Illustration de la symétrisation

SFDS 2015 Lille

Stéphane Boucheron
LPMA Université Paris-Diderot et DMA Ens Ulm

4 Juin 2015

Symmétrisation avec une illustration

Une technique ancienne

Quand on s'intéresse aux *sommes de variables aléatoires indépendantes*, vectorielles ou non, on est amené à considérer des versions *symétrisées* de ces variables

Définition

Une variable aléatoire X est *symétrique* si X et $-X$ ont même loi.

Si X' a même loi que X et est indépendante de X, $X - X'$ est symétrique.

Observation

Les normes de sommes de vecteurs aléatoires symétriques vérifient les *inégalités de Lévy* : les probabilités de déviation des normes des sommes partielles peuvent être contrôlées par les probabilités de déviation de norme de la somme finale.

Une technique facile et applicable

La symétrisation consiste à remplacer l'étude d'une fonctionnelle d'une processus par l'étude d'une fonctionnelle d'une version symmétrisée de ce processus.

Pour justifier cette technique, il faut montrer que la fonctionnelle du processus symétrisée est liée (en espérance, en distribution) à la fonctionnelle du processus de départ.

Les inégalités de symétrisation permettent de réaliser à peu de frais cet objectif

Un exemple simple

La statistique de Kolmogorov-Smirnov

F_n fonction de répartition empirique d'un n-échantillon d'une loi F supposée diffuse,

Statistique de Kolmogorov-Smirnov

$$D_n = \sqrt{n} \sup_{s \in [0,1]} |F_n(s) - F(s)| \ ,$$

on vérifie que la loi de D_n ne dépend pas de F et qu'on peut sans se restreindre supposer que F est la loi uniforme sur $[0, 1]$.

Le comportement asymptotique de D_n (limite en loi) est connu

La loi limite est celle du supremum du pont Brownien réfléchi

$$\sup_{t \in [0,1]} |B(t) - tB(1)|$$

où $B(\cdot)$ est un mouvement Brownien standard.

Des résultats non-asymptotiques

La formulation du comportement limite de D_n est savante. Les justifications modernes (à partir du principe de Donsker) demandent l'introduction d'outils trop généraux pour un enseignement général.

Avec des outils simples, on peut cependant établir des résultats non-asymptotiques (sous-optimaux) qui donnent une idée qualitativement correcte du comportement de D_n

Ce sont des applications répétées des techniques de symétrisation qui permettent d'établir des majorations des queue de probabilité pour $(D_n)_n$

Les techniques de symétrisation mises en oeuvre à cette occasion sont aussi un départ commode pour la dérivation des inégalités de Vapnik-Chervonenkis.

Inégalité de Dvoretsky-Kiefer-Wolfowitz-Massart

L'objectif de l'exposé n'est pas d'établir :

Corollary 1 in Massart (1990)

Pour tout entier $n > 0$, tout $\epsilon > 0$

$$\mathbb{P}\{D_n \geq \epsilon\} \leq 2\mathrm{e}^{-2\epsilon^2}.$$

Les constantes dans et devant l'exposant ne peuvent être améliorées :

$$2\mathrm{e}^{-2\epsilon^2} = \lim_{n\to\infty} \mathbb{P}\{D_n \geq \epsilon\}$$

La preuve est délicate : elle repose sur une comparaison explicite des temps de passage du pont Brownien empirique et pont Brownien à un niveau donné.

Une version affaiblie

On se contente d'établir un résultat sous-optimal

Plus facile

Pour tout entier $n > 0$, tout $\epsilon > 0$

$$\mathbb{P}\{D_n \geq \epsilon\} \leq 4e^{-\epsilon^2/8}.$$

avec des moyens modestes...

