine à sous avait $N \geqslant 3$ bras ; la formulation de la plupart des stratégies étudiées s'étend naturellement à ce cas, mais on peut aussi proposer de mener C_N^2 batailles deux à deux entre des couples de bras.

22.2.2 Conseils aux étudiants

Ce texte présente un problème éminemment statistique (ou presque, puisque l'objectif n'est pas tant l'estimation que l'utilisation d'estimations pour atteindre un objectif de gain, mais passons) ; le traitement est probabiliste, certes, mais il ne faut pas pour autant éluder la partie de modélisation et de motivation statistique. En particulier, il est important, au début, de définir le modèle statistique, c'est-à-dire l'ensemble des lois possibles. Ici, cet ensemble est identifié aux couples de paramètres de Bernoulli possibles et est indexé par $\Theta =]0, 1[^2$. On précisera donc que le modèle est paramétrique.

Même si l'on ne se sent pas encore à l'aise en statistique, il ne faut pas avoir peur de prononcer le mot d'estimateur (c'est simplement une variable aléatoire mesurable en les observations, par définition !), et encore moins de les qualifier de (fortement) consistants le cas échéant. Ces derniers sont souvent donnés, dans les cas simples, par une certaine fonction, typiquement la moyenne, d'un échantillon ; ce n'est pas le cas ici, on a affaire à des observations X_t dépendantes, comme rappelé ci-dessus. Il est aisé de briller à la fin du texte en expliquant comment, et surtout, pourquoi, d'une simple convergence en loi (énoncé probabiliste), on tire un intervalle de confiance asymptotique (notion statistique) : on veut indiquer l'ensemble des valeurs typiques des gains moyens G_n.

Toujours dans le cadre de la présentation des résultats, il est important de ne pas escamoter ou pire, de laisser sous silence, la définition précise d'une stratégie (un processus adapté par rapport à une filtration précisée au paragraphe 22.1.1). Le jury ne saurait se contenter de définitions vagues ou heuristiques comme « une stratégie associe à tout historique une action ». Les mathématiques conceptualisent de façon précise et rigoureuse.

La preuve de la loi des grands nombres pour les martingales peut être passablement pénible ou ennuyeuse, surtout si l'on tâche de prêter attention, comme ici (en introduisant le temps d'arrêt τ), à ne pas diviser par 0 en écrivant $1/V_t$. On pourra soulever le détail à l'oral et faire mine de l'oublier aussitôt, en n'écrivant que des sommations de 1 à n ; l'auditoire aura compris que si l'on avait le temps, on pourrait tout écrire proprement comme dans ce corrigé. De même, on peut ne pas exécuter la preuve du lemme de Kronecker et mentionner simplement lors du passage que sa démonstration repose sur une transformation d'Abel. Une telle précision montre clairement que l'on connaît les ressorts de la preuve et qu'on l'a préparée ; et le public sera fort aise de ne pas subir la démonstration et son lot de calculs, surtout si c'est pour conserver du temps pour discuter la modélisation et la pertinence du modèle ou se pencher plus avant sur les simulations.

Enfin, dans le cadre de la précision des résultats des simulations, s'agissant d'approximations de Monte-Carlo, on ne donnera pas de nombres plus précis qu'au centième (les différentes probabilités ou espérances de gains moyens) ; la précision est en effet de l'ordre de $1/\sqrt{m}$, où m est la taille de l'échantillon tiré pour approximer les quantités d'intérêt. On indiquera en revanche que telle est cette précision, et on commentera que, par exemple, on ne saurait tirer des valeurs 0.88 et 0.87 (trop proches) du tableau 22.1 une quelconque conclusion quant aux performances respectives de EN et X.

22.2.3 Bibliographie

Les problèmes de bandits à plusieurs (beaucoup plus que deux !) bras sont un domaine de recherche très actif en probabilités, théorie des jeux, statistiques et apprentissage séquentiel. La littérature est pléthorique et plutôt que d'en donner une vue d'ensemble, on raconte, pour l'anecdote, la genèse de ce texte. Il est issu d'une discussion après un exposé donné à l'université Paul-Sabatier de Toulouse. Gilles Stoltz y parlait de bandits à plusieurs bras dans un cadre non stochastique ; Bernard Bercu (merci !) a alors mentionné l'ouvrage de synthèse [Duf90] de Marie Duflo, et a bien voulu lui montrer le sujet d'un examen de DEA qu'il en avait tiré. Quelques simplifications (adieu les théorèmes de la limite centrale pour les martingales !) ou extensions, reposant notamment sur la littérature de l'apprentissage séquentielle, ont enrichi la matière ; comme seule référence, qui pointe vers les classiques du domaine (d'autres articles et de nombreux livres), on citera l'article [ACBF02], introduisant la stratégie UCB, de Peter Auer, Nicolò Cesa-Bianchi et Paul Fisher.

22.3 Commentaires sur les simulations `Matlab`

22.3.1 Architecture générale des programmes

Il y a six fonctions implémentant chacune une des six stratégies : `Algo_N`, `Algo_EN`, `Algo_S`, `Algo_X`, `Algo_UCB` et `Algo_E`, respectivement pour les stratégies N, EN, S, X, U et E présentées au paragraphe 22.1.9. Deux fonctions auxiliaires, `meanV` (pour calculer des moyennes de vecteurs éventuellement vides) et `Paiements` (pour générer les paiements des bras), sont utilisées respectivement par les six fonctions de stratégies et par les trois fichiers scripts. Le code de ces derniers est placé tout à la fin des codes ; le premier script sert à la calibration des instants d'exploration (il rend la figure 22.1), le second au tracé de l'histogramme de la figure 22.2, et le troisième, à l'étude empirique du paragraphe 22.1.9.

22.3.2 Commentaires sur le code `Matlab`

Les fonctions implémentant les stratégies ne sont pas forcément optimales en termes de temps d'implémentation mais leur code est compact et lisible. Une implémentation naturelle serait de maintenir des variables informatiques pour les estimateurs de chacun des deux paramètres, ainsi que du nombre de fois qu'est joué chaque bras. Puis, des boucles `if` décideraient quel bras tirer.

Ici, on a choisi de collecter dans le vecteur `C` la suite des bras tirés, ce qui nous permet d'en retrouver facilement le nombre, et, par extraction de la matrice de paiements `G`, le gain moyen. Les bras sont indexés par 1 et 2 et des lignes comme

```
mA = meanV(G(C == 1));
mB = meanV(G(C == 2));
c = 1 + ((mB > mA) + (rand < 1/2) * (mB == mA));
```

permettent de calculer rapidement l'indice `c` du bras à tirer, et ce, sans boucle `if`.

Ce n'est que pour la stratégie X que l'on a dû avoir recours à `if` ; pour éviter d'imbriquer les `if`, on a utilisé un environnement `if ... then ... elseif ... else`, dans lequel les

conditions sont testées séquentiellement et où la première condition vraie détermine l'action (à défaut de conditions vraies, l'action qui suit `else` est effectuée). En règle générale, les implémentations que l'on peut voir proposées ici par les étudiants sont beaucoup plus longues et imbriquent de nombreuses boucles `if` et `for`. De plus, il nous a paru plus naturel et plus lisible de séparer la génération des paiements et leur traitement par chaque algorithme.

On a besoin de la convention que la moyenne d'un vecteur vide `[]` est 0 ; c'est pour l'implémenter que l'on définit une fonction `meanV` (voir les codes en annexe, au chapitre 28) ; `mean` renvoie quant à elle l'erreur `NaN` (« not a number ») lorsque l'on demande la moyenne de `[]`.

Comme indiqué au paragraphe 20.3.2, la fonction `any` permet de vérifier qu'une condition est vraie pour au moins un élément d'un vecteur, on l'utilise pour tester s'il faut explorer ou exploiter dans l'algorithme X. Sa fonction jumelle est `all`, elle teste si une condition est vraie pour tous les éléments d'un vecteur.

On a souvent besoin de grouper des chaînes de caractères dans un même objet et de faire référence à la k–ième chaîne de cet objet, notamment pour donner les légendes ou titres de graphiques. Ce dernier ne peut être de type `array` (une matrice), sauf si toutes les chaînes ont même longueur. Le bon objet est de type `cell`, voir le paragraphe 21.3.2. On pourra regarder le traitement effectué avec `gtext` (cf. la fin du paragraphe 28.2.7 en annexe) pour placer les noms des algorithmes sur les différentes figures de ce chapitre. On rappelle que `gtext` permet à l'utilisateur de placer lui-même, interactivement, le texte sur une figure. (Cela distrait toujours et met un peu d'animation dans une présentation.)

22.3.3 Codes des simulations

Ils sont disponibles en annexe, au chapitre 28, page 443.

CHAPITRE 23

Étude du nombre de renouvellements

23.1 Préliminaires : questions possibles et conseils aux étudiants

23.1.1 Questions éventuelles d'un enseignant

Avant de lire le corrigé, nous ne saurions que trop conseiller aux étudiants de réfléchir aux questions suivantes, qu'aurait pu poser un enseignant. Elles sont de difficulté très variable bien entendu, mais elles prolongent naturellement les suggestions qui se trouvent à la fin du texte. Chacune de ces questions trouve sa réponse dans le corrigé qui suit.

1. Quelle est la forme de la fonction $t \mapsto N(t)$?
2. Si les durées de vie $(T_i)_{i \geqslant 1}$ suivent une loi exponentielle de paramètre λ, quelle est la loi de $N(t)$ pour $t > 0$?
3. On introduit la loi de Weibull de paramètres $a > 0$ et $\lambda > 0$, à support dans $\mathbb{R}_+$, en explicitant sa fonction de répartition : pour $t \geqslant 0$,

$$F_{a,\lambda}(t) = 1 - \exp\bigl(-(\lambda t)^a\bigr) .$$

   A quelles conditions sur a et λ cette loi est-elle NBU ?
4. Dans le texte, on suppose connus les paramètres μ et σ^2. Comment feriez-vous en pratique pour vous affranchir de cette hypothèse ?
5. Pouvez-vous donner une interprétation de la fonction B ?
6. Pouvez-vous encadrer m à l'aide de fonctions de F ?
7. Que pouvez-vous dire du processus $(S_n - n\mu)$? Déduisez-en une preuve directe de l'identité de Wald, qui n'utilise pas l'indication du texte.

23.1.2 Conseils aux étudiants

Le texte proposé traite d'un problème de fiabilité ; à première vue (en ce qui concerne les techniques employées), il se situe dans le champ des probabilités plutôt que dans celui de la statistique. Les objectifs à atteindre, à savoir, la construction d'intervalles de prévision, ainsi que des remarques que l'on peut formuler tout au long du texte, mettent en revanche en jeu des connaissances de statistique. Ce texte, comme celui du chapitre 10, est une autre illustration de la possibilité et de l'intérêt des références statistiques dans un texte de probabilité, mises particulièrement en valeur au paragraphe 23.2.9.

On manipule ici un objet très classique et que l'on retrouve par exemple dans de nombreux textes proposés par le jury de l'agrégation : un processus de renouvellement. Les résultats

asymptotiques du paragraphe 13.2 font partie du folklore et sont vus en général pendant les années de master. Les résultats des paragraphes ultérieurs sont moins classiques, mais les outils pour les démontrer le sont. On exploite notamment à de très nombreux endroits le conditionnement, et en particulier la formule

$$\mathbb{E}\big[\phi(X,Y)\big] = \mathbb{E}\Big[\mathbb{E}\big[\phi(X,Y)\,|\,Y\big]\Big] = \int \mathbb{E}\big[\phi(X,y)\,|\,Y=y\big]\, f_Y(y)\,\mathrm{d}y\,,$$

valable pour toutes variables aléatoires X et Y, cette dernière ayant pour densité f_Y, et pour toute fonction borélienne ϕ telle que $\phi(X,Y)$ est intégrable.

Le texte proposé est très long et il est hors de question d'en présenter tous les résultats lors d'un oral. Il ne comporte pas de pièges particuliers (sauf peut-être dans la démonstration de la convergence en loi du paragraphe 13.2), la modélisation est simple et les simulations relativement aisées à mettre en œuvre pour un étudiant entraîné. La difficulté réside donc dans les démonstrations des résultats. Celles-ci doivent être présentées avec rigueur et on ne doit pas se contenter d'arguments heuristiques. Nous fournissons dans le corrigé ci-dessous une démonstration minutieuse de chacun des résultats.

Comme nous l'avons souligné, la longueur du texte impose de faire des choix et c'est peut-être une autre difficulté du texte. Il est judicieux d'admettre les résultats d'analyse présentés en introduction pour se concentrer sur les résultats de probabilités. Enfin, comme on l'a souvent répété, un étudiant ne doit pas oublier de discuter la modélisation du texte (dont les aspects les plus intéressants se trouvent au paragraphe 13.4), ni de proposer des illustrations informatiques.

23.2 Corrigé possible du texte

23.2.1 Modèle statistique

On rappelle ici la modélisation et on indique les variables que l'on observe et les paramètres du problème. Des composants sont utilisés séquentiellement, et leurs durées de vie sont données par des variables aléatoires indépendantes et identiquement distribuées T_i, pour $i = 1, 2, \ldots$, que l'on observe. Leur loi commune ν est de densité f par rapport à la mesure de Lebesgue ; s'agissant de durées de vie, on a évidemment que $f = 0$ sur $\mathbb{R}_-^*$. On fait l'hypothèse dans le texte du chapitre 13 que f est strictement positive sur $\mathbb{R}_+$, et qu'elle est continue et bornée. Enfin, on suppose aussi que ν admet un moment d'ordre deux et on note μ et σ^2 ses espérance et variance.

On s'intéresse ci-dessous au nombre $N(t)$ de composants utilisés depuis l'instant de départ jusqu'à un instant t donné ; il est défini par

$$N(t) = \max\big\{n \geqslant 0 :\ S_n \leqslant t\big\}$$

où $S_0 = 0$ et $S_n = T_1 + \ldots + T_n$ pour $n \geqslant 1$. La loi de $N(t)$ ne dépend que de ν (et de t), et elle est portée par $\mathbb{N}$, i.e., ce nombre $N(t)$ est fini presque sûrement. En effet, on aurait sinon que $T_i \leqslant t$ pour tout $i \geqslant 1$, ce qui forme une événement de probabilité nulle, par indépendance des T_i et vu que leur loi charge tout $\mathbb{R}$.

Au final, le problème est décrit par la donnée de la loi ν, qui sera caractérisée dans ce qui suit par sa densité f, sa fonction de répartition F ou sa fonction de fiabilité R.

23.2.2 Problématique : construction d'un intervalle de prévision

On s'intéresse à un problème de fiabilité. En effet, afin de prévoir le nombre de pannes que peut rencontrer le système, on cherche à contrôler précisément la variable aléatoire $N(t)$, qui est le nombre de renouvellements, en déterminant un intervalle $I = \left[0, N_{\max}\right]$ qui contienne avec forte probabilité la variable $N(t)$. Ainsi, $N_{\max}$ constituera le nombre de composants de rechange à prévoir afin de pouvoir réparer le système pendant toute la durée t. Dans la suite, nous fixerons $\alpha \in]0, 1[$ proche de 0 et nous chercherons $N_{\max}$ de manière à ce que

$$\mathbb{P}\{N(t) \in I\} = \mathbb{P}\Big\{N(t) \in \left[0, N_{\max}\right]\Big\} \geqslant 1 - \alpha \, .$$

La valeur α constitue le risque maximal de ne pas disposer de composants de rechange en nombre suffisant pour réparer toutes les pannes du système avant l'instant t. Bien entendu, pour α donné, notre but est de pouvoir choisir $N_{\max}$ le plus petit possible. L'intervalle I, recherché ici sous la forme $\left[0, N_{\max}\right]$, est appelé *intervalle de prévision*.

Il faut bien distinguer les intervalles de prévision des intervalles de confiance définis au chapitre 3. Les intervalles de confiance fournissent des encadrements de paramètres inconnus du problème tandis que les intervalles de prévision s'intéressent au comportement d'une variable aléatoire autour d'un paramètre, souvent son espérance, et il faut donc, pour les construire, estimer d'une part le paramètre, et d'autre part, le degré de variation de la variable autour du paramètre. Il faut noter que les intervalles de prévision, comme les intervalles de confiance, doivent être construits à l'aide des observations uniquement.

Ici, on cherche à majorer $N(t)$ par $N_{\max}$, qui va bien entendu dépendre de t et de α mais aussi des paramètres du problème (de f *via* μ et σ^2 notamment). Ces paramètres ne sont pas connus *a priori*, ce qui pose problème ; dans un premier temps, nous allons ainsi réaliser une étude *probabiliste* et regarder quelle serait la forme d'intervalles de prévision hypothétiques pouvant dépendre de μ et σ^2. Nous verrons comment nous affranchir de cette difficulté au paragraphe 23.2.9 et exhiberons alors un résultat *statistique*.

23.2.3 Une loi classique en fiabilité : la loi exponentielle

Une difficulté du texte est de choisir une loi appropriée pour les variables $(T_i)_{i \geqslant 1}$ afin de réaliser des simulations. La première idée est de considérer la loi exponentielle, qui est la loi naturelle dans ce contexte comme nous allons le voir à présent. La loi exponentielle de paramètre $\lambda > 0$ vérifie $R(t) = \exp(-\lambda t)$ pour $t > 0$ et on a alors pour tous $s > 0$ et $t > 0$,

$$R(s + t) = R(s)\, R(t) \, .$$

Interprétons cette propriété (comme nous l'avons déjà fait au chapitre 19) :

$$R(s + t) = R(s)R(t) \quad \Longleftrightarrow \quad \mathbb{P}\{T_1 > s + t\} = \mathbb{P}\{T_1 > s\}\, \mathbb{P}\{T_1 > t\}$$

$$\iff \quad \mathbb{P}\{T_1 > s+t \mid T_1 > s\} = \mathbb{P}\{T_1 > t\}\,;$$

c'est ce que l'on avait appelé la propriété d'absence de mémoire de la loi exponentielle : ce qui s'est passé avant l'instant s n'a pas d'incidence sur ce qui se passera après s. Elle caractérise en réalité la loi exponentielle, comme le montre le résultat suivant.

Lemme 23.1 Soit T une variable aléatoire positive telle que sa fiabilité R est strictement positive sur $\mathbb{R}_+$. Alors, on a, pour tous $s > 0$ et $t > 0$,

$$R(s+t) = R(s)\,R(t)$$

si et seulement si T suit une loi exponentielle.

Preuve Compte tenu des calculs précédents, il suffit de montrer le sens direct, que si pour tous $s > 0$ et $t > 0$,

$$R(s+t) = R(s)\,R(t)$$

alors T suit une loi exponentielle. Par récurrence, pour tout $t \in \mathbb{N}^*$, il vient

$$R(t) = R(1)^t = \exp(-\lambda t)\,,$$

où $\lambda = -\log R(1) > 0$. Cette propriété s'étend à l'ensemble des rationnels positifs. On conclut en utilisant la densité des rationnels dans $\mathbb{R}$, la continuité à droite de $R = 1 - F$, et le fait que la fonction de fiabilité R caractérise la loi (tout comme F). □

Le lemme indique donc que la loi exponentielle est (dans notre cadre continu) la seule loi sans mémoire. Nous verrons ultérieurement, au paragraphe 23.2.7, que la loi exponentielle modélise les durées de vie en l'absence des phénomènes d'usure et de rodage. Nous pourrons nous intéresser à l'un de ces phénomènes uniquement en considérant des lois de Weibull (une généralisation de la loi exponentielle, présentée au paragraphe 23.2.4) de paramètres bien choisis.

Lorsque les variables $(T_i)_{i\geqslant 1}$ suivent une loi exponentielle de paramètre $\lambda > 0$, on connaît la loi des instants de renouvellement $(S_n)_{n\geqslant 1}$: chaque S_n suit, par définition, une loi gamma de paramètres n et λ.

Remarque au passage. On peut retrouver facilement la densité de cette dernière. Pour cela, on considère une fonction borélienne bornée h et pour $n \geqslant 1$, on détermine

$$\begin{aligned}
\mathbb{E}\big[h(S_n)\big] &= \mathbb{E}\big[h(T_1 + \cdots + T_n)\big] \\
&= \int h(t_1 + \cdots + t_n) \prod_{i=1}^{n} \Big(\lambda e^{-\lambda t_i} \mathbb{I}_{\{t_i>0\}}\Big)\, \mathrm{d}t_1 \ldots \mathrm{d}t_n \\
&= \lambda^n \int h(t_1 + \cdots + t_n)\, e^{-\lambda(t_1+\ldots+t_n)}\, \mathbb{I}_{\{t_1>0\}} \ldots \mathbb{I}_{\{t_n>0\}}\, \mathrm{d}t_1 \ldots \mathrm{d}t_n
\end{aligned}$$

$$= \lambda^n \int h(u)\, e^{-\lambda u}\, \mathbb{I}_{\{0<s_1<s_2<\ldots<s_{n-1}<u\}}\, \mathrm{d}s_1 \ldots \mathrm{d}s_{n-1}\mathrm{d}u$$
$$= \lambda^n \int h(u)\, e^{-\lambda u}\, \frac{u^{n-1}}{(n-1)!}\mathbb{I}_{\{u>0\}}\, \mathrm{d}u$$

où la troisième égalité procède du changement de variables $s_k = t_1 + \ldots + t_k$ pour tout $k = 1, \ldots, n$ (et le changement de variable muette $u = s_n$), tandis que la dernière vient de ce que l'on intègre (par réarrangement) sur une fraction $1/(n-1)!$ du volume u^{n-1} de l'hypercube de côté u. On a ainsi retrouvé que la densité de la loi gamma de paramètres n et λ est

$$f_n(u) = \lambda^n e^{-\lambda u}\, \frac{u^{n-1}}{(n-1)!}\mathbb{I}_{\{u>0\}}\,.$$

On peut alors déterminer la loi de $N(t)$, en indiquant, pour tout $k \in \mathbb{N}$, la valeur de $\mathbb{P}\{N(t) = k\}$. Pour $k \in \mathbb{N}^*$, en conditionnant par S_k puis en utilisant l'indépendance de S_k et T_{k+1}, on a :

$$\begin{aligned}
\mathbb{P}\{N(t)=k\} &= \mathbb{P}\{S_k \leqslant t,\ S_{k+1} > t\} = \mathbb{P}\{S_k \leqslant t,\ S_k + T_{k+1} > t\} \\
&= \mathbb{E}\Big[\mathbb{P}\big\{S_k \leqslant t,\ S_k + T_{k+1} > t \mid S_k\big\}\Big] \\
&= \int \mathbb{I}_{\{u\leqslant t\}}\, \mathbb{P}\big\{T_{k+1} > t-u \mid S_k = u\big\} f_k(u)\, \mathrm{d}u \\
&= \int_0^t R(t-u)\, f_k(u)\, \mathrm{d}u \\
&= \int_0^t e^{-\lambda(t-u)}\, \lambda^k e^{-\lambda u}\, \frac{u^{k-1}}{(k-1)!}\, \mathrm{d}u = \frac{(\lambda t)^k\, e^{-\lambda t}}{k!}\,.
\end{aligned}$$

Comme par ailleurs $\mathbb{P}\{N(t) = 0\} = \mathbb{P}\{S_1 > t\} = e^{-\lambda t}$, il vient que $N(t)$ suit une loi de Poisson de paramètre λt. (On peut aller plus loin et montrer que N est un processus de Poisson de paramètre λ ; nous ne détaillons pas ce point et renvoyons le lecteur à [CGCDM05].) Nous nous contentons, pour la suite, de la remarque suivante ; pour $t \geqslant 0$,

$$m(t) = \mathbb{E}\big[N(t)\big] = \lambda t = \mathrm{Var}\, N(t)\,.$$

23.2.4 Une généralisation de la loi exponentielle : la loi de Weibull

Pour a et λ deux réels strictement positifs, nous noterons $\mathcal{W}(a, \lambda)$ la loi de Weibull de paramètres a et λ caractérisée de la manière suivante : son support est $\mathbb{R}_+$ et sa fonction de répartition $F_{a,\lambda}$ et sa densité $f_{a,\lambda}$ sont respectivement définies, pour $t \geqslant 0$, par

$$F_{a,\lambda}(t) = 1 - \exp\Big(-(\lambda\, t)^a\Big) \qquad \text{et} \qquad f_{a,\lambda}(t) = a\, \lambda^a\, t^{a-1}\, \exp\Big(-(\lambda\, t)^a\Big)\,.$$

En particulier, on note que $W(1, \lambda)$ est la loi exponentielle de paramètre λ.

Pour préciser l'espérance et la variance d'une loi de Weibull, on rappelle au préalable la définition de la fonction Γ,

$$\Gamma :\ x \in \mathbb{R}_+^* \longmapsto \Gamma(x) = \int_0^{+\infty} t^{x-1}\, e^{-t}\, \mathrm{d}t\ ;$$

$(a,\ \lambda)$	$(0.5,\ 2)$	$(0.5,\ 1)$	$(1,\ 1)$	$(2,\ 1)$
$\mathbb{E}\,\mathcal{W}(a,\lambda)$	1	2	1	0.8862
$\mathrm{Var}\,\mathcal{W}(a,\lambda)$	5	20	1	0.2146

TABLE 23.1. Espérance et variance de la loi de Weibull $\mathcal{W}(a,\lambda)$ pour différentes valeurs du couple de paramètres $(a,\ \lambda)$.

on a alors, par changement de variables simple, que si $X \sim \mathcal{W}(a,\lambda)$, alors

$$\mu = \mathbb{E}\big[X\big] = \frac{1}{\lambda}\,\Gamma\left(1+\frac{1}{a}\right) \quad \text{et} \quad \sigma^2 = \mathrm{Var}\,X = \frac{1}{\lambda^2}\left(\Gamma\left(1+\frac{2}{a}\right) - \Gamma^2\left(1+\frac{1}{a}\right)\right).$$

On reporte dans le tableau 23.1 quelques valeurs numériques de ces espérances et variances.

Même si la plupart des logiciels contiennent une fonction de simulation des lois de Weibull, il est très facile de programmer cette fonction à l'aide de la méthode d'inversion étudiée au chapitre 9. En effet, l'inverse de $F_{a,\lambda}$ est donnée pour $t \in\,]0,\,1[$ par

$$F^{-1}_{a,\lambda}(t) = \frac{1}{\lambda}\big(-\log(1-t)\big)^{1/a}\,,$$

de sorte que si U suit la loi uniforme sur $[0,\,1]$, alors $F^{-1}_{a,\lambda}(U)$ suit la loi $\mathcal{W}(a,\lambda)$, ce qui se traduit ici, U et $1-U$ suivant la même loi, par le fait que

$$\frac{1}{\lambda}\big(-\log U\big)^{1/a} \sim \mathcal{W}(a,\lambda)\,.$$

Nous effectuerons les illustrations informatiques en considérant des lois de Weibull.

23.2.5 Etude asymptotique

On établit ici la preuve des résultats asymptotiques du paragraphe 13.2, la convergence presque sûre de $N(t)/t$ vers $1/\mu$ d'une part, et le résultat de normalité asymptotique correspondant, d'autre part.

Enoncé et preuve de la convergence p.s.

On montre tout d'abord que presque sûrement, $N(t) \to +\infty$ lorsque $t \to +\infty$. On a pour tout $k \in \mathbb{N}$ et pour tout $t \geqslant 0$,

$$N(t) \geqslant k \quad \Longleftrightarrow \quad S_k \leqslant t\,.$$

Donc, si la fonction croissante $t \mapsto N(t)$ est bornée, il existe $k_0 \in \mathbb{N}$ tel que pour tout $t \geqslant 0$, on a $S_{k_0} > t$; ce qui est un événement de probabilité nulle, car $S_{k_0} < \infty$ p.s. On a ainsi prouvé le résultat capital suivant, qui nous rendra encore bien des services,

$$N(t) \underset{t\to+\infty}{\longrightarrow} +\infty \qquad \text{p.s.} \tag{23.1}$$

Montrons à présent la convergence presque sûre de $N(t)/t$. De la définition même de $N(t)$ découle, pour tout $t \geqslant 0$, l'encadrement fondamental

$$S_{N(t)} \leqslant t < S_{N(t)+1} ;$$

il implique ici, pour t tel que $N(t) > 0$, que

$$\frac{S_{N(t)}}{N(t)} \leqslant \frac{t}{N(t)} < \frac{S_{N(t)+1}}{N(t)+1} \ \frac{N(t)+1}{N(t)} .$$

La loi forte des grands nombres énonce que $S_n/n \to \mu$ p.s. ; par composition de limites, en utilisant (23.1), et par encadrement, il vient

$$\frac{t}{N(t)} \underset{t\to+\infty}{\longrightarrow} \mu \quad \text{p.s.} \qquad \text{soit} \qquad \frac{N(t)}{t} \underset{t\to+\infty}{\longrightarrow} \frac{1}{\mu} \quad \text{p.s.}$$

Illustration de cette convergence p.s.

Ici, nous représentons les processus $t \mapsto N(t)$ puis $t \mapsto N(t)/t$. Au premier abord, il peut sembler difficile de représenter $t \mapsto N(t)$, mais il ne faut pas oublier que c'est une fonction en escalier, car le processus est constant entre deux instants de renouvellements. Plus précisément, pour tout entier $n \geqslant 0$, on a $N(t) = n$ pour $t \in [S_n, S_{n+1}[$. Une illustration est donnée dans la partie gauche de la figure 23.1, lorsque les durées de vie $(T_i)_{i\geqslant 1}$ suivent une loi de Weibull de paramètres $a = 2$ et $\lambda = 1$. On y retrouve bien le fait que $N(t) \to +\infty$ p.s. lorsque $t \to +\infty$. La partie droite de cette figure représente $t \mapsto N(t)/t$, et on observe la convergence $N(t)/t \to 1/\mu$ p.s. (voir le tableau 23.1 pour la valeur de μ). En ce qui concerne les ordres de grandeur, on prendra garde au fait que l'espérance de la loi $\mathcal{W}(2, 1)$ étudiée ici est inférieure à 1 ; on s'attend donc, sur le chemin représenté sur la partie droite de la figure 23.1, qu'il y ait eu plus de 2 000 pannes ! D'ailleurs, le graphique de gauche montre bien qu'il y a plus d'une panne par unité de temps.

Enoncé et preuve de la convergence en loi ; construction de l'intervalle de prévision

La convergence en loi est de nature différente de la convergence presque sûre ; en particulier, on ne peut plus ici raisonner trajectoriellement comme précédemment, et utiliser un argument de composition de limites pour déduire du théorème de la limite centrale ordinaire,

$$\sqrt{n}\left(\frac{S_n}{n} - \mu\right) \rightsquigarrow \mathcal{N}\left(0, \sigma^2\right) ,$$

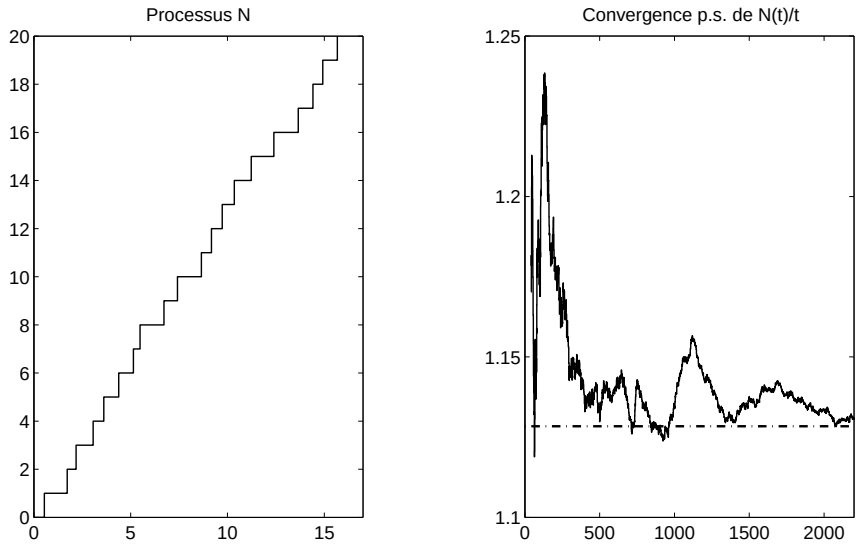

FIGURE 23.1. Processus $t \mapsto N(t)$ (à gauche) et processus $t \mapsto N(t)/t$ (à droite) pour la loi de Weibull de paramètres $a = 2$ et $\lambda = 1$.

un quelconque résultat de convergence de $\sqrt{N(t)}\left(S_{N(t)}/N(t) - \mu\right)$ vers $\mathcal{N}\left(0, \sigma^2\right)$ et en tirer la seconde partie du théorème par méthode delta. Ces considérations ne peuvent que guider notre intuition.

En fait, il faut revenir à la définition de la convergence en loi et calculer pour tout $x \in \mathbb{R}$, la limite de

$$\mathbb{P}\left\{\sqrt{t}\left(\frac{N(t)}{t} - \frac{1}{\mu}\right) > x\right\} = \mathbb{P}\left\{N(t) \geqslant n_{t,x}\right\} = \mathbb{P}\left\{S_{n_{t,x}} \leqslant t\right\},$$

lorsque $t \to +\infty$, où $n_{t,x}$ est le plus petit entier strictement supérieur à $t\left(\frac{1}{\mu} + \frac{x}{\sqrt{t}}\right)$:

$$t\left(\frac{1}{\mu} + \frac{x}{\sqrt{t}}\right) < n_{t,x} \leqslant t\left(\frac{1}{\mu} + \frac{x}{\sqrt{t}}\right) + 1\,. \tag{23.2}$$

Or, en notant Φ_n, pour tout $n \geqslant 1$, la fonction de répartition de la variable aléatoire $(S_n - n\mu)/(\sigma\sqrt{n})$,

$$\mathbb{P}\left\{\sqrt{t}\left(\frac{N(t)}{t} - \frac{1}{\mu}\right) > x\right\} = \mathbb{P}\left\{\frac{S_{n_{t,x}} - n_{t,x}\,\mu}{\sigma\sqrt{n_{t,x}}} \leqslant \frac{t - n_{t,x}\,\mu}{\sigma\sqrt{n_{t,x}}}\right\} = \Phi_{n_{t,x}}\left(\frac{t - n_{t,x}\,\mu}{\sigma\sqrt{n_{t,x}}}\right).$$

Par le théorème de la limite centrale, la suite des Φ_n converge simplement vers Φ, la fonction de répartition d'une variable normale centrée réduite. Ces fonctions Φ_n étant continues et

croissantes et la limite simple Φ étant continue, cette convergence est en fait uniforme sur tout compact, ainsi que l'énonce le second théorème de Dini (voir par exemple [Gou94, page 229]). Par ailleurs, de (23.2), on tire $n_{t,x} \sim t/\mu$ et les encadrements

$$-\mu - x\mu\sqrt{t} \leqslant t - n_{t,x}\,\mu < -x\mu\sqrt{t} \qquad \text{puis} \qquad -\frac{\mu + x\mu\sqrt{t}}{\sigma\sqrt{n_{t,x}}} \leqslant \frac{t - n_{t,x}\,\mu}{\sigma\sqrt{n_{t,x}}} < -\frac{x\mu\sqrt{t}}{\sigma\sqrt{n_{t,x}}}\,;$$

ainsi,

$$\frac{t - n_{t,x}\,\mu}{\sigma\sqrt{n_{t,x}}} \underset{t\to+\infty}{\longrightarrow} -\frac{x\mu^{3/2}}{\sigma}\,,$$

et en considérant un voisinage compact de ce dernier point, il vient

$$\Phi_{n_{t,x}}\left(\frac{t - n_{t,x}\,\mu}{\sigma\sqrt{n_{t,x}}}\right) \underset{t\to+\infty}{\longrightarrow} \Phi\left(-\frac{x\mu^{3/2}}{\sigma}\right) = 1 - \Phi\left(\frac{x\mu^{3/2}}{\sigma}\right).$$

On a finalement montré que pour tout $x \in \mathbb{R}$,

$$\mathbb{P}\left\{\sqrt{t}\left(\frac{N(t)}{t} - \frac{1}{\mu}\right) \leqslant x\right\} \underset{t\to+\infty}{\longrightarrow} \Phi\left(\frac{x\mu^{3/2}}{\sigma}\right), \quad \text{soit} \quad \sqrt{\frac{\mu^3}{\sigma^2}}\,\sqrt{t}\left(\frac{N(t)}{t} - \frac{1}{\mu}\right) \underset{t\to+\infty}{\rightsquigarrow} \mathcal{N}(0,1)\,.$$

On notera qu'on a ici utilisé l'argument de composition de limites sur des fonctions de répartition, mais pas directement sur les statistiques d'intérêt.

De cette convergence, on tire, pour $\alpha \in]0, 1[$,

$$\lim_{t\to+\infty} \mathbb{P}\left\{\sqrt{\frac{\mu^3}{\sigma^2}}\,\sqrt{t}\left(\frac{N(t)}{t} - \frac{1}{\mu}\right) \leqslant z_{1-\alpha}\right\} = \lim_{t\to+\infty} \mathbb{P}\left\{N(t) \leqslant \frac{t}{\mu} + z_{1-\alpha}\,\frac{\sigma\sqrt{t}}{\mu^{3/2}}\right\} = 1 - \alpha$$

où $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite. Ainsi, en choisissant

$$N_{\max} = \frac{t}{\mu} + z_{1-\alpha}\,\frac{\sigma\sqrt{t}}{\mu^{3/2}}\,, \tag{23.3}$$

on répond à la problématique de l'introduction mais uniquement dans un cadre asymptotique. Cela étant, $N_{\max}$ dépend de μ et σ^2, qui sont inconnus *a priori*, mais que l'on peut estimer (voir le paragraphe 23.2.9).

Illustration de cette convergence en loi

On recourt ici à la même méthodologie que pour tracer la figure 19.2. En particulier, on construit un échantillon du processus de sauts $S = (S_j)_{1\leqslant j\leqslant 5\,000}$; pour des raisons de temps de calcul, on s'est limité à un 400–échantillon, noté

$$S^{(1)},\ \ldots,\ S^{(400)}\,.$$

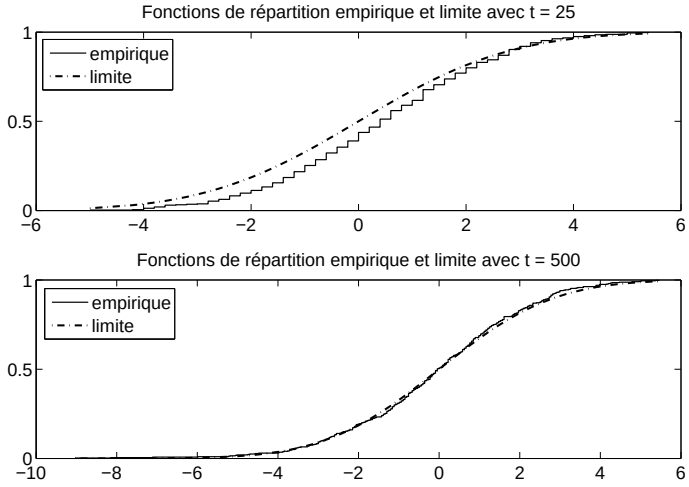

FIGURE 23.2. Pour la loi de Weibull de paramètres $a = 0.5$ et $\lambda = 2$: illustration de la convergence en loi $\sqrt{t}\left(\dfrac{N(t)}{t} - \dfrac{1}{\mu}\right) \rightsquigarrow \mathcal{N}\left(0, \dfrac{\sigma^2}{\mu^3}\right)$.

Pour tout $k \in \{1, \ldots, 400\}$, on calcule le nombre de sauts à l'instant t sur le k–ième élément de l'échantillon,

$$N^{(k)}(t) = \max\left\{j \leqslant 5\,000 :\ S_j^{(k)} \leqslant t\right\} = \sum_{j=1}^{5\,000} \mathbb{I}_{\left\{S_j^{(k)} \leqslant t\right\}}$$

pour $t = 25$ et $t = 500$. (*A priori*, c'est donc $\min\left\{N^{(k)}(t),\ 5\,000\right\}$ que l'on calcule, plutôt que $N^{(k)}(t)$. La valeur 5 000 pour le nombre de sauts a été fixée de sorte que le minimum soit atteint avec très grande probabilité en $N^{(k)}(t)$, une assertion que l'on peut et doit vérifier préalablement par simulations. On notera que l'on s'attend, vu l'espérance égale à 1 des temps de sauts $\mathcal{W}(0.5, 2)$, que $N^{(k)}(t)$ soit de l'ordre de t.) Le théorème de Glivenko–Cantelli (Théorème 8.7) assure qu'on a l'approximation, uniformément en $x \in \mathbb{R}$,

$$\frac{1}{400}\sum_{k=1}^{400} \mathbb{I}_{\left\{\sqrt{t}\left(\frac{N^{(k)}(t)}{t} - \frac{1}{\mu}\right) \leqslant x\right\}} \approx \mathbb{P}\left\{\sqrt{t}\left(\frac{N(t)}{t} - \frac{1}{\mu}\right) \leqslant x\right\} ;$$

le membre de droite convergeant vers $\mathbb{P}\{Z \leqslant x\}$ lorsque $t \to \infty$, où l'on a noté $Z \sim \mathcal{N}\big(0,\ \sigma^2/\mu^3\big)$, on pourra donc dire, d'un point de vue expérimental, que la convergence en loi

est vérifiée si

$$\frac{1}{400}\sum_{k=1}^{400}\mathbb{I}_{\left\{\sqrt{t}\left(\frac{N^{(k)}(t)}{t}-\frac{1}{\mu}\right)\leqslant x\right\}} \approx \mathbb{P}\{Z\leqslant x\}\,.$$

C'est effectivement ce que l'on observe pour $t = 500$ (mais pas encore pour $t = 25$, voir ci-dessous) sur la figure 23.2, où l'on a représenté les fonctions de répartition empiriques et limites introduites précédemment. Le côté en escalier est plus marqué pour $t = 25$, parce que dans ce cas, l'espérance de $N(25)$ étant de l'ordre de la trentaine, l'intervalle des valeurs prises par $N(25)$ est beaucoup plus restreint que celui de $N(500)$.

23.2.6 Etude non-asymptotique dans un cadre général

Pour répondre à la problématique de l'introduction et construire ici l'intervalle de prévision $\big[0,\, N_{\max}\big]$ pour $N(t)$ (lorsque $t > 0$, le cas $t = 0$ étant inintéressant), on s'appuie sur l'inégalité de Bienaymé–Tchebychev. Elle donne, pour $t > 0$ et pour tout $\varepsilon > 0$,

$$\mathbb{P}\Big\{\big|N(t)-\mathbb{E}[N(t)]\big|>\varepsilon\Big\}\leqslant\frac{\operatorname{Var}N(t)}{\varepsilon^2}\,,$$

$$\text{d'où}\qquad \mathbb{P}\left\{N(t)\leqslant\mathbb{E}[N(t)]+\sqrt{\frac{\operatorname{Var}N(t)}{\alpha}}\right\}\geqslant 1-\alpha \quad (23.4)$$

pour tout $\alpha > 0$. On va montrer ci-dessous les deux majorations suivantes,

$$m(t)=\mathbb{E}[N(t)]\leqslant M_{\mu,\sigma^2}(t)\overset{\text{not.}}{=}\frac{t}{\mu}+\frac{\sigma^2}{\mu^2}\qquad\text{et}\qquad\operatorname{Var}N(t)\leqslant m(t)+m^2(t)\,.$$

Ainsi, pour tout $t > 0$ fixé, $\mathbb{P}\big\{N(t) \leqslant N_{\max}\big\} \geqslant 1 - \alpha$ (une garantie cette fois-ci non-asymptotique, donc) pour le choix

$$N_{\max}=M_{\mu,\sigma^2}(t)+\frac{1}{\sqrt{\alpha}}\sqrt{M_{\mu,\sigma^2}(t)+M_{\mu,\sigma^2}(t)^2}\,. \quad (23.5)$$

Comme pour le calcul effectué dans le cadre asymptotique, $N_{\max}$ dépend des paramètres μ et σ^2 inconnus *a priori* (voir le paragraphe 23.2.9 pour traiter cette dépendance). L'objet du paragraphe 23.2.7 sera d'améliorer cette borne, assez grossière, en renforçant les hypothèses sur la loi des variables $(T_i)_{i\geqslant 1}$. Par ailleurs, nous allons pouvoir nous affranchir de la connaissance de σ^2.

Pour l'heure, nous prouvons d'abord quelques résultats préliminaires, puis établissons le contrôle sur l'espérance (majoration et minoration) et sur la variance.

Résultats préliminaires

Des résultats préliminaires sont affirmés dans le texte au début du paragraphe 13.3. Ils sont très classiques et ne constituent pas le cœur du sujet. Aussi, on peut tout à fait décider de ne

pas les démontrer et les admettre pour se concentrer sur les preuves des résultats ultérieurs. Nous en fournissons ici la preuve.

On établit au préalable les différentes expressions de m. On a, pour $t \geqslant 0$,

$$\begin{aligned} m(t) = \mathbb{E}\big[N(t)\big] &= \sum_{k=1}^{+\infty} k\,\mathbb{P}\{N(t)=k\} = \sum_{k=1}^{+\infty}\sum_{\ell=1}^{k} \mathbb{P}\{N(t)=k\} \\ &= \sum_{\ell=1}^{+\infty}\sum_{k=\ell}^{+\infty} \mathbb{P}\{N(t)=k\} = \sum_{\ell=1}^{+\infty} \mathbb{P}\{N(t)\geqslant \ell\} = \sum_{\ell=1}^{+\infty} \mathbb{P}\{S_\ell \leqslant t\}\,, \end{aligned}$$

les interversions étant légitimes parce que l'on ne considère que des quantités positives.

> Remarque au passage. Il est important de bien connaître ces différentes manières de calculer l'espérance d'une variable aléatoire discrète positive, car on en a souvent besoin. Ces calculs s'étendent sans difficulté au cas continu en remplaçant les sommes par des intégrales et en faisant appel au théorème de Fubini–Tonelli. En particulier, on a ici
>
> $$\mu = \mathbb{E}\big[T_1\big] = \int_0^{+\infty} R(x)\,\mathrm{d}x\,. \tag{23.6}$$

On aura également besoin par la suite du fait que pour tout $t \geqslant 0$, la suite des $\mathbb{P}\{S_\ell \leqslant t\}$ est bornée par une suite géométrique. En effet, si pour $\ell \geqslant 1$, on a $S_\ell \leqslant t$, alors en particulier, $\max\{T_1, \ldots, T_\ell\} \leqslant t$, de sorte que

$$\mathbb{P}\{S_\ell \leqslant t\} \leqslant \mathbb{P}\big\{\max\{T_1,\ldots,T_\ell\} \leqslant t\big\} = F(t)^\ell\,. \tag{23.7}$$

Pour la culture, on peut alors encadrer la fonction de renouvellement m à l'aide de la fonction de répartition F,

$$F(t) = \mathbb{P}\{S_1 \leqslant t\} \leqslant m(t) = \sum_{\ell=1}^{+\infty} \mathbb{P}\{S_\ell \leqslant t\} \leqslant \sum_{\ell=1}^{+\infty} F(t)^\ell = \frac{F(t)}{1-F(t)}\,.$$

En particulier, m ne prend que des valeurs finies.

On détermine ensuite f_k, la densité de S_k, pour tout $k \in \mathbb{N}^*$. Bien entendu, f_k est nulle sur $\mathbb{R}_-^*$ et $f_1 = f$. On détermine f_k pour tout $k \geqslant 2$ fixé par identification : pour toute fonction h borélienne et bornée, on a, toujours par théorème de Fubini–Tonelli et changement de variables,

$$\begin{aligned} \mathbb{E}\big[h(S_k)\big] &= \mathbb{E}\big[h(S_{k-1}+T_k)\big] \\ &= \iint h(u+y)\, f_{k-1}(u) f(y)\,\mathrm{d}u\,\mathrm{d}y \\ &= \int_0^{+\infty} h(x) \left(\int_0^x f_{k-1}(u) f(x-u)\,\mathrm{d}u \right) \mathrm{d}x\,. \end{aligned}$$

Donc, par identification, pour $x \geqslant 0$, et vu que $f = 0$ sur $\mathbb{R}_-^*$,

$$f_k(x) = \int_0^x f_{k-1}(u) f(x-u)\,\mathrm{d}u = \int_{-\infty}^{+\infty} f_{k-1}(u) f(x-u)\,\mathrm{d}u\ ;$$

cette égalité est encore vraie pour $x < 0$: f_k est bien le produit de convolution entre f_{k-1} et f, noté $f_{k-1} \star f$.

> Remarque au passage. Pour des rappels sur la définition de l'opérateur de convolution $\star$ et les propriétés qu'on en utilise ici et tout au long du corrigé, on pourra se référer au paragraphe 24.1.6, ainsi qu'à [Gou94, Bre99].

Enfin, on montre maintenant que la fonction m est dérivable sur $\mathbb{R}$ et on calcule l'expression de sa dérivée. A cet effet, fixons $a > 0$ et posons

$$\beta(a) = \int_0^a f(u)\,\mathrm{d}u < 1\ ,$$

l'inégalité provenant du fait que f est strictement positive sur $\mathbb{R}_+$. Pour toute fonction g, on note

$$\|g\|_\infty = \sup_{x\in\mathbb{R}} \big|g(x)\big| = \sup_{\mathbb{R}} |g|\ .$$

Pour $x \in [0, a]$,

$$f_k(x) = \int_0^x f_{k-1}(u) f(x-u)\,\mathrm{d}u \leqslant \left(\sup_{[0,a]} |f_{k-1}|\right) \int_0^x f(u)\,\mathrm{d}u \leqslant \left(\sup_{[0,a]} |f_{k-1}|\right) \beta(a)\ .$$

Par récurrence, on obtient

$$\sup_{[0,a]} |f_k| \leqslant \left(\sup_{[0,a]} |f|\right) \beta(a)^{k-1} \leqslant \|f\|_\infty\, \beta(a)^{k-1}\ ,$$

ce qui était le résultat affirmé dans le texte, pour $c = \|f\|_\infty < \infty$ (par hypothèse f est bornée). La série $\sum f_k$ converge donc normalement sur $[0, a]$ (et même sur $]-\infty, a]$). Or, on a vu que pour tout $t \geqslant 0$,

$$m(t) = \sum_{k=1}^{+\infty} \mathbb{P}\{S_k \leqslant t\} = \sum_{k=1}^{\infty} \int_0^t f_k(x)\,\mathrm{d}x = \int_0^t \left(\sum_{k=1}^{\infty} f_k(x)\right) \mathrm{d}x\ ,$$

la dernière égalité procédant de l'application du théorème de Fubini–Tonelli. Le théorème de dérivation pour les séries de fonctions montre alors que la fonction m est finie et dérivable sur $\mathbb{R}$. Sa dérivée est en outre bornée sur tout compact de $\mathbb{R}$ et vaut pour $x \in \mathbb{R}$,

$$m'(x) = \sum_{k=1}^{+\infty} f_k(x)\ . \tag{23.8}$$

Deux propositions utiles pour l'ensemble du corrigé

L'objet de ce paragraphe est de construire différentes briques de résultats, pour établir plus tard les contrôles de l'espérance et de la variance de $N(t)$ dans le cas général, mais aussi dans le cas NBU du paragraphe 23.2.7. En particulier, l'identité de Wald ne servira pas réellement avant ce moment, mais il nous semble plus naturel de la prouver dès ici, ne serait-ce que parce qu'elle permet de minorer facilement l'espérance de $N(t)$. En effet, on déduit de l'inégalité $S_{N(t)+1} > t$ pour tout $t \geqslant 0$ la minoration $m(t) \geqslant t/\mu - 1$ proposée par le Théorème 13.2.

Proposition 23.2 On a pour $t \geqslant 0$,

$$m(t) = F(t) + m \star f(t) \, . \tag{23.9}$$

Par ailleurs, on a l'égalité suivante, appelée « identité de Wald » : pour $t \geqslant 0$,

$$\mathbb{E}\big[S_{N(t)+1}\big] = \mu\big(m(t)+1\big) \, .$$

Preuve On commence par un raisonnement tout à fait similaire à celui effectué à la fin du paragraphe 23.2.3. On décompose S_k, pour $k \geqslant 2$, en $S_{k-1} + T_k$ et on utilise le fait que ces deux variables sont indépendantes ; pour $t \geqslant 0$,

$$\begin{aligned}
m(t) = \sum_{k=1}^{+\infty} \mathbb{P}\{S_k \leqslant t\} &= \mathbb{P}\{T_1 \leqslant t\} + \sum_{k=2}^{+\infty} \mathbb{P}\{S_{k-1} + T_k \leqslant t\} \\
&= F(t) + \sum_{k=2}^{+\infty} \int_0^t \mathbb{P}\big\{S_{k-1} + x \leqslant t \mid T_k = x\big\} \, f(x) \, \mathrm{d}x \\
&= F(t) + \int_0^t \sum_{k=1}^{+\infty} \mathbb{P}\{S_k \leqslant t - x\} \, f(x) \, \mathrm{d}x \\
&= F(t) + \int_0^t m(t-x) f(x) \, \mathrm{d}x \, ,
\end{aligned}$$

où l'on rappelle que F est la fonction de répartition commune aux T_k. Cela nous donne le premier résultat, $m = F + m \star f$ sur $\mathbb{R}_+$ (qui, d'ailleurs, est trivialement vrai sur $\mathbb{R}_-^*$ également).

L'identité de Wald s'obtient par le calcul suivant,

$$\begin{aligned}
\mathbb{E}\big[S_{N(t)+1}\big] &= \sum_{k=1}^{+\infty} \mathbb{E}\left[\sum_{j=1}^{k} T_j \, \mathbb{I}_{\{N(t)+1=k\}}\right] = \sum_{j=1}^{+\infty} \mathbb{E}\big[T_j \, \mathbb{I}_{\{N(t)\geqslant j-1\}}\big] \\
&= \sum_{j=1}^{+\infty} \mathbb{E}\big[T_j\big] \mathbb{P}\{N(t) \geqslant j-1\}
\end{aligned}$$

$$= \quad \mu \left(1 + \sum_{j=1}^{+\infty} \mathbb{P}\{N(t) \geqslant j\} \right) = \mu\big(1 + m(t)\big) \,,$$

qui n'a utilisé que le théorème de Fubini–Tonelli et le fait que pour $j \geqslant 1$, l'événement $\{N(t) \geqslant j - 1\} = \{S_{j-1} \leqslant t\}$ et la variable T_j sont indépendants. □

Remarque au passage. L'identité de Wald peut également être démontrée en utilisant le théorème d'arrêt des martingales (voir, par exemple, [MPB98]). En effet, on définit $M_n = S_n - n\mu$ pour tout $n \geqslant 0$, et $(\mathcal{F}_n)$ comme la filtration engendrée par la suite $T_0, T_1, \ldots, T_n$ (avec $T_0 = 0$). Alors, pour $n \geqslant 1$, on a par indépendance des T_i que (M_n) est une $(\mathcal{F}_n)$–martingale. On remarque par ailleurs que

$$N(t) + 1 = \inf\{n \geqslant 1 : \; S_n > t\}$$

est un temps d'arrêt pour la filtration $(\mathcal{F}_n)$. Par le théorème d'arrêt, on a alors que

$$\big(M_{\min\{n,\, N(t)+1\}}\big)_{n \geqslant 0}$$

est une martingale. En particulier,

$$\mathbb{E}\big[M_{\min\{n,\, N(t)+1\}}\big] = \mathbb{E}[M_0] = 0 \qquad \text{d'où} \qquad \mathbb{E}\big[S_{\min\{n,\, N(t)+1\}}\big] = \mu\, \mathbb{E}\big[\min\{n,\; N(t)+1\}\big] \,.$$

Ainsi, en faisant tendre $n \to \infty$, en utilisant le théorème de convergence monotone et le fait que $N(t) < +\infty$ p.s., on obtient finalement

$$\mathbb{E}\big[S_{N(t)+1}\big] = \mu\, \mathbb{E}\big[N(t) + 1\big] = \mu\big(m(t) + 1\big) \,.$$

On établit à présent la proposition suivante ; on rappelle que $B(t) = S_{N(t)+1} - t$ pour $t \geqslant 0$.

Proposition 23.3 Pour $t \geqslant 0$ et $x \geqslant 0$,

$$\mathbb{P}\,\{B(t) > x\} = R(t + x) + \int_0^t R(t + x - u)\, m'(u)\, \mathrm{d}u \;;$$

et par conséquent, $B(t)$ a une loi de support dans $\mathbb{R}_+$ et de densité donnée par

$$x \in \mathbb{R}_+ \;\longmapsto\; f(t + x) + \int_0^t f(t + x - u)\, m'(u)\, \mathrm{d}u \,.$$

Preuve On décompose à nouveau selon les valeurs de $N(t)$,

$$\begin{aligned}
\mathbb{P}\{B(t) > x\} &= \sum_{n=1}^{+\infty} \mathbb{P}\big\{N(t)+1 = n \text{ et } S_n > t+x\big\} \\
&= \mathbb{P}\big\{S_1 > t+x\} + \sum_{n=1}^{+\infty} \mathbb{P}\big\{N(t) = n \text{ et } S_{n+1} > t+x\big\} \\
&= R(t+x) + \sum_{n=1}^{+\infty} \mathbb{P}\big\{S_n \leqslant t \text{ et } S_n + T_{n+1} > t+x\big\} \\
&= R(t+x) + \sum_{n=1}^{+\infty} \int \mathbb{I}_{\{u \leqslant t\}}\, \mathbb{P}\big\{T_{n+1} > t+x-u \mid S_n = u\big\}\, f_n(u)\, \mathrm{d}u \\
&= R(t+x) + \sum_{n=1}^{+\infty} \int_0^t R(t+x-u)\, f_n(u)\, \mathrm{d}u \\
&= R(t+x) + \int_0^t R(t+x-u) \left(\sum_{n=1}^{+\infty} f_n(u) \right) \mathrm{d}u
\end{aligned}$$

où l'on a procédé comme dans la preuve de la Proposition 23.2 : en conditionnant par S_n (quatrième égalité), en utilisant l'indépendance de S_n et T_{n+1} (cinquième égalité), le théorème de Fubini–Tonelli (sixième égalité), ce qui nous permet de reconnaître finalement l'expression (23.8) de m'. Cela conclut la preuve de la première partie de cette proposition ; la seconde s'en déduit par dérivation (et notamment, en utilisant un théorème de dérivation des intégrales, par exemple le Théorème 26.1). □

Remarque au passage. La fonction B est la durée de vie résiduelle du système : c'est le temps d'attente entre un temps d'observation quelconque et la prochaine panne. Lorsque f est la densité d'une loi exponentielle de paramètre $\lambda > 0$ alors pour tout $x > 0$, on a $R(x) = \exp(-\lambda x)$, puis $m(x) = \lambda x$ et $m'(x) = \lambda$. La proposition montre ainsi que $B(t)$ suit également une loi exponentielle de même paramètre λ. Cela illustre encore une fois le caractère d'absence de mémoire de la loi exponentielle.

La fonction m' qui intervient dans la Proposition 23.3 est appelée la densité de renouvellement. La proposition suivante permet une interprétation intéressante de cette fonction.

Proposition 23.4 On a :

$$\frac{\mathbb{P}\{B(t) \leqslant x\}}{x} \underset{x \to 0}{\longrightarrow} m'(t)\,.$$

Quand x est proche de 0, la probabilité d'attendre la prochaine panne pendant un temps inférieur à x est environ $x\, m'(t)$: la densité de renouvellement apparaît donc comme la limite de la probabilité de panne par unité de temps.

Preuve On applique la Proposition 23.3 (deux fois) et on utilise la continuité uniforme de f sur tout compact. Soit $\varepsilon > 0$. Il existe $\delta > 0$ tel que pour tout $0 \leqslant x \leqslant \delta$,

$$\begin{aligned}
&\mathbb{P}\{B(t) > x\} \\
&\quad = \quad R(t+x) + \int_0^t R(t+x-u) m'(u)\,\mathrm{d}u \\
&\quad \leqslant \quad R(t) - x f(t) + \varepsilon x + \int_0^t m'(u)\big(R(t-u) - x f(t-u) + \varepsilon x\big)\,\mathrm{d}u \\
&\quad = \quad \mathbb{P}\{B(t) > 0\} - x\left(f(t) + \int_0^t m'(u) f(t-u)\,\mathrm{d}u\right) + \big(1 + m(t)\big)\varepsilon x \\
&\quad = \quad 1 - x\, m'(t) + \big(1 + m(t)\big)\varepsilon x
\end{aligned}$$

où l'on a utilisé que $\mathbb{P}\{B(t) = 0\} = 0$. Quant à l'égalité

$$m'(t) = f(t) + \int_0^t m'(u) f(t-u)\,\mathrm{d}u\ ,$$

elle provient de la dérivation des deux membres de (23.9) et de l'emploi du théorème de dérivation sous le signe intégrale.
On exhibe de même une minoration de $\mathbb{P}\{B(t) > x\}$ par $1 - x m'(t) - \big(1 + m(t)\big)\varepsilon x$, et on a ainsi prouvé le résultat attendu. □

Enfin, on explique maintenant comment la considération de la quantité jumelle $t - S_{N(t)}$ conduit au paradoxe dit de l'inspection.

