Boîte à outils de probabilités pour le cours d'apprentissage

Gilles Stoltz

23 février / 1er mars

Prérequis : Le cours de mesure, intégration et probabilités du premier semestre et des notions sur l'espérance conditionnelle ; mais pas besoin de suivre le cours de processus stochastiques ! C'est pourquoi, à part deux exercices facultatifs, il ne sera pas question de filtrations, de martingales, etc.

Note : Je tâche dans ce qui suit de ne pas faire de liens avec le cours de statistiques, puisque beaucoup d'entre vous ne l'ont pas encore suivi. Je reste ainsi le plus piéton possible, sans essayer un seul instant d'œuvrer à une synthèse, ce qui ne me manquera pas de subir les foudres de collègues qui tomberaient sur ces notes.

1. Contrôle en $\sqrt{n}$ de sommes de n variables aléatoires indépendantes

Lemme 1 (Lemme de Hoeffding)**.** *Soit H une variable aléatoire à valeurs dans $[a,b] \subset \mathbb{R}$. Alors pour tout $s \in \mathbb{R}$,*

$$\ln \mathbb{E}\big[e^{sH}\big] \leqslant s\,\mathbb{E}[H] + \frac{s^2}{8}(b-a)^2 \,.$$

Jeu-concours : Que donnerait l'inégalité de Jensen pour contrôler $\ln \mathbb{E}\big[e^{sH}\big]$?

Démonstration. On note $\psi_H(s) = \ln \mathbb{E}\big[e^{sH}\big]$, on dérive deux fois ψ_H, on tombe sur une variance, qu'on contrôle par $(b-a)^2/4$, et on intègre deux fois. *Jeu-concours* : Trouver ensemble le contrôle de la variance, en re-prouvant ensemble, si besoin, la formule variationnelle de la variance. □

Théorème 1 (Inégalité de Hoeffding, version probabiliste)**.** *Soit $H_1, \ldots, H_n$ des variables aléatoires indépendantes, chacune à valeurs dans $[a_t, b_t] \subset \mathbb{R}$. Alors pour tout $\varepsilon > 0$,*

$$\mathbb{P}\left\{ \sum_{t=1}^n H_t \geqslant \sum_{t=1}^n \mathbb{E}[H_t] + \varepsilon \right\} \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{t=1}^n (b_t - a_t)^2} \right).$$

Démonstration. On introduit

$$S_n = \sum_{t=1}^n \big(H_t - \mathbb{E}[H_t] \big);$$

par récurrence et lemme de Hoeffding, on contrôle $\mathbb{E}\big[e^{sS_n}\big]$ pour tout $s \in R$, puis on utilise une majoration dite de Chernoff, pour $s > 0$:

$$\mathbb{P}\{S_n \geqslant \varepsilon\} = \mathbb{P}\big\{e^{sS_n} \geqslant e^{s\varepsilon}\big\} \leqslant \frac{\mathbb{E}\big[e^{sS_n}\big]}{e^{s\varepsilon}} \,.$$

On ré-injecte la borne et on optimise en s. □

Corollaire 2 (Inégalité de Hoeffding, version statistique)**.** *Soit $H_1, \ldots, H_n$ des variables aléatoires indépendantes, chacune à valeurs dans $[a_t, b_t] \subset \mathbb{R}$. Alors pour tout $\delta \in (0,1)$, avec probabilité au moins $1-\delta$,*

$$\sum_{t=1}^n H_t < \sum_{t=1}^n \mathbb{E}[H_t] + \sqrt{\frac{1}{2} \sum_{t=1}^n (b_t - a_t)^2 \ln \frac{1}{\delta}} \,.$$

Le théorème de la limite centrale donne un contrôle asymptotique des déviations autour de la moyenne de l'ordre de $\sigma/\sqrt{n}$, pour des variables i.i.d. (cf. exercice 2) ; on précise ici des contrôles non asymptotiques de l'ordre de $\Box/\sqrt{n}$, où $\Box$ est une quantité hélas plus grande que l'écart-type... En effet :

Exemple 1. Une application typique de l'inégalité de Hoeffding est lorsque les H_t sont indépendantes et identiquement distribuées, d'étendue commune $[a,b]$ et d'espérance commune μ, auquel cas il vient qu'avec probabilité au moins $1-\delta$,

$$\frac{1}{n}\sum_{t=1}^{n} H_t < \mu + (b-a)\sqrt{\frac{1}{2n}\ln\frac{1}{\delta}} = \underbrace{\frac{(b-a)}{2}}_{\text{cf. écart-type}} \underbrace{\sqrt{2\ln\frac{1}{\delta}}}_{\text{cf. quantile}} \frac{1}{\sqrt{n}}\,.$$

Exercices simples

Exercice 1. Par symétrie et borne d'union, on peut bien sûr contrôler à la fois les déviations par le haut et par le bas : écrire ce que l'on obtient.

