« Ce que les statisticiens rêvent d'avoir dans un théorème de concentration »

(pour des sommes de variables aléatoires)

Gilles Stoltz
Édité par Yann Ollivier

1er juin 2004

Le but de ce texte est de présenter des résultats de concentration pour des sommes de variables aléatoires. Le cas le plus connu est bien sûr celui de variables indépendantes (cf. l'exposé d'introduction) qui, si ces variables sont uniformément bornées (mettons par 1), fournit de la concentration gaussienne de diamètre observable $\sqrt{n}$ pour la somme de ces variables.

On se propose ici de montrer des techniques permettant d'obtenir deux généralisations. D'une part, on veut relâcher l'hypothèse d'indépendance : au lieu de supposer que la variable X_n est indépendante des variables précédentes $X_1, \ldots, X_{n-1}$, on supposera simplement que son espérance l'est, c'est-à-dire que $\mathbb{E}(X_n | X_1, \ldots, X_{n-1}) = \mathbb{E}X_n$. L'autre généralisation consiste à remplacer le diamètre observable $\sqrt{n}$ par l'information plus précise $\sqrt{\sum \operatorname{Var} X_i}$.

La bonne hypothèse pour la première généralisation consiste à supposer qu'on a affaire à des accroissements de martingales, que nous définissons maintenant.

On donnera ensuite un exemple d'application de ces techniques à une situation « concrète » : la prévision de suites individuelles. Il s'agit d'un jeu entre un statisticien et le diable ; le statisticien essaie de prédire une suite chronologique (par exemple le temps qu'il fera demain) en utilisant des prédictions fournies par des experts. Le diable fixe la réalisation suivante de la variable (en connaissant les prédictions passées du statisticien et des experts mais pas la prédiction courante). Le but du statisticien, opposé à celui du diable, est de minimiser la somme des erreurs des prédictions. On peut montrer, en utilisant les méthodes d'accroissements martingales, que le statisticien peut faire asymptotiquement aussi bien que le meilleur expert (bien sûr, on ne sait pas au départ quel est le meilleur expert).

1 Accroissements de martingales

Définition 1.
Soit une filtration $\mathcal{F}_0 = \{\varnothing, \Omega\} \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_N \subset \ldots \subset \mathcal{F}$ *d'un espace probabilisé* $(\Omega, \mathcal{F}, \Pr)$. *Une suite de variables aléatoires* $(Y_t)_{t\in\mathbb{N}}$ *est une suite d'*accroissements de martingales *si* Y_k *est* $\mathcal{F}_k$*-mesurable et si en outre*

$$\mathbb{E}(Y_k|\mathcal{F}_{k-1}) = 0$$

Ceci équivaut bien sûr au fait que la suite de variables $X_n = \sum^n Y_k$ est une $\mathcal{F}_n$-martingale. Noter la renormalisation à 0 de l'espérance par rapport à l'hypothèse mentionnée en introduction.

Une hypothèse classique de borne sur les variations de chacun des Y_k permet alors d'obtenir de la concentration : (REF ? ? ? Hoeffding 63, Azuma 67)

Proposition 2.
Soit (Y_t) *une suite d'accroissements de martingales. Supposons de plus que pour tout* k*, il existe une variable aléatoire* a_k $\mathcal{F}_{k-1}$*-mesurable ainsi qu'une constante* $c_k \geqslant 0$ *telles que*

$$a_k \leqslant Y_k \leqslant a_k + c_k$$

presque sûrement. Alors, pour tout $t \geqslant 0$ *on a*

$$\Pr(\textstyle\sum Y_k \geqslant t) \leqslant \exp \dfrac{-2t^2}{\sum c_k^2}$$

L'important est bien sûr qu'on n'a pas fait d'hypothèse d'indépendance. L'hypothèse de variations bornées est relativement souple : en effet la borne peut, au travers de a_k, dépendre du passé.

Pour une suite (X_k) de variables aléatoires non centrées, par soustraction de $E(X_k|\mathcal{F}_{k-1})$, on obtient que si $a'_k \leqslant X_k \leqslant a'_k + c_k$ pour une certaine variable aléatoire a'_k $\mathcal{F}_k$-mesurable, alors la somme des X_k se concentre, non pas autour de 0, mais autour de

$$\sum \mathbb{E}(X_k|\mathcal{F}_{k-1})$$

c'est-à-dire qu'une bonne prédiction consiste simplement à ajouter, à chaque instant, l'espérance de la variable connaissant le passé. Cette quantité est souvent calculable dans des situations pratiques (le cas d'indépendance étant bien sûr le plus simple).