Le paradoxe de l'inspection (durées inter-pannes de loi exponentielle)

On se place également dans le cas où la loi des variables $(T_i)_{i \geqslant 1}$ est exponentielle de paramètre λ. On commence par déterminer la loi de $t - S_{N(t)}$. A cet effet, on calcule, pour tout $x \in \mathbb{R}$, la valeur de $\mathbb{P}\{t - S_{N(t)} \leqslant x\}$. Fixons $x \in [0, t[$ (ainsi, $t - x > 0$) ;

$$\begin{aligned}
\mathbb{P}\big\{t - S_{N(t)} \leqslant x\big\} \quad &= \quad \sum_{k=0}^{+\infty} \mathbb{P}\{S_k \geqslant t - x,\ N(t) = k\} \\
&= \quad \sum_{k=1}^{+\infty} \mathbb{P}\{t - x \leqslant S_k \leqslant t,\ \ S_{k+1} > t\}\ .
\end{aligned}$$

On poursuit le calcul comme aux Propositions 23.2 et 23.3 (en décomposant S_{k+1} comme $S_k + T_{k+1}$ et en utilisant l'indépendance entre S_k et T_{k+1}), puis on injecte l'expression explicite de la densité f_k de S_k donnée au paragraphe 23.2.3,

$$\mathbb{P}\big\{t - S_{N(t)} \leqslant x\big\} \quad = \quad \sum_{k=1}^{+\infty} \int_{t-x}^{t} \mathbb{P}\{T_{k+1} > t - u\}\, f_k(u)\,\mathrm{d}u$$

$$
\begin{aligned}
&= \sum_{k=1}^{+\infty} \int_{t-x}^{t} e^{-\lambda(t-u)}\, \lambda^k e^{-\lambda u} \frac{u^{k-1}}{(k-1)!}\, \mathrm{d}u \\
&= e^{-\lambda t} \left(\sum_{k=1}^{+\infty} \int_{t-x}^{t} \lambda^k \frac{u^{k-1}}{(k-1)!}\, \mathrm{d}u \right) \\
&= e^{-\lambda t} \int_{t-x}^{t} \lambda e^{\lambda u}\, \mathrm{d}u = 1 - e^{-\lambda x}\,.
\end{aligned}
$$

On a donc

$$
\mathbb{P}\big\{t - S_{N(t)} \leqslant x\big\} = \begin{cases} 0 & \text{si } x \leqslant 0, \\ 1 - e^{-\lambda x} & \text{si } x \in [0, t[, \\ 1 & \text{si } x \geqslant t. \end{cases}
$$

En particulier, $\mathbb{P}\big\{t - S_{N(t)} > x\big\} = e^{-\lambda x}$ pour $x \in [0, t[$ (et $= 0$ pour $x \geqslant t$). On en déduit, par l'égalité générale (23.6), que

$$
\mathbb{E}\big[t - S_{N(t)}\big] = \int_0^{+\infty} \mathbb{P}\big\{t - S_{N(t)} > x\big\}\, \mathrm{d}x = \frac{1}{\lambda} \Big[-e^{-\lambda x}\Big]_0^t = \frac{1 - e^{-\lambda t}}{\lambda}\,.
$$

Par ailleurs, on a vu à la page 346 que $B(t)$ suivait une loi exponentielle de paramètre λ, d'où

$$
\mathbb{E}\big[S_{N(t)+1} - t\big] = \mathbb{E}\big[B(t)\big] = \frac{1}{\lambda}\,.
$$

En particulier,

$$
\mathbb{E}\big[T_{N(t)+1}\big] = \mathbb{E}\big[S_{N(t)+1} - t\big] + \mathbb{E}\big[t - S_{N(t)}\big] = \frac{2 - e^{-\lambda t}}{\lambda}
$$

alors que pour tout $i \geqslant 1$, on avait $\mathbb{E}[T_i] = 1/\lambda$.

C'est ceci qui forme le paradoxe de l'inspection : la durée entre deux pannes qu'observe un inspecteur arrivé au temps t n'est pas la durée inter-pannes du processus. Cela provient évidemment du fait que la loi de $S_{N(t)+1} - S_{N(t)}$ n'est pas la même que celle des $S_{n+1} - S_n$ (à savoir, la loi exponentielle de paramètre λ). Cela s'explique de manière intuitive par le fait que t appartient obligatoirement à $\big[S_{N(t)}, S_{N(t)+1}\big]$ et que ce dernier doit donc être plus grand.

Majoration de $m(t) = E[N(t)]$

On s'appuie sur la fonction m_0 définie par

$$
m_0(t) = \begin{cases} m(t) + 1 & \text{si } t > 0, \\ 0 & \text{sinon.} \end{cases}
$$

Puisque $S_0 = 0$, on a pour tout $t > 0$,

$$m_0(t) = 1 + \sum_{k=1}^{+\infty} \mathbb{P}\{S_k \leqslant t\} = \sum_{k=0}^{+\infty} \mathbb{P}\{S_k \leqslant t\} .$$

Lemme 23.5 La fonction m_0 est sous-additive, *id est*, pour tous $t \in \mathbb{R}$ et $h \in \mathbb{R}$,

$$m_0(t+h) \leqslant m_0(h) + m_0(t) .$$

Preuve Considérons au préalable le cas $t > 0$ et $h > 0$; comme l'on n'a affaire qu'à des séries absolument convergentes, on peut regrouper, décomposer et échanger des sommations comme suit (pour les trois premières égalités),

$$\begin{aligned}
m_0(t+h) - m_0(t) &= \sum_{k=1}^{+\infty} \mathbb{P}\left\{t < S_k \leqslant t+h\right\} \\
&= \sum_{k=1}^{+\infty} \sum_{\ell=1}^{k} \mathbb{P}\Big\{t < S_k \leqslant t+h \text{ et } S_{\ell-1} \leqslant t < S_\ell\Big\} \\
&= \sum_{\ell=1}^{+\infty} \sum_{k=\ell}^{+\infty} \mathbb{P}\Big\{S_k - S_\ell \leqslant t+h-S_\ell, \quad S_{\ell-1} \leqslant t \text{ et } t < S_\ell \leqslant t+h\Big\} \\
&= \sum_{\ell=1}^{+\infty} \sum_{k=\ell}^{+\infty} \mathbb{E}\Big[\mathbb{I}_{\{S_{\ell-1} \leqslant t < S_{\ell-1}+T_\ell \leqslant t+h\}}\, \mathbb{P}\Big\{S_k - S_\ell \leqslant t+h-(S_{\ell-1}+T_\ell) \mid S_{\ell-1},\ T_\ell\Big\}\Big] \\
&= \sum_{\ell=1}^{+\infty} \sum_{k=\ell}^{+\infty} \int_0^t \int_{t-u}^{t+h-u} \mathbb{P}\Big\{T_{\ell+1} + \ldots + T_k \leqslant t+h-(u+v)\Big\}\, f_{\ell-1}(u) f(v)\, \mathrm{d}u\mathrm{d}v \\
&= \sum_{\ell=1}^{+\infty} \int_0^t \int_{t-u}^{t+h-u} m_0\big(t+h-(u+v)\big)\, f_{\ell-1}(u) f(v)\, \mathrm{d}u\mathrm{d}v \\
&\leqslant m_0(h) \sum_{\ell=1}^{+\infty} \int_0^t \int_{t-u}^{t+h-u} f_{\ell-1}(u) f(v)\, \mathrm{d}u\mathrm{d}v \\
&= m_0(h) \sum_{\ell=1}^{+\infty} \mathbb{P}\Big\{S_{\ell-1} \leqslant t \text{ et } t < S_{\ell-1} + T_\ell \leqslant t+h\Big\} \\
&\leqslant m_0(h) \sum_{\ell=1}^{+\infty} \mathbb{P}\Big\{S_{\ell-1} \leqslant t < S_{\ell-1} + T_\ell\Big\} \;=\; m_0(h)
\end{aligned}$$

où les ingrédients successifs ont été : un conditionnement par $S_{\ell-1}$ et T_ℓ (quatrième égalité), une utilisation de l'indépendance de $T_{\ell+1}, \ldots, T_k$ par rapport à $S_{\ell-1}$ et T_ℓ (cinquième égalité), la définition de m_0 (sixième égalité), le caractère croissant de m_0 et le fait que $u + v \geqslant t$ vu

les intervalles de sommation (septième ligne, première inégalité) ; les deux dernières lignes reposent quant à elles sur les mêmes techniques de découpages, regroupements, et échanges de sommations que les trois premières.

Cette propriété de sous-additivité s'étend à toutes valeurs de t et h. En effet, si t et h sont tels que $t + h \leqslant 0$, la propriété est vérifiée puisque $m_0(t + h) = 0$. Si $t \leqslant 0$ et $h > 0$ avec $t + h > 0$, on a, vu que m_0 est croissante,

$$m_0(t + h) \leqslant m_0(h) = m_0(t) + m_0(h) .$$

Le cas $h \leqslant 0$ et $t > 0$ avec $t + h > 0$ est symétrique. □

Preuve de la majoration du Théorème 13.2 On introduit deux variables U et U' de même loi et indépendantes des variables $(T_i)_i$. On suppose que la loi de U a pour densité la fonction

$$u : x \in \mathbb{R} \mapsto \frac{1}{\mu} R(x) \mathbb{I}_{\{x>0\}} .$$

(On utilise (23.6) pour établir que u est effectivement une densité de probabilité.) Or, par sous-additivité de m_0, pour $t \in \mathbb{R}_+$,

$$m_0(t) = \mathbb{E}\big[m_0(t + U - U' + U' - U)\big] \leqslant \mathbb{E}\big[m_0(t + U - U')\big] + \mathbb{E}\big[m_0(U' - U)\big] .$$

Comme m_0 est nulle sur $\mathbb{R}_-$, pour tout $s \in \mathbb{R}_+$, on a

$$\mathbb{E}\big[m_0(s - U)\big] = \int_0^s m_0(s - x)u(x)\,\mathrm{d}x = \frac{1}{\mu}\left(\int_0^s R(x)\,\mathrm{d}x + \int_0^{+\infty} m(s - x)R(x)\,\mathrm{d}x\right) .$$

La Proposition 23.3 et un changement de variables montrent que pour tout $t \in \mathbb{R}_+$,

$$1 = \mathbb{P}\{B(t) > 0\} = R(t) + \int_0^t m'(t - u)R(u)\,\mathrm{d}u = R(t) + \int_0^{+\infty} m'(t - u)R(u)\,\mathrm{d}u$$

où la première égalité se justifie en distinguant selon les valeurs de $N(t)$ et en arguant que la loi de chaque S_{k+1} est continue et ne charge donc pas t.

On en déduit, par intégration et par échange d'intégrations, légitime par positivité des intégrandes et théorème de Fubini–Tonelli, que pour tout $s \in \mathbb{R}_+$,

$$\begin{aligned} s &= \int_0^s R(x)\,\mathrm{d}x + \int_0^{+\infty} \big(m(s - u) - m(-u)\big)R(u)\,\mathrm{d}u \\ &= \int_0^s R(x)\,\mathrm{d}x + \int_0^{+\infty} m(s - u)R(u)\,\mathrm{d}u , \end{aligned}$$

ce qui permet de conclure que pour tout $s \in \mathbb{R}_+$,

$$\mathbb{E}\big[m_0(s - U)\big] = \frac{s}{\mu} .$$

On tire de cette égalité, appliquée en $t + U \geqslant 0$,

$$\mathbb{E}\big[m_0(t+U-U')\big] = \mathbb{E}\Big[\mathbb{E}\big[m_0(t+U-U') \mid U\big]\Big] = \mathbb{E}\left[\frac{t+U}{\mu}\right] = \frac{t+\mathbb{E}[U]}{\mu}$$

et, de même,

$$\mathbb{E}\big[m_0(U'-U)\big] = \mathbb{E}\Big[\mathbb{E}\big[m_0(U'-U) \mid U'\big]\Big] = \frac{\mathbb{E}[U']}{\mu}\,.$$

Un calcul élémentaire, reposant sur la définition de l'espérance et les densités en jeu, donne finalement, par le théorème de Fubini–Tonelli,

$$\begin{aligned}\mathbb{E}[U] = \int s\,u(s)\,\mathrm{d}s &= \frac{1}{\mu}\int_0^{+\infty} s\,R(s)\,\mathrm{d}s = \frac{1}{\mu}\int_0^{+\infty} s\left(\int_s^{+\infty} f(x)\,\mathrm{d}x\right)\mathrm{d}s \\ &= \frac{1}{\mu}\int_0^{+\infty} f(x)\left(\int_0^x s\,\mathrm{d}s\right)\mathrm{d}x = \frac{1}{2\mu}\mathbb{E}\big[T_1^2\big]\,.\end{aligned}$$

En injectant ceci plus haut, il vient, pour $t > 0$,

$$m_0(t) \leqslant \frac{1}{\mu}\left(t + \frac{1}{\mu}\,\mathbb{E}\big[T_1^2\big]\right) \qquad \text{puis finalement} \qquad m(t) \leqslant \frac{t}{\mu} + \frac{\mathbb{E}\big[T_1^2\big]}{\mu^2} - 1 = \frac{t}{\mu} + \frac{\sigma^2}{\mu^2}\,,$$

ce qui était le résultat désiré. □

Remarque au passage. On peut remarquer que le Théorème 13.2 nous donne un contrôle asymptotique de la fonction de renouvellement. On obtient en effet que lorsque $t \to +\infty$,

$$m(t) = \frac{t}{\mu} + O(1)\,.$$

Il est possible de préciser encore cette convergence : dans le cadre des hypothèses du texte, on peut établir que

$$\lim_{t\to+\infty} m(t) - \frac{t}{\mu} = \frac{\sigma^2}{2\mu^2} - \frac{1}{2}$$

(cf. [Bon95, page 137]).

Majoration de $\mathrm{Var}\,N(t)$

Il s'agit de prouver ici que $\mathrm{Var}\,N(t) \leqslant m(t) + m^2(t)$. La proposition 23.6 ci-dessous, alliée au caractère croissant de m, conduit à ce résultat :

$$\begin{aligned}\mathrm{Var}\,N(t) &= 2\int_0^t m(t-x)m'(x)\,\mathrm{d}x + m(t) - m^2(t) \\ &\leqslant 2m(t)\int_0^t m'(x)\,\mathrm{d}x + m(t) - m^2(t) = m^2(t) + m(t)\,.\end{aligned}$$

Remarque au passage. On peut encore affiner ce contrôle et montrer l'inégalité suivante, valable pour $t \geqslant \mu$:

$$\operatorname{Var} N(t) \leqslant \frac{t}{\mu}\left(\frac{4\sigma^2}{\mu^2}+3\right)+\frac{\sigma^2}{\mu^2}\left(\frac{2\sigma^2}{\mu^2}+1\right)-1\,, \tag{23.10}$$

On l'obtient en injectant les majorations et minorations de $m(t)$ et de $m(t-x)$ données par le Théorème 13.2 tout en intégrant par parties. Plus précisément, pour $t \geqslant \mu$,

$$\begin{aligned}
\operatorname{Var} N(t) &= 2\int_0^t m(t-x)m'(x)\,\mathrm{d}x + m(t) - m^2(t) \\
&\leqslant 2\int_0^t \left(\frac{t-x}{\mu}+\frac{\sigma^2}{\mu^2}\right)m'(x)\,\mathrm{d}x + m(t) - m^2(t) \\
&= 2\left(\frac{t}{\mu}+\frac{\sigma^2}{\mu^2}\right)m(t) - \frac{2}{\mu}\int_0^t x\,m'(x)\,\mathrm{d}x + m(t) - m^2(t) \\
&= 2\left(\frac{t}{\mu}+\frac{\sigma^2}{\mu^2}\right)m(t) - \frac{2}{\mu}\left(t\,m(t) - \int_0^t m(x)\,\mathrm{d}x\right) + m(t) - m^2(t) \\
&\leqslant \left(\frac{2\sigma^2}{\mu^2}+1\right)m(t) + \frac{2}{\mu}\int_0^t \left(\frac{x}{\mu}+\frac{\sigma^2}{\mu^2}\right)\mathrm{d}x - \left(\frac{t}{\mu}-1\right)^2 \\
&\leqslant \left(\frac{2\sigma^2}{\mu^2}+1\right)\left(\frac{t}{\mu}+\frac{\sigma^2}{\mu^2}\right) + \frac{2}{\mu}\left(\frac{t^2}{2\mu}+\frac{\sigma^2 t}{\mu^2}\right) - \left(\frac{t}{\mu}-1\right)^2 \\
&\leqslant \frac{t}{\mu}\left(\frac{4\sigma^2}{\mu^2}+3\right)+\frac{\sigma^2}{\mu^2}\left(\frac{2\sigma^2}{\mu^2}+1\right)-1\,.
\end{aligned}$$

La proposition qui suit est donc le dernier résultat technique à prouver avant de conclure la preuve de l'énoncé de l'intervalle de prévision général.

Proposition 23.6 On a, pour tout $t > 0$,

$$\operatorname{Var} N(t) = 2\,m \star m'(t) + m(t) - m^2(t)\,.$$

Preuve Fixons $t > 0$. On sait, par (23.7), que les probabilités $\mathbb{P}\{S_k \leqslant t\}$ sont majorées par des quantités décroissant exponentiellement en k ; cela justifie ci-dessous que $N^2(t)$ soit intégrable et valide les échanges ou coupures de sommations :

$$\begin{aligned}
\operatorname{Var} N(t) &= \mathbb{E}\Big[N^2(t)\Big] - m^2(t) \\
&= \sum_{k=1}^{+\infty} k^2\,\mathbb{P}\big\{N(t)=k\big\} - m^2(t) = \sum_{k=1}^{+\infty} k^2\,\mathbb{P}\{S_k \leqslant t \text{ et } S_{k+1} > t\} - m^2(t) \\
&= \sum_{k=1}^{+\infty} k^2\left(\mathbb{P}\{S_k \leqslant t\} - \mathbb{P}\{S_{k+1} \leqslant t\}\right) - m^2(t)
\end{aligned}$$

$$= \sum_{k=1}^{+\infty} k^2 \, \mathbb{P}\{S_k \leqslant t\} - \sum_{k=1}^{+\infty} (k-1)^2 \, \mathbb{P}\{S_k \leqslant t\} - m^2(t)$$
$$= \sum_{k=1}^{+\infty} (2k-1) \, \mathbb{P}\{S_k \leqslant t\} - m^2(t) \, .$$

En posant, pour $u > 0$,

$$W(u) = \sum_{k=1}^{+\infty} k \, \mathbb{P}\{S_k \leqslant u\}$$

(et $W = 0$ sur $\mathbb{R}_-$), on obtient bien, vu la réécriture de $m(t)$ proposée en préliminaires,

$$\operatorname{Var} N(t) = 2W(t) - m(t) - m^2(t) \, .$$

Par ailleurs,

$$\begin{aligned} W(t) &= F(t) + \sum_{k=2}^{+\infty} k \, \mathbb{P}\{S_{k-1} + T_k \leqslant t\} = F(t) + \sum_{k=1}^{+\infty} (k+1) \, \mathbb{P}\{S_k + T_{k+1} \leqslant t\} \\ &= F(t) + \sum_{k=1}^{+\infty} (k+1) \int_0^{+\infty} \mathbb{P}\big\{S_k + u \leqslant t \mid T_{k+1} = u\big\} \, f(u) \, \mathrm{d}u \\ &= F(t) + \sum_{k=1}^{+\infty} k \int_0^{+\infty} \mathbb{P}\{S_k \leqslant t-u\} \, f(u) \, \mathrm{d}u + \sum_{k=1}^{+\infty} \int_0^{+\infty} \mathbb{P}\{S_k \leqslant t-u\} \, f(u) \, \mathrm{d}u \\ &= F(t) + m \star f(t) + W \star f(t) = m(t) + W \star f(t) \end{aligned}$$

où l'on a utilisé l'indépendance de S_k et T_{k+1} à la quatrième égalité, la définition de m, W et $\star$ à la cinquième, et (23.9) pour la dernière d'entre elles. Ainsi,

$$\operatorname{Var} N(t) = 2\big(m(t) + W \star f(t)\big) - m(t) - m^2(t) = 2W \star f(t) + m(t) - m^2(t) \, .$$

D'autre part, en utilisant encore une fois $W = m + W \star f$, le théorème de Fubini–Tonelli, puis (23.9),

$$W \star m = m \star m + (W \star f) \star m = m \star m + W \star (f \star m) = m \star m + W \star (m - F) \, ,$$

soit

$$W \star F = m \star m \, , \qquad \text{qui implique} \qquad W \star f = W \star F' = m \star m' \, ,$$

en utilisant là encore les propriétés de dérivation de la convolution. Finalement, $\operatorname{Var} N(t) = 2\, m \star m'(t) + m(t) - m^2(t)$ comme attendu. □

23.2.7 Etude non-asymptotique dans le cadre NBU

Discussion du cadre NBU

On veut se placer ici sous l'hypothèse NBU, qui assure que pour tous $s > 0$ et $t > 0$,

$$R(s+t) \leqslant R(s)\, R(t)\ ,$$

où l'on rappelle que R est la fonction fiabilité, $R = 1 - F$. Interprétons cette propriété lorsque $R(s) > 0$:

$$\begin{aligned} R(s+t) \leqslant R(s)\, R(t) \quad &\Longleftrightarrow \quad \mathbb{P}\{T_1 > s+t\} \leqslant \mathbb{P}\{T_1 > s\}\, \mathbb{P}\{T_1 > t\} \\ &\Longleftrightarrow \quad \mathbb{P}\big\{T_1 > s+t \,|\, T_1 > s\big\} \leqslant \mathbb{P}\{T_1 > t\}\ , \end{aligned}$$

ce qui signifie que lorsqu' un composant a fonctionné pendant la durée s, la probabilité qu'il fonctionne encore pendant une durée t est plus faible que la probabilité qu'il fonctionne pendant la durée t à partir de l'instant initial. La propriété NBU modélise donc le phénomène d'usure d'un composant, ce que l'on retrouve en explicitant le sigle NBU en anglais : « New Better than Used ». C'est en cela que la propriété NBU est naturelle pour les problèmes de fiabilité. Bien entendu, on peut définir la propriété NWU : « New Worse than Used », qui signifie mathématiquement que pour tout $s > 0$ et pour tout $t > 0$,

$$R(s+t) \geqslant R(s)\, R(t)\ .$$

La propriété NWU modélise le phénomène de rodage d'un composant. Il est bien entendu important de connaître des lois pour lesquelles les propriétés NBU et NWU sont vraies. Par exemple, on a rappelé au paragraphe 23.2.3 que la loi exponentielle est à la fois NBU et NWU.

D'autres exemples naturels de lois NBU ou NWU sont obtenus avec la loi de Weibull, discutée au paragraphe 23.2.4. Vu l'expression de la fonction de fiabilité de la loi $\mathcal{W}(a, \lambda)$, elle est NBU si et seulement si $(t+s)^a \geqslant t^a + s^a$ pour tous $t > 0$ et $s > 0$, soit pour $a \geqslant 1$. De même, si $a \leqslant 1$, la loi est NWU. (Et pour $a = 1$, on retrouve la loi exponentielle de paramètre λ.) En conclusion, les lois de Weibull modélisent naturellement les propriétés NBU et NWU, *via* le choix du paramètre a.

Construction de l'intervalle de prévision à partir du Théorème 13.6

Nous reprenons l'inégalité (23.4). On considère ici, vu le Théorème 13.6, le majorant commun $M'_{\mu,\sigma^2}(t) = t/\mu$. Alors, l'intervalle

$$\big[0,\, N_{\max}\big] \qquad \text{où} \qquad N_{\max} = \frac{t}{\mu} + \frac{1}{\sqrt{\alpha}} \sqrt{\frac{t}{\mu}} \tag{23.11}$$

est un intervalle de prévision de niveau $1 - \alpha$.

Cette valeur de $N_{\max}$ est équivalente à celle obtenue en (23.3) dans le cadre asymptotique ; mais ici, elle est valable pour tout t et ne dépend pas de σ^2. Le coût de ce résultat a été la formulation de l'hypothèse NBU (mais celle-ci est naturelle dans un grand nombre de problèmes de fiabilité).

Preuve du Théorème 13.6

Preuve On commence par établir la majoration de $\mathbb{P}\big\{B(t) > x\big\}$ lorsque $t > 0$ et $x > 0$. Dans le cade NBU, en appliquant d'abord la Proposition 23.3,

$$\begin{aligned}\mathbb{P}\{B(t) > x\} &= R(t+x) + \int_0^t R(t+x-u)\, m'(u)\, \mathrm{d}u \\ &\leqslant R(t)R(x) + R(x)\int_0^t R(t-u)\, m'(u)\, \mathrm{d}u = R(x)\,\mathbb{P}\big\{B(t) > 0\big\} = R(x)\,,\end{aligned}$$

soit, de manière équivalente,

$$\mathbb{P}\big\{B(t) > 0\big\} \geqslant F(x)\,. \tag{23.12}$$

On note que cette inégalité vaut en fait pour tout $x \in \mathbb{R}$.

On majore alors $m(t)$ en utilisant l'identité de Wald (Proposition 23.2) exprimée en fonction de $\mathbb{E}\big[B(t)\big]$,

$$\mu\big(m(t)+1\big) = t + \mathbb{E}\big[B(t)\big] = t + \int_0^{+\infty} \mathbb{P}\{B(t) > x\}\,\mathrm{d}x \leqslant t + \int_0^{+\infty} R(x)\,\mathrm{d}x = t + \mu\,.$$

On en tire qu'effectivement $m(t) \leqslant t/\mu$.

Le dernier point est la majoration de la variance. Grâce au Lemme 23.7 énoncé et prouvé ci-dessous,

$$m \star m'(t) = \int_0^t m'(u) m(t-u)\,\mathrm{d}u \leqslant \int_0^t m'(u)\big(m(t) - m(u)\big)\,\mathrm{d}u = \frac{m^2(t)}{2}\,,$$

de sorte que, *via* la Proposition 23.6, il vient que la variance de $N(t)$ est plus petite que son espérance : $\mathrm{Var}\, N(t) \leqslant m(t)$. □

Lemme 23.7 Dans le cadre NBU, la fonction de renouvellement m est sur-additive sur $\mathbb{R}_+$, i.e., pour tous $a \geqslant 0$ et $b \geqslant 0$,

$$m(a+b) \geqslant m(a) + m(b)\,.$$

Preuve Il suffit de montrer la propriété pour $a > 0$ et $b > 0$; notons auparavant que ce qui découle de la sous-additivité de m_0 (vu la définition de m_0 sur $\mathbb{R}_+^*$), elle-même prouvée au Lemme 23.5, est que pour tous $a \geqslant 0$ et $b \geqslant 0$,

$$m(a+b) \leqslant 1 + m(a) + m(b)\,,$$

ce n'est donc pas ce que l'on cherche à montrer ici. La sur-additivité de m va évidemment mettre en jeu l'hypothèse NBU, par l'utilisation de (23.12).

La preuve commence par l'emploi des mêmes techniques que précédemment, de découpage, conditionnement, échange de sommations et d'intégrations,

$$
\begin{aligned}
& m(a+b) - m(a) \\
&= \sum_{n=1}^{+\infty} \mathbb{P}\{a < S_n \leqslant a+b\} \\
&= \sum_{n=1}^{+\infty} \sum_{k=0}^{n-1} \mathbb{P}\big\{a < S_n \leqslant a+b \text{ et } N(a) = k\big\} \\
&= \sum_{n=1}^{+\infty} \mathbb{P}\big\{a < S_n \leqslant a+b \text{ et } N(a) = n-1\big\} \\
&\quad + \sum_{n=2}^{+\infty} \sum_{k=0}^{n-2} \mathbb{P}\big\{a < S_k + T_{k+1} + (T_{k+2} + \ldots + T_n) \leqslant a+b \text{ et } N(a) = k\big\} \\
&= \mathbb{P}\{B(a) \leqslant b\} + \sum_{n=2}^{+\infty} \sum_{k=0}^{n-2} \int \mathbb{P}\big\{B(a) \leqslant b-u \text{ et } N(a) = k\big\}\, f_{n-k-1}(u)\, \mathrm{d}u \\
&= \mathbb{P}\{B(a) \leqslant b\} + \sum_{\ell=1}^{+\infty} \int \sum_{n=\ell+1}^{+\infty} \mathbb{P}\big\{B(a) \leqslant b-u \text{ et } N(a) = n-\ell-1\big\}\, f_\ell(u)\, \mathrm{d}u \\
&= \mathbb{P}\{B(a) \leqslant b\} + \int \mathbb{P}\{B(a) \leqslant b-u\} \left(\sum_{\ell=1}^{+\infty} f_\ell(u) \right) \mathrm{d}u \\
&= \mathbb{P}\{B(a) \leqslant b\} + \int \mathbb{P}\{B(a) \leqslant b-u\}\, m'(u)\, \mathrm{d}u\,.
\end{aligned}
$$

On utilise maintenant (23.12) pour obtenir la minoration

$$m(a+b) - m(a) \geqslant F(b) + F \star m'(b) = F(b) + f \star m(b) = m(b)$$

où l'on a utilisé pour l'égalité centrale un théorème d'échange d'intégration et de limite, qui en termes de convolution assure que $F \star m' = (F \star m)' = F' \star m = f \star m$, et l'on a conclu par l'égalité (23.9). □

23.2.8 Simulations probabilistes

Ce paragraphe est consacré à des simulations probabilistes : on entend par là des simulations où nous choisissons la loi des durées inter-pannes et évaluons les procédures précédentes à l'aune de notre connaissance de cette loi.

Comparaison des longueurs des trois intervalles de prévision

On s'intéresse à la longueur et au niveau des intervalles de prévision, pris tous sous la forme $\big[0,\ N_{\max}\big]$. On vient de construire trois valeurs possibles pour $N_{\max}$,

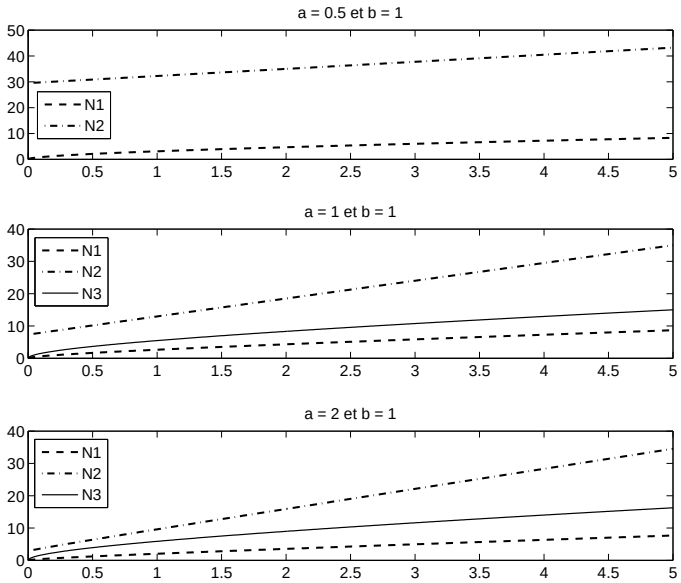

FIGURE 23.3. Tracé des fonctions $t \mapsto N_{\max}^{(1)}(t)$ et $t \mapsto N_{\max}^{(2)}(t)$, ainsi que, dans le cas NBU, de $t \mapsto N_{\max}^{(3)}(t)$, pour différents paramètres de la loi de Weibull $\mathcal{W}(a, b)$, sur l'intervalle $[0, 5]$, et pour $\alpha = 0.05$.

– dans le cadre asymptotique, en (23.3),

$$N_{\max} = N_{\max}^{(1)}(t) \overset{\text{not.}}{=} \frac{t}{\mu} + z_{1-\alpha} \frac{\sigma\sqrt{t}}{\mu^{3/2}} ,$$

– dans le cadre général, en (23.5),

$$N_{\max} = N_{\max}^{(2)}(t) \overset{\text{not.}}{=} M_{\mu,\sigma^2}(t) + \frac{1}{\sqrt{\alpha}} \sqrt{M_{\mu,\sigma^2}(t) + M_{\mu,\sigma^2}(t)^2}$$

$$\text{où} \qquad M_{\mu,\sigma^2}(t) \overset{\text{not.}}{=} \frac{t}{\mu} + \frac{\sigma^2}{\mu^2} ,$$

– dans le cadre NBU, en (23.11),

$$N_{\max} = N_{\max}^{(3)}(t) \overset{\text{not.}}{=} \frac{t}{\mu} + \frac{1}{\sqrt{\alpha}} \sqrt{\frac{t}{\mu}} .$$

On prend des durées de vie suivant une loi de Weibull, $T_i \sim \mathcal{W}(a, \lambda)$. On ne considère que les deux premiers intervalles de prévision si l'on n'est pas dans le cas NBU et on compare

$$N_{\max}^{(1)}(t) , \quad N_{\max}^{(3)}(t) \quad \text{et} \quad N_{\max}^{(2)}(t)$$

$(a,\ \lambda)$	$(0.5,\ 1)$	$(1,\ 1)$	$(2,\ 1)$
$\max_{t\in[0,10]} \mathbb{P}\left\{N(t) > N^{(2)}_{\max}(t)\right\}$	0	0	0
$\max_{t\in[0,10]} \mathbb{P}\left\{N(t) > N^{(3)}_{\max}(t)\right\}$	[non NBU]	0.008	0

TABLE 23.2. Approximations empiriques par méthode de Monte-Carlo de la probabilité maximale que $N(t)$ n'appartienne pas à ses intervalles de prévision sur [0, 10], pour un risque $\alpha = 0.05$.

à la figure 23.3, pour différentes valeurs de a et λ et pour $\alpha = 0.05$. Comme attendu, nous observons que pour tout t, on a le classement

$$N^{(1)}_{\max}(t) \leqslant N^{(3)}_{\max}(t) \leqslant N^{(2)}_{\max}(t)\ .$$

Cependant, comme nous l'avons souligné à la page 352, lorsque $t \geqslant \mu$, on peut améliorer la majoration du paragraphe 23.2.6 et redéfinir le deuxième intervalle de prévision en injectant la majoration de Var $N(t)$ donnée par l'inégalité (23.10) en lieu et place de $M_{\mu,\sigma^2}(t) + M_{\mu,\sigma^2}(t)^2$. On a alors à présent

$$N^{(2)}_{\max}(t) = \frac{t}{\mu} + \frac{\sigma^2}{\mu^2} + \mathbb{I}_{\{t<\mu\}} \frac{1}{\sqrt{\alpha}} \sqrt{\frac{t}{\mu} + \frac{\sigma^2}{\mu^2} + \left(\frac{t}{\mu} + \frac{\sigma^2}{\mu^2}\right)^2} + \mathbb{I}_{\{t\geqslant\mu\}} \frac{1}{\sqrt{\alpha}} \sqrt{\frac{t}{\mu}\left(\frac{4\sigma^2}{\mu^2} + 3\right) + \frac{\sigma^2}{\mu^2}\left(\frac{2\sigma^2}{\mu^2} + 1\right) - 1}\ .$$

Avec cette nouvelle valeur, le contrôle asymptotique de $N^{(2)}_{\max}(t)$,

$$N^{(2)}_{\max}(t) = \frac{t}{\mu} + O\left(\sqrt{t}\right)\ ,$$

est bien meilleur, comme l'atteste la figure 23.4. En particulier, il y a alors désormais équivalence lorsque $t \to \infty$ entre les bornes supérieures des trois intervalles de prévision.

Plus intéressante est l'évaluation (approchée, par simulation) des niveaux de ces intervalles. Tous ont été calculés de manière à ce qu'ils contiennent $N(t)$ avec une probabilité : proche de $1 - \alpha$ dans le cadre asymptotique, et toujours supérieure à $1 - \alpha$ dans les deux cadres non-asymptotiques. Il s'agit de vérifier ce point, ce que nous réalisons dans le tableau 23.2 et à la figure 23.5 pour différentes valeurs de a et λ et pour $\alpha = 0.05$. A cet effet, nous

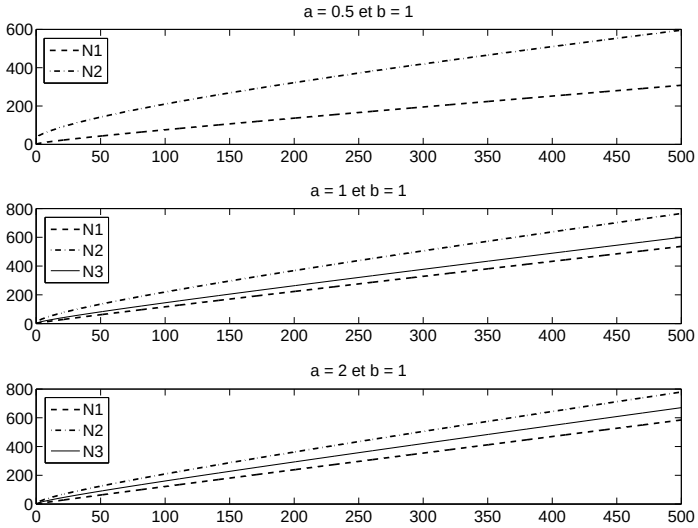

FIGURE 23.4. Tracé des fonctions $t \mapsto N_{\max}^{(1)}(t)$ et $t \mapsto N_{\max}^{(2)}(t)$ (modifiée), ainsi que, dans le cas NBU, de $t \mapsto N_{\max}^{(3)}(t)$, pour différents paramètres de la loi de Weibull $\mathcal{W}(a, b)$, sur l'intervalle [0, 500], et pour $\alpha = 0.05$.

calculons de manière approchée la probabilité $\mathbb{P}\{N(t) > N_{\max}\}$ pour toute valeur de t en utilisant la méthode de Monte-Carlo, décrite au paragraphe 9.7.1. Pour cela, on forme un 1 000–échantillon $\left(N^{(k)}(t)\right)_{1 \leqslant k \leqslant 1\,000}$ et on calcule la moyenne empirique

$$\frac{1}{1\,000} \sum_{k=1}^{1\,000} \mathbb{I}_{\{N^{(k)}(t) > N_{\max}\}} \approx \mathbb{P}\{N(t) > N_{\max}\} .$$

(La précision, que l'on peut déterminer selon les indications du paragraphe 9.7.1, est excellente, de l'ordre du millième, aussi reporte-t-on trois chiffres après la virgule au tableau 23.2.)

On conclut que lorsque $N_{\max} = N_{\max}^{(2)}(t)$ ou $N_{\max} = N_{\max}^{(3)}(t)$, alors cette probabilité est quasiment nulle et est donc loin de 0.05. Ces intervalles sont issus de procédures trop précautionneuses et sont bien plus grands que nécessaires (ils sont dit trop conservatifs).

En revanche, lorsque $N_{\max} = N_{\max}^{(1)}(t)$, la fonction $t \mapsto \mathbb{P}\{N(t) > N_{\max}\}$ oscille autour de 0.05. L'intervalle de prévision donné par $N_{\max}^{(1)}(t)$ n'est bien valable que dans un cadre asymptotique.

Notre dernier travail de simulation reprend la dernière suggestion du texte. Nous illustrons le fait que lorsque la loi des durées de vie est NBU, alors on a bien pour tout $t > 0$, le classement $\mathrm{Var}\, N(t) \leqslant m(t) \leqslant t/\mu$. Nous utilisons encore une fois la méthode de Monte-Carlo pour

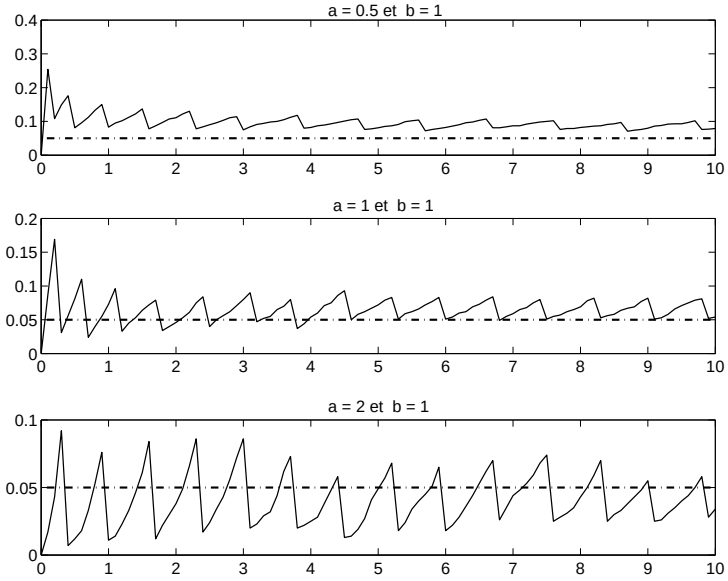

FIGURE 23.5. Tracé de l'approximation par méthode de Monte-Carlo de la fonction $t \mapsto \mathbb{P}\left\{N(t) > N_{\max}^{(1)}(t)\right\}$, pour différents paramètres de la loi de Weibull, sur l'intervalle [0, 10] et pour $\alpha = 0.05$.

obtenir cette fois les approximations suivantes de quantités inconnues par leurs équivalents empiriques,

$$m(t) = \mathbb{E}\big[N(t)\big] \approx \frac{1}{1\,000} \sum_{k=1}^{1\,000} N^{(k)}(t)$$

$$\text{et} \qquad \operatorname{Var} N(t) \approx \frac{1}{1\,000} \sum_{k=1}^{1\,000} \left(N^{(k)}(t)\right)^2 - \left(\frac{1}{1\,000} \sum_{k=1}^{1\,000} N^{(k)}(t)\right)^2 .$$

Nous considérons d'abord la loi de Weibull de paramètres $a = 2$ et $\lambda = 1$, qui est NBU. Les inégalités étudiées sont bien vérifiées, comme l'atteste le troisième tracé de la figure 23.6. Vient ensuite le cas $a = 1$ et $\lambda = 1$, celui de la loi exponentielle de paramètre 1. On a alors pour tout $t > 0$ les égalités $\operatorname{Var} N(t) = m(t) = t/\mu$, comme l'a rappelé le paragraphe 23.2.3 ; c'est, encore une fois, ce que l'on observe graphiquement, sur le tracé du milieu. Enfin, nous considérons le cas $a = 1/2$ et $\lambda = 1$, qui n'est pas NBU mais NWU. Le premier tracé de la figure 23.6 montre que pour ce troisième cas, l'inégalité précédente n'est pas vérifiée, et même qu'elle semble inversée : pour tout $t > 0$, on a le classement $\operatorname{Var} N(t) \geqslant m(t) \geqslant t/\mu$. On

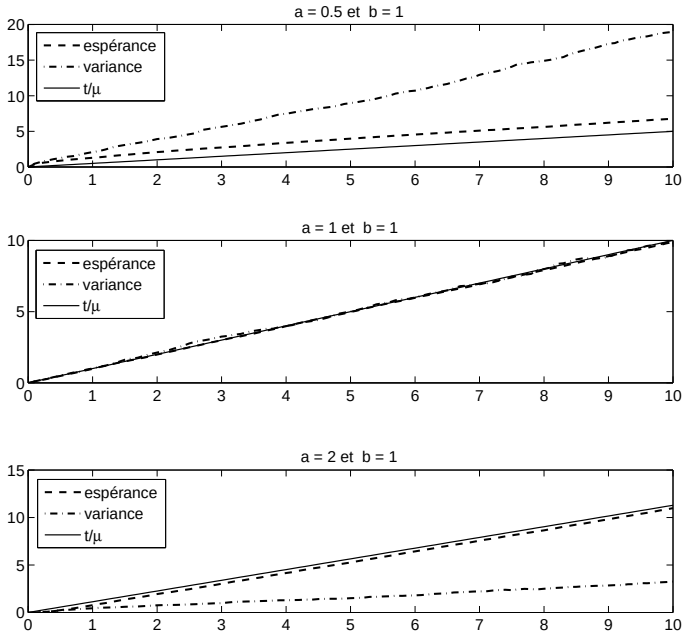

FIGURE 23.6. Tracé des fonctions $t \mapsto t/\mu$, $t \mapsto m(t)$ et $t \mapsto \mathrm{Var}\, N(t)$ (ou de leurs approximations par méthode de Monte-Carlo) pour différents paramètres de la loi de Weibull $\mathcal{W}(a, b)$.

peut en fait prouver que c'est bien le cas pour toute loi NWU, voir pour plus de détails [CT97, Proposition 6.30].

23.2.9 Application statistique

Estimation préalable des paramètres

Les intervalles de prévision rappelés au début du paragraphe 23.2.8 dépendent de μ et σ^2, qui sont généralement inconnus ; en réalité, ils ne sont donc pas encore réellement des intervalles de prévision. Nous traitons ci-dessous cette dépendance, par estimation préalable. On suppose en effet avoir recueilli dans le passé un jeu de données $D_1, D_2, \ldots, D_m$, correspondant à des observations antérieures du phénomène à contrôler. Elles nous servent à construire des intervalles de confiance (unilatères) sur μ et σ^2. On remplace alors μ et σ^2 par le majorant ou le minorant ainsi exhibé dans les expressions précédentes des intervalles de prévision. En effet, si on considère un intervalle de prévision, dépendant de μ et σ^2, de niveau $1 - \alpha/3$ et que chacun des intervalles de confiance sur les paramètres est de niveau $1 - \alpha/3$, alors le nouvel intervalle de prévision ainsi créé, qui ne dépend plus de μ et σ^2, est de niveau $1 - \alpha$.

Pour illustrer cela de manière plus concrète, nous nous restreignons au cas d'un modèle où les durées entre les pannes suivent une loi exponentielle de paramètre λ. On rappelle que cette dernière est d'espérance $1/\lambda$ et de variance $1/\lambda^2$, de sorte que λ est le seul paramètre à estimer. Nous construisons donc un intervalle de confiance pour λ de niveau $1-\alpha/2$, puis considérons un intervalle de prévision, dépendant de λ, de niveau $1-\alpha/2$. On applique le théorème de la limite centrale ordinaire, puis la méthode de stabilisation de la variance (voir le paragraphe 3.4.3), pour obtenir les convergences en loi

$$\sqrt{m}\left(\overline{D}_m - \frac{1}{\lambda}\right) \rightsquigarrow \mathcal{N}\left(0,\ \frac{1}{\lambda^2}\right) \qquad \text{puis} \qquad \sqrt{m}\left(\log \overline{D}_m - \log \frac{1}{\lambda}\right) \rightsquigarrow \mathcal{N}(0,1)\ .$$

On a noté ici, comme d'habitude, $\overline{D}_m$ la moyenne empirique de $D_1, \ldots, D_m$. On en déduit un intervalle de confiance de niveau asymptotiquement $1-\alpha/2$ pour λ,

$$\left[0,\ \widehat{\lambda}_m\right] \qquad \text{où} \quad \widehat{\lambda}_m = \exp\left(-\left(\log \overline{D}_m - \frac{z_{1-\alpha/2}}{\sqrt{m}}\right)\right). \tag{23.13}$$

Puis, par réinjection, les intervalles $\left[0,\ N^{(1)}_{\max}(t)\right]$ et $\left[0,\ N^{(3)}_{\max}(t)\right]$, où

$$N^{(1)}_{\max}(t) \overset{\text{not.}}{=} \widehat{\lambda}_m t + z_{1-\alpha/2}\sqrt{\widehat{\lambda}_m t} \qquad \text{et} \qquad N^{(3)}_{\max}(t) \overset{\text{not.}}{=} \widehat{\lambda}_m t + \sqrt{\frac{2}{\alpha}}\sqrt{\widehat{\lambda}_m t} \tag{23.14}$$

et qui sont issus respectivement de (23.3) et (23.11), sont de vrais intervalles de prévision pour un processus de nouvelles pannes $T_1, T_2, \ldots$ de même loi que l'échantillon dont on disposait. Le niveau de ces intervalles de prévision est asymptotiquement d'au moins $1-\alpha$. (On ne considère pas les intervalles du deuxième type, parce qu'on a vu qu'ils étaient plus gros que nécessaires dans le cas exponentiel, voir le paragraphe 23.2.8.)

Mise en œuvre de la méthode sur des données réelles

Notre objectif dans ce paragraphe est de mettre en application les résultats précédents lorsque nous disposons de données réelles. Pour cela, nous allons nous appuyer sur le jeu de données fourni par [Lyu96]. Ces données, au nombre de $n = 136$, représentent les durées de vie de composants identiques d'un système militaire (dans une unité de mesure qui n'est pas précisée).

Nous coupons ces données en deux sous-échantillons de taille égale, contenant donc chacun 68 observations. Le premier sera considéré comme formé de données antérieures et sera utilisé pour le test de l'ajustement exponentiel et l'estimation (la majoration) du paramètre. On pourra alors construire les intervalles de prévision, et étudier ensuite leur pertinence en les confrontant au second sous-échantillon (qu'on appelle sous-échantillon de validation).

Le découpage est effectué ici de manière aléatoire ; il apparaît en effet, à l'inspection des données, que les valeurs atypiques (ici, celles supérieures à 1 500) ne sont présentes que dans la seconde partie de l'échantillon. Afin de mieux répartir leur impact, on utilise donc un réarrangement aléatoire préliminaire. On note au passage que ces valeurs atypiques ont un impact important, puisque la moyenne des observations vaut 651 et la médiane, 286 seulement.

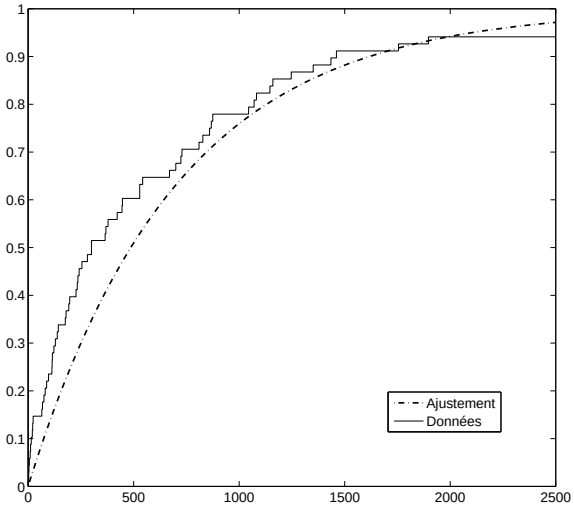

FIGURE 23.7. Ajustement graphique : tracés de la fonction de répartition empirique des 68 données et de la fonction de répartition de la loi exponentielle de paramètre l'inverse de la moyenne empirique des données.

Nous avons représenté à la figure 23.7 les données retenues pour le premier sous-échantillon servant à l'estimation. On a tracé leur fonction de répartition empirique contre celle de la loi exponentielle de paramètre l'inverse de leur moyenne empirique. La différence entre les deux courbes forme la statistique de test du caractère exponentiel mentionné au paragraphe 8.2.2. Nous ne l'avons pas mis en œuvre précisément, mais l'ajustement semble raisonnable, ou en tout cas, le serait si l'on enlevait quelques valeurs atypiques. (Nous ne le faisons pas dans la suite.)

Vient maintenant le temps de la validation. Nous retenons la valeur $\alpha = 0.10$ et traçons à la figure 23.8 les fonctions $N_{\max}^{(1)}$ et $N_{\max}^{(3)}$ obtenues par la réinjection d'estimées présentée en (23.14). On a considéré deux méthodes d'estimation possibles pour obtenir ces estimées :

- celle par intervalle, qui considère $\widehat{\lambda}_{68}$, la borne supérieure de l'intervalle de confiance de niveau $1 - \alpha/2$ donnée en (23.13) ;
- l'estimation ponctuelle, qui propose l'inverse de la moyenne empirique, $1/\overline{D}_{68}$.

L'estimation par intervalle semble garantir des résultats plus sûrs, comme cela est naturel ; l'injection du résultat d'une estimation ponctuelle, même consistante et sans biais, ne semble pas suffire, en particulier pour les petites valeurs de t.

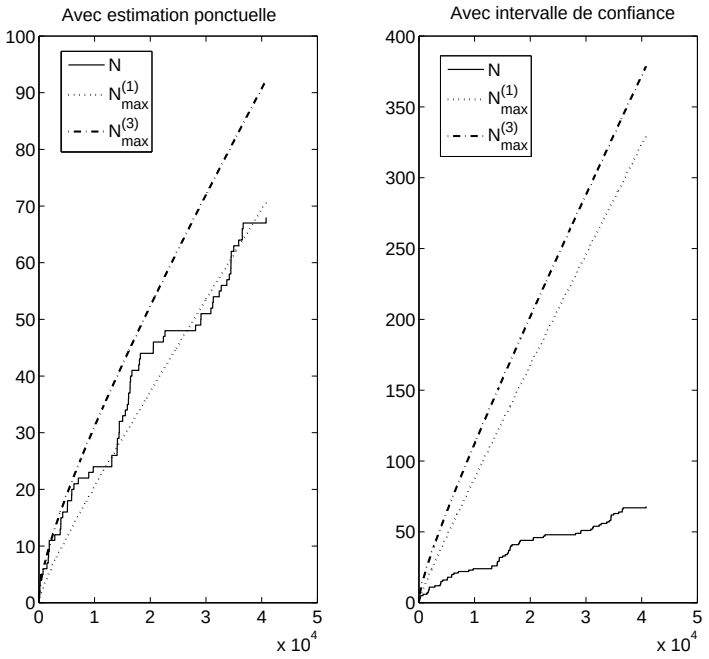

FIGURE 23.8. Fonctions $N_{\max}^{(1)}$ et $N_{\max}^{(3)}$ obtenues par réinjection d'estimées ponctuelles (à gauche) et par intervalle (à droite) et tracé du processus N de nombre de pannes.

Pour mesurer précisément la qualité de la procédure en deux temps décrite ci-dessus, on pourrait bien sûr calculer préalablement son niveau en générant des données artificielles, distribuées selon une loi exponentielle et en recourant à la méthode de Monte-Carlo.

On conclut ce paragraphe en soulignant que la fonction $N_{\max}^{(2)}$ aurait bien entendu pu être utilisée. Cette dernière a le mérite de fournir des intervalles de prévision robustes au sens où leur niveau ne dépend pas de la forme de la loi d'échantillon sous-jacente. Le lecteur pourra vérifier qu'elle est cependant, pour ce jeu de données, beaucoup trop conservative et propose des intervalles beaucoup plus grands que nécessaires.

23.2.10 Remarques de conclusion

En général, ce texte est assez bien traité par les étudiants qui maîtrisent les outils élémentaires de la théorie des probabilités (loi forte des grands nombres, théorème de la limite centrale). Certains tombent dans les pièges du paragraphe 13.2, à propos des résultats asymptotiques, en utilisant des arguments de composition de limites fallacieux. D'autres se laissent impressionner par la longueur du texte, qu'ils souhaitent traiter entièrement, au détriment d'une offre

conséquente d'illustrations informatiques et/ou d'une réflexion sur l'application statistique des résultats de probabilités. La plupart des étudiants ont l'idée d'utiliser la loi exponentielle pour modéliser les durées de vie des composants. Malheureusement, il semble que la loi de Weibull, qui généralise de façon naturelle la loi exponentielle, soit mal connue (malgré le nom de la fonction `Matlab` permettant de simuler des variables exponentielles : `rexpweib`). Elle intervient pourtant dans de nombreux problèmes de fiabilité (voir également le chapitre 12).

23.2.11 Bibliographie

Comme nous l'avons déjà souligné, ce texte de fiabilité présente des résultats extrêmement classiques, que l'on retrouve par exemple dans l'ouvrage écrit par Jean-Louis Bon [Bon95], une excellente référence du domaine. Notre lecteur y trouvera un prolongement naturel des résultats présentés ici, comme, par exemple, d'autres notions de vieillissement, et des problèmes de statistique connexes. En particulier, l'auteur expose les outils adaptés pour les modèles censurés (que nous étudions dans le chapitre 12), et notamment, les estimateurs de Kaplan–Meier et de Nelson–Aalen. Une autre excellente référence est l'ouvrage écrit par Christiane Cocozza-Thivent [CT97], qui prolonge les résultats de fiabilité de [Bon95] notamment par une analyse markovienne. Enfin, [CGCDM05, appendice au chapitre 3] expose de manière simple les résultats classiques sur les processus de Poisson.

23.3 Commentaires sur les simulations `Matlab`

23.3.1 Architecture générale des programmes

Pour réaliser les simulations de ce texte, nous faisons appel à trois fonctions et cinq scripts. Les deux premières fonctions, `weibull` et `Nmax`, permettent respectivement la construction d'un tableau de variables de Weibull indépendantes et identiquement distribuées et le calcul des intervalles de prévision. Notons qu'une fonction de la `StixBox` génère déjà des variables aléatoires distribuées selon la loi de Weibull, en la personne de `rexpweib`, mais il est très simple de la reprogrammer, et c'est ce que nous avons fait ici pour des raisons pédagogiques.

Les quatre premiers programmes maîtres sont appelés `CvPs`, `CvLoi`, `TraceNmax`, `IPetInegNBU`. `CvPs` illustre les résultats de convergence presque sûre, $N(t) \to +\infty$ et $N(t)/t \to 1/\mu$, tandis que `CvLoi` s'intéresse à la convergence en loi précisée dans la deuxième partie du Théorème 13.1. `TraceNmax` calcule les trois valeurs $N_{\max}$ proposées par le texte et les compare graphiquement. Le programme le plus important, `IPetInegNBU`, calcule d'une part, pour chacune des valeurs de $N_{\max}$, la probabilité que $N(t)$ n'appartienne pas à l'intervalle de prévision calculé (au risque 5 %), et d'autre part, compare graphiquement les trois quantités (Var $N(t)$, $m(t)$ et t/μ) classées par le Théorème 13.6, en fonction de l'existence d'une propriété NBU de la loi des durées de vie.

Enfin, s'agissant de l'étude des données réelles, la fonction `SYS1` les charge dans un ordre aléatoire tandis que le fichier maître `VraiesDonnees` les traite (en prenant la première moitié pour représenter l'ajustement à une loi exponentielle et construire des estimées et des intervalles de prévision, et la seconde, pour valider ou invalider ces intervalles).

23.3.2 Commentaires sur le code `Matlab`

On pourra se référer aux commentaires du chapitre 19 pour tout ce qui concerne le tracé de fonctions en escaliers par l'appel à la fonction `stairs`, comme, par exemple, $t \mapsto N(t)$ ou les fonctions de répartition empirique du paragraphe 23.2.5.

Ici, il est à noter que le calcul des variables $(S_j)_{1 \leqslant j \leqslant n}$ est réalisé automatiquement à l'aide de la fonction `cumsum` à partir du vecteur des variables $(T_j)_{1 \leqslant j \leqslant n}$ (il est inutile d'utiliser une boucle), et de même, les $N(t)$ se déduisent facilement des S_j en regardant combien d'entre eux sont plus petits que t et en les comptant, grâce à des lignes comme

```
N = sum(S <= t(k), 2);
```

(ligne extraite du script `CvLoi`).

23.3.3 Sorties numériques

```
Erreur maximale pour N2 et N3 (a = 0.5 et b = 1) : 0 et 0.131
Erreur maximale pour N2 et N3 (a = 1 et b = 1) : 0 et 0.008
Erreur maximale pour N2 et N3 (a = 2 et b = 1) : 0 et 0
```

(On a noté ici `b` le paramètre λ.) Lorsque $a < 1$, la loi de Weibull de paramètres a et $\lambda = 1$ n'est pas NBU ; c'est ce qui explique la mauvaise performance de l'intervalle de prévision construit sur $N_{\max}^{(3)}$ dans la première ligne des résultats, voir le paragraphe 23.2.8.

23.3.4 Codes des simulations

Ils sont disponibles en annexe, au chapitre 31, page 467.

CHAPITRE 24

Estimation de densité de probabilité

24.1 Corrigé possible du texte

24.1.1 Position du problème et théorème de Scheffé

On définit tout d'abord le modèle statistique considéré et on va montrer comment on est conduit à formuler une petite restriction sur l'ensemble des lois considérées.

On dispose d'un vecteur $X_1^n = (X_1, \ldots, X_n)$ d'observations à valeurs dans $\mathbb{R}$, indépendantes et identiquement distribuées selon une loi $\mu \in \mathcal{P}$ inconnue. On verra ci-dessous pourquoi l'ensemble $\mathcal{P}$ des lois considérées est l'ensemble des lois boréliennes ν sur $\mathbb{R}$ absolument continues par rapport à la mesure de Lebesgue. C'est cette requête d'absolue continuité qui forme notre restriction par rapport au modèle global qui considérerait toutes les lois boréliennes possibles. Cela impose en particulier que μ ne charge aucun point, mais cette conséquence n'est pas suffisante pour caractériser l'absolue continuité par rapport à la mesure de Lebesgue. $\mathcal{P}$ ne pouvant être paramétré par un espace de dimension finie, on a affaire ici à un modèle non-paramétrique.

Un estimateur de la loi inconnue μ est toute fonction $\widehat{\mu}_n$ mesurable [1] par rapport à X_1^n et à valeurs dans $\mathcal{P}$. Tout le problème est, comme toujours, de définir ce qu'est un bon estimateur. Il faut définir à cet effet une mesure de l'écart entre deux lois, à appliquer entre notre estimation $\widehat{\mu}_n$ et la loi sous-jacente μ. Pour cela, il est nécessaire de déterminer en premier lieu notre but. Selon la complexité de ce dernier, des méthodes d'estimation sophistiquées seront nécessaires, ou pas. Dans les cas les plus simples, l'estimateur donné par la mesure empirique

$$\mu_n = \frac{1}{n} \sum_{t=1}^{n} \delta_{X_t}$$

(où δ_x désigne la masse de Dirac en un point $x \in \mathbb{R}$) est en effet satisfaisant, comme nous le rappelons ci-dessous.

Ainsi, s'il ne s'agit par exemple que de déterminer $\mu(A)$ pour un événement d'intérêt A fixé, on l'estime par la méthode des moments avec

$$\frac{1}{n} \sum_{t=1}^{n} \mathbb{I}_{\{X_t \in A\}} = \mu_n(A) \, .$$

[1] Pour bien définir ce caractère mesurable, il faut préciser quelle tribu on met sur $\mathcal{P}$: comme il apparaîtra clairement ci-dessous, $\mathcal{P}$ peut être identifié à un sous-ensemble fermé de $\mathbb{L}^1(\lambda)$, qui est un espace topologique.

En fait, puisque l'on dispose alors d'observations indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre $\mu(A)$, on peut recourir aux techniques du chapitre 3 pour estimer ce paramètre par un intervalle de confiance symétrique autour de $\mu_n(A)$. (On y propose plusieurs intervalles possibles, asymptotiques ou non.)

De même, si l'objectif est de bien estimer simultanément les probabilités

$$\mu\big(]-\infty,\ x]\big) \qquad \text{pour tout } x \in \mathbb{R}\,,$$

alors la mesure empirique μ_n est encore un bon estimateur, comme l'énonce le théorème de Glivenko–Cantelli (Théorème 8.7).

Cependant, on exige davantage ici : une estimation simultanée et de qualité de la probabilité de tous les boréliens A. Comme on veut une proximité uniforme de tous les estimateurs $\widehat{\mu}_n(A)$ aux $\mu(A)$, on considère pour cela le critère dit de la distance en variation totale. Cette dernière est définie pour deux probabilités boréliennes ν_1 et ν_2 par

$$\mathrm{d_{VT}}(\nu_1, \nu_2) = \sup_{A \in \mathcal{B}(\mathbb{R})} \big|\nu_1(A) - \nu_2(A)\big|\,,$$

où $\mathcal{B}(\mathbb{R})$ est l'ensemble des boréliens de $\mathbb{R}$. On note que $\mathrm{d_{VT}}$ vérifie bien les trois axiomes de la définition d'une distance (symétrie, séparation, inégalité triangulaire) et qu'elle est toujours inférieure ou égale à 1.

