Symétrisation, une illustration

Stéphane Boucheron [1]

[1] *LPMA Université Paris-Diderot & DMA ENs ULM*
stephane.boucheron at univ-paris-diderot.fr

Résumé.
La symétrisation est une technique très 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 (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). Les normes de sommes de vecteurs aléatoires symétriques vérifient en effet 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. Ces inégalités donnent des critères simples de convergence pour les séries aléatoires (voir par exemple, Ledoux et Talagrand (1991), Chapitre 2).

En statistiques (processus empiriques), en théorie de l'apprentissage, la symétrisation apparaît naturellement dans la démonstration des inégalités de Vapnik et Chervonenkis. Le classique de Van de Vaart et Wellner (1996) y consacre un chapître. Elle permet de réduire l'étude de suprema de processus empiriques à l'analyse de questions combinatoires. On peut illustrer la puissance et la simplicité de la technique sur la statistique de Kolmogorov-Smirnov : $D_n = \sqrt{n} \sup_x |F_n(x) - F(x)|$ et montrer sans efforts

$$\mathbb{P}\left\{D_n \geq \epsilon\right\} \leq 4\mathrm{e}^{-\frac{n\epsilon^2}{8}},$$

en perdant donc un facteur 2 devant l'exponentielle, et un facteur 16 dans l'exposant, par rapport à la borne optimale (et délicate) de Dvoretsky-Kieffer-Wolfovitz-Massart.

Les inégalités de symétrisation sont aussi la justification de plusieurs mesures de complexité empiriques utilisées en sélection de modèles comme les moyennes de Rademacher (Koltchinskii, Annals of Statistics, 2006). Elles permettent même de développer des inégalités de type Bernstein auto-normalisées pour les suprema de processus empiriques (Panchenko, 2003).

Mots-clés. Symétrisation

Abstract. Symmetrization is an old technique: when concerned with sums of independent random vectors, considering symmetrized versions of those random vectors is fruitful (a random variable X is symmetric if X and $-X$ have the same distribution. If X' is distributed like X and independent from X, $X - X'$ is symmetric). As norms of sums of independent symmetric random vectors satisfy Lévy inequalities, tail bounds for

norms of partial sums may be compared with tail bounds for the norm of the last sum. Those inequalities provide us with simple criteria for the convergence of random series (see for example, Ledoux and Talagrand (1991), Chapter 2).

In Statistics (empirical processes), in Learning Theory, symmetrization is a natural ingredient in the proof of the Vapnik-Chervonenkis inequalities. van der Vaart and Wellner [1996] dedicate a whole chapter to this topic. Thanks to symmetrization, investigating suprema of empirical processes boils down to combinatorial questions. This can be illustrated while studying the Kolmogorov-Smirnov statistics: $D_n = \sqrt{n} \sup_x |F_n(x) - F(x)|$, showing in a painless way

$$\mathbb{P}\{D_n \geq \epsilon\} \leq 4\mathrm{e}^{-\frac{n\epsilon^2}{8}},$$

loosing a factor of 2 in front of the exponential, and a factor 16 inside the exponent, with respect to the difficult optimal bound due to Dvoretsky-Kieffer-Wolfovitz-Massart.

Symmetrization inequalities are at the core of empirical complexity indices used in model selection such as Rademacher averages [Koltchinskii, 2008, 2006]. They are also at the root of some Bernstein inequalities for self-normalized suprema of empirical processes [Panchenko, 2003].

Keywords. Symmetrization

1 Une technique ancienne

La symétrisation, consiste (entre autres choses) à réduire l'étude d'une suite de variables aléatoires vectorielles $(X_i)_i$ à celle d'une suite de variables aléatoires $(X'_i)_i$ symétriques. Une variable aléatoire X est dite symétrique si X et $-X$ ont même loi. Ledoux and Talagrand [1991] utilisent ce type de techniques pour caractériser la convergence presque sûre de sommes de vecteurs aléatoires indépendants. Ils renvoient aux travaux de Paul Lévy. En statistique, la symétrisation est utilisée sans être toujours mentionnée. Les moyennes de Rademacher, conditionnelles ou non, sont une variante du bootstrap à poids qui repose sur la symétrisation, Les tests de permutation peuvent aussi être considérés comme des techniques de symétrisation. Ici, nous voulons illustrer la symétrisation comme technique de preuve. Dans la section suivante, la symétrisation et quelques arguments élémentaires permettent d'étudier le comportement de la statistique de Kolmogorov-Smirnov sous l'hypothèse nulle sans invoquer d'arguments asymptotiques (érudits), ni de raisonnements trop difficiles.