Démonstration.
La preuve de cette proposition utilise l'outil classique de l'évaluation de la

transformée de Laplace [ou "borne à la Chernoff] $\mathbb{E}\exp\lambda\sum Y_i$. En intégrant d'abord sur la dernière coordonnée, on a

$$\mathbb{E}e^{\lambda\sum^k Y_i} = \mathbb{E}\left(e^{\lambda\sum^{k-1} Y_i}\,\mathbb{E}(e^{\lambda Y_k}|\mathcal{F}_{k-1})\right)$$

On utilise alors la propriété suivante (qui se démontre élémentairement de manière identique à l'inégalité de Hoeffding dont elle n'est qu'une version conditionnée, par une étude de fonction en λ) : si X est une variable aléatoire et $\mathcal{G}$ une tribu telles que $\mathbb{E}(X|\mathcal{G}) = 0$, et telles que $a \leqslant X \leqslant b$ pour des variables a et b $\mathcal{G}$-mesurables, alors pour tout λ on a

$$\mathbb{E}\left(e^{\lambda X}|\mathcal{G}\right) \leqslant e^{\lambda^2(b-a)^2/8}$$

et donc dans notre cas on a $\mathbb{E}(e^{\lambda Y_k}|\mathcal{F}_{k-1}) \leqslant e^{\lambda^2 c_k^2/8}$.

De proche en proche on obtient

$$\mathbb{E}\exp\lambda\sum^k Y_i \leqslant e^{\lambda^2\sum^k c_i^2/8}$$

et on conclut de la manière habituelle dans la méthode de Laplace, par $\Pr(\sum Y_i \geqslant t) \leqslant e^{-\lambda t}\mathbb{E}e^{\lambda\sum Y_i}$ pour le bon choix de λ soit $\lambda = 4t/\sum c_i^2$. □

On voit dans cette démonstration que l'hypothèse qui fait fonctionner la méthode de Laplace est naturellement une hypothèse d'accroissements de martingales.

On obtient comme corollaire un théorème énoncé d'abord par Talagrand dans le cas du cube discret, et amélioré (indépendamment ?) par McDiarmid, qui énonce de la concentration pour les fonctions définies sur les espaces produits, même lorsque ces fonctions ne sont pas du tout des sommes de variables indépendantes. On dit qu'une fonction de n variables est c-lipschitzienne en la k-ième variable si pour tous $x_1, \ldots, x_n$ et pour tout x'_k, on a

$$\left|f(x_1, \ldots, x_k, \ldots, x_n) - f(x_1, \ldots, x'_k, \ldots, x_n)\right| \leqslant c$$

Corollaire 3.
Soit $\Omega_1, \ldots, \Omega_n$ des ensembles probabilisés et soit $f : \prod\Omega_i \to \mathbb{R}$ une fonction sur le produit de ces espaces (muni de la probabilité produit). Supposons que f est c_k-lipschitzienne en la k-ième coordonnée, pour tout $1 \leqslant k \leqslant n$. Alors

$$\Pr\left(f - \mathbb{E}f \geqslant t\right) \leqslant \exp\frac{-2t^2}{\sum c_k^2}$$

Ce corollaire se démontre très simplement en projetant f sur la filtration engendrée par les produits partiels $\Omega_1, \Omega_1 \times \Omega_2, \ldots$ Par construction, les différences des projections successives formeront une suite d'accroissements de martingales aux variations bornées par les c_k.

2 Application : les suites individuelles

Supposons maintenant qu'on observe séquentiellement des variables $y_1, \ldots, y_k, \ldots$ à valeurs dans un espace $\mathcal{Y}$. L'objectif est, à chaque instant, de faire une prédiction I_k pour la valeur de y_k, connaissant les valeurs précédentes $y_1, \ldots, y_{k-1}$. Plus souvent, on s'intéressera en fait seulement à certains aspects de y_k (par exemple, une décision binaire à prendre) : on peut poser que I_k appartient à un espace $\mathcal{X}$, et qu'on a une fonction $\ell : \mathcal{X} \times \mathcal{Y} \to [0; 1]$ à minimiser, qui décrit le coût subi lorsqu'on fait une prédiction erronée.