Exercice 2. Ecrire que ce donne le théorème de la limite centrale dans le cas de l'exemple 1 ; comparer à ce qu'on y a obtenu. Quels sont les avantages et inconvénients de chacun des deux résultats (théorème de la limite centrale contre inégalité de Hoeffding) ?

A coder 1. Illustrer les bornes précédentes sur données simulées (par exemple sur des lois de Bernoulli ou des lois uniformes) : qu'est-il plus simple de considérer, la version probabiliste ou la version statistique ?

Exercices moins simples, voire difficiles

Exercice 3. Supposons que l'on ait une suite de variables aléatoires H_1, H_2, ... indépendantes et identiquement distribuées selon une loi à support compact, disons $[a,b]$. On note $\overline{H}_n$ la suite des moyennes des n premiers termes. Appliquer le lemme de Borel-Cantellli pour obtenir un résultat asymptotique :

$$\limsup_{n\to\infty} \sqrt{n}\,\frac{\overline{H}_n-\mu}{(b-a)\sqrt{\ln n}} \leqslant 1 \quad \text{p.s.}$$

Selon vos connaissances en théorie des probabilités : Comparer le résultat obtenu à la loi du logarithme itéré et constater qu'une amélioration est nécessaire.

L'exercice qui suit, ainsi que les liens à la loi du logarithme itéré dans l'exercice précédent, ne seront bien sûr pas vus en cours...

Exercice 4 (Facultatif). *Selon l'avancement du cours de processus stochastiques* : S_n est une martingale, donc e^{sS_n} est une sous-martingale positive. Utiliser l'inégalité de Doob pour obtenir une version dite maximale de l'inégalité de Hoeffding.
Question supplémentaire difficile : Utiliser cette version maximale avec les méthodes de l'exercice précédent afin d'obtenir cette fois-ci un résultat comparable à la loi du logarithme itéré.

2. Contrôle en $\sqrt{n}$ de sommes de n variables aléatoires hélas non indépendantes

Voici un renforcement "automatique" du lemme de Hoeffding.

Lemme 2 (Lemme de Hoeffding conditionnel). *Soit H une variable aléatoire à valeurs dans $[a,b]\subset\mathbb{R}$. Alors pour toute sous-tribu $\mathcal{F}$ des boréliens, pour tout $s\in\mathbb{R}$,*

$$\ln\mathbb{E}\big[e^{sH}\,\big|\,\mathcal{F}\big] \leqslant s\,\mathbb{E}\big[H\,\big|\,\mathcal{F}\big] + \frac{s^2}{8}(b-a)^2\,.$$

Jeu-concours : Rappeler la caractérisation de l'espérance conditionnelle à $\mathcal{F}$. A savoir, que pour tout $A\in\mathcal{F}$ tel que $\mathbb{P}(A)>0$, et en notant $\mathbb{E}_A$ l'espérance conditionnelle à A,

$$\mathbb{E}_A[H] = \mathbb{E}_A\big[\mathbb{E}[H\,|\,\mathcal{F}]\big]\,.$$

Démonstration. On introduit l'événement $A\in\mathcal{F}$ suivant :

$$A = \Big\{\mathbb{E}\big[e^{sH}\,\big|\,\mathcal{F}\big] > e^{s\,\mathbb{E}[H\,|\,\mathcal{F}]}\,e^{s^2(b-a)^2/8}\Big\}\,;$$

si A n'était pas de probabilité nulle, on pourrait écrire le lemme de Hoeffding avec H et $\mathbb{E}_A$ et on obtiendrait une contradiction avec la définition de A (nécessite une application de l'inégalité de Jensen au passage). □