2 Une borne simple pour la statistique de Kolmogorov-Smirnov

Si F_n désigne la fonction de répartition empirique d'un n-échantillon d'une loi F supposée diffuse, la statistique de Kolmogorov-Smirnov s'écrit

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

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 $[0,1]$.

L'inégalité de Dvoretzky, Kiefer, and Wolfowitz [1956], Massart [1990] s'énonce ainsi : pour tout entier $n > 0$, tout $\epsilon > 0$

$$\mathbb{P}\left\{D_n \geq \epsilon\right\} \leq 2\mathrm{e}^{-2\epsilon^2} \ . \tag{1}$$

Ce résultat difficile ne peut être amélioré. Une version affaiblie, peut être enseignée :

$$\mathbb{P}\left\{D_n \geq \epsilon\right\} \leq 4\mathrm{e}^{-\frac{\epsilon^2}{8}} \ . \tag{2}$$

C'est l'occasion d'introduire la symétrisation, d'explorer les premières étapes de la preuve des inégalités de Vapnik-Chervonenkis, et d'appliquer en statistique des outils élégants issus de la théorie des marches aléatoires.

Preuve Dans la suite les Y_i sont i.i.d. uniformément sur $[0,1]$. On introduit par souci de commodité

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

où $X_{i,s} = 1$ si $Y_i \leq s$ et $X_{i,s} = 0$ sinon. Dans la suite pour $i \leq n$, Y_i' est une copie indépendante de Y_i, et $X'_{i,s} = \mathbb{I}_{Y_i' \leq s}$. On peut réécrire Z en

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

Les ϵ_i sont des variables de Rademacher ($\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$) indépendantes des Y_i, Y_i'. On utilisera le fait que $\epsilon_i(X_i - X_i') \sim (X_i - X_i')$. Dans la suite 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, donc

$$\mathbb{E}\left[\sup_{s\in[0,1]} \mathrm{e}^{2\lambda\left|\sum_{i=1}^n \epsilon_i X_{i,s}\right|} \mid Y_1, \ldots, Y_n\right] = \mathbb{E}\left[\max_{k\leq n} \mathrm{e}^{2\lambda\left|\sum_{i\leq k} \epsilon_i\right|}\right]. \tag{3}$$

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

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

On note $A_k = \left\{\left|\sum_{i\leq k} \epsilon_i\right| \geq a, \text{ et } \left|\sum_{i\leq j} \epsilon_i\right| < a \text{ pour } j < k\right\}$.

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

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

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

D'où

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

On aboutit à $\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right] \leq 4\mathrm{e}^{2n\lambda^2}$ et l'inégalité (2) s'obtient en choisissant $\lambda = \epsilon/\sqrt{n}$ et invoquant le Lemme de Markov. □

3 Quelques développements

Si au lieu de considérer un processus empirique indexé par les demi-droites de $\mathbb{R}$, on considère des processus indexés par des classes plus générales de parties, on ne peut plus utiliser une identité comme (3). On peut cependant, exploiter la symétrisation en adoptant la démarche de Vapnik et Chervonenkis [Massart, 2006] .

Cette démarche permet même d'obtenir des résultats fins et applicables comme cette inégalité issue de Panchenko [2003] qui s'affranchit des contraintes habituelles de bornitude.

Soient $X_1', \ldots, X_n'$ des copies i.i.d. des vecteurs indépendants $X_1, \ldots, X_n$. Soit

$$W = \mathbb{E}\left[\sup_{s\in\mathcal{T}} \sum_{i=1}^{n} (X_{i,s} - X_{i,s}')^2 \middle| X_1, \ldots, X_n\right] .$$

Alors pour tout l $t \geq 0$,

$$\mathbb{P}\left\{Z \geq \mathbb{E}Z + 2\sqrt{tW}\right\} \leq 4e^{-t/4}$$

et

$$\mathbb{P}\left\{Z \leq \mathbb{E}Z - 2\sqrt{tW}\right\} \leq 4e^{-t/4} .$$

References

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 33:642–669, 1956.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 36:00–00, 2006.

V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems. Lectures from the 38th Probability Summer School, Saint-Flour*, volume 2033 of *Lecture Notes in Mathematics*. Springer, 2008.

M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18:1269–1283, 1990.

P. Massart. *Concentration inequalities and model selection.* Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics. Springer, 2006.

D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *The Annals of Probability*, 31:2068–2081, 2003.

A. van der Vaart and J. Wellner. *Weak convergence and empirical processes.* Springer-Verlag, New York, 1996.