Dans notre problème, il s'agit donc de garantir que l'estimateur $\widehat{\mu}_n$ proposé est tel que $\mathrm{d_{VT}}\big(\widehat{\mu}_n,\ \mu\big)$ est petit. Or, il est aisé de calculer la distance en variation totale entre la mesure empirique μ_n et μ : elle vaut 1. En effet, $A = \{X_1, \ldots, X_n\}$ est un borélien tel que $\mu_n(A) = 1$ mais $\mu(A) = 0$. La mesure empirique ne convient donc plus au vu de l'objectif plus ambitieux d'estimation simultanée et il faut recourir à des techniques spécifiques.

Pour rendre les choses plus agréables et éviter tout phénomène de mesures étrangères, comme ceux rencontrés avec μ_n, on se restreint d'une part au modèle $\mathcal{P}$ où toutes les probabilités sont dominées par la mesure de Lebesgue et on impose d'autre part que l'estimateur $\widehat{\mu}_n$ est à valeurs dans $\mathcal{P}$. En d'autres termes, on ne considérera que des lois $\widehat{\mu}_n$ admettant une densité $\widehat{f_n}$ par rapport à la mesure de Lebesgue. La donnée d'un élément μ de $\mathcal{P}$ étant équivalente à sa densité f par rapport à la mesure de Lebesgue, on ne parlera plus que de densités (de probabilité) désormais. En particulier, dans le reste du corrigé, on s'attachera à estimer f par un estimateur $\widehat{f_n}$. Au passage et à un petit abus de notation près (l'identification d'une loi à sa densité), on a désormais une description agréable du modèle, en termes d'un sous-ensemble fermé de $\mathbb{L}^1(\lambda)$,

$$\mathcal{P} = \left\{ f \in \mathbb{L}^1(\lambda) : \quad f \geqslant 0 \text{ et } \int_{\mathbb{R}} f \,\mathrm{d}\lambda = 1 \right\}\,.$$

Une première clé est que la distance en variation totale entre deux éléments ν_1 et ν_2 de $\mathcal{P}$ s'exprime en fonction de leurs densités f_1 et f_2. Ce résultat forme le théorème de Scheffé. On note ci-dessous $x_+ = \max\{x,\ 0\}$ et $x_- = \max\{-x,\ 0\}$ les parties positive et négative d'un réel x, comme on l'a déjà fait au chapitre 8.

Théorème 24.1 (de Scheffé) *Pour deux probabilités ν_1 et ν_2, de densités respectives f_1 et f_2 par rapport à la mesure de Lebesgue,*

$$\mathrm{d_{VT}}(\nu_1, \nu_2) = \int_{\mathbb{R}} (f_1 - f_2)_+ \, \mathrm{d}\lambda = \frac{1}{2} \int_{\mathbb{R}} |f_1 - f_2| \, \mathrm{d}\lambda \overset{\text{not.}}{=} \frac{1}{2} \|f_1 - f_2\|_1 \overset{\text{not.}}{=} \mathrm{d_{VT}}(f_1, f_2) \, .$$

Preuve Il ne s'agit que de prouver les deux premières égalités, les deux dernières ne faisant que définir des notations.

Pour la première égalité, on va montrer que le pire ensemble dans la définition de la distance en variation totale est donné par l'ensemble

$$B = \big\{ x \in \mathbb{R} \text{ tels que } f_1(x) \geqslant f_2(x) \big\} \overset{\text{not.}}{=} \{ f_1 \geqslant f_2 \} \, .$$

B est borélien, car f_1 et f_2 sont par définition des fonctions boréliennes. On a alors, pour tout borélien A,

$$\nu_1(A) - \nu_2(A) = \int_A (f_1 - f_2) \, \mathrm{d}\lambda \leqslant \int_{A \cap B} (f_1 - f_2) \, \mathrm{d}\lambda \leqslant \int_B (f_1 - f_2) \, \mathrm{d}\lambda = \int_{\mathbb{R}} (f_1 - f_2)_+ \, \mathrm{d}\lambda \, ,$$

la première inégalité venant du fait que $f_1 - f_2$ est négatif sur $A \cap B^c$, et la seconde, de ce que $f_1 - f_2$ étant positif sur B, son intégrale est plus grande sur un ensemble plus grand, en l'occurrence sur B tout entier, que sur $A \cap B$; l'égalité finale n'est qu'une réécriture de la définition de B. Il suffit de passer au supremum en $A \in \mathcal{B}(\mathbb{R})$ pour obtenir la première égalité du théorème.

La deuxième égalité est fondée sur l'observation que les densités s'intègrent chacune à 1, soit

$$\int_{\mathbb{R}} (f_1 - f_2) \, \mathrm{d}\lambda = 1 - 1 = 0$$

et utilise ensuite les écritures $x = x_+ - x_-$ et $|x| = x_+ + x_-$. Précisément, l'égalité des intégrales des parties positives et négatives,

$$\int_{\mathbb{R}} (f_1 - f_2) \, \mathrm{d}\lambda = 0 = \int_{\mathbb{R}} (f_1 - f_2)_+ \, \mathrm{d}\lambda - \int_{\mathbb{R}} (f_1 - f_2)_- \, \mathrm{d}\lambda \, ,$$

fournit le résultat désiré. □

24.1.2 Objectif : estimation consistante de f

Parmi toutes les qualités que peut avoir un estimateur (ou une suite d'estimateurs), nous ne nous intéresserons qu'à la consistance, *id est*, au fait qu'une suite d'estimateurs $\widehat{f_n}$ converge en probabilité ou non vers f en distance en variation totale : on souhaite obtenir des convergences de la forme

$$\mathrm{d_{VT}}\Big(\widehat{f_n}, \, f\Big) \xrightarrow{\mathbb{P}} 0 \, .$$

On pourrait proposer d'autres critères, comme la robustesse (être peu sensible à la vraie distribution μ en un sens à préciser) ou l'optimalité minimax (cette dernière étant discutée au paragraphe 32.1.4 en annexe).

On va proposer trois familles d'estimateurs consistants ; tous régularisent en un certain sens la mesure empirique μ_n pour la rendre absolument continue par rapport à la mesure de Lebesgue. Les estimateurs par histogrammes sont ceux qu'on a spontanément tendance à considérer pour une illustration informatique (ils sont donnés par les fonctions `hist` et `histo` de `Matlab`). Ils reposent, comme on le verra, sur une suite de partitions de $\mathbb{R}$. Les hypothèses à poser sur la régularité de f et sur cette suite de partitions pour en tirer des estimateurs consistants ne sont pas données dans le texte du chapitre 14, mais on peut en proposer facilement un jeu en s'inspirant du traitement de l'estimateur par fenêtres glissantes. C'est pourquoi l'honneur du traitement mathématique revient à ce dernier.

24.1.3 Estimateur par fenêtres glissantes, preuve de la consistance

La définition de cet estimateur repose sur le fait que la fonction de répartition F de μ étant l'intégrale de la fonction $f \in \mathbb{L}^1(\lambda)$, elle est dérivable presque partout, de dérivée f. Cela procède en effet du théorème fondamental du calcul, rappelé au chapitre 26. Or le théorème de Glivenko-Cantelli (Théorème 8.7) assure que F est bien estimée par la fonction de répartition empirique F_n, où, pour tout $x \in \mathbb{R}$,

$$F_n(x) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}_{\{X_t \leqslant x\}} \, .$$

C'est ce qui motive la définition de $\widehat{f_n}$, où l'on fait une approximation symétrique de la dérivée,

$$\widehat{f_n} = \frac{F_n(\cdot + h_n) - F_n(\cdot - h_n)}{2h_n}$$

et non pas simplement $F_n(\cdot + h_n)/h_n$ par exemple. Pour que l'heuristique soit justifiée, il faut clairement $h_n \to 0$. Le texte indique qu'il est préférable que cette convergence ne soit pas trop rapide : $h_n \to 0$ tout en ayant $nh_n \to \infty$. Intuitivement, cette dernière condition assure qu'il y a un nombre suffisant d'observations X_t dans chaque intervalle $]x - h_n,\ x + h_n]$.

Avant de s'intéresser aux propriétés de $\widehat{f_n}$, on va prouver qu'il prend bien ses valeurs dans la classe des densités de probabilité (on en aura besoin ci-dessous au cours d'une preuve).

Lemme 24.2 Pour tout $h_n > 0$, l'estimateur $\widehat{f_n}$ est une densité de probabilité : $\widehat{f_n} \in \mathcal{P}$.

Preuve $\widehat{f_n}$ est bien une fonction à valeurs positives et

$$\begin{aligned} \int_{\mathbb{R}} \widehat{f_n} \, \mathrm{d}\lambda &= \int_{\mathbb{R}} \frac{F_n(x + h_n) - F_n(x - h_n)}{2h_n} \, \mathrm{d}x \\ &= \frac{1}{2nh_n} \sum_{t=1}^{n} \int_{\mathbb{R}} \mathbb{I}_{\{X_t \in \,]x - h_n,\, x + h_n]\}} \, \mathrm{d}x \end{aligned}$$

$$= \frac{1}{2nh_n} \sum_{t=1}^{n} \lambda\big([X_t - h_n,\ X_t + h_n[\,\big) = \frac{1}{2nh_n}\ 2nh_n = 1\ ,$$

ce qui montre qu'elle s'intègre bien à 1 et conclut la preuve. □

On prouve maintenant le Théorème 14.1 en suivant le schéma de preuve que l'on avait indiqué au chapitre 14. On commence notamment par la décomposition du risque de l'estimateur $\widehat{f_n}(x)$ de $f(x)$ selon

$$\mathbb{E}\left[\left(\widehat{f_n}(x) - f(x)\right)^2\right] = \left(\mathbb{E}\left[\widehat{f_n}(x)\right] - f(x)\right)^2 + \operatorname{Var}\widehat{f_n}(x)\ ,$$

dont on rappelle qu'elle se nomme la décomposition biais–variance et qu'elle a été vue au paragraphe 2.5.1. Le premier terme de la décomposition est appelé terme de biais parce qu'il mesure l'écart d'approximation de la méthode d'estimation : cette dernière se donne un objectif d'estimation plus simple à réaliser que $f(x)$. Le second terme, le terme de variance, correspond à la difficulté d'estimer cet objectif simplifié.

On continue par le résultat préliminaire suivant, que l'on utilisera deux fois par la suite : pour presque tout x,

$$\mathbb{E}\big[\widehat{f_n}(x)\big] = \frac{F(x+h_n) - F(x-h_n)}{2h_n} \longrightarrow f(x) \qquad \text{lorsque} \quad h_n \to 0. \tag{24.1}$$

En effet, ainsi que rappelé au début de ce paragraphe, F est dérivable presque partout, de dérivée f.

On a donc déjà que pour presque tout x, le terme de biais tend vers 0 dès lors que $h_n \to 0$. Quant au terme de variance, on a déjà utilisé plus haut l'égalité

$$2nh_n\,\widehat{f_n}(x) = \sum_{t=1}^{n} \mathbb{I}_{\{X_t \in\,]x-h,\, x+h]\}}\ ,$$

qui montre que l'on a affaire à une loi binomiale $\mathrm{Bin}(n,\ p_{n,x})$ de paramètres n et

$$p_{n,x} = \mu\big(]x - h_n,\ x + h_n]\big) = F(x+h_n) - F(x-h_n)\ .$$

La variance de cette loi est $np_{n,x}(1 - p_{n,x})$. On a donc pour presque tout x,

$$\operatorname{Var}\widehat{f_n}(x) = \frac{1}{4n^2h_n^2}\ \operatorname{Var}\big(2nh_n\,\widehat{f_n}(x)\big) = \frac{np_{n,x}(1-p_{n,x})}{4n^2h_n^2} \leqslant \frac{1}{4nh_n}\,\frac{p_{n,x}}{h_n} \longrightarrow 0$$

sous l'hypothèse $nh_n \to \infty$ pour le premier terme et par (24.1) pour le second terme.

On a ainsi prouvé que pour presque tout $x \in \mathbb{R}$,

$$\mathbb{E}\left[\left(\widehat{f_n}(x) - f(x)\right)^2\right] \longrightarrow 0\ .$$

La convergence dans $\mathbb{L}^2(\mathbb{P})$ entraînant celle en $\mathbb{P}$–probabilité, on aboutit à la proposition suivante. On notera que vu la preuve, elle vaut en fait dès que $nh_n \to \infty$ et $p_{n,x}/(2h_n) \to f(x)$ pour presque tout x lorsque $h_n \to 0$ et $n \to \infty$. (Nous aurons besoin de ce fait au paragraphe 24.1.5.)

Proposition 24.3 Lorsque la suite (h_n) est telle que $h_n \to 0$ et $nh_n \to \infty$, on a pour l'estimateur par fenêtres glissantes défini ci-dessus que pour presque tout x,

$$\widehat{f}_n(x) \xrightarrow{\mathbb{L}^2(\mathbb{P})} f(x) \qquad \text{d'où} \qquad \widehat{f}_n(x) \xrightarrow{\mathbb{P}} f(x)\,.$$

La seconde partie du Théorème 14.1 énonce que

$$\mathrm{d_{VT}}\Big(\widehat{f}_n,\, f\Big) = \int_{\mathbb{R}} \Big(f - \widehat{f}_n\Big)_+ \mathrm{d}\lambda \xrightarrow{\mathbb{P}} 0\,,$$

où l'égalité procède du théorème de Scheffé conjointement au fait, essentiel, que f et les $\widehat{f}_n$ sont des densités de probabilité (voir Lemme 24.2). Cette convergence en probabilité découle par exemple de la convergence

$$\mathbb{E}\bigg[\int_{\mathbb{R}} \Big(f - \widehat{f}_n\Big)_+ \mathrm{d}\lambda\bigg] \longrightarrow 0\,.$$

(En réalité, les deux assertions de convergence sont ici équivalentes au vu du caractère positif et borné de la distance en variation totale, comme le notera la preuve du Théorème 32.2.)

Nous prouvons la seconde assertion : premièrement,

$$\varphi_n(x) \overset{\text{not.}}{=} \mathbb{E}\Big[\Big(f(x) - \widehat{f}_n(x)\Big)_+\Big] \leqslant \sqrt{\mathbb{E}\Big[\Big(f(x) - \widehat{f}_n(x)\Big)^2\Big]} \longrightarrow 0$$

pour presque tout $x \in \mathbb{R}$, où nous avons appliqué successivement l'inégalité de Cauchy–Schwarz puis la convergence dans $\mathbb{L}^2(\mathbb{P})$ garantie par la Proposition 24.3. Or, on a par définition la domination $\varphi_n \leqslant f$. Le théorème de Fubini–Tonelli et le théorème de convergence dominée de Lebesgue assurent donc la convergence désirée,

$$\mathbb{E}\bigg[\int_{\mathbb{R}} \Big|\widehat{f}_n - f\Big| \mathrm{d}\lambda\bigg] = 2\,\mathbb{E}\bigg[\int_{\mathbb{R}} \Big(f - \widehat{f}_n\Big)_+ \mathrm{d}\lambda\bigg] = \int_{\mathbb{R}} \varphi_n \,\mathrm{d}\lambda \longrightarrow 0\,.$$

Cela conclut la preuve du Théorème 14.1.

En fait, la technique de preuve que nous venons d'employer permet de prouver une généralisation du lemme de Scheffé aux densités de probabilité, appelé théorème de Glick. Nous l'énonçons et le prouvons en annexe (au paragraphe 32.1.1).

24.1.4 Estimateur par fenêtres glissantes, normalité asymptotique

On continue l'étude des estimateurs par fenêtres glissantes avec un résultat de normalité asymptotique ponctuelle, suivi de sa conséquence : la construction d'intervalles de confiance ponctuels sur les $f(x)$, pour tout $x \in \mathbb{R}$, sous une hypothèse de régularité de f. On quantifie ainsi l'incertitude que l'on a de l'estimation de $f(x)$ par $\widehat{f_n}(x)$. On va voir que la taille de ces intervalles est proportionnelle à $1/\sqrt{nh_n}$ alors que dans les cas habituels (paramétriques), la décroissance est plus forte, en $1/\sqrt{n}$ (voir les résultats du chapitre 3). Cela souligne le caractère plus difficile du problème (non-paramétrique) que nous considérons.

On utilise comme outils principaux ici le résultat de convergence suivant, ainsi que plusieurs applications du lemme de Slutzky. Lorsque $np_n \to \lambda$, on a la convergence $S_n \rightsquigarrow \mathcal{P}(\lambda)$, qu'on appelle l'approximation binomiale–Poisson (on en a présenté un cas particulier à l'exemple 9.13, voir le Lemme 25.4 pour le résultat général). Le lemme suivant montre que lorsque $np_n \to \infty$, on dispose en un certain sens d'une approximation binomiale–normale.

Lemme 24.4 Si (S_n) est une suite de variables aléatoires de lois binomiales de paramètres respectifs n et p_n, alors, lorsque

$$np_n \to \infty \qquad \text{et} \qquad \limsup p_n < 1 \,,$$

on a la convergence en loi

$$S_n^* \overset{\text{not.}}{=} \frac{S_n - np_n}{\sqrt{np_n(1-p_n)}} \rightsquigarrow \mathcal{N}(0,1) \,.$$

Preuve On utilise la caractérisation de la convergence en loi par convergence simple des fonctions caractéristiques. On fixe à cet effet $t \in \mathbb{R}$ et pour raccourcir les écritures, on pose $t_n = t/\sqrt{np_n(1-p_n)}$. On note tout d'abord que

$$p_n t_n^2 = \frac{t^2}{n(1-p_n)} = O\left(\frac{1}{n}\right) \qquad \text{et donc} \qquad p_n^2 t_n^2 = O\left(\frac{1}{n}\right) \,.$$

On a utilisé ici de manière cruciale l'hypothèse que les p_n sont asymptotiquement plus petits qu'une constante strictement plus petite que 1. On écrit alors deux développements limités et on les multiplie ; le premier procède du fait que $p_n t_n \to 0$ et le second de ce que $t_n \to 0$ (et c'est là que l'on utilise l'hypothèse $np_n \to \infty$) :

$$\begin{aligned}
\mathbb{E}\left[e^{\mathrm{i}tS_n^*}\right] &= e^{-\mathrm{i}t_n np_n}\,\mathbb{E}\left[e^{\mathrm{i}t_n S_n}\right] = \left(e^{-\mathrm{i}t_n p_n}\left(p_n e^{\mathrm{i}t_n} + (1-p_n)\right)\right)^n \\
&= \left(1 - \mathrm{i}p_n t_n - \frac{1}{2}\,p_n^2 t_n^2 + o\left(\frac{1}{n}\right)\right)^n \left(1 + \mathrm{i}p_n t_n - \frac{1}{2}\,p_n t_n^2 + o\left(\frac{1}{n}\right)\right)^n \\
&= \left(1 - \frac{1}{2}\,p_n(1-p_n)\,t_n^2 + o\left(\frac{1}{n}\right)\right)^n \qquad \longrightarrow e^{-t^2/2} \,,
\end{aligned}$$

où pour la convergence finale on a notamment utilisé que $p_n(1-p_n)t_n^2 = t^2/n$. Cela conclut la preuve, vu la forme des fonctions caractéristiques pour la loi normale (voir le Théorème 5.2). □

Remarque au passage. Le résultat du Lemme 24.4 énonce un résultat de normalité asymptotique du type de celui donné par le théorème de la limite centrale. En fait, c'est un cas particulier du théorème de Lindeberg–Feller, qui généralise le théorème de la limite centrale en s'affranchissant de l'hypothèse d'identique distribution. On en énonce une version simplifiée : si pour tout $n \geqslant 1$, les $(X_{n,j})_{1 \leqslant j \leqslant n}$ sont des suites finies de n variables aléatoires indépendantes et bornées uniformément en n et j, et si, notant $S_n = X_{n,1} + \ldots + X_{n,n}$, on a

$$\sigma_n^2 = \operatorname{Var} S_n = \operatorname{Var} X_{n,1} + \operatorname{Var} X_{n,2} + \ldots + \operatorname{Var} X_{n,n} \longrightarrow \infty ,$$

alors

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\sigma_n^2}} \rightsquigarrow \mathcal{N}(0,1) .$$

Voir par exemple [vdV98, Proposition 2.27] pour un énoncé général.

Proposition 24.5 On suppose que la suite (h_n) est telle que $nh_n^3 \to 0$ et $nh_n \to \infty$. En un point $x \in \mathbb{R}$ tel que $f(x) > 0$ et f est dérivable sur un voisinage de x avec une dérivée bornée, on a le résultat suivant de normalité asymptotique pour l'estimateur par fenêtres glissantes défini ci-dessus :

$$\sqrt{2nh_n}\left(\frac{\widehat{f_n}(x) - f(x)}{\sqrt{\widehat{f_n}(x)}}\right) \rightsquigarrow \mathcal{N}(0,1) .$$

Preuve On fixe dans la suite un réel x vérifiant les conditions proposées. En particulier, comme f est continue en x, la convergence (24.1) et donc le résultat de la Proposition 24.3 valent en ce point. Avec les notations précédentes, on a l'équivalence $p_{n,x} \sim 2h_n f(x)$. En particulier, $np_{n,x} \to \infty$ vu que $nh_n \to \infty$. Par ailleurs, comme $nh_n^3 \to 0$, on a $h_n \to 0$ et donc $p_{n,x} \to 0$. On note $D_x \in \mathbb{R}_+$ un majorant de f' sur le voisinage V_x considéré. Les énoncés suivants étant tous asymptotiques et ayant $h_n \to 0$, on suppose sans perte de généralité que les h_n–voisinages de x sont inclus dans V_x.

L'application du Lemme 24.4 à $S_n = 2nh_n \widehat{f_n}(x)$ conduit alors à la convergence

$$\frac{2nh_n\left(\widehat{f_n}(x) - \mathbb{E}\left[\widehat{f_n}(x)\right]\right)}{\sqrt{np_{n,x}(1-p_{n,x})}} \rightsquigarrow \mathcal{N}(0,1) .$$

Par le lemme de Slutzky, au vu des convergences déterministes $2h_n f(x)/p_{n,x} \to 1$ et $1 - p_{n,x} \to 1$, il vient alors

$$\sqrt{2nh_n}\,\frac{\left(\widehat{f_n}(x) - \mathbb{E}\left[\widehat{f_n}(x)\right]\right)}{\sqrt{f(x)}} \rightsquigarrow \mathcal{N}(0,1) . \tag{24.2}$$

On essaie maintenant de remplacer l'espérance par $f(x)$ dans la formule précédente. L'inégalité des accroissements finis permet la majoration

$$\left| \mathbb{E}\Big[\widehat{f_n}(x)\Big] - f(x) \right| = \left| f(x) - \frac{1}{2h_n} \int_{x-h_n}^{x+h_n} f(t)\,\mathrm{d}t \right| \leqslant \frac{D_x}{2h_n} \int_{x-h_n}^{x+h_n} |x-t|\,\mathrm{d}t = \frac{D_x\, h_n}{2}\,.$$

On a donc en particulier la convergence déterministe

$$\sqrt{2nh_n}\, \frac{\left| \mathbb{E}\Big[\widehat{f_n}(x)\Big] - f(x) \right|}{\sqrt{f(x)}} \leqslant \frac{D_x}{\sqrt{2f(x)}} \sqrt{nh_n^3} \longrightarrow 0\,.$$

De cette convergence déterministe et de (24.2) découle, à nouveau par le lemme de Slutzky, la convergence en loi

$$\sqrt{2nh_n} \left(\frac{\widehat{f_n}(x) - f(x)}{\sqrt{f(x)}} \right) \rightsquigarrow \mathcal{N}(0,1)\,.$$

La Proposition 24.3 garantit la convergence $\widehat{f_n}(x) \to f(x)$ en $\mathbb{P}$–probabilité. Une troisième application du lemme de Slutzky conclut donc au résultat annoncé. □

Pour tout x et (h_n) en vérifiant ses hypothèses, la Proposition 24.5 conduit à des intervalles de confiance asymptotiques sur $f(x)$, par exemple de la forme

$$I_{1-\alpha}\big(x,\ X_1^n\big) = \left[\widehat{f_n}(x) - z_{1-\alpha/2}\sqrt{\frac{\widehat{f_n}(x)}{2nh_n}},\ \ \widehat{f_n}(x) + z_{1-\alpha/2}\sqrt{\frac{\widehat{f_n}(x)}{2nh_n}} \right]\,,$$

de niveau asymptotique $1-\alpha$ par choix du quantile $z_{1-\alpha/2}$ de la loi normale standard d'ordre $1-\alpha/2$.

On note qu'ici, on a affaire à une famille d'intervalles de confiance ponctuels : chacun des intervalles $I_{1-\alpha}\big(x,\ X_1^n\big)$ contient, avec probabilité asymptotiquement au moins $1-\alpha$, la valeur $f(x)$ sous les conditions précédemment énoncées. Mais on n'a pas de bande de confiance, *id est*, d'uniformité en x, contrairement au cas de l'estimation des fonctions de répartition : avec les notations du chapitre 8, F étant ici continue, on a qu'avec probabilité exactement $1-\alpha$,

$$\forall\, x \in \mathbb{R}, \qquad F(x) \in I'_{1-\alpha}(x,\ X_1^n) \overset{\text{not.}}{=} \big[F_n(x) - \xi_{n,1-\alpha},\ F_n(x) - \xi_{n,1-\alpha}\big]\,.$$

Les quantificateurs sont inversés dans le cas des intervalles de confiance pour les densités de probabilité : pour tout x vérifiant les conditions de la Proposition 24.5, on a, avec probabilité tendant vers $1-\alpha$, que $f(x) \in I_{1-\alpha}\big(x, X_1^n\big)$.

> Remarque au passage. Les intervalles de confiance sont utilisés pour exhiber de manière visuelle et de manière facilement calculable numériquement, un ensemble de candidats possibles pour le paramètre à estimer. Par exemple, un intervalle de confiance en norme

du supremum ou un ensemble (que l'on devrait prendre fini) d'intervalles de confiance ponctuels en des points $x_1, \ldots, x_N$ se représente facilement. Ici, en revanche, les régions de confiance que l'on tirerait de résultats de normalité asymptotique pour la distance en variation totale (le mieux que l'on puisse espérer ici en termes d'uniformité) seraient des boules en norme $\mathbb{L}^1(\lambda)$; ces dernières n'admettent pas de représentation graphique simple. On note également que les hypothèses de régularité formulées à la Proposition 24.5 sont insatisfaisantes au sens où il est difficile de vérifier (tester) si la densité sous-jacente f les remplit.

Dans nos simulations, malgré l'absence de garanties sur le fait qu'ils puissent être vrais simultanément, on a quand même tracé les intervalles de confiance ponctuels en tout point x de l'intervalle d'étude.

On a considéré ici une loi ν_u donnée par un mélange de lois uniformes, de densité par rapport à la mesure de Lebesgue égale à

$$f_u = \sum_{j=1}^{3} p_j f_{a_j,b_j} , \qquad \text{où} \quad f_{a,b} = \frac{1}{b-a} \mathbb{I}_{[a,b]}$$

avec $(p_1, p_2, p_3) = (0.25,\ 0.45,\ 0.30)$ et

$$[a_1, b_1] = [-1,\ 1] , \qquad [a_2, b_2] = [0,\ 1] , \qquad [a_3, b_3] = [2.5,\ 3.5] .$$

On trace à la figure 24.1 les intervalles de confiance obtenus avec un échantillon de taille $n = 800$ et les paramètres de bande $h \in \{0.05,\ 0.15,\ 0.25,\ 0.35\}$. Pour comparaison, des paramètres de bande choisis de manière automatique par la théorie seraient (avec le choix d'une constante multiplicative égale à 1 dans chaque cas)

$$h_{800} = 1/\sqrt{800} \approx 0.04 , \qquad h'_{800} = 800^{-1/3} \approx 0.11 \qquad \text{et} \qquad h''_{800} = 800^{-1/5} \approx 0.26 .$$

On justifierait ces valeurs en notant que le choix $h_n = 1/\sqrt{n}$ vérifie les conditions de la Proposition 24.5 tandis que le choix de $h'_n = n^{-1/3}$ apparaît comme cas limite de la condition $nh_n^3 \to 0$ de la Proposition 24.5. Enfin, l'intérêt de prendre un paramètre de bande de l'ordre de $h''_n = n^{-1/5}$ est discuté à la fin du paragraphe 24.1.7.

La densité f_u vérifie les hypothèses de la Proposition 24.5 sur l'union d'intervalles ouverts suivante $]-1, 0[\cup]0, 1[\cup]1, 2.5[\cup]2.5, 3.5[$. En ces points, l'estimation semble d'autant plus raisonnable qu'ils sont éloignés des points -1, 0, 1, 2.5 et 3.5, où f_u n'est pas dérivable. Cependant, la taille du voisinage de ces points où l'estimation n'est pas satisfaisante dépend fortement du paramètre de bande h. On sent dès à présent que le choix d'une valeur non asymptotique est crucial. Ce choix est appelé la calibration du paramètre de bande et nous y reviendrons au paragraphe 24.1.7.

24.1.5 Retour rapide sur les estimateurs par histogrammes

On peut maintenant revenir sur les estimateurs par histogrammes. Il reposent sur la considération d'une suite de partitions de $\mathbb{R}$ en un nombre dénombrable de sous-ensembles et on note

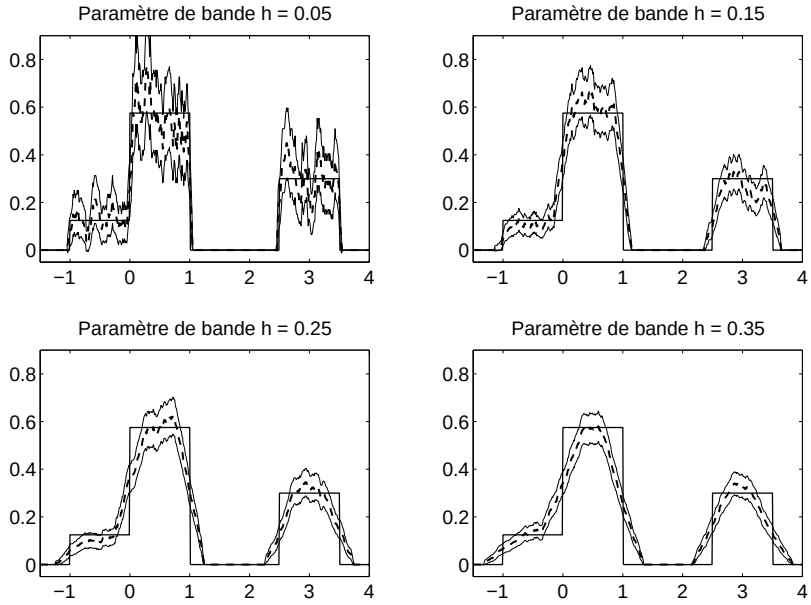

FIGURE 24.1. Estimation par fenêtres glissantes : $n = 800$ et différents paramètres de bande h.

$(A_{n,k})_{k\in\mathbb{Z}}$ la partition considérée au rang $n \geqslant 1$. A partir de cette partition et en utilisant les notations précédentes, on définit la fonction étagée $\widehat{f_n}$ selon

$$\forall k \in \mathbb{Z}, \quad \forall x \in A_{n,k}, \qquad \widehat{f_n}(x) = \frac{\mu_n(A_{n,k})}{\lambda(A_{n,k})} \,. \tag{24.3}$$

Notons que si les partitions $(A_{n,k})_{k\geqslant 1}$ ne dépendaient pas de n et étaient prises égales à chaque rang à une partition $A = (A_k)_{k\in\mathbb{Z}}$ fixée, alors, par la loi forte des grands nombres, la suite d'estimateurs convergerait vers la densité constante par morceaux f_A définie par (24.3) en remplaçant μ_n par μ. Cette densité f_A est généralement différente de f et d'autre part, dépend fortement de la partition choisie A. Pour éliminer cette dépendance et converger vers f, il faut considérer des partitions s'adaptant au nombre d'observations.

On restreint essentiellement notre attention aux partitions régulières, même si une remarque ci-dessous explique la liberté que l'on pourrait avoir sur le choix de leur forme. Une suite de telles partitions est paramétrée par une suite de pas (h_n) : au rang n, les éléments de la partition sont les $A_{n,k} =]kh_n,\ (k+1)h_n]$, où k décrit les entiers naturels $\mathbb{Z}$. Pour tout $x \in \mathbb{R}$

et $n \geqslant 1$, on note $k_{n,x}$ l'entier k tel que $x \in A_{n,k}$. On a alors

$$\widehat{f_n}(x) = \frac{\mu_n\big(A_{n,k_{n,x}}\big)}{h_n}\,.$$

On retrouve la même heuristique que celle discutée au début du paragraphe 24.1.3, à ceci près que cette fois-ci la fenêtre est fixe et non plus glissante.

Il s'agit de prouver dans un premier temps que l'analogue suivant de la convergence (24.1) est vérifié, pour presque tout $x \in \mathbb{R}$:

$$\mathbb{E}\big[\widehat{f_n}(x)\big] = \frac{F\Big(\big(k_{n,x}+1\big)h_n\Big) - F\big(k_{n,x}h_n\big)}{h_n} \longrightarrow f(x) \qquad \text{lorsque} \quad h_n \to 0. \tag{24.4}$$

Ici, on ne peut pas se contenter d'appliquer le théorème fondamental du calcul mais on va utiliser un résultat préliminaire à ce dernier : le fait que pour une fonction $f \in \mathbb{L}^1(\lambda)$, presque tout point $x \in \mathbb{R}$ est un point de Lebesgue de f (voir [Rud98, paragraphes 7.6 et 7.7]). On rappelle que x est un point de Lebesgue si, par définition,

$$\lim_{r\to 0}\ \frac{1}{r}\int_{x-r}^{x+r} \big|f(y) - f(x)\big|\,\mathrm{d}y = 0\,.$$

On note alors que

$$A_{n,k_{n,x}} = \ \big]k_{n,x}h_n,\ \big(k_{n,x}+1\big)h_n\big] \subseteq [x-h_n,\ x+h_n]$$

et que donc

$$\begin{aligned}\left|f(x) - \frac{F\Big(\big(k_{n,x}+1\big)h_n\Big) - F\big(k_{n,x}h_n\big)}{h_n}\right| &\leqslant \frac{1}{h_n}\int_{k_{n,x}h_n}^{(k_{n,x}+1)h_n} \big|f(y) - f(x)\big|\,\mathrm{d}y \\ &\leqslant \frac{1}{h_n}\int_{x-h_n}^{x+h_n} \big|f(y) - f(x)\big|\,\mathrm{d}y \quad \longrightarrow 0\end{aligned}$$

en tout point de Lebesgue x de f, ce qui achève de prouver (24.4).

Or, les Proposition 24.3 et Théorème 14.1 ne reposaient que sur l'analogue (24.1) et le fait que les $\widehat{f_n}$ soient des densités. En particulier, ils ne nécessitaient aucune hypothèse sur la densité de probabilité sous-jacente f. Ils valent donc tous deux pour les estimateurs par histogrammes définis ci-dessus.

Théorème 24.6 *Pour les estimateurs par histogrammes réguliers définis ci-dessus, lorsque $h_n \to 0$ et $nh_n \to \infty$, on a pour presque tout x,*

$$\widehat{f_n}(x) \xrightarrow{\mathbb{L}^2(\mathbb{P})} f(x)$$

et par conséquent,

$$\mathrm{d}_{\mathrm{VT}}\Big(\widehat{f_n},\ f\Big) \xrightarrow{\mathbb{P}} 0\,.$$

De même, il est aisé de voir que le résultat de normalité asymptotique de la Proposition 24.5 s'étend lui aussi aux estimateurs par histogrammes réguliers.

> Remarque au passage. Le lecteur pointilleux pourrait se demander si l'on a vraiment besoin de cette régularité des partitions et imaginer des situations où des informations *a priori* pourraient conduire à des partitions de forme *ad hoc*. L'hypothèse dont nous avons en fait besoin pour établir (24.4) est que les ensembles $A_{n,k_{n,x}}$ rétrécissent convenablement en tout point $x \in \mathbb{R}$, au sens de la définition donnée par [Rud98, paragraphes 7.9 et 7.10].

24.1.6 Consistance des estimateurs par noyaux

On pose qu'un noyau K est toute fonction mesurable $\mathbb{R} \to \mathbb{R}_+$ (positive donc) s'intégrant à 1. On n'impose pas ici de condition de carré intégrable. Les noyaux définissent donc des densités de probabilité, mais ce n'est pas sous cet angle que nous les verrons : nous les utiliserons comme des fonctions. Ainsi, lorsque nous supposerons qu'un noyau K admet un moment d'ordre deux, cela signifiera que

$$S = \int_{\mathbb{R}} K^2 \, \mathrm{d}\lambda < +\infty \, .$$

En particulier, l'intuition soutenant les estimateurs par noyaux est plutôt la suivante. La famille des noyaux

$$K_h : x \in \mathbb{R} \mapsto \frac{1}{h} \, K\left(\frac{x}{h}\right) ,$$

construits à partir d'un noyau K donné, approche, lorsque $h \to 0$, la masse de Dirac en 0. En effet, pour toute fonction f prise par exemple bornée et continue en 0, il vient, par changement de variables puis par théorème de convergence dominée,

$$\int_{\mathbb{R}} f \, K_h \, \mathrm{d}\lambda = \int_{\mathbb{R}} f(hx) \, K(x) \, \mathrm{d}x \longrightarrow f(0) \int_{\mathbb{R}} K(x) \, \mathrm{d}x = f(0)$$

lorsque $h \to 0$. Ainsi, les noyaux $K_h(\cdot - X_i)$ approchent la masse de Dirac en X_i et au final, les estimateurs proposés

$$\widehat{f}_{n,h} : x \in \mathbb{R} \mapsto \widehat{f}_{n,h}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

forment des approximations régulières de la mesure empirique μ_n. Les $K_h(\cdot - X_i)$ étant positives et s'intégrant à 1, ces estimateurs sont bien tous des densités de probabilité.

Dans la suite, on va considérer notamment les deux familles de noyaux suivantes : les noyaux gaussiens d'une part,

$$K_{(m,\,v)} = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(\cdot - m)^2}{2v}\right)$$

(où l'indexation est en lettres romaines m et v plutôt qu'en lettres grecques, parce que m et v sont des paramètres d'utilisateur, choisis par le statisticien, et non des paramètres inconnus), et les noyaux uniformes d'autre part,

$$K_{(a,b)} = \frac{1}{b-a}\,\mathbb{I}_{[a,b]}\,,$$

que l'on prendra le plus souvent symétriques : $a = -b$.

> Remarque au passage. On note que l'estimateur $\widehat{f}_{n,h}$ par noyau uniforme de paramètres $a = -b = 1$ est presque sûrement égal à l'estimateur par fenêtres glissantes de pas h. Il y aurait égalité exacte pour le choix du noyau $K = \mathbb{I}_{[-1,1[}/2$. Le Théorème 24.7 ci-dessous formera donc une généralisation du Théorème 14.1. Cependant, leurs preuves ne mettent pas en jeu les mêmes techniques et c'est pourquoi nous les présentons toutes les deux. En outre, la preuve effectuée pour le Théorème 14.1 était essentiellement ce qu'il nous fallait pour prouver le Théorème 24.6, qu'en revanche, il n'est pas possible de relier aux estimateurs par noyaux.

L'intérêt des estimateurs par noyaux est que les $\widehat{f}_{n,h}$ ont la même régularité que K comme fonctions de $\mathbb{R}$ dans $\mathbb{R}$. Les estimateurs construits sur les noyaux gaussiens sont par exemple très réguliers, de classe C^∞, et donnent par conséquent des graphiques bien plus agréables à visualiser que ceux obtenus par histogrammes ou même par fenêtres glissantes.

Il s'agit donc de prendre pour K un noyau de régularité déterminée en fonction des informations *a priori* que l'on a sur f. Si par exemple on sait que f est continue, il s'agira de prendre K continu également ; si f est à support compact, alors, lorsque K est également à support compact, les $\widehat{f}_{n,h}$ sont à support compact, inclus dans les h–voisinages du support de f.

On obtient la régularité maximale en prenant pour K une fonction ρ de classe C^∞ et de support compact, inclus dans $[-1,\ 1]$; il en existe, par exemple

$$\rho(x) = \begin{cases} \exp\left(\dfrac{1}{x^2-1}\right) & \text{si } |x| < 1, \\ 0 & \text{sinon.} \end{cases}$$

Mais cela peut dégrader les performances pratiques si la vraie densité n'est pas aussi régulière. La suite des K_{h_n} est parfois appelée suite régularisante (on pourra voir [Bre99, chapitre IV.4]).

Les choix classiques sont les noyaux gaussiens, les fonctions trigonométriques, et le noyau d'Epanechnikov

$$x \in \mathbb{R} \mapsto \frac{3}{4}\left(1-x^2\right)_+$$

(c'est bien une fonction positive s'intégrant à 1). Ce dernier est optimal en un certain sens pour la perte $\mathbb{L}^2(\lambda)$ mais pas pour la perte $\mathbb{L}^1(\lambda)$ considérée ici. On renvoie le lecteur curieux de détails à [Tsy04, chapitre 1] et [DL01, paragraphe 9.6]. Dans ce corrigé, nous ne considérons

que deux familles simples de noyaux, les noyaux gaussiens et les noyaux uniformes, ainsi que le noyau d'Epanechnikov.

Il s'agit ici de prouver le résultat suivant (que l'on énonce avec les notations du texte).

Théorème 24.7 *Pour tout noyau K admettant un moment d'ordre deux et pour toute suite déterministe (h_n) telle que $h_n \to 0$ et $nh_n \to \infty$, on a*

$$\mathbb{E}\left[\left\| \widehat{f}_{n,h_n} - f \right\|_1 \right] \longrightarrow 0$$

soit, de manière équivalente,

$$\mathrm{d}_{\mathrm{VT}}\left(\widehat{f}_{n,h_n},\ f \right) \xrightarrow{\mathbb{P}} 0\,.$$

Rappels préliminaires sur la convolution

Avant de détailler la preuve du Théorème 24.7, on fixe précisément les notations du texte, en définissant proprement la convolution de deux fonctions puis celle d'une probabilité avec une fonction et en rappelant leurs propriétés. Une référence pour retrouver la preuve de ces résultats est proposée à la fin du paragraphe.

Lorsque f et g sont deux fonctions de $\mathbb{L}^1(\lambda)$, pour presque tout x, la fonction $y \mapsto g(x-y)\,f(y)$ est intégrable ; on définit donc légitimement une fonction borélienne $f \star g$ par

$$f \star g(x) = \int_{\mathbb{R}} f(x-y)\,g(y)\,\mathrm{d}y\,.$$

En outre, on a encore $f \star g \in \mathbb{L}^1(\lambda)$, avec $\|f \star g\|_1 \leqslant \|f\|_1\ \|g\|_1$ (on appellera cette dernière inégalité l'inégalité de convolution) et la formule de symétrie $f \star g = g \star f$.

On veut généraliser cette définition à la convolution par des mesures. A cet effet, pour toute fonction g telle que $g(x - \cdot\,)$ est μ–intégrable pour presque tout $x \in \mathbb{R}$, on définit

$$\mu \star g(x) = \int_{\mathbb{R}} g(x-y)\,\mathrm{d}\mu(y)$$

en ces points x. En particulier, lorsque la mesure borélienne μ par laquelle on convole est donnée par une densité f par rapport à la mesure de Lebesgue, i.e., $\mathrm{d}\mu = f\,\mathrm{d}\lambda$, alors on a, pour toute fonction $g \in \mathbb{L}^1(\lambda)$, que $\mu \star g = f \star g\ (= g \star f)$.

On réinterprète alors l'estimateur $\widehat{f}_{n,h}$ comme la convolée $\mu_n \star K_h$:

$$\widehat{f}_{n,h} = \frac{1}{n}\sum_{t=1}^{n} K_h(\cdot - X_t) = \mu_n \star K_h$$

et on rappelle qu'il définit bien une densité de probabilité. Il procède immédiatement par linéarité de l'espérance que

$$\mathbb{E}\Big[\widehat{f}_{n,h}\Big] = \mathbb{E}\big[\mu_n \star K_h\big] = \mathbb{E}\big[K_h(\,\cdot\, - X_1)\big] = \mu \star K_h = f \star K_h\,,$$

qui est également une densité de probabilité.

> Remarque au passage. Pour des rappels sur la convolution de fonctions dans les espaces $\mathbb{L}^p(\lambda)$, avec $1 \leqslant p \leqslant \infty$, et notamment la preuve des résultats avancés ci-dessus, on pourra utilement consulter [Bre99, chapitre IV.4]. On y discute également la régularité des fonctions objectifs $f \star K_h$ (Propositions IV.19 et IV.20) et on y prouve même un cas particulier (Théorème IV.22) du fait que le terme de biais (défini formellement au paragraphe suivant)
>
> $$\big\| f \star K_h - f \big\|_1$$
>
> tend vers 0 lorsque $h \to 0$ et que K est pris égal à la fonction régularisante ρ.

Décomposition biais–variance

On a noté précédemment que $\mu_n \star K_{h_n}$ a pour espérance $f \star K_{h_n}$ et c'est pourquoi on s'attend à ce que $\mu_n \star K_{h_n}$ soit proche de $f \star K_{h_n}$ plutôt que de f. La différence entre ces deux derniers termes forme le terme dit de biais ou d'erreur d'approximation. Il mesure l'écart entre l'objectif plus simple de l'estimation, donné par $f \star K_{h_n}$, par rapport à l'objectif réel f. On fait apparaître ce terme de biais par inégalité triangulaire,

$$\mathbb{E}\Big[\,\big\|\widehat{f}_{n,h_n} - f\big\|_1\Big] \leqslant \big\|f \star K_{h_n} - f\big\|_1 + \mathbb{E}\Big[\,\big\|\mu_n \star K_{h_n} - f \star K_{h_n}\big\|_1\Big]\,. \tag{24.5}$$

Cette décomposition est réminiscente de celle vue au paragraphe 20.1.6. Le second terme de la borne correspond à un terme de variance (un terme lié à la difficulté d'estimer l'objectif $f \star K_{h_n}$). Dans la preuve ci-dessous, nous le majorerons essentiellement de la manière suivante, qui fait effectivement apparaître une variance :

$$\begin{aligned}\mathbb{E}\Big[\,\big\|\mu_n \star K_{h_n} - f \star K_{h_n}\big\|_1\Big] &= \int_{\mathbb{R}} \mathbb{E}\Big[\big|\mu_n \star K_{h_n}(x) - f \star K_{h_n}(x)\big|\Big]\,\mathrm{d}x \\ &\leqslant \int_{\mathbb{R}} \sqrt{\mathbb{E}\Big[\big(\mu_n \star K_{h_n}(x) - f \star K_{h_n}(x)\big)^2\Big]}\,\mathrm{d}x = \int_{\mathbb{R}} \sqrt{\mathrm{Var}\big(\mu_n \star K_{h_n}(x)\big)}\;\mathrm{d}x\end{aligned}$$

où nous avons utilisé le théorème de Fubini–Tonelli puis l'inégalité de Cauchy–Schwarz.

On prouve dans les deux paragraphes ci-dessous que chacun des termes du membre de droite de (24.5) tend vers 0. On a ainsi la première partie du Théorème 24.7, soit

$$\mathbb{E}\Big[\,\big\|\widehat{f}_{n,h_n} - f\big\|_1\Big] \longrightarrow 0\,.$$

Par le théorème de Scheffé (Théorème 24.1), les $\widehat{f}_{n,h_n}$ étant des densités de probabilité, cette convergence équivaut à

$$\mathbb{E}\Big[\mathrm{d}_{\mathrm{VT}}\Big(\widehat{f}_{n,h_n},\ f\Big)\Big] \longrightarrow 0\,, \qquad \text{qui implique} \qquad \mathrm{d}_{\mathrm{VT}}\Big(\widehat{f}_{n,h_n},\ f\Big) \xrightarrow{\mathbb{P}} 0\,.$$

Au vu du caractère borné des variables aléatoires en jeu, il y a en fait équivalence entre la dernière convergence, en probabilité, et les deux précédentes. (On pourra consulter la fin de la preuve du Théorème 32.2 pour le voir.)

Un résultat d'analyse : le terme de biais tend vers 0

Ici, on prouve dans un premier temps que, quelle que soit la fonction $f \in \mathbb{L}^1(\lambda)$ et pour tout noyau K, dont on n'utilise pour l'instant que l'appartenance $K \in \mathbb{L}^1(\lambda)$,

$$\|f \star K_h - f\|_1 = \int_{\mathbb{R}} \big|f \star K_h(x) - f(x)\big|\,\mathrm{d}x \longrightarrow 0 \qquad \text{quand } h \to 0. \tag{24.6}$$

On simplifie le problème en remarquant au préalable qu'il suffit de montrer cette convergence pour un sous-ensemble dense $\mathcal{F}$ bien choisi dans $\mathbb{L}^1(\lambda)$, par exemple celui formé par les fonctions lipschitziennes à support compact. (Le caractère dense de $\mathcal{F}$ est une conséquence d'un résultat encore plus fort, la densité des fonctions de classe C^∞ à support compact dans $\mathbb{L}^1(\lambda)$, voir [Bre99, Corollaire IV.23].)

En effet, pour tout $f \in \mathbb{L}^1(\lambda)$ et tout $\varepsilon > 0$, il existe alors $g \in \mathcal{F}$ telle que $\|f - g\|_1 \leqslant \varepsilon$. Par changement de variables, on a $\|K_h\|_1 = \|K\|_1 = 1$. Par inégalité triangulaire, en remarquant que par le théorème de Fubini $f \star K_h - g \star K_h = (f - g) \star K_h$ et en appliquant l'inégalité de convolution, il vient alors

$$\begin{aligned} \|f \star K_h - f\|_1 &\leqslant \|f \star K_h - g \star K_h\|_1 + \|g \star K_h - g\|_1 + \|g - f\|_1 \\ &\leqslant \big(\|K_h\|_1 + 1\big)\,\|f - g\|_1 + \|g \star K_h - g\|_1\ , \end{aligned}$$

soit

$$\limsup_{h\to 0} \|f \star K_h - f\|_1 \leqslant 2\varepsilon + \limsup_{h\to 0} \|g \star K_h - g\|_1 = 2\varepsilon\ .$$

Cette majoration par 2ε vaut pour tout ε, pourvu que l'on prouve la propriété (24.6) pour tous les éléments de $\mathcal{F}$. Dans ce dernier cas, elle vaut donc également pour tout élément de $\mathbb{L}^1(\lambda)$.

On fixe désormais une fonction g lipschitzienne de constante L et de support compact inclus dans $[-M,\ M]$, et on montre qu'elle vérifie (24.6). On commence par une majoration simple et un découpage,

$$\begin{aligned} \|g \star K_h - g\|_1 &= \int_{\mathbb{R}} \big|g \star K_h(x) - g(x)\big|\,\mathrm{d}x = \int_{\mathbb{R}} \left|\int_{\mathbb{R}} \big(g(x-y) - g(x)\big) K_h(y)\,\mathrm{d}y\right| \mathrm{d}x \\ &\leqslant \iint_{\mathbb{R}^2} \big|g(x-y) - g(x)\big|\,K_h(y)\,\mathrm{d}x\,\mathrm{d}y \end{aligned}$$

$$
\begin{aligned}
&= \iint_{\mathbb{R}^2} \big|g(x-y)-g(x)\big| K_h(y)\, \mathbb{I}_{[-hr,\,hr]}(y)\, \mathrm{d}x\, \mathrm{d}y \\
&\quad + \iint_{\mathbb{R}^2} \big|g(x-y)-g(x)\big| K_h(y)\, \mathbb{I}_{[-hr,\,hr]^c}(y)\, \mathrm{d}x\, \mathrm{d}y \qquad (24.7)
\end{aligned}
$$

et on s'intéresse à la convergence vers 0 de chacun des deux termes de la majoration finale. Le seuil r qui est apparu est un paramètre de l'analyse destiné à tendre vers $+\infty$.

Pour majorer le premier terme de (24.7), on note tout d'abord que pour $0 < h \leqslant 1$, les conditions $y \in [-hr, hr]$ (qui entraîne $y \in [-r, r]$) et $x \notin [-M-r,\ M+r]$ sont des conditions suffisantes pour que $x-y$ et x soient simultanément hors du support $[-M,\ M]$ de g. On a alors, en utilisant en outre le caractère lipschitzien de g,

$$
\begin{aligned}
&\iint_{\mathbb{R}^2} \big|g(x-y)-g(x)\big| K_h(y)\, \mathbb{I}_{[-hr,\,hr]}(y)\, \mathrm{d}x\, \mathrm{d}y \\
&\quad \leqslant \quad L \iint_{\mathbb{R}^2} |y|\, K_h(y)\, \mathbb{I}_{[-hr,\,hr]}(y)\, \mathbb{I}_{[-M-r,\,M+r]}(x)\, \mathrm{d}x\, \mathrm{d}y \\
&\quad \leqslant \quad 2L(M+r) \int_{\mathbb{R}} |y|\, K_h(y)\, \mathbb{I}_{[-r,\,r]}\left(\frac{y}{h}\right) \mathrm{d}y \\
&\quad = \quad 2L(M+r)h \int_{\mathbb{R}} \left|\frac{y}{h}\right| K_h(y)\, \mathbb{I}_{[-r,\,r]}\left(\frac{y}{h}\right) \mathrm{d}y \\
&\quad \leqslant \quad 2L(M+r)hr \int_{\mathbb{R}} K_h(y)\, \mathrm{d}y = 2L(M+r)r\ h\ .
\end{aligned}
$$

Le majorant obtenu tend vers 0 lorsque $h \to 0$.

D'autre part, pour le second terme de (24.7), par application du changement de variables $y' = y/h$,

$$
\begin{aligned}
&\iint_{\mathbb{R}^2} \big|g(x-y)-g(x)\big| K_h(y)\, \mathbb{I}_{[-hr,\,hr]^c}(y)\, \mathrm{d}x\, \mathrm{d}y \\
&\quad \leqslant \quad \iint_{\mathbb{R}^2} \Big(\big|g(x-y)\big| + \big|g(x)\big|\Big) K_h(y)\, \mathbb{I}_{[-hr,\,hr]^c}(y)\, \mathrm{d}x\, \mathrm{d}y \\
&\quad = \quad 2\,\|g\|_1 \int_{\mathbb{R}} K_h(y)\ \mathbb{I}_{[-r,\,r]^c}\left(\frac{y}{h}\right) \mathrm{d}y = 2\,\|g\|_1 \int_{\mathbb{R}} K(y')\ \mathbb{I}_{[-r,\,r]^c}(y')\, \mathrm{d}y'.
\end{aligned}
$$

Le majorant obtenu est indépendant de h.

En combinant les deux majorants et prenant la limite supérieure lorsque $h \to 0$, on a finalement, pour tout $r > 0$ puis en prenant la limite lorsque $r \to +\infty$,

$$
\limsup_{h \to 0} \|g \star K_h - g\|_1 \leqslant 2\,\|g\|_1 \left(\int_{-\infty}^{-r} K\, \mathrm{d}\lambda + \int_r^{+\infty} K\, \mathrm{d}\lambda\right) \longrightarrow 0 \quad \text{quand } r \to +\infty.
$$

La convergence procède d'un résultat de convergence monotone, K étant intégrable positive. Cela clôt la preuve du résultat préliminaire d'analyse.

Le terme de variance tend également vers 0

Il s'agit maintenant de montrer que

$$\lim_{n\to\infty} \mathbb{E}\Big[\big\| \mu_n \star K_{h_n} - f \star K_{h_n} \big\|_1 \Big] = 0 \,.$$

Ce terme reflète une difficulté d'estimation et c'est par abus qu'on l'appelle terme de variance. Cependant, la preuve ci-dessous exhibera une majoration en termes d'une intégrale d'écart-types, ainsi que le suggéraient les commentaires après la décomposition (24.5).

On fixe $\varepsilon > 0$ et on montre ci-dessous que

$$\limsup_{n\to\infty} \mathbb{E}\Big[\big\| \mu_n \star K_{h_n} - f \star K_{h_n} \big\|_1 \Big] \leqslant \varepsilon \,; \tag{24.8}$$

ceci valant pour tout $\varepsilon > 0$, on a bien la convergence vers 0 souhaitée.

On rappelle que $\mu_n \star K_{h_n}$ et $\mu \star K_{h_n} = f \star K_{h_n}$ sont des densités de probabilité. Par conséquent, le théorème de Scheffé (Théorème 24.1) assure que

$$\begin{aligned} \big\| \mu_n \star K_{h_n} - f \star K_{h_n} \big\|_1 &= 2 \int_{\mathbb{R}} \big(f \star K_{h_n} - \mu_n \star K_{h_n} \big)_+ \, \mathrm{d}\lambda \\ &\leqslant 2 \int_{[-M,\,M]^c} f \star K_{h_n} \, \mathrm{d}\lambda + 2 \int_{[-M,\,M]} \big(f \star K_{h_n} - \mu_n \star K_{h_n} \big)_+ \, \mathrm{d}\lambda \end{aligned}$$

où, par intégrabilité de f, on a choisi le seuil $M > 0$ tel que

$$\int_{[-M,\,M]^c} f \, \mathrm{d}\lambda \leqslant \varepsilon \,.$$

Par le théorème de Fubini–Tonelli,

$$\begin{aligned} &\mathbb{E}\Big[\big\| \mu_n \star K_{h_n} - f \star K_{h_n} \big\|_1 \Big] \\ &\qquad \leqslant 2 \int_{[-M,\,M]^c} f \star K_{h_n} \, \mathrm{d}\lambda + 2 \int_{[-M,\,M]} \mathbb{E}\Big[\big(f \star K_{h_n}(x) - \mu_n \star K_{h_n}(x) \big)_+ \Big] \, \mathrm{d}x \,. \end{aligned}$$

Or, on a vu au paragraphe précédent que la suite $(f \star K_{h_n})$ converge vers f dans $\mathbb{L}^1(\lambda)$, ce qui entraîne en particulier

$$\int_{[-M,\,M]^c} f \star K_{h_n} \, \mathrm{d}\lambda \underset{n\to\infty}{\longrightarrow} \int_{[-M,\,M]^c} f \, \mathrm{d}\lambda \leqslant \varepsilon \,.$$

Il suffit donc de montrer que

$$\int_{[-M,\,M]} \mathbb{E}\Big[\big(f \star K_{h_n}(x) - \mu_n \star K_{h_n}(x) \big)_+ \Big] \, \mathrm{d}x \underset{n\to\infty}{\longrightarrow} 0$$

pour prouver (24.8).

On s'intéresse à cet effet aux fonctions

$$x \mapsto \varphi_n(x) = \mathbb{E}\Big[\big(f \star K_{h_n}(x) - \mu_n \star K_{h_n}(x)\big)_+\Big] .$$

D'une part, l'inégalité de Cauchy–Schwarz assure que pour tout x,

$$\varphi_n(x) \leqslant \sqrt{\mathbb{E}\Big[\big(f \star K_{h_n}(x) - \mu_n \star K_{h_n}(x)\big)^2\Big]}$$

et l'on reconnaît

$$\begin{aligned}\mathbb{E}\Big[\big(f \star K_{h_n}(x) - \mu_n \star K_{h_n}(x)\big)^2\Big] &= \operatorname{Var}\big(\mu_n \star K_{h_n}(x)\big) \\ &= \operatorname{Var}\left(\frac{1}{n}\Big(Y_1^{n,x} + \ldots + Y_n^{n,x}\Big)\right) = \frac{1}{n}\operatorname{Var} Y_1^{n,x}\end{aligned}$$

où, pour n et x fixés, les variables aléatoires $Y_j^{n,x} = K_{h_n}(x - X_j)$ sont indépendantes de même loi lorsque $j = 1, \ldots, n$. Or,

$$\begin{aligned}\operatorname{Var} Y_1^{n,x} \leqslant \mathbb{E}\Big[\big(Y_1^{n,x}\big)^2\Big] &= \int_{\mathbb{R}} \big(K_{h_n}(x-y)\big)^2 f(y)\,\mathrm{d}y = \frac{1}{h_n}\int_{\mathbb{R}} \frac{1}{h_n} K^2\left(\frac{x-y}{h_n}\right) f(y)\,\mathrm{d}y \\ &= \frac{S}{h_n}\int_{\mathbb{R}} K'_{h_n}(x-y) f(y)\,\mathrm{d}y = \frac{S}{h_n} K'_{h_n} \star f(x) ,\end{aligned}$$

où l'on a noté $K' = K^2/S$ (on rappelle que S a été défini comme l'intégrale de K^2). La fonction K' est bien un noyau, car elle est positive et s'intègre à 1.

Après substitution de ces inégalités, il vient la majoration suivante du terme d'intérêt,

$$\begin{aligned}\int_{-M}^{M} \varphi_n(x)\,\mathrm{d}x &\leqslant \frac{1}{\sqrt{n}}\int_{-M}^{M} \sqrt{\operatorname{Var} Y_1^{n,x}}\,\mathrm{d}x \leqslant \sqrt{\frac{S}{nh_n}}\int_{-M}^{M} \sqrt{K'_{h_n} \star f(x)}\,\mathrm{d}x \\ &\leqslant \sqrt{\frac{S}{nh_n}}\ \sqrt{2M}\sqrt{\int_{-M}^{M} K'_{h_n} \star f(x)\,\mathrm{d}x} \leqslant \sqrt{\frac{2MS}{nh_n}} \underset{n\to\infty}{\longrightarrow} 0 ,\end{aligned}$$

où l'on a appliqué d'abord l'inégalité de Jensen à la fonction concave $\sqrt{\cdot}$ et à la probabilité uniforme sur $[-M, M]$, puis l'on a utilisé le fait que les $K'_{h_n} \star f$, étant des densités de probabilité, sont d'intégrales sur $\mathbb{R}$ égales à 1, et enfin, le fait que $nh_n \to \infty$.

Remarques et commentaires sur le Théorème 24.7 et sa preuve

Renforcement du résultat [Dev87, chapitre 3] montre de manière générique que si une procédure d'estimation par noyaux est consistante pour l'estimation d'une densité de probabilité f, alors elle est également fortement consistante pour l'estimation de toutes les densités de probabilité et qu'une condition nécessaire et suffisante pour cela est que $h_n \to 0$ et $nh_n \to \infty$.

Ces équivalences sont tout à fait remarquables. En particulier, le Théorème 24.7 peut être renforcé et énoncer une convergence $\mathbb{P}$–presque sûre des distances en variation totale. Les techniques utilisées sont des techniques de concentration de la mesure, appliquées à des lois multinomiales.