Théorème 3 (Inégalité de Hoeffding–Azuma, version probabiliste)**.** *Soit* $\mathcal{F}_0 = \{\mathbb{R}, \emptyset\}$ *et soit* $\mathcal{F}_0 \subset \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}_n$ *des sous-tribus boréliennes. Soit alors* $H_1, \ldots, H_n$ *des variables aléatoires, chacune à valeurs dans* $[a_t, b_t] \subset \mathbb{R}$ *et telle que* H_t *soit* $\mathcal{F}_t$*–mesurable. Alors pour tout* $\varepsilon > 0$,

$$\mathbb{P}\left\{ \sum_{t=1}^{n} H_t \geqslant \sum_{t=1}^{n} \mathbb{E}\big[H_t \,\big|\, \mathcal{F}_{t-1}\big] + \varepsilon \right\} \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{t=1}^{n}(b_t - a_t)^2} \right).$$

Démonstration. Très similaire à celle de l'inégalité de Hoeffding ; on utilise seulement comme ingrédient supplémentaire que l'espérance de l'espérance conditionnelle est l'espérance tout court. □

Exercice

Exercice 5 (Facultatif). *Selon l'avancement du cours de processus stochastiques* : On appelle la suite des $\Delta_t = H_t - \mathbb{E}\big[H_t \,\big|\, \mathcal{F}_{t-1}\big]$ une suite d'accroissements de martingale. En effet, $M_n = \Delta_1 + \ldots + \Delta_n$ est une martingale. Réciproquement, toute martingale s'écrit comme la somme de tels accroissements.

3. Aparté : liens entre les sommes de variables aléatoires et la loi normale

Si $N \sim \mathcal{N}\big(0, \sigma^2\big)$, alors pour tout $s \in \mathbb{R}$,

$$\mathbb{E}\big[e^{sN}\big] = e^{s^2\sigma^2/2}.$$

Cette observation motive la définition suivante.

Définition 1. Une variable aléatoire S telle que pour tout $\lambda > 0$,

$$\mathbb{E}\big[e^{\lambda S}\big] \leqslant e^{\lambda^2 v/2}$$

est dite sous-gaussienne de paramètre v.

La preuve de l'inégalité de Hoeffding montre le fait suivant, qui en était la clé. On notera que le paramètre obtenu est exactement ce par quoi on avait majoré les variances dans le lemme de Hoeffding.

Théorème 4 (Inégalité de Hoeffding, cœur de l'argument)**.** *Soit* $H_1, \ldots, H_n$ *des variables aléatoires indépendantes, chacune à valeurs dans* $[a_t, b_t] \subset \mathbb{R}$*. Alors la somme*

$$S_n = \sum_{t=1}^{n} \big(H_t - \mathbb{E}[H_t]\big)$$

est une variable aléatoire sous-gaussienne de paramètre $\sum_{t=1}^{n}(b_t - a_t)^2/4$.

Or, on a le résultat suivant, qui procède d'une majoration de Chernoff, et cela nous ramène à la version originelle de l'inégalité de Hoeffding.

Corollaire 5. *Si* S *est une variable aléatoire sous-gaussienne de paramètre* v*, alors pour tout* $\varepsilon > 0$,

$$\mathbb{P}\{S \geqslant \varepsilon\} \leqslant e^{-\varepsilon^2/(2v)}.$$

Exercices

Exercice 6. Prouver le Corollaire 5.

Exercice 7. Montrer par un calcul direct simple (majoration de $-(t+x)^2$ par $-x^2 - t^2$) que lorsque $N \sim \mathcal{N}(0, 1)$, on a

$$\mathbb{P}\{N \geqslant x\} \leqslant \frac{e^{-x^2/2}}{2}.$$

En fait, on a même mieux pour les x grands en tenant compte du double produit et en négligeant cette fois t^2 : effectuer le calcul.
Enfin, quel contrôle obtient-on en revanche en appliquant une majoration de Chernoff, comme au Corollaire 5 ?