Une preuve (de la version affaiblie)

Le décor (I)

$Y_i \sim_{i.i.d.}$ uniformément sur $[0,1]$.

$$Z = \sqrt{n} D_n = \sup_{s \in [0,1]} \left| \sum_{i=1}^{n} X_{i,s} - \mathbb{E} X_{i,s} \right|$$

où $X_{i,s} = 1$ si $Y_i \leq s$ et $X_{i,s} = 0$ sinon.

Pour $i \leq n$, Y_i' est une copie indépendante de Y_i, et $X'_{i,s} = \mathbb{I}_{Y_i' \leq s}$.

$$Z = \sup_{s \in [0,1]} \left| \sum_{i=1}^{n} X_{i,s} - \mathbb{E} X'_{i,s} \right| .$$

Le décor (II)

- Les ϵ_i sont des variables de Rademacher indépendantes des Y_i, Y_i'.

$$\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$$

- Pour tout ϵ_i

$$\epsilon_i(X_i - X_i') \sim (X_i - X_i')$$

Les espérances sont prises par rapport à $X_1, \ldots, X_n, X_1', \ldots, X_n', \epsilon_1, \ldots, \epsilon_n$.

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right] &= \mathbb{E}\left[\sup_{s\in[0,1]} \mathrm{e}^{\lambda\left|\sum_{i=1}^n X_{i,s}-\mathbb{E}X'_{i,s}\right|}\right] \\
&\leq \mathbb{E}\left[\sup_{s\in[0,1]} \mathrm{e}^{\lambda\left|\sum_{i=1}^n X_{i,s}-X'_{i,s}\right|}\right] && \text{(Inégalité de Jensen)} \\
&= \mathbb{E}\left[\sup_{s\in[0,1]} \mathrm{e}^{\lambda\left|\sum_{i=1}^n \epsilon_i(X_{i,s}-X'_{i,s})\right|}\right] && (X_i - X'_i \text{ est symétrique}) \\
&\leq \mathbb{E}\left[\sup_{s\in[0,1]} \frac{1}{2}\left(\mathrm{e}^{2\lambda\left|\sum_{i=1}^n \epsilon_i X_{i,s}\right|} + \mathrm{e}^{2\lambda\left|-\sum_{i=1}^n \epsilon_i X'_{i,s}\right|}\right)\right] && \text{(Inégalité de Jensen)} \\
&\leq \mathbb{E}\left[\sup_{s\in[0,1]} \mathrm{e}^{2\lambda\left|\sum_{i=1}^n \epsilon_i X_{i,s}\right|}\right] .
\end{aligned}
$$

Presque sûrement les Y_i sont deux à deux distincts

$$\mathbb{E}\left[\sup_{s\in[0,1]} \mathrm{e}^{2\lambda\left|\sum_{i=1}^{n}\epsilon_i X_{i,s}\right|} \mid Y_1,\ldots,Y_n\right] = \mathbb{E}\left[\max_{k\le n} \mathrm{e}^{2\lambda\left|\sum_{i\le k}\epsilon_i\right|}\right]. \tag{1}$$

Le membre droit est un moment exponentiel de la marche aléatoire symétrique réfléchie

Pour toute variable aléatoire positive W,

$$\mathbb{E}W = \int_0^t \mathbb{P}\{W > t\}\mathrm{d}t$$

Soit

$$\mathbb{E}\left[\max_{k\le n} \mathrm{e}^{2\lambda\left|\sum_{i\le k}\epsilon_i\right|}\right] = \int_0^\infty \mathbb{P}\left\{\max_{k\le n} \mathrm{e}^{2\lambda\left|\sum_{i\le k}\epsilon_i\right|} \ge a\right\}\mathrm{d}a$$

Principe de réflexion

Pour $k \leq n$

$$A_k = \left\{ \exp\left(2\lambda \left|\sum_{i\leq k} \epsilon_i\right|\right) \geq a, \text{ et } \exp\left(2\lambda \left|\sum_{i\leq j} \epsilon_i\right|\right) < a \text{ pour } j < k \right\}$$