Par ailleurs, [Dev87, chapitre 3] indique également que même si l'on choisit des tailles de fenêtre H_n aléatoires, mesurables en les $X_1, \ldots, X_n$ et telles que $H_n \to 0$ et $nH_n \to \infty$ en $\mathbb{P}$–probabilité, alors le résultat de convergence du Théorème 24.7 est encore vérifié. L'extension n'est pas aussi simple qu'il n'y paraît ; on peut avoir une idée des problèmes qui se posent en notant par exemple que dans ce cas, les $Y_j^{n,x} = K_{H_n}(x - X_j)$ ne sont plus indépendantes.

À propos des hypothèses On pourrait considérer dans ce qui précède des fonctions K signées (i.e., à valeurs dans $\mathbb{R}$ et non seulement dans $\mathbb{R}_+$) et intégrables, $K \in \mathbb{L}^1(\lambda)$, d'intégrale $\int K \, d\lambda = 1$. Dans ce cas, on peut toujours prouver la convergence dans $\mathbb{L}^1(\mathbb{P} \otimes \lambda)$ des densités,

$$\mathbb{E}\left[\left\| \widehat{f}_{n,h_n} - f \right\|_1 \right] \longrightarrow 0 \, ;$$

il suffit de modifier légèrement la preuve en remplaçant K par $|K|$ lorsqu'il le faut (essentiellement, lorsque l'on majore des intégrales). On perd cependant le caractère de densité de probabilité des estimateurs $\widehat{f}_{n,h_n}$: ces derniers définissent une mesure signée de masse 1 par rapport à la mesure de Lebesgue. Il est facile de vérifier que le théorème de Scheffé (Théorème 24.1) s'étend à ce cadre, de sorte que l'on conserve bien ici l'interprétation en termes de convergence vers 0 en $\mathbb{P}$–probabilité de la suite des distances en variation totale.

Dans le même ordre d'idées, on a énoncé le Théorème 24.7 pour des noyaux K ayant un moment d'ordre deux, que l'on avait noté S. Cette dernière condition est superflue : on peut le voir en approchant K à ε près au sens $\mathbb{L}^1$ par une fonction L bornée à support compact, comme le propose [DL01, chapitres 9.1 et 9.3]. Ici encore, c'est une simple modification de la preuve précédente, qui consiste essentiellement à ajouter des facteurs correctifs $\int |K - L| \, d\lambda$ partout où cela est nécessaire.

Un autre schéma de preuve pour le Théorème 24.7 On pourra le trouver au paragraphe 32.1.2 en annexe.

24.1.7 Comparaison par simulations des différentes méthodes d'estimation

En plus de la loi ν_u introduite à la fin du paragraphe 24.1.4 comme un mélange de lois uniformes, nous considérons ici la loi ν_g donnée par un mélange de lois normales, de densité par rapport à la mesure de Lebesgue égale à

$$f_g = \sum_{j=1}^{3} q_j \Phi_{m_j, s_j}$$

h	Histogrammes	Fenêtres glissantes	Noyaux : gaussien et d'Epanechnikov	
0.02	0.236	0.168	0.131	0.184
0.05	0.152	0.117	<u>0.101</u>	0.124
0.15	0.119	<u>0.108</u>	0.125	<u>0.100</u>
0.20	<u>0.099</u>	0.119	0.149	0.104
0.30	0.133	0.149	0.208	0.123
0.40	0.146	0.183	0.270	0.146

TABLE 24.1. Estimées des distances en variation totale des estimateurs au mélange des densités uniformes ν_u en fonction des tailles de bande h et pour $n = 400$: moyennes de Monte-Carlo sur $K = 1\,000$ répétitions. Dans chaque colonne, la meilleure valeur est soulignée.

où $\Phi_{m,s}$ est la densité de la loi $\mathcal{N}(m, s^2)$ et où l'on a effectué les choix de paramètres $(q_1, q_2, q_3) = (0.45,\ 0.20,\ 0.35)$ et

$$(m_1, s_1) = (-1.5,\ 0.75)\ , \qquad (m_2, s_2) = (0.5,\ 0.60)\ , \qquad (m_3, s_3) = (2.5,\ 0.45)\ .$$

Nous étudions ici quatre méthodes d'estimation : par histogrammes, par utilisation du noyau d'Epanechnikov et du noyau gaussien standard $(m,\ v) = (0,\ 1)$, et par fenêtres glissantes (cette dernière correspondant presque sûrement à l'utilisation du noyau uniforme $\mathbb{I}_{[-1,1]}/2$). Nous avons réalisé ces simulations sur des n–échantillons de données artificielles, avec $n = 400$. Comme à la fin du paragraphe 24.1.4, nous donnons à titre informatif ce que pourraient être des paramètres choisis de manière automatique :

$$h_{400} = 1/\sqrt{400} = 0.05\ , \qquad h'_{400} = 400^{-1/3} \approx 0.14 \qquad \text{et} \qquad h''_{400} = 400^{-1/5} \approx 0.30\ .$$

Nous rappelons une fois encore que ces choix ont été faits à constante multiplicative près. La valeur de cette dernière n'a pas d'influence sur les résultats asymptotiques mais elle change de manière cruciale les performances non asymptotiques. Les tableaux 24.1 et 24.2 mettent cela en lumière : les performances pratiques sont fort différentes selon que l'on prend h_{400} ou $3h_{400}$.

Ces tableaux représentent des estimées des espérances des distances en variation totale entre les estimateurs considérés (définis par une méthode d'estimation et un paramètre de bande h) et les vraies densités. On a obtenu ces estimées en répétant $N = 1\,000$ fois la procédure qui consiste à simuler un 400–échantillon, calculer tous les estimateurs et déterminer leur distance en variation totale respective à la vraie densité (par calcul d'une intégrale selon la méthode des trapèzes). On a alors reporté la moyenne des 1 000 valeurs obtenues. Ce faisant, on a réalisé une instance de l'estimation par méthode de Monte-Carlo, voir le paragraphe 9.7.1.

Avant de commenter ces résultats, on explique pourquoi on les a donnés avec trois chiffres après la virgule : il s'agit de discuter la précision de nos estimées. En même temps que l'on

h	Histogrammes	Fenêtres glissantes	Noyaux : gaussien et d'Epanechnikov	
0.02	0.358	0.255	0.193	0.280
0.05	0.230	0.162	0.121	0.177
0.15	0.134	0.093	0.070	0.102
0.20	0.118	0.081	<u>0.064</u>	0.088
0.30	0.101	0.066	0.067	0.070
0.40	<u>0.097</u>	<u>0.064</u>	0.087	0.064
0.50	0.099	0.069	0.110	<u>0.062</u>
0.60	0.103	0.081	0.134	0.065

TABLE 24.2. Estimées des distances en variation totale des estimateurs au mélange des densités gaussiennes v_g en fonction des tailles de bande h et pour $n = 400$: moyennes de Monte-Carlo sur $K = 1\,000$ répétitions. Dans chaque colonne, la meilleure valeur est soulignée.

a calculé les moyennes, on a aussi regardé les écarts-types associés à chacune d'elles, afin de déterminer la demi-largeur des réalisations des intervalles de confiance (9.2). Ils étaient tous inférieurs à 0.015, ce qui montre que chaque valeur numérique des tableaux est associée (individuellement) à la réalisation d'un intervalle de confiance à 95 % de demi-largeur inférieure à la réalisation de

$$z_{97.5\,\%} \frac{\widehat{s}_N}{\sqrt{N}}\,, \qquad \text{soit} \qquad 1.96\,\frac{0.015}{\sqrt{1\,000}} \leqslant 0.001\,.$$

C'est à cause de cette précision à ± 0.001 que nous avons reporté trois chiffres significatifs dans les résultats numériques.

Dans les tableaux 24.1 et 24.2, nous avons choisi des paramètres de bande tressant une grille autour de h_{400}, h'_{400} et h''_{400}, et de sorte que l'on observe dans chaque colonne un comportement en forme de puits : d'abord une décroissance des (estimées des) distances en variation totale lorsque h augmente puis une croissance.

On passe maintenant à l'interprétation proprement dite des résultats. On remarque que pour l'estimation du mélange de lois uniformes, les méthodes ont toutes des performances similaires, sauf peut-être l'estimation par fenêtres glissantes qui est un tout petit peu en retrait. Pour le cas du mélanges de lois gaussiennes, trois méthodes obtiennent des performances similaires mais cette fois-ci la quatrième méthode, celle par histogrammes, est en net retrait. Dans tous les cas, le paramètre de bande optimal h dépend fortement de la méthode. Il n'est donc pas seulement fonction de la taille d'échantillon n et de la densité f à estimer. On a représenté aux figures 24.2 et 24.3 une réalisation de chacun des estimateurs sur le même jeu de données et pour des valeurs de h choisies selon les tableaux 24.1 et 24.2 de sorte que toutes les méthodes aient des performances d'estimation raisonnables. (Notons au passage

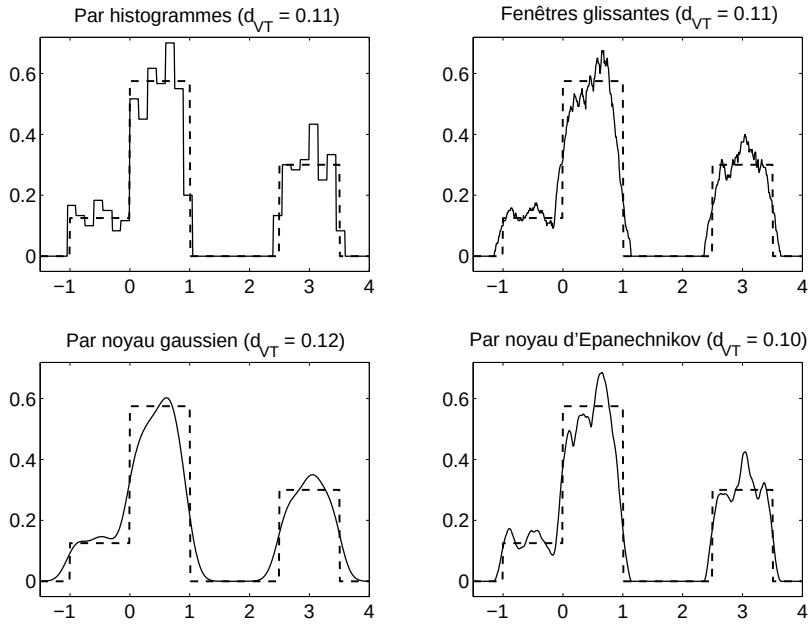

FIGURE 24.2. Une réalisation des estimations pour le mélange de lois uniformes, $n = 400$ et $h = 0.15$. On précise entre parenthèses les distances en variation totale réalisées.

que nous avons reprogrammé les méthodes d'estimation par histogrammes, déjà implémentés dans `Matlab` par les fonctions `hist` et `histo`, pour avoir une plus grande uniformité dans les affichages graphiques et pour permettre le calcul facile des distances en variation totale, voir les codes des programmes informatiques au paragraphe 32.2.)

D'un point de vue visuel, l'estimateur le plus agréable car le plus régulier est fourni, comme la théorie le garantit, par l'utilisation d'un noyau gaussien et, dans une moindre mesure, d'un noyau d'Epanechnikov. On rappelle que les estimateurs par fenêtres glissantes et par histogrammes ne sont pas continus en général.

On le voit, la calibration (la détermination d'une bonne valeur) de h sur des données, réelles ou artificielles, est délicate. Ici, les tableaux 24.1 et 24.2 représentent des performances auxquelles on a eu accès uniquement parce que l'on connaissait les vraies densités sous-jacentes et les valeurs qu'ils recommandent ne valent que pour le cadre étudié. On ne pourrait pas reproduire cette méthodologie de choix de h sur des données pour lesquelles on ne connaîtrait pas la vraie densité f. C'est là le point sensible des méthodes d'estimation vues ici : la calibration de h à rang fini n (le choix de l'ordre de grandeur asymptotique des h_n ayant été traité par la théorie).

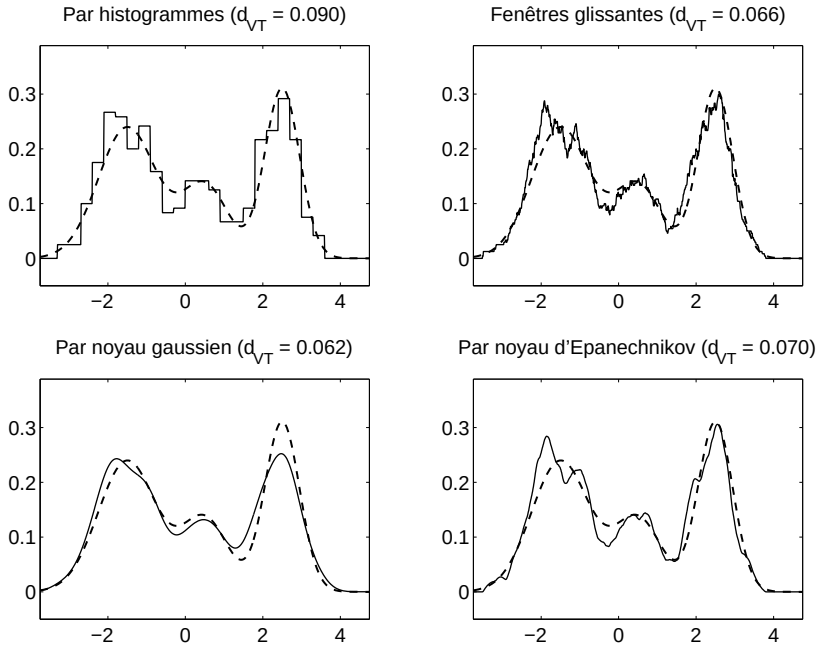

FIGURE 24.3. Une réalisation des estimations pour le mélange de lois gaussiennes, $n = 400$ et $h = 0.30$. On précise entre parenthèses les distances en variation totale réalisées.

Remarque au passage. Nous présentons brièvement deux méthodes pour réaliser cette calibration dans un cadre $\mathbb{L}^2$. Les techniques pour prouver la première sont similaires à celles que nous avons employées pour établir le résultat de normalité asymptotique du paragraphe 24.1.4. On définit un noyau centré et admettant un moment d'ordre deux (tant comme fonction que comme densité de probabilité) comme un noyau K tel que

$$S = \int_{\mathbb{R}} K^2 \, \mathrm{d}\lambda < \infty \,, \qquad s = \int_{\mathbb{R}} x^2 K(x) \, \mathrm{d}x < \infty \quad \text{et} \quad \int_{\mathbb{R}} x \, K(x) \, \mathrm{d}x = 0 \,.$$

Théorème 24.8 (cf. [Was06], Théorème 6.28) *On considère un noyau K centré et tel que ses moments d'ordre deux s et S sont finis. On suppose que la densité f est deux fois dérivable, de dérivée seconde f'' absolument continue et vérifiant*

$$A(f) \overset{\text{not.}}{=} \int \left(f''\right)^2 \mathrm{d}\lambda < \infty \,.$$

On a alors la majoration

$$\mathbb{E}\left[\left\| \widehat{f}_{n,h_n} - f \right\|_2^2 \right] = \frac{s^2 A(f)}{4} h_n^4 + \frac{S}{n h_n} + O\left(\frac{1}{n}\right) + O\left(h_n^6\right) .$$

À la lumière de ce résultat, on propose la valeur suivante pour h_n, qui minimise la somme des deux termes principaux de la borne théorique exhibée :

$$h_n = \left(\frac{S}{s^2 A(f)}\right)^{1/5} n^{-1/5} .$$

Pour déterminer la constante optimale devant le facteur $n^{-1/5}$ dans le cas du noyau gaussien, l'idée est de considérer que f est la densité d'une variable gaussienne de variance σ^2. Dans ce cas,

$$h_n = 1.06\,\sigma\, n^{-1/5} .$$

Le choix final de h_n s'obtient en substituant à σ une estimée obtenue à l'aide de l'échantillon (par exemple par les techniques du chapitre 2). Cette règle, dite « règle du pouce », donne de bons résultats lorsque la densité sous-jacente f est très régulière.

La seconde technique est la méthode de la validation croisée. Nous présentons ci-dessous des résultats pour la version dite « leave one out » ; on a vu au paragraphe 20.1.11 la version « m–fold ». Ici, pour tout $i \in \{1, \ldots, n\}$ et tout $h > 0$, on définit l'estimateur par noyau $\widehat{f}_{n,h,-i}$ construit sur toutes les données sauf la i–ième :

$$\widehat{f}_{n,h,-i} = \frac{1}{(n-1)\,h} \sum_{j \neq i} K\left(\frac{X_j - \cdot}{h}\right) .$$

A partir de ces estimateurs, on peut estimer sans biais le risque $\mathbb{L}^2$ de la procédure par noyaux s'appuyant sur l'ensemble de l'échantillon. Plus précisément, on a le résultat suivant.

Théorème 24.9 (cf. [Tsy04], Proposition 1.9) *On suppose que $f \in \mathbb{L}^2(\lambda)$ et que pour tout $h > 0$,*

$$\iint_{\mathbb{R}^2} \left| K\left(\frac{x-y}{h}\right) \right| f(x) f(y)\, \mathrm{d}x\, \mathrm{d}y < \infty .$$

Alors, pour tout $h > 0$,

$$\mathbb{E}\left[\| \widehat{f}_{n,h} - f \|_2^2 \right] = \int_{\mathbb{R}} f^2\, \mathrm{d}\lambda + \mathbb{E}\left[\int_{\mathbb{R}} \widehat{f}_{n,h}^{\,2}\, \mathrm{d}\lambda - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{n,h,-i}(X_i) \right] .$$

La méthode de la validation croisée propose alors de choisir un paramètre $h_n > 0$ minimisant l'estimateur sans biais du risque :

$$h_n \in \operatorname*{argmin}_{h>0} \left\{ \int_{\mathbb{R}} \widehat{f}_{n,h}^{\,2}\, \mathrm{d}\lambda - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{n,h,-i}(X_i) \right\} .$$

Il sera fait allusion à une troisième technique à la fin du paragraphe 32.1.4, issue de [DL01]. Cette dernière, comme la règle du pouce et contrairement à la validation croisée, est spécifique au cas du choix du paramètre de bande des estimateurs à noyaux ; elle repose sur des méthodes combinatoires.

24.1.8 L'exemple de la reconnaissance de formes

Le nom de l'exemple est justifié par l'application suivante. Avec les notations du texte, on prend par exemple $M = 10$ et on identifie $\{1, \ldots, 10\}$ à $\{0, \ldots, 9\}$: il s'agit de reconnaître des chiffres. C'est ce que fait par exemple une machine de tri postal, quand on lui soumet des enveloppes sur lesquelles il y a cinq cases pour les cinq chiffres du code postal ; elle associe à une image X_t (qui n'est rien d'autre qu'un élément d'un certain ensemble multidimensionnel $\mathbb{R}^d$) une étiquette dans $\{0, \ldots, 9\}$, qui peut être égale à la vraie étiquette Y_t ou être différente d'elle. Le chapitre 20 est entièrement consacré à cette application et explique plus en détails d'où vient la modélisation stochastique que nous rappelons ci-dessous uniquement sur le plan mathématique et pas tellement du point de vue des intuitions ou des heuristiques qui la sous-tendent.

Ici, pour simplifier, on suppose que les éléments X_t à classer vivent dans $\mathbb{R}$; à chacun d'entre eux correspond une étiquette $Y_t \in \{1, \ldots, M\}$, aléatoire (le paragraphe 20.1.1 explique la raison de l'existence de cet aléa). Les couples (X_t, Y_t) sont indépendants et identiquement distribués selon une loi ν, inconnue, qui n'est évidemment pas une loi produit, mais une loi jointe, non dégénérée (sans quoi, la connaissance de X_t n'apporterait aucune information sur la valeur de Y_t et aucune inférence statistique ne pourrait avoir lieu).

Une fonction de classification est toute fonction mesurable $g : \mathbb{R} \to \{1, \ldots, M\}$, qui nous indique donc, étant donnée une nouvelle observation X, quelle étiquette $g(X)$ nous prédisons. La qualité d'une fonction est mesurée en espérance, par l'erreur moyenne

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

On cherche à minimiser $L(g)$ alors qu'on ne connaît pas ν, mais qu'on a simplement accès à un n–échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de loi ν. On va évidemment estimer la loi jointe ν, avec l'intuition qu'une bonne estimation de ν nous permettra de choisir une bonne fonction. Pour estimer ν, il faut estimer la loi de Y (ce qui est aisé puisque Y est à valeurs discrètes) de même que les différentes marginales conditionnelles de X (conditionnellement aux $Y = i$, pour $i = 1, \ldots, M$).

On note μ la première marginale de ν (celle selon X) et on suppose, pour pouvoir appliquer les résultats précédents, qu'elle est absolument continue par rapport à la mesure de Lebesgue, de densité f. On a alors que les marginales selon X conditionnellement à $Y = i$ sont toutes absolument continues par rapport à la mesure de Lebesgue, de densités respectives notées f_i, pour $i = 1, \ldots, M$. En effet, si un événement borélien A est de mesure nulle pour la mesure de Lebesgue, alors, par absolue continuité de μ, il est tel que $\mu(A) = \mathbb{P}\{X \in A\} = 0$, et en particulier, $\mathbb{P}\{X \in A \mid Y = i\} = 0$.

Supposons dans un premier temps qu'on connaisse ν, c'est-à-dire les $p_i = \mathbb{P}\{Y = i\}$ et les f_i : quelle serait la fonction optimale en fonction des p_i et f_i ? On réécrit l'erreur moyenne en fonction de g,

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

$$= \quad 1 - \sum_{i=1}^{M} \mathbb{P}\big\{g(X) = i \text{ et } Y = i\big\} = 1 - \sum_{i=1}^{M} p_i \, \mathbb{P}\big\{g(X) = i \mid Y = i\big\}$$

$$= \quad 1 - \sum_{i=1}^{M} p_i \int_{\mathbb{R}} \mathbb{I}_{\{g(x)=i\}} f_i(x) \, \mathrm{d}x = 1 - \int_{\mathbb{R}} p_{g(x)} f_{g(x)}(x) \, \mathrm{d}x \, .$$

Cela prouve qu'une meilleure fonction g^* (dite de Bayes, voir le paragraphe 20.1.2) est définie par

$$g^*(x) \in \underset{i=1,\ldots,N}{\operatorname{argmax}} \; p_i f_i(x) \, .$$

On remplace alors les p_i et f_i dans la définition de la fonction optimale par des estimateurs $\widehat{p}_{i,n}$ et $\widehat{f}_{i,n}$ (précisés ci-dessous) pour construire la suite de fonctions (la règle) donnée par

$$\widehat{g}_n(x) \in \underset{i=1,\ldots,N}{\operatorname{argmax}} \; \widehat{p}_{i,n} \widehat{f}_{i,n}(x) \, .$$

Les estimateurs $\widehat{p}_{i,n}$ sont simplement la proportion de Y_t valant i,

$$\widehat{p}_{i,n} = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}_{\{Y_t = i\}} \, ;$$

ils sont fortement consistants par loi des grands nombres.

Les $\widehat{f}_{i,n}$ sont obtenus par les techniques précédentes, par exemple celles à noyaux, en ne considérant pour chacun d'eux que les X_t tels que $Y_t = i$. On note $N_{i,n} = n\widehat{p}_{i,n}$ la taille du sous-échantillon associé à l'étiquette i. Ainsi, les définitions précédentes donnent ici, pour un noyau K et une étiquette i telle que $N_{i,n} > 0$,

$$\widehat{f}_{i,n} = \frac{1}{N_{i,n} h_n} \sum_{t : Y_t = i} K\left(\frac{\cdot - X_t}{h_n}\right) .$$

On mène d'abord l'étude conditionnellement à la suite infinie

$$Y^\infty = \big(Y_1, \, Y_2, \, \ldots, \, Y_n, \, \ldots\big) \, .$$

Le Théorème 24.7 assure que les $\widehat{f}_{i,n}$ estiment de manière consistante f_i en distance en variation totale dès que $h_n \to 0$ et $N_{i,n} h_n \to \infty$; pour les réalisations y^∞ de Y^∞ vérifiant ces conditions,

$$\mathbb{E}\left[\left\| \widehat{f}_{i,n} - f_i \right\|_1 \mid Y^\infty = y^\infty \right] \longrightarrow 0 \, .$$

Or, $n h_n \to \infty$ de manière déterministe et $\mathbb{P}$–presque sûrement, $\widehat{p}_{i,n} \to p_i$ (où l'on suppose que $p_i > 0$, quitte à ne pas considérer i), de sorte que $\mathbb{P}$–presque sûrement, $N_{i,n} h_n \to \infty$. Ainsi,

$$\mathbb{E}\left[\left\| \widehat{f}_{i,n} - f_i \right\|_1 \mid Y^\infty \right] \longrightarrow 0 \qquad \text{p.s.}$$

Par application du théorème de convergence dominée et intégration selon Y^∞, il vient la consistance en distance en variation totale suivante :

$$\mathbb{E}\left[\left\| \widehat{f}_{i,n} - f_i \right\|_1 \right] \longrightarrow 0 \,.$$

Or, l'expression précédente pour les erreurs moyennes conduit à

$$\begin{aligned} L\big(\widehat{g}_n\big) - L(g^*) &= \int_{\mathbb{R}} p_{g^*} f_{g^*} \,\mathrm{d}\lambda - \int_{\mathbb{R}} p_{\hat{g}_n} f_{\hat{g}_n} \,\mathrm{d}\lambda \\ &= \int_{\mathbb{R}} \big(p_{g^*} f_{g^*} - \widehat{p}_{\hat{g}_n,n} \widehat{f}_{\hat{g}_n,n}\big) \,\mathrm{d}\lambda + \int_{\mathbb{R}} \big(\widehat{p}_{\hat{g}_n,n} \widehat{f}_{\hat{g}_n,n} - p_{\hat{g}_n} f_{\hat{g}_n}\big) \,\mathrm{d}\lambda \\ &= \int_{\mathbb{R}} \left(\max_{i=1,\dots,N} p_i f_i - \max_{i=1,\dots,N} \widehat{p}_{i,n} \widehat{f}_{i,n} \right) \mathrm{d}\lambda + \int_{\mathbb{R}} \big(\widehat{p}_{\hat{g}_n,n} \widehat{f}_{\hat{g}_n,n} - p_{\hat{g}_n} f_{\hat{g}_n}\big) \,\mathrm{d}\lambda \\ &\leqslant 2 \sum_{i=1}^{M} \int_{\mathbb{R}} \big| p_i f_i - \widehat{p}_{i,n} \widehat{f}_{i,n} \big| \,\mathrm{d}\lambda \,, \end{aligned}$$

où la majoration, assez grossière, procède d'une part de ce que la différence entre deux maxima peut être bornée par le maximum des différences, et donc leur somme, et d'autre part, d'une décomposition selon les valeurs de $\widehat{g}_n$.

On continue le calcul pour faire apparaître séparément les p_i et $\widehat{p}_{i,n}$ d'un côté et les f_i et $\widehat{f}_{i,n}$ de l'autre : en utilisant par ailleurs que les $\widehat{f}_{i,n}$ sont des densités de probabilité, on écrit

$$\begin{aligned} L\big(\widehat{g}_n\big) - L(g^*) &\leqslant 2 \sum_{i=1}^{M} \int_{\mathbb{R}} \big| p_i f_i - \widehat{p}_{i,n} \widehat{f}_{i,n} \big| \,\mathrm{d}\lambda \\ &\leqslant 2 \sum_{i=1}^{M} \left(\int_{\mathbb{R}} p_i \,\big| f_i - \widehat{f}_{i,n} \big| \,\mathrm{d}\lambda + \int_{\mathbb{R}} \big| p_i - \widehat{p}_{i,n} \big| \widehat{f}_{i,n} \,\mathrm{d}\lambda \right) \\ &\leqslant 2 \sum_{i=1}^{M} \left(\left\| f_i - \widehat{f}_{i,n} \right\|_1 + \big| p_i - \widehat{p}_{i,n} \big| \right) \xrightarrow{\mathbb{P}} 0 \end{aligned}$$

où la convergence en $\mathbb{P}$–probabilité a lieu par consistance de tous les estimateurs.

En conclusion, le problème de reconnaissance de formes a été essentiellement résolu par l'estimation de M densités (les M densités conditionnelles de la loi de X). En général, sans restriction sur la loi ν, il n'est pas possible de préciser les vitesses de convergence vers 0. Cela est illustré dans ce chapitre au paragraphe 32.1.4 et est également reformulé pour le problème de la classification par le Théorème 20.6.

D'autres applications de l'estimation de densité peuvent être trouvées au paragraphe 32.1.3 en annexe.

24.2 Questions possibles, conseils aux étudiants et bibliographie

24.2.1 Questions éventuelles d'un enseignant

Une question classique, mais qui, surtout en fin d'oral, est source de nombreuses erreurs, est la suivante :

> Est-ce qu'une loi de fonction de répartition continue est nécessairement absolument continue par rapport à la mesure de Lebesgue ?

La réponse est non : le contre-exemple est donné par la fonction de Borel (également appelée escalier du diable), qui est la fonction de répartition d'une probabilité dont le support est l'ensemble triadique de Cantor (de mesure de Lebesgue nulle). La bonne hypothèse, c'est que la fonction de répartition doit être dérivable presque partout et être égale à l'intégrale de sa dérivée, c'est-à-dire que de manière équivalente, elle doit être absolument continue, voir le chapitre 26.

Les notes historiques à la fin de [Rud98, chapitre 7] montrent une discussion par Riesz et Nagy de la preuve d'un résultat frappant de Lebesgue : toute fonction monotone est dérivable presque partout. En particulier, toute fonction de répartition F est dérivable presque partout. Mais, encore une fois, F n'est l'intégrale de sa dérivée que si elle est absolument continue (au sens des fonctions réelles), ce qui correspond à l'absolue continuité de la probabilité sous-jacente par rapport à la mesure de Lebesgue.

Un enseignant aura également à cœur de revenir sur les liens entre convergence des densités et convergence en loi, si l'étudiant ne les a pas spontanément traitées (voir la discussion en annexe au paragraphe 32.1.3). Il est à noter que beaucoup d'étudiants ignorent la définition de la convergence en loi, et encore plus les caractérisations données par le lemme de Portmanteau (voir là encore le chapitre 26).

24.2.2 Conseils aux étudiants

Présentation et synthèse des résultats

Lors des oraux blancs d'agrégation que nous avons eu à faire passer sur ce texte, nous avons constaté que peu d'étudiants prenaient le temps de définir le modèle statistique : qui sont les observations, selon quelle loi sont-elles distribuées et dans quel ensemble ($\mathcal{P}$, ici) pense-t-on que vit cette loi ? L'ensemble $\mathcal{P}$ des lois possibles est crucial, sa taille montre la difficulté du problème statistique. En particulier, il est ici non-paramétrique.

Il est important de donner des heuristiques et de préciser les intuitions en jeu. C'est aisé à faire, mais encore faut-il penser à le faire, pour les estimateurs par histogrammes et ceux par fenêtres glissantes (puisque la densité est la dérivée de la fonction de répartition), et un peu plus délicat pour ceux à noyaux. On indiquera aussi pourquoi il est intuitivement nécessaire que $h_n \to 0$ (approximation de la dérivée ou des mesures de Dirac, selon les cas) mais $nh_n \to \infty$ (avoir des fenêtres pas trop petites, de sorte qu'elles contiennent un nombre suffisant de points).

Il est encore plus essentiel, s'agissant d'une épreuve de mathématiques, de bien définir tous les objets que l'on manipule, même et surtout lorsque le texte ne le fait pas. La convolution $\star$

n'est pas une opération suffisamment classique pour qu'on puisse ne pas en rappeler son expression. C'est vrai pour la convolution de deux fonctions, mais cela l'est encore davantage pour celle entre une mesure et une fonction.

Ce qui relève de la définition de distance en variation totale et ce qui a trait à ses propriétés n'est (à dessein) pas clairement distingué dans le texte ; l'exposé doit mettre cela en forme. Il faut également préciser dans quels espaces vivent les objets, en particulier les estimateurs de la densité : il s'agit de vérifier qu'ils définissent bien des densités, cela n'est pas évident *a priori*.

Enfin, comme toujours, il ne faut pas oublier de donner un plan au début de l'exposé, en indiquant clairement les résultats que l'on va prouver. Si l'on a pu préparer beaucoup de choses par ailleurs, il ne faut pas hésiter à prouver très brièvement le théorème de Scheffé en se contenant d'indiquer l'élément-clé consiste à considérer l'ensemble $A = \{f \geqslant g\}$. C'est un excellent compromis entre ne rien dire sur la preuve (ce qui peut être dangereux si l'on dit beaucoup d'erreurs par la suite) et y passer trop de temps, alors que la plupart des étudiants vont la présenter. En outre, cela indique une bonne distance par rapport à la preuve : on en a extrait le substantifique argument.

Deux commentaires plus mathématiques

Les étudiants à l'aise avec les preuves peuvent bien entendu étendre les résultats aux probabilités sur $\mathbb{R}^d$; les conditions $h_n \to 0$ et $nh_n \to \infty$ deviennent alors $h_n \to 0$ et $nh_n^d \to \infty$ (avec toujours la même intuition, qu'il y ait suffisamment de points dans un h_n–voisinage). Il vaut mieux faire des remarques le moment venu sur les extensions que de faire directement tout le traitement dans $\mathbb{R}^d$, c'est plus sûr.

Encore une fois, mais on l'a déjà souligné dans le texte, il ne faut pas hésiter à avancer coûte que coûte, quitte, par exemple pour la preuve de la consistance des estimateurs à noyaux,

- à admettre le résultat d'analyse $\|K_h \star f - f\|_1 \to 0$, surtout si c'est par manque de temps pour le prouver ;
- à faire des hypothèses plus fortes sur les noyaux K considérés (K borné et à support compact, voire de régularité C^∞ : il en existe, voir [Bre99, chapitre IV.4]) ;
- on peut faire également des hypothèses sur f, mais en dernier recours seulement, car on ne peut généralement pas tester si la vraie densité f les vérifie. Prendre f continue nous ramène par exemple aux intégrales de Riemann, ce qui peut aider les candidats peu à l'aise avec la théorie de la mesure et de l'intégration de Lebesgue. On peut aussi vouloir supposer f bornée et expliquer pourquoi c'est raisonnable.

Dans ce corrigé, on a travaillé dans un cadre minimal, sans hypothèse particulière sur K (puisque l'on a indiqué comment l'hypothèse de moment d'ordre deux pouvait être levée) et sans aucune hypothèse sur la densité de probabilité f. En contrepartie, les résultats exhibés restent relativement faibles, comme le souligne notamment le paragraphe 32.1.4 en annexe.

24.2.3 Bibliographie

Les sources principales pour ce texte sont les traités de Luc Devroye et de ses co-auteurs László Györfi et Gábor Lugosi [DG85, Dev87, DL01] (mais le lecteur pourra également se référer

à [Tsy04]). Devroye et ses co-auteurs défendent le critère de la distance $\mathbb{L}^1$ pour évaluer la qualité de l'estimation de la densité, en montrant qu'il est plus intrinsèque que des distances $\mathbb{L}^p$ pour $p > 1$ et notamment, $p = 2$. En effet, une convergence au sens $\mathbb{L}^p$ implique toujours la convergence $\mathbb{L}^1$ (mais la réciproque n'est pas vraie). Cela étant, nous avons vu que certains résultats sont partiellement établis en majorant des écarts en norme $\mathbb{L}^2$, comme par exemple la Proposition 24.3 et le Théorème 24.7. Ils insistent également sur la stabilité du critère $\mathbb{L}^1$ d'écarts entre deux densités par rapport à des transformations sur les lois dont elles issues (voir [Dev87, paragraphe 1.2]), comme le changement d'échelle affine. Enfin, ils soulignent que le critère $\mathbb{L}^1$ est plus graphique, comme nous l'avons indiqué plusieurs fois dans le corrigé, puisqu'il se lit en termes de différences d'aires entre deux courbes.

Bien entendu, le critère $\mathbb{L}^2$ est largement utilisé. Comme nous l'avons souligné dans le chapitre 18, les calculs dans $\mathbb{L}^2$ sont rendus agréables par sa structure d'espace euclidien, qui permet en outre d'établir des résultats plus fins. Les connexions sont d'ailleurs fortes entre le problème de régression non-paramétrique traité dans le chapitre 18 et celui de l'estimation de densité. Dans le cadre $\mathbb{L}^2$, nous renvoyons le lecteur aux méthodes classiques pour l'estimation de densité que l'on trouvera dans [HKPT98, Tsy04, Mas07]. Pour compléter cette bibliographie, citons l'ouvrage fondamental écrit par Bernard Silverman [Sil86] qui traite de l'estimation de densité sous de multiples facettes. Bien qu'ancien, ce livre reste une référence sur ce sujet.

Début 2008, le jury de l'agrégation a rendu public un texte sur l'estimation de densité. Il est sensiblement différent du texte du chapitre 14, que nous avions écrit de manière totalement indépendante. Le texte du jury montre notamment l'optimalité en un sens $\mathbb{L}^2$ des estimateurs à noyaux sur les classes de régularités höldériennes et étudie l'implémentation pratique de ces estimateurs en utilisant une méthode de validation croisée (voir le paragraphe 20.1.11 pour un énoncé de cette dernière). Il est donc parfaitement complémentaire de celui que nous avons proposé ici.

24.3 Commentaires sur les simulations `Matlab`

24.3.1 Architecture générale des programmes

On recourt à trois fonctions de calcul des estimations, `FenGliss` (fenêtres glissantes), `EstNoyau` (par noyaux) et `Histogr` (par histogrammes réguliers). Bien que les estimateurs par fenêtres glissantes soient un cas particulier d'estimateurs par noyaux, nous avons préféré leur réserver une fonction à part. Le noyau d'Epanechnikov est représenté par la fonction `KEpa`. Les données à partir desquelles effectuer les estimations sont simulées par les fonctions `MelUnif` et `MelGauss`. Les densités correspondantes peuvent être représentées avec `dunif` et `dgauss`. Les distances en variation totale entre les densités estimées et la vraie densité sont calculées (de manière approchée, par méthode des trapèzes) par `dvt`. Le script principal `MainDensite` appelle une dernière fonction, `compareGr`, qui effectue, sur un jeu de données artificielles, les tracés des quatre graphiques et le calcul des distances en variation totale associées.

24.3.2 Commentaires sur le code Matlab

On s'arrêtera sur deux fonctions. `trapz` est l'une des nombreuses fonctions d'intégration numérique de `Matlab` (taper `help trapz` et suivre les liens hypertextes pour en connaître d'autres) ; elle implémente la méthode des trapèzes pour l'intégration approchée. Comme toujours, les méthodes mathématiques fondamentales d'approximation sont déjà implémentées dans `Matlab` et permettent de gagner un temps précieux lors de la programmation.

`str2func` permet d'associer à une chaîne de caractères la fonction qui porte le nom donnée par la chaîne. Par exemple, `h = str2func('dnorm')` fait référence à la fonction `dnorm`. On définit ainsi un pointeur `h` vers `dnorm` ; en particulier, `h(0)` retourne la même valeur que `dnorm(0)`, soit environ 0.40. On rappelle que les pointeurs sont donnés par `@` sous `Matlab`, `h` est donc égal à `@dnorm`. La fonction `str2func` est utile dès lors que l'on veut entrer une chaîne de caractères donnant le nom de la fonction, plutôt que directement un pointeur. On pourra également se référer au paragraphe 19.3.2, qui présente également certains des aspects ci-dessus dans le cadre de l'utilisation de `fplot`.

La boîte à outils `StixBox` pour `Matlab` contient une fonction `histo` qui trace de manière intelligente les histogrammes, en déterminant la taille de fenêtre h_n en fonction du nombre de données. (Elle prend par défaut $h_n = 4\lceil n^{-1/4} \rceil$.) En outre, elle permet d'assurer que l'aire sous l'histogramme est 1. C'est donc une implémentation efficace de la procédure d'estimation consistante par histogrammes. Elle évite, contrairement à l'utilisation de `hist`, de devoir effectuer soi-même les renormalisations à 1. Quand on a affaire à des lois discrètes, on recourt plutôt aux fonctions `bar` ou `stem`.

On peut visualiser les performances de `histo` avec le code suivant, qui produit un film (une suite d'images avec effet d'animation) montrant la convergence des estimateurs par histogrammes vers la densité sous-jacente normale standard :

```
M = [];
X = randn(1,10^6);
for i = 1:6
  clf
  hold on
  axis([-6 6 0 .45]);
  histo(X(1:10^i), [], [], 1);
  fplot('dnorm', [-6 6], 'r');
  M = [M; getframe];
end
movie(M, [], 1)
```

L'effet est surprenant !

24.3.3 Codes des simulations

Ils sont disponibles en annexe, au chapitre 32, page 473.

CHAPITRE 25

Jeux de grattage

25.1 Corrigé possible du texte

25.1.1 Construction d'un modèle statistique général

Dans ce chapitre, on considère deux modèles. Un modèle probabiliste, que nous présenterons dans un second temps, sera celui de répartition uniformément au hasard des lots. Un modèle statistique formera le modèle général, où l'on ne suppose rien de l'uniformité ou de la non-uniformité de cette répartition. Parce qu'ici il est plus facile d'aller du général au particulier, nous construisons d'abord le modèle général (statistique). Nous spécifierons ensuite le modèle particulier (probabiliste) de répartition uniforme.

On part de 100 observations $c_1, \dots, c_{100}$, représentant le nombre de lots significatifs obtenus dans 100 carnets achetés à des moments aléatoires d'une année. Leurs valeurs sont

$$\Big|\big\{j \in [\![1, 100]\!] : \ c_j = 0\big\}\Big| = 33 \qquad \text{et} \qquad \Big|\big\{j \in [\![1, 100]\!] : \ c_j = 1\big\}\Big| = 67 \, .$$

On modélise maintenant ces observations comme la réalisation d'un vecteur aléatoire $C = (C_1, \dots, C_{100})$.

Loi du nombre de lots C_j dans un seul carnet

A cet effet, il faut expliquer davantage comment un carnet se retrouve chez un buraliste : on suppose que la société de jeux conçoit d'un coup l'ensemble d'une tranche d'émission, i.e., 10 000 carnets, contenant exactement les lots discutés au tableau 17.1. Ces carnets sont alors répartis uniformément au hasard entre les buralistes du pays. Même en achetant un carnet chez un buraliste fixé, on prélève donc un carnet uniformément au hasard dans une tranche d'émission.

On suppose également pour le moment que chaque tranche comporte les mêmes proportions de carnets contenant 0, 1, 2, etc., et au plus 50 lots significatifs. (Comme le carnet contient 50 tickets, il ne peut y avoir plus de 50 lots significatifs.) On étudiera le cas général de proportions aléatoires ci-dessous. On note le vecteur de ces proportions déterministes par ν, qui est une probabilité sur $[\![0, 50]\!]$; il est évidemment inconnu et on cherche à l'inférer.

Avec ces notations et hypothèses, lorsque l'on prélève un carnet j au hasard dans une tranche, son nombre de lots significatifs C_j est distribué selon ν (où l'aléa est engendré uniquement par le choix aléatoire d'un élément dans la population donnée par l'ensemble des carnets d'une tranche).

Loi du vecteur de nombre de lots C

On a supposé que pour ce jeu, 10 tranches d'émission étaient réalisées par an et que l'on prélevait 100 carnets, de manière étalée dans le temps. Cela assure qu'on prélève peu de carnets par tranche, de l'ordre d'une dizaine, à comparer à l'ensemble des 10 000 carnets de la tranche. Comme on le rappelle ci-dessous, on peut donc approximer ces tirages, *a priori* sans remise, par des tirages avec remise.

En conséquence, les observations C_j sont indépendantes et identiquement distribuées : le vecteur aléatoire $C = (C_1, \ldots, C_{100})$ est un 100–échantillon de loi commune ν.

> Remarque au passage. On définit alors le vecteur du décompte des occurrences $N = (N_0, \ldots, N_{50})$ de la manière suivante : pour tout $k \in [\![0, 50]\!]$,
>
> $$N_k = \Big|\big\{j \in [\![1, 100]\!] :\ C_j = k\big\}\Big| .$$
>
> Comme on le rappelle ci-dessous, par définition, la loi de N est alors la loi multinomiale $\mathcal{M}(100, \nu)$ de paramètres 100 et ν. Elle charge uniquement les 51–uplets $(n_0, \ldots, n_{50})$ d'entiers tels que $n_0 + \ldots + n_{50} = 100$, et pour un tel 51–uplet,
>
> $$\mathbb{P}\big\{(N_0, \ldots, N_{50}) = (n_0, \ldots, n_{50})\big\} = \frac{50!}{n_0! \ldots n_{50}!} \prod_{j=0}^{50} \nu_j^{n_j} .$$
>
> La loi multinomiale est la généralisation de la loi binomiale au décompte d'occurrences dans un échantillon dont les éléments peuvent prendre plus de deux valeurs différentes. Pour $k \in [\![0, 50]\!]$ fixé, la loi de N_k est la loi binomiale $\mathrm{Bin}(100, \nu_k)$ de paramètres 100 et ν_k.

Ici, on a donc affaire à un modèle paramétrique, puisque l'ensemble des lois possibles (l'ensemble des ν possibles) est formé par le simplexe des probabilités de $\mathbb{R}^{51}$, noté $\mathcal{P}\big([\![0, 50]\!]\big)$:

$$\mathcal{P}\big([\![0, 50]\!]\big) = \left\{\nu = (\nu_0, \ldots, \nu_{50}) \in \mathbb{R}^{51} :\quad \sum_{j=0}^{50} \nu_j = 1 \quad \text{et} \quad \nu_i \geqslant 0 \text{ pour tout } i \in [\![0,\ 50]\!]\right\} .$$

Pour $\nu \in \mathcal{P}\big([\![0, 50]\!]\big)$ et $j \in [\![0,\ 50]\!]$, on notera indifféremment la probabilité que ν associe à j par ν_j ou $\nu\{j\}$. Le contexte (la considération de ν comme un vecteur ou comme une probabilité) déterminera quelle est l'écriture naturelle.

Il y a cinquante paramètres, ce qui constitue un nombre fini mais plutôt élevé. Par ailleurs, même si l'on ne rapporte pas d'observation c_j ayant une valeur plus grande que 2, on ne peut pas encore, à ce moment de la modélisation, partir du principe que ν met une probabilité nulle à l'ensemble $[\![3, 50]\!]$ et que le modèle est donc décrit par trois paramètres uniquement. Cela viendra ultérieurement, au paragraphe 25.1.3 et sous une hypothèse supplémentaire à discuter.

Retour sur l'approximation d'un tirage sans remise par un tirage avec remise

On commence par définir les lois associées respectivement aux tirages avec et sans remise de m éléments parmi M. On suppose que les observations peuvent prendre ici k modalités dont les valeurs sont représentées par $\{1, \ldots, k\}$ et qu'il y a M_j éléments parmi les M pour la modalité j. Ainsi, $M = M_1 + \ldots + M_k$.

On note $X_1, \ldots, X_m$ les résultats des tirages, avec, pour $t = 1, \ldots, m$, les valeurs possibles $X_t \in \{1, \ldots, k\}$. On définit le vecteur des occurrences $N = (N_1, \ldots, N_k)$ de la manière suivante : pour tout $j \in \{1, \ldots, k\}$,

$$N_j = \sum_{t=1}^{m} \mathbb{I}_{\{X_t = j\}} \, .$$

On montre par des arguments élémentaires de dénombrement les deux résultats suivants. Pour deux entiers $0 \leqslant r \leqslant s$, on notera

$$\binom{s}{r} = \frac{s!}{r!(s-r)!}$$

le coefficient binomial représentant le nombre de choix de r éléments parmi s.

Définition–Théorème 25.1 Lorsque dans l'expérience définie ci-dessus, les tirages s'effectuent avec remise, la loi de N est appelée la loi multinomiale de paramètres m et

$$\boldsymbol{p} = (p_1, \ldots, p_k) = \left(\frac{M_1}{M}, \ \ldots, \ \frac{M_k}{M} \right) \, .$$

On la note $\mathcal{M}(m, \boldsymbol{p})$. Elle admet pour support les k–uplets d'entiers $(\ell_1, \ldots, \ell_k)$ tels que $\ell_1 + \ldots + \ell_k = m$. Elle affecte une probabilité

$$\frac{m!}{\ell_1! \ldots \ell_k!} \ p_1^{\ell_1} \ \ldots \ p_k^{\ell_k}$$

à un tel k–uplet.

Définition–Théorème 25.2 Lorsque dans l'expérience définie ci-dessus, les tirages s'effectuent sans remise, la loi de N est appelée la loi hypergéométrique de paramètres m et $(M_1, \ldots, M_k)$. On la note $\mathcal{H}\big(m, \ (M_1, \ldots, M_k)\big)$. Elle admet pour support les k–uplets d'entiers $(\ell_1, \ldots, \ell_k)$ tels que $\ell_1 + \ldots + \ell_k = m$. Elle affecte une probabilité

$$\frac{\displaystyle\prod_{j=1}^{k} \binom{M_j}{\ell_j}}{\displaystyle\binom{M_1 + \ldots + M_k}{m}}$$

à un tel k–uplet.

Le résultat de convergence ci-dessous justifie l'approximation précédemment considérée d'une expérience de tirages de carnets sans remise par des tirages de carnets avec remise.

Lemme 25.3 Lorsque $M_j \to \infty$ pour tout $j \in \{1, \ldots, k\}$ de sorte qu'il existe un vecteur de probabilité $\boldsymbol{p} = (p_1, \ldots, p_k)$ avec $p_j > 0$ et

$$\frac{M_j}{M_1 + \ldots + M_k} \longrightarrow p_j \tag{25.1}$$

pour tout $j \in \{1, \ldots, k\}$, alors

$$\mathcal{H}\big(m,\ (M_1, \ldots, M_k)\big) \rightsquigarrow \mathcal{M}(m, \boldsymbol{p})\ .$$

Preuve Il s'agit de prouver que pour tout k–uplet d'entiers $(\ell_1, \ldots, \ell_k)$ tel que $\ell_1 + \ldots + \ell_k = m$,

$$\frac{\displaystyle\prod_{j=1}^{k} \binom{M_j}{\ell_j}}{\displaystyle\binom{M_1 + \ldots + M_k}{m}} \longrightarrow \frac{m!}{\ell_1! \ldots \ell_k!}\ p_1^{\ell_1} \ldots p_k^{\ell_k}\ .$$

Or, on note que puisque les $M_j \to \infty$, on a les équivalences

$$\binom{M_j}{\ell_j} = \frac{M_j!}{\ell_j!(M_j - \ell_j)!} \sim \frac{M_j^{\ell_j}}{\ell_j!} \quad \text{et} \quad \binom{M_1 + \ldots + M_k}{m} \sim \frac{(M_1 + \ldots + M_k)^m}{m!}\ .$$

La preuve est conclue en injectant les équivalences données par (25.1). □

Extension à des proportions aléatoires dans une tranche

On discute ici le cas où les proportions ne sont plus déterministes et varient d'une tranche à l'autre, selon un certain aléa. C'est le cas notamment lorsque la génération de la position des lots est faite selon un hasard total, ainsi qu'expliqué au paragraphe 25.1.2.

Dans ce cas, on note $\mu_1, \mu_2, \ldots, \mu_{10}$ les proportions ν correspondant aux dix émissions dans lesquelles on a prélevé des carnets. On suppose que ces proportions sont les réalisations de variables aléatoires indépendantes et identiquement distribuées selon une loi $\mathcal{V}$ sur $\mathcal{P}\big(\llbracket 0, 50 \rrbracket\big)$. En effet, cela revient simplement à supposer qu'elles ont été engendrées par le même processus aléatoire de fabrication. Nous donnons un exemple d'une telle situation au paragraphe 25.1.2.

Lorsque l'on s'intéresse au nombre C_j de lots significatifs contenus dans un seul carnet, on peut imaginer que le tirage de ce nombre est effectué en deux temps : on tire d'abord la valeur du vecteur de proportions μ selon $\mathcal{V}$, puis on tire C_j selon μ. Il est équivalent d'effectuer directement un tirage de C_j selon la loi

$$\nu = \int_{\mathcal{P}\left(\llbracket 0,50 \rrbracket\right)} \mu\, \mathrm{d}\mathcal{V}(\mu)\ . \tag{25.2}$$

Cela montre que la modélisation précédente, introduite pour des proportions déterministes, est encore valable dans ce cadre de proportions aléatoires, au sens où C est bien encore un échantillon de loi ν. On prendra cependant garde qu'on fera œuvre d'inférence ici sur ν et non sur $\mathcal{V}$. Dit autrement, le vrai modèle sous-jacent (décrit par $\mathcal{V}$) est non-paramétrique et nous ne nous intéressons qu'à ν, qui est une loi image de $\mathcal{V}$.

25.1.2 Construction et rejet d'un modèle probabiliste de référence : répartition totalement aléatoire des lots significatifs

Ici, on définit un cas particulier du modèle statistique général exhibé précédemment et qui nous servira de modèle de référence : on étudie la méthode de génération $\mathcal{V}_{\text{unif}}$ qui place les lots significatifs totalement au hasard dans la tranche d'émission.

Dans un premier temps, on construit précisément ce modèle et on détermine le vecteur ν associé selon la formule (25.2). Dans un second temps, on montre que les observations recueillies $(c_1, \ldots, c_{100})$ contredisent ce modèle (au sens des tests statistiques).

Construction du modèle probabiliste de répartition totalement aléatoire

On suppose ici que l'ensemble des lots (et des tickets perdants) est obtenu par considération d'un arrangement aléatoire des lots mineurs et significatifs décrits au tableau 17.1 parmi les 500 000 tickets d'une tranche. Cela définit les gains ou absences de gains $T_1, \ldots, T_{500\,000}$ portés par les tickets. Ces derniers sont alors séquentiellement groupés 50 par 50 pour former les carnets. Ainsi, pour $1 \leqslant k \leqslant 10\,000$, le k–ième carnet de la tranche contient les lots $T_{50(k-1)+1}, \ldots, T_{50k}$.

Cette procédure décrit en particulier la loi $\mathcal{V}_{\text{unif}}$ dont il a été question plus haut. Nous calculons maintenant la loi ν_{ref} induite selon la formule (25.2), ou tout du moins, dans un premier temps, une bonne approximation. Cette loi formera la loi de référence lors de la construction d'un test ci-dessous.

La méthode de génération selon $\mathcal{V}_{\text{unif}}$ est équivalente au fait de tirer séquentiellement et sans remise les valeurs $T_1, \ldots, T_{500\,000}$. On s'intéresse alors à la loi du nombre de lots significatifs présents dans le premier carnet : par symétrie, sa loi est identique à celle du nombre de lots contenus dans tout autre carnet et est donc égale à ν.

Or, comme il découle de la Définition–théorème 25.2, ce nombre suit la loi ν_{ref} donnée par la première marginale de $\mathcal{H}\big(50,\ (7\,697,\ 500\,000 - 7\,697)\big)$. Selon les mêmes arguments que précédemment (taille d'échantillon prélevé faible par rapport à la population), ce tirage peut être approximé par un tirage avec remise, voir le Lemme 25.3. Ainsi,

$$\mathcal{H}\big(50,\ (7\,697,\ 500\,000 - 7\,697)\big) \approx \mathcal{M}\big(50,\ (p_0, 1 - p_0)\big)$$

de sorte que $\nu_{\text{ref}} \approx \text{Bin}(50, p_0)$, où l'on a noté $p_0 = 7\,697/500\,000$ la proportion de lots significatifs dans une tranche d'émission. Un petit abus (qu'on quantifiera par le calcul) nous conduira à écrire dans la suite $\nu_{\text{ref}} = \text{Bin}(50, p_0)$.

$\mathbb{P}\{\text{Bin}(50,\ p_0) = 0\}$	$\mathbb{P}\{\text{Bin}(50,\ p_0) = 1\}$	$\mathbb{P}\{\text{Bin}(50,\ p_0) \geqslant 2\}$
0.4604	0.3599	0.1797

TABLE 25.1. Probabilités de quelques événements sous la loi $\nu_{\text{ref}} = \text{Bin}(50,\ p_0)$, arrondies à 10^{-4} près.

Quelle que soit la méthode de génération $\mathcal{V}$ des carnets, un ticket pris au hasard dans un carnet pris au hasard porte un lot significatif avec probabilité $p_0 = 7\,697/500\,000$, puisqu'il y a 7 697 lots significatifs dans une tranche de 500 000 tickets. On vient de prouver que la méthode de génération aléatoire $\mathcal{V}_{\text{unif}}$ étend en un sens ce fait à l'ensemble des tickets d'un carnet donné, car le nombre de lots significatifs dans un carnet est donné par la loi binomiale $\nu_{\text{ref}} = \text{Bin}(50,\ p_0)$: autrement dit, il y a indépendance entre les gains portés par des tickets dans un carnet. C'était bien sûr attendu au vu de la méthode de génération (et au vu de l'approximation des tirages sans remise par des tirages avec remise).

Test d'ajustement au modèle de répartition totalement aléatoire

On souhaite tester ici si l'allocation des lots significatifs par carnet est réalisée par la société de jeux selon le modèle de répartition totalement aléatoire (c'est notre hypothèse H_0) ou si des contraintes sont imposées sur leur répartition par carnet (c'est l'hypothèse alternative H_1). Ces hypothèses se traduisent mathématiquement, avec les notations précédentes, par

$$H_0 : \nu = \nu_{\text{ref}} \big(= \text{Bin}(50,\ p_0)\big) \qquad \text{contre} \qquad H_1 : \nu \neq \nu_{\text{ref}}\,.$$

En particulier, sous H_0, les nombres de lots significatifs observés $C = (C_1, \ldots, C_{100})$ forment un 100–échantillon de loi $\text{Bin}(50,\ p_0)$.

La probabilité de quelques événements sous la loi $\text{Bin}(50,\ p_0)$ est rappelée dans le tableau 25.1. Il y a en particulier une probabilité environ égale à 18 % qu'une variable aléatoire de loi $\text{Bin}(50,\ p_0)$ prenne des valeurs supérieures ou égales à 2. On note

$$S_j = \mathbb{I}_{\{C_j \geqslant 2\}}\,, \qquad \text{pour } \ j = 1, \ldots, 100.$$

Ces variables aléatoires indiquent s'il y a ou non au moins deux lots significatifs dans un carnet. Sous H_0, on attend donc, par la loi des grands nombres :

$$\overline{S}_{100} = \frac{S_1 + \ldots + S_{100}}{100} \approx 0.18\,.$$

Or, sur les observations, la réalisation $\overline{s}_{100}$ de $\overline{S}_{100}$ vaut $\overline{s}_{100} = 0$. On soupçonne que H_0 est infondée et l'on va maintenant quantifier mathématiquement ce sentiment, en exhibant une p–valeur.

D'une manière générale (i.e., sous H_0 comme sous H_1), les variables aléatoires $S_1, \ldots, S_{100}$ sont indépendantes et identiquement distribuées selon une loi de Bernoulli $\mathrm{Ber}(q)$ de paramètre

$$q = \nu\big([\![2,\ 50]\!]\big) .$$

Sous H_0, on a

$$q = q_{\text{ref}} \overset{\text{not.}}{=} \mathbb{P}\big\{\mathrm{Bin}(50,\ p_0) \geqslant 2\big\} \approx 0.18 .$$

Un test de H_0 et H_1 est donc donné en particulier par un test de

$$H_0' : q = q_{\text{ref}} \qquad \text{contre} \qquad H_1' : q < q_{\text{ref}} .$$

On a pris ici un test unilatère parce que l'on suppose que le détective avait recueilli au préalable des témoignages de buralistes qui laissaient entendre qu'ils observaient rarement, voire jamais, l'obtention de deux lots significatifs dans un carnet. Bien entendu, si cette information externe n'avait pas été disponible, il aurait fallu mener un test bilatère.

Les tests de conformité pour le paramètre d'une loi de Bernoulli ont été vus notamment au paragraphe 4.4. Ils utilisent comme statistique de test la moyenne empirique correctement centrée et renormalisée par le paramètre à tester, ici, q_{ref}. On les met en œuvre au paragraphe suivant.

On va en effet effectuer d'abord un calcul direct, pour changer et parce que l'on a accès à une détermination informatique aisée des quantiles. Ce calcul direct met en jeu une statistique similaire, la somme $S = S_1 + \ldots + S_{100}$, dont on sait qu'elle suit la loi $\mathrm{Bin}(100, q)$; en particulier, sous H_0, elle suit la loi $\mathrm{Bin}(100, q_{\text{ref}})$. Vu la forme des hypothèses, un test de niveau $\alpha \in]0,\ 1[$ rejette H_0 lorsque S est trop petite, *id est*, utilise une zone de rejet sur S de la forme $[\![0,\ r_\alpha]\!]$.

Le seuil r_α est fixé le plus grand possible de telle sorte que l'erreur de première espèce soit inférieure ou égale à α, soit ici,

$$\mathbb{P}_{H_0}\left\{S \in [\![0,\ r_\alpha]\!]\right\} = \mathbb{P}\left\{\mathrm{Bin}(100,\ q_{\text{ref}}) \leqslant r_\alpha\right\} \leqslant \alpha .$$

Le tableau 25.2 permet de calculer r_α pour certaines valeurs de α. Par exemple, pour les niveaux usuels $\alpha = 5\ \%$ et $\alpha = 1\ \%$, on a $r_{5\%} = 11$ et $r_{1\%} = 9$; on a également, entre autres, que $r_{0.05\,\%} = 6$. On a même $r_\alpha = 1$ pour $3 \times 10^{-9} \leqslant \alpha \leqslant 6 \times 10^{-8}$.

Or, les observations sont $s_1 = \ldots = s_{100} = 0$, de sorte que la réalisation s de S vaut $s = 0$. Le test précédent rejette ainsi H_0 à tout niveau α, pour $\alpha \geqslant 3 \times 10^{-9}$. Ce n'est que lorsque l'on s'oblige à avoir une erreur de première espèce inférieure à 3×10^{-9} que le test ne rejette plus H_0. Cette valeur seuil est appelée la *p*–valeur du test, voir le paragraphe 4.2 ; elle est ici très inférieure aux seuils usuels de 5 % et 1 %.

Conclusion intermédiaire

Sans aucun doute, on rejette H_0 et on peut affirmer que cette dernière contredit gravement les données recueillies : le paramètre q est significativement plus petit que la valeur $q_{\text{ref}} = 18\ \%$.

k	12	11	10	9
$\mathbb{P}\{\mathrm{Bin}(100,\ q_{\mathrm{ref}}) \leqslant k\}$	7×10^{-2}	4×10^{-2}	2×10^{-2}	10^{-2}

k	6	3	1	0
$\mathbb{P}\{\mathrm{Bin}(100,\ q_{\mathrm{ref}}) \leqslant k\}$	5×10^{-4}	5×10^{-6}	6×10^{-8}	3×10^{-9}

TABLE 25.2. Fonction de répartition de la loi $\mathrm{Bin}(100,\ q_{\mathrm{ref}})$; valeurs données avec un seul chiffre significatif.