4. Contrôle de maxima de variables aléatoires sous-gaussiennes

Pour simplifier le propos, on considère dans cet ensemble de paragraphes des variables aléatoires sous-gaussiennes $N_1, \ldots, N_K$ toutes de même paramètre v (quitte à prendre le maximum des paramètres individuels). *Elles ne sont pas supposées indépendantes.* On s'intéresse alors au comportement du maximum de ces variables,

$$M = \max_{j=1,\ldots,K} N_j .$$

Remarque 1. Le comportement de juste une variable N_1 est connue : l'inégalité de déviation est donnée par le Corollaire 5, et pour l'espérance, on se doute et on peut montrer, cf. Proposition 3 avec $K = 1$, que les variables sous-gaussiennes sont nécessairement d'espérance négative ou nulle. (Si la condition de sous-normalité pour N_1 est vérifiée pour tout $\lambda \in \mathbb{R}$, on a donc que cette espérance est nulle.)

4.1. Bornes de déviation, par union d'événements

Cette borne procède simplement du fait que la probabilité d'une union d'événements est plus petite que la somme de leurs probabilités. Ainsi, par le Corollaire 5,

$$\begin{aligned}\mathbb{P}\left\{\max_{j=1,\ldots,K} N_j \geqslant \varepsilon\right\} &\leqslant \mathbb{P}\big\{\exists\, j \in \{1,\ldots,K\}, \quad N_j \geqslant \varepsilon\big\} \\ &\leqslant \sum_{j=1}^{K} \mathbb{P}\{N_j \geqslant \varepsilon\} \leqslant K\, e^{-\varepsilon^2/(2v)} . \qquad (1)\end{aligned}$$

En intégrant cette borne en $\varepsilon > 0$, on pourrait (cf. exercices à la fin de ce chapitre) en déduire une borne sur l'espérance de M_+, la partie positive du maximum, et donc sur l'espérance de M elle-même. Mais il y a mieux à faire ! Revenons un cran en arrière, à la définition du caractère sous-gaussien.

4.2. Contrôle en espérance

Proposition 3. *Si $N_1, \ldots, N_K$ sont des variables aléatoires sous-gaussiennes toutes de même paramètre v, non nécessairement indépendantes, alors*

$$\mathbb{E}\left[\max_{j=1,\ldots,K} N_j\right] \leqslant \sqrt{2v \ln K} .$$

La preuve est très amusante, parce qu'elle semble à première vue complètement grossière. Or, la borne est optimale, comme le montre le théorème énoncé dans le paragraphe suivant.

Démonstration. Par l'inégalité de Jensen puis par un argument dit à la Pisier (le maximum $e^{\lambda M}$ de termes positifs est plus petit que... leur somme $e^{\lambda N_1} + \ldots + e^{\lambda N_K}$), on a

$$e^{\lambda \mathbb{E}[M]} \leqslant \mathbb{E}\big[e^{\lambda M}\big] = \mathbb{E}\left[\max_{j=1,\ldots,K} e^{\lambda N_j}\right] \leqslant \sum_{j=1}^{K} \mathbb{E}\big[e^{\lambda N_j}\big] \leqslant K\, e^{\lambda^2 v/2} .$$

Il suffit ensuite de passer au logarithme, de diviser les deux membres par λ et d'optimiser celui de droite en la variable d'analyse $\lambda > 0$. $\square$

Exercices : Optimalité du contrôle en espérance

A coder 2. Illustrer la borne de la Proposition 3 sur données simulées (on pourra partir de lois bornées, d'une part, de lois normales, d'autre part).

Théorème 6. *Si $N_1, \ldots, N_K$ sont des variables aléatoires indépendantes et identiquement distribuées selon la loi $\mathcal{N}(0,1)$, qui sont donc sous-gaussiennes de paramètre 1, alors on a l'équivalent, lorsque $K \to \infty$,*

$$\mathbb{E}\left[\max_{j=1,\ldots,K} N_j\right] \sim \sqrt{2 \ln K} .$$

Exercice 8. Prouver le théorème ci-dessus. Un schéma de preuve, avec trous à remplir, sera proposé. (La preuve utilise un argument de simulation de variables aléatoires : on calcule l'inverse généralisée du maximum, qui admet une forme simple, et on écrit que l'espérance du maximum est l'espérance de l'inverse généralisée prise en une variable aléatoire U uniforme sur $[0,1]$.)