Décomposition en événements disjoints

$$\left\{ \max_{k\leq n} \mathrm{e}^{2\lambda\left|\sum_{i\leq k} \epsilon_i\right|} \geq a \right\} = \cup_{k\leq n} A_k$$

Pour $k \leq n$

$$\mathbb{P}\left\{ \mathrm{e}^{2\lambda\left|\sum_{i\leq n} \epsilon_i\right|} \geq a \mid A_k \right\} \quad \geq \quad \frac{1}{2},$$

Exploitation

$$\begin{aligned}
\mathbb{P}\left\{\max_{k\le n} \mathrm{e}^{2\lambda\left|\sum_{i\le k}\epsilon_i\right|} \ge a\right\} &= \sum_{k\le n}\mathbb{P}\{A_k\}\mathbb{P}\\
&\le 2\sum_{k\le n}\mathbb{P}\{A_k\}\mathbb{P}\left\{\mathrm{e}^{2\lambda\left|\sum_{i\le n}\epsilon_i\right|} \ge a \mid A_k\right\}\\
&= 2\mathbb{P}\left\{\mathrm{e}^{2\lambda\left|\sum_{i\le n}\epsilon_i\right|} \ge a\right\}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\max_{k\le n} \mathrm{e}^{2\lambda\left|\sum_{i\le k}\epsilon_i\right|}\right] &\le 2\mathbb{E}\left[\mathrm{e}^{2\lambda\left|\sum_{i\le n}\epsilon_i\right|}\right]\\
&\le 4\mathbb{E}\left[\mathrm{e}^{2\lambda\sum_{i\le n}\epsilon_i}\right] && \text{(Inégalité triangulaire)}\\
&\le 4\mathrm{e}^{2n\lambda^2} && \text{(Lemme de Hoeffding).}
\end{aligned}$$

On aboutit à

$$\mathbb{E}\left[e^{\lambda Z}\right] \leq 4e^{2n\lambda^2}$$

L'inégalité recherchée s'obtient en choisissant $\lambda = \epsilon/\sqrt{n}$ et invoquant le Lemme de Markov.

Et après ?

Inégalité de symmétrisation générale

La symmétrisation fonctionne pour les processus empiriques en général.

$$\frac{1}{2}\mathbb{E}\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\epsilon_i X_{i,s}\right| \leq \mathbb{E}\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}(X_{i,s}-\mathbb{E}X_{i,s})\right| \leq 2\mathbb{E}\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\epsilon_i X_{i,s}\right|$$

Inégalités Vapnik-Chervonenkis

Dans la seconde partie de la preuve, nous avons profité de la forme très spéciale du processus empirique.

Si ...

$\mathcal{T}$ désigne une collection de parties d'un univers probabilisé.
Si $X_{i,s} = \mathbb{I}_{Y_i \in s}$ où $s \in \mathcal{T}$, et
si $\mathrm{Tr}(Y_1, \ldots, Y_n)$ désigne la trace de $\mathcal{T}$ dans $(Y_1, \ldots, Y_n)$, soit l'ensemble des parties de $(Y_1, \ldots, Y_n)$ qui s'écrivent $\{Y_1, \ldots, Y_n\} \cap s$ avec $s \in \mathcal{T}$,

$$\begin{aligned}
\mathbb{E}\left[\sup_{s\in\mathcal{T}} \mathrm{e}^{2\lambda\left|\sum_{i=1}^n \epsilon_i X_{i,s}\right|} \mid Y_1, \ldots, Y_n\right] &\leq \mathbb{E}\left|\mathrm{Tr}(Y_1, \ldots, Y_n)\right| \mathbb{E}\left[\mathrm{e}^{2\lambda\left|\sum_{i\leq n} \epsilon_i\right|}\right] \\
&\leq \mathbb{E}\left|\mathrm{Tr}(Y_1, \ldots, Y_n)\right| \mathrm{e}^{2n\lambda^2}
\end{aligned}$$