Si l'on veut reproduire l'expérience (par exemple, pour faire foi, sous contrôle d'huissier), faut-il vraiment acheter 100 carnets, et combien cela coûtera-t-il (hors frais d'huissier) ? Cela revient à se demander quel est le plus petit n tel qu'une réalisation égale à 0 pour $S_1 + \ldots + S_n$ entraîne le rejet de H_0, avec une erreur de première espèce contrôlée, disons, par le seuil très faible 1 ‰. Avec le même raisonnement que ci-dessus, cela consiste à déterminer le plus petit n tel que

$$\mathbb{P}\{\mathrm{Bin}(n,\ q_0) = 0\} \leqslant 1\,‰\,.$$

Un calcul informatique montre qu'il s'agit de $n = 35$. Or, il est facile et bon marché d'acheter 35 carnets (dans plusieurs tranches d'émission, pour conserver la qualité de l'approximation de tirage avec remise, donc, pas tous à la fois). En effet, le paragraphe 25.1.4 indique que l'on peut compter sur un solde (différence entre le gain procuré par le ticket et la mise de 3 €) typique moyen de l'ordre de -1.15 € par ticket. L'achat d'un carnet entier revient donc environ à 57.50 €, et celui de 35 carnets, à environ 2 000 €. C'est un budget raisonnable pour un enquêteur (détective ou journaliste). Pour une garantie théorique moindre, par exemple une majoration de l'erreur de première espèce par le seuil typique de 5 %, un nombre $n = 16$ d'achats de carnets suffit.

Variante du test précédent avec davantage d'approximations

A titre informatif, on regarde ce qu'aurait donné le test asymptotique du paragraphe 4.4 : sa considération consiste essentiellement à ne plus utiliser les quantiles calculés informatiquement et reportés dans le tableau 25.2, mais à prendre à la place les quantiles de la loi normale. On utilise la convergence

$$\frac{\sqrt{n}}{\sqrt{q(1-q)}}\left(\overline{S}_n - q\right) \rightsquigarrow \mathcal{N}(0,1)$$

pour tout $q \in]0, 1[$, assurée par le théorème de la limite centrale. En particulier, sous H_0,

$$\frac{\sqrt{n}}{\sqrt{q_{\text{ref}}(1-q_{\text{ref}})}} \left(\overline{S}_n - q_{\text{ref}} \right) \rightsquigarrow \mathcal{N}(0, 1) ,$$

où l'on note par ailleurs que la statistique donnée par le membre de gauche a tendance à être plus petite sous H_1 que sous H_0. Un test asymptotiquement de taille $\alpha \in]0, 1[$ pour H_0 et H_1 est donc donné par

$$\phi(C) = \mathbb{I}_{\left\{ \overline{S}_n < q_{\text{ref}} - z_{1-\alpha} \sqrt{q_{\text{ref}}(1-q_{\text{ref}})/n} \right\}} .$$

Ici, on effectue l'hypothèse que l'approximation donnée par la convergence en loi est déjà bonne lorsque $n = 100$. La détermination de la p–valeur au vu des observations est alors effectuée selon les indications de la remarque 4.11. Elle vaut

$$p = \mathbb{P}\left\{ N \leqslant \frac{\sqrt{100}}{\sqrt{q_{\text{ref}}(1-q_{\text{ref}})}} \left(\overline{s}_{100} - q_{\text{ref}} \right) \right\} = \Phi(-4.68) = 1.43 \times 10^{-6} ,$$

où l'on a noté Φ la fonction de répartition de la loi normale standard et où l'on a précisé les valeurs numériques avec trois chiffres significatifs uniquement. Elle est supérieure à celle précédemment calculée mais elle conduit toujours à un rejet clair de H_0.

> Remarque au passage. On se gardera cependant de dire que la procédure précédente fondée sur les quantiles de la loi binomiale est plus exacte ou plus satisfaisante. Toute modélisation comporte une part d'imprécision ou d'erreur et il est difficile de quantifier quelle modélisation est la moins mauvaise face à un problème concret. On gardera cela en mémoire en lisant les autres variantes discutées ci-dessous.

Une autre approximation possible dans la procédure fondée sur les lois binomiales aurait été de continuer la chaîne d'approximations pour la loi de référence ν_{ref} selon

$$\nu_{\text{ref}} \approx \text{Bin}(50, p_0) \approx \mathcal{P}(50\, p_0) ,$$

en effectuant l'approximation dite binomiale–Poisson. On rappelle que cette dernière est justifiée par le résultat suivant.

Lemme 25.4 Soit (p_n) une suite dans $[0, 1]$ telle que $np_n \to \lambda$, pour un réel $\lambda > 0$. Alors,

$$\text{Bin}(n, p_n) \rightsquigarrow \mathcal{P}(\lambda) .$$

Des conditions empiriques assurent la validité de l'approximation $\text{Bin}(n, p) \approx \mathcal{P}(np)$ correspondante : que $n \geqslant 30$ et $np \in [1, 10]$. Ici, $n = 50$ et $\lambda_{\text{ref}} = 50\, p_0 = 0.7697$, de sorte que les conditions sont presque respectées.

On a besoin, pour mettre en œuvre le test discuté plus haut, de la valeur de $q_{\text{ref}} = \nu\big([\![2,\ 50]\!]\big)$, que l'on se propose d'approcher par

$$q_{\text{ref}} \approx \mathbb{P}\{P \geqslant 2\} = 1 - \Big(e^{-\lambda_{\text{ref}}} + e^{-\lambda_{\text{ref}}}\lambda_{\text{ref}}\Big) = 0.1804 \qquad \text{où} \qquad P \sim \mathcal{P}(50\, p_0)\ ,$$

à comparer à la valeur 0.1797 proposée par le tableau 25.1. La différence ténue entre ces deux valeurs conduit aux mêmes ordres de grandeur pour la p–valeur associée au test et aux observations.

Remarque 25.5 L'intérêt des deux approximations présentées (utilisation des quantiles de la loi normale, approximation binomiale–Poisson) est qu'avec elles, une calculatrice ordinaire et une table de la loi normale suffisent à réaliser le test. On n'a alors pas besoin de recourir à un logiciel de statistique comme `Matlab`.

Variante du test précédent avec moins d'approximations

Etant donné le modèle idéal décrit par la génération $\mathcal{V}_{\text{unif}}$, on peut exhiber un test exact, donné par des considérations de dénombrement. Nous allons ici ne faire qu'une allusion brève à sa mise en œuvre et nous concentrer sur le calcul de sa p–valeur. On l'effectue sous l'hypothèse (qui précise la réalité) qu'on a tiré exactement 10 carnets dans 10 tranches d'émission successives.

Comme précédemment, dans notre cadre et au vu de nos observations, la p–valeur est donnée par la probabilité qu'avait de se produire le fait que dans tous les carnets tirés, on ait eu aucun ou un seul lot significatif. Soit o la probabilité que dans une tranche d'émission, les 10 carnets tirés aient chacun contenu au plus un lot significatif : on a alors $p = o^{10}$.

On calcule maintenant o de manière exacte, en recourant à des lois hypergéométriques. Vu la symétrie du processus de génération, la probabilité o est la probabilité des enchaînements suivants :

- on tire sans remise 50 tickets parmi les 500 000 tickets et il y en a $k_1 \in \{0, 1\}$ contenant un lot significatif ;
- on tire sans remise 50 tickets parmi les $500\,000 - 50$ tickets restants (dont $7\,697 - k_1$ portent un lot significatif) et parmi eux, $k_2 \in \{0, 1\}$ portent un lot significatif ;
- plus généralement, pour $t \in [\![2, 10]\!]$, on tire sans remise 50 tickets parmi les $500\,000 - 50(t-1)$ tickets restants (dont $7\,697 - (k_1 + \ldots + k_{t-1})$ portent un lot significatif) et parmi eux, $k_t \in \{0, 1\}$ portent un lot significatif.

Soit $h(k, 50, M', M)$ la probabilité qu'une variable aléatoire de loi $\mathcal{H}\big(50,\ (M', M)\big)$ vale $(k, 50-k)$. On réécrit alors o comme

$$o = \prod_{(k_1,\ldots,k_{10})\in\{0,1\}^{10}} \ \prod_{t=1}^{10} h\big(k_t,\ 50,\ 7\,697 - (k_1 + \ldots + k_{t-1}),\ 500\,000 - 50(t-1)\big)\ .$$

Ainsi formulée, la probabilité o est calculable par un logiciel informatique aux problèmes de précision près et bien que cela requière de l'ordre de 2^{10}, donc plusieurs milliers, de calculs. Les

méthodes précédentes consistaient à dire que les termes h dans le produit ci-dessus définissant o ne dépendaient que de leur premier argument k_t et non des suivants : c'était l'approximation de la loi hypergéométrique par la loi binomiale. On pourrait également déterminer par le calcul un majorant commun à tous ces termes. Dans les deux cas, le lecteur se persuadera avec l'utilisation de la fonction `dhypg` de la boîte à outils `StixBox` de `Matlab` (voir le paragraphe 25.3.2) que l'ordre de grandeur 10^{-9} pour la p–valeur est préservé.

Nous ne recommandons pas ce schéma exact, qui effectue en réalité un dénombrement efficace des arrangements des lots significatifs parmi les tickets tels que les dix premiers carnets ainsi construits ne contiennent pas plus d'un lot significatif. Il présente donc un argument combinatoire. Nous trouvons, mais c'est là une question de goût, que le schéma par approximation binomiale a l'avantage de nous ramener à un cadre parfaitement stochastique. Par ailleurs, et c'est une remarque plus objective, s'il est facile de calculer ici la p–valeur au vu des observations, il ne serait pas du tout aisé de préciser de manière un peu automatique un test de taille α, pour toute valeur α.

25.1.3 Inférence sur ν

Majoration de la probabilité q d'avoir deux lots significatifs dans un carnet tiré au hasard

Les tests du paragraphe précédent nous ont appris quelque chose, puisque tous, et sans hésitation, ont rejeté l'hypothèse H_0. On sait comment il ne faut pas modéliser la répartition ν : elle n'est pas issue du seul hasard $\mathcal{V}_{\text{unif}}$. Cela ne nous donne pas encore la loi $\mathcal{V}$ de son engendrement. En particulier, le paramètre

$$q = \nu\big([\![2,\ 50]\!]\big)$$

est certes significativement plus petit que $q_{\text{ref}} \approx 18\,\%$, mais on voudrait quantifier cette assertion et en donner un majorant précis.

On utilise à cet effet la dualité entre tests et intervalles de confiance, vue au paragraphe 4.4. L'intervalle de confiance (au niveau $1-\alpha$) sur le paramètre q est donné par tous les paramètres q' qui seraient acceptés par un test unilatère donné de

$$H_0 : q = q' \qquad \text{contre} \qquad H_1 : q < q' \,,$$

par exemple celui utilisant la statistique $S = S_1 + \ldots + S_{100}$, où $S \sim \text{Bin}(100,\ q')$ sous H_0.

Le même raisonnement qu'au paragraphe précédent montre que H_0 est rejetée au vu des observations $c_1, \ldots, c_{100}$ et au niveau α lorsque

$$\mathbb{P}\big\{\text{Bin}(100,\ q') = 0\big\} = (1-q')^{100} < \alpha \,,$$

soit lorsque

$$q' > q_\alpha = 1 - \exp\left(\frac{\log \alpha}{100}\right) .$$

La réalisation de l'intervalle de confiance au niveau $1-\alpha$ sur q est alors $[0,\ q_\alpha]$. Pour l'erreur de première espèce typique $\alpha = 5\,\%$, on a la réalisation $q_{5\%} = 0.030 \approx 3\,\%$.

Remarque au passage. On ne peut évidemment jamais montrer par analyse statistique que cette valeur de q est nulle. Même avec davantage d'observations rapportant qu'aucun carnet gratté ne contenait strictement plus d'un lot significatif, l'intervalle de confiance se rétrécirait autour de 0, à vitesse $1/\sqrt{n}$, où n est le nombre de telles observations, mais ne sera jamais réduit au singleton $\{0\}$. Cependant, la borne de 3 % exhibée ci-dessus est à apprécier à l'aune de la valeur $q_{\text{ref}} \approx 18\,\%$ correspondant au modèle de répartition totalement aléatoire.

Nécessité (au vu des observations) de l'existence de carnets contenant au moins deux lots significatifs

On montre ici qu'il est improbable, au vu des observations, que la valeur $q_{\text{ref}} = 0$ convienne pour la probabilité q qu'un carnet tiré au hasard contienne deux lots significatifs ou plus. A cet effet, on va construire un test ϕ de

$$H_0 : q = 0 \qquad \text{contre} \qquad H_1 : q > 0\,.$$

Dans la suite, on regardera essentiellement l'image suivante de ν,

$$(a, b, q) = \Big(\nu\{0\},\ \nu\{1\},\ \nu\big([\![0, 50]\!]\big)\Big)\,.$$

On voit qu'ici on va donner une place particulière à deux paramètres a et b.

Le cas $q = 0$ est celui où tous les carnets contiennent aucun ou un unique lot significatif. Le tableau 17.1 indique que cela est équivalent à

$$b = \frac{7\,697}{10\,000} \overset{\text{not.}}{=} b_0 \qquad \text{et} \qquad a = 1 - b_0\,.$$

Pour construire le test ϕ, on se ramène alors, par exemple, à tester l'ajustement de a à la valeur de référence $a_{\text{ref}} = 1 - b_0$, i.e., à construire un test ϕ' de

$$H_0' : a = 1 - b_0 \qquad \text{contre} \qquad H_1 : a > 1 - b_0\,.$$

La forme de l'alternative est justifiée par les faits suivants. Le nombre total de lots significatifs étant constant, si des carnets en contiennent deux ou plus (cas $q > 0$), c'est nécessairement qu'une proportion b plus faible de l'ensemble des carnets en contient un exactement. Cette proportion b diffère de la proportion b_0 d'un facteur additif $-2q$ au moins. A cause de la contrainte $a + b + q = 1$, c'est finalement que a est plus grand, d'au moins un facteur additif q. (On considère ici a plutôt que b pour des raisons qui deviendront claires ci-dessous : essentiellement, on veut exhiber un minorant de a par dualité entre tests et intervalles de confiance.)

Le test ϕ est alors donné par $\phi = \phi'$. Comme H_0 implique H_0', la taille de ϕ est plus petite que celle de ϕ' ; en particulier, si ϕ' est de niveau α, alors c'est également le cas pour ϕ.

On note

$$T_j = \mathbb{I}_{\{C_j = 0\}}\,, \qquad \text{pour } j = 1, \ldots, 100,$$

et on procède comme au paragraphe 25.1.2, en considérant cette fois-ci la statistique de test $T = T_1 + \ldots + T_{100}$. Sous H_0, elle suit une loi $\mathrm{Bin}(100,\ 1 - b_0)$ tandis que sous H_1, elle tend à prendre des valeurs plus grandes. L'intervalle de rejet est de la forme $[\![r_\alpha,\ 100]\!]$, où, pour une majoration $\alpha \in\ ]0,\ 1[$ de l'erreur de première espèce, r_α est choisi le plus petit possible tel que

$$\mathbb{P}\big\{\mathrm{Bin}(100,\ 1 - b_0) \geqslant r_\alpha\big\} \leqslant \alpha\ .$$

Par exemple, pour $\alpha = 5\ \%$, on a $r_{5\%} = 31$.

Or, les observations indiquent que la réalisation t de T vaut $t = 33$. On rejette donc H_0' puis H_0 au niveau 5 %. On peut quantifier ce rejet en calculant la p–valeur. La valeur seuil p est égale à

$$p = \mathbb{P}\big\{\mathrm{Bin}(100,\ 1 - b_0) \geqslant 33\big\} \approx 1.5\,\%\ ,$$

la p–valeur est donc de 1.5 %. H_0 est contredite, mais pas trop gravement. On suspecte que la vraie proportion a de carnets ne contenant aucun unique lot significatif est un peu supérieure, mais pas dramatiquement supérieure, à $1 - b_0$.

En fait, en utilisant ici encore la dualité entre tests et intervalles du paragraphe 4.4, on montre que la réalisation d'un intervalle de confiance à 95 % sur a est $[0.252,\ 1]$. En effet, le test de

$$H_0' : a = a' \qquad \text{contre} \qquad H_1 : a > a'$$

fondé sur la statistique T conserve H_0 au vu des observations et au niveau 5 % si et seulement si

$$\mathbb{P}\big\{\mathrm{Bin}(100,\ a') \geqslant 33\big\} > 5\,\%\ ;$$

avec l'aide des fonctions quantiles implémentées dans `Matlab`, on voit que cette inégalité est vérifiée pour les a' tels que $a' \geqslant 0.252 \approx 25\,\%$.

Minoration de q

Il n'est pas facile de minorer directement q par une valeur non triviale, i.e., strictement positive, car les observations ne comportent aucun carnet avec deux lots significatifs ou plus. On va utiliser une méthode indirecte utilisant des encadrements pour a et b, ce qui, combiné à un système de contraintes linéaires que l'on peut écrire sur a, b et q, suffira pour conclure.

On commence par majorer b : cela se fait, comme pour a et q, par dualité entre tests et intervalles. On omet les détails et indique simplement qu'au niveau 5 % l'intervalle est donné par $[0,\ 0.748]$, où cette valeur $0.748 \approx 75\,\%$ est obtenue comme la valeur seuil b' telle que

$$\mathbb{P}\big\{\mathrm{Bin}(100,\ b') \leqslant 67\big\} = 5\,\%\ .$$

On a donc dans un premier temps les trois encadrements suivants, chacun issu de la réalisation d'un intervalle de confiance de niveau 95 %,

$$\left\{\begin{array}{lcl} a & \geqslant & 25.2\,\%\ , \\ b & \leqslant & 74.8\,\%\ , \\ q & \leqslant & 3.0\,\%\ . \end{array}\right.$$

Le système formé par ces derniers est la réalisation d'une région de confiance de niveau au moins 85 % sur (a, b, q).

On veut écrire maintenant un système de contraintes linéaires sur a, b et q. A cet effet, on suppose qu'il n'existe pas de carnet contenant trois lots significatifs ou plus. On a déjà vu, au paragraphe 25.1.2 que la société de jeux émettait des contraintes techniques sur la répartition, puisque le mode de génération selon $\mathcal{V}_{\text{unif}}$ avait été clairement rejeté.

On peut imaginer avec raison qu'elle le fait pour étaler les lots dans le plus grand nombre de carnets possibles et qu'à cause de ce souhait, elle élude également le cas de carnets contenant trois lots significatifs ou plus. Ainsi, on peut réécrire la loi ν comme

$$\nu = (a,\ b,\ q,\ 0,\ \ldots,\ 0)\ .$$

On aboutit alors au système de contraintes

$$\left\{\begin{array}{rcl} a,\ b,\ q & \geqslant & 0\ , \\ a+b+q & = & 1\ , \\ b+2q & = & b_0\ , \end{array}\right.$$

où les deux premières contraintes viennent de ce que ν est une probabilité, et la troisième assure qu'il y a bien le bon nombre de lots significatifs dans une tranche d'émission de 10 000 carnets.

Ce système est équivalent à

$$\left\{\begin{array}{rcl} a,\ b,\ q & \geqslant & 0 \\ q & = & (b_0-1)+a\ , \\ 2q & = & b_0-b\ . \end{array}\right.$$

En le combinant à la réalisation de la région de confiance et en réinjectant les valeurs numériques, on aboutit à l'encadrement de q suivant,

$$2.2\,\% \leqslant q \leqslant 3.0\,\%\ .$$

En fait, ici, on utilise la minoration de a, qui est plus efficace que la majoration de b pour exhiber le minorant sur q. De même, de $a = 1-b_0+q$ pour la majoration et de la minoration initiale issue de la région de confiance, on déduit

$$25.2\,\% \leqslant a \leqslant 26.0\,\%\ ;$$

tandis que de $b = b_0 - 2q$ ou $b = 1-a-q$ et des encadrements précédents vient

$$71.0\,\% \leqslant b \leqslant 72.6\,\%\ .$$

On est ainsi parvenu à

$$\left\{\begin{array}{rcccl} a+b+q & = & 1\ , & & \\ 25.2\,\% & \leqslant & a & \leqslant & 26.0\,\%\ , \\ 71.0\,\% & \leqslant & b & \leqslant & 72.6\,\%\ , \\ 2.2\,\% & \leqslant & q & \leqslant & 3.0\,\%\ . \end{array}\right.$$

L'excellente nouvelle est qu'en particulier il existe une modélisation qui soit en accord avec les données ; en fait, il en existe toute une famille. Pour fixer les idées, et parce qu'il n'y a de toute façon une incertitude qu'à quelques dixièmes de pour cent, on supposera par la suite, pour étudier les espaces d'opportunités, que l'on a affaire à une génération aléatoire sous-jacente telle que

$$(a, b, q) = \big(25.6\,\%,\ 71.8\,\%, 2.6\,\%\big)\ . \tag{25.3}$$

Ce choix est motivé par le fait que la valeur retenue pour chacun des trois paramètres est au centre de l'intervalle exhibé plus haut sur sa vraie valeur.

> Remarque au passage. Le résultat justifié ci-dessus peut sembler refléter la dictature éclairée du niveau de confiance 95 %, à ceci près bien sûr que la réalisation de la région finalement exhibée sur (a, b, q) est issue d'une région de confiance de niveau simplement au moins égal à 85 %. Cette région n'étant pas limitée à un point, on pourrait réduire un peu la valeur de α et obtenir encore des résultats raisonnables. Cela étant, il faut bien reconnaître ici que nous avons joué de chance avec les données pour obtenir un système final donnant une image aussi précise de ν. La pratique statistique est parfois plus rude et peut conduire à des images plus floues des phénomènes en jeu. Par ailleurs, il ne faut pas oublier qu'une fois sur vingt, on doit jouer de malchance avec une région de confiance à 95 %.

25.1.4 Etude de la redistribution des sommes mises en jeu

Les jeux à gratter sont des jeux dits par redistribution : les sommes misées sont rassemblées, une fraction d'entre elles revient aux joueurs sous la forme des lots, et la société de jeux empoche le reste. Dans le cas du jeu considéré ici, le tableau 17.1 indique que 1 020 500 € sont reversés pour un montant total joué (correspondant au prix des tickets) de $500\,000 \times 3 = 1\,500\,000$ €. Cela forme un taux de redistribution de 68 % (et donc un bénéfice brut de 32 %). Du point de vue des joueurs, cela signifie que l'espérance de gain, pour un ticket, soit 3 € misés, est de 2.04 €, d'où un solde moyen égal à -0.96 €. (Un mathématicien commenterait froidement : le prix du frisson du jeu avoisine donc 1 €.)

Mais regardons plus précisément le ressenti des joueurs : l'analyse précédente tient trop compte d'événements rares, comme le fait de gagner le gros lot de 40 000 €, ce qui arrive avec une probabilité très faible, mais change malgré tout significativement l'espérance du gain. Nous expliquons ici pourquoi et comment un joueur typique ne doit pas compter sur des gains supérieurs ou égaux à 500 €. Prenons-en un qui achète par exemple deux tickets par semaine.

On commence par déterminer la loi du gain G d'un ticket pris au hasard. On va appliquer des arguments similaires à ceux du paragraphe 25.1.2. Un ticket acheté même chez un buraliste fixé est tiré au hasard dans la tranche d'émission (enfin, c'est le cas pour peu que le buraliste ne mette pas en œuvre les stratégies de joueur informé dont il sera question au paragraphe 25.1.5). La loi du gain de ce ticket est donc une loi multinomiale $\mathcal{M}(1, T)$ de paramètres 1 et T, où

$$T = \big(T_0,\ T_3,\ T_4,\ T_6,\ \ldots,\ T_{40\,000}\big)$$

est un vecteur de probabilité donné par le tableau 17.1 : par exemple,

$$\mathbb{P}\{G=0\} = 1 - \frac{117\,197}{500\,000} \overset{\text{not.}}{=} T_0\,, \qquad \mathbb{P}\{G=3\} = \frac{47\,500}{500\,000} \overset{\text{not.}}{=} T_3\,,$$

$$\mathbb{P}\{G=4\} = \frac{25\,000}{500\,000} \overset{\text{not.}}{=} T_4\,, \qquad \dots\,, \qquad \mathbb{P}\{G=40\,000\} = \frac{1}{500\,000} \overset{\text{not.}}{=} T_{40\,000}\,.$$

On note G_j le gain du j–ième ticket acheté par le joueur. Comme le joueur prélève peu de tickets dans une tranche, puisqu'il n'achète que deux tickets par semaine, les gains $G_1,\ G_2, \dots$ qu'il obtient sont indépendants et identiquement distribués selon $\mathcal{M}(1, T)$. On utilise ici les arguments désormais habituels d'approximation de tirage *a priori* sans remise par un tirage avec remise. Ils sont en fait valables pour la plupart des gains, sauf les plus gros d'entre eux pour lesquels les T_j sont très petits. En revenant à la preuve du Lemme 25.3, on note que pour eux l'approximation n'est pas satisfaisante. De manière équivalente et plus intuitive, il est clair que si $G_1 = 40\,000$ par exemple, alors G_2 et quelques tickets suivants ne pourront prendre cette valeur, puisque le gros lot est unique. Il faut attendre une nouvelle émission pour que la valeur 40 000 puisse être à nouveau prise. Ces très gros gains arrivant rarement (une assertion que nous quantifions plus bas), nous négligeons leur impact dans la modélisation et menons donc l'approximation citée.

En particulier, sous cette modélisation d'échantillon indépendant et identiquement distribué, la loi du gain le plus élevé parmi les n premiers tickets a pour fonction de répartition

$$k \longmapsto F_n(k) \overset{\text{not.}}{=} \mathbb{P}\left\{\max_{t=1,\dots,n} G_t \leqslant k\right\} = \big(\mathbb{P}\{G_1 \leqslant k\}\big)^n\,.$$

La loi s'en déduit : $P_n(0) = F_n(0)$ et pour tout $k = 3,\ 4,\ 6,\ \dots,\ 40\,000$,

$$P_n(k) \overset{\text{not.}}{=} \mathbb{P}\left\{\max_{t=1,\dots,n} G_t = k\right\} = F_n(k) - F_n(k-1)$$

(la formule étant valable parce que les gains sont séparés d'au moins 1 €).

On calcule cette loi pour $n = 100$ (un an) et $n = 1\,000$ (dix ans) dans le tableau 25.3. On la représente à la figure 25.1, *via* sa fonction de survie

$$k \longmapsto R_n(k) = \mathbb{P}\left\{\max_{t=1,\dots,n} G_t \geqslant k\right\} = 1 - F_n(k-1)$$

(où pour la dernière égalité, on a de nouveau utilisé que les gains sont séparés d'au moins 1 €). On remarque qu'en un an, le joueur a toutes les chances (probabilité de 78.8 %) de voir passer un gain d'au moins 20 €, mais qu'il n'a presque aucune chance (probabilité 0.54 %) de gagner 500 € ou plus. Sur une période de dix ans, le gain maximal typique (probabilité 75.64 %) est de 200 €, mais le joueur a toujours peu de chances (probabilité 5.26 %) de gagner 500 € ou plus.

On retient donc que les gains d'un joueur typique varient entre 0 et 200 €. En particulier, à destination du paragraphe 25.1.5, on note qu'il ne faut pas s'attendre à obtenir des

Gain max.	sur 1 an	sur 10 ans	Gain max.	sur 1 an	sur 10 ans
0	0.00	0.00	100	13.28	15.63
3	0.00	0.00	200	14.72	75.64
4	0.01	0.00	500	0.36	3.47
6	3.34	0.00	1 000	0.10	0.99
10	17.85	0.00	10 000	0.04	0.40
20	26.41	0.06	20 000	0.02	0.20
50	23.86	3.41	40 000	0.02	0.20

TABLE 25.3. Table donnant les lois P_{100} et $P_{1\,000}$, i.e., les fréquences (en %) qu'un joueur grattant deux tickets par semaine gagne au moins une fois une certaine somme. Valeurs arrondies à 0.01 % près.

gains supérieurs ou égaux à 500 € lorsque l'on mènera notre analyse d'existence d'espaces d'opportunité.

Un autre argument montre qu'il est intéressant de se restreindre à ces gains et de ne pas considérer ceux qui sont supérieurs. Dans la suite, au paragraphe 25.1.5, nous effectuerons un raisonnement en espérance. Pour qu'il reflète le comportement réel, il faut utiliser un argument indiquant que les moyennes empiriques d'échantillons sont proches de l'espérance de la loi les ayant générés. Cette proximité est d'autant plus grande que la variance est petite, comme l'indiquent les intervalles de confiance du chapitre 3.

Or, la loi $\mathcal{M}(1, T)$ du gain d'un ticket admet pour espérance $\mu \approx 2.04$ tandis que son écart-type est $\sigma = 67.2$; cette valeur est très grande par rapport à l'espérance, il y a un facteur multiplicatif de plus de 30. Mais au vu des gains typiques, il semble plus intéressant de considérer la loi du gain d'un ticket conditionnellement au fait que ce dernier soit inférieur ou égal à 200 €. Cette loi conditionnelle admet $\mu' = 1.85$ et $\sigma' = 9.81$ comme espérance et écart-type. Cette dernière n'est en particulier plus que 5 fois plus grande que l'espérance, ce qui est beaucoup plus satisfaisant.

En résumé, un joueur typique perdra donc plutôt $3 - \mu' = 1.15$ € par ticket ; un joueur atypique gagnera un gros lot et aura éventuellement un solde total positif pendant un certain temps. Les courbes des soldes mobiles de la figure 25.2 l'illustrent : un seul gain supérieur ou égal à 500 € apparaît, pour un total de 2 400 tickets grattés, et encore a-t-il fallu relancer les simulations plusieurs fois ! A l'opposé, plusieurs tickets étaient porteurs d'un lot de 200 €.

Pour s'amuser et dans un but pédagogique envers les joueurs invétérés, on peut alors exhiber un intervalle de confiance sur le montant perdu par un joueur typique (i.e., un joueur se trouvant parmi les 94.74 % de joueurs ne gagnant jamais 500 € ou plus en $n = 1\,000$

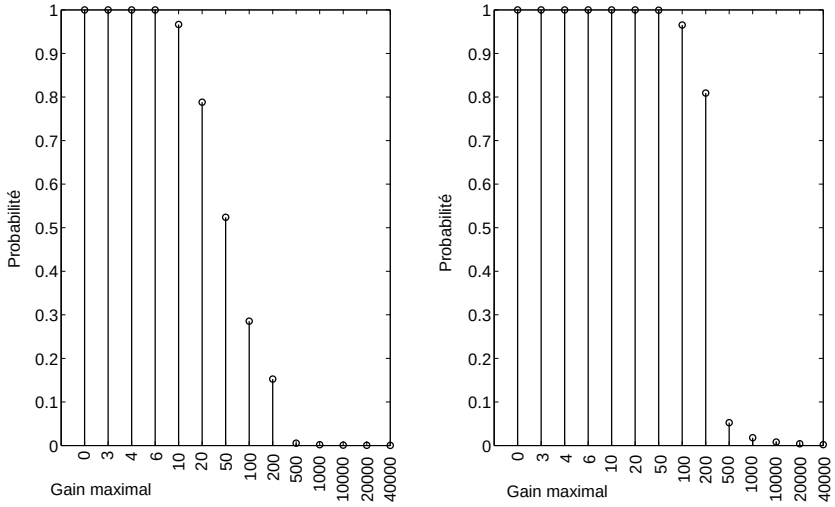

FIGURE 25.1. Fonction de survie du montant du gain maximal, pour un joueur grattant deux tickets par semaine pendant un an (à gauche), pendant dix ans (à droite).

grattages). Le théorème de la limite centrale assure que

$$\sqrt{n}\left(\frac{G_1\,\mathbb{I}_{\{G_1\leqslant 200\}}+\ldots+G_n\,\mathbb{I}_{\{G_n\leqslant 200\}}}{n}-\mu''\right)\rightsquigarrow\mathcal{N}\big(0,\ (\sigma'')^2\big)\,,$$

où μ'' et σ'' sont l'espérance et l'écart-type de $G\,\mathbb{I}_{\{G\leqslant 200\}}$ pour $G\sim\mathcal{M}(1,T)$; un calcul simple montre que $\mu'\approx\mu''\approx 1.85$ et $\sigma'\approx\sigma''$ (il y a égalité à 2×10^{-4} près, ce qui est moins que la précision avec laquelle on avait donné μ' et σ'). Ainsi, avec une probabilité pas trop éloignée de 95 % (l'écart étant induit par le caractère asymptotique de la convergence en loi ci-dessus),

$$G_1\,\mathbb{I}_{\{G_1\leqslant 200\}}+\ldots+G_n\,\mathbb{I}_{\{G_n\leqslant 200\}}\leqslant n\mu''+z_{95\,\%}\,\sigma''\sqrt{n}$$

où $z_{95\,\%}\approx 1.65$ est le quantile à 95 % de la loi normale standard.

Par union d'événements, avec probabilité proche de 90 %, pour n modérément grand (de l'ordre d'une ou plusieurs centaines, mais plus petit que 1 000), on a la majoration pour les soldes de gains eux-mêmes,

$$G_1+\ldots+G_n\leqslant n(\mu''-3)+z_{95\,\%}\,\sigma''\sqrt{n}\,.$$

Pour $n\geqslant 200$, cette borne est négative, i.e., les déviations non improbables envers la perte moyenne $-1.15\,n$ ne permettent même pas d'espérer un gain total positif. Autrement dit, environ 90 % de la population perd à jouer régulièrement pendant deux ans, certains perdant beaucoup d'argent, en moyenne ici, 330 €. Seuls 10 % au mieux réussissent à maintenir, pour

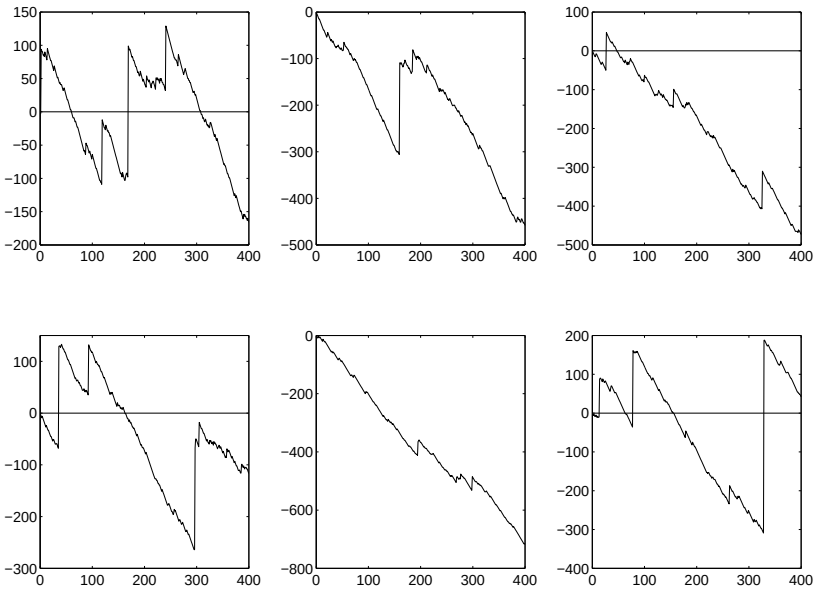

FIGURE 25.2. Soldes mobiles (en ordonnée) des gains et mises de joueurs individuels, en fonction du temps (en abscisse) : six réalisations aléatoires.

un temps, un solde positif entre leurs mises et leurs gains. Mais ce dernier est voué, par loi des grands nombres, à devenir négatif.

25.1.5 Sur l'existence et la pertinence des stratégies de joueur informé

On étudie ici si l'on peut utiliser l'inférence (25.3) pour construire une stratégie qui aurait un solde positif en espérance. On rappelle que l'espérance du solde de la stratégie qui consiste à acheter n tickets au hasard est de $-0.96\,n$ (ou même, plutôt typiquement de l'ordre de $-1.15\,n$), ainsi qu'on l'a vu au paragraphe 25.1.4 On suppose à cet effet deux choses.

Premièrement, même s'il existe des contraintes techniques fortes qui réduisent l'aléa et entraînent qu'un carnet ne peut contenir qu'au plus deux lots significatifs, aucune manipulation n'est effectuée sur les lots mineurs, qui sont répartis dans la tranche d'émission selon un hasard uniforme. De plus, la société de jeu tâche de conserver autant d'aléa que possible et place les lots, et en particulier, le ou les lot(s) significatif(s), au hasard dans le carnet. Autrement dit, une fois l'ensemble des lots constitué pour un carnet, la place de ces derniers à l'intérieur du carnet est donnée par un arrangement aléatoire.

Deuxièmement, le buraliste retenu pour mettre en œuvre notre stratégie, à réception d'un

carnet de 50 tickets, ne le découpe pas pour mettre les tickets individuels sur un présentoir, mais accroche le carnet avec une pince à linge et le débite au fur et à mesure des ventes. En outre, il suit le résultat des gains de ses clients. Ainsi, à tout moment, lorsqu'il reste x tickets dans le carnet, il peut déterminer s'il y avait eu un lot significatif dans les $50 - x$ tickets déjà vendus.

La stratégie proposée est de la forme : s'il reste x tickets dans un carnet dans lequel aucun lot significatif n'est encore tombé, alors on achète ces x tickets. Nous allons tout d'abord prouver qu'il existe des valeurs de x (celles plus petites que 9) telles que l'espérance du solde de cette stratégie est strictement positif. Nous montrerons ensuite qu'il arrive suffisamment fréquemment qu'on soit dans les conditions de mise en œuvre de la stratégie.

Espérance du solde de gain de la stratégie en fonction de x

On dispose de trois types de carnets, selon qu'ils contiennent en leur sein zéro, un, ou deux lots significatifs. A un carnet c donné, on associe alors son nombre de lots significatifs $N(c) \in \{0, 1, 2\}$. On a vu en (25.3) qu'un carnet C tiré au hasard a un type $N(C)$ distribué selon (a, b, q) :

$$\mathbb{P}\big\{N(C) = k\big\} = \begin{cases} a & \text{si } k = 0, \\ b & \text{si } k = 1, \\ q & \text{si } k = 2. \end{cases}$$

On note qu'ici on ne fait plus l'identification entre C et $N(C)$ qu'on faisait au début du corrigé, par exemple au paragraphe 25.1.1.

On note E_x l'événement selon lequel aucun des $50 - x$ premiers tickets d'un carnet C pris au hasard n'a été porteur d'un lot significatif. C'est une observation que le buraliste peut effectuer. On rappelle que l'emplacement du ou des lots significatifs, quand ils existent dans un carnet, est aléatoire. Etant donné un carnet avec un seul lot significatif, la probabilité que celui-ci soit dans les x derniers tickets est de $x/50$. De même, si le carnet contient deux lots significatifs, la probabilité que ces deux lots se trouvent dans les x derniers tickets est $d_1 = 0$ si $x = 1$ et est donné par le rapport

$$d_x \overset{\text{not.}}{=} \binom{x}{2} \Big/ \binom{50}{2}$$

pour $x \geqslant 2$. Cela entraîne que conditionnellement à $N(C) = 1$, l'événement E_x a une probabilité $x/50$, tandis que conditionnellement à $N(C) = 2$, cette probabilité est d_x. Enfin, il est clair que lorsque $N(C) = 0$, l'événement E_x est certain. Il vient alors

$$r_x \overset{\text{not.}}{=} \mathbb{P}(E_x) = \sum_{k \in \{0,1,2\}} \mathbb{P}\big\{N(C) = k\big\}\ \mathbb{P}\Big(E_x \,\Big|\, \big\{N(C) = k\big\}\Big) = a + b\,\frac{x}{50} + q\,d_x\ .$$

On détermine alors la loi conditionnelle de C sachant E_x (il s'agit d'une loi dite *a posteriori*). Par la formule de Bayes et en réinjectant les valeurs précédentes, on a

$$\mathbb{P}\big\{N(C)=k \mid E_x\big\} = \frac{\mathbb{P}\big\{N(C)=k\big\}\ \mathbb{P}\Big(E_x \,\Big|\, \big\{N(C)=k\big\}\Big)}{\mathbb{P}(E_x)}\,,$$

soit

$$\mathbb{P}\big\{N(C)=k \mid E_x\big\} = \begin{cases} a/r_x & \text{si } k=0,\\ (b/r_x)\,(x/50) & \text{si } k=1,\\ (q/r_x)\,d_x & \text{si } k=2. \end{cases}$$

On calcule maintenant, dans chaque cas, l'espérance du gain typique. Avec les notations du paragraphe 25.1.4, et en considérant toujours une variable aléatoire $G \sim \mathcal{M}(1,T)$ donnant le gain d'un ticket pris au hasard, on note respectivement

$$m = \mathbb{E}\big[G \mid G \leqslant 10\big] \approx 1.02 \qquad \text{et} \qquad M = \mathbb{E}\big[G \mid G \in \{20,50,100,200\}\big] \approx 55.54$$

l'espérance du gain conditionnellement au fait que le ticket gagne au mieux un lot mineur ou est perdant et à celui que le ticket porte un lot significatif typique. On élimine donc ici également les gains atypiques, i.e., ceux supérieurs ou égaux à 500 € : on les a caractérisés au paragraphe 25.1.4 comme suffisamment rares pour que même en jouant régulièrement, la probabilité qu'on les rencontre soit extrêmement faible.

Vu les hypothèses effectuées sur la fabrication des carnets, lorsque E_x est réalisé sur un carnet C aléatoire, l'espérance des gains portés par les x tickets restants est, conditionnellement à $N(C)$ et au fait qu'il n'y a pas de gains atypiques, de

$$\big(x - N(C)\big)m + N(C)\,M$$

pour $x \geqslant 1$ (on rappelle que $N(C)=2$ est incompatible avec E_1).

En combinant avec la loi *a posteriori*, il vient que l'espérance des gains d'un carnet pris au hasard, conditionnellement à E_x avec $x \geqslant 2$ (et toujours au fait qu'il n'y a pas de gains atypiques), vaut

$$e_x = \frac{1}{r_x}\Big(axm + b(x/50)\big(M + (x-1)m\big) + qd_x\big(2M + (x-2)m\big)\Big)\,.$$

Cette formule est valable également pour $x=1$, puisque l'on a défini $d_1 = 0$.

Il faut bien entendu s'intéresser en pratique au solde de jeu, c'est-à-dire aux gains déduction faite des mises, à savoir : $s_x = e_x - 3x$. La figure 25.3 donne le tracé de s_x comme fonction de x. On y voit, sans surprise, et conformément aux résultats du paragraphe 25.1.4, que la perte moyenne lorsque l'on achète 50 tickets vaut (aux erreurs d'arrondis près) $50 \times 1.15 = 57.50$ €.

La valeur de s_x est maximale pour $x=4$ et vaut $s_4 = 2.10$ € (aux arrondis près).

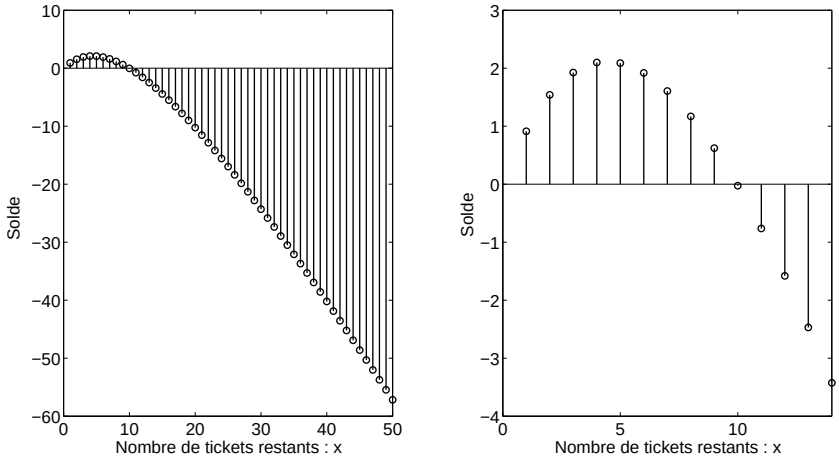

FIGURE 25.3. Espérance du solde de jeu (en ordonnée) pour un détaillant fictif mettant en œuvre la stratégie de joueur informé, en fonction (en abscisse) du nombre de tickets restant dans un carnet dans lequel aucun gain significatif n'est sorti.

Pertinence de la stratégie précédente

Le solde admet une espérance strictement positive : chaque fois que E_4 survient, on gagne donc en moyenne 2.10 €. La question est de déterminer la fréquence de survenance de E_4. Si c'est une fois l'an, l'investissement en temps et en énergie du buraliste (qui, rappelons-le, doit suivre les gains ou absences de gains des joueurs) n'est par exemple pas rentable. Le calcul a déjà été effectué : cette fréquence vaut

$$\mathbb{P}(E_4) = r_4 \approx 31.7\,\% \,.$$

En termes simples : plus de 30 % des carnets conduisent à la réalisation de E_4. Or, on peut supposer que le buraliste vend plusieurs carnets par semaine, disons, vingt. Dans ce cas, la stratégie précédente gagne en moyenne

$$20\, r_4\, e_4 \approx 20 \times 0.317 \times 2.10 = 13.31\,€$$

par semaine. (Le fait que la stratégie de joueur informé puisse être appliquée régulièrement justifie, par la loi des grands nombres et le théorème de la limite centrale, que le solde moyen des gains soit proche de s_4 avec grande probabilité.)

En réalité, on peut faire un tout petit peu mieux. La meilleure valeur de x a été déterminée uniquement pour maximiser l'espérance de gain lors d'une réalisation de la stratégie, alors que le buraliste est amené à la mettre en œuvre plusieurs fois par semaine. Il s'agit en fait de maximiser la quantité $r_x\, e_x$, qui tient compte non seulement des gains réalisés à chaque fois,

x	1	2	3	4	5	6	7	8	9	10 et plus
$r_x\, s_x$	0.25	0.44	0.58	0.66	0.69	0.66	0.57	0.43	0.24	$\leqslant 0$

TABLE 25.4. Valeurs de la quantité $r_x\, s_x$ en fonction de x (présentées avec deux chiffres significatifs) ; pour $x \geqslant 10$, les valeurs sont négatives.

mais aussi du nombre de fois où on les réalise. Le tableau 25.4 reporte les valeurs prises par $r_x\, e_x$ en fonction de x. On en déduit que la valeur $x = 5$ obtient des performances légèrement meilleures en situation de répétitions que la valeur $x = 4$ calibrée pour une mise en œuvre unitaire de la stratégie.

Dans tout ce paragraphe, nous avons écarté les gains atypiques. En fait, il se trouve que le buraliste mettant en œuvre la stratégie pour $x = 4$ ou $x = 5$ achète de nombreux tickets par semaine : en moyenne, $x\, r_x$ fois le nombre de carnets vendus. Lorsque ce dernier vaut 20 comme supposé précédemment et que $x = 5$, cela consiste en plus de 30 tickets par semaine. En un an, il effectue un peu plus d'achats que le joueur invétéré du paragraphe 25.1.4 en dix ans. L'hypothèse que les gains typiques sont tous inférieurs ou égaux à 200 serait peut-être à revoir ; mais nous préférons la conserver afin de montrer que même sous cette hypothèse pessimiste, il y a bien opportunité de gain.

25.1.6 Conclusions ?

La conclusion est laissée au lecteur : d'après vous, l'investissement du buraliste pour gagner ces 13 € supplémentaires par semaine en vaut-il la peine, ou le temps et l'énergie consacrés au suivi trouveraient-ils une meilleure rémunération dans une autre activité accessoire ?

Par ailleurs, si une telle affaire était portée devant un tribunal et que vous y soyez cité comme expert, que diriez-vous ? Le hasard dans la répartition des lots, à défaut d'être total, vous semble-t-il suffisant ? Si l'affaire avait lieu en France, on porterait notamment à la défense de la société de jeux que selon un décret de 2002, il n'est pas demandé une répartition des lots selon un hasard total, mais un « hasard prépondérant » – une notion floue qu'il nous est délicate de quantifier mathématiquement. On peut sans doute dire, par exemple, que le résultat d'un tirage selon une loi de Bernoulli Ber(0.49) est causé essentiellement par le hasard, mais que ce n'est pas le cas pour une réalisation de Ber(0.80). On pourrait relier la place prépondérante du hasard au fait que les déviations de part et d'autre de la valeur moyenne rendent difficile la prédiction des réalisations. Dans ce cas, la quantité-clé serait le rapport entre cette espérance et l'écart-type de la distribution, c'est elle qui mesurerait la prépondérance ou non du hasard.

25.2 Questions possibles, conseils aux étudiants et bibliographie

25.2.1 Questions éventuelles d'un enseignant

S'agissant ici d'effectuer des tests de conformité sur les paramètres de lois multinomiales, un enseignant ne pourrait réprimer une question sur les tests du χ^2 ; comme par exemple, rappeler la forme générale de la statistique du χ^2 et l'énoncé de sa convergence en loi vers une loi du χ^2 (de degrés de liberté à préciser). Il reviendra également sur tous les points mal maîtrisés dans la méthodologie des tests.

On peut également s'attendre à une discussion sur la forme du tirage stochastique sous-jacent des lots, qui risque de tourner un peu au débat d'opinions, dans la mesure où l'on n'a que peu de données précises et où l'on est obligé, à plusieurs endroits du texte, d'imaginer comment procède la société de jeux et d'émettre des hypothèses supplémentaires sur la bonne génération aléatoire de la répartition des lots. Dans tous les cas, il est important d'argumenter son opinion et de rester honnête dans l'échange en reconnaissant une contradiction soulevée par un enseignant et en essayant, en retour, de mettre en difficulté, si cela est possible, son idée de la modélisation.

Enfin, sur un tout autre plan, à propos de jeux de hasard du monde réel, on pourrait demander s'il existe des stratégies de joueur informé au Loto par exemple, qui est un célèbre jeu organisé par la Française des Jeux. Il est clair que compte tenu de la méthode de tirage, on ne peut pas augmenter sa probabilité de gagner. En revanche, comme il s'agit d'un jeu par répartition et que les gains sont divisés entre les vainqueurs, il vaut mieux être seul ou peu nombreux à gagner le gros lot. Pour assurer cela, il faut jouer des nombres peu retenus par les autres joueurs (par exemple les nombres plus grands que 32, qui ne correspondent ni à des jours ni à des mois de naissance). Ainsi, on augmente l'espérance de son gain.

Pour mettre cette stratégie en œuvre, il faudrait cependant disposer d'une étude statistique donnant les répartitions des numéros joués. Or, la Française des Jeux ne communique pas ces données. En revanche, elle procure les résultats des tirages passés et tous les tests statistiques montrent que les résultats obtenus semaines après semaines sont bien totalement indépendants et uniformément répartis.

25.2.2 Conseils aux étudiants

Le texte paraît inoffensif au premier abord, mais il teste l'étendue des connaissances dans la méthodologie des tests. Tous les étudiants raisonnablement à l'aise avec le cours de statistique doivent pouvoir rejeter l'hypothèse de répartition totalement uniforme.

On rappelle que lorsque l'on rejette une hypothèse H_0 aussi franchement que celle de répartition uniforme ici, la forme de H_1 importe peu, et c'est pourquoi nous avons parfois pris H_1 fantôme. La plupart des tests suivants (de conformité) ont été également pris unilatères : là, en revanche, la forme unilatère de H_1 est fort importante pour avoir une meilleure p–valeur (souvent, pour la diminuer d'un facteur 2 par rapport au cas bilatère) et il faut la justifier précisément au cas par cas.

Au passage, rappelons que la notion de p–valeur (voir chapitre 3) n'est pas au programme de l'agrégation ; elle est pourtant bien utile pour quantifier l'attachement à H_0, surtout que l'on obtient ici une p–valeur, lors du test de la répartition totalement aléatoire, inhabituellement petite. On pourra utilement, lors du calcul d'une p–valeur, faire un dessin pour montrer la forme de la zone de rejet (déterminée en fonction de l'hypothèse alternative H_1), la valeur observée de la statistique de test, et la p–valeur (l'aire située sous une certaine partie de la courbe de densité de la loi de la statistique).

Le point crucial du corrigé est la modélisation stochastique sur laquelle s'ouvre le texte ; il faut bien la détailler, et notamment expliquer précisément pourquoi et comment l'on approxime, à deux moments, des tirages sans remise par des tirages avec remise. On ne peut pas et on ne doit pas commencer de but en blanc en supposant que l'on a affaire à des échantillons d'observations indépendantes et identiquement distribuées : c'est là tout le sel de la modélisation, passer d'une situation concrète à une situation mathématiquement agréable. Nous n'avons vu ce point que trop peu souvent suffisamment détaillé par nos étudiants.

25.2.3 Bibliographie

Ce texte est *inspiré* de faits réels, mais ne prétend en aucun cas les reconstituer. Au contraire, des modifications et simplifications notoires ont été considérées, pour rendre le propos plus exploitable dans le cadre d'un manuel de niveau master. Gilles Stoltz est reconnaissant aux sources qui l'avaient contacté pour lui présenter l'histoire réelle sous-jacente. Certaines idées de simulation proviennent d'un énoncé de travaux pratiques d'Alain Trouvé.

25.3 Commentaires sur les simulations `Matlab`

25.3.1 Architecture générale des programmes

Une fois n'est pas coutume, le texte utilise plutôt `Matlab` comme une calculatrice géante (de quantiles et de fonctions de répartition). On a groupé dans le fichier maître ci-dessous une partie seulement des calculs informatiques menés pour écrire le corrigé ; on a essentiellement repris ceux qui ont conduit aux tableaux et graphiques des paragraphes 25.1.4 et 25.1.5. Une seule fonction auxiliaire, `Ticket`, est utilisée, elle simule un échantillon de tickets distribués selon la loi multinomiale d'intérêt.

25.3.2 Commentaires sur le code `Matlab`

La boîte à outils `StixBox`, dont nous avons déjà parlé aux chapitres 9.6 et 19, nous est très utile ici ; elle permet notamment, pour les lois les plus communes, par exemple la loi binomiale, de simuler un échantillon (`rbinom`), de calculer la probabilité d'une valeur (`dbinom`) ou la densité en un point donné (lorsque la loi est continue), la valeur de la fonction de répartition (`pbinom`) ou de déterminer les quantiles (`qbinom`). Autres exemples, pour la loi hypergéométrique (pour $k = 2$ modalités seulement), on utilise respectivement `rhypg`, `dhypg`, `phypg` et `rhypg`. Pour voir la liste de toutes les fonctions implémentées dans la boîte à outils `StixBox`, qui, on le

rappelle, est disponible à l'oral de l'agrégation, on pourra taper `help Contents` dans une fenêtre de commande `Matlab` (en vérifiant au préalable que le répertoire contenant les fichiers de `StixBox` est dans la liste des chemins disponibles). Ainsi, on a par exemple obtenu le tableau 25.1 en calculant un vecteur `B` selon

```
B = dbinom([0 1], 50, 7697/500000);
B = [B 1-sum(B)];
```

Quant au tableau 25.2, il procède de

```
pbinom([12 11 10 9 6 3 1 0], 100, B(3));
```

En particulier, on n'a pas besoin de déterminer les différentes probabilités à la main en calculant des coefficients binomiaux ; ceci étant, il est utile de savoir que ceux-ci sont obtenus par la commande `nchoosek` (c'est comme cela que l'on appelle les coefficients binomiaux en anglais, cela réfère au nombre de choix possibles de k éléments dans un ensemble en contenant n).

D'autres remarques plus rapides sont que d'une part, au lieu de décrire explicitement notre matrice `T` de répartition dans le code, on aurait pu la lire à partir d'un fichier (par exemple, un fichier texte, avec deux colonnes séparées, de manière cohérente, par des virgules, espaces ou tabulations), en employant la commande `dlmread`. La commande inverse permettant d'écrire une matrice dans un fichier texte est `dlmwrite`.

Du point de vue des graphiques, `stem` semble ici plus adapté que `plot` et fournit des figures souvent plus jolies que `bar` ; cette dernière commande est vraiment à réserver aux cas où l'on est intéressé par l'aire des bâtons (comme c'est le cas pour les histogrammes estimant les fonctions de densité, au chapitre 24, et encore dans ce cas lui préfère-t-on `histo`). Les lignes

```
set(gca, 'XTick', [1:length(R)]);
set(gca, 'XTickLabel', [0; T(end:-1:1, 1)]);
```

sont très techniques et permettent de changer la légende des bâtons en abscisses ; ceux-ci sont par défaut indexés par 0, 1, 2, etc, et l'on veut plutôt remplacer cette numérotation par la valeur des lots, 0, 3, 4, 6, 10, etc. `gca` désigne l'objet figure. `XTick` et `XTickLabel` sont les propriétés de cet objet que l'on veut modifier, et le troisième argument de `set` donne les nouvelles valeurs. On voit apparaître là le côté orienté objet de `Matlab`, que l'on avait déjà évoqué au paragraphe 20.3.2.

25.3.3 Codes des simulations

Ils sont disponibles en annexe, au chapitre 35, page 519.

Cinquième partie

Compléments techniques

CHAPITRE 26

Rappels de théorie de l'intégration et des probabilités

26.1 Résultats de théorie de l'intégration

26.1.1 Théorème de dérivation des intégrales à paramètre

On en énonce une version lisible et qui est suffisante pour la plupart des applications. Les hypothèses peuvent cependant être affaiblies, voir par exemple [LG06, Théorème 2.3.2].

Théorème 26.1 *Soit $(\mathcal{X}, \mathcal{F}, \mu)$ un espace mesuré et soit I un intervalle ouvert de $\mathbb{R}$. On considère une application $f : I \times \mathcal{X} \to \mathbb{R}$ vérifiant les conditions suivantes :*

(1) pour tout $t \in \mathbb{R}$, l'application $f(t, \cdot) : \mathcal{X} \to \mathbb{R}$ appartient à $\mathbb{L}^1(\mu)$;

(2) pour μ–presque tout x, l'application $f(\cdot, x) : I \to R$ est dérivable ;

(3) il existe une fonction $g \in \mathbb{L}^1(\mu)$ telle que pour μ–presque tout x,

$$\forall t \in I, \qquad \left| \frac{\partial f}{\partial t}(t, x) \right| \leqslant g(x) \, .$$

Alors l'application

$$F : t \in I \longmapsto \int_{\mathcal{X}} f(t, x) \, \mathrm{d}\mu(x)$$

est bien définie et est dérivable sur I, de dérivée en $t \in I$ égale à

$$F'(t) = \int_{\mathcal{X}} \frac{\partial f}{\partial t}(t, x) \, \mathrm{d}\mu(x) \, .$$

26.1.2 Le théorème fondamental du calcul différentiel et intégral

Ce qui suit est essentiellement tiré de [Rud98, chapitre 7] et s'intéresse aux liens réciproques entre intégration et dérivation. On redonne ici brièvement la progression logique et l'enchaînement des résultats. Dans tout ce qui suit, on note λ la mesure de Lebesgue.

La première partie du théorème fondamental indique qu'en dérivant une primitive, on obtient l'intégrande. Sa preuve repose sur la considération des points de Lebesgue de la fonction, dont on a rappelé la définition au paragraphe 24.1.5.

Théorème 26.2 *Soit $f \in \mathbb{L}^1(\lambda)$. On note F la primitive définie pour tout $x \in \mathbb{R}$ par*

$$F(x) = \int_{-\infty}^{x} f \, \mathrm{d}\lambda \, .$$

Alors l'ensemble $\mathcal{D}$ des points où F est dérivable est le complémentaire d'un ensemble de mesure nulle et l'on a $F'(x) = f(x)$ pour $x \in \mathcal{D}$.

La seconde partie du théorème fondamental généralise ce fait et indique une condition nécessaire et suffisante pour qu'une fonction soit l'intégrale de sa dérivée.

Définition 26.3 Soit $I = [a, b]$ un intervalle compact de $\mathbb{R}$. Une fonction $f : I \to \mathbb{R}$ est dite *absolument continue* sur I si pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour tout entier $n \geqslant 1$ et toute suite finie $\big([a_k, b_k]\big)_{1 \leqslant k \leqslant n}$ de sous-intervalles de I d'intérieurs disjoints,

$$\sum_{k=1}^{n} |b_k - a_k| < \delta \qquad \text{implique} \qquad \sum_{k=1}^{n} \big| f(a_k) - f(b_k) \big| < \varepsilon \, .$$

Une fonction $\mathbb{R} \to \mathbb{R}$ est dite absolument continue si elle est absolument continue sur tout sous-intervalle compact de $\mathbb{R}$.

En particulier, en prenant $n = 1$, on note qu'une fonction absolument continue est continue. Le paragraphe 24.2.1 explique les liens entre absolue continuité au sens des fonctions réelles (énoncée ci-dessus) et absolue continuité $\nu \ll \lambda$ d'une loi ν sur $\mathbb{R}$ par rapport à la mesure de Lebesgue : le théorème suivant indique qu'une condition nécessaire et suffisante pour $\nu \ll \lambda$ est l'absolue continuité (au sens des fonctions réelles) de la fonction de répartition F_ν de ν.

Théorème 26.4 *Soit $I = [a, b]$ un intervalle compact de $\mathbb{R}$. Une fonction $f : I \to \mathbb{R}$ est absolument continue sur I si et seulement si f est dérivable en presque tout point de I, de dérivée (ainsi définie presque partout) notée f' vérifiant en outre $f' \in \mathbb{L}^1(\lambda)$ et*

$$\forall x \in I, \qquad f(x) - f(a) = \int_a^x f' \, \mathrm{d}\lambda \, .$$

Pour la culture, voici, dans la lignée de ce théorème, un autre résultat utile (et dont la preuve est fort différente). On note qu'on requiert ici une dérivabilité en tout point et non pas seulement en presque tout point.

Théorème 26.5 *Soient $I = [a, b]$ un intervalle compact de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction dérivable en* tout *point de I, de dérivée vérifiant $f' \in \mathbb{L}^1(\lambda)$. Alors*

$$\forall x \in I, \qquad f(x) - f(a) = \int_a^x f' \, \mathrm{d}\lambda \, .$$

26.1.3 Le théorème de changement de variables

On fixe ici deux ouverts U et D de $\mathbb{R}^d$, où $d \in \mathbb{N}^*$. On rappelle qu'une application $\varphi : U \to D$ est un difféomorphisme de classe C^1 si φ est bijective et que φ et φ^{-1} sont de classe C^1 respectivement sur U et D. On sait alors que la différentielle $D\varphi(u)$ de φ est inversible en tout point $u \in U$. On appelle jacobien de φ en u le déterminant $\det D\varphi(u)$.