On suppose qu'on a à notre disposition N experts fournissant chacun une prédiction ; soit $f_{n,k}(y_1, \ldots, y_{k-1})$ la prédiction de l'expert n pour y_k connaissant les valeurs précédemment observées $y_1, \ldots, y_{k-1}$. On ne sait pas, a priori, si les experts sont plutôt bons ou mauvais.

On cherche à minimiser la fonction ℓ dans le pire des cas : on joue contre la nature, qui à chaque instant peut choisir la valeur de y_k en fonction de tout le passé, et aussi en fonction des prédictions des experts et de nos propres prédictions passées $I_1, \ldots, I_{k-1}$ (ce qui est indispensable par exemple pour des cours de bourse et dans toute situation où notre comportement influence ce qui se passe). Noter qu'en particulier, il n'y a pas de modèle aléatoire implicite pour la manière dont la nature engendre les variables observées : y_k est une fonction mesurable arbitraire du passé y compris de $I_1, \ldots, I_{k-1}$.

Cela revient à dire que l'on joue contre le diable : on cherche une stratégie valable quel que soit le choix des y_i par la nature. Intuitivement, dans un tel contexte, il est clair qu'on a intérêt à randomiser les stratégies, car par exemple suivre toujours un des experts conduit le diable à choisir systématiquement le contraire.

Dans ce contexte, le but consiste à faire asymptotiquement aussi bien que le meilleur expert (lequel n'est pas connu à l'avance). L'objectif consistera à obtenir une inégalité sur la perte cumulée au temps t (qui est a priori bornée par t), par exemple

$$\sum_1^t \ell(I_k, y_k) - \inf_j \sum_1^t \ell(f_{j,k}, y_k) \leqslant o(t)$$

presque sûrement.

Une bonne stratégie consiste, à chaque instant, à tirer au hasard un expert $J_t \in 1, \ldots, N$ suivant une loi p_t sur l'ensemble des experts, puis à utiliser la prédiction de cet expert : $I_t := f_{J_t,t}$. (Dans certains cas où on dispose de convexité, on peut avoir intérêt à prendre aussi des combinaisons linéaires des experts.) On prendra

$$p_t(j) := \exp(-\eta_t L_{j,t-1})/Z$$

où Z est la constante de normalisation, avec

$$L_{j,t-1} := \sum_{1}^{t-1} \ell(f_{j,k}, y_k)$$

l'erreur passée totale de l'expert j, et

$$\eta_t := \sqrt{\frac{8 \ln N}{t}}$$

où N est le nombre d'experts (ici il est important de noter qu'on a supposé $0 \leqslant \ell \leqslant 1$, sinon mettre des constantes numériques en facteur de L_t n'a pas grand sens ; en particulier, si la borne supérieure pour ℓ est mal évaluée, le η_t obtenu sera moins performant).

On prouve alors assez facilement la proposition suivante.

Proposition 4.
Avec la stratégie précédente, pour tout choix de $y_1, \ldots, y_t, \ldots$, pour tout t on a

$$\sum_{1}^{t} \mathbb{E}[\ell(I_k, y_k) \,|\, I_1, \ldots, I_{t-1}] - \min_j \sum_{1}^{t} \ell(f_{j,k}, y_k) \leqslant \sqrt{\frac{t}{2} \ln N}$$

Remarquons que les seules variables aléatoires ici sont les I_k, et donc un choix de la nature est, pour chaque t, une fonction y_t qui soit $(I_1, \ldots, I_{t-1})$-mesurable.