5. Obtention d'un contrôle sur l'espérance à partir de celui des déviations

On suppose qu'on étudie une variable aléatoire S pour qui on a réussi à écrire une inégalité de déviations [1] : pour tout $\varepsilon > 0$,

$$\mathbb{P}\{S \geqslant \varepsilon\} \leqslant f(\varepsilon)\,,$$

où f est une fonction décroissante, bien sûr. Que peut-on dire alors de $\mathbb{E}[S]$, en fonction de f ?

5.1. Méthode rapide dans le cas d'une variable aléatoire bornée

Le manière la plus simple et rapide d'avoir une borne (mais rarement celle qui est optimale) est de procéder ainsi, lorsque S est bornée (par une quantité notée $\|S\|_\infty$) et que f est continue et strictement décroissante, donc injective (ce qui est souvent le cas).

La version statistique de la borne de déviations est alors qu'avec probabilité au moins $1-\delta$, on a $S < f^{-1}(\delta)$. Sur l'ensemble complémentaire où $S > f^{-1}(\delta)$, on majore S par $\|S\|_\infty$. Au final, en notant de plus que δ est un paramètre d'analyse, il vient

$$\mathbb{E}[S] \leqslant \inf_{\delta \in (0,1)} \Big\{ (1-\delta)\, f^{-1}(\delta) + \delta\, \|S\|_\infty \Big\}\,.$$

Souvent, on choisit la valeur particulière $\delta = 1/n$; typiquement, le coût de cette méthode directe et brutale par rapport à des méthodes plus sophistiquées est un facteur $\ln n$ (voir l'exercice ci-dessous).

5.2. Méthode plus efficace : utilisation du théorème de Fubini–Tonelli

On emploie ici le résultat suivant, qui découle du théorème de Fubini–Tonelli. Si T est une variable aléatoire *positive*, alors

$$\mathbb{E}[T] = \int_0^\infty \mathbb{P}\{T \geqslant \varepsilon\}\, \mathrm{d}\varepsilon\,. \tag{2}$$

Ici, S n'est pas nécessairement à valeurs positives, mais on majore $\mathbb{E}[S]$ par $\mathbb{E}\big[S_+\big]$, où $S_+ = \max\{S, 0\}$ est la partie positive de S. Or, S_+ vérifie la même inégalité de déviation que S, puisque cette inégalité est formulée uniquement pour des $\varepsilon > 0$. Ainsi,

$$\mathbb{E}[S] \leqslant \mathbb{E}\big[S_+\big] = \int_0^\infty \mathbb{P}\{S \geqslant \varepsilon\}\, \mathrm{d}\varepsilon \leqslant \int_0^\infty f(\varepsilon)\, \mathrm{d}\varepsilon\,.$$

En fait, pour de petites valeurs de ε, disons, pour $\varepsilon \leqslant \varepsilon_0$, il peut arriver que $f(\varepsilon)$ soit plus grand que 1. On gagne alors à remplacer $f(\varepsilon)$ par 1 et on obtient le contrôle

$$\mathbb{E}[S] \leqslant \varepsilon_0 + \int_{\varepsilon_0}^\infty f(\varepsilon)\, \mathrm{d}\varepsilon\,.$$

Exercices

Exercice 9. Prouver (2).

Exercice 10. Qu'obtient-on comme contrôle sur l'espérance à partir de (1) en employant la méthode fondée sur Fubini–Tonelli ? Comparer au résultat optimal procuré par la Proposition 3 et voir ce que l'on perd : (au plus) un facteur additif de $2v$. D'après vous, quel endroit dans les majorations ci-dessus est la cause de ce terme supplémentaire ?

Exercice 11. Qu'obtient-on comme contrôle sur l'espérance à partir de (1) en employant la méthode rapide et directe ? (On pensera que les N_j sont issues de l'inégalité de Hoeffding et sont données par des moyennes de n termes.) Comparer au résultat optimal procuré par la Proposition 3 et voir ce que l'on perd : le terme $\sqrt{2v \ln K}$ est fortement détérioré car remplacé par un terme de l'ordre de $\sqrt{v \ln n}$. Conclusion : cette méthode rapide est bonne seulement en toute première approximation, avant de se lancer dans un calcul un peu plus long du genre de celui mené à l'exercice précédent.

1. ... mais pour qui on n'aurait pas vu, au cours de l'obtention de cette inégalité de déviations, comme aux paragraphes précédents, de moyen simple d'avoir une borne sur son espérance