Théorème 26.6 *Soit $\varphi : U \to D$ un difféomorphisme de classe C^1. Alors, pour toute fonction borélienne $f : D \to \mathbb{R}_+$,*

$$\int_D f \, \mathrm{d}\lambda = \int_U f\big(\varphi(u)\big) \, \big|\det D\varphi(u)\big| \, \mathrm{d}u \, .$$

La version qui précède est proposée dans tous les ouvrages classiques de théorie de l'intégration, voir par exemple [LG06]. On invite le lecteur à consulter également une version légèrement plus générale proposée par [Rud98, Théorème 7.26].

26.2 Résultats de la théorie des probabilités

26.2.1 Convergences

Définitions équivalentes de la convergence en loi $X_n \rightsquigarrow X$

Le lemme suivant donne plusieurs définitions équivalentes de la convergence en loi. On pourra en trouver une preuve dans [vdV98, chapitre 2]. On l'énonce pour un vecteur aléatoire X défini sur un espace $(\Omega, \mathcal{A}, \mathbb{P})$ et à valeurs dans $\mathbb{R}^d$, pour $d \in \mathbb{N}^*$, et des vecteurs aléatoires X_n définis chacun sur un espace $(\Omega_n, \mathcal{A}_n, \mathbb{P}_n)$ et à valeurs dans $\mathbb{R}^d$, pour $n \geqslant 1$. On note les espérances correspondantes $\mathbb{E}$ et $\mathbb{E}_n$. Lorsque x et y sont deux vecteurs, la notation $x \leqslant y$ signifie que l'inégalité $x_j \leqslant y_j$ vaut pour chacune des composantes de x et y.

Lemme 26.7 (Portmanteau) Les assertions suivantes sont équivalentes et définissent la convergence en loi $X_n \rightsquigarrow X$ d'une suite de vecteurs aléatoires (X_n) vers X :

(1) en tout point de continuité de $x \mapsto \mathbb{P}\{X \leqslant x\}$, on a la convergence simple

$$\mathbb{P}_n\{X_n \leqslant x\} \longrightarrow \mathbb{P}\{X \leqslant x\} \, ;$$

(2) pour toute fonction continue bornée f,

$$\mathbb{E}_n\big[f(X_n)\big] \longrightarrow \mathbb{E}\big[f(X)\big] \, ;$$

(3) pour toute fonction lipschitzienne bornée f,

$$\mathbb{E}_n\big[f(X_n)\big] \longrightarrow \mathbb{E}\big[f(X)\big] \, ;$$

(4) pour toute fonction f continue et positive,

$$\liminf \ \mathbb{E}_n\big[f(X_n)\big] \geqslant \mathbb{E}\big[f(X)\big] ;$$

(5) pour tout ensemble ouvert U,

$$\liminf \ \mathbb{P}_n\{X_n \in U\} \geqslant \mathbb{P}\{X \in U\} ;$$

(6) pour tout ensemble fermé F,

$$\limsup \ \mathbb{P}_n\{X_n \in F\} \leqslant \mathbb{P}\{X \in F\} ;$$

(7) pour tout ensemble borélien B, de frontière $\delta B = \overline{B} \setminus \overset{\circ}{B}$ telle que $\mathbb{P}\{X \in \delta B\} = 0$,

$$\mathbb{P}_n\{X_n \in B\} \longrightarrow \mathbb{P}\{X \in B\} .$$

Ces définitions équivalentes de la convergence en loi montrent en particulier que cette dernière est stable par passage aux fonctions continues.

Lemme de Slutzky

Il n'est en général pas vrai que les convergences marginales $X_n \rightsquigarrow X$ et $Y_n \rightsquigarrow Y$ entraînent la convergence jointe $(X_n, Y_n) \rightsquigarrow (X, Y)$. C'est vrai lorsque pour tout n, les variables aléatoires X_n et Y_n sont indépendantes et que X et Y le sont également. C'est aussi vrai lorsque $X_n \to X$ et $Y_n \to Y$ en probabilité, auquel cas (X_n, Y_n) converge vers (X, Y) en probabilité, et donc en loi.

Un troisième cas plus intéressant que ces deux cas trop particuliers est fourni par le lemme de Slutzky : le cas où Y est presque sûrement constante, égale à c. On rappelle que dans ce cas, la convergence en loi $Y_n \rightsquigarrow c$ est équivalente à la convergence en probabilité $Y_n \to c$. Pour simplifier l'énoncé, on suppose ici que toutes les variables aléatoires sont définies sur le même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. On pourrait évidemment lever cette hypothèse en indiçant correctement toutes les probabilités et espérances comme nous l'avons fait plus haut.

Lemme 26.8 (Slutzky) Soient (X_n) et (Y_n) deux suites de vecteurs aléatoires prenant leurs valeurs respectivement dans $\mathbb{R}^d$ et $\mathbb{R}^p$, où p et d sont deux éléments de $\mathbb{N}^*$ éventuellement différents. Si $X_n \rightsquigarrow X$ pour un certain vecteur aléatoire X et $Y_n \to c$ en probabilité pour un vecteur déterministe c, alors $(X_n, Y_n) \rightsquigarrow (X, c)$.

Le plus souvent, on applique à cette convergence jointe $(X_n, Y_n) \rightsquigarrow (X, c)$ une fonction continue g et l'on en tire $g(X_n, Y_n) \rightsquigarrow g(X, c)$. Des fonctions souvent considérées sont, dans le cas où $d = p = 1$, la somme $g(x, y) = x + y$, le produit $g(x, y) = xy$ ou la division $g(x, y) = x/y$, que l'on réservera toutefois au cas où $c \neq 0$. Ces formules pour la somme et le produit s'étendent de manière naturelle au cas multidimensionnel lorsque $d = p$.

Preuve On va utiliser la caractérisation (3) du lemme de Portmanteau (Lemme 26.7) et on considère à cet effet une fonction lipschitzienne bornée $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ arbitraire. On note L sa constante de Lipschitz et B un majorant uniforme sur les valeurs qu'elle prend. Il s'agit de montrer que le terme suivant converge vers 0,

$$\Big|\mathbb{E}\big[f(X_n, Y_n)\big] - \mathbb{E}\big[f(X, c)\big]\Big| \leqslant \Big|\mathbb{E}\big[f(X_n, Y_n) - f(X_n, c)\big]\Big| + \Big|\mathbb{E}\big[f(X_n, c)\big] - \mathbb{E}\big[f(X, c)\big]\Big| .$$

Le second terme dans la majoration tend vers 0 par le point (3) du lemme de Portmanteau appliqué à la fonction $f(\cdot, c)$ et à la convergence $X_n \rightsquigarrow X$. On montre maintenant que c'est également le cas du premier terme. Pour cela, on fixe dans un premier temps $\varepsilon > 0$ et on décompose ce premier terme en fonction de la valeur de $\|Y_n - c\|$:

$$\Big|\mathbb{E}\big[f(X_n, Y_n) - f(X_n, c)\big]\Big| \leqslant 2B\ \mathbb{P}\big\{\|Y_n - c\| > \varepsilon\big\} + L\varepsilon\ \mathbb{P}\big\{\|Y_n - c\| \leqslant \varepsilon\big\} .$$

Comme $Y_n \to c$ en probabilité, il vient

$$\limsup \Big|\mathbb{E}\big[f(X_n, Y_n) - f(X_n, c)\big]\Big| \leqslant L\varepsilon .$$

Ceci étant vrai pour tout $\varepsilon > 0$, cela montre la convergence vers 0 du membre de gauche et conclut la preuve. □

Extension du théorème de convergence dominée à la convergence en probabilité

On énonce enfin une extension souvent méconnue du théorème de convergence dominée au cas où les variables aléatoires ne convergent pas presque sûrement mais en probabilité. Ici, il est important que toutes les variables aléatoires soient définies sur le même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. On l'énonce pour des variables aléatoires réelles unidimensionnelles mais, quitte à travailler composante par composante, le résultat s'étend à des vecteurs aléatoires.

Théorème 26.9 *Soit (X_n) une suite de variables aléatoires réelles convergeant en $\mathbb{P}$–probabilité vers X et telle que les X_n sont toutes dominées par $Y \in \mathbb{L}^1(\mathbb{P})$. Alors $X \in \mathbb{L}^1(\mathbb{P})$ et $X_n \to X$ dans $\mathbb{L}^1(\mathbb{P})$.*

Preuve On montre que 0 est la seule valeur d'adhérence possible pour la suite

$$(r_n)_{n\geqslant 1} = \Big(\mathbb{E}\big[|X_n - X|\big]\Big)_{n\geqslant 1} .$$

Soit $\ell \in [-\infty, +\infty]$ une de ses valeurs d'adhérence (*a priori*, éventuellement infinie) et soit une sous-suite convergeant vers ℓ, donnée par $\phi : \mathbb{N} \to \mathbb{N}$, i.e., $r_{\phi(n)} \to \ell$.

La suite $\big(X_{\phi(n)}\big)$ converge en particulier en probabilité vers X, on peut donc en extraire une sous-suite convergeant $\mathbb{P}$–presque sûrement vers X, donnée par $\phi \circ \psi$ où $\psi : \mathbb{N} \to \mathbb{N}$:

$$X_{\phi(\psi(n))} \longrightarrow X \qquad \mathbb{P}\text{–p.s.}$$

A la sous-suite $X_{\phi(\psi(n))}$, dominée par Y, on peut alors appliquer le théorème de convergence dominée *ordinaire* et voir qu'elle converge vers X dans $\mathbb{L}^1(\mathbb{P})$, soit $r_{\phi(\psi(n))} \to 0$. Nécessairement, $\ell = 0$ et l'extension est prouvée. □

26.2.2 Inégalités utiles

Retour sur l'inégalité de Hoeffding

On prouve dans un premier temps l'inégalité de Hoeffding (Lemme 3.7). Elle repose sur le lemme suivant, dit lemme de Hoeffding.

Lemme 26.10 (Lemme de Hoeffding) Soit Y une variable aléatoire bornée : il existe deux réels a et b tels que $a \leqslant Y \leqslant b$ presque sûrement. Alors pour tout $s > 0$,

$$\log \mathbb{E}\big[e^{sY}\big] \leqslant s\,\mathbb{E}[Y] + \frac{s^2}{8}(b-a)^2 \,.$$

On propose une preuve élégante mais plus longue que la preuve plus classique et calculatoire présentée, par exemple, par [GS01].

Preuve On note, pour tout $s \geqslant 0$,

$$\psi(s) = \log \mathbb{E}\big[e^{sY}\big] \,.$$

Comme Y est bornée, des applications répétées du Théorème 26.1 montrent que les fonctions $s \in \mathbb{R}_+^* \mapsto \mathbb{E}\big[e^{sY}\big]$ et ψ sont bien définies et de classe C^∞ sur $\mathbb{R}_+^*$. Par ailleurs, les dérivées premières et secondes sont données, pour $s > 0$, par

$$\psi'(s) = \frac{\mathbb{E}\big[Ye^{sY}\big]}{\mathbb{E}\big[e^{sY}\big]} \qquad \text{et} \qquad \psi''(s) = \frac{\mathbb{E}\big[Y^2e^{sY}\big]\,\mathbb{E}\big[e^{sY}\big] - \Big(\mathbb{E}\big[Ye^{sY}\big]\Big)^2}{\Big(\mathbb{E}\big[e^{sY}\big]\Big)^2} \,.$$

On prolonge ces formules par continuité en $s = 0$.

Interprétons $\psi''(s)$ comme une certaine variance :

$$\psi''(s) = \mathbb{E}_{\mathbb{Q}}\big[Y^2\big] - \big(\mathbb{E}_{\mathbb{Q}}[Y]\big)^2 = \mathrm{Var}_{\mathbb{Q}}\, Y \,,$$

où la loi $\mathbb{Q}$ est absolument continue par rapport à $\mathbb{P}$, de densité donnée par

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\omega) = \frac{1}{\mathbb{E}\big[e^{sY}\big]} e^{sY(\omega)} \,,$$

pour tout $\omega \in \Omega$, où Ω est l'espace probabilisé sur lequel est définie Y. En particulier, en utilisant la remarque de la page 24, il vient

$$\psi''(s) = \mathrm{Var}_{\mathbb{Q}}\, Y \leqslant \frac{(b-a)^2}{4} \,.$$

En intégrant deux fois et en utilisant le fait que $\psi'(0) = \mathbb{E}[Y]$ et $\psi(0) = 0$, on obtient que pour tout $s > 0$,

$$\psi(s) \leqslant \frac{(b-a)^2}{4}\,\frac{s^2}{2} + s\,\mathbb{E}[Y] \,,$$

qui est le résultat attendu. □

Preuve du Lemme 3.7 La preuve de l'inégalité de Hoeffding se déduit du lemme de Hoeffding par une majoration de Chernoff (i.e., une application de l'inégalité de Markov à l'exponentielle de la somme de variables aléatoires d'intérêt) : pour tout $s > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n} Y_i \geqslant t\right\} = \mathbb{P}\left\{e^{s(Y_1+\ldots+Y_n)} \geqslant e^{st}\right\} \leqslant \frac{\mathbb{E}\left[e^{s(Y_1+\ldots+Y_n)}\right]}{e^{st}} .$$

Or, par indépendance des Y_i, leur caractère centré et par le lemme de Hoeffding (Lemme 26.10),

$$\mathbb{E}\left[e^{s(Y_1+\ldots+Y_n)}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{sY_i}\right] \leqslant \prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8} .$$

Ainsi, pour tout $s > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n} Y_i \geqslant t\right\} \leqslant \exp\left(-st + \frac{s^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right) .$$

En considérant

$$s = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2} ,$$

on optimise le membre de droite de la dernière inégalité et on conclut la preuve. □

Le théorème de Berry–Esséen

On énonce enfin le théorème de Berry–Esséen, auquel on avait fait allusion au début du paragraphe 3.4. Ce théorème donne une borne uniforme sur la différence entre deux fonctions de répartition, celle de la statistique considérée dans le théorème de la limite centrale et Φ, celle de la loi limite $\mathcal{N}(0, 1)$.

Théorème 26.11 (Berry–Esséen) *Soit $(Y_1, \ldots, Y_n)$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées, de loi commune admettant un moment d'ordre deux et dont on note respectivement μ et σ^2 la moyenne et la variance. On a alors pour tout $n \geqslant 1$,*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left\{\frac{\sqrt{n}}{\sigma}\left(\overline{Y}_n - \mu\right) \leqslant t\right\} - \Phi(t)\right| \leqslant 3\,\frac{\mathbb{E}\left[\left|Y_1 - \mu\right|^3\right]}{\sigma^3\sqrt{n}} .$$

On ne propose pas la preuve de ce résultat, mais le lecteur pourra se référer par exemple à [Dur96, paragraphe 2.4.d], qui fournit une preuve complète donnant la constante 3 et indique que cette dernière peut être améliorée (dans l'état actuel de la recherche, sans connaître sa valeur optimale, on sait qu'elle est plus petite que 0.8).

CHAPITRE 27

Compléments sur l'estimation

Résumé. Ce chapitre complète le chapitre 2. Plus précisément, ce qui suit prouve deux assertions énoncées à l'exemple 2.16 et complète la fin du pararaphe 2.5.2 (dont on reprend ici les notations) en étudiant l'estimation optimale dans le cadre des modèles exponentiels.

27.1 Preuve des assertions de l'exemple 2.16

On souhaite estimer θ uniquement à l'aide de l'observation de S_n, qui suit une loi binomiale de paramètres $n \in \mathbb{N}^*$ et $\theta \in \Theta =]0, 1[$. Dans ce cadre, on souhaite montrer dans un premier temps que $\widehat{g} = S_n/n$ est l'unique estimateur sans biais construit sur S_n.

Soit $T_n = h(S_n)$ un autre estimateur sans biais de θ, où $h : \{0, \dots, n\} \to \mathbb{R}$. On note f l'application définie sur $\{0, \dots, n\}$ par $f(x) = x/n - h(x)$. Il s'agit de voir que, dans notre cadre, f est la fonction nulle. Or, $\widehat{g}$ et T_n étant sans biais, on a que pour tout $\theta \in]0, 1[$,

$$\mathbb{E}_\theta\big[f(S_n)\big] = \mathbb{E}_\theta\big[\widehat{g} - T_n\big] = 0\,, \qquad \text{soit} \qquad \sum_{k=0}^{n} f(k)\, C_n^k \theta^k (1-\theta)^{n-k} = 0\,.$$

Pour tout θ, en développant $(1-\theta)^{n-k}$ selon la formule du binôme et en regroupant selon les puissances de θ, il vient

$$\begin{aligned}\sum_{k=0}^{n} f(k)\, C_n^k\, \theta^k (1-\theta)^{n-k} &= \sum_{k=0}^{n} \theta^k\, f(k) C_n^k \left(\sum_{j=0}^{n-k} C_{n-k}^j (-\theta)^j \right) \\ &= \sum_{m=0}^{n} \theta^m \left(\sum_{j,k:\, k+j=m} C_n^k C_{n-k}^j (-1)^j f(k) \right)\,.\end{aligned}$$

Cette fonction polynomiale étant nulle sur l'intervalle $]0, 1[$, on a par identification que tous les coefficients sont nuls : pour tout $m \in \{0, \dots, n\}$,

$$\sum_{j,k:\, k+j=m} C_n^k C_{n-k}^j (-1)^j f(k) = 0\,.$$

On en tire bien alors par récurrence que f est la fonction nulle.

On montre de manière similaire que pour l'estimation de $\sqrt{\theta}$, il n'existe pas d'estimateur sans biais construit uniquement sur l'observation de S_n. En effet, cela reviendrait à écrire $\sqrt{\theta}$ comme un polynôme $P(\theta)$ sur l'intervalle $]0, 1[$.

27.2 Estimation optimale dans les modèles exponentiels

On commence par donner la définition de ces derniers dans le cas unidimensionnel, $\Theta \subset \mathbb{R}$. Elle se généralise naturellement au cas multidimensionnel, voir [Fou02].

Définition 27.1 On dit que $(\Omega,\ \mathcal{A},\ \mathbb{P}_\theta,\ \theta \in \Theta)$ est un modèle exponentiel s'il existe une mesure dominante commune μ et quatre applications mesurables, ϕ à valeurs dans $\mathbb{R}_+$, d'une part, et h, a et b à valeurs dans $\mathbb{R}$, d'autre part, telles que pour tout $\theta \in \Theta$, la densité f_θ de $\mathbb{P}_\theta$ par rapport à μ s'écrit, pour tout $x \in \Omega$,

$$f_\theta(x) = \phi(x)\ \exp\big(a(\theta)h(x) - b(\theta)\big)\ .$$

Exemple 27.2 Un modèle poissonnien forme un modèle exponentiel, puisque pour la mesure de comptage μ sur $\mathbb{N}$, pour tout paramètre $\theta > 0$ et pour tout point $x \in \mathbb{N}$,

$$f_\theta(x) = e^{-\theta}\ \frac{\theta^x}{x!} = \frac{1}{x!}\ \exp\big(x \log\theta - \theta\big)\ .$$

C'est également le cas d'un modèle associé à un n–échantillon poissonnien : pour la mesure de comptage μ sur $\mathbb{N}^n$, pour tout paramètre $\theta > 0$ et pour toute réalisation $x = (x_1, \ldots, x_n) \in \mathbb{N}^n$,

$$f_\theta(x) = \frac{1}{\prod_{i=1}^n x_i!}\ \exp\left(\sum_{i=1}^n x_i \log\theta - n\theta\right)\ .$$

Les modèles exponentiels fournissent naturellement des estimateurs UMVU, ainsi que le montre le résultat suivant. Ce dernier justifie également le choix, dans la définition des modèles exponentiels, de la notation h jusqu'à présent dévolue à la définition d'un estimateur.

Théorème 27.3 *Soit une observation X associée à un modèle exponentiel ayant la forme indiquée à la définition 27.1. On suppose en outre que*

- *$h(X) \in \mathbb{L}^2(\mathbb{P}_\theta)$, pour tout $\theta \in \Theta$,*
- *a et b sont de classe C^1 sur l'ouvert Θ et a' ne s'annule pas sur Θ.*

Alors l'estimateur $\widehat{g} = h(X)$ est un estimateur UMVU *de $g(\theta) = b'(\theta)/a'(\theta)$.*

Preuve On veut appliquer le Théorème 2.22. Les fonctions a et b étant dérivables sur Θ, l'application $\theta \mapsto f_\theta$ l'est également. Les supports des fonctions f_θ ne dépendent que de ϕ et sont en particulier indépendants de θ ; ils sont égaux au support S de ϕ. On prouve maintenant la condition primitive indiquée dans la remarque située après l'énoncé de l'hypothèse 2.18. On fixe un point θ et un voisinage compact $K \subset \Theta$. Par continuité de a, il existe θ^* et θ_* tels que $a(K) \subset \big[a(\theta^*),\ a(\theta_*)\big]$. Par ailleurs, si on note

$$M = \max\left\{e^{b(\theta_*)},\ e^{b(\theta^*)}\right\}\ \max_{\theta\in K} e^{-b(\theta)}\ ,$$

on peut majorer les densités, pour $\theta \in K$, par

$$f_\theta(x) \leqslant \phi(x)\exp\bigl(a(\theta^*)h(x)-b(\theta)\bigr)+\phi(x)\exp\bigl(a(\theta_*)h(x)-b(\theta)\bigr) \leqslant M\bigl(f_{\theta^*}(x)+f_{\theta_*}(x)\bigr)\,.$$

Ainsi, notant A et B les maxima de $|a'|$ et $|b'|$ (qui sont des fonctions continues) sur le compact K, il vient pour tout $\theta \in K$,

$$\left|\frac{\partial}{\partial\theta} f_\theta(x)\right| = \left|\bigl(a'(\theta)h(x)-b'(\theta)\bigr)\, f_\theta(x)\right| \leqslant M\Bigl(A\,|h(x)|+B\Bigr)\,\bigl(f_{\theta^*}(x)+f_{\theta_*}(x)\bigr)\,.$$

On en déduit, puisque $h(x) \in \mathbb{L}^2(\mathbb{P}_{\theta'})$ pour tout θ',

$$h\,\sup_{\theta'\in K}\frac{\partial}{\partial\theta} f_{\theta'} \in \mathbb{L}^1(\mu) \qquad \text{et} \qquad \sup_{\theta'\in K}\frac{\partial}{\partial\theta} f_{\theta'} \in \mathbb{L}^1(\mu)\,.$$

Le théorème de dérivation des intégrales à paramètre (Théorème 26.1) assure alors que l'hypothèse 2.18 est vérifiée notamment pour h, $h\,\mathbb{I}_S$, $\mathbb{I}_\Omega$ et $\mathbb{I}_S$, et en fait pour toute variable aléatoire appartenant à tous les $\mathbb{L}^2(\mathbb{P}_{\theta'})$.

En particulier, la remarque 2.20 montre que $h(X)$ est sans biais pour l'estimation de $b'(\theta)/a'(\theta)$:

$$\begin{aligned} 0 = \int\left(\frac{\partial}{\partial\theta} f_\theta(x)\right)\mathbb{I}_S(x)\,\mathrm{d}\mu(x) &= \int\bigl(a'(\theta)h(x)-b'(\theta)\bigr)\, f_\theta(x)\,\mathrm{d}\mu(x)\\ &= a'(\theta)\,\mathbb{E}_\theta\bigl[h(X)\bigr]-b'(\theta)\,. \end{aligned}$$

Ensuite, le Théorème 2.22 s'applique. Il suffit donc de montrer que

$$\mathrm{Var}_\theta\bigl(h(X)\bigr) = \frac{\bigl(g'(\theta)\bigr)^2}{I(\theta)}\,.$$

Or, cela découle d'une part de la remarque 2.20, qui indique que

$$I(\theta) = \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right) = a'(\theta)^2\,\mathrm{Var}_\theta\bigl(h(X)\bigr)\,,$$

et d'autre part de l'égalité suivante, qui procède du fait que l'hypothèse 2.18 est vérifiée pour $\mathbb{I}_\Omega$:

$$\begin{aligned} g'(\theta) = \frac{\partial}{\partial\theta}\int h(x)\, f_\theta(x)\,\mathrm{d}\mu(x) &= \int h(x)\,\bigl(a'(\theta)\,h(x)-b'(\theta)\bigr) f_\theta(x)\,\mathrm{d}\mu(x)\\ &= a'(\theta)\,\mathbb{E}_\theta\Bigl[h^2(X)\Bigr]-b'(\theta)\,\mathbb{E}_\theta\bigl[h(X)\bigr] = a'(\theta)\,\mathrm{Var}_\theta\bigl(h(X)\bigr)\,. \end{aligned}$$

□

Exemple 27.4 Dans un modèle poissonnien (voir les exemples 2.24 et 27.2) où l'on dispose de l'observation d'un n–échantillon $X = (X_1, \ldots, X_n)$, l'estimateur $\overline{X}_n$ est bien un estimateur UMVU de $g(\theta) = b'(\theta)/a'(\theta) = \theta$.

Le Théorème 27.3 admet la réciproque suivante.

Théorème 27.5 *On suppose qu'un modèle vérifie les conditions de la remarque 2.20, que l'hypothèse 2.18 est en outre vérifiée pour $h\mathbb{I}_S$, que*

$$\theta \in \Theta \longmapsto \frac{\partial}{\partial\theta} \log f_\theta$$

est continue sur le support S commun aux lois du modèle et qu'il existe un estimateur $h(X)$ sans biais et non trivial sur S, au sens où pour tout ensemble μ–mesurable A,

$$\mu\big(S \setminus A\big) = 0 \qquad \text{implique} \qquad \text{il existe } x_1,\ x_2 \in S \cap A \text{ tels que } h(x_1) \neq h(x_2)\ .$$

Si pour tout $\theta \in \Theta$, cet estimateur $h(X)$ atteint la minoration prescrite par le Théorème 2.22, alors le modèle statistique considéré est un modèle exponentiel associé à h ; c'est-à-dire qu'il existe trois applications mesurables, ϕ à valeurs dans $\mathbb{R}_+$ et a et b à valeurs dans $\mathbb{R}$, telles que pour tout $\theta \in \Theta$ et pour tout x,

$$f_\theta(x) = \phi(x) \exp\big(a(\theta)h(x) - b(\theta)\big)\ .$$

Preuve Si la minoration du Théorème 2.22 est atteinte pour un certain θ, alors sa preuve montre que l'on est dans le cas d'égalité de l'inégalité de Cauchy-Schwarz et que donc pour ce θ, il existe deux constantes notées $a^*(\theta)$ et $b^*(\theta)$ telles que

$$\frac{\partial}{\partial\theta} \log f_\theta(X) = a^*(\theta)h(X) - b^*(\theta) \qquad \mathbb{P}_\theta\text{–p.s.}$$

Ceci est vrai pour tout θ par hypothèse. On note

$$A_\theta = \left\{ x :\ \frac{\partial}{\partial\theta} \log f_\theta(x) = a^*(\theta)h(x) - b^*(\theta) \right\}\ ;$$

cet ensemble est de probabilité pleine pour $\mathbb{P}_\theta$.

On va montrer que a^* et b^* sont continues, donc mesurables. Il s'agira également de voir que

$$A^{**} = \left(\bigcap_{\theta\in\Theta} A_\theta \right) \cap S$$

est tel que son complémentaire à S est de μ–mesure nulle, $\mu\big(S \setminus A^{**}\big) = 0$. On aura alors le résultat attendu par intégration.

Or, pour tout $\theta \in \Theta$,

$$0 = \mathbb{P}_\theta\big(A_\theta^c\big) = \mathbb{P}_\theta\big(A_\theta^c \cap S\big)\ , \qquad \text{soit} \qquad \mu\big(A_\theta^c \cap S\big) = 0$$

vu que μ–presque partout, $f_\theta > 0$ sur le support commun S.

Fixons $\theta_0 \in \Theta$ et montrons à présent la continuité de a et b en θ_0. On considère une suite $(\theta_k)_{k\geqslant 1}$ de points de Θ convergeant vers θ_0 et on pose

$$A^* = \bigcap_{k\geqslant 0} A_{\theta_k} \ .$$

D'après ce qui précède, et par union dénombrable, $\mu\big(S \setminus A^*\big) = \mu\left(\bigcup_{k\geqslant 0} S \setminus A_{\theta_k}\right) = 0$. Par hypothèse de non-trivialité de h sur S, il existe alors x_1 et x_2 dans $S \cap A^*$ tels que $h(x_1) \neq h(x_2)$. En écrivant le système d'équations vérifié pour tout $\theta = \theta_k$ (avec $k \geqslant 0$),

$$\left\{ \begin{array}{rcl} \dfrac{\partial}{\partial\theta} \log f_\theta(x_1) & = & a^*(\theta)h(x_1) - b^*(\theta) \\ \dfrac{\partial}{\partial\theta} \log f_\theta(x_2) & = & a^*(\theta)h(x_2) - b^*(\theta) \end{array} \right.$$

et en le résolvant en $a^*(\theta)$ et $b^*(\theta)$, on écrit chacune de ces dernières comme une combinaison linéaire de la forme

$$\gamma_1 \frac{\partial}{\partial\theta} \log f_\theta(x_1) + \gamma_2 \frac{\partial}{\partial\theta} \log f_\theta(x_2)$$

où les coefficients γ_1 et γ_2 dépendent de $h(x_1)$ et $h(x_2)$, mais pas de θ_k. Ces égalités sont donc vérifiées pour les θ_k (avec $k \geqslant 1$) et pour θ_0 : cela montre, par continuité séquentielle des fonctions

$$\theta \in \Theta \mapsto \frac{\partial}{\partial\theta} \log f_\theta(x_1) \qquad \text{et} \qquad \theta \in \Theta \mapsto \frac{\partial}{\partial\theta} \log f_\theta(x_2) \ ,$$

que a^* et b^* sont elles-mêmes séquentiellement continues et donc continues en θ_0. Ceci étant vrai pour tout θ_0, ces fonctions sont donc continues sur Θ.

Θ étant séparable, on en fixe un sous-ensemble dénombrable dense Γ. Un argument de continuité permet alors de conclure à l'égalité

$$A^{**} = \left(\bigcap_{\theta\in\Gamma} A_\theta\right) \cap S$$

et donc, par union dénombrable, au fait que $\mu\big(S \setminus A^{**}\big) = 0$, ce qui conclut la preuve. □

Le Théorème 27.3 fournit un estimateur optimal pour une quantité que le statisticien ne choisit pas et qui se déduit de la forme du modèle. Cela étant, l'objectif naturel d'estimer directement le paramètre $g(\theta) = \theta$ de manière optimale reste possible, mais dans un cadre asymptotique. On recourt à cet effet à l'estimateur du maximum de vraisemblance. On notera qu'une hypothèse de régularité supplémentaire s'ajoute aux conditions du Théorème 27.3.

Théorème 27.6 *Soit une observation X associée à un modèle exponentiel ayant la forme indiquée à la définition 27.1. On suppose en outre que*

- *$h(X) \in \mathbb{L}^2(\mathbb{P}_\theta)$, pour tout $\theta \in \Theta$,*
- *a et b sont de classe C^1 sur l'ouvert Θ et a' ne s'annule pas sur Θ,*

– *la fonction $\varphi = b'/a'$ est de classe C^1 sur Θ, de dérivée ne s'annulant en aucun point.*

Si $(X_1, \ldots, X_n)$ est un échantillon loi $\mathbb{P}_\theta$ et si l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ construit à partir de cet échantillon existe, alors il vérifie

$$\sqrt{n}\Big(\widehat{\theta}_n - \theta\Big) \rightsquigarrow \mathcal{N}\left(0,\ \frac{1}{I(\theta)}\right) .$$

Preuve La log–vraisemblance associée au modèle est

$$L_{X_1,\ldots,X_n}(\theta) = \sum_{i=1}^{n} \big(\log \phi(X_i) + a(\theta)h(X_i) - b(\theta)\big) .$$

L'estimateur du maximum de vraisemblance (qui existe par hypothèse) annule sa dérivée, soit

$$a'\Big(\widehat{\theta}_n\Big)\,\overline{h}_n - b'\Big(\widehat{\theta}_n\Big) = 0 , \qquad \text{où} \quad \overline{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i) .$$

Vu les hypothèses de régularité sur φ, cette dernière admet un inverse, noté φ^{-1}, ce qui conduit à l'expression

$$\widehat{\theta}_n = \varphi^{-1}\Big(\overline{h}_n\Big) .$$

Or, le théorème de limite centrale assure que

$$\sqrt{n}\Big(\overline{h}_n - \mathbb{E}_\theta\big[h(X_1)\big]\Big) \rightsquigarrow \mathcal{N}\Big(0,\ \mathrm{Var}_\theta\big(h(X_1)\big)\Big) ,$$

où la preuve du Théorème 27.3 montre que $\mathbb{E}_\theta\big[h(X_1)\big] = b'(\theta)/a'(\theta) = \varphi(\theta)$. Les hypothèses de régularité sur φ montrent que φ^{-1} est également de classe C^1 sur l'ouvert Θ, de sorte qu'une application de la méthode delta (Théorème 2.9) conduit à la convergence en loi

$$\sqrt{n}\Big(\widehat{\theta}_n - \theta\Big) = \sqrt{n}\left(\varphi^{-1}\Big(\overline{h}_n\Big) - \theta\right) \rightsquigarrow \mathcal{N}\left(0,\ \frac{\mathrm{Var}_\theta\big(h(X_1)\big)}{\varphi'(\theta)^2}\right) = \mathcal{N}\left(0,\ \frac{1}{I(\theta)}\right) .$$

Le calcul de la variance asymptotique procède des résultats donnés par les deux dernières égalités de la preuve du Théorème 27.3, que nous réécrivons pour mémoire,

$$I(\theta) = a'(\theta)^2\,\mathrm{Var}_\theta\big(h(X)\big) \qquad \text{et} \qquad \varphi'(\theta) = a'(\theta)\,\mathrm{Var}_\theta\big(h(X)\big) .$$

□

CHAPITRE 28

Machines à sous (compléments)

Résumé. Ce qui suit complète le chapitre 22. On explique ici brièvement comment rendre non-asymptotiques les résultats de convergence qui reposaient sur la loi des grands nombres pour les martingales ; au passage, nous allons même exhiber des vitesses de convergence. Pour cela, nous allons recourir aux inégalités de Hoeffding–Azuma et de Bernstein pour les martingales. (Tout ceci n'a évidemment qu'une visée uniquement culturelle et est d'un niveau master 2.) On conclut en donnant les codes `Matlab` des simulations effectuées au chapitre 22.

28.1 Compléments mathématiques

28.1.1 Inégalité de Hoeffding–Azuma

Enoncé

Nous allons employer cette inégalité pour préciser (22.2). A cet effet, nous l'énonçons tout d'abord, en soulignant qu'elle forme une généralisation de l'inégalité de Hoeffding (présentée au Lemme 3.7).

Théorème 28.1 (Inégalité de Hoeffding–Azuma) *Soit $(Y_t)_{t\geqslant 1}$ une suite d'accroissements de martingale par rapport à la filtration $(\mathcal{F}_t)_{t\geqslant 1}$, tels que pour tout $t \geqslant 1$, il existe a_t, b_t avec $a_t \leqslant Y_t \leqslant b_t$ p.s. Alors, pour tout $\lambda > 0$,*

$$\mathbb{P}\left\{\sum_{t=1}^{n} Y_t \geqslant \lambda\right\} \leqslant \exp\left(-\frac{2\lambda^2}{\sum_{t=1}^{n}(b_t - a_t)^2}\right) .$$

En particulier, sous les hypothèses du théorème, si $a_t = a$ et $b_t = b$ pour tout $t \geqslant 1$, on a, pour tout $n \geqslant 1$ et avec probabilité au moins $1 - \delta$,

$$\sum_{t=1}^{n} Y_t \leqslant (b-a)\sqrt{\frac{n}{2}\log\frac{1}{\delta}} .$$

Application

On reprend les notations de la preuve du Corollaire 22.3 et il s'agit de préciser l'énoncer $M_n = o(n)$ p.s. : on va montrer que

$$M_n = O\left(\sqrt{n\log n}\right) \qquad \text{p.s.}$$

Preuve On a $-1 \leqslant Y_t \leqslant 1$ p.s., de sorte que $M_n \leqslant \sqrt{2n\log(2/\delta)}$ avec probabilité au moins $1-\delta/2$. Symétriquement, en appliquant l'inégalité aux $-Y_t$, on a également $M_n \geqslant -\sqrt{2n\log(2/\delta)}$ avec probabilité au moins $1-\delta/2$. Au final, avec probabilité au moins $1-\delta$,

$$\frac{|M_n|}{\sqrt{2n\log\dfrac{2}{\delta}}} \leqslant 1\,.$$

En choisissant $\delta = \delta_n = 1/n^2$, le lemme de Borel–Cantelli indique que

$$\mathbb{P}\left\{\limsup \frac{|M_n|}{\sqrt{2n\log\left(2n^2\right)}} > 1\right\} \leqslant \mathbb{P}\left(\limsup\left\{\frac{|M_n|}{\sqrt{2n\log\left(2n^2\right)}} > 1\right\}\right) = 0\,.$$

Ainsi, on vient de montrer que

$$\limsup\ \frac{|M_n|}{\sqrt{2n\log\left(2n^2\right)}} \leqslant 1 \quad \text{p.s.},$$

ce qui était le résultat recherché. □

Cette vitesse, qui est essentiellement $\sqrt{n\log n}$, ne doit pas être surprenante : c'est, à facteur logarithmique près, celle du théorème de la limite centrale et celle de la loi du logarithme itéré. A vrai dire, une formulation dite maximale de l'inégalité de Hoeffding–Azuma permettrait de renforcer le résultat précédent en remplaçant la vitesse $\sqrt{n\log n}$ obtenue par $\sqrt{n\log\log n}$, qui est cette fois-ci exactement celle de la loi du logarithme itéré. Cette dernière est bien connue pour des sommes de variables aléatoires indépendantes et identiquement distribuées, voir par exemple [vdV98, Proposition 2.26], mais elle est également vraie pour les martingales.

28.1.2 Inégalité de Bernstein–Freedman pour les martingales

Enoncé

Nous allons l'employer pour préciser les vitesses de convergence des estimateurs $\widehat{\theta}_n^A$ et $\widehat{\theta}_n^B$ vers, respectivement, θ^A et θ^B. Etant donnée une filtration $(\mathcal{F}_t)$, on rappelle qu'on note $\mathbb{E}_t$ l'espérance conditionnelle par rapport à $\mathcal{F}_{t-1}$.

Théorème 28.2 (Inégalité de Bernstein–Freedman) *Soient $(Y_t)_{t\geqslant 1}$ une suite d'accroissements de martingale par rapport à la filtration $(\mathcal{F}_t)_{t\geqslant 1}$ et K une constante telles que $Y_t \leqslant K$ p.s. pour tout $t \geqslant 1$. On note $\mathrm{Var}_t\, Y_t$ la variance conditionnelle de Y_t,*

$$\mathrm{Var}_t\, Y_t = \mathbb{E}_t\left[Y_t^2\right].$$

Alors, pour tout $n \geqslant 1$, avec probabilité au moins $1-\delta$,

$$\sum_{t=1}^{n} Y_t \leqslant \square\sqrt{\left(K^2 + \sum_{t=1}^{n}\mathrm{Var}_t\, Y_t\right)\log\frac{n}{\delta}} + \square\, K\log\frac{n}{\delta}$$

où les $\square$ désignent des constantes universelles que l'on n'a pas cru bon d'expliciter.

Notons que l'inégalité de Bernstein est généralement plus fine que celle de Hoeffding–Azuma, en tout cas, dès que

$$\sum_{t=1}^{n} \mathrm{Var}_t\, Y_t \ll \frac{n}{\log n} \qquad \text{p.s.}$$

Application

On reprend les notations de la preuve du Lemme 22.4 et il s'agit d'en préciser l'assertion $M_n^A = o\big(N_n^A\big)$ située à sa fin ; on va montrer que sous l'hypothèse $N_n^A \gg \log n$, on a

$$M_n^A = O\left(\sqrt{N_n^A \log n}\right) \qquad \text{p.s.}$$

L'hypothèse est vérifiée dès que (c_k) croît sous-exponentiellement, par exemple, polynomialement : on a vu en effet au paragraphe 22.1.7 que $N_n^A \geqslant c_A^{-1}(n) \geqslant \big(c^{-1}(n) - 1\big)/2$.

> Remarque au passage. Avant de prouver ce raffinement par application de l'inégalité de Bernstein, on peut se demander pourquoi l'on n'appliquerait pas l'inégalité de Hoeffding–Azuma ; c'est parce que cette dernière donnerait, de la même manière qu'au paragraphe précédent, $M_n^A = O\big(\sqrt{n \log n}\big)$ p.s. Or, cela n'est pas assez précis, puisque l'on a $N_n^A \ll n$ dès que A n'est pas le bon bras et que la stratégie remplit l'objectif du paragraphe 22.1.4.

Preuve L'inégalité de Bernstein pour les martingales livre que pour tout n fixé, avec probabilité au moins $1 - \delta$,

$$M_n^A \leqslant \square \sqrt{\big(1 + N_n^A\big) \log \frac{n}{\delta}} + \square \log \frac{n}{\delta}$$

(où les $\square$ désignent des constantes universelles), et une inégalité symétrique sur $-M_n^A$. En appliquant le lemme de Borel–Cantelli avec $\delta = \delta_n = 1/n^2$, il vient

$$\limsup \frac{M_n^A}{\log n + \sqrt{N_n^A \log n}} \leqslant \square \qquad \text{p.s.}$$

où l'on a utilisé que N_n^A domine 1, et où $\square$ désigne toujours une constante universelle. Ainsi, $M_n^A = O\big(\log n + \sqrt{N_n^A \log n}\big)$ est $O\big(\sqrt{N_n^A \log n}\big) = o\big(N_n^A\big)$ dès que la condition $N_n^A \gg \log n$ est remplie. □

28.2 Codes des simulations

28.2.1 Fonctions auxiliaires : génération des paiements, moyenne d'un vecteur éventuellement vide

```
function P = Paiements(p, q, n)
```

```
%
% Entrées :
%   p est le paramètre de Bernoulli du premier bras,
%   q celui du second,
%   n est le nombre de tours de jeu
% Sortie :
%   Une matrice P à deux lignes et n colonnes,
%   composée de 0 et de 1
%

  P = [rand(1,n) < p;
       rand(1,n) < q];

end



function m = meanV(P)
%
% Entrées :
%   P est un vecteur de nombres réels
% Sortie :
%   La moyenne m des éléments de P si P est non vide,
%   m = 0 sinon
%
  if (sum(P) == 0)
    m = 0;
  else
    m = mean(P);
  end

end
```

28.2.2 Stratégie N, et stratégies dérivées EN et S

```
function [G, C] = Algo_N(P, C, G)
%
% Entrées :
%   P est la matrice des paiements, à deux lignes et n colonnes
%   Le vecteur C contient la suite des bras déjà tirés (index 1 ou 2)
%   en cas d'exploration préalable, C = [] sinon lors de l'appel
%   Le vecteur G contient la suite des gains déjà tirés
%   en cas d'exploration préalable, G = [] sinon lors de l'appel
% Sortie :
%   G, le vecteur des paiements obtenus à chaque tour
%   par l'algorithme naif de pure exploitation (N)
```

```
%   C, la suite des bras choisis
%

  n = size(P, 2);
  % Exploitation pure (à tous les tours de jeu après l'exploration
  % éventuelle incluse dans C)
  for j = (1+length(C)):n
    % On calcule les moyennes des bras et on les compare,
    % ce qui donne l'action c (aléatoire en cas d'égalité)
    mA = meanV(G(C == 1));
    mB = meanV(G(C == 2));
    c = 1 + ((mB > mA) + (rand < 1/2) * (mB == mA));
    C = [C c];
    G = [G P(c, j)];
  end

end



function [G, C] = Algo_EN(P)
%
% Entrées :
%   P est la matrice des paiements, à deux lignes et n colonnes
% Sortie :
%   G, le vecteur des paiements obtenus à chaque tour
%   par l'algorithme (EN) composé de 2 x 10 coups
%   d'exploration (non alternés), suivis par une procédure
%   du type algorithme naif de pure exploitation
%   C, la suite des bras choisis
%

  n = size(P, 2);
  G = [P(1, 1:10), P(2, 11:20)];
  C = [ones(1,10), ones(1,10)*2];
  [G, C] = Algo_N(P, C, G);

end



function [G, C] = Algo_S(P)
%
% Entrées :
%   P est la matrice des paiements, à deux lignes et n colonnes
%   (pour cet algo, il est bon qu'elle soit composée de 0 et 1 uniquement)
```

```
% Sortie :
%   G, le vecteur des paiements obtenus à chaque tour
%   par l'algorithme de type théorème du scrutin (S)
%   C, la suite des bras choisis
%

  % On convertit les paiements 0/1 en paiements -1/1,
  % on lance l'algorithme naïf de décision par exploitation
  % pure, et on re-convertit les paiements
  [G, C] = Algo_N(2*P-1, [], []);
  G = (G + 1)/2;

end
```

28.2.3 Stratégie excitée X

```
function [G, C] = Algo_X(P, EA, EB)
%
% Entrées :
%   P est la matrice des paiements, à deux lignes et n colonnes
%   EA, temps d'exploration de A (éléments entre 1 et n)
%   EB, temps d'exploration de B (éléments entre 1 et n)
% Sortie :
%   G, le vecteur des paiements obtenus à chaque tour
%   par l'algorithme avec excitation (X)
%   C, la suite des bras choisis
%

  n = size(P, 2);
  C = [];
  G = [];
  for j = 1:n
    if any(EA == j)
      c = 1;
    elseif any(EB == j)
      c = 2;
    else
      % On calcule les moyennes des bras et on les compare,
      % ce qui donne l'action c (aléatoire en cas d'égalité)
      mA = meanV(G(C == 1));
      mB = meanV(G(C == 2));
      c = 1 + ((mB > mA) + (rand < 1/2) * (mB == mA));
    end
    C = [C c];
    G = [G P(c, j)];
  end
```

```
end
```

28.2.4 Stratégies U et E

```
function [G, C] = Algo_UCB(P)
%
% Entrées :
%   P est la matrice des paiements, à deux lignes et n colonnes
% Sortie :
%   G, le vecteur des paiements obtenus à chaque tour
%   par l'algorithme naif de pure exploitation (N)
%   C, la suite des bras choisis
%

  n = size(P, 2);
  % On joue d'abord chaque bras, puis on exploite
  % en considérant des bornes supérieures d'intervalles de confiance
  C = [1 2];
  G = [P(1,1) P(2,2)];
  for j = 3:n
    nA = (C == 1);
    mA = mean(G(nA)) + sqrt(2*log(j-1)/length(nA));
    nB = (C == 2);
    mB = mean(G(nB)) + sqrt(2*log(j-1)/length(nB));
    c = 1 + ((mB > mA) + (rand < 1/2) * (mB == mA));
    C = [C c];
    G = [G P(c, j)];
  end

end




function [G, crit] = Algo_E(P, eps)
%
% Entrées :
%   P est la matrice des paiements, à deux lignes et n colonnes
%   eps > 0 est le paramètre d'approximation
% Sortie :
%   G, le vecteur des paiements obtenus à chaque tour
%   par l'algorithme de pure exploration (E)
%   crit, variable booléenne indiquant si l'exploration
%   a eu le temps de prendre fin
%

  n = size(P, 2);
```

```
  C = [];
  G = [];
  % Phase préalable d'exploitation pure
  % crit est une variable booléenne indiquant si l'exploration doit
  % être poursuivie
  crit = 1;
  % Pas de temps courant
  t = 1;
  while crit & (t <= n-1)
    C = [C 1 2];
    G = [G P(1,t) P(2,t+1)];
    t = t + 2;
    mA = meanV(G(C == 1));
    mB = meanV(G(C == 2));
    crit = (abs(mA - mB) <= sqrt((2/t)*log(pi^2*t^2/(3*eps))));
  end
  % On est ensuite en exploitation pure
  if not(crit)
    c = 1 + (mB > mA);
    G = [G P(c, t:n)];
  end

end
```

28.2.5 Fichier maître (script) pour le choix de la forme des instants d'excitation

```
m = 1000;
n = 300;

% Différentes suites d'instants d'excitation
% On joue le premier bras aux temps définis par
% les Ej et le second aux temps Ej+1
E1 = (1:sqrt(n)).^2;
E2 = [(1:4)*2-1 8+(1:sqrt(n)).^2];
E3 = 2.^(1:(log2(n)+1));
E4 = (1:n^(1/3)).^3;

% On refait m fois la simulation pour calculer
% une approximation de l'espérance du gain moyen
R = zeros(4, n);
for j = 1:m
  P = Paiements(0.45, 0.30, n);
  R(1,:) = R(1,:) + Algo_X(P, E1, E1+1);
  R(2,:) = R(2,:) + Algo_X(P, E2, E2+1);
  R(3,:) = R(3,:) + Algo_X(P, E3, E3+1);
  R(4,:) = R(4,:) + Algo_X(P, E4, E4+1);
end
```

```
% Suite des approximations des gains moyens en fonction du temps
Z = cumsum(R, 2)./repmat(m*(1:n), [4 1]);
% On élimine les premiers pas de temps et on
% commence le dessin au 50ème tour
d = 50;
plotstyle = {'--', '-', '-.', ':'};
clf; hold on;
for k = 1:4
  plot(d:n, Z(k, d:n), plotstyle{k}, 'LineWidth', 1.7);
end
legend('c_k = k^2', 'Expl. puis c_k = k^2', '2^k', 'k^3')
xlabel('Pas de temps')
ylabel('Gain moyen')
```

28.2.6 Fichier maître (script) pour le tracé de la loi de la statistique asymptotiquement normale

```
clf, clc, clear;

m = 10000;
n = 300;

G = [];
% On refait simplement m simulations de l'algorithme
% par excitation
for j = 1:m
  P = Paiements(0.45, 0.30, n);
  E = 2.^(1:(log2(n)+1));
  G = [G sum(Algo_X(P, E, E+1))];
end

% Et on construit alors l'histogramme de la statistique d'intérêt
G = (G - 0.45*n)/sqrt(n);
histo(G, 100, [], 1);
title(strcat([ 'Variance (obs. non seuillées) ', num2str(var(G),3), ...
               ' -- Variance (obs. seuillées) ', ...
               num2str(var(G(G >= -1.5)), 3) ]));
```

28.2.7 Fichier maître (script) pour la comparaison des six stratégies

```
clc, clear;

m = 1000; % Ou 1, ou 100, ou 10000, selon les simulations
n = 300;

% On initialise les variables
```

```
R = zeros(6, n);
prb = zeros(6, 1);
M = zeros(6, 1);

% On répète m fois la simulation : on tire une suite de couples
% de paiements, on lance les algorithmes et on évalue leur performance :
% suites des gains moyens, probabilité de jouer le bon bras au tour n,
% probabilité d'être le meilleur algorithme
for j = 1:m
  % Cf. les paramètres doivent être < 1/2, sinon le casino est perdant
  % On a fait varier les couples de paramètres : (0.45, 0.30) mais
  % aussi (0.45, 0.05) et (0.45, 0.40)
  P = Paiements(0.45, 0.05, n);
  [R1, C1] = Algo_N(P, [], []);
  [R2, C2] = Algo_EN(P);
  E = 2.^(1:(log2(n)+1));
  [R3, C3] = Algo_X(P, E, E+1);
  [R4, crit] = Algo_E(P, 0.10);
  [R5, C5] = Algo_S(P);
  [R6, C6] = Algo_UCB(P);
  % Pour le calcul de l'espérance des gains
  R = R + [R1; R2; R3; R4; R5; R6];
  % Pour le calcul de la probabilité, on prend celle
  % de jouer le bon bras au dernier tour (i.e., de ne pas
  % être dans un piège et/ou d'avoir convergé)
  prb = prb + [C1(end) == 1;
               C2(end) == 1;
               C3(end) == 1;
               1 - crit;
               C5(end) == 1;
               C6(end) == 1];
  % Détermination du meilleur algorithme
  % (on traite les cas d'égalité en mettant +1 à tous les ex-aequo !)
  r = [sum(R1); sum(R2); sum(R3); sum(R4); sum(R5); sum(R6)];
  M(r == max(r)) = M(r == max(r)) + 1;
end

% Tracés des gains moyens
Z = cumsum(R, 2)./repmat(m*(1:n), [6 1]);
% On ne représente pas les premiers pas
d = 21;
textalgo = {'N', 'EN', 'X', 'E', 'S', 'U'};
clf; hold on;
for k = 1:6
  plot(d:n, Z(k, d:n), 'b', 'LineWidth', 1.7);
  % Permet d'annoter les courbes à la main
  gtext(textalgo{k});
```

```
end
xlabel('Pas de temps')
ylabel('Gains moyens')
```

CHAPITRE 29

Estimation non-paramétrique pour le modèle de régression (compléments)

Résumé. Nous présentons ici les codes des simulations informatiques relatives au chapitre 18.

29.1 Codes des simulations

29.1.1 Construction des données artificielles

```
function Y = observations_regression(f, sigma)
%
% Entrées :
%   sigma, un réel > 0 donnant un écart-type
%   f, un vecteur de taille n représentant un un signal f échantillonné
%      aux points  t_k = k/n pour k = 1,...,n
% Sortie :
%   Y, un vecteur d'observations de taille n
%

  n = length(f);
  Y = f + sigma * randn(1,n);

end
```

29.1.2 Calcul des estimateurs

```
function [hat_theta, Phi] = FourierSurDesign(Y)
%
% Entrée :
%   Y, un vecteur d'observations de taille n
% Sorties :
%   hat_theta, un vecteur de taille n-1 donnant les estimées
%              des n-1 premiers coefficients de Fourier
%   Phi, une matrice de taille n-1 x n dont la j-ième ligne est
%        le vecteur des valeurs du j-ième élément de la base de Fourier
%        sur le design
%

  n = length(Y);
```

```
  t = (1:n)/n;     % design uniforme
  N = n-1;
  % Calcul des valeurs des vecteurs de la base sur les points du design
  Phi = ones(N, n);
  for j = 2:N
    % Note : rem(a,b) donne le reste de le division euclidienne de a par b
    %        de sorte que rem(j+1,2) est p.ex. une indicatrice du
    %        fait que j+1 est impair
    Phi(j, :) =  sqrt(2)*(cos(pi*j*t)*rem(j+1,2)+sin(pi*(j-1)*t)*rem(j,2));
  end;
  % Calcul du vecteur hat_theta des estimées des coefficients de Fourier
  for j = 1:N
    hat_theta(j) = mean(Y .* Phi(j, :));
  end;

end



function hat_f = estimation_regression(Y, N);
%
% Entrées :
%   Y, un vecteur d'observations
%   N, le nombre de coefficients de Fourier à estimer
% Sortie :
%   hat_f, un vecteur de taille n donnant la valeur de l'estimée de f
%          en les points du design
%

  n = length(Y);
  [hat_theta, Phi] = FourierSurDesign(Y);
  for j = 1:n
    hat_f(j) = hat_theta(1:N) * Phi(1:N, j);
  end

end



function N_mallows = mallows_regression(Y, sigma);
%
% Entrées :
%   Y, un vecteur d'observations de taille n
%   sigma, un niveau de bruit
% Sortie :
%   N_mallows, un nombre entier entre 1 et n-1 donné par le critère
```

```
%              de Mallows
%

  n = length(Y);
  [hat_theta, Phi] = FourierSurDesign(Y);
  CritMallows = 2*sigma^2/n * (1:(n-1)) - cumsum(hat_theta.^2);
  [C, N_mallows] = min(CritMallows);
  disp(['Valeur de N donnée par le critère de Mallows --> N = ', ...
        num2str(N_mallows)]);

end
```

29.1.3 Fonction principale de tracé et de comparaison des méthodes

```
function principal_regression(sigma, n, t, f, p)
%
% Entrées :
%   sigma, un réel donnant un niveau de bruit
%   n, un nombre d'observations,
%   t, le design
%   f, la fonction à estimer échantillonnée sur t et passée en argument
%      comme un vecteur de taille n
%   p, un nombre donnant le numéro du signal
% Sortie :
%   Aucune sortie numérique mais des tracés graphiques
%

  disp([char(13), 'Pour la fonction numéro ', num2str(p)]);
  % Calcul approché de la norme L2 du signal par méthode des rectangles
  energie = sqrt(1/n * sum(f.^2));
  disp(['La norme L2 du signal vaut environ : ', num2str(energie)]);
  % On simule le vecteur des observations
  Y = observations_regression(f, sigma);

  % Tracé du signal p et des observations associées
  figure(1); subplot(2,2,p);
  plot(t, f, 'k', t, Y, 'k--');
  title(['n = ', num2str(n), ' et \sigma = ', num2str(sigma)]);

  % Calcul des estimateurs pour toutes les valeurs N = 1, ... n-1
  % (on en a besoin pour déterminer N_opt)
  % Au passage, on calcule également leur erreur quadratique
  hat_f = zeros(n-1, n);
  e = zeros(1, n-1);
  for N = 1:(n-1)
    hat_f(N, :) = estimation_regression(Y, N);
    e(N) = sum( (f-hat_f(N,:)) .^2 ) /n;
```

```
  end;

  % Tracé de la courbe de l'erreur quadratique en fonction de N
  figure(2); subplot(2,2,p);
  plot([1:n-1], e, 'k');
  xlabel('dimensionnalité N'); ylabel('erreur');

  % Détermination de N_opt
  [e_min, N_opt] = min(e);
  disp(['Valeur optimale de N --> N_opt = ', num2str(N_opt)]);
  % Détermination de N_mallows
  N_mallows = mallows_regression(Y,sigma);
  % Autres valeurs de N
  N_floor = floor(n^(1/3));
  N_max = n-1;

  % Tracés des reconstructions associées aux différents N retenus
  N = [N_opt N_floor N_max N_mallows];
  Tt = {' N_{opt}', '\lfloor n^{1/3} \rfloor', 'n-1', 'N_{mallows}'};
  for k = 1:4
    tr_f = hat_f(N(k), :);
    figure(p+2); subplot(2,2,k);
    plot(t, f, 'k', t, tr_f, 'k--');
    title(['signal et estimateur avec N = ', Tt{k}, ' = ', num2str(N(k))]);
  end;
```

29.1.4 Programme maître de simulation (script)

```
clc; clf;

% Signal #1
clear;
p = 1; n = 100; sigma = 0.1;
t = (1:n)/n;
f = (t >= 0.2) .* (t <= 0.3);
principal_regression(sigma, n, t, f, p);

% Signal #2
clear;
p = 2; n = 100; sigma = 0.1;
t = (1:n)/n;
f = exp(-t) .* sin(4*pi*t);
principal_regression(sigma, n, t, f, p);

% Signal #3
clear;
p = 3; n = 250; sigma = 1;
```

```
t = (1:n)/n;
f = min(10, 1./abs(cos(4*pi*t)));
principal_regression(sigma, n, t, f, p);

% Signal #4
clear;
p = 4; n = 200; sigma = 3;
t = (1:n)/n;
f = 4*sin(4*pi*t) + 3*cos(6*pi*t) - 2*sin(6*pi*t);
principal_regression(sigma, n, t, f, p);
```

29.1.5 Traitement des données réelles

Fonction pour l'estimation et le tracé de l'estimateur de l'écart-type

```
function s = estimation_ec(Y);
%
% Entrée :
%   Y, un vecteur d'observations (de taille n)
% Sortie :
%   s, le vecteur des estimées de sigma (de taille n-1)
%

  n = length(Y);
  [hat_theta, Phi] = FourierSurDesign(Y);
  % On détermine la trajectoire de l'estimateur de l'écart-type
  P = cumsum(hat_theta([end:-1:1]).^2);
  s = sqrt(n * P ./ (1:(n-1)));

end
```

Fichier maître (script) pour les données réelles

```
% Etude des performances de la procédure associée au critère de Mallows
% sur le jeu de vraies données  *Sunspot_month*
clc; clear;

% Chargement des données
load Sunspot_month.mat;
Y = Sunspot_month;
% Tracé des observations
n = length(Y);
t = [1:n];
figure(1); subplot(2,2,1);
plot(t, Y, '-');
title('Observations');
```

```
% Tracé de la trajectoire de l'estimateur de l'écart-type
s = estimation_ec(Y);
figure(1); subplot(2,2,2);
plot([1:n-1], s, '-');
title('Tracé de la trajectoire de l''estimateur de \sigma');

% On va donner deux fois la main à l'utilisateur pour le choix
% délicat des estimées de l'écart-type
for j = 1:2
  sigma = input('Estimée de sigma ?  ');
  N_mallows = mallows_regression(Y, sigma);
  hat_f = estimation_regression(Y, N_mallows);
  % Tracé de l'estimateur hat_f
  figure(1); subplot(2,2,2+j);
  plot(t, hat_f, '-');
  title(['Estimateur pour \sigma = ', num2str(sigma), ...
         '  (N_{mallows} = ', num2str(N_mallows), ')']);
  % Tracés des résidus
  figure(2); subplot(2,3,j);
  plot(t, Y-hat_f);
  title(['Vraies données : \sigma = ', num2str(sigma)]);
  subplot(2,3,3+j);
  plot(hat_f,Y-hat_f,'.');
  title(['Vraies données : \sigma = ', num2str(sigma)]);
end;

% On renouvelle l'expérience avec un signal connu
Y = [];
n = 200;
sigma = 3;
t = (1:n)/n;
f = 4*sin(4*pi*t) + 3*cos(6*pi*t) - 2*sin(6*pi*t);
Y = observations_regression(f,sigma);
N_mallows = mallows_regression(Y, sigma);
hat_f = estimation_regression(Y, N_mallows);
% Tracé des résidus
figure(2); subplot(2,3,3);
plot(t, Y-hat_f);
title(['Données simulées : \sigma = ', num2str(sigma)]);
subplot(2,3,6);
plot(hat_f, Y-hat_f, '.');
title(['Données simulées : \sigma = ',num2str(sigma)]);
```

CHAPITRE 30

Inférence statistique pour des modèles censurés (compléments)

Résumé. Nous présentons ici le code des simulations informatiques relatives au chapitre 19.