On peut ensuite essayer de passer de l'évaluation "en moyenne" ci-dessus à une évaluation presque sûre. En effet, d'après les résultats de la première partie, on a concentration gaussienne de $\sum \ell(I_k, y_k)$ autour de $\sum \mathbb{E}[\ell(I_k, y_k) \,|\, I_1, \ldots, I_{t-1}]$; c'est le cas puisque ℓ est bornée. On montre ainsi qu'avec probabilité $\geqslant 1 - \delta_t$, on a $\sum_k \ell(I_k, y_k) - \min_j L_{j,t} \leqslant \gamma \sqrt{t \ln \frac{N}{\delta_t}}$ pour une certaine constante γ. Alors, en posant par exemple $\delta_t = 1/t^2$, par Borel–Cantelli on obtient que presque sûrement, à partir d'un certain rang (aléatoire) on a $\sum_k \ell(I_k, y_k) - \min_j L_{j,t} \leqslant \gamma \sqrt{t(\ln N + 2 \ln t)}$ et en particulier cette quantité est $o(t)$ presque sûrement.

3 Introduction de termes de type variance empirique

Dans les arguments ci-dessus, il est très important que la fonction de coût ℓ soit bornée (par 1). Or, si les variations typiques de ℓ sont beaucoup plus petites que la borne utilisée, il est intuitivement clair que les estimations ci-dessus seront assez mauvaises. On aimerait, grossièrement, remplacer le

terme en $\sqrt{t}$ par un terme impliquant la racine d'une variance empirique au temps t, par exemple $\sqrt{\min_j L_{j,t}}$.

Pour cela, on va maintenant se restreindre au cas où les y_t sont fixés à l'avance, indépendamment des prévisions effectuées I_k. C'est le cas, par exemple, en météo mais pas pour les cours de bourse, où les actions passées influencent les cours.

Ces résultats seront intéressants seulement lorsque $\sqrt{\min_j L_{j,t}} \ll t$, autrement dit lorsqu'un des experts (à chaque instant) est suffisamment bon. On a besoin des propositions suivantes.

Proposition 5.
Soient $X_1, \dots, X_n$ des variables aléatoires $\mathcal{F}_t$-adaptées telles que $\operatorname{Var}(X_t \mid \mathcal{F}_{t-1}) \leqslant V_t$ pour un certain $V_t \in \mathbb{R}$ (ne dépendant pas de $\mathcal{F}_{t-1}$). Posons $V = \sum_1^n V_t$. Soit $\delta \in \mathbb{R}$ et supposons que pour tout t on a $X_t - \mathbb{E}[X_t|\mathcal{F}_{t-1}] \leqslant \sqrt{\frac{V}{\ln 1/\delta}}$. Alors

$$\Pr\left(\sum_1^n X_t - \sum_1^n \mathbb{E}[X_t|\mathcal{F}_{t-1}] \geqslant \sqrt{\frac{\ln 1/\delta}{V}}\right) \leqslant \delta$$

(La démonstration est similaire à celles ci-dessus avec les bornes de Chernoff, et la remarque que $e^x \leqslant 1 + x + x^2$ pour $x \leqslant 1$.)

Proposition 6 (Bennett–Bernstein).
Soient $X_1, \dots, X_n$ des variables aléatoires $\mathcal{F}_t$-adaptées telles que $\operatorname{Var}(X_t \mid \mathcal{F}_{t-1}) \leqslant V_t$ pour un certain $V_t \in \mathbb{R}$ (ne dépendant pas de $\mathcal{F}_{t-1}$). Posons $V = \sum_1^n V_t$. Supposons que pour un certain $b \in \mathbb{R}$, pour tout t on a $Y_t \leqslant b$ presque sûrement. Alors

$$\Pr\left(\sum_1^n X_t - \sum_1^n \mathbb{E}[X_t|\mathcal{F}_{t-1}] \geqslant x\right) \leqslant \exp\left(-\frac{x^2}{2(V + bx/3)}\right)$$

Le résultat final obtenu est le suivant.

Proposition 7.
Posons

$$\hat{L}_t = \sum_1^t \ell(I_k, y_k) \quad \text{et} \quad L_t^* = \min_j \sum_1^t \ell(f_{j,k}, y_k)$$

Alors, pour tout $y_1, \dots, y_t, \dots$ fixé à l'avance, il existe un certain choix de η_t tel que si l'on applique la stratégie ci-dessus, avec probabilité supérieure à $1 - \delta$ on a

$$\hat{L}_t - L_t^* \leqslant 2\sqrt{L_t^* \ln N} + 2\sqrt{\ln 1/\delta}\, \max\left(\sqrt{\ln 1/\delta}\,,\, \sqrt{L_t^*} + \sqrt{\ln N}\right)$$