30.1 Codes des simulations

30.1.1 Fonctions auxiliaires

```
function [Y, d, hl] = donnees_censure(n)
%
% Entrée :
%   n est la taille d'échantillon désirée
% Sorties :
%   Y  est le vecteur des observations censurées
%   d  est un vecteur dont chaque coordonnée vaut 0 ou 1 selon
%      qu'il y a eu censure (Y_i = C_i) ou non (Y_i = T_i)
%   hl est le vecteur des estimateurs de lambda, le t-ième
%      élément du vecteur étant construit uniquement
%      avec les t premiers éléments de Y et d
%

  % Génération de n variables exponentielles de
  % paramètres 1 (durées de vie) puis 0.5 (censure)
  T = -log(rand(1,n));
  C = -2*log(rand(1,n));
  % Observations des données censurées uniquement
  Y = min(C,T);
  d = (Y == T);
  % On utilise ici le fait que presque sûrement, Y(1) est non nul,
  % pour ne pas avoir de problème de division par 0 dans la ligne suivante
  hl = cumsum(d)./cumsum(Y);

end



function [KM1, KM2] = Kaplan_Meier(Y, d)
%
```

```
% Entrée :
%    Y et d, vecteurs d'observations (de même longueur n)
% Sortie :
%         Estimateur de Kaplan-Meier, sous la forme suivante.
%         Etant une fonction en escalier, on donne
%    KM1, les instants de sauts (vecteur de taille n) et
%    KM2, les valeurs de l'estimateur sur les n+1 intervalles
%         ainsi définis (vecteur de taille n+1)
%

  n = length(Y);
  % Réordonnement des observation et détermination de KM1
  [KM1, I] = sort(Y);
  dord = d(I);
  % Calcul de KM2
  KM2 = ones(1,n+1);
  for i = 1:n
    KM2(i+1) = KM2(i)*(1 - dord(i)/(n+1-i));
  end;

end
```

30.1.2 Illustration de la normalité asymptotique de l'estimateur de Kaplan–Meier

```
% Illustration de la normalité asymptotique de l'estimateur de Kaplan-Meier
clf, clc, clear;

N = 1000;
figure(3);
for k = 1:2
  % Valeur de n
  n = 10*(k==1) + 1000*(k==2);
  for t = 1:2
    sigma = 2/3 * exp(-2*t) * (exp(3/2*t)-1);
    % Calcul d'un N-échantillon
    Knorm = zeros(1, N);
    for m = 1:N
      [Y, d] = donnees_censure(n);
      [KM1, KM2] = Kaplan_Meier(Y, d);
      J = max(find(KM1 <= t));
      Knorm(m) = sqrt(n) * (KM2(J+1) - exp(-t)) / sqrt(sigma);
    end;
    % Tracés de la distribution empirique associée à Knorm
    % et de la distribution limite sur l'intervalle [-3 3]
    subplot(2, 2, t+2*(k-1));
    hold on;
    stairs(sort(Knorm), [1:N]/N, 'k--');
```

```
    axis([-3 3 0 1]);
    fplot('pnorm', [-3 3], 'k');
    legend('KM','limite');
    title(['t = ', num2str(t), ' et n = ',num2str(n)]);
  end;
end;
```

30.1.3 Illustration des propriétés asymptotiques des $\widehat{\lambda}_n$

```
% Illustration 1. de la consistance et 2. de la normalité asymptotique
% de \hat\lambda_n vers \lambda, dans le cas du modèle avec lois
% exponentielles
clf, clc, clear;

% 1. Illustration de la convergence presque-sûre
n = 1000;
[Y, d, hl] = donnees_censure(n);
figure(1);
plot(10:n, hl(10:n), 'k', [10, n], [1, 1], 'k:');
legend('\lambda_n', '\lambda = 1');
title('Convergence ps de l''estimateur de \lambda');

% 2. Illustration de la normalité asymptotique
%    Pour les valeurs n = 10, n = 100, n = 1000, on calcule un
%    échantillon de taille N = 1000 de \hat\lambda_n, puis
%    les deux variables aléatoires d'intérêt
N = 1000;
figure(2);
for k = [1, 2, 3]          %n sera 10^k
  hlnorm = zeros(N, 2);
  for m = 1:N
    [Y, d, hl_all] = donnees_censure(10^k);
    hl = hl_all(end);
    hlnorm(m, 1) = sqrt(2/3*10^k) * (hl - 1);
    hlnorm(m, 2) = sqrt(sum(d)) * (hl - 1) / hl;
  end;
  % Tracé des distributions empiriques et de la distribution limite
  % sur l'intervalle [-3, 3]
  hlnorm = sort(hlnorm, 1);
  figure(2);
  subplot(3,1,k);
  hold on;
  stairs(hlnorm(:,1), [1:N]/N, 'k--');
  stairs(hlnorm(:,2), [1:N]/N, 'k-.');
  axis([-2.5 2.5 0 1]);
  fplot('pnorm', [-2.5 2.5], 'k');
  title(['n = ', num2str(10^k)]);
```

```
end;
legend('Empirique 1','Empirique 2','Limite');
```

30.1.4 Comparaison entre estimateurs de Kaplan–Meier et par plug-in

```
% Comparaison entre estimateurs de Kaplan-Meier et par plug-in
% 1.  Consistance de ces estimateurs
% 2.  Régions de confiance
%     2.1  Consistance de l'estimateur de Greenwood
%     2.2  Comparaison des régions de confiance
clc, clear, clf;

% 1. Illustration de la consistance
n = 200;
[Y, d, hl] = donnees_censure(n);
hl = hl(end);
[KM1, KM2] = Kaplan_Meier(Y, d);
figure(4);
stairs([0,KM1], [KM2], 'k--');
hold on;
fplot(@(t)exp(-hl*t), [0 KM1(end)], 'k-.')
fplot('exp(-t)', [0 KM1(end)], 'k');
legend('Kaplan-Meier', 'plug-in', 'limite');
title(['Consistance pour n = ', num2str(n)]);


% 2.1 Calcul de l'estimateur de Greenwood pour la variance asymptotique
%     (et représentation de sa convergence ps)
[Yord, I] = sort(Y);
dord = d(I);
% G contiendra les n+1 valeurs associées aux n+1 intervalles
% définis par les n points de KM1
G = zeros(1,n+1);
% Estimation des intégrales I(t)
for i = 1:n
  G(i+1) = G(i) + dord(i)/((n-i+1)*(n-i));
end;
G = n * KM2.^2 .* G;
figure(5);
stairs([0,KM1], G, 'k-.');
hold on;
fplot(@(t)(2/3*(exp(3/2*t)-1)*exp(-2*t)), [0 KM1(end)], 'k');
legend('Greenwood', 'Var W(t)')
title(['n = ', num2str(n)]);

% 2.2 Tracés
%     - de la vraie fonction de survie
```

```
%     - de la bande de confiance (niveau = 95%) associée à l'estimateur
%       plug-in
%     - de la collection des intervalles de confiance associés
%       à l'estimateur de Kaplan-Meier
figure(6);
hold on;
IKM = qnorm(.975)*sqrt(G/n);
stairs([0,KM1], KM2 - IKM ,'k--');
IPI = qnorm(.975)/sqrt(sum(d));
fplot(@(t)(exp(-hl*t) - IPI), [0 KM1(end)], 'k-.')
fplot('exp(-x)', [0 KM1(end)], 'k');
title(['Niveau 0.95 et n = ',num2str(n)]);
legend('Kaplan-Meier', 'plug-in', 'limite');
stairs([0,KM1], KM2 + IKM ,'k--');
fplot(@(t)(exp(-hl*t) + IPI), [0 KM1(end)], 'k-.')
```

30.1.5 Illustration sur un jeu de vraies données

```
clc, clear, clf;

% On exploite les données fournies par l'ouvrage  "Survival Analysis" de
% Miller qui donne les données de durée de vie partiellement censurées pour
% 23 patients atteints de leucémie aiguë myéloïde
D = [9 13 13 18 23 28 31 34 45 48 161 5 5 8 8 12 16 23 27 30 33 43 45;
     1  1  0  1  1  0  1  1  0  1   0 1 1 1 1  1  0  1  1  1  1  1  1];
% Tri des données
Y = D(1,:);
[Y,S] = sort(Y);
d = D(2,S);
% Tracé des données
hold on;
I = find(d == 0); J = find(d == 1);
stem(I, Y(I), 'filled');
stem(J, Y(J));
legend('Non censurées', 'Censurées');
xlabel('Patients'); ylabel('Durées de survie');

% 2. Calcul de l'estimateur de Greenwood pour la variance asymptotique
n = length(Y);
hl = sum(d)/sum(Y);
[KM1, KM2] = Kaplan_Meier(Y, d);
%    Cf. le script principal :
[Yord, I] = sort(Y);
dord = d(I);
G = zeros(1,n+1);
for i = 1:n
  G(i+1) = G(i) + dord(i)/((n-i+1)*(n-i));
```

```
end;
G = n * KM2.^2 .* G;
%
figure;
stairs([0,KM1], G, 'k-.');
legend('Greenwood')
title(['n = ', num2str(n)]);

% 2.2 Tracés
%     - de la bande de confiance (niveau = 95%) associée à  l'estimateur
%       plug-in
%     - de la collection des intervalles de confiance associés
%       à  l'estimateur de Kaplan-Meier
figure;  hold on;
title(['Niveau 0.95 et n = ',num2str(n)]);
stairs([0,KM1], [KM2], 'k');
IKM = qnorm(.975)*sqrt(G/n);
stairs([0,KM1], KM2 - IKM ,'k--');
stairs([0,KM1], KM2 + IKM ,'k--');
fplot(@(t)exp(-hl*t), [0 KM1(end)], 'k-.')
IPI = qnorm(.975)/sqrt(sum(d));
fplot(@(t)(exp(-hl*t) - IPI), [0 KM1(end)], 'k:')
fplot(@(t)(exp(-hl*t) + IPI), [0 KM1(end)], 'k:')
hold off;
```

CHAPITRE 31

Étude du nombre de renouvellements (compléments)

Résumé. Nous présentons ici le code des simulations informatiques relatives au chapitre 23.

31.1 Codes des simulations

31.1.1 Fonctions auxiliaires

```
function W = weibull(a, b, m, n)
%
% Entrées :
%   a et b sont les paramètres de la loi de Weibull
%   m et n désignent la taille du tableau de variables à obtenir
% Sortie :
%   W, une matrice m x n dont les éléments sont iid selon la
%   loi de Weibull de paramètres a et b
%

  % On simule un tableau de taille (m,n) de variables uniformes sur [0,1]
  U = rand(m, n);
  % Et on en déduit W en appliquant l'inverse de la fonction de répartition
  W = b^(-1) * (-log(U)).^(1/a);

end
```

```
function [N1, N2, N3] = Nmax(t, m, v, alpha)
%
% Entrées :
%   t est un vecteur de points où l'on veut calculer les Nmax
%   m et v sont respectivement la moyenne et la variance
%   1-alpha est le niveau de confiance
% Sortie :
%   N1, N2 et N3 sont trois vecteurs de même taille que t
%

  N1 = t/m + qnorm(1-alpha)*sqrt(v)*sqrt(t)*m^(-3/2);
```

```
  M = t/m + v/(m^2);
  N21 = M + alpha^(-1/2) * sqrt(M.^2 + M);
  N22 = M + alpha^(-1/2) * sqrt(t/m*(4*v/(m^2)+3)+v/(m^2)*(2*v/(m^2)+1)-1);
  N2 = N21.*(t<m) + N22.*(t>=m);
  N3 = t/m + alpha^(-1/2)*sqrt(t/m);
end
```

31.1.2 Illustration des convergences presque sûres

```
% On illustre la convergence presque sûre de N(t) vers l'infini
clear, clc, clf;
a = 2; b = 1;
m = gamma(1+1/a)/b;
% On construit la suite de nos temps de saut
N = 2500;
W = weibull(a, b, 1, N);
S = [0 cumsum(W)];
% On représente les 20 premiers sauts du processus N
subplot(1, 2, 1);
stairs(S(1:21), 0:20)
title('Processus N');
% On représente t --> N(t)/t aux instants de saut
% et on trace la droite d'ordonnée 1/mu
subplot(1, 2, 2);
% On omet les premiers pas de temps, pour que ce soit plus clair
n = 50;
stairs(S((n+1):end), (n:N)./S((n+1):end));
hold on;
plot([0 S(end)], [1/m 1/m], '-.');
legend('N(t)/t', 'y = 1/\mu');
title('Convergence p.s. de N(t)/t');
```

31.1.3 Illustration de la convergence en loi

```
% On illustre la convergence en loi énoncée au début du texte
clear, clc, clf;
a = 0.5; b = 2;
m = gamma(1+1/a)/b;
v = (gamma(1+2/a) - gamma(1+1/a)^2)/(b^2);
% On calcule 400 échantillons de vecteurs de 5000 sauts
T = 400;
Nmax = 5000;
W = weibull(a, b, T, Nmax);
S = cumsum(W, 2);
% On choisit des t suffisamment petits pour que N(t) soit plus
% petit que l'instant du dernier (N-ième, où N = 5000) saut
t = [25, 500];
```

```
for k = 1:2
  % Calcul du 400-échantillon de N(t)
  N = sum(S <= t(k), 2);
  % Tracé de la fonction de répartition de la variable d'intérêt,
  % approchée par méthode de Monte-Carlo
  A = sort(sqrt(t(k))*(N/t(k)-1/m));
  subplot(2, 1, k);
  stairs(A, (1:T)/T);
  % Tracé de la loi limite
  hold on;
  plot(A, pnorm(sqrt(m^3/v)*A), '-.');
  legend('empirique', 'limite');
  title(['Fonctions de répartition empirique et limite avec t = ', ...
         num2str(t(k))]);
end;
```

31.1.4 Tracé des intervalles de prévision

```
clc, clf, clear;
% Calcul et tracé des différentes valeurs de N_max pour différentes lois
% de Weibull -- sur l'intervalle [0,500] et pour alpha = 0.05
t = 0:0.1:500;
alpha = 0.05;
% On considère trois valeurs possibles de a
a = [1/2, 1, 2];
b = 1;
for k = 1:3
  % Calcul de la moyenne et de la variance
  m = gamma(1+1/a(k))/b;
  v = (gamma(1+2/a(k))-gamma(1+1/a(k))^2)/(b^2);
  % Calcul des trois valeurs de N_max
  [N1, N2, N3] = Nmax(t, m, v, alpha);
  % Tracé des fonctions
  subplot(3, 1, k);
  hold on;
  plot(t, N1, '--');
  plot(t, N2, '-.');
  if (a(k) >= 1)
    plot(t, N3, '-');
  end
  title(['a = ', num2str(a(k)), ' et b = ', num2str(b)]);
  legend('N1', 'N2', 'N3');
end
```

31.1.5 Calcul ou tracé des niveaux des intervalles de prévision ; illustration des inégalités NBU entre variance, espérance et borne déterministe

```
% 1. Calcul d'une estimation par méthode de Monte-Carlo de la probabilité
%    empirique d'appartenir (niveau) aux intervalles de prévision
%    construits avec alpha = 0.05 et pour t dans [0, 10]
%  2. Classement entre les trois quantités du théorème du paragraphe NBU
%     (variance et espérance de N(t), et borne t/mu)
clc, clf, clear;
t = 0:0.1:10;
alpha = 0.05;
% On considère trois valeurs possibles de a
a = [1/2, 1, 2];
b = 1;
for j = 1:3
  % Calcul de la moyenne et de la variance
  m = gamma(1+1/a(j))/b;
  v = (gamma(1+2/a(j))-gamma(1+1/a(j))^2)/(b^2);
  % On calcule 1000 échantillons de suites de 5000 sauts
  W = weibull(a(j), b, 1000, 5000);
  S = cumsum(W, 2);
  % On construit un 1000-échantillon de N évalué régulièrement
  % sur l'intervalle [0, 10] et on regarde combien de valeurs sont
  % dans les intervalles de prévision en moyenne
  R1 = []; R2 = []; R3 = []; N = [];
  for k = t
    % 1000-échantillon du processus N au pas de temps k
    Nk = sum(S <= k, 2);
    N = [N Nk];
    % Calcul des intervalles de prévision en k
    [N1, N2, N3] = Nmax(k, m, v, alpha);
    % Probabilités empiriques en k
    R1 = [R1 mean(Nk > N1)];
    R2 = [R2 mean(Nk > N2)];
    R3 = [R3 mean(Nk > N3)];
  end;
  % 1. Niveaux (ou plutôt, 1 - niveaux)
  %    Tracé de R1, sorties numériques pour R2 et R3
  figure(1);
  subplot(3, 1, j);
  disp(['Erreur maximale pour N2 et N3 (a = ', ...
        num2str(a(j)), ' et b = ', num2str(b), ') : ', ...
        num2str(max(R2)), ' et ', num2str(max(R3))]);
  plot(t, R1, '-');
  hold on;
  title(['a = ', num2str(a(j)), ' et  b = ', num2str(b)]);
  plot([min(t) max(t)], [alpha alpha], '-.');
```

```
  % 2. Tracé des trois quantités à classer
  figure(2);
  subplot(3, 1, j);
  plot(t, mean(N), 'b--', t, var(N), 'b-.', t, t/m, 'b-')
  title(['a = ', num2str(a(j)), ' et  b = ', num2str(b)]);
  legend('espérance', 'variance', 't/\mu');
end
```

31.1.6 Illustration de nos résultats sur un jeu de vraies données

```
clc, clear, clf;
x = SYS1;
alpha = 0.05; % correspond à 0.10/2
n = length(x);

% Utilisation d'un jeu de données antérieures pour
% effectuer une estimation
D = x(1:(n/2)); m = length(D);
figure(1); hold on;
fplot(@(t)(1-exp(-t/mean(D))),[0 2500],'b');
stairs(sort(D),(1:m)/m)
% Estimations : ponctuelle, par intervalle
l = [1/mean(D) exp(-(log(mean(D))-qnorm(1-alpha)))];

% Jeu de données d'évaluation
T = x((1+n/2):end);
t = [0:1:sum(T)];
figure(2);
for k = 1:2
  subplot(1,2,k);
  hold on;
  [N1,N2,N3] = Nmax(t,1/l(k),(1/l(k))^2,alpha);
  stairs(cumsum([0 T]), [0:(length(T))]);
  plot(t,N1,'--b');
  plot(t,N3,'-.b');
  legend('N','N1','N3');
end;
```

```
function x = SYS1;
%
% Voici la fonction chargeant les vraies données.
%
% Sortie :
%   x, le vecteur des vraies données (rangées dans un ordre aléatoire)
```

```
%

x = [3, 30, 113, 81, 115, 9, 2, 20, 20, 15, 138, 50, 77, 24, 108, 88, ...
    670, 120, 26, 114, 325, 55, 242, 68, 422, 180, 10, 1146, 600, 15, ...
    36, 4, 0, 8, 227, 65, 176, 58, 457,300, 97, 263, 452, 255, 197, ...
    193, 6, 79, 816, 1351, 148, 21, 233, 134, 357, 193, 236, 31, 369, ...
    748, 0, 232, 330, 365, 1222, 543, 10, 16, 529, 379, 44, 129, 810, ...
    290, 300, 529, 281, 160, 828, 1011, 445, 296, 1755, 1064, 1783, 860, ...
    983, 707, 33, 868, 724, 2323, 2930, 1461, 843, 12, 261, 1800, 865, ...
    1435, 30, 143, 108, 0, 3110, 1247, 943, 700, 875, 245, 729, 1897, ...
    447, 386, 446, 122, 990, 948, 1082, 22, 75, 482, 5509, 100, 10, ...
    1071, 371, 790, 6150, 3321, 1045, 648, 5485, 1160, 1864, 4116];

[X,I] = sort(rand(1,length(x)));
x = x(I);

end
```

CHAPITRE 32

Estimation de densité de probabilité (compléments)

RÉSUMÉ. Nous fournissons ici quelques extensions et commentaires quant aux résultats du chapitre 24 : l'énoncé et la preuve du lemme de Scheffé et du théorème de Glick ; d'autres applications de l'estimation de densité ; ainsi qu'une discussion des vitesses de convergence possibles dans les résultats d'estimation consistante des densités. Nous présentons enfin les codes des simulations informatiques.

32.1 Compléments mathématiques

32.1.1 Lemme de Scheffé et théorème de Glick

Le lemme de Scheffé ne doit pas être confondu avec le théorème de Scheffé prouvé au début du chapitre 24 (le Théorème 24.1). C'est un résultat analogue au théorème de convergence dominée de Lebesgue, puisqu'il renforce une convergence presque partout ou presque sûre en une convergence $\mathbb{L}^1$.

Lemme 32.1 (de Scheffé) Soit μ une mesure positive sur un espace mesurable $(\Omega, \mathcal{A})$. On considère une fonction $\ell \in \mathbb{L}^1(\mu)$ et une suite (ℓ_n) dans $\mathbb{L}^1(\mu)$ vérifiant d'une part la convergence μ–presque partout $\ell_n \to \ell$ et d'autre part,

$$\int_\Omega |\ell_n| \, \mathrm{d}\mu \longrightarrow \int_\Omega |\ell| \, \mathrm{d}\mu \, .$$

Alors $\ell_n \to \ell$ dans $\mathbb{L}^1(\mu)$.

Les situations les plus souvent considérées sont celle où $\mu = \lambda$, la mesure de Lebesgue, et celles où μ est une probabilité (bien entendu, on parle alors de convergence presque sûre). Dans ce dernier cas, on montre comme dans la preuve du Théorème 26.9 que le résultat du lemme de Scheffé vaut encore lorsque les ℓ_n convergent vers ℓ en $\mathbb{P}$–probabilité plutôt que $\mathbb{P}$–presque sûrement.

Preuve (générale) Il suffit d'appliquer le lemme de Fatou à la suite (g_n) définie par $g_n = |\ell_n| + |\ell| - |\ell_n - \ell| \geqslant 0$: d'une part, μ–presque partout, $g_n \to 2\,|\ell|$, et ainsi

$$2 \int_\Omega |\ell| \, \mathrm{d}\mu = \int_\Omega \big(\liminf g_n\big) \, \mathrm{d}\mu \leqslant \liminf \int_\Omega g_n \, \mathrm{d}\mu = 2 \int_\Omega |\ell| \, \mathrm{d}\mu - \limsup \int_\Omega |\ell_n - \ell| \, \mathrm{d}\mu \, ,$$

soit

$$\|\ell_n - \ell\|_1 = \int_\Omega |\ell_n - \ell| \, \mathrm{d}\mu \longrightarrow 0$$

après simplification des intégrales en $|\ell|$ (qui ont bien une valeur finie). □

Preuve (dans le cas des densités de probabilité) Lorsque $\mu = \lambda$ et que les ℓ_n et ℓ sont des densités de probabilité, la preuve est encore plus directe. On applique le théorème de convergence dominée aux $g_n = (\ell - \ell_n)_+$, dominées par ℓ et convergeant presque partout vers 0 :

$$\|\ell_n - \ell\|_1 = 2 \int_{\mathbb{R}} (\ell - \ell_n)_+ \, \mathrm{d}\lambda \longrightarrow 0 \,,$$

où l'on a utilisé également une des égalités données par le théorème de Scheffé (Théorème 24.1). □

On généralise cette preuve de la manière suivante. Dans le théorème suivant, il y a une double intégration, selon la mesure de Lebesgue λ et la probabilité $\mathbb{P}$.

Théorème 32.2 (de Glick) *Si l'on dispose d'une suite d'estimateurs $\widehat{f_n}$ définissant chacun une densité de probabilité et convergeant en presque tout point en $\mathbb{P}$–probabilité vers une densité de probabilité f, alors $\widehat{f_n} \to f$ dans $\mathbb{L}^1(\mathbb{P} \otimes \lambda)$. De manière équivalente,*

$$\mathrm{d}_{\mathrm{VT}}\Big(\widehat{f_n},\ f\Big) \xrightarrow{\mathbb{P}} 0 \,.$$

Preuve On emploie la même technique de preuve que pour le lemme de Scheffé dans le cas des densités de probabilité : deux applications du théorème de convergence dominée. D'une part, pour x tel que $\widehat{f_n}(x) \to f(x)$ en $\mathbb{P}$–probabilité, on a, par l'extension du théorème de convergence dominée aux convergences en probabilité (voir chapitre 26) et par domination par la constante $f(x)$,

$$\varphi_n(x) \overset{\text{not.}}{=} \mathbb{E}\left[\Big(f(x) - \widehat{f_n}(x)\Big)_+\right] \longrightarrow 0 \,.$$

Ainsi, φ_n tend presque partout vers 0 tout en étant dominée par la fonction intégrable f. Par le théorème de Scheffé (on utilise là que f et les $\widehat{f_n}$ sont des densités de probabilité), le théorème de Fubini–Tonelli puis une seconde application du théorème de convergence dominée, il vient finalement

$$\frac{1}{2}\,\mathbb{E}\left[\int_{\mathbb{R}} \left|f - \widehat{f_n}\right| \mathrm{d}\lambda\right] = \mathbb{E}\left[\int_{\mathbb{R}} \Big(f(x) - \widehat{f_n}(x)\Big)_+ \mathrm{d}x\right] = \int_{\mathbb{R}} \varphi_n(x)\,\mathrm{d}x \longrightarrow 0 \,.$$

On vient de prouver la convergence $\widehat{f_n} \to f$ dans $\mathbb{L}^1(\mathbb{P} \otimes \lambda)$. Elle entraîne évidemment la convergence en $\mathbb{P}$–probabilité des distances en variation totale, simplement parce que de manière générale, la convergence $\mathbb{L}^1(\mathbb{P})$ entraîne la convergence en $\mathbb{P}$–probabilité. Il y a ici équivalence entre les deux convergences car on manipule des variables aléatoires bornées (par 1, s'agissant de distances en variation totale ; on pourra effectuer un raisonnement direct ou recourir au Théorème 26.9). □

32.1.2 Un autre schéma de preuve pour le Théorème 24.7

Sans surprise, il est dans l'esprit des démonstrations concernant les estimateurs par fenêtres glissantes (qui sont, on le rappelle, un cas particulier d'estimateurs par noyaux) et consiste à établir des résultats ponctuels.

On commence par remarquer qu'en tout point de Lebesgue x de f, donc presque partout, $f \star K_h(x) \to f(x)$ lorsque $h \to 0$. Une démonstration possible de ce fait, suggérée par [DL01, exercice 5.9], repose sur une approximation de K en norme $\mathbb{L}^1$ par une combinaison linéaire d'indicatrices de segments. On montre alors qu'en tous ces points, $\mu_n \star K_{h_n}(x) \to f(x)$ en $\mathbb{P}$–probabilité, lorsque $n \to \infty$ avec $h_n \to 0$: on écrit ici la décomposition biais–variance ponctuelle

$$\mathbb{E}\Big[\big(\mu_n \star K_{h_n}(x) - f(x)\big)^2\Big] = \big(f \star K_{h_n}(x) - f(x)\big)^2 + \mathrm{Var}\big(\mu_n \star K_{h_n}(x)\big)$$

et on a indiqué ci-dessus que le premier terme (de biais) tendait vers 0 presque partout. On a montré à la fin du paragraphe 24.1.6 que le second terme (de variance) était égal à $\big(\mathrm{Var}\, Y_1^{n,x}\big)/n$ et qu'on pouvait le majorer, avec les notations utilisées là-bas, par $\big(K'_{h_n} \star f(x)\big)\,\sigma^2/nh_n$. Ce majorant tend vers 0 dès que $nh_n \to \infty$ car en un point de Lebesgue de f, on a également $f \star K'_h(x) \to f(x)$ lorsque $h \to 0$. Le théorème de Glick (Théorème 32.2) permet alors de passer de la convergence $\mu_n \star K_{h_n}(x) \to f(x)$ en $\mathbb{P}$–probabilité en tous les points de Lebesgue à la convergence

$$\mathbb{E}\Big[\,\big\|\,\mu_n \star K_{h_n} - f\,\big\|_1\Big] \to 0\,.$$

Ce schéma de preuve, proposé par [DL01, Théorème 9.2], a l'avantage de ressembler fort au schéma de l'estimateur par fenêtres glissantes mais a l'inconvénient majeur de reposer sur deux lemmes assez techniques, le théorème de densité de Lebesgue (sa version forte avec les ensembles rétrécissant convenablement, voir [Rud98, paragraphes 7.9 et 7.10]) et sa conséquence, que $f \star K_{h_n}(x) \to f(x)$ en tout point de Lebesgue x de f. Nous n'avons pas montré cette dernière mais comme on s'en doute, la preuve n'est pas difficile, c'est essentiellement une question d'écriture et d'emploi des bons théorèmes (inégalité de convolution) et définitions (pour choisir l'approximation de K, il faut se souvenir des caractérisations et constructions de l'intégrale de Lebesgue). Dans la preuve détaillée au paragraphe 24.1.6, nous n'avions pas eu à utiliser le théorème de densité de Lebesgue et des techniques tout à fait élémentaires ont suffi à prouver le théorème de consistance pour les estimateurs par noyaux.

32.1.3 A propos des autres applications de l'estimation de densité

Simulation de variables

Imaginons que l'on dispose d'un n–échantillon $X_1, \ldots, X_n$ (de taille modérée) distribué selon une loi μ de densité f et que l'on veuille simuler de nouvelles variables aléatoires $X_{n+1}, \ldots, X_m$ indépendantes et identiquement distribuées selon une loi aussi proche que possible de celle de μ. On requiert cela à des fins d'estimation de l'intégrale sous μ d'une fonction compliquée, pour laquelle il est nécessaire d'avoir une grande taille d'échantillon m, voir la fin du chapitre 9.

On estime alors f par $\widehat{f_n}$ sur le n–échantillon initial, puis, en utilisant une méthode de simulation de variables aléatoires mettant en jeu la densité (la plus connue est la méthode de rejet, voir le paragraphe 9.3), on obtient de nouvelles variables aléatoires, qui, conditionnellement à $X_1, \ldots, X_n$, sont indépendantes et identiquement distribuées selon la loi de densité $\widehat{f_n}$. Par cette méthode, on a un contrôle asymptotique de l'écart de la distribution des nouvelles observations à celle de l'échantillon initial et par conséquent, pour peu que par exemple la fonction à intégrer soit bornée, de l'écart des intégrales sous μ et sous ses estimées. Cet écart est en effet quantifié par des distances en variation totale.

Notons que la méthode qui aurait consisté à estimer la fonction de répartition (par la fonction de répartition empirique) pour simuler les nouvelles variables par méthode d'inversion (voir chapitre 9) aurait été décevante, puisqu'elle consiste à simplement tirer au hasard et avec remise les $X_{n+1}, \ldots, X_m$ parmi les $X_1, \ldots, X_n$. On retrouve le fait que la mesure empirique est concentrée en un nombre fini de points. En particulier, dans cette méthode naïve, les queues de distribution posent problème. La loi associée à la fonction de répartition empirique canonique a en effet un support borné, alors que ce n'est pas nécessairement le cas de la loi $\widehat{\mu}_n$ donnée par un bon estimateur de la densité $\widehat{f_n}$, surtout si l'on a une idée *a priori* de la forme ou du poids de la queue et qu'on l'exploite. C'est le cas notamment si l'on est dans un cadre d'estimation paramétrique et que l'on connaît la forme de f.

Illustrations informatiques de convergence en loi

Dans les illustrations informatiques, on représente souvent la convergence en loi par une convergence de densités. Typiquement, si on établit, par exemple, qu'une suite d'estimateurs converge en loi, $X^{(n)} \rightsquigarrow \mathcal{N}(0, 1)$ pour fixer les idées, alors, pour illustrer cette convergence, on va fixer une grande valeur de n et simuler un N_n–échantillon distribué selon la loi de $X^{(n)}$, noté

$$X_1^{(n)}, \ldots, X_{N_n}^{(n)};$$

après quoi, on établira l'histogramme des valeurs du N_n–échantillon et on essaiera de se convaincre qu'il a une forme proche d'une courbe en cloche, qui est précisément la forme de la densité de la loi normale standard. C'est ce que nous avons réalisé par exemple dans les simulation des chapitres 19 et 23.

Mais ces histogrammes, nous l'avons vu, sont en fait des estimateurs de la densité de la loi de $X^{(n)}$ lorsque celle-ci existe et que la taille des classes est bien choisie en fonction de N_n. On essaie donc de représenter une convergence en loi par une proximité de densités. Or, *a priori*, la convergence en loi est équivalente à la convergence des fonctions de répartitions, c'est-à-dire, à celle des intégrales des densités (lorsque ces dernières existent). Cela ne signifie généralement pas que les densités elles-mêmes convergent ponctuellement. C'est, en conclusion, une illustration dangereuse…

… Sauf qu'elle est beaucoup plus agréable et plus parlante visuellement qu'une convergence de fonctions de répartitions : plus parlante parce que les différences d'aires entre la densité (estimée sur le N_n–échantillon) au rang n et la densité de la loi limite forment exactement

la distance en variation totale. On peut donc lire cette dernière graphiquement et quantifier facilement les écarts entre les deux distributions.

Il arrive en fait que la convergence des lois s'accompagne de celles des densités de probabilité associées. Il est clair que si ces densités convergent dans $\mathbb{L}^1(\lambda)$ (ce qui découle d'une convergence ponctuelle, *via* le lemme de Scheffé), alors les fonctions de répartition convergent ponctuellement et il y a convergence des lois. Réciproquement, une version plus forte du théorème de la limite centrale aboutissant à une convergence $\mathbb{L}^1$ des densités (et donc une convergence en variation totale des lois) s'appelle le théorème de la limite locale ; il est présenté par [vdV98, paragraphe 2.9].

32.1.4 Quelques commentaires sur l'optimalité des procédures d'estimation

On donne ici quelques indications sur la vitesse de convergence des estimateurs par noyaux. Les preuves des résultats avancés ne sont pas si difficiles mais elles sont trop longues pour pouvoir être reproduites ici. L'objectif de ce qui suit est de mettre en lumière la difficulté intrinsèque du problème d'estimation de densité.

En termes de performances individuelles, *id est*, pour une densité f donnée, on a le résultat suivant.

Théorème 32.3 *Pour toute densité f et pour tout noyau K, il existe une constante $A(f, K) \geqslant 0.86$ telle que les estimateurs par noyaux construits sur K vérifient*

$$\liminf_{n\to\infty} \inf_{h>0} \; n^{2/5}\, \mathbb{E}\left[\left\| \widehat{f}_{n,h} - f \right\|_1 \right] \;\geqslant\; A(f, K)\,.$$

En particulier, la distance en variation totale entre tout estimateur par noyau $\widehat{f}_{n,h_n}$ et la vraie densité f est (asymptotiquement) plus grande que $0.43\, n^{-2/5}$, où n désigne toujours la taille de l'échantillon. Si par exemple, on veut que les probabilités de tous les événements soient estimées avec un écart (absolu) d'au plus 1 % par rapport à leur vraie fréquence, il faudra que n soit (c'est ce que suggère l'asymptotique) tel que $0.43 n^{-2/5} \leqslant 0.01$, i.e., $n \geqslant 12\,000$. C'est un nombre d'observations très grand, que l'on rencontre peu souvent en pratique. Et ce résultat vaut pour toutes les densités de probabilité, même les plus régulières !

> Remarque au passage. Lorsque l'on a affaire à des probabilités discrètes, on peut estimer simultanément, avec les fréquences empiriques, la probabilité de tous les événements à un écart $n^{-1/2}$ près. Le théorème de la limite centrale en donne l'intuition et on peut le prouver en utilisant des inégalités de concentration, par exemple celle de McDiarmid (voir par exemple [DL01, Théorème 2.2]). Ici, on a affaire à des lois continues et estimer leur densité se fait au mieux à la vitesse non pas $n^{-1/2}$ mais $n^{-2/5}$: on perd donc au moins un ordre de grandeur $n^{1/2-2/5} = n^{1/10}$.

Est-ce que cette vitesse $n^{-2/5}$ est atteignable ? En général, non, et il faudra restreindre la classe des densités considérées au vu du résultat suivant. On peut en effet montrer que pour

toute suite (a_n) tendant vers 0, pour toute stratégie d'estimation (f_n) de la densité, il existe une densité f (que l'on peut prendre bornée par 3, à support dans $[0, 1]$ et infiniment dérivable sur $[0, 1[$) telle que

$$\limsup_{n\to\infty} \frac{1}{a_n} \mathbb{E}\Big[\|f_n - f\|_1\Big] = \infty .$$

On a vu plus haut qu'on pouvait toujours faire en sorte que la distance en variation totale tende vers 0, quelle que soit la densité de probabilité sous-jacente f, en retenant des estimateurs de la forme

$$f_n = \widehat{f}_{n,h_n} .$$

Nous apprenons ici que cette convergence peut être arbitrairement lente. Plus lente, par exemple, que $a_n = 1/(\log\log\log n)\ldots$ pour laquelle il faudrait recueillir un nombre phénoménal d'observations pour obtenir un résultat non trivial.

Par conséquent, pour garantir des vitesses de convergence uniformes, on doit se restreindre à des sous-classes de densités régulières. Nous présentons quelques exemples de résultats que l'on peut exhiber en ce sens. Considérons par exemple la classe $\mathcal{F}$ des densités f à support compact, absolument continues de dérivée f' également absolument continue, de dérivée f'' vérifiant $c(f) = \int |f''| \, \mathrm{d}\lambda < \infty$. Pour $f \in \mathcal{F}$, on note $s(f)$ la mesure de Lebesgue du support de f. Pour tout $C > 0$, on introduit alors la sous-classe $\mathcal{F}_C$ des densités $f \in \mathcal{F}$ telles que $s^2(f)\, c(f) \leqslant C$. Pour cette sous-classe, on a le résultat de convergence suivant.

Théorème 32.4 *Pour tout noyau K borné, symétrique et à support dans $[-1, 1]$, il existe une constante $B(K)$ telle que pour la suite des estimateurs par noyaux construits sur K :*

$$\sup_{f\in\mathcal{F}_C} \inf_{h>0} \mathbb{E}\left[\left\|\widehat{f}_{n,h} - f\right\|_1\right] \leqslant \frac{B(K)\, C^{1/5}}{n^{2/5}} .$$

Le théorème précédent n'indique pas directement comment choisir h. Un coup d'œil à la preuve montre en revanche qu'on a les majorations suivantes pour les termes de biais et variance respectivement,

$$h^2 \frac{c(f)}{2} \int_{\mathbb{R}} x^2 K(x)\, \mathrm{d}x \qquad \text{et} \qquad \sqrt{\frac{s(f) + 2h}{nh}} \sqrt{\int_{\mathbb{R}} K^2 \, \mathrm{d}\lambda} .$$

Ainsi, le choix $h_n = n^{-1/5}$ conduit bien à la vitesse uniforme $n^{-2/5}$. Cependant, si l'on voulait optimiser le majorant, il s'agirait de prendre h_n égal à $n^{-1/5}$ à une constante multiplicative près, cette dernière dépendant de $c(f)$ et $s(f)$, que l'on ne connaît pas. C'est ce qui explique que dans le théorème on ne puisse échanger le supremum et l'infimum.

On atteint également des vitesses uniformes sur des classes comme celle des densités lipschitiziennes de constante de Lipschitz et de support compact fixés (vitesse $n^{-1/3}$), celle des densités monotones et bornées par c fixé sur un intervalle $[a, b]$ fixé (vitesse $n^{-1/3}$), etc.

Le Théorème 32.4 est, en un sens, inefficace, parce qu'il ne donne pas de formule explicite pour la fenêtre h_n en fonction de $\mathcal{F}_C$. C'est pourquoi, au vu des bonnes performances des estimateurs par noyaux et pour se débarrasser de ce problème délicat qu'est le choix de h, [DL01] résout le problème suivant : trouver une stratégie d'estimation (f_n) telle que pour toute densité f,

$$\mathbb{E}\Big[\,\|\,f_n - f\,\|_1\Big] \leqslant \big(3 + o(1)\big) \inf_{h>0} \mathbb{E}\Big[\,\|\,\widehat{f}_{n,h} - f\,\|_1\Big] + o\big(n^{-2/5}\big)\,.$$

C'est ce que l'on appelle une inégalité d'oracle. En particulier, sur la classe $\mathcal{F}_C$, elle conduit à

$$\mathbb{E}\Big[\,\|\,f_n - f\,\|_1\Big] \leqslant \big(3 + o(1)\big)\frac{B(K)\,C^{1/5}}{n^{2/5}} + o\big(n^{-2/5}\big) = \big(3 + o(1)\big)\frac{B(K)\,C^{1/5}}{n^{2/5}}\,.$$

La suite d'estimateurs (f_n) n'est pas nécessairement à noyaux, mais elle est définie à partir des briques fondamentales que sont les estimateurs à noyaux $\widehat{f}_{n,h}$.

32.2 Codes des simulations

32.2.1 Calcul des estimations

```
function [f, f_inf, f_sup] = FenGliss(x, D, h)
%
% Calcul de l'estimateur par fenêtres glissantes,
% et des intervalles de confiance ponctuels associés
%
% Entrées :
%   x est le vecteur des observations
%   D (pour design) est l'ensemble des points où l'on veut
%     calculer l'estimateur de la densité
%   h est un réel positif donnant la taille de fenêtre
% Sortie :
%   f, f_inf, f_sup sont des vecteurs de même taille que D
%   f est le vecteur des estimées de la densité
%   f_inf, f_sup sont les bornes des intervalles de confiance
%                ponctuels à 95 %
%

  N = length(D);
  n = length(x);

  f = zeros(1,N);
  for j = 1:N
    f(j) = sum((D(j)-h < x) & (D(j)+h >= x));
  end
  f = f / (2*n*h);
```

```
  q = qnorm(.975);
  f_sup = f + q * sqrt(f/(2*n*h));
  f_inf = f - q * sqrt(f/(2*n*h));

end



function f = EstNoyau(x, h, D, Kfun)
%
% Estimation par noyaux
%
% Entrées :
%   x est le vecteur des observations,
%   h est un réel donnant le paramètre de bande,
%   D (pour design) est l'ensemble des points où l'on veut
%     calculer l'estimateur de la densité,
%   Kfun est le nom d'une fonction MATLAB donnant un noyau
%        (p.ex. 'dnorm' pour le noyau gaussien standard)
% Sortie :
%   Un vecteur f de même taille que D, où chaque élément
%   est l'estimée de la densité au point correspondant de D
%

  n = length(D);
  K = str2func(Kfun);
  f = zeros(1,n);
  for j = 1:n
    f(j) = sum(K((D(j)-x)/h));
  end
  f = f/(length(x)*h);

end



function f = Histogr(x, D, h)
%
% Estimation par histogrammes
%
% Entrées :
%   x est le vecteur des observations
%   D (pour design) est l'ensemble des points où l'on veut
%     calculer l'estimateur de la densité
%   h est un réel donnant le paramètre de bande
```

```
% Sortie :
%   Un vecteur f de même taille que D, où chaque élément
%   est l'estimée de la densité au point correspondant de D
%   sur la partition régulière de pas h
%

  N = length(D);
  n = length(x);

  f = zeros(1,N);
  for j = 1:N
    kh = B(D(j),h);
    f(j) = sum((kh <= x) & (kh+h > x));
  end
  f = f / (n*h);

end

function b = B(y,h)
  % Calcul de l'index de la partition régulière de pas h
  % où se trouve y
  b = floor(y/h)*h;
end
```

32.2.2 Densités et noyaux

```
function d = dEpa(x)
%
% Noyau d'Epanechnikov
%
% Entrée :
%   x un vecteur
% Sortie :
%   d vecteur de même taille que x
%

  d = 3/4 * max(1-x.^2, 0);

end
```

```
function d = dgauss(D, p, m, s)
%
% Mélange de densités gaussiennes
%
```

```
% Entrées :
%   D un vecteur
%   p, m, s trois vecteurs de même taille précisant respectivement
%            les coefficients du mélange, les moyennes des lois
%            normales et leurs écarts-types
% Sortie :
%   d vecteur de même taille que D donnant la densité en chaque point
%

  d = zeros(1, length(D));
  for j = 1:length(p)
    d = d + p(j) * dnorm(D, m(j), s(j));
  end;

end



function d = dunif(D, p, a, b)
%
% Mélange de densités uniformes
%
% Entrées :
%   D un vecteur
%   p, a, b trois vecteurs de même taille précisant respectivement
%            les coefficients du mélange, les bornes inférieures et
%            supérieures des intervalles de lois uniformes
% Sortie :
%   d vecteur de même taille que D donnant la densité en chaque point
%

  d = zeros(1, length(D));
  for j = 1:length(a)
    ind = (D >= a(j)) & (D <= b(j));
    d(ind) = d(ind) + p(j)/(b(j)-a(j));
  end;

end
```

32.2.3 Tirages aléatoires

```
function X = MelGauss(n, p, m, s)
%
% Tirage aléatoire selon un mélange de lois gaussiennes
%
% Entrées :
%   n, un entier
```

```
%   p, m, s trois vecteurs de même taille précisant respectivement
%           les coefficients du mélange, les moyennes des lois
%           normales et leurs écarts-types
% Sortie :
%   X, un vecteur de taille n échantillonné selon la loi de mélange
%

  y = rand(1,n);
  X = zeros(1,n);
  q = [0 cumsum(p)];
  for j = 1:length(p)
    ind = (y <= q(j+1)) & (y > q(j));
    X(ind) = rnorm(sum(ind+0), m(j), s(j));
  end;

end




function X = MelUnif(n, p, a, b)
%
% Tirage aléatoire selon un mélange de lois uniformes
%
% Entrées :
%   n, un entier
%   p, a, b trois vecteurs de même taille précisant respectivement
%           les coefficients du mélange, les bornes inférieures et
%           supérieures des intervalles de lois uniformes
% Sortie :
%   X, un vecteur de taille n échantillonné selon la loi de mélange
%

  y = rand(1,n);
  X = zeros(1,n);
  q = [0 cumsum(p)];
  for j = 1:length(p)
    ind = (y <= q(j+1)) & (y > q(j));
    X(ind) = runif(sum(ind+0), a(j), b(j));
  end;

end

function d = runif(k, a, b)
  d = a + (b-a)*rand(1,k);
end
```

32.2.4 Fonctions auxiliaires

```
function d = dvt(D, delta, f, g)
%
% Approximation de la distance en variation
% entre la vraie densité et une estimation
%
% Entrées :
%   D doit être un vecteur de points régulièrement espacés
%   delta est cet espacement régulier
%   g est la vraie densité
%   f est un vecteur d'estimations (de g, calculées en les éléments de D)
% Sortie :
%   d réel positif
%

  d = delta * trapz(abs(f-g(D))) /2;

end
```

32.2.5 Fonction de comparaison des quatre méthodes

```
function R = compareGr(x, D, delta, true_f, h, trace)
%
% Comparaison des quatres méthodes d'estimation de densité
%
% Entrées :
%   x est le vecteur des observations,
%   D (pour design) est un ensemble de points régulièrement espacés
%     où l'on veut calculer l'estimateur de la densité,
%   delta est l'espacement uniforme entre les points de D,
%   true_f est une fonction donnant la vraie densité,
%   h est le paramètre réel précisant la largeur de bande
%   trace est un booléen indiquant si un tracé graphique est souhaité
% Sortie :
%   R est un vecteur à quatre éléments précisant quatre distances en
%     variation
%

  if trace
    clf;
    dmin = min(D);
    dmax = max(D);
    ymax = max(true_f(D))*1.25;
  end;

  f = Histogr(x, D, h);
```

```
  dH = dvt(D, delta, f, true_f);
  if trace
    subplot(2,2,1)
    box on; hold on;
    plot(D, f, 'k');
    plot(D, true_f(D), 'k-.');
    axis([dmin dmax -0.05 ymax]);
    title(strcat(['Par histogrammes (d_{VT} = '  num2str(dH,2)  ')']));
  end;

  [f, f_inf, f_sup] = FenGliss(x, D, h);
  dF = dvt(D, delta, f, true_f);
  if trace
    subplot(2,2,2)
    box on; hold on;
    plot(D, f, 'k');
    plot(D, true_f(D), 'k-.');
    axis([dmin dmax -0.05 ymax]);
    title(strcat(['Fenêtres glissantes (d_{VT} = ' num2str(dF,2) ')']));
  end;

  f = EstNoyau(x, h, D, 'dnorm');
  dN1 = dvt(D, delta, f, true_f);
  if trace
    subplot(2,2,3)
    box on; hold on;
    plot(D, f, 'k');
    plot(D, true_f(D), 'k-.');
    axis([dmin dmax -0.05 ymax]);
    title(strcat(['Par noyau gaussien (d_{VT} = ' num2str(dN1,2) ')']));
  end;

  f = EstNoyau(x, h, D, 'KEpa');
  dN2 = dvt(D, delta, f, true_f);
  if trace
    subplot(2,2,4)
    box on; hold on;
    plot(D, f, 'k');
    plot(D, true_f(D), 'k-.');
    axis([dmin dmax -0.05 ymax]);
    title(strcat(['Par noyau d''Epanechnikov (d_{VT} = ', ...
                  num2str(dN2,2), ')']));
  end;

  R = [dH dF dN1 dN2];

end
```

32.2.6 Programme maître (script)

```
clc, clf, clear;

% Paramètres du mélange uniforme
p = [0.25 0.45 0.30];
a = [-1 0 2.5];
b = [ 1 1 3.5];
% Paramètres du mélange gaussien
q = [0.45 0.20 0.35];
m = [-1.5 0.5 2.5];
s = [ 0.75 0.60 0.45];

% 1. Intervalles de confiance pour fenêtres glissantes
%    et mélange de lois uniformes

N = 800;
h = [0.05 0.15 0.25 0.35];

figure(1);
clf;
delta = 0.01;
D = (min(a)-0.5):delta:(max(b)+0.5);
for i = 1:4;
  subplot(2,2,i)
  box on; hold on;
  x = MelUnif(N, p, a, b)';
  [f, f_inf, f_sup] = FenGliss(x, D, h(i));
  plot(D, dunif(D, p, a, b), 'k');
  plot(D, f, 'k:', D, f_inf, 'k', D,f_sup, 'k');
  title(strcat(['Paramètre de bande h = ' num2str(h(i))]));
  axis([min(a)-0.5 max(b)+0.5 -0.05 0.9]);
end

% 2. Comparaisons graphiques entre les quatre méthodes

N = 400;

% Mélange de lois uniformes

h = 0.15;
figure(2);
x = MelUnif(N, p, a, b)';
f = @(x)dunif(x, p, a, b);
compareGr(x, D, delta, f, h, true);

% Mélange de lois gaussiennes
```

```
h = 0.30;
figure(3);
ymin = min(m)-3*max(s);
ymax = max(m)+3*max(s);
D = ymin:delta:ymax;
x = MelGauss(N, q, m, s)';
g = @(x)dgauss(x, q, m, s);
compareGr(x, D, delta, g, h, true);

% 3. Comparaison numérique

% Selon que l'on veut comparer les performances pour l'estimation
% du mélange de densités uniformes ou de densités gaussiennes
u = false;

N = 400;      % Taille de chaque échantillon
K = 1000;     % Nombre d'itérations de Monte-Carlo
h = [0.02 0.05 0.15 0.20 0.30 0.40 0.50 0.60];

% Dans les deux cas, on augmente D pour mesurer les écarts
% sur toute la droite réelle
if u
  D = (-3):delta:(5.5);
  ff = f;
else
  D = (-5):delta:5;
  ff = g;
end;

% Boucle d'estimation par méthode de Monte-Carlo
R = zeros(length(h), 4, K);
for j = 1:length(h)
  for k = 1:K
    if u
      x = MelUnif(N, p, a, b)';
    else
      x = MelGauss(N, q, m, s)';
    end;
    R(j,:,k) = compareGr(x, D, delta, ff, h(j), false);
  end;
end;

% Affichage des résultats : estimations des moyennes et écart-types
disp(mean(R, 3));
disp(std(R, [], 3));
```

CHAPITRE 33

Classification de données (compléments)

Résumé. Dans cette annexe au chapitre 20, on s'attache à montrer que le problème de classification considéré, qui est un problème d'apprentissage, est en fait beaucoup plus facile que le problème auxiliaire d'estimation de la fonction de régression et qui donne lieu à certaines règles de classification. On énonce en premier lieu un argument général, puis on se place dans le cadre paramétrique : on y oppose la vitesse d'estimation $1/\sqrt{n}$ à la vitesse d'apprentissage $1/n$. Plus précisément, on prouve alors le Théorème 20.7 et on donne des éléments de preuve pour le Théorème 20.8. On procure également les codes des simulations `Matlab`.

33.1 Compléments mathématiques

33.1.1 Il est plus facile de classer que d'estimer la fonction de régression

On étoffe ici la discussion de la fin du paragraphe 20.1.4. On note $\mathbb{L}^1(\mu)$ et $\mathbb{L}^2(\mu)$ les classes de fonctions sur $\mathbb{R}^d$ respectivement intégrables et de carré intégrable par rapport à la loi μ de X. L'inégalité (20.3) entraîne en particulier que

$$L(g_n) - L(g^*) \leqslant 2\,\mathbb{E}\Big[\big|\widehat{\eta}_n(X) - \eta(X)\big| \;\Big|\; Z_1^n\Big] = 2\int_{\mathbb{R}^d} \big|\widehat{\eta}_n(x) - \eta(x)\big|\,\mathrm{d}\mu(x)$$
$$\overset{\text{not.}}{=} 2\,\big\|\widehat{\eta}_n - \eta\big\|_{\mathbb{L}^1(\mu)} \leqslant 2\,\big\|\widehat{\eta}_n - \eta\big\|_{\mathbb{L}^2(\mu)} \tag{33.1}$$

(par l'inégalité de Cauchy–Schwarz). La majoration en norme $\mathbb{L}^2(\mu)$ est proposée essentiellement parce que s'agissant d'une fonction de régression, caractérisée par une propriété de meilleure approximation au sens $\mathbb{L}^2$, il est naturel de mesurer l'écart de l'estimation par $\widehat{\eta}_n$ (construite sur les données) à η en norme $\mathbb{L}^2(\mu)$.

Il découle de (33.1) une majoration de l'espérance du risque de classification,

$$\mathbb{E}\big[L(g_n)\big] - L(g^*) \leqslant 2\,\mathbb{E}\Big[\,\big\|\widehat{\eta}_n - \eta\big\|_{\mathbb{L}^2(\mu)}\Big]\,;$$

et, par une seconde application de l'inégalité de Jensen,

$$\mathbb{E}\Big[\,\big\|\widehat{\eta}_n - \eta\big\|_{\mathbb{L}^2(\mu)}\Big] \leqslant \sqrt{\mathbb{E}\Big[\big(\widehat{\eta}_n(X) - \eta(X)\big)^2\Big]}\,. \tag{33.2}$$

C'est la vitesse de convergence vers 0 de ce majorant, qui est une quantité naturelle à étudier s'agissant d'estimateurs définis par une propriété de meilleure approximation $\mathbb{L}^2$, que nous comparons à celle du risque de classification.

Le membre de droite de (33.2) tend vers 0 si et seulement si $\|\widehat{\eta}_n - \eta\|_{\mathbb{L}^2(\mu)} \to 0$ en $\mathbb{P}$–probabilité : au vu de (33.2), il suffit de prouver le sens direct, ce qui se fait en remarquant, comme dans la preuve de la Définition–Théorème 20.2, qu'une suite de variables aléatoires positives bornées converge en $\mathbb{P}$–probabilité vers 0 si et seulement si elle converge vers 0 dans $\mathbb{L}^2(\mathbb{P})$. L'hypothèse sur les suites considérées dans le théorème suivant est donc nécessaire pour que le résultat ait un sens.

Théorème 33.1 *Soient $(\widehat{\eta}_n)$ une suite d'estimateurs $\mathbb{P}$–consistants en norme $\mathbb{L}^2(\mu)$ de la fonction de régression η,* id est,

$$\|\widehat{\eta}_n - \eta\|_{\mathbb{L}^2(\mu)} \longrightarrow 0 \qquad \text{en } \mathbb{P}\text{–probabilité,}$$

et (g_n) la règle de classification associée par plug-in. Alors

$$\lim_{n\to\infty} \frac{\mathbb{E}\big[L(g_n)\big] - L(g^*)}{\sqrt{\mathbb{E}\left[\big(\widehat{\eta}_n(X) - \eta(X)\big)^2\right]}} = 0\,.$$

Ce théorème indique ainsi que le risque de classification converge plus rapidement vers 0 que l'erreur d'estimation de la fonction de régression : le problème de classification est plus simple que le problème d'estimation (non-paramétrique) de la fonction de régression. On en rappelle l'explication heuristique donnée dans le corrigé principal : pour bien classifier, il suffit d'estimer η de manière précise dans la zone où η est proche de $1/2$ alors qu'on peut s'autoriser une estimation moins précise autre part, tant qu'elle ne nous fait pas franchir la dite barre des $1/2$.

Preuve On utilise l'inégalité de Cauchy–Schwarz et un découpage du membre de droite de (20.3) selon que $|\eta - 1/2|$ est plus petit ou plus grand que ε :

$$\begin{aligned}
&\mathbb{E}\big[L(g_n)\big] - L(g^*)\\
&\leqslant\ 2\sqrt{\mathbb{E}\left[\big(\widehat{\eta}_n(X) - \eta(X)\big)^2\right]}\quad \sqrt{\mathbb{P}\Big\{g_n(X) \neq g^*(X),\ \ \eta(X) \neq 1/2\Big\}}\\
&\leqslant\ 2\sqrt{\mathbb{E}\left[\big(\widehat{\eta}_n(X) - \eta(X)\big)^2\right]}\quad \left(\sqrt{\mathbb{P}\Big\{\big|\eta(X) - 1/2\big| \leqslant \varepsilon,\ \ \eta(X) \neq 1/2\Big\}}\right.\\
&\qquad\qquad \left. + \sqrt{\mathbb{P}\Big\{\big|\eta(X) - 1/2\big| > \varepsilon,\ \ g_n(X) \neq g^*(X)\Big\}}\right)\,.
\end{aligned}$$

Il suffit de montrer la convergence vers 0 de la somme entre parenthèses, lorsque $n \to +\infty$ puis $\varepsilon \to 0$. D'une part, par convergence monotone,

$$\lim_{\varepsilon\to 0} \mathbb{P}\Big\{\big|\eta(X) - 1/2\big| \leqslant \varepsilon,\ \eta(X) \neq 1/2\Big\} = 0\,.$$

D'autre part, par un fait déjà utilisé en (20.3),

$$\left\{ \left| \eta(X) - 1/2 \right| > \varepsilon,\ g_n(X) \neq g^*(X) \right\} \subseteq \left\{ \left| \eta(X) - \widehat{\eta}_n(X) \right| > \varepsilon \right\},$$

qui est de $\mathbb{P}$–probabilité tendant vers 0 lorsque $n \to +\infty$, puisque la suite des $\widehat{\eta}_n(X)$ converge dans $\mathbb{L}^2(\mathbb{P})$ et donc en $\mathbb{P}$–probabilité vers $\eta(X)$. □

On voit donc qu'il est plus simple de classifier que d'estimer la fonction de régression ; et paradoxalement, nous nous sommes plutôt intéressés dans le corrigé du chapitre 20 à obtenir des estimateurs de la fonction de régression universellement consistants (*id est*, consistants dans le modèle où toutes les lois jointes ν sont possibles). Nous en avons construit par histogrammes et par noyaux. Nous avons procédé ainsi simplement parce qu'il est à la portée d'un étudiant de master 1 ou d'un agrégatif d'en montrer l'universelle consistance. Ce n'est pas le cas de règles plus efficaces.

33.1.2 Preuve du Théorème 20.7

Le Théorème 20.8 améliore le résultat du Théorème 20.7, mais c'est au prix d'une preuve plus complexe. Nous prouvons donc d'abord ce dernier et étudions ensuite comment l'améliorer pour trouver la vitesse d'apprentissage $1/n$. Dans ce paragraphe et le suivant, on reprend les notations du paragraphe 20.1.10.

Les techniques employées ici sont exactement les mêmes que celles du Lemme 19.5 et du Théorème 19.6.

Preuve Dans ce modèle paramétrique, il vient, par (20.1), (20.6) et le rappel fait au début du paragraphe 20.1.10 selon lequel la loi μ de X est absolument continue par rapport à la mesure de Lebesgue de densité $(1-p)\, f_{\lambda_0} + p\, f_{\lambda_1}$, que l'erreur de classification de toute fonction de classification g vaut

$$\begin{aligned} L(g) &= \int_{\mathbb{R}} \Big(\eta(x)\, \mathbb{I}_{\{g(x)=0\}} + \big(1-\eta(x)\big)\, \mathbb{I}_{\{g(x)=1\}} \Big)\, \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}} \Big(p\, f_{\lambda_1}(x)\, \mathbb{I}_{\{g(x)=0\}} + (1-p)\, f_{\lambda_0}(x)\, \mathbb{I}_{\{g(x)=1\}} \Big)\, \mathrm{d}x \\ &= 1 - \int_{\mathbb{R}} \Big(p\, f_{\lambda_1}(x)\, \mathbb{I}_{\{g(x)=1\}} + (1-p)\, f_{\lambda_0}(x)\, \mathbb{I}_{\{g(x)=0\}} \Big)\, \mathrm{d}x\ . \end{aligned}$$

Par un calcul déjà effectué page 395 (pour la classification multi-classes), on voit que la différence entre les erreurs de classification de g_n et g^* est donnée par

$$\begin{aligned} & L\big(g_n\big) - L(g^*) \\ &= \int_{\mathbb{R}} \Big(p\, f_{\lambda_1}(x)\, \mathbb{I}_{\{g^*(x)=1\}} + (1-p)\, f_{\lambda_0}(x)\, \mathbb{I}_{\{g^*(x)=0\}} \Big)\, \mathrm{d}x \\ &\quad - \int_{\mathbb{R}} \Big(\widehat{p}_n\, f_{\hat{\lambda}_{1,n}}(x)\, \mathbb{I}_{\{g_n(x)=1\}} + \big(1-\widehat{p}_n\big)\, f_{\hat{\lambda}_{0,n}}(x)\, \mathbb{I}_{\{g_n(x)=0\}} \Big)\, \mathrm{d}x \end{aligned}$$

$$\leqslant \quad 4\left|\widehat{p}_n - p\right| + 2\int_{\mathbb{R}} \left|f_{\lambda_0}(x) - f_{\hat{\lambda}_{0,n}}(x)\right| \mathrm{d}x + 2\int_{\mathbb{R}} \left|f_{\lambda_1}(x) - f_{\hat{\lambda}_{1,n}}(x)\right| \mathrm{d}x \ . \quad (33.3)$$

On montre que (deux fois) la distance en variation entre deux lois exponentielles de paramètres $\lambda < \lambda'$ est majorée par l'écart entre λ et λ' ; par le théorème de Scheffé (Théorème 24.1) et la forme des densités considérées, il vient

$$\begin{aligned}\int_{\mathbb{R}} \left|f_\lambda(x) - f_{\lambda'}(x)\right| \mathrm{d}x &= 2\int_0^{\varphi(\lambda,\,\lambda')} \left(f_\lambda(x) - f_\mu(x)\right) \mathrm{d}x \\ &= 2\left[e^{-\lambda x} - e^{-\lambda' x}\right]_0^{\varphi(\lambda,\,\lambda')} = 2\,e^{-\lambda'\,\varphi(\lambda,\,\lambda')}\left(e^{-(\lambda-\lambda')\,\varphi(\lambda,\,\lambda')} - 1\right)\end{aligned}$$

où

$$\varphi(\lambda,\,\lambda') = \psi(\lambda,\,\lambda',\,1/2) = \frac{1}{\lambda - \lambda'}\,\log\frac{\lambda}{\lambda'} \geqslant 0$$

est le point x où les densités sont égales, $f_\lambda(x) = f'_\lambda(x)$. En substituant cette valeur et en majorant le facteur exponentiel par 1, il reste

$$\int_{\mathbb{R}} \left|f_\lambda(x) - f_{\lambda'}(x)\right| \mathrm{d}x \leqslant 2\left(\frac{\lambda'}{\lambda} - 1\right) \ .$$

Par symétrie dans le raisonnement précédent entre λ et λ', on a donc majoré l'erreur de classification selon

$$L\big(g_n\big) - L(g^*) \leqslant M_n \overset{\text{not.}}{=} 4\left|\widehat{p}_n - p\right| + 4\left|\frac{\widehat{\lambda}_{0,n} - \lambda_0}{\min\left\{\lambda_0,\,\widehat{\lambda}_{0,n}\right\}}\right| + 4\left|\frac{\widehat{\lambda}_{1,n} - \lambda_1}{\min\left\{\lambda_1,\,\widehat{\lambda}_{1,n}\right\}}\right| \ . \quad (33.4)$$

On montre maintenant que $\sqrt{n}\,M_n$, le membre de droite de cette inégalité multiplié par $\sqrt{n}$, converge en loi (vers une loi que l'on va préciser par l'analyse).

On applique pour cela le théorème de la limite centrale dans sa version vectorielle, en considérant les vecteurs $U_1,\,\ldots,\,U_n$, indépendants et identiquement distribués, où pour $j = 1,\,\ldots,\,n$,

$$U_j = \begin{bmatrix} U_{j,1} \\ U_{j,2} \\ U_{j,3} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\{Y_j=1\}} \\ X_j\,\mathbb{I}_{\{Y_j=0\}} \\ X_j\,\mathbb{I}_{\{Y_j=1\}} \end{bmatrix} \ .$$

On note $\overline{U}_n$ la moyenne empirique des U_j, c'est-à-dire, $\overline{U}_n = (U_1 + \ldots + U_n)/n$; on a

$$\sqrt{n}\left(\overline{U}_n - \underline{m}\right) \rightsquigarrow \mathcal{N}(0,\,\Gamma) \quad (33.5)$$

où

$$\underline{m} = \begin{bmatrix} p \\ (1-p)/\lambda_0 \\ p/\lambda_1 \end{bmatrix}$$

est l'espérance commune des U_j et Γ leur matrice de variance–covariance :

$$\Gamma = \begin{bmatrix} p(1-p) & -p(1-p)/\lambda_0 & p(1-p)/\lambda_1 \\ -p(1-p)/\lambda_0 & (1-p)(1+p)/\lambda_0^2 & -p(1-p)/(\lambda_0\lambda_1) \\ p(1-p)/\lambda_1 & -p(1-p)/(\lambda_0\lambda_1) & p(2-p)/\lambda_1^2 \end{bmatrix}.$$

On indique les éléments qui ont permis de calculer Γ : l'espérance et la variance d'une loi exponentielle $\mathcal{E}(\lambda)$ sont respectivement $1/\lambda$ et $2/\lambda^2$, celle d'une loi de Bernoulli de paramètre p est $p(1-p)$. Alors d'une part pour les éléments diagonaux, on a par exemple

$$\begin{aligned}\Gamma_{2,2} = \mathbb{E}\Big[X^2\,\mathbb{I}_{\{Y=0\}}\Big] - \Big(\mathbb{E}\big[X\,\mathbb{I}_{\{Y=0\}}\big]\Big)^2 &= \mathbb{P}\{Y=0\}\;\mathbb{E}\Big[X^2 \mid Y=0\Big] - \left(\frac{1-p}{\lambda_0}\right)^2 \\ &= (1-p)\,\frac{2}{\lambda_0^2} - \frac{(1-p)^2}{\lambda_0^2} = \frac{(1-p)(1+p)}{\lambda_0^2}\,,\end{aligned}$$

de quoi on déduit la valeur de $\Gamma_{3,3}$ par symétrie. D'autre part, pour les éléments hors diagonale, on utilise les valeurs des composantes de $\underline{m}$ et le fait que vu les produits de fonctions indicatrices,

$$\mathbb{E}\big[U_{j,1}\,U_{j,2}\big] = 0\,, \qquad \mathbb{E}\big[U_{j,1}\,U_{j,3}\big] = \mathbb{E}\big[U_{j,3}\big] = p/\lambda_1\,, \qquad \mathbb{E}\big[U_{j,2}\,U_{j,3}\big] = 0\,.$$

On note $\Phi : \mathbb{R}^3 \to \mathbb{R}^3$ la fonction vectorielle

$$\Phi : \begin{bmatrix} x \\ y \\ z \end{bmatrix} \longmapsto \begin{bmatrix} x \\ (1-x)/y \\ x/z \end{bmatrix};$$

elle a été prise de telle sorte que

$$\Phi(\underline{m}) = \begin{bmatrix} p \\ \lambda_0 \\ \lambda_1 \end{bmatrix} \qquad \text{et} \qquad \Phi\big(\overline{U}_n\big) = \begin{bmatrix} \widehat{p}_n \\ \widehat{\lambda}_{0,n} \\ \widehat{\lambda}_{1,n} \end{bmatrix}.$$

Φ est différentiable sur $\mathbb{R} \times \mathbb{R}_+^* \times \mathbb{R}_+^*$, de matrice différentielle égale à

$$\mathrm{D}\Phi(x,y,z) = \begin{bmatrix} 1 & 0 & 0 \\ -1/y & x/y^2 & 0 \\ 1/z & 0 & -x/z^2 \end{bmatrix}.$$

La méthode delta (voir le Lemme 2.9) permet alors de passer de la convergence (33.5) à

$$\sqrt{n}\left(\begin{bmatrix} \widehat{p}_n \\ \widehat{\lambda}_{0,n} \\ \widehat{\lambda}_{1,n} \end{bmatrix} - \begin{bmatrix} p \\ \lambda_0 \\ \lambda_1 \end{bmatrix}\right) \rightsquigarrow \mathrm{D}\Phi(\underline{m})\,\mathcal{N}\big(0,\,\Gamma\big)\,. \tag{33.6}$$

En particulier, $\widehat{\lambda}_{0,n}$ et $\widehat{\lambda}_{1,n}$ sont des estimateurs consistants de λ_0 et λ_1, convergent en probabilité vers eux, de sorte que le lemme de Slutzky (utilisé dans sa forme multidimensionnelle, Lemme 26.8) garantit finalement que

$$\sqrt{n}\left(\begin{bmatrix} \widehat{p}_n \\ \widehat{\lambda}_{0,n}/\min\left\{\lambda_0,\, \widehat{\lambda}_{0,n}\right\} \\ \widehat{\lambda}_{1,n}/\min\left\{\lambda_1,\, \widehat{\lambda}_{1,n}\right\} \end{bmatrix} - \begin{bmatrix} p \\ \lambda_0/\min\left\{\lambda_0,\, \widehat{\lambda}_{0,n}\right\} \\ \lambda_1/\min\left\{\lambda_1,\, \widehat{\lambda}_{1,n}\right\} \end{bmatrix}\right) \rightsquigarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\lambda_0 & 0 \\ 0 & 0 & 1/\lambda_1 \end{bmatrix} \mathrm{D}\Phi(\underline{m})\, \mathcal{N}\big(0,\, \Gamma\big)$$

Enfin, l'application de la fonction continue $g : (x, y, z) \mapsto 4\big(|x| + |y| + |z|\big)$ montre que $\sqrt{n}\, M_n$ converge en loi,

$$\sqrt{n}\, M_n = 4\sqrt{n}\left(|\widehat{p}_n - p| + \left|\frac{\widehat{\lambda}_{0,n} - \lambda_0}{\min\left\{\widehat{\lambda}_{0,n},\, \lambda_0\right\}}\right| + \left|\frac{\widehat{\lambda}_{1,n} - \lambda_1}{\min\left\{\widehat{\lambda}_{1,n},\, \lambda_1\right\}}\right|\right) \rightsquigarrow \mathcal{L}$$

où $\mathcal{L}$ est la loi de

$$g\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\lambda_0 & 0 \\ 0 & 0 & 1/\lambda_1 \end{bmatrix} \mathrm{D}\Phi(\underline{m})\, \mathcal{N}\big(0,\, \Gamma\big)\right),$$

et est donc une loi dérivée de la loi normale. □

33.1.3 Eléments de preuve pour le Théorème 20.8

Si l'on regarde attentivement comment (33.3) a été obtenue, on constate qu'on a négligé un terme $\mathbb{I}_{\{g^*(x)\neq g_n(x)\}}$, et c'est pourquoi on est retombé sur un problème d'estimation pure et que l'on a obtenu la vitesse paramétrique $1/\sqrt{n}$. S'agissant d'un problème d'apprentissage, la vitesse est plus rapide : nous allons montrer comment considérer le terme $\mathbb{I}_{\{g^*(x)\neq g_n(x)\}}$ permet de montrer que la vitesse est en fait au moins en $1/n$.

On suppose ici $\lambda_0 \neq \lambda_1$; on peut traiter le cas $\lambda_0 = \lambda_1$ de manière similaire à ce qui suit et nous laissons ce travail en exercice.

Plus précisément, les majorations des intégrales de la page 395 procèdent de majoration des intégrandes uniquement ; or, lorsque $g^*(x)$ et $g_n(x)$ sont égales, l'erreur commise est nulle. On aboutit donc par cet argument plus fin à la majoration

$$\begin{aligned} L\big(g_n\big) - L(g^*) \;\leqslant\; M_n'' \;\overset{\text{not.}}{=}\; & 4\left|\widehat{p}_n - p\right| \int_{\mathbb{R}} \mathbb{I}_{\{g^*(x)\neq g_n(x)\}}\, \mathrm{d}x \\ & + 2\int_{\mathbb{R}} \left|f_{\lambda_0}(x) - f_{\hat{\lambda}_{0,n}}(x)\right| \mathbb{I}_{\{g^*(x)\neq g_n(x)\}}\, \mathrm{d}x \\ & + 2\int_{\mathbb{R}} \left|f_{\lambda_1}(x) - f_{\hat{\lambda}_{1,n}}(x)\right| \mathbb{I}_{\{g^*(x)\neq g_n(x)\}}\, \mathrm{d}x\,. \end{aligned}$$

On appelle M_n'' ce majorant car on va le majorer à son tour par la quantité M_n' apparaissant dans le Théorème 20.8.

Vu (20.7) et l'expression similaire pour g_n en remplaçant les trois paramètres par leurs estimateurs, on a deux situations :

– un cas très défavorable où $\lambda_0 - \lambda_1$ et $\widehat{\lambda}_{0,n} - \widehat{\lambda}_{1,n}$ sont de signes contraires ; les fonctions de classification g^* et g_n sont alors différentes partout, sauf sur l'intervalle

$$I_n = \left[\min\left\{ \psi_0, \widehat{\psi}_n \right\}, \ \max\left\{ \psi_0, \widehat{\psi}_n \right\} \right]$$

$$\text{où} \qquad \psi_0 = \psi(\lambda_0, \lambda_1, p) \qquad \text{et} \qquad \widehat{\psi}_n = \psi\left(\widehat{\lambda}_{0,n}, \widehat{\lambda}_{1,n}, \widehat{p}_n \right),$$

ce qui conduit à une grosse erreur de classification ; cette situation arrive heureusement exponentiellement peu souvent ;

– le cas le plus fréquent où les deux fonctions de classification ne diffèrent que sur l'intervalle I_n.

L'heuristique est que la longueur de I_n est de l'ordre de $1/\sqrt{n}$ et que les différences $f_{\lambda_0} - f_{\hat{\lambda}_{0,n}}$ et $f_{\lambda_1} - f_{\hat{\lambda}_{1,n}}$ sont aussi de ce même ordre de grandeur sur cet intervalle, ce qui donne la vitesse $1/n$ au final. On va maintenant détailler ce raisonnement.

On note à cet effet D_n l'événement donné par le cas très défavorable et on va majorer notre borne intermédiaire M_n'' sur D_n et D_n^c pour obtenir la borne M_n' du Théorème 20.8.

Premièrement, une étude de fonctions rapide montre que pour $0 < \lambda < \lambda'$,

$$\left\| f_\lambda - f_{\lambda'} \right\|_\infty = e^{-\lambda y^*} \left(\lambda - \lambda' e^{-(\lambda' - \lambda) y^*} \right) \leqslant \frac{\lambda}{\lambda'} \left(\lambda' - \lambda \right) \leqslant \left| \lambda' - \lambda \right|$$

où le supremum dans le membre de gauche est atteint en $y^* > 0$ tel que

$$e^{-(\lambda' - \lambda) y^*} = \left(\frac{\lambda}{\lambda'} \right)^2 .$$

Ainsi, sur l'événement D_n^c, comme g^* et g_n ne diffèrent que sur I_n, de longueur $\left| \widehat{\psi}_n - \psi_0 \right|$ (et en majorant des facteurs 2 par un facteur 4),

$$M_n'' \leqslant 4 \left| \widehat{\psi}_n - \psi_0 \right| \left(\left| \widehat{p}_n - p \right| + \left| \widehat{\lambda}_{0,n} - \lambda_0 \right| + \left| \widehat{\lambda}_{1,n} - \lambda_1 \right| \right) ;$$

les deux termes du produit présent dans la majoration étant fonctions continues de $p - \widehat{p}_n$, $\lambda_0 - \widehat{\lambda}_{0,n}$ et $\lambda - \widehat{\lambda}_{1,n}$, et puisque la convergence en loi passe aux fonctions continues, on déduit, comme précédemment, de (33.6) une convergence en loi

$$\sqrt{n} \begin{bmatrix} \left| \widehat{\psi}_n - \psi_0 \right| \\ \left| \widehat{p}_n - p \right| + \left| \widehat{\lambda}_{0,n} - \lambda_0 \right| + \left| \widehat{\lambda}_{1,n} - \lambda_1 \right| \end{bmatrix} \rightsquigarrow \ \ldots$$

(vers une loi qu'il ne serait pas difficile, mais pénible, d'exhiber explicitement) et en particulier, on a également la convergence en loi

$$4n \left| \widehat{\psi}_n - \psi_0 \right| \left(\left| \widehat{p}_n - p \right| + \left| \widehat{\lambda}_{0,n} - \lambda_0 \right| + \left| \widehat{\lambda}_{1,n} - \lambda_1 \right| \right) \rightsquigarrow \ldots$$

Comme l'erreur de classification est toujours majorée par 1, on procède à une majoration grossière sur D_n et on pose donc

$$M'_n = \mathbb{I}_{D_n} + 4\,\left|\widehat{\psi}_n - \psi_0\right|\,\left(\left|\widehat{p}_n - p\right| + \left|\widehat{\lambda}_{0,n} - \lambda_0\right| + \left|\widehat{\lambda}_{1,n} - \lambda_1\right|\right)\mathbb{I}_{D_n^c}\,.$$

La preuve sera conclue par lemme de Slutzky pour peu que $\mathbb{P}\big(D_n\big) = o(1/n)$.

À cet effet, on rappelle qu'au vu de (20.7) le mauvais cas indiqué par D_n arrive seulement lorsque

– λ_0 et λ_1 d'une part,
– $\widehat{\lambda}_{0,n}$ et $\widehat{\lambda}_{1,n}$ d'autre part,

sont classés dans des ordres différents. C'est équivalent au fait que

$$(\lambda_1 - \lambda_0)\Big(\widehat{\lambda}_{1,n} - \widehat{\lambda}_{0,n}\Big) < 0\,, \qquad \text{ou encore,} \qquad \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_0}\right)\left(\frac{1}{\widehat{\lambda}_{1,n}} - \frac{1}{\widehat{\lambda}_{0,n}}\right) < 0\,.$$

En particulier, une condition nécessaire est que

$$\text{ou} \quad \left|\frac{1}{\widehat{\lambda}_{1,n}} - \frac{1}{\lambda_1}\right| \geqslant \frac{1}{2}\left|\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right|\,, \qquad \text{ou} \quad \left|\frac{1}{\widehat{\lambda}_{0,n}} - \frac{1}{\lambda_0}\right| \geqslant \frac{1}{2}\left|\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right|\,.$$

Nous allons utiliser une inégalité de Bernstein (voir le Théorème 28.2, qui se place cependant dans le cadre plus général des martingales), à ceci près qu'il nous en faut une version pour des variables non bornées : ici, nous avons essentiellement affaire à des lois exponentielles. Nous recourons par conséquent à la version avec condition de moments exponentiels énoncée, par exemple, dans [Mas07, chapitre 2]. Elle indique qu'avec probabilité au moins $1 - 2\,\delta$,

$$\left|\frac{np}{\lambda_1} - \sum_{t=1}^{n} X_t\,\mathbb{I}_{\{Y_t=1\}}\right| \leqslant \square\sqrt{n\log\frac{1}{\delta}} + \square\,\log\frac{1}{\delta}$$

où les $\square$ désignent des quantités dépendant des paramètres p et λ_1, mais pas de n. De même, l'inégalité de Hoeffding (voir le Lemme 3.7) donne qu'avec probabilité au moins $1 - 2\,\delta$,

$$\left|np - \sum_{t=1}^{n}\mathbb{I}_{\{Y_t=1\}}\right| \leqslant \sqrt{\frac{n}{2}\,\log\frac{2}{\delta}}\,.$$

Au final, en combinant ces deux inégalités, on a avec probabilité au moins $1 - 4\,\delta$ que

$$\left|\frac{1}{\widehat{\lambda}_{1,n}} - \frac{1}{\lambda_1}\right| \leqslant \square\sqrt{\frac{\log 1/\delta}{n}} + \square\,\frac{\log 1/\delta}{n}\,;$$

en particulier, la probabilité que

$$\left|\frac{1}{\widehat{\lambda}_{1,n}} - \frac{1}{\lambda_1}\right| \geqslant \frac{1}{2}\left|\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right| > 0$$

est plus petite que $\square\, e^{-\square n} = o(1/n)$, où les constantes $\square$ dépendent de $\lambda_0,\ \lambda_1,\ p$ mais pas de n.

De même pour l'écart entre $1\big/\widehat{\lambda}_{0,n}$ et $1/\lambda_0$; cela conclut cette preuve (qui est plutôt du niveau master 2, comme on l'aura remarqué par le recours à l'inégalité de Bernstein).

33.2 Codes des simulations

33.2.1 Fonctions de simulations des données artificielles

```
function [X, Y] = SimExp(N)
%
% Simulation de couples observation--étiquette
% selon le modèle exponentiel
%
% Entrée :
%   N, un entier naturel
% Sorties :
%   X et Y, deux vecteurs de taille N contenant
%           respectivement des observations (selon
%           des lois exponentielles) et des étiquettes
%

  X = zeros(1,N);
  Y = [rand(1,N) <= 0.5];

  c = (Y == 1);
  n = sum(c);

  X(c) = rexpweib([1 n], 1);
  X(not(c)) = rexpweib([1 N-n], 0.4);

end



function [X, Y] = SimGauss(N)
%
% Simulation de couples observation--étiquette
% selon le modèle gaussien
%
% Entrée :
%   N, un entier naturel
% Sorties :
%   X et Y, deux vecteurs de taille 2 x N contenant
%           respectivement des observations (selon des lois
%           normales bi-dimensionnelles) et des étiquettes
```

```
%

  X = zeros(2,N);
  Y = [rand(1,N) <= 0.4];

  c = (Y == 1);
  n = sum(c);

  X(:,c) = [2 0; 0.5 1] * randn([2 n]);
  X(1,c) = X(1,c) + 2;
  X(2,c) = X(2,c) - 1.5;

  X(:,not(c)) = [1 -1; 0 1] * randn([2 N-n]);

end
```

33.2.2 Fonction de classification par histogrammes cubiques

```
function D = CHistoCubiq(X, x, y, h)
%
% Classification par histogrammes cubiques
%
% Entrées :
%   X, des observations à classer
%   x, des observations pour lesquelles on a déjà :
%   y, le vecteur des étiquettes des éléments de x
%   h, le paramètre de grille
% Sortie :
%   D, le vecteur des étiquettes pour les éléments de X
%      (valeurs possibles : 0, 0.5, 1)
%

  n = size(x); n = n(end);
  c = floor(X/h);
  [U, I, J] = unique(c','rows');
  D = [];
  for u = U'
    ind = (repmat(u*h, 1, n) <= x) & (x < repmat((u+1)*h, 1, n));
    ind = logical(prod(ind+0, 1));
    D = [D compare(sum(y(ind)),sum(1-y(ind)))];
  end;

  if (length(D) >= 2)
    D = D(J);
  else
    D = D(J)';
  end;
```

```
end
```

33.2.3 Fonction auxiliaire utile pour les méthodes décidant selon une majorité

```
function r = compare(a, b)
%
% Fonction auxiliaire nécessitée lors de toute décision majoritaire
% Il y a indétermination si les deux termes à comparer sont issus
% tous deux de sommes sur aucun élément (tous deux nuls) ou, plus
% généralement, s'ils sont égaux
%
% Entrées :
%   a et b deux vecteurs de scalaires positifs
% Sortie :
%   r un vecteur comparant a et b terme à terme
%

  r = (a > b) + 0.5 * (a == b);

end
```

33.2.4 Fonction de classification par méthodes de noyaux

```
function D = CNoyaux(X, x, y, k, h)
%
% Classification par méthodes de noyaux (gaussien ou uniforme symétrique)
%
% Entrées :
%   X, des observations à classer
%   x, des observations pour lesquelles on a déjà :
%   y, le vecteur des étiquettes des éléments de x
%   k, booléen, indique le type de noyau
%      (k = true donne le noyau uniforme, k = false le gaussien)
%   h, le paramètre de bande
% Sortie :
%   D, le vecteur des étiquettes pour les éléments de X
%      (valeurs possibles : 0, 0.5, 1)
%

  n = size(x);
  n = n(end);

  b = (y == 1);
  D = [];
  for u = X
    M = (repmat(u, 1, n) - x)/h;
```

```
    if k
      K = logical(prod(0+(abs(M) <= 1), 1));
    else
      K = exp(-sum(M.^2, 1)/2);
    end;
    D = [D compare(sum(K(b)), sum(K(not(b))))];
  end;

end
```

33.2.5 Fonctions de classification par plug-in paramétrique

```
function D = PluginExp(X, x, y, estimer)
%
% Classification par plug-in paramétrique (cas du modèle exponentiel)
%
% Entrées :
%   X, des observations à classer
%   x, des observations pour lesquelles on a déjà :
%   y, le vecteur des étiquettes des éléments de x
%   estimer, true si l'on ne connaît pas les paramètres et qu'il faut
%            les estimer (cas de la construction d'une fonction de
%            classification); false si l'on veut approximer L^*
% Sortie :
%   D, le vecteur des étiquettes pour les éléments de X
%      (valeurs possibles : 0 et 1)
%

  if estimer
    l0 = 1/mean(x(:, y == 0));
    l1 = 1/mean(x(:, y == 1));
    p = mean(y);
  else
    l0 = 0.4;
    l1 = 1;
    p = 0.5;
  end;

  D = compare(p*dexp(X,l1), (1-p)*dexp(X,l0));

end

function d = dexp(X, l)
% Densité de la loi exponentielle de paramètre l en X

  d = l * exp(-l * X);
```

```
end



function D = PluginExpSeuil(X, x, y)
%
% Classification par détermination du meilleur seuil empirique
% (cas du modèle exponentiel)
%
% Entrées :
%   X, des observations à classer
%   x, des observations pour lesquelles on a déjà :
%   y, le vecteur des étiquettes des éléments de x
% Sortie :
%   D, le vecteur des étiquettes pour les éléments de X
%      (valeurs possibles : 0 et 1, ici, pas d'indétermination : p.s. les
%       nouvelles observations X ne tombent pas pile sur le seuil)
%

  [x, I] = sort(x);
  y = y(I);
  P = [0 (x(1:(end-1))+x(2:end))/2 x(end)+1];

  R = [];
  for p = P
    d = (x <= p);
    R = [R; sum(abs(d-y)) sum(abs((1-d)-y))];
  end;

  R = R';
  [m, J] = min(R(:));
  p = P(floor((J+1)/2));

  if mod(J,2)
    D = (X <= p);
  else
    D = (X >= p);
  end;

end



function D = PluginGaussien(X, x, y, estimer)
%
% Classification par plug-in paramétrique (cas du modèle gaussien)
```

```
%
% Entrées :
%   X, des observations à classer
%   x, des observations pour lesquelles on a déjà :
%   y, le vecteur des étiquettes des éléments de x
%   estimer, true si l'on ne connaît pas les paramètres et qu'il faut
%            les estimer (cas de la construction d'une fonction de
%            classification); false si l'on veut approximer L^*
% Sortie :
%   D, le vecteur des étiquettes pour les éléments de X
%      (valeurs possibles : 0 et 1)
%

  if estimer
    C1 = cov(x(:,y==1)');
    m1 = mean(x(:,y==1), 2);
    C0 = cov(x(:,y==0)');
    m0 = mean(x(:,y==0), 2);
    p = mean(y);
  else
    C1 = [2 0; 0.5 1]; C1 = C1*C1';
    m1 = [2; -1.5];
    C0 = [1 -1; 0 1]; C0 = C0*C0';
    m0 = [0; 0];
    p = 0.4;
  end;

  D = compare(p*dnorm2(X,m1,C1), (1-p)*dnorm2(X,m0,C0));

end

function d = dnorm2(X, m, C)
% A un facteur multiplicatif près, la densité de la loi normale
% de moyenne m et de matrice de covariance C aux points X

  I = inv(C);
  isd = 1/sqrt(det(C));
  d = [];
  for x = X
    d = [d isd * exp(-(x-m)'*I*(x-m)/2)];
  end;

end
```

33.2.6 Fonction de classification par plus proches voisins

```
function D = kNN(X, x, y, k)
```

```
%
% Classification par les plus proches voisins
%
% Entrées :
%   X, des observations à classer
%   x, des observations pour lesquelles on a déjà :
%   y, le vecteur des étiquettes des éléments de x
%   k, le nombre de voisins à considérer
% Sortie :
%   D, le vecteur des étiquettes pour les éléments de X
%      (valeurs possibles : 0 et 1 ; on prendra garde à lancer
%       cette fonction avec k impair pour éviter les indéterminations)
%

  n = size(x);
  n = n(end);
  %
  D = [];
  for u = X
    dist = sum((x-repmat(u, 1, n)).^2, 1);
    [dist, I] = sort(dist);
    D = [D (mean(y(I(1:k))) >= 1/2)];
  end;

end
```

33.2.7 Fonction de tracé des résultats

```
function [] = TraceEG(unidim, x, y, X, Y, D, grille)
%
% Fonction de tracé des résultats d'une classification
%
% Entrées :
%   unidim, vaut true si on a affaire au modèle exponentiel et false pour
%           le modèle gaussien
%   x, des observations pour lesquelles on a déjà vu :
%   y, le vecteur des étiquettes des éléments de x
%   X, des observations classées par une fonction de classification en :
%   D, le vecteur des étiquettes proposées pour les éléments de X
%   Y, le vecteur des vraies étiquettes pour les éléments de X
%   grille, true si une grille doit être tracée et false sinon
% Sorties :
%   Deux figures, une pour les vraies étiquettes 1, et l'autre, pour les 0
%

  symbol = {'^', 'x', '+'};
  taille = {6, 7, 6};
```

```
  % '^' = triangle : quand bien classifié
  % 'x' = croix diagonale : quand mal classifié
  % '+' = croix verticale : quand indétermination

  if unidim
    X = [X; jitter(length(X))];
    x = [x; jitter(length(x))];
  end;

  for i = [0 1]
    figure(1+i);
    clf;
    hold on;
    for k = [0 1]
      gca1 = plot(x(1, y == k), x(2, y == k), 'ok', 'MarkerSize', 5);
      if k
        set(gca1, 'MarkerFaceColor', 'k');
      end;
    end;
    for d = [0 1 0.5]
      p = (Y == i) & (D == d);
      if not(isempty(p))
        gca2 = plot(X(1,p), X(2,p), [symbol{indice(i,d)} 'k'], ...
                    'MarkerSize', taille{indice(i,d)});
        if grille
            grid on;
        end;
        if i
          set(gca2, 'MarkerFaceColor', 'k');
        end;
      end;
    end;
    if unidim
      axis([-0.5 6 -0.2 0.2]);
      plot([-0.5 6], [0 0], 'k');
      set(get(gca1, 'Parent'), 'YTick', [])
    end;
  end;

end

function V = jitter(n)
% Choix d'une ordonnée aléatoire

  jitter = 0.1;
  V = 2 * jitter * rand(1, n) - jitter;
```

```
end

function I = indice(i,d)
% Choix du symbole selon l'étiquette lue 0, 0.5 ou 1, et sa
% correspondance à la vraie étiquette i

  I = 3*(d==0.5) + (d~=0.5)*(abs(i-d)+1);

end
```

33.2.8 Fichier script permettant de réaliser des simulations et tracés

```
% Exemple de fonction maître pour réaliser les tracés
clc, clear;
N = 50;
M = 50;
unidim = false;
if unidim
    sim = @SimExp;
else
    sim = @SimGauss;
end;
[x, y] = sim(N);
[X, Y] = sim(M);

  D = CHistoCubiq(X, x, y, 1);   grille = true;
% D = CNoyaux(X, x, y, true, 1); grille = false;
% D = PluginExp(X, x, y, true);  grille = false;

TraceEG(unidim, x, y, X, Y, D, grille);
```

33.2.9 Fichier script permettant une comparaison numérique des performances des différentes méthodes

```
% Fichier maître pour l'évaluation et la comparaison des performances
% numériques des différentes méthodes de classification

clc, clear;
unidim = true;
N = 50;
M = 200;
m = 200;

tic;
E = [];
% Double moyenne de Monte-Carlo pour calculer l'erreur moyenne
% Une boucle sur j pour la seconde moyenne, la première
% moyenne est effectuée lors du calcul des composantes de R
```

```
for j = 1:m
  % Simulation des données
  if unidim
    sim = @SimExp;
  else
    sim = @SimGauss;
  end;
  [x, y] = sim(N);
  [X, Y] = sim(M);
  %
  % Vecteur des résultats pour un j donné
  R = [];
  % Histogrammes cubiques
  for h = [0.5 1 2 3]
    D = CHistoCubiq(X, x, y, h);
    R = [R mean(abs(D-Y))];
  end;
  for h = [0.5 1 2 3]
  % Par noyau uniforme
    D = CNoyaux(X, x, y, true, h);
    R = [R mean(abs(D-Y))];
    % Par noyau gaussien
    D = CNoyaux(X, x, y, false, h);
    R = [R mean(abs(D-Y))];
  end;
  % Méthode k-NN
  for k = [1 3 11]
    D = kNN(X, x, y, k);
    R = [R mean(abs(D-Y))];
  end;
  % Par plug-in
  % Et on calcule au passage une estimation de L^*
  if unidim
    D1 = PluginExp(X, x, y, true);
    D2 = PluginExpSeuil(X, x, y);
    Dtheo = PluginExp(X, x, y, false);
    R = [R mean(abs(D1-Y)) mean(abs(D2-Y)) mean(abs(Dtheo-Y))];
  else
    D = PluginGaussien(X, x, y, true);
    Dtheo = PluginGaussien(X, x, y, false);
    R = [R mean(abs(D-Y)) mean(abs(Dtheo-Y))];
  end;
  %
  E = [E; R];
end;
disp(mean(E));
toc;
```

CHAPITRE 34

Compression de données (compléments)

Résumé. Ce chapitre complète le chapitre 21. On étudie ici le modèle formé par une source émettant des symboles selon une loi stationnaire. On commence notamment par définir le taux d'entropie d'un processus stationnaire. On procure également les codes des simulations `Matlab`.

34.1 Compléments mathématiques

34.1.1 Sous-additivité de l'entropie, taux d'entropie d'un processus stationnaire

Dans ce paragraphe, on veut définir ce qu'est le taux d'entropie d'un processus stationnaire $X = (X_1, X_2, \ldots)$. On rappelle qu'on note $\mathbb{P}_N$ la loi des N premières marginales X_1^N de ce processus et $\mathbb{P}_\infty$ sa loi sur $\mathcal{A}^{\mathbb{N}}$.

Définition 34.1 On appelle taux d'entropie H_∞ du processus X la limite des entropies moyennes $H(\mathbb{P}_N)/N$, lorsque cette dernière existe :

$$H_\infty = \lim_{N\to\infty} \frac{H(\mathbb{P}_N)}{N} .$$

Par exemple, quand le processus est formé de variables aléatoires indépendantes et identiquement distribuées selon une loi commune $\boldsymbol{p}$, on a pour tout entier $N \geqslant 1$ l'égalité $H(\mathbb{P}_N)/N = H(\boldsymbol{p})$. Le taux d'entropie existe alors et est égal à cette valeur commune : $H_\infty = H(\boldsymbol{p})$.

On va montrer que plus généralement, le taux d'entropie est défini pour tout processus stationnaire. A cet effet, on se propose de montrer au préalable que la suite formée par les entropies $H(\mathbb{P}_N)$ est sous-additive.

Sous-additivité de l'entropie

On l'établit dans un cadre général. On considère deux ensembles finis $\mathcal{X}$ et $\mathcal{Y}$, et une probabilité $\boldsymbol{p}$ sur $\mathcal{X} \times \mathcal{Y}$. Les marginales sur $\mathcal{X}$ et $\mathcal{Y}$ sont notées respectivement $\boldsymbol{p}_\mathcal{X}$ et $\boldsymbol{p}_\mathcal{Y}$.

Proposition 34.2 L'entropie de $\boldsymbol{p}$ est majorée par la somme des entropies de ses marginales $\boldsymbol{p}_\mathcal{X}$ et $\boldsymbol{p}_\mathcal{Y}$,

$$H(\boldsymbol{p}) \leqslant H(\boldsymbol{p}_\mathcal{X}) + H(\boldsymbol{p}_\mathcal{Y}) ,$$

avec égalité si et seulement si $\boldsymbol{p}$ est une loi–produit, *id est*, $\boldsymbol{p} = \boldsymbol{p}_{\mathcal{X}} \otimes \boldsymbol{p}_{\mathcal{Y}}$ est le produit de ses lois marginales.

Preuve L'entropie de la loi jointe $H(\boldsymbol{p})$ vaut, par définition,

$$\begin{aligned} H(\boldsymbol{p}) &= -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{(x,y)} \log_2 p_{(x,y)} \\ &= -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{(x,y)} \log_2 \frac{p_{(x,y)}}{p_x p_y} - \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{(x,y)} \big(\log_2 p_x + \log_2 p_y\big) \\ &= -\mathcal{K}\big(\boldsymbol{p},\ \boldsymbol{p}_{\mathcal{X}} \otimes \boldsymbol{p}_{\mathcal{Y}}\big) + H(\boldsymbol{p}_{\mathcal{X}}) + H(\boldsymbol{p}_{\mathcal{Y}}) \,. \end{aligned}$$

La troisième égalité procède de la définition des marginales, en particulier, on a par exemple $\sum_y p_{(x,y)} = p_x$. La deuxième égalité est légitimée par le fait que si $p_{(x,y)} > 0$, alors $p_x > 0$ et $p_y > 0$ (soit $\boldsymbol{p} \ll \boldsymbol{p}_{\mathcal{X}} \otimes \boldsymbol{p}_{\mathcal{Y}}$) ; il n'y a pas d'introduction de formes indéterminées. En utilisant la positivité de $\mathcal{K}$ (voir le Lemme 21.6), il vient alors

$$H(\boldsymbol{p}) \leqslant H(\boldsymbol{p}_{\mathcal{X}}) + H(\boldsymbol{p}_{\mathcal{Y}}) \,.$$

On note que l'égalité a lieu si et seulement si $\mathcal{K}\big(\boldsymbol{p},\ \boldsymbol{p}_{\mathcal{X}} \otimes \boldsymbol{p}_{\mathcal{Y}}\big) = 0$, soit, si et seulement si $\boldsymbol{p} = \boldsymbol{p}_{\mathcal{X}} \otimes \boldsymbol{p}_{\mathcal{Y}}$ (voir, là encore, le Lemme 21.6). □

> Remarque au passage. La quantité $\mathcal{K}\big(\boldsymbol{p},\ \boldsymbol{p}_{\mathcal{X}} \otimes \boldsymbol{p}_{\mathcal{Y}}\big)$ est appelée information mutuelle (entre les deux marginales de $\boldsymbol{p}$). On définit également une notion d'entropie conditionnelle : la différence $H(\boldsymbol{p}) - H(\boldsymbol{p}_{\mathcal{X}})$ est l'entropie conditionnelle de la loi jointe par rapport à sa première marginale. Pour les liens entre entropie d'une loi jointe, entropie conditionnelle et information mutuelle, on lira avec profit [CT91, chapitre 2].

Taux d'entropie d'un processus stationnaire

On rappelle qu'on dit que X est stationnaire si pour tous entiers $t \geqslant 1$ et $k \geqslant 1$, on a l'égalité des lois

$$(X_1, \ldots, X_k) \overset{(\mathrm{d})}{=} (X_{t+1}, \ldots, X_{t+k}) \,.$$

Dans ce cas, la suite des entropies moyennes $H(\mathbb{P}_N)/N$ admet un comportement asymptotique similaire à celui d'une suite décroissante.

Proposition 34.3 Si X est stationnaire, alors il admet un taux d'entropie, égal à

$$H_\infty = \lim_{N\to\infty} \frac{H(\mathbb{P}_N)}{N} = \inf_{N\in\mathbb{N}^*} \frac{H(\mathbb{P}_N)}{N} \,.$$

Preuve On rappelle que $\mathbb{P}_N$ désigne la loi $\mathbb{P}^{X_1^N}$ des N premières marginales X_1^N. Par la Proposition 34.2, pour N et M deux entiers non nuls, on a

$$H(\mathbb{P}_{N+M}) = H\left(\mathbb{P}^{\left(X_1^N,\, X_{N+1}^{N+M}\right)}\right) \leqslant H\left(\mathbb{P}^{X_1^N}\right) + H\left(\mathbb{P}^{X_{N+1}^{N+M}}\right) = H(\mathbb{P}_N) + H(\mathbb{P}_M)\,,$$

la dernière égalité utilisant l'hypothèse de stationnarité. La suite de terme général $u_N = H(\mathbb{P}_N)$ est donc sous-additive, et par conséquent, (u_N/N) est convergente. On note H_∞ sa limite, qui est égale à $\inf_k\ u_k/k$. □

Remarque au passage. Rappelons la preuve de ce résultat élémentaire d'analyse. On fixe un entier $n \geqslant 1$ et on considère un entier $N \geqslant n$, que l'on écrit sous la forme $N = qn + r$ où $q \geqslant 1$ et $r \leqslant n-1$ sont des entiers. Par récurrence immédiate, il vient $u_N \leqslant q u_n + u_r$, d'où

$$\frac{u_N}{N} \leqslant \frac{q}{N}\, u_n + \frac{u_r}{N} \leqslant \frac{u_n}{n} + \frac{u_r}{N} \qquad \text{puis} \qquad \limsup_{N\to\infty} \frac{u_N}{N} \leqslant \frac{u_n}{n}\,,$$

et ce, pour tout entier n. Ainsi,

$$\limsup_{N\to\infty} \frac{u_N}{N} \leqslant \inf_{n\geqslant 1} \frac{u_n}{n} \leqslant \liminf_{N\to\infty} \frac{u_N}{N}\,,$$

ce qui montre que

$$\frac{u_N}{N} \underset{N\to\infty}{\longrightarrow} \inf_{n\geqslant 1} \frac{u_n}{n}\,.$$

Exemple 34.4 On considère une chaîne de Markov X sur un espace d'états fini E, de matrice de transition Q irréductible. La chaîne est donc irréductible récurrente positive et admet une unique probabilité stationnaire π. On va voir que pour que le taux d'entropie soit défini, on n'a en fait pas besoin ici de supposer que X est stationnaire, i.e., que X_1 admet pour loi π. Pour tout $N \geqslant 1$, on note en effet p_N la densité de $\mathbb{P}_N$ par rapport à la mesure de comptage sur E^N et on a alors l'écriture

$$\begin{aligned} H(\mathbb{P}_{N+1}) &= -\sum_{x_1^{N+1}\in E^{N+1}} p_{N+1}\big(x_1^{N+1}\big)\log_2 p_{N+1}\big(x_1^{N+1}\big) \\ &= -\sum_{x_1^N\in E^N}\sum_{x_{N+1}\in E} \Big(p_N\big(x_1^N\big)\, Q(x_N, x_{N+1})\Big)\log_2\Big(p_N\big(x_1^N\big)\, Q(x_N, x_{N+1})\Big) \\ &= -\sum_{x_1^N\in E^N} p_N\big(x_1^N\big)\log_2 p_N\big(x_1^N\big) + \sum_{x_1^N\in E^N} p_N\big(x_1^N\big)\, h_Q(x_N) \end{aligned}$$

où l'on a défini pour tout $x \in E$,

$$h_Q(x) = -\sum_{y\in E} Q(x,y)\log_2 Q(x,y)\,.$$

Avec ces notations, on a ainsi prouvé la décomposition

$$H(\mathbb{P}_{N+1}) = H(\mathbb{P}_N) + \mathbb{E}\big[h_Q(X_N)\big] .$$

Par conséquent, pour tout entier $N \geqslant 2$,

$$H(\mathbb{P}_N) = H(\mathbb{P}_1) + \sum_{t=1}^{N-1} \mathbb{E}\big[h_Q(X_t)\big] .$$

La fonction h_Q étant bornée (voir figure 21.1) donc intégrable pour la probabilité stationnaire π, les théorèmes de type ergodique pour les chaînes de Markov (voir par exemple [MPB98, Théorème 4.25]) assurent la convergence

$$\frac{1}{N}\sum_{t=1}^{N} h_Q(X_t) \underset{N\to\infty}{\longrightarrow} \sum_{x\in E} \pi(x) h_Q(x) \qquad \mathbb{P}\text{–p.s.},$$

qui entraîne également une convergence dans $\mathbb{L}^1(\mathbb{P})$ par théorème de convergence dominée. On a ainsi en particulier

$$\frac{H(\mathbb{P}_N)}{N} = \frac{H(\mathbb{P}_1)}{N} + \frac{1}{N}\sum_{t=1}^{N-1} \mathbb{E}\big[h_Q(X_t)\big] \underset{N\to\infty}{\longrightarrow} H_\infty \overset{\text{not.}}{=} \sum_{x\in E} \pi(x) h_Q(x) .$$

En reprenant les calculs précédents, on remarque que la limite ainsi obtenue est une entropie conditionnelle : si $Y = (Y_1, Y_2, \ldots)$ est une chaîne de Markov stationnaire de transition Q et de loi initiale π, alors H_∞ est égale à l'entropie conditionnelle de la loi de (Y_1, Y_2) par rapport à la loi de Y_1.

34.1.2 Codes universels (un aperçu de la recherche en codage)

Au paragraphe 16.1, on avait parlé de deux situations menant à une modélisation stochastique naturelle, le cas du codage de données déjà disponibles et qu'on peut lire une première fois avant de les coder, d'une part, et d'autre part, celui de données émises de manière séquentielle par une source et qu'on doit coder à la volée par blocs de N caractères. Le chapitre 21 s'est essentiellement intéressé au premier cas et nous indiquons maintenant quel est le type de résultats établis pour le second modèle. Pour cela, on suppose avoir affaire à un processus stationnaire $X = (X_1, X_2, \ldots)$, pour lequel on reprend les notations du paragraphe 34.1.1.

Le Théorème 16.5 et la Proposition 34.3 assurent que lorsque $\mathbb{P}_\infty$ et donc tous les $\mathbb{P}_N$ sont connus, il existe une suite de codes $c_N^*(\mathbb{P}_N)$ construits sur les $\mathbb{P}_N$ et tels que les facteurs de compression vérifient

$$\gamma\big(c_N^*(\mathbb{P}_N),\, \mathbb{P}_N\big) \longrightarrow H_\infty \qquad \text{lorsque} \quad N \to \infty ,$$

ce qui est le résultat limite optimal. Ici, pour bien indiquer la dépendance des facteurs de compression en la loi $\mathbb{P}_\infty$, on les a indexés d'une part par le code considéré et d'autre part, par la loi des mots.

La question est de savoir si l'on peut se passer de la connaissance de $\mathbb{P}_\infty$ et faire asymptotiquement aussi bien que précédemment, i.e., exhiber une suite de codes $\widehat{c}_N$ ne requérant aucune connaissance préliminaire et telle que leurs facteurs de compression vérifient que pour toute loi stationnaire $\mathbb{P}_\infty$,

$$\gamma\big(\widehat{c}_N,\, \mathbb{P}_N\big) \longrightarrow H_\infty \qquad \text{lorsque} \quad N \to \infty\,.$$

Au vu des commentaires autour de la remarque 21.9, cela revient à exhiber une suite de probabilités $\widehat{\mathbb{Q}}_N$, chacune définie sur $\mathcal{A}^N$, telle que pour toute loi stationnaire $\mathbb{P}_\infty$,

$$\frac{1}{N}\,\mathcal{K}\Big(\widehat{\mathbb{Q}}_N,\, \mathbb{P}_N\Big) \longrightarrow 0 \qquad \text{lorsque} \quad N \to \infty\,.$$

On appelle code universel un code associé à une telle suite de probabilités.

Le fait remarquable est qu'il existe de tels codes. On peut même les construire explicitement et de manière efficace, par double mélange utilisant des arbres de contexte (voir par exemple [Cat04, chapitre 1] pour un résumé de plusieurs décennies de recherches à ce sujet).

34.2 Codes des simulations

34.2.1 Fonctions de conversion ANSI – chaînes de caractères

```
function A = str2ansi(E)
%
% Entrée :
%   E, une chaîne de caractères de '0' et de '1', dont les éléments sont
%      lus 8 par 8 ; sa longueur doit donc être multiple de 8
% Sortie :
%   A, un vecteur de nombres entre 0 et 255, chacun des nombres correspondant
%      à un groupe de 8 caractères '0' ou '1'

  A = bin2dec(reshape(E, 8, length(E)/8)');

end
```

```
function S = ansi2str(A)
%
% Entrée :
%   A, un vecteur de nombres entre 0 et 255
% Sortie :
%   S, une chaîne de caractères de '0' et de '1', dont les éléments sont
%      lus 8 par 8 ; chaque groupe de 8 correspondont à un élément de A

  x = dec2bin(A, 8)';
```

```
  S = num2str(x(:))';

end
```

34.2.2 Construction du code de Huffman et représentation de l'arbre associé

```
function T = HuffmanTree(M, P)
%
% Entrées :
%   M, objet de type cell, contenant les mots du dictionnaire
%   P, vecteur de nombres, contenant les probabilités *ou* les nombres
%      d'occurrences associés à chacun des mots (P n'a pas besoin d'être
%      renormalisé de sorte que ses composantes se somment à 1)
% Sortie :
%   T, un arbre, i.e., un objet de type cell contenant ou
%      des objets de type cell ou des mots

  [P, I] = sort(P);
  % S'il ne reste plus que deux éléments dans M,
  % alors l'arbre associé est M lui-même.
  if length(M) == 2
    T = M(I);
  % Sinon on utilise la description récursive du code
  % de Huffman
  else
    T = HuffmanTree({{M{I(1)}, M{I(2)}}, M{I(3:end)}}, [P(1)+P(2) P(3:end)]);
  end;

end
```

```
function R = drawTree(T, varargin)
%
% Entrée :
%   T, l'arbre à représenter, ne contenant que des mots d'une et une seule
%      lettre *ou* que des nombres entre 0 et 255 (on représentera alors
%      les caractères ANSI associés)
%   L'objet de type cell nommé varargin contient tous les éventuels
%   arguments optionnels ; c'est ainsi que l'on peut transmettre un
%   nombre variable d'arguments sous Matlab ; seuls les appels récursifs
%   utilisent ces arguments optionnels
% Sortie :
%   R, un vecteur de chaînes de caractères toutes de même longueur

  % Lorsque lancé par l'utilisateur, il n'y a pas encore d'arguments
```

```
  % optionnels, on les détermine en vue des appels récursifs
  if isempty(varargin)
    depth = profondeur(T);
    current = 0;
    varargin = {depth, current};
  end;

  if not(iscell(T))
    d = varargin{1} - varargin{2};
    S = repmat(' ', 1, 2^d-1);
    R = [S char(T) S;
         repmat(' ', d, 2^(d+1)-1)];
  else
    Rgauche = drawTree(T{1}, varargin{1}, varargin{2}+1);
    Rdroite = drawTree(T{2}, varargin{1}, varargin{2}+1);
    d = varargin{1} - varargin{2};
    R = [repmat(' ', 1, 2^d-1) '*' repmat(' ', 1, 2^d-1);
         Rgauche, repmat(' ', d, 1), Rdroite];
  end;

end

function depth = profondeur(T)
% Calcul récursif de la profondeur d'un arbre

  if not(iscell(T))
    depth = 0;
  else
    depth = max(profondeur(T{1}), profondeur(T{2})) + 1;
  end;

end
```

34.2.3 Fonction pour le codage et le décodage d'une chaîne de caractères

```
function [C, D] = convert(T)
%
% Entrée :
%   T, un arbre de mots
% Sorties :
%   C, un objet de type cell contenant la suite des codes
%   D, un objet de type cell contenant la suite des mots

  LR = ['0', '1'];
  C = {};
  D = {};
  for k = 1:2
```

```
    if iscell(T{k})
      [Cchild, Dchild] = convert(T{k});
    else
      Cchild = '';
      Dchild = {T{k}};
    end;
    C = [C strcat(LR(k), Cchild)];
    D = [D Dchild];
  end;

end
```

```
function D = decode(S, T)
%
% Entrées :
%   S, la chaîne des caractères à décoder
%   T, l'arbre de codage (contenant pour mots ou uniquement des chaînes
%      de caractères, ou uniquement des nombres)
% Sortie :
%   D, une chaîne de caractères ou un vecteur de nombres, contenant
%      la suite des mots décodés

  t = T;
  D = [];
  for i = 1:length(S)
    % Lire la chaîne de gauche à droite fait parcourir l'arbre
    t = t{1+str2num(S(i))};
    if not(iscell(t))
      % Chaque fois que l'on atteint une feuille, on note sa valeur
      D = [D t];
      t = T;
    end;
  end;

end
```

34.2.4 Fonction pour encoder et décoder un arbre de code au début d'un fichier

```
function [S, W] = encodeTree(T);
%
% Entrée :
%   T, l'arbre à encoder
% Sortie :
%   S, une chaîne de caractères contenant des 0 et des 1 et
```

```
%      encodant la structure de T
%   W, un objet de type cell dressant la liste des mots présents
%      aux feuilles, de gauche à droite

  if iscell(T)
    [Sleft, Wleft] = encodeTree(T{1});
    [Sright, Wright] = encodeTree(T{2});
    S = ['1' Sleft Sright];
    W = [Wleft Wright];
  else
    S = '0';
    W = { T };
  end;

end



function [T, C, W] = decodeTree(S, M);
%
% Entrées :
%   S, une chaîne de caractères '0' et '1' contenant en préfixe la
%      structure de l'arbre à décoder
%   M, un objet de type cell contenant au moins autant d'éléments que
%      l'arbre a de feuilles ; ses éléments, lus de gauche à droite, seront
%      les feuilles de l'arbre
% Sortie :
%   T, l'arbre correspondant à S et M
%   C, le reste de la chaîne S une fois le préfixe codant l'arbre lu
%   W, les éléments de M qui étaient en trop (dans le cas où M avait
%      plus d'éléments que T n'a de feuilles)

  if S(1) == '0'
    T = M{1};
    W = M(2:end);
    C = S(2:end);
  else
    [Tleft, Cright, Wright] = decodeTree(S(2:end), M);
    [Tright, C, W] = decodeTree(Cright, Wright);
    T = {Tleft, Tright};
  end;

end
```

34.2.5 Script de codage d'un fichier

```
clc, clear;
```

```
% Déterminer le fichier à coder
[fname, pname] = uigetfile('*.*', 'Choisissez le fichier à coder');

% Lire fichier (sous la forme du vecteur contenant
% la suite des caractères ANSI du fichier)
fid = fopen([pname fname]);
ansi = fread(fid, inf);
fclose(fid);

% Construction de la table des fréquences des codes ANSI,
% non renormalisées
freqtable = histc(ansi, -0.5 + 0:256);
% Matlab ajoutant la classe [255.5 -> infini] :
freqtable = freqtable(1:256);
garde = (freqtable >= 1);
freqtable = freqtable(garde);
% Calcul de l'entropie et indication d'un taux de compression prévisionnel
f = freqtable/sum(freqtable);
H = - sum(f .* log(f))/log(2);
disp(['Taux de codage attendu (hors encodage de la table) : entre ', ...
      num2str(100*H/8, 2), ' % et ', num2str(min(100,100*(H+1)/8), 2), ' %']);

% Pour obtenir l'arbre, en plus de ces nombres d'occurrences,
% on doit transmettre un objet de type cell, contenant
% les index des codes ANSI ; ce sont eux qui forment nos mots
X = 0:255;
M = num2cell(X(garde), 1);
T = HuffmanTree(M, freqtable');

% Pour coder (mais pas pour décoder), il faut convertir l'arbre
% en deux objets cell, contenant d'une part, les codes (dans C),
% et d'autre part, les mots correspondants (dans D)
[C, D] = convert(T);
s = cell(1,256);
s(cell2mat(D)+1) = C;
% La chaîne codant le fichier étant donné l'arbre T est alors :
R = strcat(s{ansi+1});

% On code maintenant l'arbre
[S, W] = encodeTree(T);
W = ansi2str([W{:}]);

% On met les codes bout à bout, et on ajoute autant de bits qu'il faut
% (maximum 7) pour que la longueur du tout soit multiple de 8
% On indique par les trois premiers bits combien on en a ajouté à la fin
E = [S W R];
r = mod(length(E) + 3, 8);
```

```
d = (8 - r) * (r > 0);
E = [num2str(dec2bin(d, 3)) E repmat('0',1,d)];

Final = str2ansi(E);

% Déterminer sous quel nom on veut sauvergarder le fichier encodé
[fname, pname] = uiputfile('*.huf', 'Enregistrer le fichier encodé sous...');
fid2 = fopen([pname fname], 'w');
fwrite(fid2, Final);
fclose(fid2);

disp(['Taux de codage final (y compris encodage de la table) : ', ...
    num2str(100*length(Final)/length(ansi),2), ' %']);
```

34.2.6 Script de décodage d'un fichier

```
clc, clear;
% Déterminer le fichier à décoder
[fname, pname] = uigetfile('*.huf', 'Choisissez le fichier à décoder');
fid = fopen([pname fname]);
ansi = fread(fid, inf);
fclose(fid);

% On convertit la suite des caractères ANSI en une chaîne
% de 0 et de 1
S = ansi2str(ansi);
% Les 3 premiers 0 ou 1 nous indiquent combien de caractères jeter à la fin
% de la chaîne
d = bin2dec(S(1:3));
S = S(4:(end-d));

% On parcourt une première fois le début de la chaîne pour
% identifier le préfixe correspondant au code de l'arbre T
% (de longueur n)
[T, C, W] = decodeTree(S, num2cell(0:255));
n = length(S) - length(C);
TreeCode = S(1:n);
% Il y a un mot par feuille, les feuilles sont codées par 0
numberWords = sum(TreeCode == '0');
% Les mots ont été placés à la suite de l'arbre
TreeWords = S((n+1):(n+8*numberWords));
M = str2ansi(TreeWords);
% On peut maintenant décoder l'arbre, puisqu'on sait la vraie étiquette
% de chaque feuille
T = decodeTree(S(1:n), num2cell(M));
disp(['Table lue, format de fichier correct -- ', ...
      'Décodage en cours, veuillez patienter']);
```

```
% Et il suffit de décoder le reste de la chaîne en utilisant l'arbre obtenu
Res = decode(S((n+8*numberWords+1):end), T);

% Déterminer sous quel nom on veut sauvergarder le fichier décodé
[fname, pname] = uiputfile('*.*', 'Enregistrer le fichier décodé sous...');
fid2 = fopen([pname fname], 'w');
fwrite(fid2, Res);
fclose(fid2);
```

CHAPITRE 35

Jeux de grattage (compléments)

Résumé. Nous présentons ici le code des simulations informatiques relatives au chapitre 25.

35.1 Codes des simulations

35.1.1 Fonction auxiliaire de tirage d'un échantillon de tickets

```
function G = Ticket(T, n)
%
% Tirage du lot d'un ticket selon une loi multinomiale
%
% Entrée :
%   T un tableau de lots possibles
%   n taille de l'échantillon désiré
% Sortie :
%   G vecteur de lots de taille n
%

  Z = ceil(rand(1,n) * 500000);
  R = cumsum(T(:,2));
  G = zeros(1,n);
  for k = R'
    G = G + (Z > k);
  end
  J = (G <= 12);
  if any(J)
    G(J) = T(G(J)+1, 1)';
  end;
  G(G == 13) = 0;

end
```

35.1.2 Fichier maître de calculs de différentes valeurs numériques (probabilités et espérances) et de tracés de graphiques

```
T = [40000, 1;
     20000, 1;
     10000, 2;
```

```
      1000, 5;
       500, 18;
       200, 800;
       100, 850;
        50, 2020;
        20, 4000;
        10, 9000;
         6, 28000;
         4, 25000;
         3, 47500];

% Calcul du nombre L (= 7697) de lots significatifs
L = sum(T((T(:,1) >= 20),2));
p = L / 500000;

N = 500000;

% A propos des calculs autour de l'espérance des gains
% Pourcentage de redistribution
r = sum(T(:,1) .* T(:,2))/(N*3);
% Ecart-type du gain d'un ticket
s = sqrt(sum(T(:,1).^2 .* T(:,2))/N - (sum(T(:,1) .* T(:,2))/N)^2);
% Montant moyen et écart-type des gains typiques (i.e., inférieurs
% ou égaux à 200 euros)
J = (T(:,1) <= 200);
NT = N - sum(T(not(J),2));
mT = sum(T(J,1) .* T(J,2))/NT;
sT = sqrt(sum(T(J,1).^2 .* T(J,2))/NT - mT^2);
mA = sum(T(J,1) .* T(J,2))/N;
sA = sqrt(sum(T(J,1).^2 .* T(J,2))/N - mA^2);


% Fonction de répartition du montant du gain maximal
F = (1 - cumsum(T(:,2))/N).^100;
F = [F(end:-1:1); 1]; % sur 1 an
FF = F.^10; % sur 10 ans
% Fonctions de 'survie'
R = [1; 1-F(1:(end-1))]; % sur 1 an
RR = [1; 1-FF(1:(end-1))]; % sur 10 ans
RRR = [R RR];
for k = 1:2
  subplot(1,2,k)
  stem(RRR(:,k));
  % On effectue une nouvelle graduation de l'axe des abscisses, passant de
  % 0, 1, 2, 3, 4 ... à 0, 3, 4, 6, 10, 20, 50 ... (i.e., les gains
  % potentiels)
  set(gca, 'XTick', [1:length(RRR(:,k))])
```

```
  set(gca, 'XTickLabel', [0; T(end:-1:1, 1)]);
  % Une version modifiée de xticklabel_rotate90 (de Denis Gilbert, Ph.D.)
  % permet de mieux présenter la graduation des abscisses
  % en la tournant de 90 degrés
  % xticklabel_rotate90(get(gca,'XTick'), get(gca,'XTickLabel'));
  xlabel('Gain maximal');
  ylabel('Probabilité');
end;

% Calcul des probabilités qu'un joueur grattant deux tickets par
% semaine gagne au moins une fois une certaine somme
G = [F(1); F(2:end) - F(1:(end-1))];
GG = [FF(1); FF(2:end) - FF(1:(end-1))];
disp([ G GG ]);

% Tracés des six graphiques de soldes totaux des gains et mises
% de joueurs individuels, en fonction du temps
m = 400;
for k = 1:6
  subplot(2,3,k);
  hold on;
  plot(cumsum(Ticket(T, m)) - 3*(1:m), 'color', 'k');
  line([0 400], [0 0], 'color', 'k');
end

% Calculs relatifs à l'existence de stratégies gagnantes
% Montant moyen d'un lot significatif typique
JM = (T(:,1) <= 200) & (T(:,1) >= 20);
NM = sum(T(JM,2));
M = sum(T(JM,1) .* T(JM,2))/NM;
% Montant moyen gagné par un ticket perdant ou comportant
% une misère
Jm = (T(:,1) <= 10);
Nm = N - sum(T(:,2)) + sum(T(Jm,2));
m = sum(T(Jm,1) .* T(Jm,2))/Nm;
% Tentative de modélisation proposée dans le corrigé
a = 0.256; b = 0.718; q = 0.026;
e = []; r = [];
% Calcul de l'espérance de gain de la stratégie consistant
% en l'achat des x derniers tickets d'un carnet où aucun lot
% significatif n'est encore sorti
% Pour x >= 2
for x = (2:50)
  r = [r, a + b*x/50 + q*nchoosek(x,2)/nchoosek(50,2)];
  e = [e, (a*x*m + b*(x/50)*((x-1)*m+M) + ...
           q*(nchoosek(x,2)/nchoosek(50,2))*((x-2)*m+2*M))/r(end)];
end
```

```
% Pour x = 1
r = [a + b/50, r];
e = [(a*m + (b/50)*M)/(a + b/50), e];
% Et on enlève les mises (prix des tickets)
e = e - 3*(1:50);
subplot(1,2,1)
stem(e);
subplot(1,2,2)
stem(e(1:14));
% Si l'on tient compte de la répétition
gr = r.*e;
I = (gr > 0);
ind = 1:50;
disp(ind(I));
disp(gr(I));
```